Adversarial Machine Learning in Generative Models

Jiyao Li

Doctor of Philosophy

University of Technology Sydney

School of Computer Science

Australia

May 2025

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Jiyao Li* declare that this thesis is submitted in fulfillment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science* at the *University of Technology Sydney*. This thesis is wholly my work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note: Signature removed prior to publication.

Jiyao Li

DATE: 11th May, 2025

Abstract

Generative artificial intelligence (AI) models, such as natural language processing (NLP) and computer vision (CV) models, have demonstrated remarkable progress in recent years, enabling significant advancements in applications such as question answering, summarization, and image captioning. However, recent studies highlight the vulnerability of these models to adversarial attacks, raising critical concerns about AI system reliability and real-world deployment challenges. This thesis investigates the vulnerabilities of three widely used models: question answering, summarization, and image captioning models, and it introduces innovative techniques for generating highly effective yet imperceptible adversarial examples. The research presents three novel attack methodologies. First, the Paraphrasing-Based Summarization Attack (SAP) addresses abstractive summarization models by ranking sentences based on their importance to the summarization outcome. This approach employs sophisticated paraphrasing mechanisms to craft adversarial examples that preserve semantic coherence while inducing incorrect summaries. Second, AICAttack (Attention-Based Image Captioning Attack) introduces a novel black-box adversarial attack for image captioning models. The method leverages an attention-based mechanism to identify critical image regions. It employs customized differential evolution algorithms to optimize pixel perturbations, achieving highly effective adversarial captions while maintaining minimal visual perturbation. Finally, the QA-Attack method presents a comprehensive approach to crafting practical adversarial examples for question answering models, addressing both boolean and informative queries. The method adopts a Hybrid Ranking Fusion algorithm that combines attention-based and removal-based ranking mechanisms, demonstrating high success rates across diverse question answering systems. This thesis advances the field of AI security through these three innovative attack methodologies, each demonstrating superior effectiveness while maintaining imperceptibility and revealing critical vulnerabilities in today's generative AI landscape. These innovative approaches not only expose fundamental limitations across diverse AI applications but also establish a crucial foundation for developing more robust and interpretable next-generation systems, ultimately fostering greater trust in AI technologies increasingly deployed in sensitive real-world contexts.

Acknowledgments

The completion of this thesis represents a milestone that would not have been possible without the support of many individuals. I owe my deepest gratitude to my supervisor, Associate Professor Wei Liu, and my co-supervisor Associate Professor Jianlong Zhou, whose exceptional mentorship has been the cornerstone of my academic journey. His deep expertise in adversarial learning, coupled with his thoughtful guidance and unwavering support, has profoundly shaped not only this thesis but also my evolution as a researcher. His insightful feedback, innovative perspectives, and commitment to excellence have consistently inspired me to explore new frontiers in my research endeavors.

I am deeply indebted to my parents, whose unconditional love and steadfast support have been my anchor throughout this journey. Their faith in my abilities and constant encouragement have provided me with the strength to persevere through challenges and celebrate achievements.

My heartfelt appreciation extends to my friends, whose invaluable camaraderie has enriched this academic pursuit. Their understanding, encouragement, and timely words of wisdom have been instrumental in helping me navigate the complexities of doctoral research while maintaining balance and determination. Their presence has made this journey not only more manageable but also more meaningful.

Finally, I dedicate this thesis to everyone who has contributed to my academic journey. Your collective support, guidance, and encouragement have left an indelible impact on both this work and my personal growth. This achievement stands as a testament to the power of collaborative spirit and shared dedication to academic excellence.

List of Publications

1. The paper titled *"Summarization Attack via Paraphrasing (Student Abstract)"* has been published in the **37th Proceedings of the AAAI Conference on Artificial Intelligence**.

2. The paper titled *"AICAttack: Adversarial Image Captioning Attack with Attention-Based Optimization"* has been accepted for publication and is awaiting release in the **Machine Intelligence Research** journal.

3. The paper titled *"Deceiving Question-Answering Models: A Hybrid Word-Level Adversarial Approach"* has been submitted to the journal **Neural Networks** and is currently under revision.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

In recent years, deep learning models have achieved remarkable success across a variety of generative tasks in Computer Vision (CV) and Natural Language Processing (NLP), such as abstractive summarization, image captioning, and question answering. These tasks require models to classify or retrieve information and generate coherent and contextually appropriate outputs. Transformer-based architectures, which often combine self-attention mechanisms and encoder-decoder frameworks, have replaced earlier Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) as the dominant approach, producing state-of-the-art results across these domains. Despite these advances, deep learning models remain vulnerable to adversarial attacks, which introduce subtle perturbations to the input data. Such perturbations can lead to misleading or incorrect outputs while remaining imperceptible to human observers, posing significant risks to the reliability and robustness of AI systems deployed in real-world applications [2].

1.1 Background

As AI systems spread throughout language and image processing, growing concerns about how well they hold up against attacks and unexpected situations have become a key focus for the field. While state-of-the-art neural architectures, such as transformers and convolutional neural networks (CNNs), have achieved groundbreaking performance across a wide range of tasks, they remain susceptible to adversarial attacks-carefully crafted perturbations designed to deceive models while remaining imperceptible to humans [2]. These vulnerabilities pose significant risks, particularly as AI systems are integrated into critical real-world applications such as healthcare, autonomous driving, and financial decision-making. Understanding and addressing these weaknesses is essential for building trustworthy and reliable AI systems.

Adversarial attacks are generally categorized as white-box or black-box. White-box attacks assume complete knowledge of the target model, including its architecture, parameters, and gradients, enabling attackers to craft precise perturbations that maximize the likelihood of misclassification [9]. In contrast, black-box attacks assume a more practical setting where the attacker lacks access to the model's internal mechanics. Instead, they rely on query-based strategies or heuristic methods to infer vulnerabilities [10]. Both attacks present unique challenges in NLP and CV due to the discrete nature of textual inputs and the complex, generative behavior of modern deep learning models.

In NLP, adversarial attacks have been extensively studied in classification tasks, where small modifications, such as synonym substitutions, character-level changes, or paraphrasing, can drastically alter model predictions [11]. For example, replacing the word "excellent" with "outstanding" in a product review might cause a sentiment analysis model to misclassify a positive review as negative. However, attacking generative models, such as those used for abstractive summarization or question answering, presents additional challenges. In these tasks, adversarial perturbations must not only alter the

output but also ensure that the generated text remains fluent, coherent, and semantically consistent. This complexity makes generative models particularly difficult to attack effectively.

Similarly, adversarial attacks on image classification models have been widely explored in CV. For instance, adding imperceptible noise to an image of a stop sign can cause a model to misclassify it as a yield sign, with potentially catastrophic consequences in autonomous driving systems. However, the adversarial vulnerability of image captioning systems, which generate textual descriptions of images, remains a relatively underexplored area. These systems rely on both visual and linguistic components, making them susceptible to attacks that manipulate either modality or both simultaneously.

This thesis investigates adversarial attacks on three critical tasks at the intersection of NLP and CV: abstractive summarization, image captioning, and question answering (QA). Unlike traditional classification tasks, these involve complex sequence generation, where adversarial perturbations must not only mislead the model but also preserve the naturalness and coherence of the output. Existing adversarial methods for these tasks often struggle to balance attack effectiveness with semantic integrity, highlighting the need for more advanced and targeted strategies.

To address these challenges, this research introduces novel adversarial attack methodologies tailored to each domain:

- **Summarization Attack via Paraphrasing (SAP)**: A paraphrasing-based adversarial attack to deceive abstractive summarization models by subtly rewriting key sentences in the input text. SAP ensures that the generated summary is significantly altered while maintaining fluency and coherence, making it a stealthy and effective attack strategy.

- **Attention-based Image Captioning Attack (AICAttack)**: A black-box adversarial attack targeting image captioning models by perturbing critical regions of

the input image based on attention mechanisms. By focusing on areas that the model relies on most heavily, AICAttack can mislead the captioning process without significantly altering the image's overall appearance.

- **QA-Attack**: A hybrid ranking-based adversarial attack that exploits attention mechanisms to mislead question answering models. By identifying and perturbing high-impact tokens in the input text, QA-Attack effectively fools QA models while preserving the semantic integrity of the text.

1.1.1 Attacks for Summarization Models

Abstractive summarization has emerged as a pivotal task in natural language processing (NLP), addressing the increasing demand for condensing extensive textual information into concise, informative summaries [12]. The field has witnessed remarkable progress with the advent of deep learning techniques, particularly in abstractive summarization, which generates summaries using novel phrases and sentences rather than merely extracting existing text [13]. Modern abstractive summarization models predominantly employ encoder-decoder architectures, often based on transformer networks, which excel at capturing long-range dependencies and producing coherent summaries [14]. Previous advanced models such as BART [14], Pegasus [15], and T5 [16] have demonstrated exceptional performance on benchmark datasets, showcasing their ability to generate summaries that closely resemble human-written content [15].

Despite these advancements, abstractive summarization models grapple with several persistent challenges. These include ensuring factual consistency, mitigating hallucinations (i.e., generating content not present in the source text), and effectively processing long documents [17]. Additionally, evaluating the quality of generated summaries remains a complex endeavor, as traditional metrics like ROUGE often fail to capture nuanced aspects such as coherence, relevance, and fluency [18]. While these issues have

been extensively studied, a critical yet underexplored area is the vulnerability of abstractive summarization models to adversarial attacks. Such attacks can subtly manipulate input text to induce significant deviations in the generated summaries, posing a threat to the reliability of these systems [19].

Existing adversarial attack methods for summarization typically rely on word-level perturbations or sentence-level manipulations [6]. However, these approaches often struggle to maintain semantic coherence or produce substantial changes in the generated summaries, particularly in transformer-based models that leverage broader contextual understanding [20]. Given the generative nature of abstractive summarization, simple word substitutions may not suffice to meaningfully alter model outputs, necessitating more sophisticated and targeted attack strategies.

In this thesis, we address this research gap by introducing Summarization Attack via Paraphrasing (SAP), a novel adversarial attack framework specifically designed for abstractive summarization models. SAP employs a paraphrasing-based approach to identify and strategically modify key sentences in the input text, ensuring that the generated summary is significantly altered while preserving fluency and semantic coherence. Unlike traditional adversarial attacks, which often compromise text readability or rely on unnatural modifications, SAP replaces high-impact sentences with semantically equivalent paraphrases. This approach effectively misleads the summarization model without introducing detectable distortions, making it a more subtle and potent attack strategy.

SAP is designed to overcome the unique challenges posed by abstractive summarization models, offering a more effective and targeted method compared to existing techniques. By leveraging advanced paraphrasing mechanisms, SAP highlights critical vulnerabilities in state-of-the-art summarization systems and underscores the need for enhanced robustness in NLP applications. This work not only advances the understand-

ing of adversarial threats in abstractive summarization but also provides a foundation for developing more secure and reliable summarization systems capable of maintaining high performance even in adversarial settings. Ultimately, this research aims to contribute to the creation of NLP systems that are both accurate and resilient, ensuring their trustworthiness in real-world applications.

1.1.2 Attacks for Image Captioning Models

Adversarial attacks on image captioning models present unique challenges due to the inherent complexity of generating coherent and semantically meaningful textual descriptions from visual inputs. Unlike traditional image classification tasks, where the model outputs a single label [21], image captioning models must generate a sequence of words that accurately describe the content of an image [22]. This sequential nature of the output makes it significantly more challenging to craft effective adversarial examples, as the attacker must manipulate the input in a way that disrupts the entire generated caption rather than just a single classification decision. The encoder-decoder architecture commonly used in image captioning models further complicates the attack process. In such frameworks, the encoder processes the input image to extract visual features, while the decoder generates the corresponding caption based on these features [23]. The gradients necessary for optimizing adversarial perturbations in this setting are often difficult to compute or unavailable, particularly when attackers face black-box scenarios where can not obtain models' internal parameters or architecture [24]. Additionally, the textual output of captioning models must be evaluated not only for its semantic relevance to the image but also for its syntactic correctness, adding another layer of complexity to the attack evaluation process [25].

Most existing research on adversarial attacks for image captioning models has focused on white-box scenarios. For example, the "Show-and-Fool" attack [22] leverages gradient

information to craft perturbations that influence the generated captions. While effective in controlled settings, this approach is impractical in real-world applications where such detailed knowledge of the target model is rarely available. Another line of research has explored black-box attacks using generative adversarial networks (GANs) to generate perturbations [25]. However, these methods often suffer from instability during training and are not specifically tailored for image captioning tasks, as they fail to account for the precise pixel-level manipulations required to effectively fool captioning models [23].

We introduce AICAttack (Attention-based Image Captioning Attack), a novel adversarial method designed to target image captioning models and address existing challenges. Our approach operates in a black-box setting, meaning it does not require access to the target model's architecture, parameters, or gradients. Instead, we leverage an attention-based mechanism to identify the most critical pixels in an image for adversarial manipulation. By focusing on these high-attention areas-regions that captioning models inherently prioritize during caption generation-we maximize the impact of the perturbations while minimizing their visibility to human observers. Additionally, we introduce a differential evolution algorithm to optimize the perturbation values applied to the selected pixels, ensuring that the generated adversarial examples are both effective and imperceptible. This combination of attention-guided targeting and evolutionary optimization allows AICAttack to circumvent the limitations of gradient-dependent white-box methods and unstable GAN-based approaches.

1.1.3 Attacks for Question Answering Models

Deep learning breakthroughs have substantially enhanced question answering (QA) models' capabilities, allowing them to interpret contexts and questions with remarkable precision [26, 27]. State-of-the-art models, including BERT-based architectures and transformer variants, have demonstrated strong results across fact-based retrieval

and reasoning tasks [28, 29]. These models leverage large-scale pre-training on diverse corpora and fine-tuning on specific QA datasets, allowing them to capture intricate linguistic patterns and contextual relationships. However, despite these achievements, QA models remain vulnerable to adversarial attacks, which manipulate input text to mislead predictions while maintaining linguistic coherence [30, 31]. Such weaknesses cast doubt on QA systems' dependability in practical settings, where deliberately manipulated inputs might trigger inaccurate or potentially dangerous responses.

Existing adversarial attacks on QA models primarily fall into three categories: token-level perturbations, sentence insertion, and contextual misalignment. Token-level attacks, such as HotFlip [32] and TextFooler [11], modify critical words through synonym replacement or character-level changes. These methods aim to alter the semantic meaning of key tokens in the input text, causing the model to produce incorrect answers. Sentence insertion attacks, including AddSent [30], introduce misleading content into the context or question to confuse extractive QA models. These attacks exploit the model's reliance on specific patterns or keywords in the input. Contextual misalignment approaches, on the other hand, manipulate answer positions or alter the structure of the context to cause incorrect extractions [33]. While these methods have proven effective in degrading model performance, most existing attacks assume white-box access to model gradients, limiting their applicability in real-world black-box settings where the internal parameters of the model are inaccessible.

To address this gap, this thesis introduces QA-Attack, a novel black-box adversarial strategy targeting QA models. Our approach leverages a Hybrid Ranking Fusion (HRF) algorithm that combines attention-based and removal-based ranking techniques to identify and perturb critical words. The attention-based ranking identifies tokens that significantly influence the model's predictions by analyzing the attention weights assigned to each token. The removal-based ranking, on the other hand, evaluates the

impact of removing individual tokens on the model's output. By fusing these two rankings, HRF effectively identifies the most vulnerable tokens in the input text. Once these tokens are identified, QA-Attack replaces them with contextually appropriate synonyms generated through a masked language model, ensuring that the perturbed text remains linguistically coherent and semantically plausible. This approach allows QA-Attack to effectively fool QA models without requiring access to their internal parameters, making it applicable in real-world black-box scenarios.

1.2 Aims and Objectives

The research of my PhD candidature focuses on developing and evaluating novel adversarial attack methods for three key areas of natural language processing and computer vision: abstractive summarization, image captioning, and question-answering systems. The overarching goal is to identify vulnerabilities in state-of-the-art models and contribute to improving their robustness and reliability. The aims are as follows:

1. To develop and evaluate novel adversarial attack methods for abstractive summarization models, identifying specific vulnerabilities and providing insights into improving their robustness against adversarial attacks.

2. To design and implement AICAttack, a new adversarial attack strategy for image captioning models, focusing on identifying critical pixels influencing caption generation while ensuring imperceptibility to human observers.

3. To create and assess a question-answering adversarial attack approach that combines hybrid ranking techniques to identify vulnerable tokens, particularly under black-box scenarios.

4. To analyze the experimental results from the proposed attacks across diverse datasets and leading models, contributing to the understanding of vulnerabilities and providing recommendations for enhancing the security and reliability of these technologies.

The objectives are as follows:

1. Develop and evaluate a novel adversarial attack method for abstractive summarization models. This research will focus on creating attacks with high success rates while minimally altering the input text. We will establish a comprehensive evaluation metric to measure attack effectiveness, implement our proposed method, and

test it against leading abstractive summarization models. Through extensive experiments on diverse datasets, we aim to demonstrate the attack's broad applicability and robustness. By analyzing the results, we will uncover specific weaknesses in current summarization models, especially in challenging scenarios. These insights will guide recommendations for improving the resilience of summarization systems against adversarial attacks, ultimately advancing our understanding of how these models function and how they can be made more robust.

2. Design and evaluate AICAttack, a novel adversarial attack method for image captioning models. This research will focus on developing an attention-based mechanism to precisely identify and manipulate critical pixels that influence caption generation while ensuring the alterations remain imperceptible to human observers. We will implement a customized differential evolution algorithm to optimize these adversarial modifications, aiming for high attack success rates with minimal image perturbation. The study will assess AICAttack's performance across various state-of-the-art image captioning models and diverse real-world datasets, demonstrating its effectiveness and generalizability. By analyzing the results, we will gain insights into the vulnerabilities of current image captioning systems, particularly in challenging scenarios. These findings will contribute to our understanding of multimodal learning models and guide future improvements in their robustness against adversarial attacks, ultimately enhancing the security and reliability of image captioning technologies.

3. Develop and evaluate a question-answering (QA) attack, an innovative attack method for question-answering models. This research will focus on creating a Hybrid Ranking Fusion algorithm that combines Attention-based Ranking and Removal-based Ranking to identify vulnerable tokens in input texts. We will design the attack to operate in black-box scenarios, requiring no access to the target

model's architecture or parameters, and make it adaptable to various question types, including informative and boolean queries. The study will implement this method and assess its effectiveness across multiple datasets and victim models, comparing its performance against existing adversarial techniques. Through comprehensive analysis of the results, we aim to uncover specific weaknesses in current question-answering systems, especially in challenging cases. These insights will deepen our understanding of the vulnerabilities in QA models and guide future improvements in their robustness, ultimately contributing to the development of more secure and reliable question-answering technologies.

1.3 Research Significance

Our research holds significant importance in the rapidly evolving fields of natural language processing and computer vision. By developing novel adversarial attack methods for abstractive summarization, image captioning, and question-answering systems, this study addresses critical gaps in our understanding of the vulnerabilities and limitations of state-of-the-art models.

The significance of this research is multifaceted:

- Enhancing AI Security: By exposing weaknesses in current models, this research contributes directly to improving the security and reliability of AI systems in real-world applications. This is crucial as these technologies become increasingly integrated into various sectors, including healthcare, finance, and information systems.

- Advancing Model Robustness: The insights gained from these adversarial attacks will guide the development of more robust models, leading to AI systems that are

less susceptible to manipulation and more trustworthy in critical decision-making processes.

- Cross-Domain Insights: By addressing three distinct but related areas (summarization, image captioning, and question-answering), this research offers the potential for cross-pollination of ideas and techniques across different domains of AI.

- Ethical AI Development: Understanding and mitigating vulnerabilities in AI systems is crucial for ensuring their ethical deployment, particularly in sensitive applications where misinformation or misinterpretation could have serious consequences.

- Stimulating Future Research: The methodologies and findings from this study will likely spark new research directions in adversarial machine learning, model interpretability, and robust AI design.

1.4 Research Contribution

This thesis presents the following key research contributions:

1. Summarization Attack (SAP): This attack targets abstractive summarization models by identifying critical sentences within input articles. Using a ranking mechanism based on their impact on generated summaries, measured by ROUGE score differences, the vital sentences are replaced with semantically equivalent paraphrases. This preserves the original meaning while deceiving the summarization model. The contribution of this approach, as presented in the paper, lies in its innovative ranking strategy and paraphrasing technique, which outperform baseline methods on real-world datasets, achieving the highest success rate and minimal semantic drift.

2. Attention-based Image Captioning Attack (AICAttack): AICAttack introduces a novel black-box adversarial framework for attacking image captioning models. It employs an attention-based mechanism to identify critical image regions, enabling selective perturbation of influential pixels. Through a differential evolution optimization process, the framework determines optimal perturbations that maximize impact on caption generation while maintaining visual imperceptibility. The paper's significant contribution includes leveraging attention mechanisms to identify susceptible regions and customizing the differential evolution algorithm, achieving superior success rates against state-of-the-art models across diverse datasets.

3. Question Answering Attack (QA-Attack): QA-Attack is a word-level adversarial strategy designed for deceiving QA models. It introduces the Hybrid Ranking Fusion (HRF) algorithm, combining Attention-Based Ranking (ABR) and Removal-Based Ranking (RBR) to identify critical tokens within the context. These tokens are replaced with contextually appropriate synonyms generated by a masked language model. The method ensures adversarial examples maintain semantic coherence and grammatical correctness while effectively fooling QA models. The study contributes to adversarial research by presenting a unified attack strategy applicable to both Informative and Boolean queries, with extensive experiments demonstrating its robustness and effectiveness across multiple QA models and datasets.

1.5 Thesis Organization

This thesis is organized as follows:

- Chapter 2: Literature Review

  This chapter presents a comprehensive survey of adversarial learning in both natural language processing (NLP) and computer vision (CV).

- Chapter 3: Summarization Attack via Paraphrasing

  This chapter introduces a summarization attack with a paraphrasing technique (SAP) and presents experimental results.

- Chapter 4: AICAttack: Adversarial Image Captioning Attack with Attention-Based Optimization

  This chapter introduces an adversarial image captioning attack with attention-based optimization (AICAttack) and experimental analysis.

- Chapter 5: Deceiving Question-Answering Models: A Hybrid Word-Level Adversarial Approach

  This chapter introduces a hybrid word-level adversarial approach for deceiving textual question answering models (QA-Attack) and experimental analysis.

- Chapter 6: Conclusion and Future Works

  This chapter concludes the thesis and presents directions for future work.

CHAPTER 2

LITERATURE REVIEW

Adversarial attacks have emerged as a significant vulnerability in machine learning systems, exploiting inherent weaknesses in models to deceive and mislead them. As machine learning is increasingly deployed in sensitive domains, understanding and mitigating these vulnerabilities has become a crucial challenge. This chapter provides a comprehensive overview of adversarial attacks, with a focus on their implications in two key areas: Natural Language Processing (NLP) and Computer Vision (CV). By examining some of the most popular tasks in these domains, we delve into the fundamentals of adversarial attacks and their specialized applications, discussing both traditional and state-of-the-art attack methodologies along with their strengths and limitations.

In this chapter, we organize the review by: the overview of adversarial attack (Section 2.1), adversarial attack for NLP (Section 2.2), adversarial attack for CV (Section 2.3), discussion and challenges (Section 2.4).

2.1 Overview of Adversarial Attack

In general, adversarial attacks aim to exploit vulnerabilities in systems, algorithms, or processes. by dividing it into different mechanisms, adversarial attacks present with perturbation-based attacks, poisoning-based attacks, and other real-world attacks [34].

Perturbation-based attacks represent a significant vulnerability in machine learning systems, operating through subtle modifications to input data that can dramatically affect model predictions while remaining virtually imperceptible to human observers [35]. These attacks are particularly prominent in computer vision, where minimal pixel-level alterations can successfully mislead classifiers, and in Natural Language Processing (NLP), where subtle text modifications such as strategic synonym substitutions can fundamentally alter model outputs. The effectiveness of these attacks stems from exploiting models' inherent sensitivity to input variations, revealing critical weaknesses in systems that rely on learned decision boundaries. The practical significance and cross-model transferability of perturbation-based techniques, including the Fast Gradient Sign Method (FGSM) [2] and Projected Gradient Descent (PGD) [36], have made them a central focus in adversarial machine learning research.

Poisoning-based attacks target the training process of machine learning models by injecting malicious or biased data into the training dataset [37–40]. By compromising the quality or integrity of the data, these attacks aim to influence the model's learning process, leading to suboptimal or harmful behavior during inference. Examples include flipping labels in a classification dataset, introducing outliers to distort regression models, or biasing pre-trained models through carefully crafted data samples. Poisoning attacks are particularly dangerous because they exploit the foundational training step, potentially affecting models deployed across multiple applications. They are also difficult to detect, as the injected data can be designed to blend seamlessly with the original dataset, making robust data curation and sanitation critical defenses.

Physical adversarial attacks extend beyond the digital space, exploiting physical systems, environments, and infrastructure [41–45]. These attacks often manipulate real-world inputs to deceive AI systems in situ, such as altering road signs to mislead autonomous vehicles, spoofing GPS signals to misdirect navigation systems, or overloading sensors with noise. Unlike digital perturbation or poisoning attacks, these real-world attacks often require careful engineering to ensure they remain effective under varying physical conditions, such as changes in lighting or perspective. They emphasize the challenge of bridging the gap between theoretical adversarial vulnerabilities and practical, real-world scenarios, raising significant concerns for the deployment of AI in critical fields like transportation, healthcare and security.

In this thesis, we investigate perturbation-based attacks against Deep Neural Networks (DNN) in machine learning systems. In the following sections, we review attack baselines for both Natural Language Processing (NLP) (Section. 2.2) and Computer Vision (CV) (Section. 2.3) models, addressing the unique characteristics of textual and visual inputs.

2.2 Adversarial Attack for NLP

This section reviews adversarial attacks in NLP, where vulnerabilities in models are exploited to induce errors in various textual tasks. We focus on four categories of NLP systems: textual classification models, translation models, abstract summarization models, and question answering models. Each subsequent subsection provides a broad analysis of the unique challenges posed by adversarial attacks in these domains, examines notable attack strategies, and explores their contributions to enhancing model robustness.

2.2.1 Textual Classification Models

Textual classification stands as a fundamental task in natural language processing (NLP), with widespread applications ranging from sentiment analysis [46] and spam detection [47] to topic categorization [48]. While the field has evolved from traditional machine learning approaches with handcrafted features to sophisticated deep learning architectures powered by pre-trained language models, a critical vulnerability persists: these classification models remain susceptible to adversarial attacks. Through subtle modifications to input text, these attacks can manipulate models into making incorrect predictions. The attack methodologies operate at multiple linguistic levels, encompassing character-level, word-level, and sentence-level approaches.

Traditional textual classification relied on machine learning methods like Naive Bayes [49], Support Vector Machines (SVMs) [50], and logistic regression [51]. These models employed handcrafted features such as bag-of-words, tf-idf, and n-grams, which captured shallow statistical properties of text but failed to model complex semantic relationships.

The development of deep learning networks like Recurrent Neural Networks (RNNs) [52], Long Short-Term Memory (LSTM) [53] networks, and Gated Recurrent Units (GRUs) [54], which captured sequential dependencies in text. Attention-based models, such as Hierarchical Attention Networks (HAN) [55], further improved the ability to focus on relevant input portions for classification tasks.

Pre-trained language models revolutionized textual classification by leveraging transformers. BERT (Bidirectional Encoder Representations from Transformers) [28], GPT (Generative Pre-trained Transformer) [56], and Roberta [57] achieved state-of-the-art results by generating contextualized embeddings that captured both local and global semantic information. These models fine-tuned on specific classification tasks, are now standard for achieving high performance across diverse NLP benchmarks.

Attacks against these models can be categorized into three distinct levels: character-level, word-level, and sentence-level modifications. Character-level adversarial attacks focus on modifying individual characters within the input text. These modifications can include typos, misspellings, or character substitutions using visually or semantically similar alternatives. For instance, Ebrahimi et al. [32] proposed HotFlip, a gradient-based method that identifies the most influential characters in a text and replaces them to maximize the likelihood of misclassification. Examples include changing “great” to “gr8” or “movie” to “m0vie”. These attacks highlight the sensitivity of models to small perturbations in the text, especially in tasks such as spam detection or sentiment analysis. TextBugger [58] further demonstrated that combining character-level changes with heuristic-based rules can achieve high attack success rates in both white-box and black-box settings.

Word-level attacks target individual words in the text, replacing, inserting, or deleting them to mislead the model. Synonym substitution is one of the most common techniques, where keywords in a sentence are replaced with semantically similar alternatives. TextFooler [11] uses gradient information to identify the most important words in a text and substitutes them with synonyms from embedding spaces or lexical databases, ensuring minimal change to the input’s semantics. Zang et al. [59] extended this approach with a particle swarm optimization (PSO) algorithm to optimize word substitutions and maximize attack effectiveness. Other methods, such as deletion-based attacks [60], remove specific words from the text to disrupt the classifier’s decision-making process. These attacks demonstrate that even minor changes at the word level can significantly affect the model’s predictions. Embedding-based attacks also fall under word-level methods. Michel et al. [61] showed that small perturbations in word embeddings, which represent words in a continuous vector space, can lead to misclassifications. These attacks exploit the semantic relationships captured by embeddings, revealing vulnerabilities in models

that rely heavily on these representations.

Sentence-level adversarial attacks involve more extensive modifications, such as rephrasing or restructuring sentences while preserving their semantic meaning. Sentence-level attacks, such as the ISPED algorithm proposed by Dong et al. [62], perturb key sentences within a text to mislead classifiers while maintaining semantic similarity. ISPED utilizes a semantic-aware similarity function to ensure that the adversarial examples remain imperceptible to human readers while effectively degrading model performance. This approach highlights the challenges of sentence-level attacks, particularly in tasks that rely heavily on context and linguistic structure. Syntax-aware attacks, proposed by Zhou et al. [63], focus on altering the grammatical structure of sentences, such as switching from active to passive voice or changing word order, to mislead classifiers. Another category of sentence-level attacks involves combining multiple perturbations at the character and word levels to create adversarial examples that affect entire sentences. Black-box attacks, such as TextBugger [58], leverage heuristic-based rules and query models to identify sentences that can be perturbed to maximize misclassification rates. These methods are particularly effective in real-world applications, where attackers lack access to model internals.

2.2.2 Translation Models

Machine translation aims to convert text from one language to another. Over the years, translation systems have evolved significantly, from rule-based approaches to modern neural network-based architectures [64]. Neural machine translation (NMT) has become the standard for high-quality translations, but it remains vulnerable to adversarial attacks that exploit its reliance on linguistic patterns and context [65].

Early translation systems, such as rule-based machine translation (RBMT), relied on predefined grammatical rules and dictionaries, which limited their flexibility and

scalability. The introduction of Statistical Machine Translation (SMT), exemplified by systems like Moses [66], marked a major advancement by leveraging probabilistic models trained on bilingual corpora. However, SMT struggled with long-range dependencies and reordering issues in languages with significant syntactic differences.

Neural Machine Translation (NMT) revolutionized the field with sequence-to-sequence (Seq2Seq) architectures [67], which encode input sentences into a latent representation and decode them into the target language. Attention mechanisms [68] enhanced translation quality by dynamically focusing on relevant input tokens during decoding. More recently, transformer-based architectures [69], such as MarianMT [67] and mBART [70], have pushed the boundaries of translation quality with their scalability and ability to model long-range dependencies.

Despite these advancements, translation models are highly susceptible to adversarial attacks that introduce subtle modifications to the input or exploit model weaknesses to degrade translation quality.

Character-level attacks, such as HotFlip [32], modify individual characters in the source text to mislead translation models. For example, changing “hello” to “helo” can cause significant translation errors, especially in models reliant on subword tokenization methods like Byte Pair Encoding (BPE). These seemingly minor typographical errors can propagate into substantial semantic inconsistencies in the output.

Word-level perturbations are another common form of attack. Seq2Sick [71] generates adversarial examples by replacing or inserting words in the source sentence, aiming to create misleading translations while preserving semantic similarity in the source language. For instance, substituting “president” with “leader” might lead to a translation that alters the intended meaning subtly but effectively.

Sentence-level attacks focus on rephrasing or restructuring the input sentence to confuse translation systems. Belinkov and Bisk [72] showed that syntactic variations,

such as converting “The cat is on the mat” to “On the mat is the cat,” can disrupt the coherence and accuracy of the translated output.

Context manipulation attacks exploit the dependency of translation models on the surrounding textual context. Zhao et al. [73] demonstrated that appending misleading or ambiguous clauses to the input sentence, such as “according to rumors” can severely affect the quality and accuracy of the translation.

Cross-lingual adversarial attacks, as described by Chen et al. [74], target multilingual translation models by introducing adversarial examples in one language to degrade performance across multiple target languages. For instance, a perturbation in an English sentence might result in errors in translations to French, German, and Spanish simultaneously, revealing systemic vulnerabilities in multilingual models.

These adversarial attacks expose critical weaknesses in translation models, from their reliance on tokenization and context to their sensitivity to linguistic variations. They highlight the need for robust defenses and improved architectures that can handle adversarial perturbations effectively.

2.2.3 Abstract Summarization Models

The increasing prevalence of network information in modern society has led to an explosion of textual content across various formats, including academic articles, novels, books, and reviews. This exponential growth in text length has created a significant challenge for readers who need to quickly access and understand relevant information [75]. For long articles, in particular, the traditional approach of manual summarization has become increasingly impractical and time-consuming. This challenge has sparked renewed interest in automated approaches to text summarization, especially in recent years.

Abstractive summarization has emerged as a promising solution within the field of natural language processing (NLP). Unlike extractive summarization, which simply

selects and copies existing segments from the source text, abstractive summarization aims to generate concise and coherent summaries by paraphrasing and rephrasing the original content. This approach allows for more flexible and natural summaries that can effectively capture the essential meaning of the source document while presenting it in a new form.

Early abstractive summarization systems were largely based on templates and manually designed linguistic rules, which limited their flexibility and scalability. For example, Radev et al. (2004) [76] introduced centroid-based multi-document summarization methods, but these approaches primarily focused on extractive techniques with limited abstractive capabilities.

The introduction of neural networks revolutionized abstractive summarization. Chopra [77] developed a convolutional attention-based conditional recurrent neural network (RNN) model for the same problem of abstractive sentence summarization. The framework consists of CNN [78] with an attention mechanism for the encoder and LSTM for the decoder. Sequence-to-sequence (Seq2Seq) models, originally designed for machine translation, were adapted for summarization. Sutskever et al. (2014) [67] demonstrated the potential of Seq2Seq architectures in text generation tasks. Further improvements were made by incorporating attention mechanisms, as proposed by Bahdanau et al. (2015) [68], enabling models to focus on relevant parts of the input text during decoding.

Despite their success, Seq2Seq models often faced issues such as repetition and lack of factual accuracy. To address these challenges, Paulus et al. (2018) [79] employed reinforcement learning to optimize models for non-differentiable objectives, such as ROUGE scores, which improved the fluency and relevance of generated summaries.

The advent of transformer architectures further advanced abstractive summarization. Pre-trained models such as BERT [28] and GPT [80] significantly improved the quality of text generation by leveraging self-attention mechanisms [77], as shown in Figure 2.1, and

large-scale pre-training. Transformer-based models like BART [14] and PEGASUS [15] introduced specialized pre-training objectives tailored for summarization. For instance, PEGASUS utilized a gap-sentence generation task to enhance its ability to identify and rephrase key information.

However, abstractive summarization still faces critical challenges. One major issue is factual inconsistency: models may hallucinate content that does not exist in the source text. Cao et al. (2018) [81] highlighted this problem and proposed fact-aware summarization methods to improve the fidelity of generated summaries. Another challenge lies in evaluation metrics. Commonly used metrics like ROUGE focus on surface-level overlap and fail to capture semantic correctness. Sellam et al. (2020) [82] proposed BLEURT, a learned metric designed to provide a more robust evaluation of text generation quality.

Future research aims to enhance factual consistency, improve evaluation methods, and address domain-specific summarization challenges. Additionally, extending abstractive summarization to multimodal content, such as combining text with images or videos, represents a promising frontier for the field.

In parallel, researchers have begun to investigate the robustness of abstractive summarization systems by applying adversarial attack strategies. These attacks aim to subtly perturb the input text in ways that mislead the summarization model while preserving human-level coherence and semantics.

A variety of attack frameworks have emerged based on token or sentence-level importance ranking. TextFooler [11] utilized gradient-based scoring to identify high-impact words for adversarial replacement, while PWWS [83] introduced genetic algorithms that balance attack strength and semantic similarity. Particle swarm optimization (PSO)-based methods [59] have also been used to guide word substitution by maximizing prediction divergence.

Wallace et al. [31] proposed universal adversarial triggers,Äîshort token sequences

that can consistently fool models regardless of the input,Äîdemonstrating the potential for highly transferable attacks. In BERT-Attack [60], contextual embeddings were leveraged to rank and substitute tokens, whereas BAE [84] combined word- and sentence-level transformations in a hierarchical manner.

To enhance fluency and semantic preservation, recent efforts have integrated paraphrasing into attack pipelines. Methods such as those by Iyyer et al. [85], Kassem et al. [86], and Ter Hoeve et al. [87] use transformer-based generators (e.g., T5, PEGASUS) to rewrite sentences in ways that maintain meaning while altering model behavior. These methods reveal significant vulnerabilities in summarization models, particularly their sensitivity to surface-level variation and their limited understanding of factual grounding.

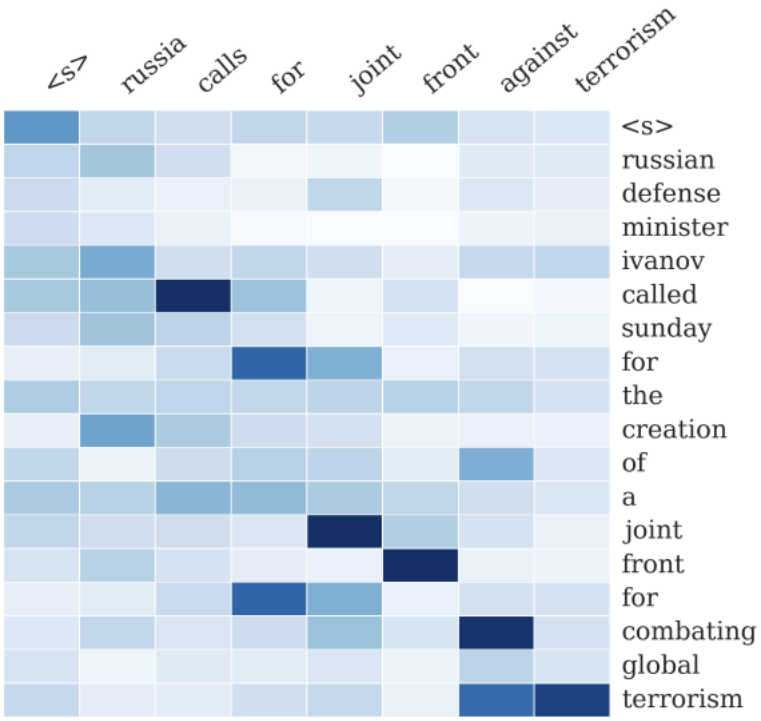

Figure 2.1: Attention Based Summarization by [1]

2.2.4 Question Answering Models

Question answering represents a complex interplay of NLP, information retrieval, and reasoning capabilities [88, 89]. These models are designed to process an input question and a context passage, extracting or generating an appropriate answer through elaborate analysis of the semantic relationships between these elements [90]. Modern QA systems typically rely on deep learning models with transformer-based architectures like BERT [28] and its variants [91–93] being particularly prevalent. These models excel at capturing contextual information and understanding nuanced relationships in the text with transformers, allowing them to perform impressively on QA tasks. In addition to these transformer models, encoder-decoder architectures such as T5 [16, 94] and BART [14], GPT [56] and PEGASUS [15] have also become prominent in QA models. These models utilize an encoder to process the input question and context, transforming them into a rich, context-aware representation, and the decoder is then used to generate a coherent and contextually appropriate answer.

With the development of NLP techniques, recent research has increasingly focused on developing sophisticated textual adversarial examples for QA systems [4]. The inherent differences between “informative queries” and “boolean queries” necessitate distinct attacking diversities due to their unique answer structures [31]. Attacks on boolean QA pairs closely resemble methods used to mislead textual classifiers. These attacks primarily operate at the word level, aiming to manipulate the model’s binary (yes/no) output [60, 84]. In contrast, informative queries present a more complex challenge. These attacks frequently target the sentence level, requiring an approach to disrupt the model’s comprehensive understanding [95].

Boolean queries are similar to classification tasks in NLP, while the answer is based on two-way input: question and context. They are vulnerable to attacks designed for NLP classifiers when question and context are simply encoded and concatenated.

Approaches such as [11, 36, 59, 60, 83, 84] concentrate on altering individual words based on their influence on model predictions. These methods typically employ carefully selected synonyms for word substitution. The process of word replacement is guided either by the direct use of BERT Masked Language Model (MLM) [28] or by leveraging gradient information to determine optimal substitution candidates. While effectively fool classifiers (boolean queries), these attacks were initially designed for classification tasks and have shown limited efficacy when applied to the question-and-context format of QA systems. To address this limitation, some attack methods for Seq2Seq models have been adapted for QA models. UAT [31], which averages gradients and modifies input data to maximize the model's loss, has been adapted for QA but still struggles with boolean queries due to their simplicity. Similarly, TextBugger [58], which focuses on character-level perturbations, also faces challenges in handling the deeper semantic understanding required in QA, especially for multi-sentence reasoning. Liang's approach [96], relying on confidence-based manipulations, has difficulty reducing the model's certainty in boolean queries where the binary answers leave less room for variation in confidence. Although these approaches offer improved accuracy in attacking informative questions with minor modifications, they struggle with boolean queries. We argue that these methods face challenges in identifying the most vulnerable words when dealing with concatenated question-context input relationships. The MLQA attack [97] attempts to bridge this gap by utilizing attention weights to identify and alter influential words. However, this method, developed specifically for multi-language BERT models, may not fully address QA-specific vulnerabilities.

In contrast to boolean queries, adversarial attacks on informative queries within QA systems share fundamental similarities with attacks on other Seq2Seq models [24, 33, 68, 71, 98, 99], concentrating more on the inter-relationship between question and context. Mechanisms like RobustQA [5] have been developed to enhance model resilience

through improved training methods, and sophisticated attacks continue to successfully compromise these systems, especially when employing subtle manipulations of key input elements. Character-level attack methods, notably HotFlip [32], have demonstrated significant success by strategically flipping critical characters based on gradient information, leading to misinterpreting informative inputs. In the multilingual domain, MLQA [100] leverages attention weights to identify and target crucial words, though its attention mechanism, primarily designed for multilingual functionality, may not fully exploit the intricate vulnerabilities within the model's attention architecture. Advanced techniques have emerged to target the influence that answers have on QA systems. Position Bias and Entropy Maximization methods exploit model weaknesses by manipulating contextual patterns and answer positioning, particularly effective in scenarios involving complex, lengthy responses. Syntactically Controlled Paraphrase Networks (SCPNs) [85] generate adversarial examples through strategic syntactic alterations while preserving semantic meaning. TASA (Targeted Adversarial Sentence Analysis) [3] primarily relies on manipulating the answer sentences to mislead QA models, making it particularly effective for informative queries where complex responses provide more opportunities for subtle modifications. However, this approach is not suitable for boolean queries, as the simplicity of yes/no answers limits the sentence-level manipulations that TASA depends on.

2.3 Adversarial Attack for CV

Computer vision (CV) is a popular field of machine learning that leverages deep neural networks (DNNs) to interpret and analyze visual data such as images and videos [101]. Its applications span diverse domains, from autonomous driving [102] and healthcare [103] to security [104] and entertainment [105], where robust and accurate interpretation of visual information is paramount. While DNNs have revolutionized modern

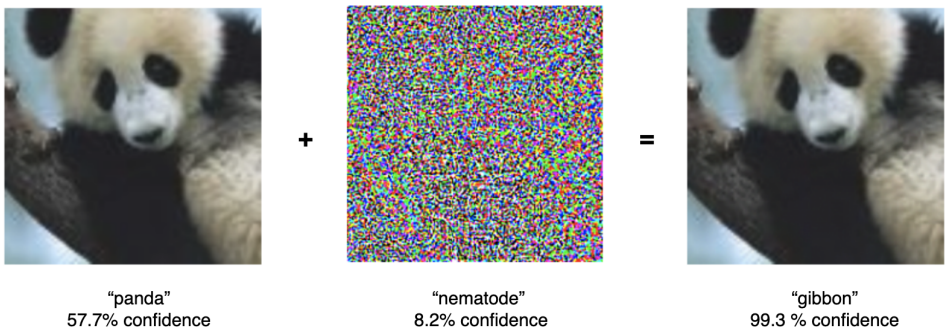

Figure 2.2: The figure illustration for Fast Gradient Sign Method [2]. It demonstrates adversarial attack generation on ImageNet. The attack modifies the original image by adding a minimal perturbation vector, computed as the sign of the cost function’s gradient.

computer vision systems through their exceptional ability to learn hierarchical representations from raw data, their vulnerability to adversarial attacks presents a critical security concern, particularly in safety-sensitive applications.

According to different tasks in computer vision, such as image classification, object detection, and image segmentation, image captioning. This task-based organization provides insights into the unique challenges and vulnerabilities associated with each application, offering a comprehensive view of adversarial attacks in CV.

2.3.1 Image Classification Models

Image classification differs from text classification as it processes visual data rather than textual input, requiring models to extract and interpret features directly from pixel-based representations. Deep learning approaches, particularly convolutional neural networks (CNNs) [106–109], have transformed this field by achieving unprecedented accuracy on benchmark datasets like ImageNet [110] and CIFAR-10 [111].

Despite their impressive capabilities, these powerful models have a critical vulnerability: they are susceptible to adversarial attacks. In these attacks, minute and often imperceptible changes to pixel values can cause the models to dramatically misclassify

inputs. These attack strategies can be categorized into two main types: black-box attacks, where the attacker has no access to model's internal parameters (gradients), and white-box attacks, where the attacker has part or full access to these parameters.

$$
\begin{aligned}
& x^{adv} = x - \varepsilon \cdot sign(\bigtriangledown_x J(x, y_{target})) \\
& x_0^{adv} = x,\ x_{N+1}^{adv} = clip_{x,\epsilon}\left\{ x_t^{adv} + \alpha \cdot sign(\bigtriangledown_x J(x_t^{adv}, y)) \right\}
\end{aligned}
\tag{2.1}
$$

The representative method of white-box attacks is FGSM, as illustrated in Figure 2.2, a carefully crafted noise pattern added to an image of a panda can cause the model to misclassify it as a gibbon with high confidence [2]. This vulnerability arises from fundamental properties of neural networks: their locally linear behavior in high-dimensional feature spaces makes them susceptible to adversarial manipulation. Building on this insight, the Basic Iterative Method (BIM) [112] enhanced the attack effectiveness by applying FGSM iteratively, generating stronger adversarial examples through cumulative perturbations. Projected Gradient Descent (PGD) [36], widely considered one of the most potent attack methods, further refined this approach by projecting perturbations into a constrained space after each iteration, ensuring bounded modifications while maximizing attack success. The comparison of the two methods is shown in Equation 2.1. The experiment showed that FGSM outperformed BIM in the white-box attack, while BIM had a higher success rate than FGSM in the black-box attack. The potential reason is that iterative methods could cause overfitting to a specific model. The Momentum Iterative Method (MIM) [113] boosts gradient-based attacks by adding momentum, which helps accumulate gradients across iterations, avoids getting stuck in local maxima, and ultimately improves upon standard FGSM techniques.

The collection of white-box attacks has expanded to include increasingly sophisticated methods. The DeepFool attack [24] generates minimal adversarial perturbations by iteratively projecting inputs toward the nearest decision boundary. Taking a different approach, the Jacobian-based saliency Map Attack (JSMA) [114] achieves targeted mis-

classification by identifying and modifying the most influential pixels through saliency maps. Universal Adversarial Perturbations [115] demonstrate a particularly concerning capability: generating a single perturbation pattern that can successfully mislead multiple inputs across different datasets and models. The One Pixel Attack [116] reveals the extreme brittleness of neural networks by achieving misclassification through the modification of just a single pixel. Combining different optimization objectives, the Elastic-Net Attack (EAD) [117] employs both "L1" and "L2" regularization to craft adversarial examples that balance effectiveness with imperceptibility.

Beyond white-box or gradient-based approaches, the field has evolved to encompass more sophisticated attack methodologies. The Carlini-Wagner (CW) attack [9] represents a significant advancement through its optimization-based approach, which carefully balances attack effectiveness with perturbation imperceptibility using a custom loss function. Its notable ability to circumvent many defensive measures has established it as a benchmark for evaluating model robustness. The development of black-box attacks has further expanded the threat landscape. Zeroth-Order Optimization (ZOO) [118] demonstrates that attackers can craft adversarial examples without access to model architecture or gradients, relying solely on input-output queries to approximate gradients. This capability makes such attacks particularly relevant for real-world systems deployed as APIs or cloud services. GenAttack [119] harnesses evolutionary algorithms for efficient adversarial example generation, while Boundary Attack [10] employs decision boundary exploration through iterative refinement. Transfer-based techniques [120] exploit a crucial vulnerability: adversarial examples crafted for one model often successfully deceive other models trained on similar tasks.

2.3.2 Objective Detection Models

Object detection is a branch of computer vision that involves identifying and localizing multiple objects within an image. Modern object detection models, such as Faster R-CNN [121], YOLO (You Only Look Once) [122], and SSD (Single Shot MultiBox Detector) [123], have achieved remarkable success. Despite these advancements, adversarial attacks pose a critical threat to the robustness and reliability of these systems. Unlike image classification, adversarial attacks on object detection must not only manipulate class predictions but also disrupt bounding-box coordinates, making them more complex and challenging.

Early research on adversarial attacks adapted gradient-based techniques from image classification to object detection. Liu et al. [124] demonstrated that gradient-based methods like FGSM and BIM could be extended to disrupt classification and localization in detection models. These attacks typically generate perturbations to misclassify objects, remove detected objects, or create false positives. For example, an adversarial perturbation could make a detection model fail to recognize a stop sign or falsely identify an object in the scene, compromising the system's functionality. Zhao et al. [125] introduced targeted adversarial attacks for object detection, where perturbations are crafted to make the model predict specific incorrect labels and bounding box coordinates for objects. These targeted attacks require precise optimization to achieve the desired outcomes while maintaining imperceptibility. More recent research has focused on developing black-box attack methods that do not require access to the model's architecture or parameters. Duan et al. [126] proposed a transfer-based approach, leveraging the transferability property of adversarial examples. By generating adversarial perturbations on surrogate models, these attacks can deceive target detection models with limited knowledge of their internals.

Optimization-based black-box attacks, such as those using zeroth-order optimiza-

tion, have also been explored for object detection. These methods estimate gradients by querying the model and iteratively refining the perturbations to achieve the desired adversarial effect. Cheng et al. [127] introduced a query-efficient black-box attack specifically designed for detection models, reducing the number of queries needed to craft successful adversarial examples.

Adversarial patch attacks represent a unique category of attacks on object detection models. These attacks involve creating small, localized perturbations, often patches, that can be physically placed in the environment to fool detection systems. Eykholt et al. [41] demonstrated the effectiveness of adversarial patches in causing object detection models to ignore or misclassify objects in real-world settings. Unlike traditional pixel-level perturbations, adversarial patches are designed to work under various physical conditions, such as changes in lighting and perspective.

2.3.3 Image Captioning Models

Image captioning serves as a vital bridge between computer vision (CV) and natural language processing (NLP). It involves generating descriptive sentences in natural language that accurately reflect the visual content of an image. This task requires a deep understanding of both the objects and the relationships within the image, as well as the ability to articulate these observations into coherent sentences.

Early models for image captioning primarily relied on encoder-decoder architectures, where convolutional neural networks (CNNs) serve as encoders to extract features from the image, and recurrent neural networks (RNNs) function as decoders to generate the caption as a sequence of words. One of the foundational models in this domain is the "Show and Tell" model, introduced by Vinyals et al. [128], which uses a CNN to encode image features and an RNN to decode these features into a sentence.

In recent years, attention mechanisms have dramatically improved the quality of

image captioning models. Attention mechanisms allow models to focus on different parts of the image while generating each word in the caption, enhancing the model's ability to capture fine-grained details and contextual information. One prominent example is the "Show, Attend, and Tell" model [129], which introduced an attention-based approach that enabled the model to selectively focus on different image regions during the generation process, resulting in more accurate and contextually relevant captions. This attention-based approach has since become a standard in many modern image captioning models, with extensions that further improve the precision of caption generation by incorporating hierarchical attention [130] and semantic scene graphs [131].

Other advancements in image captioning have included models that incorporate reinforcement learning [132], which optimize the generated captions based on reward functions like BLEU or CIDEr scores, leading to captions that are more semantically meaningful and aligned with human preferences. Additionally, transformer-based models, such as the Vision Transformer (ViT) [133], have been applied to image captioning tasks, leveraging the transformer's ability to model long-range dependencies and complex relationships between visual elements in images.

Adversarial attacks in image captioning represent a more complex challenge than attacks in traditional image classification or NLP tasks. Image captioning models, which generate descriptive sentences from images, are vulnerable to adversarial examples that can manipulate both the visual input and the textual output. These attacks can lead to the generation of captions that either misdescribe the image or fail to capture its critical details.

Early research in this area focused on adapting existing adversarial attack techniques from image classification to image captioning. With the advent of the fast gradient sign method (FGSM) [2], deep neural network models (DNNs) have exhibited vulnerabilities to adversarial examples. A multitude of gradient-based attack techniques have

emerged, such as the Basic Iterative Method (BIM) [112] and Projected Gradient Descent (PGD) [36], which enhanced FGSM by iteratively updating perturbations, resulting in more potent attacks. The "Show-and-Fool" method [134] is one of the first approaches specifically targeting image captioning models by applying these gradient-based attacks to modify pixel values in the input image, thereby altering the generated caption. However, like many early approaches, Show-and-Fool operates in a white-box setting, assuming full access to the model's architecture and parameters, which is often impractical in real-world scenarios. Moreover, these white-box attacks only target specific models and have relatively weak transferability [135].

To transcend this constraint, black-box attacks based on generative adversarial networks (GANs) have been developed [136, 137]. These attacks deploy GANs to generate perturbations that can mislead the model. However, training GANs is often fraught with instability, and tasks that target an image, such as image captioning, cannot meet the requirement of modifying pixels. More recent research has aimed to develop alternative black-box attack methods for image captioning that are more applicable to real-world scenarios. Optimization-based attacks [24, 115] offer a potential solution by casting the creation of adversarial examples as an optimization problem to identify the minimal perturbation that induces misclassification [116]. For example, the work by Aafaq et al. [138] introduces a grey-box adversarial attack on image captioning that manipulates attention mechanisms within the model. Similar strategies have also been applied in domains like road sign recognition [139] and Stackelberg adversarial games [140, 141].

Visual attention also plays a pivotal role in adversarial attacks. By identifying regions of interest within an image, attackers can precisely target critical areas for perturbation, ensuring maximum impact while keeping the modifications imperceptible to the human eye [142–146]. This guided perturbation approach is particularly effective in crafting adversarial examples for models that rely on spatial and contextual information, such as

image captioning systems. These mechanisms, initially introduced to enhance caption generation, have been successfully repurposed to increase the efficiency and stealth of adversarial attacks.

The differential evolution (DE) algorithm [147] has emerged as a powerful black-box optimization technique for adversarial example generation. DE is especially useful in scenarios involving non-differentiable, high-dimensional objective functions,Äîcharacteristics commonly found in image-based attacks. It works by maintaining a population of candidate solutions and iteratively refining them through mutation, crossover, and selection operations. In the context of image captioning, DE can optimize perturbations over attention-highlighted regions, allowing attacks to focus on areas most influential to caption generation. Prior work has demonstrated that DE-based methods can effectively balance attack success and imperceptibility, without requiring access to model gradients [116].

Despite these advances, attacking image captioning models remains inherently difficult due to their generative nature and sequential outputs [148, 149]. Unlike classification tasks, where a single label is the output, captioning models produce word sequences, introducing dependencies among tokens and complicating loss function computation. One proposed solution is to treat the entire sentence as a single unit when computing adversarial loss [150], simplifying the optimization process.

Recent approaches have also explored targeted attacks that attempt to force the model to generate specific vocabulary within the caption [151, 152]. These attacks are formulated as constrained optimization problems and have shown success under white-box assumptions. However, such methods often require access to internal model parameters, limiting their applicability in real-world black-box scenarios.

Another significant challenge lies in evaluating the effectiveness of adversarial attacks in captioning tasks. Unlike classification, where accuracy drop is a clear metric,

evaluating sequence generation is less straightforward. Commonly used metrics like BLEU and ROUGE capture surface-level overlap but fail to reflect semantic shifts or factual inconsistencies caused by perturbations. Consequently, there is growing interest in developing more robust, semantically-aware evaluation metrics tailored to the nature of generative models [153].

2.4 Discussion and Challenges

In this section, we summarize our findings and discuss the main challenges in attacking three types of models spanning both Natural Language Processing (NLP) and Computer Vision (CV): abstractive summarization, image captioning, and question answering systems.

Abstractive summarization poses perhaps the most complex challenges for adversarial attacks due to its document-level nature. While most existing adversarial methods excel at word-level or sentence-level tasks, they struggle when applied to document-level summarization. The fundamental challenge lies in the need to manipulate entire sections of text rather than isolated tokens while maintaining coherence throughout the document. This creates a delicate balance between perturbation extent and attack effectiveness - too much perturbation makes the attack detectable, while too little fails to deceive the model. Current paraphrasing approaches, while promising, face significant limitations in ensuring semantically correct but contextually inconsistent replacements, reducing the attack's quality and impactally inconsistent, reducing both the quality and impact of the attack.

In image captioning attacks, the primary limitations stem from both white-box and black-box approaches. Traditional gradient-based attacks like FGSM, BIM, and PGD were originally designed for classification tasks and heavily depend on white-box access to model parameters, significantly limiting their real-world applicability. These methods

also demonstrate poor transferability across different models, further restricting their effectiveness. When attempting black-box approaches, GAN-based attacks offer potential solutions but face significant challenges in training stability and direct application to captioning tasks. The fundamental difficulty lies in handling discrete sequence outputs and the reduced effectiveness of pixel-level perturbations in generative contexts. The complex relationships between words in generated sequences make it particularly challenging to apply conventional adversarial methods while maintaining semantic coherence.

Question answering systems present their own unique set of challenges for adversarial attacks. Traditional classification-based attacks like BERT-based approaches, TextBugger, and UAT struggle to address the specific requirements of QA tasks. However, the heavy reliance on question-context relationships presents distinct vulnerabilities that classification-centric methods do not address effectively. The challenge becomes even more complex with informative queries, which depend on intricate interrelations between questions and contexts. Cross-language applications add another layer of complexity, as existing methods like MLQA and RobustQA struggle to address vulnerabilities in multilingual settings comprehensively. The heavy reliance on attention mechanisms in multilingual QA models creates specific vulnerabilities, yet current attack methods often fail to combine gradient-based perturbations with attention-weight exploitation effectively.

The practical implementation of these attacks faces additional challenges in real-world scenarios. While adversarial attacks are typically evaluated on benchmark datasets, their effectiveness in real-world applications remains largely unexplored. This is particularly crucial in domains like news summarization, legal document processing, and research article generation, where model robustness against adversarial inputs is critical. The lack of comprehensive evaluation methods for real-world scenarios and the need for domain-specific validation approaches create significant barriers to practical

implementation. Furthermore, the development of effective black-box attack methods remains an open challenge across all these tasks, as most current strategies rely heavily on access to model gradients or internal parameters.

Common across all these tasks is the fundamental challenge of balancing attack effectiveness with semantic preservation. As the complexity increases from basic classification tasks to sophisticated sequence generation, the need for more nuanced approaches becomes apparent. The development of attack methods must consider not only the technical aspects of model deception but also the maintenance of output quality and coherence. This becomes particularly crucial in long-form content generation, where maintaining consistent semantic relationships and logical flow throughout the sequence poses significant challenges for current attack methodologies.

CHAPTER 3

SUMMARIZATION ATTACK VIA PARAPHRASING

Many natural language processing models are perceived to be fragile in adversarial attacks. Recent work on adversarial attacks has demonstrated a high success rate in sentiment analysis and classification models. However, attacks on summarization models have not been well-studied. Summarization tasks are rarely influenced by word substitution since advanced abstractive summary models utilize sentence-level information. In this chapter, we propose a paraphrasing-based attack method called Summarization Attack via Paraphrasing (SAP) aiming at abstractive summarization models. We first rank the sentences in the document according to their impacts on summarization. Then, we apply the paraphrasing procedure to generate adversarial samples. Finally, we test our algorithm on benchmark datasets against other methods. Our approach achieved the highest success rate and the lowest sentence substitution rate. In addition, the adversarial samples have high semantic similarity with the original sentences.

3.1 Introduction

Abstractive summarization, which generates concise summaries by synthesizing new sentences from source documents, represents a complex challenge in natural language processing (NLP). While models have achieved remarkable progress in this domain, they remain vulnerable to adversarial attacks - subtle manipulations of input text that can induce the generation of inaccurate or misleading summaries. Although adversarial attacks have been extensively studied for tasks like text classification [60, 154] and machine translation [95, 99], document-level abstractive summarization presents unique challenges that remain underexplored. State-of-the-art abstractive summarization models, such as T5 [16] and Bart [14], leverage sophisticated encoder-decoder architectures with attention mechanisms to generate summaries that capture key information while potentially differing lexically from the source text. These systems have found widespread application across diverse domains, including news [155], legal document processing [156], education [157], social media content analysis [158], and healthcare records summarization [159]. However, these advanced models exhibit high sensitivity to adversarial perturbations, which can significantly disrupt their sequence-level generation process. Earlier attempts to study adversarial attacks in summarization relied primarily on handcrafted or heuristic approaches, but these methods proved limited in both scalability and efficiency [4, 160]. Traditional NLP adversarial attacks have predominantly focused on word-level perturbations, including synonym substitution and token-level replacements [32, 60]. While these approaches prove effective for shorter texts like sentence classification, they face significant limitations when applied to document-level summarization, where perturbations must account for broader context and complex sequence relationships. The challenge lies in maintaining a delicate balance: perturbations must be substantial enough to mislead the model while preserving semantic integrity and remaining imperceptible to human readers. Additionally, identifying optimal target sen-

tences within lengthy documents adds another layer of complexity to the attack strategy. To address these challenges, we propose a novel adversarial attack framework specifically designed for abstractive summarization models. Our approach employs paraphrasing techniques to generate subtle, contextually coherent perturbations while preserving the document's overall meaning. The framework incorporates a ranking mechanism to identify and target the most influential sentences, ensuring that generated adversarial examples effectively deceive the model while remaining undetectable to human evaluators. The key innovation of our framework lies in its sentence-level operation, moving beyond traditional token-level substitutions or gradient-based perturbations. By leveraging advanced paraphrasing techniques, our method maintains textual fluency and coherence while manipulating key elements to mislead the summarization model. This approach is well-suited for document-level summarization tasks, where preserving context and sequence relationships is crucial for generating effective adversarial examples.

In summary, the contribution of this work is:

- We propose a novel adversarial attack system Summarization Attack via Paraphrasing (SAP) for document-level abstractive summarization models, utilizing sentence-level paraphrasing to generate adversarial examples.
- SAP introduces a ranking mechanism to identify and target the most influential sentences in the input document, enabling precise and effective perturbations.
- We conduct extensive experiments on real-world datasets and state-of-the-art summarization models, demonstrating the efficacy of our method in generating imperceptible yet impactful adversarial examples.

The subsequent sections of this chapter are structured as follows. In Section 3.2, we introduce the background on importance ranking and paraphrasing. Section 3.3

details the methodology of the proposed framework. We present experimental results and analysis in Section 3.4. Finally, we conclude the chapter and discuss potential future directions in Section 3.5.

3.2 Preliminary

In this section, we introduce the foundational concepts that support our adversarial summarization framework, specifically focusing on importance-based ranking and paraphrasing as the core components.

3.2.1 Importance Ranking

Importance ranking is a fundamental strategy in textual adversarial attacks to identify the most influential components within a document. This involves evaluating the importance of sentences, phrases, or tokens based on their contribution to model predictions. Traditional approaches often use metrics like model's loss for classification tasks, ROUGE [25] or BLEU [161] scores for Seq2Seq tasks, which measure the overlap between generated outputs and reference data, to assess the significance of each sentence in a summarization task.

Advanced strategies further leverage gradient-based techniques, semantic similarity metrics, or prediction probabilities to refine the ranking process. This ranking procedure serves as the basis for selecting candidate regions in the input to be perturbed, guiding the design of effective and minimal adversarial attacks.

3.2.2 Paraphrasing Models

Paraphrasing involves crafting alternative textual formulations that maintain the original semantic content [162]. In the context of adversarial NLP tasks, it is particularly

useful for generating input perturbations that preserve human readability and fluency while challenging model robustness.

Modern paraphrasing techniques span from rule-based and statistical methods to deep neural models and large-scale transformer-based architectures. Neural sequence-to-sequence (Seq2Seq) models encode an input and decode a semantically equivalent variant, often through techniques such as back-translation [163]. Recent transformer-based models like T5 [164] and Pegasus [15] achieve state-of-the-art performance in generating high-quality paraphrases that preserve contextual meaning.

In adversarial contexts, these paraphrasing tools allow controlled, semantics-preserving modifications to the input that exploit model vulnerabilities without introducing unnatural or noisy text.

3.2.3 Potential Social Impact

The deployment of adversarial summarization attacks represents a significant threat to information integrity across critical sectors. In financial markets, algorithmic trading systems increasingly depend on automated news summarization for decision-making [165]. Malicious actors could exploit this dependency by subtly modifying earnings report summaries—transforming "strong growth ahead" into "moderate progress expected"—thereby manipulating market sentiment without detection. Such seemingly minor alterations could trigger algorithmic trading cascades, potentially destabilizing markets and impacting institutional portfolios and individual retirement accounts.

Healthcare systems face comparable vulnerabilities as medical professionals increasingly utilize AI-powered summarization tools to synthesize clinical literature [166]. Adversarial manipulation of research summaries concerning vaccine efficacy or drug interactions could compromise clinical decision-making while maintaining semantic plausibility. The COVID-19 pandemic demonstrated how information distortion can

rapidly influence public health outcomes [167]. Sophisticated paraphrasing attacks could systematically erode confidence in evidence-based medicine precisely when accurate information dissemination is most critical.

Political communication systems exhibit particular susceptibility to such attacks. Evidence indicates that social media users predominantly engage with headlines and automated summaries rather than full articles [168]. This behavior pattern creates opportunities for adversaries to manipulate public discourse through systematic paraphrasing of political content, subtly altering the perceived positions of candidates or the implications of policies. Unlike conventional disinformation campaigns that rely on fabrication, paraphrasing attacks preserve factual accuracy while modifying rhetorical emphasis—a characteristic that renders them resistant to traditional fact-checking mechanisms. Given the established influence of framing effects on electoral behavior [169], coordinated paraphrasing campaigns could potentially influence democratic processes while evading existing content moderation frameworks.

These vulnerabilities are amplified by the expanding integration of large language models into information infrastructure, encompassing applications from legal document analysis to academic peer review [170]. As automated summarization becomes increasingly embedded in knowledge dissemination systems, the potential for adversarial manipulation to influence collective decision-making processes escalates correspondingly. This emerging threat landscape necessitates the development of robust detection mechanisms and defensive strategies to preserve information integrity in AI-mediated communication systems.

3.3 Proposed Method

In this section, we introduce the SAP algorithm. The algorithm can be divided into two steps: importance ranking and sentence replacement. The completion steps of our

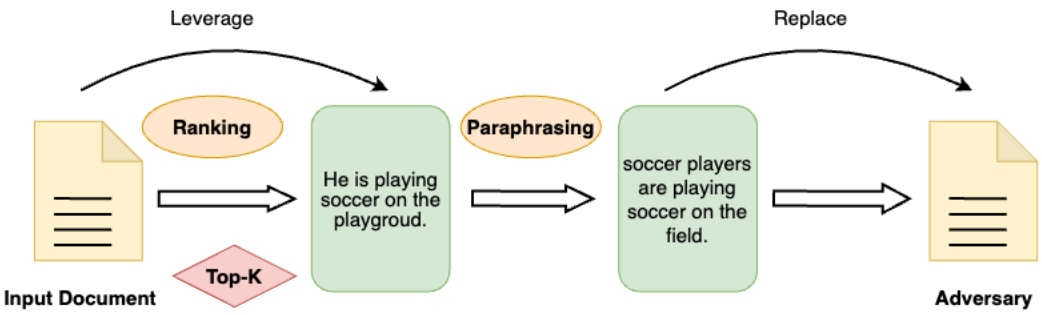

Figure 3.1: The brief workflow of SAP approach. For each target document, we rank out the sentences in reverse order and rebuild them by replacing the $top-k$ sentences with sentences produced with the paraphrasing model.

proposed method are shown in Algorithm 1.

3.3.1 Importance Ranking

The first step of our method involves sentence-level importance ranking, which identifies sentences that significantly influence summarization outputs. Evaluation in summarization predominantly depends on the ROUGE score [25], which quantifies similarity between machine-generated and reference summaries. Building upon this widely adopted metric, we propose a ranking procedure integral to enhancing the effectiveness of our summarization attack via paraphrasing.

Our ranking procedure systematically evaluates the importance of individual sentences within a document by examining the impact of their removal on the summarization model's performance. Specifically, sentences are deleted one at a time to generate modified versions of the input document. These modified documents are then fed into the summarization model to produce summaries. The importance of each sentence is quantified by analyzing the change in the ROUGE score before and after the deletion of the target sentence.

Let $a = [s_1, \ldots, s_i, \ldots]$ represent the input document, where each s_i is a sentence. Let

F_s denote the summarization model employed in the ranking process, R represent the ROUGE score function, and D_0 and D_i denote the original document and the document with sentence s_i removed, respectively. The importance of a sentence, G_i, is calculated as:

$$G_i = R(F_s(D_i)) - R(F_s(D_0)) \tag{3.1}$$

In this formulation, G_i captures the impact of removing sentence s_i on the quality of the generated summary. A higher G_i value indicates that sentence s_i is more critical to the summarization output.

3.3.2 Sentence Replacement

The next step is Sentence Replacement. Previous text attack methods obtain substitutes from the prediction of Masked Language Model. It requires a sophisticated selection strategy from predicted words. However, summarization models mine the semantic relationship between sentences and substituting words has a low effect on misleading summary models. Therefore, a hypothesis to build an adversary document is to simply remove influential sequences, but lacking a bunch of topic sentences would also mislead human readers. Hence, in our approach, we tested $top-k$ from 10% to 30% of document length to balance the attack cost and semantics of documents. The chosen paraphrasing model uses hierarchical sketches to build semantically preserved sentences with various vocabulary and grammar. It is known that the paraphrasing approach maintains the grammatical and semantic features of each sentence with a combination sentence structure.

Algorithm 1: Summarization Attack with Paraphrasing

Input: Dataset D with article inputs a; Pegasus model (victim model F); ROUGE score for output $F(a)$ is R_0; K is the percentage of attacked sentences in a document; α is the maximum number of sentences to perturb; Function to get adversarial sample F_{adv}. ;
Input: Adversarial samples adv ;
// **Sentence Importance Ranking**;
Initialize sentence importance ranking list: $sentence\_rank \leftarrow [\,]$;
for *each* s_i *in* a **do**
    Build document without sentence s_i;
    Calculate ROUGE score R_i without s_i;
    Append ROUGE difference to ranking: $sentence\_rank.append(R_0 - R_i)$;
end
Compute dataset size: $size \leftarrow len(D)$;
Initialize total length: $length \leftarrow 0$;
for *each* a *in* D **do**
    $length \leftarrow length + len(a)$;
end
Compute average document length: $ave\_length \leftarrow length/size$;
Compute threshold: $threshold \leftarrow \min(ave\_length \cdot K, \alpha)$;
Select top-ranked sentences: $ranked \leftarrow sentence\_rank[:threshold]$;
// **Generating Adversarial Example**;
Initialize adversarial samples: $adv \leftarrow [\,]$;
for *each* s_i *in* $ranked$ **do**
    Generate adversarial sentence: $s_{adv} \leftarrow F_{adv}(s_i)$;
    Construct adversarial document: $a_{adv} \leftarrow [s_1, \ldots, s_{adv}, \ldots, s_n]$;
    Append adversarial document to adv;
end
return adv

3.4 Experiment Results and Analysis

This section presents a comprehensive evaluation of our SAP performance compared to baselines. Our analysis encompasses multiple dimensions, utilizing diverse metrics to thoroughly understand the method's effectiveness and robustness across various contexts. We begin by introducing the datasets and victim models used in our experiments (Sec. 3.4.1). Next, we analyze the experimental outcomes (Sec. 3.4.4). Then, we conduct ablation studies (Sec. 3.4.6) to analyze the contribution of individual components and

assess the computational efficiency of our attack strategy. We then examine the transferability of attacks (Sec. 3.4.7) to demonstrate SAP's performance on other Seq2Seq tasks. Finally, we explore how adversarial retraining helps deepen our understanding of summarization models (Sec. 4.4.7).

3.4.1 Datasets

We evaluate our SAP method on five real-world abstractive summarization datasets, each presenting distinct summarization challenges. The XSum [171] dataset consists of BBC news articles paired with highly abstractive single-sentence summaries, aiming to test the model's ability to produce focused and concise outputs. The CNN/Daily Mail [172] dataset includes news articles accompanied by multi-sentence summaries written by human annotators, providing a large-scale benchmark for document-level summarization. WikiHow [173] contains instructional articles and their concise summaries, designed to evaluate the summarization of procedural and step-by-step content. Multi-News [174] offers a multi-document summarization setting in which each instance contains multiple news articles and a single reference summary, challenging models to synthesize information from diverse sources. Reddit TIFU [175] is constructed from user-generated posts on Reddit, featuring informal and conversational language that tests model robustness in handling noisy and colloquial input.

3.4.2 Victim Models

We evaluate our approach on three stunning abstractive summarization models: Pegasus [15], T5 [16], and Bart [14]. Pegasus is specifically designed for abstractive summarization, employing a gap-sentence pretraining objective that predicts masked sentences. T5 approaches summarization as a text-to-text task, offering enhanced flexibility, while Bart leverages a denoising autoencoder architecture to handle noisy inputs and gener-

Table 3.1: Dataset distribution and corresponding baseline performance (ROUGE-1).

Datasets	Data Distribution				Model Performance		
	Total	Train	Validation	Test	Pegasus	T5	Bart
XSum	226,711	204,045	11,332	11,334	47.60	49.83	51.21
CNN/Daily Mail	311,971	287,227	13,368	11,376	64.16	61.56	62.11
WikiHow	230,843	200,000	15,000	15,843	66.39	67.12	65.82
Multi-News	56,216	44,972	5,622	5,622	67.65	68.75	64.13
Reddit TIFU	120,000	95,000	12,500	12,500	57.99	59.81	57.33

ate coherent summaries. Table 3.1 presents the dataset distributions and performance results across all models. To assess the effectiveness of paraphrasing attacks, we implement both translation and deletion methods for generating candidate adversaries. For the translation approach, we employ the T5 [16] transformer encoder-decoder model to translate input sentences to German and back to English. In the deletion method, we remove the $top-k$ candidates from the input text.

3.4.3 Experiment Settings and Evaluation Metrics

In the base setting for our experiment, 1000 samples were randomly selected from the train split of the dataset; the $top-k$ was finally set to 20% sentences with 5 sentences as an upper bound. We provide our code for reproducibility of the experiments[1].

We measure attack success by the decrease in ROUGE-1 scores when summarizing the modified documents. We also report ROUGE-2 and BLEU-1,2 scores, with their equations shown in Equation 3.2, where:

- p_n represents the ratio of matching n-grams between the generated and reference summaries to the total n-gram count in the generated summary.
- $w_n = \frac{1}{N}$ is the weight for each n-gram order, typically used to give equal weight to both unigram and bigram.

[1] https://github.com/UTSJiyaoLi/Summarization-Attack-via-Paraphrasing

- The brevity penalty ((BP)) penalizes summaries that are shorter than the reference summary. It is defined as:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - r/c) & \text{if } c \leq r \end{cases}$$

where c is the length of generated summary and r is the length of reference summary.

Additionally, the semantic similarity (sim) between original and adversarial sentences is evaluated using BERT [28] word vector comparisons.

$$\begin{aligned} \text{ROUGE-N} &= \frac{\sum_{s \text{ in reference}} \sum_{n\text{-grams in } s} \text{count of common n-grams}}{\sum_{s \text{ in reference}} \sum_{n\text{-grams in } s} \text{count of all n-grams}} \\ \text{BLEU} &= \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right) \\ \text{BP} &= \begin{cases} 1 & \text{if } c > r \\ \exp(1 - r/c) & \text{if } c \leq r \end{cases} \end{aligned} \tag{3.2}$$

3.4.4 Experiment Analysis

The comparative analysis across Pegasus, T5, and Bart demonstrates SAP's (ours) consistent superiority over other attack methods (AdSent, UAT, HotFlip, and TextFooler) across all datasets (XSum, CNN & Daily Mail, WikiHow, Multi-News, and Reddit TIFU), shown in Tables 3.2, 3.3, 3.4. SAP achieves the highest BLEU and ROUGE scores, indicating its effectiveness in generating adversarial samples that maintain high summarization quality while minimizing semantic deviation, as evidenced by its superior SIM scores. While TextFooler occasionally outperforms other baselines like AdSent and UAT, particularly on Pegasus and T5, it consistently falls short of SAP's performance. T5 shows higher vulnerability to UAT and HotFlip attacks, while Bart demonstrates greater resilience, yet SAP maintains robust performance across all models. These results confirm that

Table 3.2: Comparative analysis of attack effectiveness across datasets and baselines targeting Pegasus, where higher values indicate stronger performance.

Dataset	Attack Method	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	SIM
XSum	AdSent	17.87	16.21	23.95	24.63	65.46
	UAT	23.14	19.50	19.12	23.81	68.38
	HotFlip	19.08	17.48	19.08	14.77	66.22
	Textfooler	25.92	23.35	27.89	26.59	70.01
	SAP (ours)	**27.38**	**24.62**	**31.61**	**28.09**	**72.80**
CNN & Daily Mail	AdSent	19.87	18.31	20.45	17.62	65.54
	UAT	21.12	19.23	21.15	18.41	67.29
	HotFlip	20.03	18.72	20.58	17.89	66.43
	Textfooler	23.56	20.87	22.15	19.12	68.74
	SAP (ours)	**24.33**	**21.67**	**22.71**	**19.34**	**69.31**
WikiHow	AdSent	18.54	16.87	22.11	20.45	64.35
	UAT	20.12	17.92	23.45	21.08	66.14
	HotFlip	19.83	17.65	22.98	20.78	65.76
	Textfooler	22.45	19.87	25.12	23.54	68.21
	SAP (ours)	**23.89**	**21.12**	**26.34**	**24.12**	**69.89**
Multi-news	AdSent	20.23	18.45	24.12	22.03	65.87
	UAT	21.78	19.65	25.45	23.12	67.54
	HotFlip	21.03	18.92	24.56	22.65	66.89
	Textfooler	23.56	20.45	26.89	24.89	68.90
	SAP (ours)	**25.14**	**21.87**	**28.34**	**26.12**	**70.54**
Reddit TIFU	AdSent	17.45	15.89	21.78	20.21	63.21
	UAT	19.12	17.34	23.45	21.89	65.67
	HotFlip	18.76	16.98	22.89	21.45	64.89
	Textfooler	21.56	19.45	25.34	23.67	67.23
	SAP (ours)	**23.34**	**20.89**	**27.12**	**24.89**	**69.45**

SAP not only generates more effective adversarial samples but also better preserves semantic alignment with the original text compared to other methods. We listed two pairs of visualised examples in Table 3.5 that were generated by Pegasus. In Table 3.5, the attacked summary in the second sample introduces the phrase “it could be cursed ” - a new word from both the original document and our adversarial input. This demonstrates a critical vulnerability: our paraphrasing attack not only degrades summarization quality but actively induces the model to fabricate emotionally charged content. Such an unrelated generation of vocabulary that adversarial perturbations can fool the model’s

Table 3.3: Comparative analysis of attack effectiveness across datasets and baselines targeting T5, where higher values indicate stronger performance.

Dataset	Attack Method	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	SIM
XSum	AdSent	22.13	21.10	24.07	17.99	69.05
	UAT	24.92	21.43	30.97	21.93	72.49
	HotFlip	19.78	23.13	24.15	15.42	72.44
	Textfooler	20.74	22.47	28.51	21.49	70.06
	SAP (ours)	**26.34**	**25.37**	**32.01**	**24.15**	**74.96**
CNN & Daily Mail	AdSent	21.08	19.56	22.13	16.87	66.24
	UAT	22.41	20.12	24.59	18.75	67.89
	HotFlip	20.53	19.92	23.17	16.22	66.71
	Textfooler	23.42	21.06	25.12	19.98	69.21
	SAP (ours)	**24.76**	**22.11**	**27.43**	**21.51**	**71.38**
WikiHow	AdSent	19.87	18.12	21.49	19.06	65.45
	UAT	21.21	19.34	23.15	20.42	66.83
	HotFlip	20.14	18.89	22.51	19.56	65.92
	Textfooler	22.34	20.15	24.78	21.07	68.34
	SAP (ours)	**24.01**	**21.67**	**26.93**	**23.15**	**70.82**
Multi-news	AdSent	20.23	19.11	23.59	20.45	66.15
	UAT	21.76	20.14	24.61	22.11	67.53
	HotFlip	20.87	19.54	23.78	21.23	66.89
	Textfooler	23.03	21.12	25.92	23.09	69.24
	SAP (ours)	**25.54**	**22.34**	**28.13**	**25.02**	**71.65**
Reddit TIFU	AdSent	17.98	16.52	20.34	19.05	63.94
	UAT	19.45	17.89	22.11	20.21	65.62
	HotFlip	18.76	17.31	21.45	19.98	64.83
	Textfooler	21.15	19.14	24.31	22.45	67.32
	SAP (ours)	**23.67**	**21.41**	**26.92**	**24.58**	**69.94**

semantic understanding, causing it to project interpretations that diverge from any textual evidence while maintaining the source document's surface-level coherence.

3.4.5 Parameter Study

In this section, we analyze how the parameter *top-k* influences attack performance. We conduct experiments across five datasets with varying *top-k* settings and report the ROUGE-1 scores decreases in Figure 3.2. The results demonstrate that larger *top-k* values (representing the number of sentences attacked) more effectively expose model

Table 3.4: Comparative analysis of attack effectiveness across datasets and baselines targeting Bart, where higher values indicate stronger performance.

Dataset	Attack Method	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	SIM
XSum	AdSent	17.15	14.92	20.47	16.03	61.74
	UAT	21.76	16.45	24.03	19.92	63.92
	HotFlip	19.13	15.98	21.34	18.65	64.21
	Textfooler	20.89	18.42	25.48	22.12	65.84
	SAP (ours)	**28.56**	**23.87**	**33.14**	**29.92**	**72.45**
CNN & Daily Mail	AdSent	18.54	15.78	21.19	16.42	62.33
	UAT	19.97	16.85	23.14	18.74	63.21
	HotFlip	18.92	15.34	22.08	17.54	62.54
	Textfooler	22.13	19.45	24.58	20.96	65.11
	SAP (ours)	**27.43**	**23.12**	**30.87**	**27.45**	**71.02**
WikiHow	AdSent	17.34	15.09	20.12	17.13	60.84
	UAT	19.04	16.45	22.37	19.24	62.41
	HotFlip	18.23	14.87	21.12	18.03	61.39
	Textfooler	21.15	18.31	24.75	21.36	64.12
	SAP (ours)	**26.87**	**22.49**	**29.43**	**26.12**	**69.87**
Multi-news	AdSent	19.87	16.24	23.12	20.03	61.93
	UAT	20.34	17.14	24.05	21.34	63.04
	HotFlip	19.41	16.03	22.89	20.15	62.38
	Textfooler	22.18	19.41	25.34	23.07	65.87
	SAP (ours)	**28.45**	**23.76**	**31.78**	**28.12**	**72.23**
Reddit TIFU	AdSent	16.89	14.87	19.78	16.89	60.23
	UAT	18.15	16.12	21.54	18.94	61.92
	HotFlip	17.42	15.34	20.23	17.98	60.85
	Textfooler	20.09	18.23	23.67	21.42	63.74
	SAP (ours)	**25.93**	**21.54**	**28.78**	**26.24**	**69.42**

vulnerabilities. However, the ROUGE-1 score stabilizes when $top\text{-}k$ exceeds 20% of the document length. Based on this observation, we select 20% as our baseline setting to optimize both attack effectiveness and computational cost.

3.4.6 Ablation Study

This section presents a comprehensive ablation study examining different ranking strategies and attack methods, validating each component's contribution to our attack framework. We evaluate our importance ranking procedure against tf-idf and Textrank

Table 3.5: Comparison of original and adversarial contexts. The table highlights the differences between the original and adversarial contexts, as well as the corresponding abstractive summaries provided by the model before and after the attack.

Context	Scientists have made a groundbreaking discovery in Antarctica. A vast subglacial lake, named Lake Enigma, has been detected beneath 2 miles of ice, spanning 100 square miles, thought to be isolated from the surface for over 15 million years, is expected to harbor ancient microbial life. This finding could shed light on extraterrestrial life forms, as similar conditions exist on Jupiter's moon Europa. Researchers used radar imaging to confirm the lake's existence, and an expedition is planned to drill into the ice to study its contents.
Adversary	Scientists have made a groundbreaking discovery in Antarctica. Lake Enigma, a massive subglacial body of water spanning 100 square miles, has been discovered under 2 miles of ice and is believed to contain ancient microbial organisms. This finding could shed light on extraterrestrial life forms, as similar conditions exist on Jupiter's moon Europa. Researchers used radar imaging to confirm the lake's existence, and an expedition is planned to drill into the ice to study its contents.
Original Output	Scientists have discovered a massive subglacial lake in Antarctica, which could offer insights into ancient microbial life.
Attacked Output	Scientists discovered a subglacial lake in Antarctica but warned of risks in studying it due to contamination and unknown dangers.
Context	A historical treasure trove has been unearthed in the Scottish Highlands. Archaeologists discovered a Viking burial site containing weapons, jewelry, and a preserved longship, believed to date back to the 10th century. The burial site is thought to belong to a Viking chieftain and offers new insights into the Norse influence in Scotland. The artifacts, including a golden brooch and an elaborately decorated sword, are in remarkably good condition. Researchers hope this find will deepen understanding of Viking migration and cultural integration.
Adversary	A historical treasure trove has been unearthed in the Scottish Highlands. A Viking burial site, featuring weapons, jewelry, and a well-preserved longship, has been uncovered by archaeologists and is thought to originate from the 10th century. The burial site is thought to belong to a Viking chieftain and offers new insights into the Norse influence in Scotland. The artifacts, including a golden brooch and an elaborately decorated sword, are in remarkably good condition. Researchers hope this find will deepen understanding of Viking migration and cultural integration.
Original Output	Archaeologists have discovered a Viking burial site in Scotland, offering insights into Norse history.
Attacked Output	Archaeologists have discovered a Viking burial site in Scotland, with some researchers warning it could be cursed.

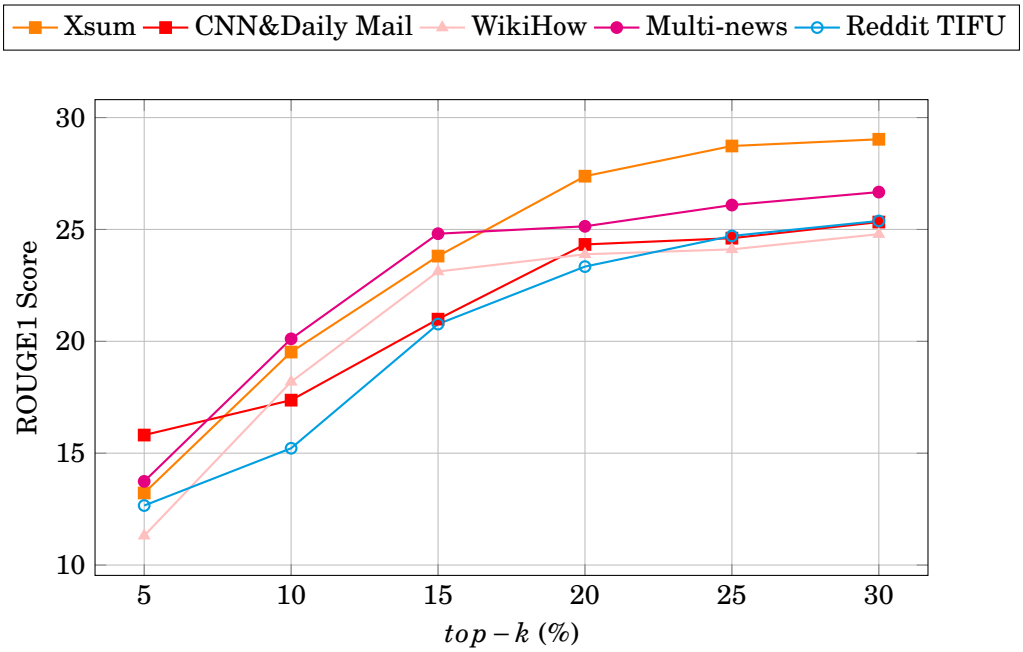

Figure 3.2: Impact of $top-k$ parameter settings on ROUGE-1 scores when attacking Pegasus model across five datasets.

baselines, while also comparing various approaches for generating substitution sentences. The evaluation uses Pegasus [15] on 1000 samples from the XSum test set under our base experimental configuration. As shown in Table 3.6, our approach which combines paraphrasing with the Pegasus ranking procedure yields optimal attack performance.

3.4.7 Transferability of Attacks

In this section, we evaluate the transferability of our attack scenarios across different Neural Machine Translation (NMT) models. We compare our SAP attack scheme against four baseline methods, testing their effectiveness on T5 [16], MarianNMT [176], and Bart [14], each specifically trained on the Commentary translation dataset for English-to-Arabic translation. As demonstrated in Figure 3.3, our SAP consistently outperforms other attack methods across all three translation models, indicating the strong transferability of our approach.

Table 3.6: Comparison of different importance ranking methods (tf-idf, Textrank, and Pegasus) across three attack strategies: translation, deleting, and paraphrasing. Higher values indicate better performance. "N/A" indicates metrics not applicable for the deletion attack method.

Attack Method	Ranking Method	Similarity (sim)	ROUGE
Translation	tf-idf	68.2	10.3
	Texrank	**76.5**	11.3
	Pegasus (ours)	75.8	14.3
Deleting	tf-idf	N/A	13.2
	Texrank	N/A	12.9
	Pegasus (ours)	N/A	17.8
Paraphrasing	tf-idf	67.8	13.3
	Texrank	71.5	12.0
	Pegasus (ours)	72.8	**18.4**

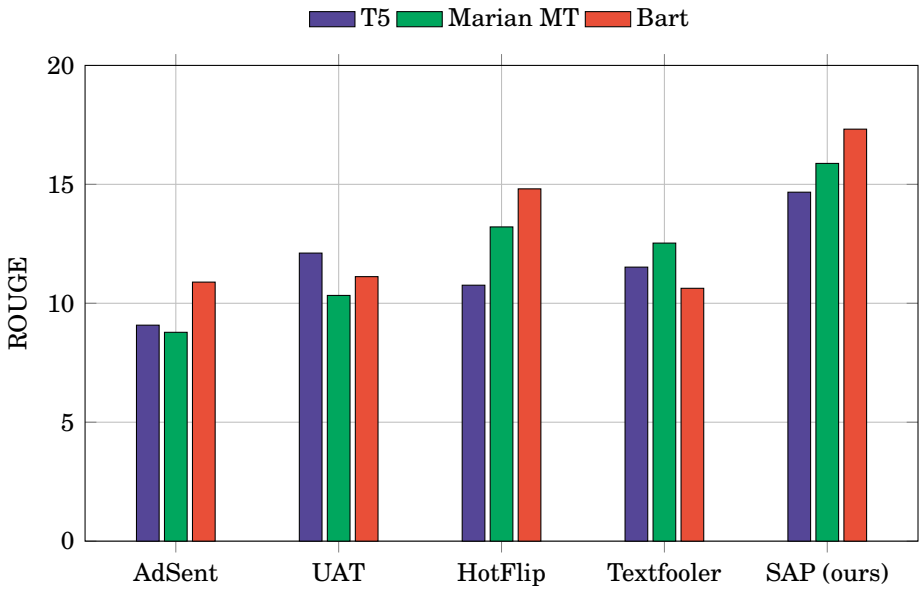

Figure 3.3: Outcomes of attacking NMT models (T5 Marian MT and Bart) across different attack methods. A higher ROUGE score indicates better performance.

3.4.8 Adversarial Retraining

This section explores how SAP can be leveraged to improve the performance of downstream models. We generate adversarial examples from Xsum training sets and integrate

them into the training data as augmentation. By reconstructing the training set with various proportions of adversarial examples combined with the original data, we investigate the relationship between test accuracy and adversarial content. Figure 3.4 reveals that incorporating adversarial examples moderately enhances model performance when they constitute less than 30% of the training data, beyond which performance deteriorates. This observation suggests that determining the optimal ratio of adversarial examples requires empirical validation, consistent with findings from prior attack methods. To assess the defensive benefits of adversarial training, we evaluate the robustness of Pegasus models trained with different proportions of adversarial examples (ranging from 0% to 40%) generated by various attack methods, as depicted in Figure 3.5. Using F1 score as an inverse indicator of model vulnerability to adversarial attacks, our results demonstrate that adversarial training consistently strengthens model robustness across all attack methods. SAP proves particularly effective in this context, achieving superior performance compared to alternative approaches, with its advantages becoming more pronounced as the proportion of adversarial training data increases.

3.5 Summary and Discussion

This chapter introduced Summarization Attack via Paraphrasing (SAP), a novel framework for adversarial attacks on document-level abstractive summarization models. Using sentence-level paraphrasing, we can craft adversarial examples that effectively mislead models while preserving semantic integrity and coherence. Our key contributions include introducing a ranking mechanism to identify critical sentences that influence summarization outputs, developing a paraphrasing-based attack strategy that generates fluent and semantically aligned adversarial samples, and demonstrating SAP,Äôs effectiveness through extensive evaluations of multiple datasets and models. Additionally, we showed SAP's strong transferability across NMT tasks and its ability to improve model robust-

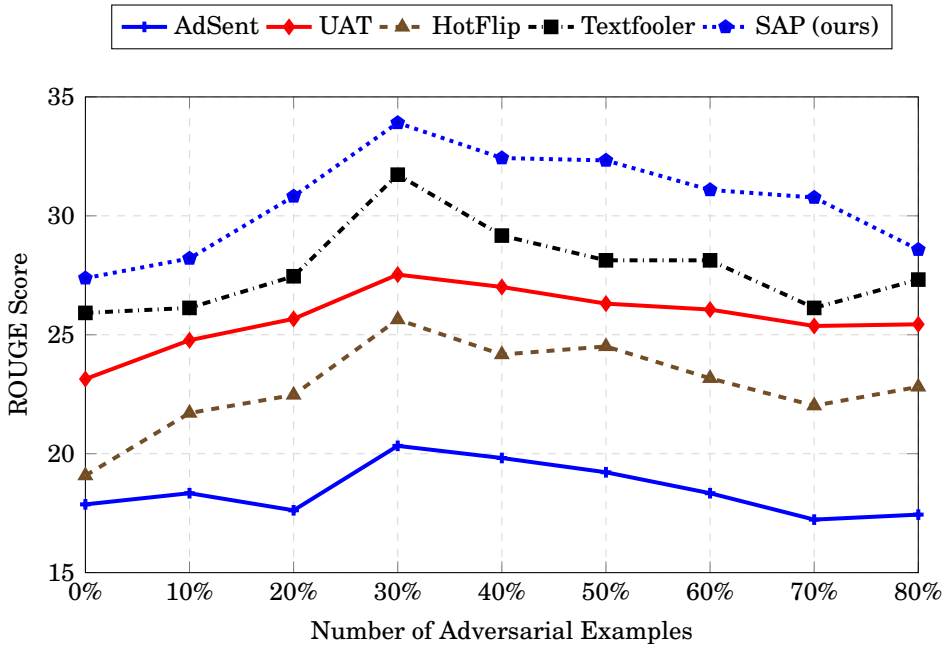

Figure 3.4: Pegasus model performance after retraining on Xsum dataset incorporating diverse adversarial examples from AdSent, UAT, HotFlip, Textfooler, and our novel SAP approach.

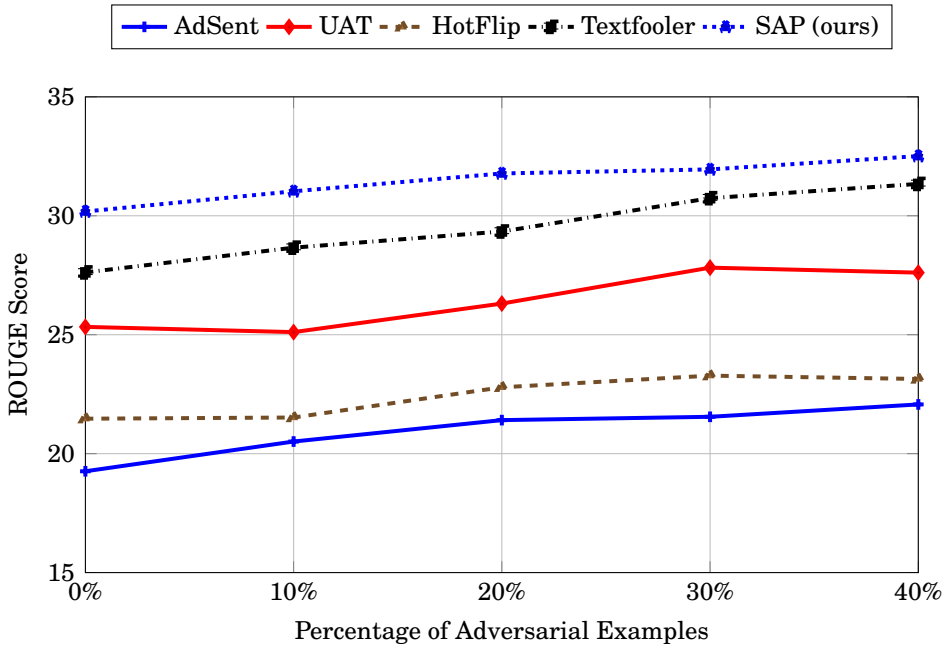

Figure 3.5: ROUGE score of attacking Pegasus models retrained with increasing proportions of adversarial examples generated by baseline methods (AdSent, UAT, HotFlip, Textfooler, and our SAP.

ness through adversarial retraining. These contributions provide new insights into the vulnerabilities of summarization models and offer valuable tools for enhancing their resilience. While SAP advances the state-of-the-art in summarization attacks, future work can address scalability, metric optimization, and real-world applicability, further solidifying its impact on developing robust NLP systems.

CHAPTER 4

AICAttack: Adversarial Image Captioning Attack with Attention-Based Optimization

Recent advances in deep learning research have shown remarkable achievements across many tasks in computer vision (CV) and natural language processing (NLP). At the intersection of CV and NLP is the problem of image captioning, where the related models' robustness against adversarial attacks has not been well studied. This section presents a novel adversarial attack strategy, AICAttack (Attention-based Image Captioning Attack), designed to attack image captioning models through subtle perturbations on images. Operating within a black-box attack scenario, our algorithm requires no access to the target model's architecture, parameters, or gradient information. We introduce an attention-based candidate selection mechanism that identifies the optimal pixels to attack, followed by a customized differential evolution method to optimize the perturbations of pixels' RGB values. We demonstrate AICAttack's effectiveness through extensive experiments on benchmark datasets against multiple victim models. The experimental results demonstrate that our method outperforms current leading-edge techniques by achieving consistently higher attack success rates.

4.1 Introduction

In recent years, deep learning models, particularly Convolutional Neural Networks (CNNs), have showcased remarkable achievements across diverse computer vision tasks, notably image classification. These models have attained human-level or surpassed human performance [21], thus opening avenues for their practical integration into real-world applications. Nevertheless, this swift advancement has brought to light a critical vulnerability inherent in these models - their susceptibility to adversarial attacks.

In computer vision tasks, adversarial attacks aim to introduce meticulously crafted perturbations to input images, thereby causing models to yield erroneous or misleading predictions [107]. These perturbations can profoundly influence model outputs despite being imperceptible to human observers. Adversarial image attacks predominantly target tasks rooted in CNNs, with image classification as common examples [106, 107, 110].

In practice, attackers can use these methods to fool content filters by making violent or sexual images appear harmless to AI systems [177]. Self-driving cars face similar risks-small changes to camera inputs could make the AI misread stop signs or pedestrians, leading to dangerous driving decisions [178]. These vulnerabilities underscore the imperative of developing robust image captioning architectures for safety-critical deployments, where model failures translate directly to human harm [179].

Examining attacks in image classification tasks from an input perspective, the conventional approach involves injecting perturbations into the original image to prompt the model to generate an incorrect classification label. Computations involving gradients often play a crucial part in directing attacks aimed at image classification problems. In the context of white-box attacks, access to the model's gradient is feasible, allowing researchers to derive perturbations by minimizing the redefined objective function [2]. Image captioning represents a closely related domain, encompassing the generation of coherent and intelligible captions for images through meticulous analysis. The predom-

inant attacking methodology to captioning entails the deployment of CNNs to extract image features and RNNs (Recurrent Neural Networks) to formulate descriptive captions [22, 180–182], which is commonly denoted as an Encoder-Decoder architecture.

Crafting adversarial attacks against image captioning models poses unique challenges that surpass those in image classification attacks. These difficulties primarily arise from the complexities of leveraging gradients within the Encoder-Decoder framework [152]. The field's main hurdles can be distilled into two key issues: the impracticality of using internal model information for attacks, and the intricacy of accurately evaluating attack effectiveness on generated text [138, 183]. Most existing studies on adversarial attacks targeting image captioning systems have concentrated on white-box scenarios [184], assuming complete attacker knowledge of the model's architecture and parameters. However, this assumption often proves unrealistic in practical settings, where attackers rarely have comprehensive access to the target model's inner workings. To address this disconnect between theoretical assumptions and real-world applications, we introduce a novel approach that better reflects authentic adversarial conditions.

Our proposed methodology, AICAttack, integrates an attention mechanism to precisely identify and target the most susceptible pixels in an image for adversarial manipulation. We then utilize a differential evolution algorithm to optimize the attack's effectiveness, ensuring that the generated adversarial samples are impactful and plausible. This strategy not only tackles the practical constraints of previous methods but also enhances the viability and applicability of adversarial attacks in image captioning.

Our work presents the following major contributions:

- We present AICAttack, an adversarial attack method employing an attention mechanism to accurately locate the pixels most critical to caption generation. This strategy enables us to target our efforts on areas with the greatest potential to influence captions, thus enhancing attack efficiency without relying on gradient

computations.

- A differential evolution algorithm further refines our approach. This algorithm is customized to precisely adjust adversarial modifications on the identified key pixels, ensuring the alterations are both imperceptible and highly effective.

- We perform extensive testing across diverse real-world datasets, targeting various image captioning models as potential victims. These comprehensive experiments demonstrate the efficacy of our approach in creating adversarial examples that effectively undermine the reliability of image captioning systems.

The subsequent sections of this chapter are structured as follows. Section 4.2 reviews the fundamental concepts of visual attention and the differential evolution algorithm. Section 4.3 presents our proposed AICAttack methodology in detail. We then evaluate AICAttack's performance through well-structured experiments in Section 4.4. Finally, Section 4.5 concludes the chapter by discussing our findings and future works.

4.2 Preliminary

In this section, we present the foundational concepts and methodologies that are integral to understanding the AICAttack framework. Specifically, we introduce the visual attention mechanism, which guides our region selection strategy, and the differential evolution algorithm, which serves as the core optimization engine for generating adversarial examples.

4.2.1 Visual Attention

Visual attention mechanisms have emerged as a critical component in modern computer vision systems, enabling models to dynamically prioritize relevant spatial or semantic

regions in input images. Inspired by the human visual system, these mechanisms simulate the way humans focus selectively on important visual stimuli while ignoring less relevant background information. This capacity is particularly advantageous in dense visual scenes where salient information may be distributed across multiple locations.

In the context of image captioning, attention mechanisms enhance the quality of generated captions by assigning weights to different spatial regions of the image at each decoding time step. This allows the model to focus sequentially on different objects or areas when generating each word, resulting in more descriptive and contextually appropriate captions. Typical implementations include spatial attention, where attention weights are applied over feature maps, and channel-wise attention, which re-weights feature channels to emphasize specific semantic patterns. Hybrid attention strategies combine these dimensions to simultaneously model spatial and feature-level importance.

Transformer-based architectures, such as the Vision Transformer (ViT), further generalize attention by treating image patches as sequences, allowing global dependencies to be captured across the entire visual field. These attention-based designs have been widely adopted in state-of-the-art image captioning models due to their superior ability to generate semantically rich and fluent sentences.

In adversarial settings, visual attention provides a principled way to identify vulnerable and semantically critical regions for targeted perturbation. Rather than applying random or uniform noise across the image, attacks guided by attention can inject perturbations into areas most influential to the model‚Äôs decision-making process, thereby enhancing attack success while maintaining imperceptibility. In AICAttack, we utilize pre-extracted attention maps from the captioning model to locate these sensitive regions. By doing so, our attack framework narrows the search space and amplifies perturbation impact without significantly altering benign image perception.

4.2.2 Differential Evolution Algorithm

Differential Evolution (DE) is a population-based stochastic optimization algorithm introduced by Storn and Price [147], designed to solve complex optimization problems over continuous, high-dimensional, and multimodal landscapes. DE has gained popularity in adversarial machine learning due to its strong global search capabilities, simplicity, and gradient-free nature, making it suitable for both white-box and black-box attack scenarios.

The algorithm maintains a population of candidate solutions, each represented as a vector in the search space. During each iteration, DE applies three core operations: mutation, crossover, and selection. In the mutation phase, new candidate solutions are generated by adding the weighted difference of two randomly chosen vectors to a third vector. The crossover step combines the mutated vector with the original one to introduce diversity. In the selection phase, the candidate with better fitness-defined by an objective function-is retained for the next generation. This process continues until convergence or a predefined stopping criterion is met.

In adversarial attack settings, DE is particularly effective when the loss landscape is non-smooth or when gradient information is unavailable, such as in black-box models or models with non-differentiable components. Its population-based design allows for parallel exploration of multiple perturbation trajectories, increasing the likelihood of discovering successful adversarial examples with minimal perturbation.

AICAttack leverages DE to search for optimal perturbations constrained to selected high-attention regions of the image. Each candidate solution corresponds to a set of pixel-level modifications localized to these regions. The fitness function incorporates both the deviation between the original and adversarial captions and a perceptual distance metric to ensure visual similarity. Compared to gradient-based methods, DE offers greater flexibility and robustness, especially when targeting sequence generation tasks like

image captioning, where gradient computation is complicated by discrete output spaces and attention alignment.

By integrating visual attention with differential evolution, AICAttack constructs a guided and efficient adversarial pipeline that selectively perturbs semantically meaningful image regions through a powerful optimization strategy, yielding effective and imperceptible attacks against state-of-the-art image captioning models.

4.3 Proposed Method

This section elaborates on our proposed image captioning attack method, AICAttack (Attention-based Image Captioning Attack).

4.3.1 Problem Setting

Given a pre-trained image captioning model $F(\cdot) : \mathscr{X} \rightarrow \mathscr{Y}$, where $\mathscr{X}$ represents the image feature space and $\mathscr{Y}$ represents the textual space, an attacker seeks to generate an adversarial image $\mathbf{x}'$ by manipulating an existing image $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_P)$, where $\mathbf{x} \in \mathscr{X}$, $x_i \in \mathbb{R}$, and P is the number of pixels. The objective is to deceive the performance of $F(\cdot)$ such that $F(\mathbf{x}')$ does not match the ground truth $y \in \mathscr{Y}$.

To craft an adversarial example, a perturbation $\Delta\mathbf{x}$ is added to the image $\mathbf{x}$, resulting in the construction of the adversarial example $\mathbf{x}'$ as follows:

$$\begin{aligned} \mathbf{x}' &= \mathbf{x} + \Delta\mathbf{x}, \\ \text{where} \quad \Delta\mathbf{x} &= (\Delta x_1, \ldots, \Delta x_i, \ldots, \Delta x_P). \end{aligned} \tag{4.1}$$

In the above context, attacking the victim model F involves the process of searching for $\Delta\mathbf{x}$. To construct such a perturbation $\Delta\mathbf{x}$, the attacker first identifies m (where $m \leq P$) pixels using indices $\mathbf{I} = \{I_j\}_{j=1}^{m}$, and then optimizes the corresponding perturbation

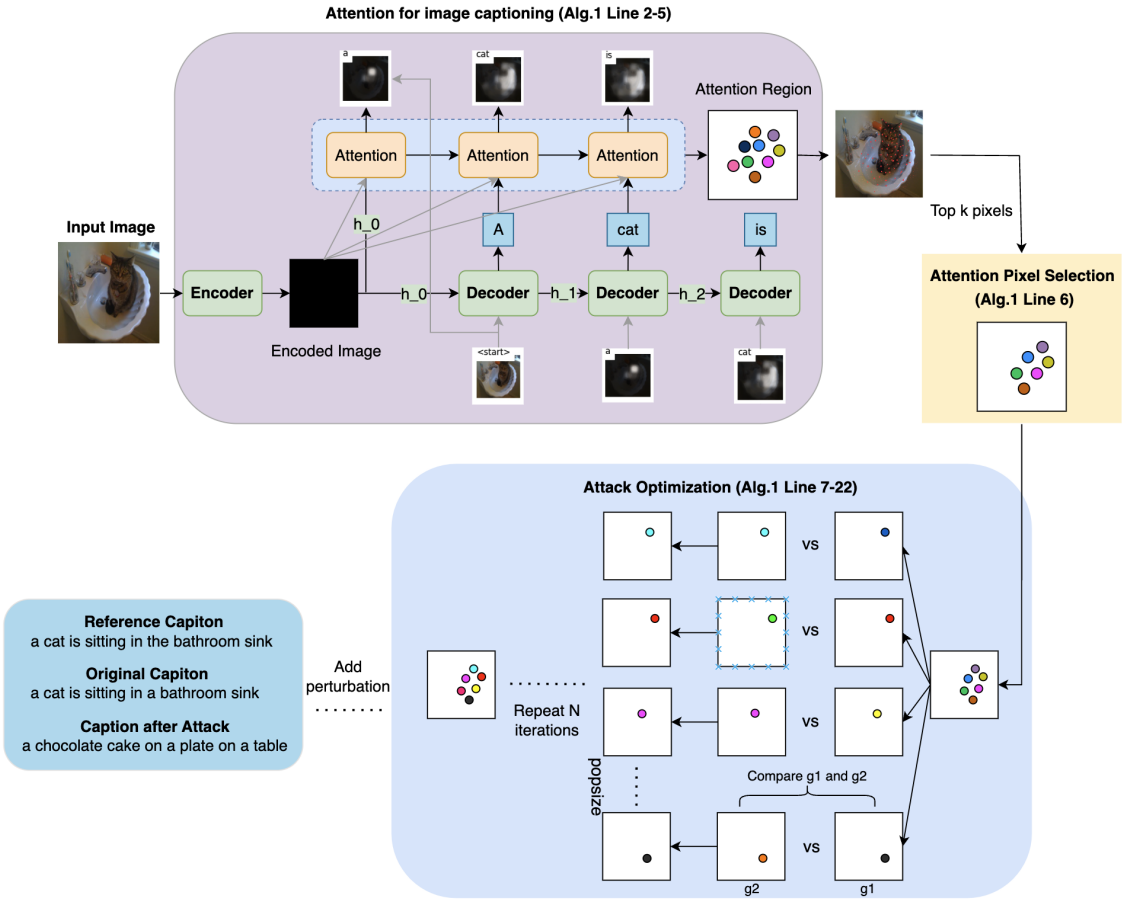

Figure 4.1: The Workflow of our AICAttack Algorithm for Image Captioning Attacks. The process begins by feeding the input image into the attention block, which generates attention scores. These scores are then used for attention pixel selection. During the attack optimization phase, the Differential Evolution (DE) algorithm searches for the most effective adversarial sample.

values $\{\Delta x_{I_j}\}_{I_j \in \mathbf{I}}$. Additionally, for the unaltered pixels, their perturbation values are set to 0 ($\Delta x_h = 0$ for $h \notin \mathbf{I}$).

In the following sections, we introduce our AICAttack algorithm in detail. An illustration figure of the AICAttack process is shown in Figure 4.1. An input image goes through an attention layer to generate attention scores. Based on that, an attack optimization step is implemented to generate the optimal adversarial example. Diverging from conventional algorithms, we transform this task into a problem of discovering the optimal solution within a specified region. Therefore, we have two main tasks to accomplish: (1)

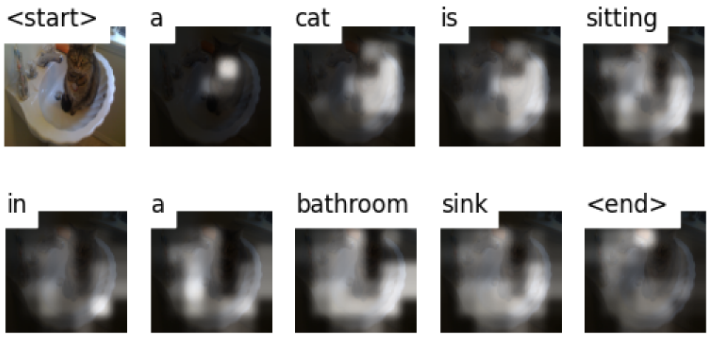

Figure 4.2: Attention Mechanism Illustration in a Small Cat Image Example. Highlighted regions denote attention concentration guiding the encoder-decoder network during word generation processes.

select the optimal locations of pixels to be attacked. (which we detail in Sec. 4.3.2), and (2) determine the optimal perturbation values for the selected pixels. (which we discuss in Sec. 4.3.3).

4.3.2 Attention for Candidate Selection

Attention-based networks enable models to choose only the parts of the encoded variables relevant to the task encountered. Bahdanau [68] used it to address the challenge of handling long-range dependencies in lengthy textual sequences within natural language processing. As a type of soft attention, it uses a learned attention function to compute attention weights for each element in the input sequence. The same mechanism can be used in other models where the encoder's output has multiple points in space or time. In image captioning, specific pixels are usually assigned higher importance than others. We consider these high-importance pixels as potential candidate regions to be attacked.

In this research, we extracted the attention score α derived from an image captioning network, such as Show, attend and tell (SAT) [129]. For each pair of input images and generated captions with a length of l, we derive attention mappings for each individual word in the caption. Figure 4.2 shows a visual example of the attention mapping for each word. Given our use of soft attention, where there are P pixels, and the pixel weights

Figure 4.3: Examples of "Sentence-based Attack" (our proposed method) and "Word-based Attack" approaches for computing attention scores. The highlighted red areas represent the region for pixel selections.

sum up to 1, for each word token t, we have: $\sum_1^P \alpha_{p,t} = 1$, where $\alpha_{p,t}$ is the attention score of pixel p for the word token t.

Two methodologies were used to utilise the attention scores we obtained. As an example displayed in Figure 4.3, "Sentence-based Attack" involves aggregating attention scores for all pixels of all words. This results in attention mapping that matches the dimensions of the original image. Subsequently, scores are ranked, and the pixel coordinates of the top k, where $k \leq P$, attention values are selected to form the candidate region. Our AICAttack algorithm is shown in Alg. 2, while the candidate region formulation is shown in line 6 of the algorithm. The second approach is referred to as "Word-based Attack", which can be considered a baseline approach. In this case, we only focus on the top k pixels ranked by attention scores for each word and then join these regions to form candidate regions.

4.3.3 Differential Evolution Optimization

After selecting the targeted pixels, our next step is to derive the best magnitudes of attacks on the pixels. We want the magnitude of the attack to be as small as possible while its impacts on the generation of captions as much as possible. To do so, we optimize the pixel perturbation values to decrease the caption's quality measured by BLEU. To this end, we apply Differential Evolution (DE) [116], a robust evolutionary algorithm, to solve this optimisation problem. This optimisation method maintains a population of

Algorithm 2: AICAttack: Adversarial Image Captioning Attack

Input: Captioning model $F(\cdot)$, image $\mathbf{x}$, attention network A, number of pixels P, attention region size k, population size $popsize$, iteration time T, attacking strength s, BLEU score calculation function B. ;
Output: Optimal adversarial sample $\mathbf{x}'$. ;
// **Attention for Candidate Selection**;
Compute attention scores: $\alpha \leftarrow A(I)$, $\alpha^* \leftarrow [\,]$;
foreach x *in* α **do**
    $\alpha^* \leftarrow \alpha^* + x$;
end
Pick top-k pixels from α^*;
// **Differential Evolution Optimization**;
for $i = 1$ **to** $popsize$ **do**
    Construct $\mathbf{x}^i$ where pixel locations and changes are determined by attention weights and attack strength, respectively;
end
$\mathbf{x}^0 \leftarrow \mathbf{x}$, $\mathbf{x}' \leftarrow \mathbf{x}^0$;
for $g = 1$ **to** T **do**
    for $j = 1$ **to** $popsize$ **do**
        Build $\mathbf{x}_j^g$ from the previous generation $\mathbf{x}^{g-1}$ using mutation;
        if $B(F(\mathbf{x}_j^g)) < B(F(\mathbf{x}'))$ **then**
            if $B(F(\mathbf{x}_j^g)) < B(F(\mathbf{x}_j^{g-1}))$ **then**
                $\mathbf{x}' \leftarrow \mathbf{x}_j^g$;
            end
        end
    end
end
return The best attack example $\mathbf{x}'$;

candidate solutions, often called individuals. The key idea behind DE is the differential mutation operator, which involves creating new candidate solutions by perturbing the difference between two other solutions from the population. New solutions are generated through mutation operation, and their fitness is evaluated. If a new solution outperforms its parent, it replaces the parent in the population.

In this research, we customize DE to obtain an optimal solution by finding the best pixel coordinates and RGB values to attack a given input image. Each candidate solution's perturbation encompasses the coordinates/locations of pixels and the changes

in the pixels' RGB values. In our configuration, the initial count of candidate solutions (population) is set to $popsize$ (which is a parameter that can be changed to adapt to different applications and scenarios). Accordingly, by the DE algorithm, every new iteration generates $popsize$ new candidate solutions (children candidates) according to Equation 4.2:

$$\begin{aligned} &\mathbf{x}_j^g \leftarrow \mathbf{x}_{r_1}^{g-1} + \lambda \cdot (\mathbf{x}_{r_2}^{g-1} - \mathbf{x}_{r_3}^{g-1}) \\ &where \quad r_1 \neq r_2 \neq r_3 \end{aligned} \tag{4.2}$$

where $\mathbf{x}_j^g$ is the candidate solution, g and j represent the indices of generation and the mutant in population, respectively. λ is a parameter for candidates weight balancing and r_1, r_2, r_3 are random positive integers.

After the attention-based candidate selection (lines 3 to 7 of Alg. 2), the algorithm initializes a population of candidate solutions (lines 9 to 13 of Alg. 2), where each solution represents a perturbed image. The DE algorithm then iteratively updates these solutions by performing differential mutation and crossover operations (lines 14 to 23 of Alg. 2). For each generation, the algorithm evaluates the fitness of candidate solutions using the BLEU score calculated with its predicted caption and compares it to the previous generation. If a candidate solution yields a lower captioning BLEU score (indicating success in fooling the victim model), it is selected as the new adversarial example.

4.4 Experiment Results and Analysis

In this section, we comprehensively evaluate the performance of our method against the current state of the art. Besides the main results of attack performance and imperceptibility (Sec. 4.4.4), we also conduct experiments on ablation studies (Sec. 4.4.5), transferability (Sec. 4.4.6), adversarial retraining (Sec. 4.4.7). We provide code for repro-

ductivity of our experiments[1].

4.4.1 Datasets

Our experiment was conducted on the COCO [185] and Flickr8k [186] datasets. Each image in the COCO dataset is accompanied by five human-generated captions, providing rich linguistic annotations that describe the visual content with varying levels of detail and perspectives.

The Flickr8k dataset is sourced from the Flickr image-sharing platform. It includes 8,000 images, each accompanied by five distinct captions, resulting in 40,000 captions.

4.4.2 Victim Models and Baselines

We use two victim image captioning models with leading-edge performance to examine our attacking algorithm. They are “Show, Attend, and Tell" (abbreviated as “SAT") [129] and “BLIP" [22]. For image captioning attack baselines, we chose “Show and Fool" [151] and “GEM" [152]. We also finetuned “One Pixel Attack" [116], initially designed for fooling image classification models to generate adversarial images for comparison to our approach.

4.4.3 Metrics

To examine and measure the performance of the attacks, we reported the attack performance of different methods using several metrics.

4.4.3.1 BLEU Score

BLEU (Bilingual Evaluation Understudy) score [161] is a commonly used metric in the evaluation of natural language processing systems, including image captioning. In our

[1]We provide our code in an anonymous setting for review: `https://github.com/UTSJiyaoLi/Adversarial-Image-Captioning-Attack`.

Table 4.1: The performance of two baseline victim models tested on COCO and Flickr8k datasets.

		BLEU1	BLEU2	BLEU3	BLEU4
SAT	COCO	71.8	50.4	36.7	25.0
	Flickr8k	67.0	45.7	31.4	21.3
BLIP	COCO	73.1	48.9	38.2	26.6
	Flickr8k	70.1	47.2	32.5	22.8

experiment, we used BLEU-1 and BLEU-2 (unigram- and bigram-based BLEU scores) and BLEU-4 to deal with four-word phrases for longer captions. The equation of BLEU-4 is shown in Equation 4.3:

$$\text{BLEU-4} = \text{BP} \times \exp\left(\frac{1}{4}\sum_{n=1}^{4}\log\left(\text{precision}_n\right)\right), \tag{4.3}$$

where $\text{precision} = \frac{Number\ of\ correct\ word\ tokens\ generated}{Number\ of\ total\ word\ tokens\ generated}$. Sometimes, candidates might be very small for longer captions and missing important information relative to the reference. So we include the Brevity Penalty (BP) to penalize predicted captions that are too short compared to the reference captions. The BP is defined as following:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

, where r refers to the length of the original caption, and c refers to the length of the generated caption.

4.4.3.2 ROUGE Score

By using BLEU only, it may not fully capture human language's semantic and contextual nuances, and it might not always correlate perfectly with human judgment. Hence, we also report ROUGE-n [25], which measures the number of matching n-grams between the model-generated captions and human-produced/ground-truth reference captions. Our

experiments used unigram and bigram ROUGE scores (i.e., ROUGE-1 and ROUGE-2).

$$\text{ROUGE-n} = \frac{\sum_{S \in \text{ReferenceCaptions}} \sum_{\text{n-gram} \in S} \text{Count}_{\text{match}}(\text{n-gram})}{\sum_{S \in \text{ReferenceCaptions}} \sum_{\text{n-gram} \in S} \text{Count}(\text{n-gram})} \tag{4.4}$$

4.4.3.3 BR-measure

Besides reporting BLEU and ROUGE individually to comprehensively represent the results, we introduced a new measure to simplify the process of evaluating attack results, which integrates the ROUGE and BLEU scores in a way similar to driving F-measure from precision and recall, which we call BR-measure:

$$\text{BR-measure} = \frac{BLEU * ROUGE}{BLEU + ROUGE} \tag{4.5}$$

The BR-measure has a desirable property where the value of the BR-measure is high if only both BLUE and ROUGE values are high. The BR-measure will be low if the BLUE or the ROUGE value is low.

Before examining our AICAttack, we present the baseline captioning performance of two target models evaluated on both COCO and Flickr8k datasets in Table 4.1.

4.4.4 Experiment Analysis

Our experiments were conducted by the following settings (the evaluation of the tuning of these parameters is studied in the later part of this section): attention region k at 50% (attacking pixels whose attention weight is above the median of all weights), with ±50 range for s (i.e., modify at most 50-pixel intensity values for each attack pixel), parameter λ is set to be 0.5 and targeting 500 to 1000 random image-caption pairs from the test dataset. For the “SAT" model, we attack 500 pixels. Considering the larger image size in the “BLIP" model, we extended the attack to 1000 pixels.

We demonstrate some attack outcomes in Figure 4.4. The results in Tables 4.2 and 4.3 highlight our attack strategies’ effectiveness across models. Notably, our “AICAttack" (i.e.,

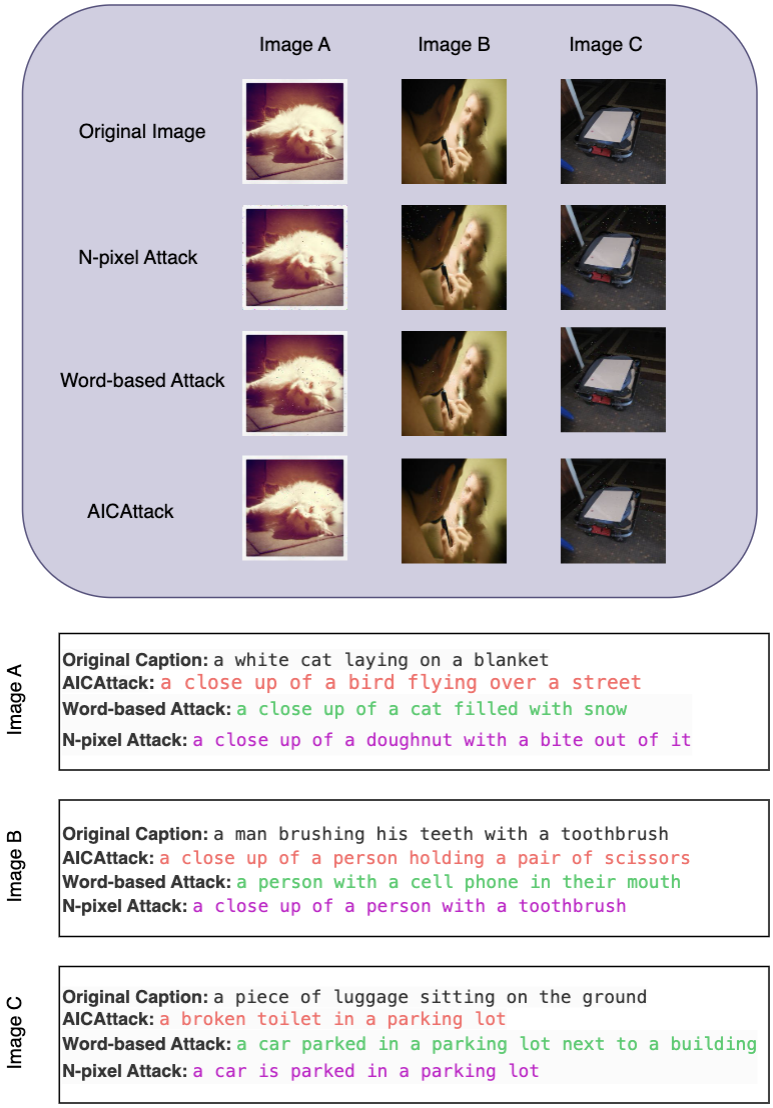

Figure 4.4: Visual examples illustrating different attack strategies, accompanied by captions.

the “Sentence-based Attack") method outperformed baseline approaches. Particularly, our AICAttack outperformed GEM and Show and Fool, which revealed the effectiveness of our “Sentence-based Attack" work. The comparison between “Sentence-based Attack" and “Word-based Attack" methods exposed a more pronounced decidable in the former. This distinction arises from the “Word-based Attack" approach’s attention selection, contrasting the more focused nature of “Sentence-based" selection. The sentence exemplifies this distinction “a cat is sitting in a bathroom sink," wherein “Word-based Attack" at-

tends to “a", “in" and “is". Consequently, a broader region from the image, encompassing non-significant elements, was incorporated into the candidate region. As visualisation in Figure 4.2, Word-based selection indicates a broader scope, incorporating boundary pixels. Conversely, the attention region depicted in Figure 4.3 for Sentence-based selection is more confined, centering exclusively on pertinent entities.

4.4.5 Ablation and Hyperparameters Studies

We introduce ablation experiments to validate the effectiveness of our AICAttack method, apart from employing baselines to substantiate attention. As shown in Figure 4.5. The figure displays the variation in BLEU2 scores under five attack methods across different pixel counts. Firstly, compared to other baselines, it can be observed that our approach consistently maintains the optimal performance under extreme conditions (attack fewer pixels), our method consistently maintains the optimal performance. This is attributed to attention and weight selection capabilities that facilitate the choice of the most optimal pixels for attack. Furthermore, we can observe that the attention method combined with weight outperforms the attention-only approach.

4.4.5.1 Number of iterations in genetics optimization versus attacking performance

To examine the impact of iteration numbers on performance, we subject 1000 samples from the COCO test set to our AICAttack method under various iteration configurations. The results are shown in Figure 4.6. The attack performance notably improves when increasing from three to five iterations. However, beyond five iterations, up to ten, the BLEU 2 score fluctuates, indicating that our methodology achieves stable performance with iterations exceeding five.

4.4.5.2 Candidate Region Analysis

We examined the impact of varying candidate region k sizes on experimental outcomes, shown in Figure 4.7. Note that this consideration differs from pixel count.

Table 4.2: The table presents the outcomes of our attack methods applied to BLIP with 1000 randomly selected samples from the COCO and Flickr8k datasets. All measures in the table denote the differences before and after the attacks (i.e., the value dropped after the attacks). N-Pixel Attack randomly selected pixels without using attention. Optimal outcomes are denoted in bold.

Datasets	Methods	Value Dropped After the Attack (Higher is Better)					
		BLEU1	BLEU2	BLEU4	ROUGE1	ROUGE2	BR
COCO	One Pixel Attack	0.005	0.003	0.002	0.002	0.004	0.001
	Show and Fool	0.054	0.086	0.033	0.006	0.038	0.005
	GEM	0.051	0.093	0.048	0.006	0.034	0.004
	N-Pixel Attack	0.059	0.073	0.062	0.006	0.020	0.032
	Word-based Attack	**0.066**	0.074	0.053	0.005	0.021	0.032
	AICAttack (**ours**)	0.060	**0.104**	**0.066**	**0.008**	**0.041**	**0.005**
Flickr8k	One Pixel Attack	0.003	0.002	0.004	0.004	0.001	0.002
	Show and Fool	0.048	0.08	0.038	0.007	0.025	0.031
	GEM	0.053	0.079	0.041	0.007	0.028	0.031
	N-Pixel Attack	0.053	0.073	0.044	0.006	0.024	0.023
	Word-based Attack	**0.057**	0.071	0.048	0.006	0.022	0.032
	AICAttack (**ours**)	0.057	**0.081**	**0.052**	**0.007**	**0.028**	**0.033**

Table 4.3: The table presents the outcomes of our attack methods applied to SAT with 1000 randomly selected samples from the COCO and Flickr8k datasets. All measures in the table denote the differences before and after the attacks (i.e., the value dropped after the attacks). N-Pixel Attack randomly selected pixels without using attention. Optimal outcomes are denoted in bold.

Datasets	Methods	Value Dropped After the Attack (Higher is Better)					
		BLEU1	BLEU2	BLEU4	ROUGE1	ROUGE2	BR
COCO	One Pixel Attack	0.009	0.004	0.003	0.005	0.002	0.002
	Show and Fool	0.119	0.163	0.109	0.037	0.045	0.063
	GEM	0.120	0.177	0.083	0.039	0.049	0.061
	N-Pixel Attack	0.125	0.179	0.130	0.074	0.070	0.064
	Word-based Attack	0.117	0.170	0.127	0.070	0.067	0.062
	AICAttack (**ours**)	**0.127**	**0.188**	**0.131**	**0.075**	**0.074**	**0.065**
Flickr8k	One Pixel Attack	0.004	0.003	0.002	0.004	0.003	0.004
	Show and Fool	0.104	0.153	0.014	0.030	0.037	0.043
	GEM	0.102	0.151	0.013	0.033	0.048	0.045
	N-Pixel Attack	0.101	0.166	0.010	0.032	0.061	0.043
	Word-based Attack	0.103	0.168	0.107	0.048	0.066	0.045
	AICAttack (**ours**)	**0.114**	**0.175**	**0.108**	**0.053**	**0.068**	**0.049**

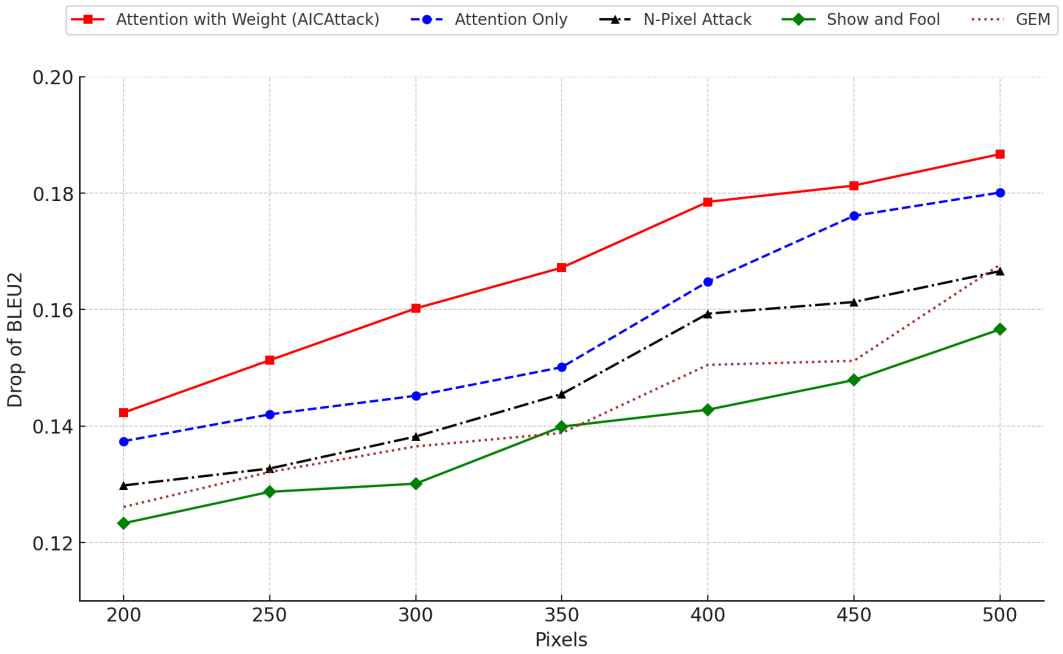

Figure 4.5: Drops of BLEU2 scores before and after five attack scenarios across different pixel counts.

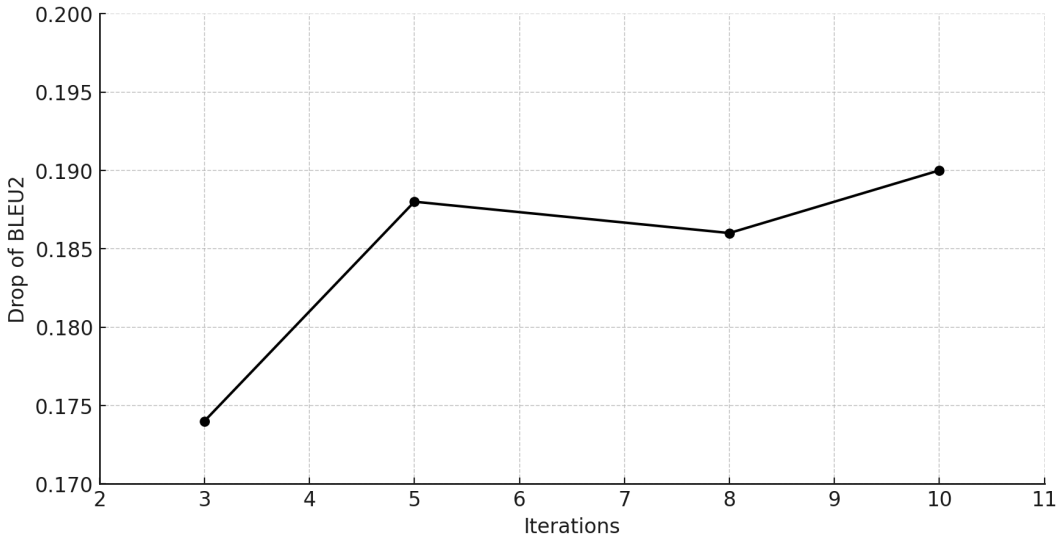

Figure 4.6: Drops in BLEU2 scores across varying iteration counts in the differential evolution algorithm.

The value of k determines the initial size of the candidate region, while the number of pixels dictates how many are selected for attacks within this area. As shown in the graph, fluctuations in both directions occur as k varies, but the overall trend is predominantly

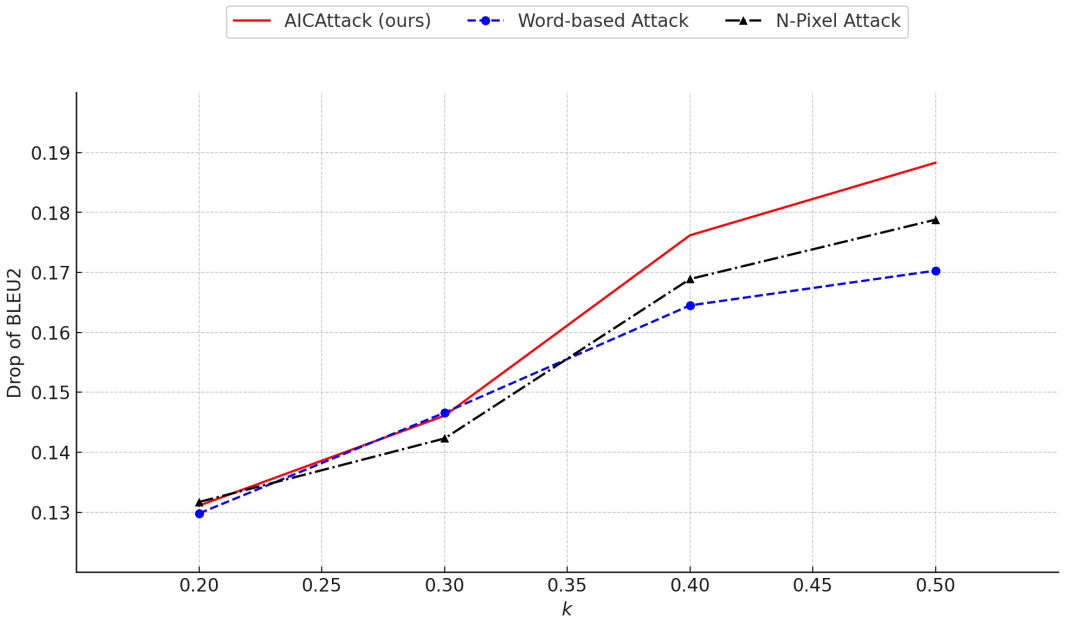

Figure 4.7: Drops of BLEU2 scores before and after attack when applying multiple attention regions k.

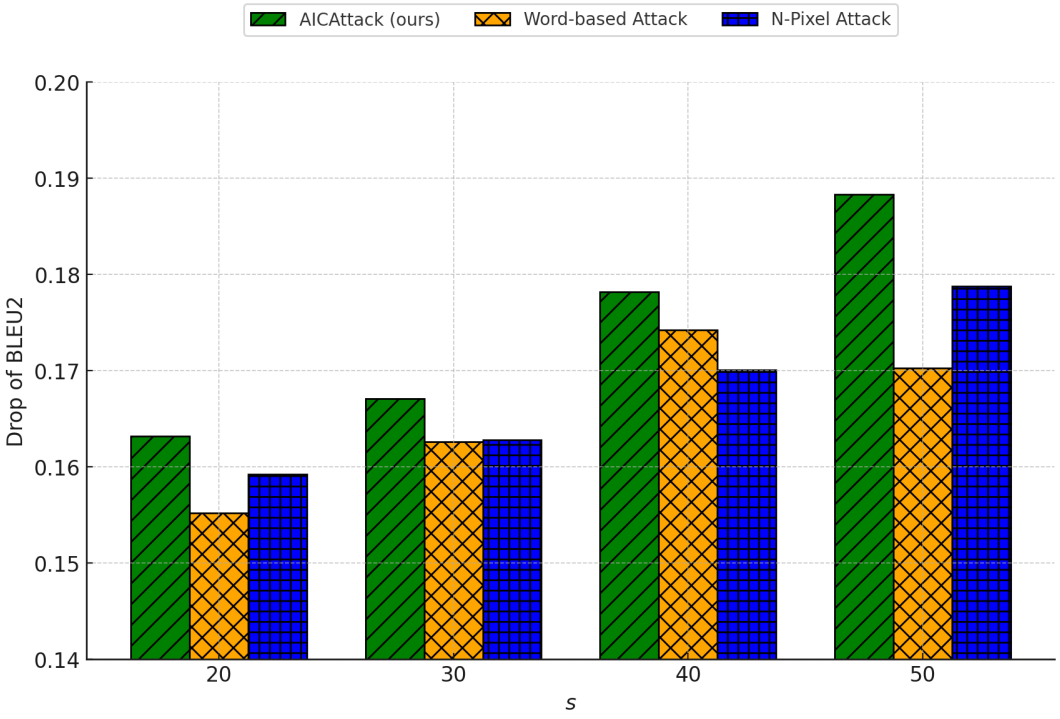

Figure 4.8: Drops of BLEU2 scores before and after attack when applying multiple attack strengths s.

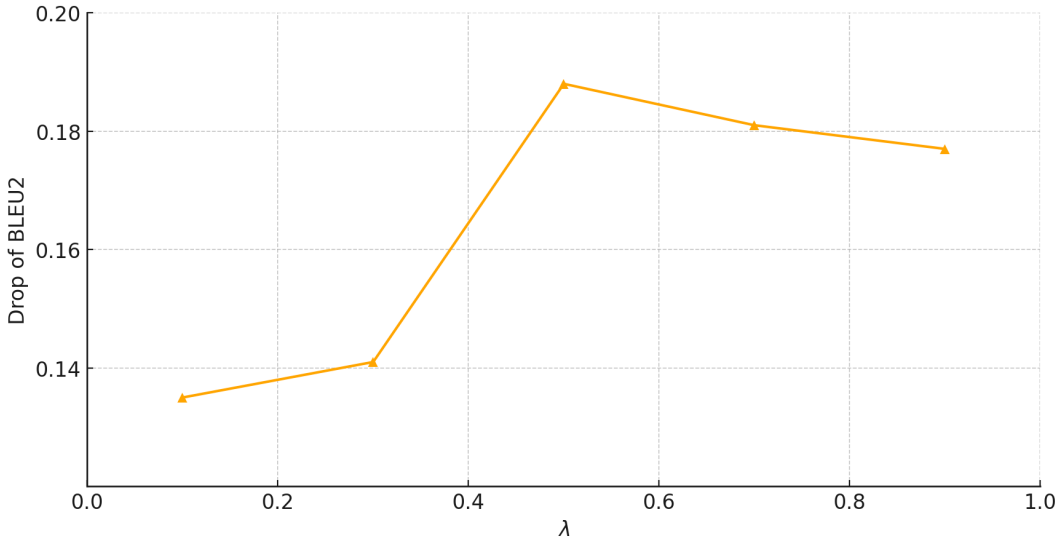

Figure 4.9: Drops in BLEU2 scores across varying λ counts in the differential evolution algorithm.

upward. This suggests that when the attention region is too narrow, the pixels targeted for attack may miss critical information. The candidate region must expand to a critical threshold before effectively capturing relevant sensitive data. This pattern underscores the importance of selecting an appropriate range for k to ensure comprehensive coverage of vulnerable areas in the input.

4.4.5.3 Attack Strength Analysis

On the other hand, we explore the impact of different attack intensities (strength s) on the results. It can be observed in Figure 4.8 that in our proposed method, larger intensities lead to more noticeable changes in BLEU scores. This can be attributed to a significant alteration in pixel coloration, which impacts the model's capacity to interpret the contents of the image accurately.

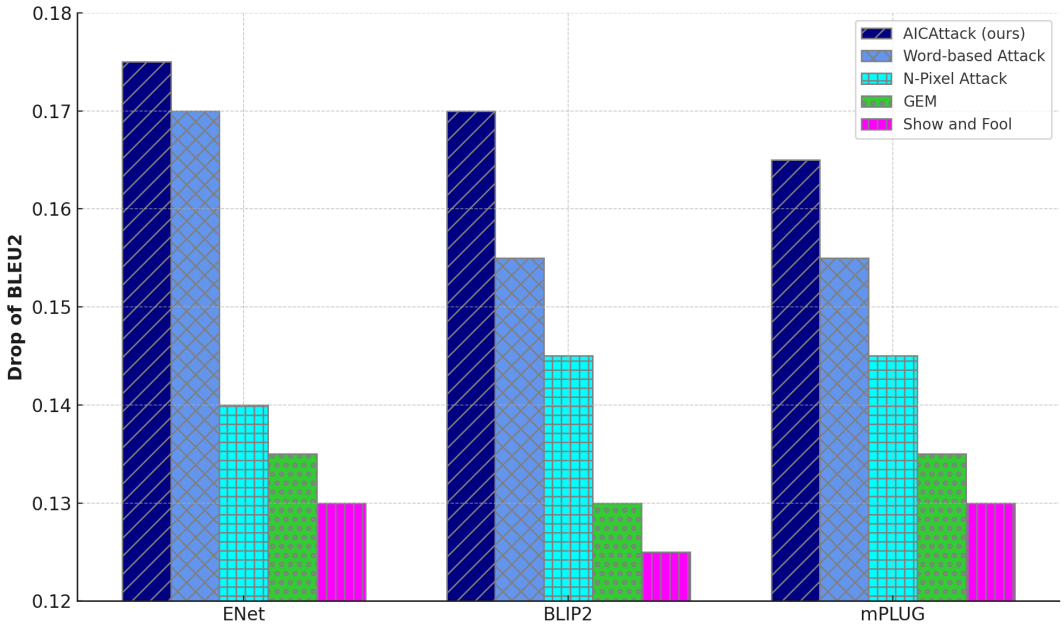

Figure 4.10: Drops of BLEU2 scores reported on multiple baseline captioning models with COCO datasets.

4.4.5.4 Scale Factor Analysis

Finally, we conduct scale factor λ experiments in the differential evolution algorithm. The result is shown in Figure 4.9. A larger value of λ enables the algorithm to conduct a wider search across the solution space. This can help avoid local minima, although it carries the risk of instability or not achieving accurate solutions. Conversely, a smaller λ value fosters exploitation, concentrating the search in a more confined area. This can be advantageous for detailed adjustments but may lead to the algorithm becoming trapped in local optima. Hence, in our AICAttack, we pick 0.5 between 0 to 1 as the λ value.

4.4.6 Transferability of Attacks

To test our model's reaction to unknown captioning models F', we conducted a set of experiments with three baseline captioning works mPLUG[187], BLIP2[23], ExpansionNet v2 [188](denote as ENet). Specifically, we selected adversarial examples designed for SAT to attack baselines across different attack methods. As results shown in Figures 4.10

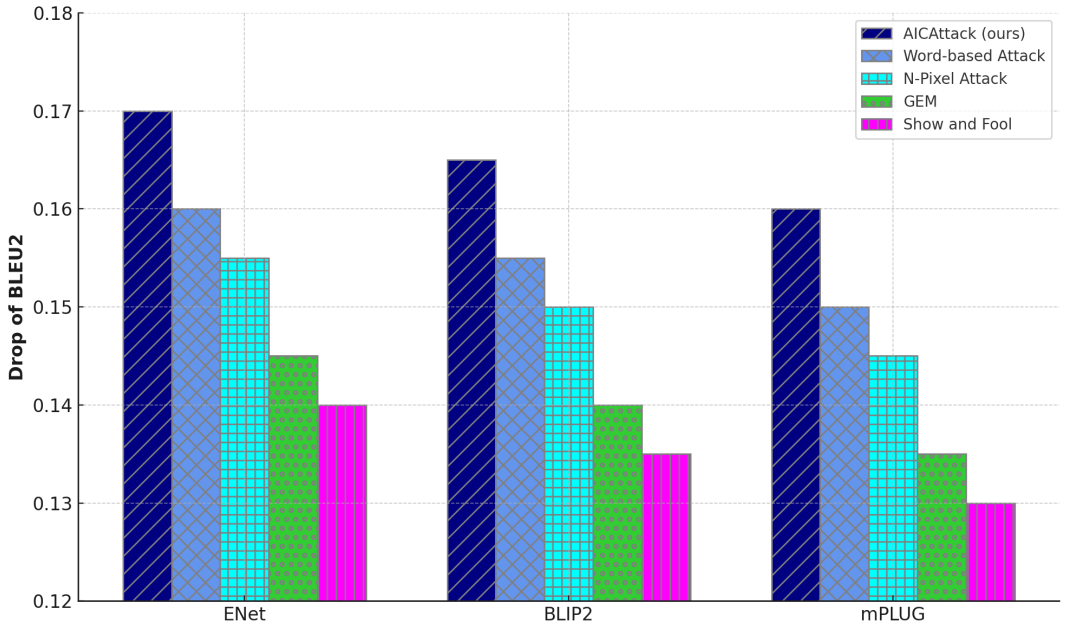

Figure 4.11: Drops of BLEU2 scores reported on multiple baseline captioning models with Flickr8k datasets.

	Data	BLEU1	BLEU2	BLEU4
Train	100% training data	71.824	50.381	25.033
	+ 1000 adversarial	-0.013	-0.017	-0.06
	5% training data	56.333	51.387	23.667
	+ 100 adversarial	+3.251	+1.33	-0.216
Attack	attacking 1000 training data	-59.124	-31.581	-11.533
	attacking 1000 adversarial	-58.273	-30.333	-9.175
	attacking 100 training data	-21.592	-19.033	-13.833
	attacking 100 adversarial	-4.261	-5.012	-1.417

Table 4.4: Adversarial Retraining and attacking results on two different scenarios. All the increases and decreases are based on training results with 100% or 5% training data, respectively.

and 4.11, our AICAttack generates adversarial examples with higher transferability among five attacking approaches.

4.4.7 Adversarial Retraining

This section discussed AICAttack's potential improvement to victim models on the BLEU score. The basic experiment setting is to generate adversarial samples with AICAttack and data from COCO and include them in additional training data. The retraining was conducted in two scenarios: training SAT with (1) full training data and 1000 adversarial data, and (2) 5% training data and 100 adversarial data to simulate low-resource cases.

4.4.7.1 The Accuracy of Retrained Model

As shown in Table 4.4, when training with all training data and 1000 adversarial samples, the BLEU scores decrease by 0.013, 0.017 and 0.06, respectively. This result shows that adding new adversarial examples to the original training data leads to worse outcomes. This indicates that adversarial examples can negatively affect how models are trained. However, the SAT model performs better under low-data scenarios with adversarial samples. This shows that adversarial examples may influence model retraining and are more likely to affect small data sets.

4.4.7.2 The Robustness Confronting Adversarial Attacks

To measure the effectiveness of the model's robustness after adversarial retraining, we use our AICAttack to foul the SAT trained with or without adversarial samples for comparison. In Table 4.4, the change in the BLEU score shows that adversarial training makes the attack less effective, with fewer BLEU scores dropping. This phenomenon is more significant in the low-resource scenario due to training with little data.

These results suggest that AICAttack can be used to improve retrained captioning models' robustness with a considerable BLEU score drop.

4.5 Summary and Discussion

In this chapter, we introduced AICAttack, a robust and versatile adversarial learning strategy for the attack of image captioning models. Our black-box approach harnesses the power of an attention mechanism and differential evolution optimization to orchestrate subtle yet effective pixel perturbations. It avoids the complex extraction of parameters from encoder-decoder models while keeping the attack cost within a minimal range. Another critical innovation of AICAttack is its attention-based candidate selection mechanism, which identifies optimal pixels for perturbation, enhancing the precision of our attacks. Through extensive experimentation on benchmark datasets and captioning models, we have demonstrated the superiority of AICAttack in achieving significantly higher attack success rates compared to state-of-the-art methods. In future, we plan to develop defensive strategies against image captioning attacks by designing more robust learning algorithms for image captioning models.

CHAPTER 5

DECEIVING QUESTION-ANSWERING MODELS: A HYBRID WORD-LEVEL ADVERSARIAL APPROACH

Deep learning underpins most of the currently advanced natural language processing (NLP) tasks such as textual classification, neural machine translation (NMT), abstractive summarization and question-answering (QA). However, the robustness of the models, particularly QA models, against adversarial attacks is a critical concern that remains insufficiently explored. This chapter introduces QA-Attack (Question Answering Attack), a novel word-level adversarial strategy that fools QA models. It demonstrates versatility across various question types, particularly when dealing with extensive long textual inputs. Extensive experiments on multiple benchmark datasets demonstrate that QA-Attack successfully deceives baseline QA models and surpasses existing adversarial techniques regarding success rate, semantics changes, BLEU score, fluency and grammar error rate.

5.1 Introduction

Question-answering (QA) models, a key task within Sequence-to-Sequence (Seq2Seq) frameworks, aim to enhance computers' ability to process and respond to natural language queries. As these models have evolved, they have been widely adopted in real-world applications such as customer service chatbots[189], search engines [190], and information retrieval in fields like medicine [191] and law [192]. However, despite the significant progress in deep learning and natural language processing (NLP), these models remain vulnerable to adversarial examples, leading to misinformation, privacy breaches, and flawed decision-making in critical areas [193–196].

Beyond technical effectiveness, QA-Attack poses significant societal risks in safety-critical applications. In healthcare systems, malicious actors could subtly alter input contexts to elicit harmful treatment recommendations. For instance, changing "The patient shows no signs of allergic reaction to penicillin" to "The patient shows known signs of allergic reaction to penicillin" through word substitution could cause the system to contraindicate a life-saving antibiotic. Research confirms such vulnerabilities, showing that medical QA systems like Med-PaLM produce clinically unsafe responses under adversarial prompts [197]. Similarity, legal advice systems exhibit similar weaknesses. An attacker could transform "The defendant must not have intended harm" to "The defendant must have intended harm" by removing a single word, potentially reversing the interpretation of criminal liability standards [198]. In financial advisory contexts, altering "investments with high risk typically yield lower returns" to "investments with low risk typically yield lower returns" could mislead users into dangerous investment strategies.

QA models are expected to comprehend given texts and questions, providing accurate and contextually relevant answers [88]. These models primarily address two types of questions: Informative Queries and Boolean Queries. The Informative Queries typically

begin with interrogative words such as “who,” “what,” “where,” “when,” “why,” or “how,” requiring detailed and specific information from the provided context. Although models like T5 [16], LongT5 [199], and BART [14], which follow an encoder-decoder structure, have demonstrated strong performance, they still suffer from the maliciously crafted adversarial examples. Initially, studies like “Trick Me If You Can” [4] primarily relied on human annotators to construct effective adversarial question-answering examples. This methodology, however, inherently constrained scalability and increased resource demands. Automated approaches for attacking textual classifiers in QA models emerged as research progressed. Gradient-based methods, as employed in Fast Gradient Sign Method (FGSM) [2], RobustQA [5], UAT [31], and HotFlip [32], were developed to identify and modify the most influential words affecting model answers. Building upon a deeper understanding of QA tasks, subsequent studies explored more targeted strategies. For instance, Position Bias [200], TASA [3], and Entropy Maximization [201] investigated the manipulation of sentence locations and the analysis of answer sentences to identify vulnerable parts of the context. These approaches refined the attack methods by applying modifications through paraphrasing or replacing original sentences, thus enhancing the effectiveness of adversarial examples. However, these methods encounter two primary challenges: 1) None of these attack methods is suitable for both “informative queries” and “boolean queries”. 2) Constraining the search space for optimal vulnerable words to answer-related sentences compromises attack effectiveness; meanwhile, targeting entire sentences proves inefficient [30].

In addition, Boolean Queries seek a simple binary “Yes” or “No” answer. Models like BERT [28], RoBERTa [92], and GPT variants [202–205], which excel at sentence-level understanding and token classification, are widely used for Boolean QA tasks. These models leverage their deep contextual understanding of language to accurately determine whether a given statement is true or false, making them state-of-the-art baselines for

the task. Researchers have proposed various approaches to target boolean classifiers in the context of Boolean Queries attacks. Attacks like [11, 59, 60, 83, 84], which involve adding, relocating, or replacing words, are based on the influence that each word has on the prediction. They retrieve word importance by the output confidence to the level or with gradient. However, gradient calculation is computationally intensive and ineffective when dealing with long context input, and knowing victim models' internal information is unrealistic in practice.

We present QA-Attack, an adversarial attack framework tailored for both Informative Queries and Boolean Queries in QA models. QA-Attack uses a Hybrid Ranking Fusion (HRF) algorithm that integrates two methods: Attention-based Ranking (ABR) and Removal-based Ranking (RBR). ABR identifies important words by analyzing the attention weights during question processing, while RBR evaluates word significance by observing changes in the model's output when specific words are removed. The HRF algorithm combines these insights to locate vulnerable tokens, replaced with carefully selected synonyms to generate adversarial examples. These examples mislead the QA system while preserving the input's meaning. This unified attack method improves performance and stealth, ensuring realistic applicability for both queries. In summary, our work makes the following key contributions:

- We present QA-Attack with a Hybrid Ranking Fusion (HRF) algorithm designed to target question-answering models. This novel approach integrates attention and removal ranking techniques, accurately locating vulnerable words and fooling the QA model with a high success rate.
- Our QA-Attack can effectively target multiple types of questions. This adaptability allows our method to exploit vulnerabilities across diverse question formats, which significantly broadens the scope of potential attacks in various real-world scenarios.
- QA-Attack generates adversarial examples by implementing subtle word-level

changes that preserve both linguistic and semantic integrity while minimizing the extent of alterations, and we conduct extensive experiments on multiple datasets and victim models to thoroughly evaluate our method's effectiveness in attacking QA models.

The rest of this section is structured as follows. We first review attention for Seq2Seq models and synonym generation in Section 5.2. Then, we detail our proposed method in Section 5.3. We evaluate the performance of the proposed method through extensive empirical analysis in Section 5.4. We conclude the chapter with suggestions for future work in Section 5.5.

5.2 Preliminary

This section provides an overview of attention mechanism for Seq2Seq models and the existing synonym generation methods.

5.2.1 Attention Mechanism

The attention mechanisms also revolutionized textual input tasks by enabling models to selectively focus on relevant tokens during decoding, significantly enhancing both performance and the ability to process longer texts [206]. These advances proved particularly transformative for sequence-to-sequence (Seq2Seq) models, which convert textual inputs into textual outputs for tasks ranging from machine translation and text summarization to interactive dialogue systems. While early Seq2Seq architectures suffered from a fundamental limitation-compressing entire input sequences into fixed-length vectors-attention mechanisms elegantly solved this bottleneck by enabling dynamic focus on input elements throughout the processing pipeline. The field has developed three primary attention variants for textual processing: Bahdanau attention, Luong attention,

and self-attention, each offering distinct approaches to the challenge of context-aware text processing.

Bahdanau Attention (Additive Attention): Proposed by Bahdanau et al. [68], calculates a weighted sum of encoder hidden states, where the weights (attention scores) are determined by the relevance of each encoder state to the current decoder state. This relevance is computed using an additive scoring function shown in Equation 5.1. By aligning the decoder's focus with specific parts of the input, this mechanism improved the performance of neural machine translation systems, especially for languages with long or syntactically complex sentences. The mechanism's ability to explicitly model input-output alignment makes it particularly useful for tasks like translation, where individual input tokens correspond closely to output tokens. However, additive attention is computationally expensive for long sequences due to its use of additional parameters and matrix operations.

$$e_{ij} = v^T \tanh\left(W_1 h_i + W_2 s_j\right) \tag{5.1}$$

where in the scoring function e_{ij}:

- h_i: Hidden state of the encoder at time step i.
- s_j: Hidden state of the decoder at time step j.
- W_1 and W_2: Trainable weight matrices to transform the hidden states.
- v^T: A trainable weight vector.
- This function scores the relevance of the encoder's hidden state (h_i) to the decoder's current state (s_j).

Luong Attention (Multiplicative Attention): Luong et al. [99] introduced multiplicative attention as a simpler and more computationally efficient alternative to additive attention. To compute attention weights, this method uses a dot-product scoring function, as

shown in Equation 5.2, between encoder and decoder states. While it lacks the flexibility of learnable parameters found in additive attention, its efficiency makes it ideal for tasks requiring real-time inference, such as speech-to-text systems.

$$e_{ij} = h_i^T W s_j \tag{5.2}$$

where:

- h_i: Encoder hidden state at time step i.
- s_j: Decoder hidden state at time step j.
- W: Trainable weight matrix.

The dot product computes the similarity between h_i and s_j. This method is computationally efficient and suitable for tasks with large input sequences.

Self-Attention (Transformer Attention): Self-attention, introduced in Transformer models by Vaswani et al. [69], solved Seq2Seq architectures by allowing all tokens in the input sequence to attend to each other, irrespective of their positions. Unlike Bahdanau or Luong attention, which only operates between encoder and decoder states, self-attention captures global dependencies within the input sequence. This mechanism is potent for tasks like summarization, where long-range dependencies across a document must be preserved. Self-attention also enables parallel computation, making it highly scalable for large datasets and longer sequences.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.3}$$

where:

- Q (Query): A representation of the current token or position in the sequence.
- K (Key): A representation of all tokens in the sequence, used to compute the relevance to Q.

- V (Value): A representation of all tokens in the sequence, used to generate the output.

- d_k: The dimensionality of the key vectors, used to scale the dot product to prevent excessively large values.

- Scaled Dot Product**: The relevance between Q (query) and K (key) is calculated using their dot product:

$$\frac{QK^\top}{\sqrt{d_k}},$$

$\sqrt{d_k}$ is a scaling factor to avoid overly large dot-product values when d_k is large.

5.2.2 Attention-related Attacks

Although attention mechanisms have enhanced the capabilities of Seq2Seq models, they also introduce distinct security vulnerabilities that adversaries can exploit. The attention mechanism's core strength-its ability to focus on specific input elements-becomes a potential weakness when malicious adversaries target either the attention scores directly or manipulate the input tokens that influence these scores. Such adversarial attacks can subtly or dramatically alter the model's output.

Token substitution exemplifies how attackers can exploit attention mechanisms by identifying and replacing high-importance tokens with synonyms or contextually similar terms. This sophisticated manipulation disrupts the crucial alignment between input and output sequences by shifting attention weights in predictable ways. Consider how replacing "bank" with "shore" in financial text can completely redirect a translation model's interpretation, fundamentally altering its output. Alzantot et al. demonstrated the elegance of this attack vector, showing how adversarial examples could preserve perfect grammatical structure while exploiting attention mechanisms' structured focus to deceive models into generating incorrect outputs.

Zhou et al. improved this with their attention-guided genetic algorithm (AGA), an innovative approach that harnesses attention scores for precise attack targeting. The method systematically identifies high-impact regions in input sequences by analyzing attention weights, then employs genetic algorithms to optimize perturbations iteratively. This sophisticated optimization process maximizes output disruption while maintaining minimal input changes. What makes AGA particularly powerful is its ability to leverage attention patterns for efficient targeting, even in black-box scenarios where model architectures remain hidden.

Ni et al. further refined adversarial techniques through their work on neural machine translation (NMT) systems, developing a hybrid attention learning method that exploits attention distributions with unprecedented precision. Their approach systematically identifies vulnerable points in translation systems by analyzing both language-specific patterns and sequence-sensitive tokens, using attention distributions to pinpoint optimal targets for manipulation. By precisely disrupting tokens that receive the highest attention weights, their method effectively degrades translation quality while maintaining the appearance of legitimate input.

The HackAttend technique developed by Liong et al. introduces a direct assault on attention mechanisms by manipulating adversarial attention masks. Rather than modifying input tokens, this sophisticated approach targets the self-attention weights fundamental to Transformer-based architectures, forcing models to misallocate focus to irrelevant or deceptive tokens. Their research demonstrates how subtle perturbations to attention distributions can trigger cascading errors throughout the model's prediction process.

Wallace et al.'s research on adversarial triggers revealed a powerful exploitation of attention mechanisms through strategically crafted trigger phrases. These carefully designed sequences exploit the model's attention patterns, drawing disproportionate focus

away from contextually relevant tokens. The technique's power lies in its universality: by understanding how attention mechanisms prioritize certain patterns, attackers can craft trigger phrases that consistently derail model outputs across diverse tasks, from machine translation to text summarization.

Finally, Shen et al. (2023) explored dynamic attention mechanisms as a countermeasure to adversarial attacks, highlighting how attackers often exploit the static nature of traditional attention systems. While this study focuses on defense, it illustrates how static attention distributions provide adversaries with predictable targets for disruption. By introducing attention rectification and dynamic modeling, Shen et al. aim to mitigate the impact of adversarial inputs. This work highlights the critical interplay between attack strategies and attention mechanisms, underscoring the need for adaptive approaches to address the inherent vulnerabilities of static attention.

5.3 Our Proposed Attack Method

In this section, we introduce the QA-Attack algorithm. It can be summarized into three main steps. First, the method effectively captures important words in context by processing pairs of questions and corresponding context using attention-based and removal-based ranking approaches. Then, attention and removal scores are combined, allowing the identification of the most influential words. At last, a masked language model [28] is utilized to identify potential synonyms that could replace the targeted words. The overall workflow of QA-Attack is shown in Figure 5.1. In the following sections, we explain our model in detail.

5.3.1 Problem Setting

Given a pre-trained question-answering model F, which receives an input of context C, question q, and outputs answer a, such that $F(q,C) = a$. The objective is to deceive the

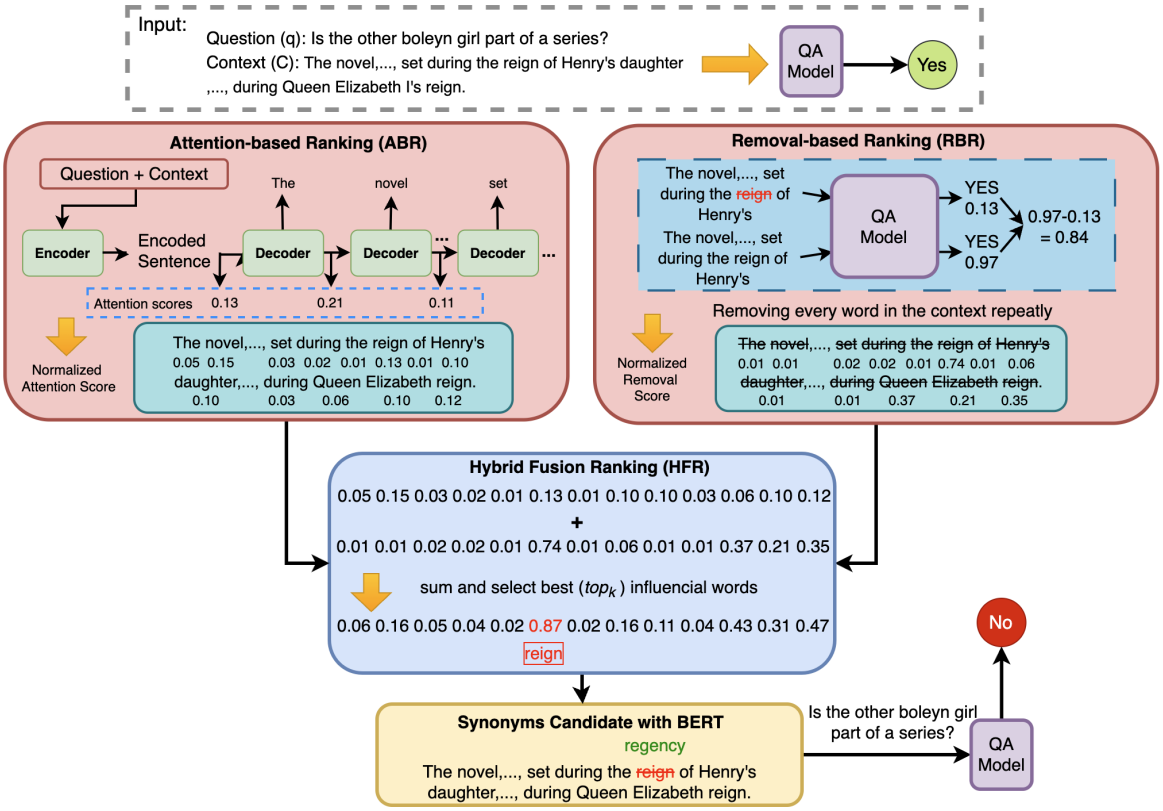

Figure 5.1: The workflow of our QA-Attack algorithm for QA models. It processes question-context pairs through two parallel modules: Attention-based Ranking (ABR) and Removal-based Ranking (RBR). These modules generate attention and removal scores respectively for each word using customized attention mechanisms and removal ranking strategies. The scores are then aggregated, and the top_k highest-scoring words are selected as candidates. Finally, these candidates are replaced with BERT-generated synonyms to create adversarial examples that can effectively mislead the QA model.

performance of F with perturbed context C' such that $F(q, C') \neq a$. To craft C', a certain number of perturbation c_{adv} is added to the context C by replacing some of its original tokens $\{c_1, c_2, ..., c_n\}$.

5.3.2 Attention-based Ranking (ABR)

Attention mechanisms were first used in image feature extraction in the computer vision field [129, 139, 207, 208]. However, they were later employed by [68, 209] to solve machine translation problems. In translation tasks, attention mechanisms enable models

Algorithm 3: QA-Attack Algorithm (Adversarial Generation)

Input: QA victim model $F(\cdot)$, logits L, question q, context C, words in the context c, reference answer a, attention network A, top k words to attack top_k, number of synonyms d, BERT MLM model $BERT$ for generating synonyms. ;
Output: Optimal adversarial sample C' ;
// **Attention-based Ranking**;
Compute attention scores: $\alpha \leftarrow [(c, A(q+C))]$;
Initialize attention score list: $attention\_scores \leftarrow [\,]$;
for *each score in* α **do**
    if $score \in C$ **then**
        Append $score$ to $attention\_scores$;
    end
end
// **Removal-based Ranking**;
Initialize importance score list: $importance\_scores \leftarrow [\,]$;
for *each c in* C **do**
    Generate modified context: $C^* \leftarrow C$ excluding c;
    Compute importance score: $importance\_scores.append(|F(q,C^*) - F(q,C)|)$;
end
// **Hybrid Ranking Fusion**;
Combine attention and importance scores:
$combined\_scores \leftarrow attention\_scores \cup importance\_scores$;
Select top_k words: $top\_k\_list \leftarrow \text{sort}(combined\_scores)[:top_k]$;
Initialize adversarial examples list: $Adv\_list \leftarrow [\,]$;
for *each t in* top_k_list **do**
    Generate adversarial token from d potential synonyms: $c_{adv} \leftarrow BERT(t)$;
    Create adversarial context: $\Delta C \leftarrow [c_1, \ldots, c_{adv}, \ldots, c_n]$;
    Append ΔC to Adv_list;
end

to prioritize and focus on the most relevant parts of the input data [99]. In question-answering tasks, attention scores are imported to examine the relationships between question and context, allowing the model to determine which words or phrases are most relevant to answering the question [210]. Hence, we leverage the attention score to identify target words for our attack. We employ the attention mechanism from T5 [16] that has been specifically optimized for question-answering tasks in UnifiedQA [94]. As shown in Fig. 5.1, the "Attention-based Ranking" begins by encoding the input context

Algorithm 4: QA-Attack Algorithm (Optimization)

Initialize maximum gap: max_gap $\leftarrow -\infty$;
Initialize optimal adversarial context: $C' \leftarrow \emptyset$;
for *each* adv *in* Adv_list **do**
    if $F(adv) \neq a$ **then**
        Compute gap: gap $\leftarrow L(F(adv)) - L(F(C))$;
        if $gap > max\_gap$ **then**
            Update maximum gap: max_gap $\leftarrow$ gap;
            Update optimal adversarial context: $C' \leftarrow adv$;
        end
    end
end
return Optimal adversarial sample C'

and question through an encoder. During the encoding process, self-attention allows the model to analyze how each word in the input relates to every other word, effectively highlighting the words that carry the most weight in understanding both question and context. In the decoding process, cross-attention further refines this by focusing on the parts of the input most relevant to generating the correct output. By averaging the attention scores of all layers and heads, we match them to each input word.

The implement details are shown in Algorithm 3. The question & context pair is fitted into attention network A, and we filter out the attention scores for context (lines 3 to 10 of Algorithm 3). Then, the attention score of each word corresponding to each layer is summed up. After averaging and normalization, the word-level attention score is obtained.

5.3.3 Removal-based Ranking (RBR)

Previous studies on adversarial attacks in the text have shown that each word's significance can be quantified using an importance score [3, 11, 60, 98]. This score is largely determined by how directly the word influences the final answer. To enhance the efficacy of ranking progress, we rank each word in the context to obtain the removal importance

score (lines 11 to 16 of Algorithm 3). Given the input context C containing n words from c_1 to c_n and question q, the importance score (removal score) of the $i\ th$ $(1 \leq i \leq n)$ word c_i is:

$$I_i = L_F(a \mid q, C) - L_F(a \mid q, C \setminus c_i), \tag{5.4}$$

where $C \setminus c_i$ represents the context after deleting c_i, and $L_F = \log P(a \mid q, C)$ refers to the probability (logits) of the label, respectively.

5.3.4 Hybrid Ranking Fusion (HRF)

The attention-based and removal-based word selection techniques offer complementary perspectives on token significance, each highlighting different aspects of word importance. Consequently, we tend to choose words that both methods consider significant. This is achieved by adding the scores from each method for every word to create a fusion score.

When generating a fusion score, we address several key factors. First, we independently normalize the attention and removal scores before adding them together. Then, to balance attack effectiveness and efficiency, we introduce a top_k parameter, a positive integer that controls the number of words targeted. Finally, we select the top_k highest-scoring words for modification (lines 17 to 20 of Algorithm 3).

5.3.5 Synonym Selection

Various synonym generation methods exist, including Word2Vec [211], Hownet [212], and WordNet [212]. We adopt BERT [28] for synonym selection due to its textual capabilities, which enable it to generate synonyms based on the complete sentence structure. Unlike Word2Vec's static embeddings or WordNet's fixed synonym lists, BERT's context-sensitive approach allows for dynamic synonym selection that preserves both semantic meaning and grammatical correctness. This contextual awareness makes BERT particularly effective for crafting natural and semantically coherent adversarial examples.

We process each selected word in the context by replacing it with the "[MASK]" token. This modified context is then input into the BERT Masked Language Model (MLM) to predict the most likely substitutions for the masked word. To expand the range of potential samples, we introduce a parameter d that controls the number of synonym substitutions considered (lines 21 to 25 of Algorithm 3). This approach allows us to generate a diverse set of imperceptible replacements while maintaining contextual relevance.

5.3.6 Candidate Selection

We define an optimal adversary as one that maximizes the divergence between the predicted and attacked answers. For boolean queries (yes/no), we follow previous successful textual classifier approaches by comparing the logits of output labels. For informative queries, we aggregate the logits across individual words in the response. The optimal adversary C' is identified from the "Adv_list" using the logits derivation function L, as detailed in Algorithm 4.

5.4 Experiment and Analysis

In this section, we present a comprehensive evaluation of QA-Attack's outcomes compared to current high-performance baselines. Our analysis covers several key aspects with various metrics, providing a thorough understanding of our method's capabilities, limitations, and performance across diverse scenarios. We provide a detailed analysis of attack performance and imperceptibility (Sec. 5.4.4). Besides, to gain deeper insights, we conduct Ablation Studies (Sec. 5.4.5) and assess attacking efficiency (Sec. 5.4.6). In addition, we examine QA-Attack's response to defense strategies (Sec. 5.4.8), exploring the effects of Adversarial Retraining (Sec. 5.4.7) and investigating the Transferability of Attacks (Sec. 5.4.9). Finally, we report the preference of our attack by investigating

Table 5.1: Dataset distribution and corresponding baseline performance (F1).

Dataset	Data Distribution				Model Performance (F1)		
	Total	Train	Validation	Test	T5	LongT5	$BERT_{base}$
SQuAD 1.1	100k	87,600	10,570	N/A	88.9	89.5	88.5
SQuAD V2.0	150k	130,319	11,873	N/A	81.3	83.2	74.8
NewsQA	119k	92,549	5,165	5,126	66.8	67.2	60.1
BoolQ	16k	9,427	3,270	3,245	85.2	86.1	80.4
NarrativeQA	45k	32,747	3,461	10,557	67.5	68.9	62.1

parts of speech preference (Sec. 5.4.10) and analyzing its robustness versus the scale of pre-trained models

5.4.1 Datasets and Victim Models

We assess QA-Attack using four informative queries datasets: SQuAD 1.1 [27], SQuAD V2.0 [213], NarrativeQA [214], and NewsQA [215], along with the boolean queries dataset BoolQ [216].

- SQuAD 1.1: Questions formulated by crowd workers based on Wikipedia articles. Answers are extracted as continuous text spans from the corresponding passages.

- SQuAD 2.0: Extension of SQuAD 1.1 incorporating unanswerable questions. These questions are designed such that no valid answer can be located within the provided passage, adding complexity to the task.

- NarrativeQA: Questions based on entire books or movie scripts. Answers are typically short and abstractive, demanding deeper comprehension and synthesis of narrative elements.

- NewsQA: Questions based on CNN news articles designed to test reading comprehension in the context of current events and journalistic writing.

- BoolQ: Dataset of boolean (yes/no) questions derived from anonymized, aggregated queries submitted to the Google search engine, reflecting real-world information-seeking behaviour.

Our experiment includes three question-answering models for comparison. They are T5[94], LongT5 [199], and $\text{BERT}_{\text{base}}$ [28]. The LongT5 is an extension of T5 with an encoder-decoder specifically for long contextual inputs. The BERT-based models are structured with bidirectional attention, meaning each word in the input sequence contributes to and receives context from both its left and right sides. Table 5.1 presents the distribution of dataset splits and F1 scores reported on each QA baseline.

5.4.2 Baseline Attacks

For our experimental baselines, we employ five leading attack methods: TASA [3], RobustQA [5], Tick Me If You Can (TMYC) [4], T3 [6], and TextFooler [11]. We utilize the official implementation of T3 in its black-box setting, while TASA, TMYC, and RobustQA are employed with their standard configurations. TextFooler, originally not designed for question-answering tasks, was adapted for our experiments. We modified it to process the context only (questions are removed).

5.4.3 Experiment Settings and Evaluation Metrics

The base setting of our experiments is to let $top_k = 5$, $d = 2$, and use a BERT-base-uncased[1] with 12 Transformer encoder layer (L) and 768 hidden layers (H) as the synonym generation model. Some visualized examples are shown in Table 5.2 and 5.3. Tables 5.4, 5.5, and 5.6 summarize the experimental results on informative queries datasets, offering a comparative analysis of our QA-Attack method against five state-of-the-art QA baselines. For boolean queries, we present the attacking results on the BoolQ

[1] `https://github.com/google-research/bert/?tab=readme-ov-file`.

dataset in Table 5.7. Besides, we provide code for the reproductivity of our experiments[2]. The metrics used in our experiment are:

- **F1**: The F1 score balances precision and recall, providing a nuanced view of how much the attacked answers match reference answers.
- **ROUGE and BLEU**: A higher BLEU [161] or ROUGE [25] score in context indicates that the adversarial context retains more of the exact phrasing, contributing to better linguistic fluency and coherence.
- **Exact Match (EM)** Measures the percentage of model predictions that exactly match the correct answers in both content and format.
- **Similarity (SIM)**: Evaluates the semantic similarity between original and adversarial context using BERT [28] embeddings. (Note: In our following experiments, EM and SIM are not only measured answers but also reflect the quality of the generated context in Sec. 5.4.5.3).
- **Modification Rate (Mod)**: Mod measures the proportion of altered tokens in the text. This metric considers each instance of replacement, insertion, or deletion as a single token modification.
- **Grammar Error (GErr)**: GErr measures the increase in grammatical inaccuracies within successful adversarial examples relative to the original text. This measurement employs LanguageTool [217] to enumerate grammatical errors.
- **Perplexity (PPL)**: PPL serves as an indicator of linguistic fluency in adversarial examples [59, 218]. The perplexity calculation utilizes a GPT-2 model with a restricted vocabulary [80].

[2]Our code is available at `https://github.com/UTSJiyaoLi/QA-Attack`.

5.4.4 Experiment Analysis

Table 5.2: Comparison of original and adversarial contexts for boolean queries. The table highlights the differences between the original and adversarial contexts, as well as the corresponding answers provided by the model before and after the attack.

Question	Was the movie "The Strangers" based on a true story?
Context	The Strangers is a 2008 American slasher film written and directed by Bryan Bertino. Kristen (Liv Tyler) and James (Scott Speedman) are expecting a relaxing weekend at a family vacation home, but their stay turns out to be anything but peaceful as three masked torturers leave Kristen and James struggling for survival. Writer-director Bertino was inspired by real-life events: the Manson family Tate murders, a multiple homicide; the Keddie Cabin Murders, that occurred in California in 1981; and a series of break-ins that occurred in his own neighborhood as a child.
Adversary	The Strangers is a 2008 American slasher thriller written and directed by Bryan Bertino. Kristen (Liv Tyler) and James (Scott Speedman) are spending a relaxing weekend at a family vacation home, but their stay turns out to be anything but peaceful as three masked torturers leave Kristen and James struggling for survival. Writer-director Bertino was influenced by real-life incidents: the Manson family Tate murders, a multiple homicide; the Keddie Cabin Murders, that occurred in California in 1981; and a series of break-ins that occurred in his own home as a child.
Original Answer	Yes
Attacked Answer	No

Our experimental results in Table 5.4, 5.5, 5.6 demonstrate that QA-Attack consistently outperforms baseline methods across all informative datasets. As shown in Table 5.7, our method achieves superior performance on the boolean dataset, surpassing all baseline approaches in degrading victim models' accuracy (note that TASA is designed only for informative queries; it is incompatible with boolean query attacks). For informative queries, comparing performance on attacking LongT5 with SQuAD 1.1 and NarrativeQA datasets (representing shortest and longest contexts) in Table 5.6, we observe that while F1 and EM scores decrease for longer contexts, QA-Attack maintains superiority over baselines. This indicates our approach's robustness and adaptability to varying context lengths, particularly in long text. The improved performance in longer

Table 5.3: Comparison of original and adversarial contexts for informative queries. The table highlights the differences between the original and adversarial contexts, as well as the corresponding answers provided by the model before and after the attack.

Question	Who ruled the Duchy of Normandy?
Context	The Normans were famed for their martial spirit ... The Duchy of Normandy, which they formed by treaty with the French crown, was a great fief of medieval France, and under Richard I of Normandy was forged into a cohesive and formidable principality in feudal tenure ... Norman adventurers founded the Kingdom of Sicily ... an expedition on behalf of their duke, William the Conqueror, led to the Norman conquest of England at the Battle of Hastings in 1066.
Adversary	The Normans were famed for their warrior spirit ... The Duchy of Normandy, which they formed by treaty with the French crown, was a great fief of medieval France, and under William I of Normandy was forged into a cohesive and formidable principality in feudal tenure ... Norman adventurers invaded the Kingdom of Sicily ... an invasion on behalf of their duke, William the Conqueror, led to the Norman conquest of England at the siege of Hastings in 1066.
Original Answer	The French crown
Attacked Answer	William I of Normandy

contexts suggests our HRF approach effectively identifies and targets vulnerable tokens. Regarding semantic consistency, QA-Attack achieves lower similarity scores compared to baseline methods, indicating that the answers generated after the attack deviate more in meaning from the ground truth responses.

Additionally, the quality of the generated adversarial samples is evident from the ROUGE and BLEU scores. Our method consistently achieves higher ROUGE and BLEU scores compared to the baselines, which suggests that the adversarial examples generated by QA-Attack are not only effective in terms of altering the model's output but also maintain a high degree of contextual and linguistic coherence. This is largely due to our synonym selection method, which ensures the replacements are contextually appropriate and semantically relevant. Moreover, the token-level replacement strategy, which only mods fewer words (typically five in the base setting), further ensures that the adversarial examples remain similar to the original context while fooling the model.

Table 5.4: Comparative analysis of QA-Attack and baseline models on T5. Drops of BLEU and ROUGE scores (uni-gram) on contexts are reported in the table, with higher values indicating better performance. For F1, EM, and SIM (i.e., similarity) metrics on answers, lower values indicate better performance.

Datasets	Methods	F1↓	EM↓	ROUGE↑	BLEU↑	SIM↓
SQuAD 1.1	TASA [3]	9.21	7.49	89.12	82.88	6.38
	TMYC [4]	7.28	8.21	81.91	78.72	8.22
	RobustQA [5]	5.89	7.52	84.23	77.41	6.03
	TextFooler [11]	10.6	10.49	83.11	76.05	6.29
	T3 [6]	5.41	6.29	86.83	73.82	7.23
	QA-Attack (ours)	**4.67**	**5.68**	**90.51**	**84.11**	**5.91**
SQuAD V2.0	TASA [3]	20.09	19.31	70.21	76.06	7.29
	TMYC [4]	17.23	20.68	65.19	69.82	9.05
	RobustQA [5]	16.37	18.73	67.71	63.19	8.14
	TextFooler [11]	21.69	24.5	65.33	65.01	9.32
	T3 [6]	11.19	19.68	69.71	73.53	8.82
	QA-Attack (ours)	**9.13**	**15.41**	**72.76**	**77.28**	**6.33**
Narrative QA	TASA [3]	11.79	15.25	68.11	70.36	6.11
	TMYC [4]	12.73	9.32	65.91	67.22	7.61
	RobustQA [5]	10.01	13.91	67.19	64.11	6.81
	TextFooler [11]	14.72	18.61	63.85	62.82	11.74
	T3 [6]	11.74	11.37	62.34	60.17	6.28
	QA-Attack (ours)	**5.61**	**7.23**	**69.18**	**75.73**	**5.23**
NewsQA	TASA [3]	8.56	29.44	77.28	69.44	7.11
	TMYC [4]	6.12	31.23	77.96	72.49	9.22
	RobustQA [5]	5.12	29.48	83.81	79.82	10.84
	TextFooler [11]	9.01	30.86	74.21	57.44	27.91
	T3 [6]	6.21	28.52	75.22	72.56	14.27
	QA-Attack (ours)	**3.61**	**24.42**	**78.85**	**82.83**	**8.92**

5.4.5 Ablation and Hyperparameters Studies

To comprehensively validate the efficacy of the proposed QA-Attack method, this section conducts a detailed ablation study, dissecting each component to assess its individual impact and overall contribution to the method,Äôs performance.

Table 5.5: Comparative analysis of QA-Attack and baseline models on $\text{Bert}_{\text{base}}$. Drops of BLEU and ROUGE scores (uni-gram) on contexts are reported in the table, with higher values indicating better performance. For F1, EM, and SIM (i.e., similarity) metrics on answers, lower values indicate better performance.

Datasets	Methods	F1↓	EM↓	ROUGE↑	BLEU↑	SIM↓
SQuAD 1.1	TASA [3]	15.27	34.33	82.87	67.22	8.19
	TMYC [4]	12.89	28.63	81.51	76.39	10.24
	RobustQA [5]	15.72	25.38	79.28	73.27	15.81
	TextFooler [11]	23.04	37.28	67.28	49.49	14.11
	T3 [6]	8.79	16.11	57.19	63.81	16.92
	QA-Attack (ours)	**6.42**	**13.31**	**91.22**	**77.16**	**7.43**
SQuAD V2.0	TASA [3]	31.22	28.9	77.06	69.05	8.22
	TMYC [4]	29.38	27.77	73.81	67.23	10.34
	RobustQA [5]	27.64	31.82	75.67	71.42	11.23
	TextFooler [11]	36.8	29.49	67.14	62.67	13.28
	T3 [6]	26.16	27.47	74.94	70.14	7.24
	QA-Attack (ours)	**22.18**	**21.5**	**80.12**	**75.23**	**4.11**
Narrative QA	TASA [3]	12.11	14.51	61.15	63.04	7.32
	TMYC [4]	8.41	10.23	52.89	69.82	10.72
	RobustQA [5]	7.24	9.43	63.81	67.43	9.53
	TextFooler [11]	13.74	18.79	56.11	56.82	14.21
	T3 [6]	8.49	15.35	65.48	67.09	7.83
	QA-Attack (ours)	**3.86**	**9.34**	**69.44**	**71.15**	**5.61**
NewsQA	TASA [3]	16.85	20.95	68.74	69.12	15.22
	TMYC [4]	15.86	31.23	77.96	72.49	9.22
	RobustQA [5]	17.72	29.48	83.81	79.82	10.84
	TextFooler [11]	24.13	22.63	59.17	61.22	31.07
	T3 [6]	21.22	22.57	65.14	67.11	18.27
	QA-Attack (ours)	**14.91**	**20.20**	**70.04**	**74.87**	**9.22**

5.4.5.1 Effectiveness of Hybrid Fusion Ranking on Multiple Question Types

We test how HRF, ABR, and RBR methods perform across different top_k values on the SQuAD and BoolQ datasets, with d remaining, shown in Fig. 5.2. HRF consistently outperforms ABR and RBR for all top_k values on both datasets. This suggests combining attention-based and removal-based ranking in HRF is more effective at generating robust adversarial examples than using either method alone. The graph also shows that as top_k increases, all methods improve, indicating that higher top_k values help identify

Table 5.6: Comparative analysis of QA-Attack and baseline models on LongT5. Drops of BLEU and ROUGE scores (uni-gram) on contexts are reported in the table, with higher values indicating better performance. For F1, EM, and SIM (i.e., similarity) metrics on answers, lower values indicate better performance.

Datasets	Methods	F1↓	EM↓	ROUGE↑	BLEU↑	SIM↓
SQuAD 1.1	TASA [3]	10.61	22.45	80.67	70.41	11.88
	TMYC [4]	12.43	29.81	75.37	63.83	13.22
	RobustQA [5]	17.22	31.11	73.11	68.29	17.64
	TextFooler [11]	35.31	44.09	57.77	49.49	25.33
	T3 [6]	9.33	24.52	49.23	60.33	20.87
	QA-Attack (ours)	**7.38**	**18.78**	**84.22**	**72.67**	**9.67**
SQuAD V2.0	TASA [3]	30.71	30.11	64.71	67.28	9.32
	TMYC [4]	34.11	33.88	64.21	65.11	14.82
	RobustQA [5]	29.01	39.59	62.91	68.22	13.09
	TextFooler [11]	38.25	34.67	60.47	64.16	15.44
	T3 [6]	30.44	30.13	65.81	63.72	8.29
	QA-Attack (ours)	**27.11**	**24.73**	**77.37**	**70.32**	**5.29**
Narrative QA	TASA [3]	8.22	10.67	69.83	65.77	9.53
	TMYC [4]	9.36	11.33	63.15	64.27	14.72
	RobustQA [5]	15.83	12.03	64.28	63.12	12.77
	TextFooler [11]	12.77	14.82	62.99	54.21	17.33
	T3 [6]	8.38	8.26	63.92	66.32	8.92
	QA-Attack (ours)	**4.62**	**5.33**	**70.33**	**68.32**	**7.44**
NewsQA	TASA [3]	16.85	24.54	64.83	66.81	14.82
	TMYC [4]	19.28	29.01	62.88	68.67	11.43
	RobustQA [5]	17.23	27.42	58.32	57.22	13.37
	TextFooler [11]	27.22	26.39	53.33	53.01	25.82
	T3 [6]	17.83	25.87	63.25	65.43	19.27
	QA-Attack (ours)	**15.32**	**24.12**	**68.23**	**70.55**	**10.48**

vulnerable tokens better and lead to more effective attacks.

Despite the better performance at higher top_k values, the study uses $top_k = 5$ as a base setting. This choice balances effectiveness with minimal text modification, ensuring that adversarial examples remain close to the original context while still being effective. The consistent trend across both SQuAD and BoolQ datasets demonstrates that HRF's superior performance holds for different question types, showing its versatility in attacking various question-answering models. This analysis highlights the practical effectiveness

Table 5.7: Attack performance comparison on baseline models using the BoolQ dataset, with top results highlighted in bold. Note that TASA [3] is not applicable to boolean questions.

Victim Models	Methods	F1↓	EM↓	ROUGE↑	BLEU↑	SIM↓
UnifiedQA	TASA [3]	N/A	N/A	N/A	N/A	N/A
	TMYC [4]	17.43	19.36	82.09	77.23	21.83
	RobustQA [5]	14.33	18.92	79.15	80.33	13.22
	TextFooler [11]	20.11	19.07	80.91	83.25	33.82
	T3 [6]	15.16	14.74	71.32	68.79	15.82
	QA-Attack (ours)	**8.64**	**13.9**	**87.31**	**86.57**	**11.42**
$Bert_{base}$	TASA [3]	N/A	N/A	N/A	N/A	N/A
	TMYC [4]	21.35	13.28	63.21	70.57	7.34
	RobustQA [5]	24.81	9.21	69.22	76.01	6.67
	TextFooler [11]	33.02	11.57	65.11	67.81	8.17
	T3 [6]	22.06	11.02	76.17	74.62	6.23
	QA-Attack (ours)	**18.39**	**6.51**	**77.21**	**78.11**	**4.66**
LongT5	TASA [3]	N/A	N/A	N/A	N/A	N/A
	TMYC [4]	29.77	9.82	67.04	73.22	7.43
	RobustQA [5]	24.56	8.21	70.49	71.83	9.33
	TextFooler [11]	33.02	11.57	65.11	67.81	8.17
	T3 [6]	22.06	11.02	76.17	74.62	6.23
	QA-Attack (ours)	**18.39**	**6.51**	**77.21**	**78.11**	**4.66**

of the HRF method and its ability to generate impactful adversarial examples across different QA tasks.

5.4.5.2 Effectiveness of Synonyms Selection

To evaluate our Synonyms Selection approach, we conduct comparisons in two aspects. We first compare our BERT-based synonym generation against two alternative methods: WordNet [219], an online database that contains sets of synonyms, and HowNet [212], which produces semantically similar words using its network structure. Using the base configuration, we evaluate the EM scores when attacking T5 and $BERT_{base}$ models across three datasets: SQuAD 1.1, NarrativeQA, and BoolQ. The results in Table 5.8 demonstrate that our QA-Attack with $BERT_{base}$ consistently achieved superior performance

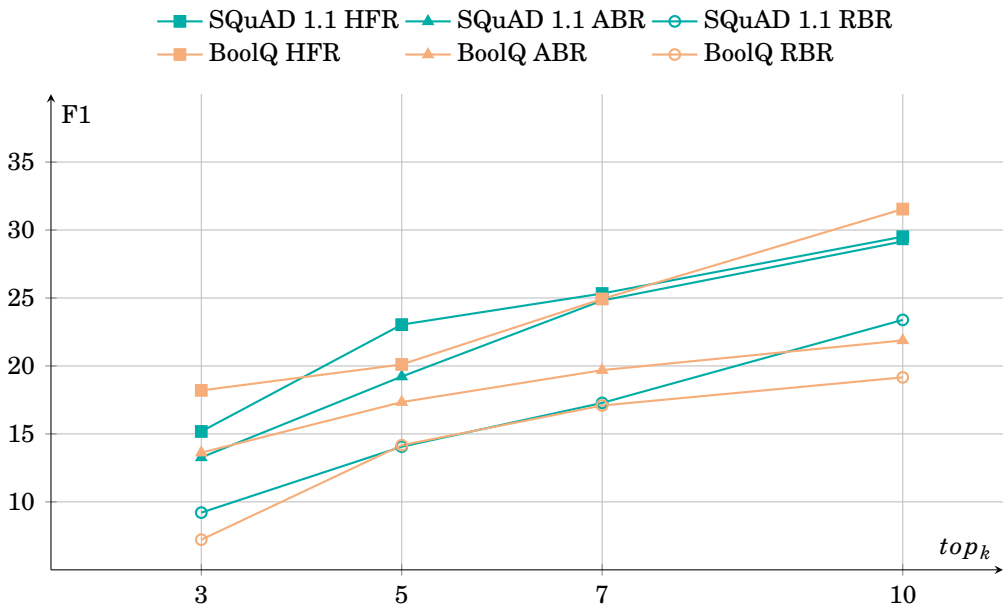

Figure 5.2: F1 score analysis for HFR, ABR, and RBR variants of QA-Attack using different top_k values, tested on datasets SQuAD 1.1 and BoolQ.

Table 5.8: EM scores for attacks on T5 and $\text{BERT}_{\text{base}}$ models using three distinct synonym generation methods. Lower scores indicate more effective attacks.

Methods	Victim Models	Datasets		
		SQuAD 1.1	NarrativeQA	BoolQ
HowNet	T5	14.22	7.25	29.08
	$\text{BERT}_{\text{base}}$	7.66	4.52	26.91
WordNet	T5	5.31	3.99	21.63
	$\text{BERT}_{\text{base}}$	7.23	5.67	19.35
$\text{BERT}_{\text{base}}$ (ours)	T5	**4.67**	**5.61**	**8.64**
	$\text{BERT}_{\text{base}}$	**6.42**	**3.86**	**18.39**

compared to other methods across all datasets and victim models.

On the other hand, we also examine the impact of parameter d in Synonym Selection, which determines the number of synonyms obtained from the Masked Language Model (MLM). Table 5.9 illustrates that as d increases from 1 to 3, F1 scores consistently decrease across all datasets, indicating improved attack performance. This trend suggests that a more aggressive setting (higher d) is more effective in compromising model accuracy across various datasets.

Table 5.9: F1 scores demonstrating QA-Attack's performance across five datasets under different d values (i.e., number of synonym candidates for substitutions).

	SQuAD 1.1	SQuAD V2.0	BoolQ	Narrative QA	NewQA
$d = 1$	8.52	14.72	19.22	7.63	10.66
$d = 2$	4.67	9.13	15.16	5.61	3.61
$d = 3$	2.17	7.26	11.43	3.71	3.27

Table 5.10: Performance metrics for different word candidate selection strategies against T5 model on SQuAD 1.1 dataset.

Methods	EM↓	SIM↑	Mod↓	PPL↓	GErr↓
TASA [3]	9.21	6.38	8.15	143	0.13
TMYC [4]	7.28	8.22	9.21	151	0.14
RobustQA [5]	5.89	6.03	8.35	147	0.15
T3 [6]	5.41	7.23	7.93	133	0.13
TextFooler [11]	10.60	6.29	8.17	136	0.14
QA-Attack (ours)	**5.68**	**5.91**	**7.24**	**125**	**0.12**

5.4.5.3 Texual Quality of Word Candidates

In our ablation study, detailed in Table 5.10, we investigate the quality of adversarial examples generated by various attack methods on the T5 model using the SQuAD 1.1 dataset. We evaluate our word replacement technique with encoder-decoder candidate generation (T3), as well as sentence-level modification methods (TASA, TMYC). The results indicate that our word-level synonym selection approach outperformed all other baselines. Notably, our word-level attack maintains a lower grammar error rate and higher linguistic fluency than alternative methods. Although RobustQA employs the same synonym selection strategy, it requires more word modifications to successfully attack the model and tends to produce more adventurous alterations.

Table 5.11: Time consumption (seconds per sample) for various methods and datasets. A lower value indicates better performance.

	Narrative QA	SQuAD 1.1	SQuAD V2.0	NewsQA	BoolQ
TASA [3]	28.77	15.82	18.25	10.72	–
TMYC [4]	25.61	12.75	16.33	9.21	7.42
RobustQA [5]	25.82	24.46	22.15	12.81	15.82
T3 [6]	26.52	21.37	28.38	14.74	7.93
QA-Attack (ours)	**23.51**	**10.61**	**12.38**	**8.32**	**7.22**

5.4.6 Platform and Efficiency Analysis

In this section, we evaluate QA-Attack's computational efficiency under base settings. We measure efficiency using time consumption per sample, expressed in seconds, where a lower value indicates superior performance. As shown in Table 5.11, the outcomes reveal that QA-Attack exhibits remarkable time efficiency, consistently outperforming baseline methods across both long-text (NarrativeQA) and short-text (SQuAD 1.1) datasets. This superior performance can be attributed to QA-Attack's innovative Hybrid Ranking Fusion (HRF) strategy, which effectively identifies vulnerable words within the text, significantly enhancing the speed of the attack process.

5.4.7 Adversarial Retraining

In this section, we investigate QA-Attack's potential for enhancing downstream models' accuracy. We employ QA-Attack to generate adversarial examples from SQuAD 1.1 training sets and incorporate them as supplementary training data. We reconstruct the training set with varying proportions of adversarial examples added to the raw training set. The retraining process with this augmented data aims to examine how test accuracy changes in response to the inclusion of adversarial examples. As illustrated in Fig. 5.3, re-training with adversarial examples slightly improves model performance when less than 30% of the training data consists of adversaries. However, performance

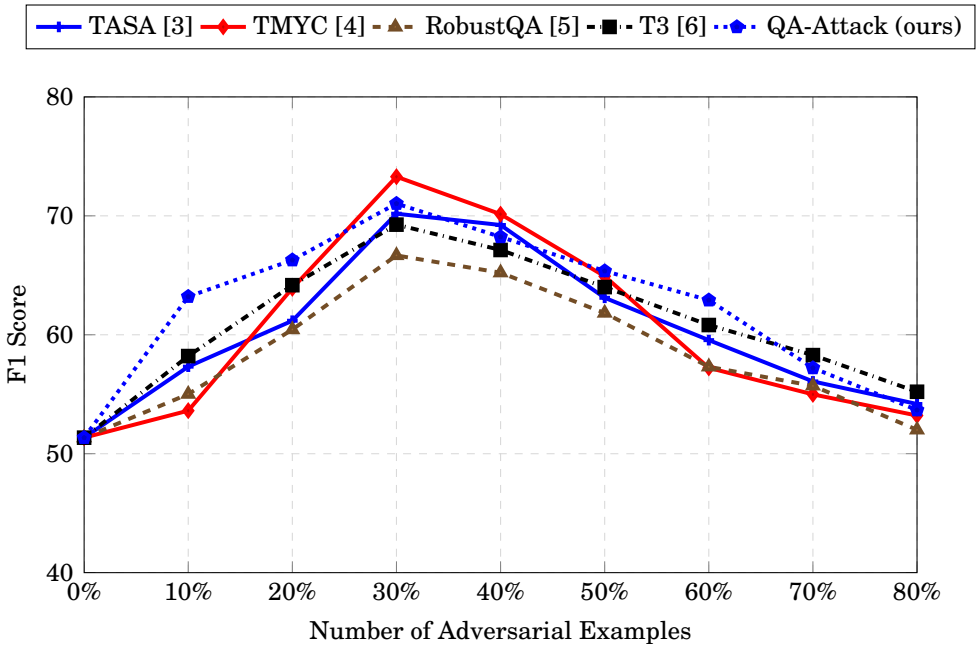

Figure 5.3: T5 model performance after retraining on SQuAD 1.1 dataset using diverse adversarial examples created by TASA [3], TMYC [4], RobustQA [5], T3 [6], and our QA-Attack method.

decreases when the proportion of adversaries exceeds 30%. This finding indicates that the optimal ratio of adversarial examples in training data needs to be determined empirically, which aligns with conclusions from previous attacking methods. To evaluate how re-training helps defend against adversarial attacks, we analyze the robustness of T5 models trained with varying proportions of adversarial examples (0%, 10%, 20%, 30%, 40%) from different attack methods, as shown in Fig. 5.3. A lower F1 score indicates higher model susceptibility to adversarial attacks. The attack performance of the re-trained model is shown in Fig. 5.4. It demonstrates that incorporating adversarial examples during training consistently improves model robustness, as evidenced by increasing F1 scores across all attack methods. Notably, QA-Attack emerges as the most effective approach, consistently outperforming other methods, with its advantage becoming particularly pronounced at higher percentages of adversarial training data.

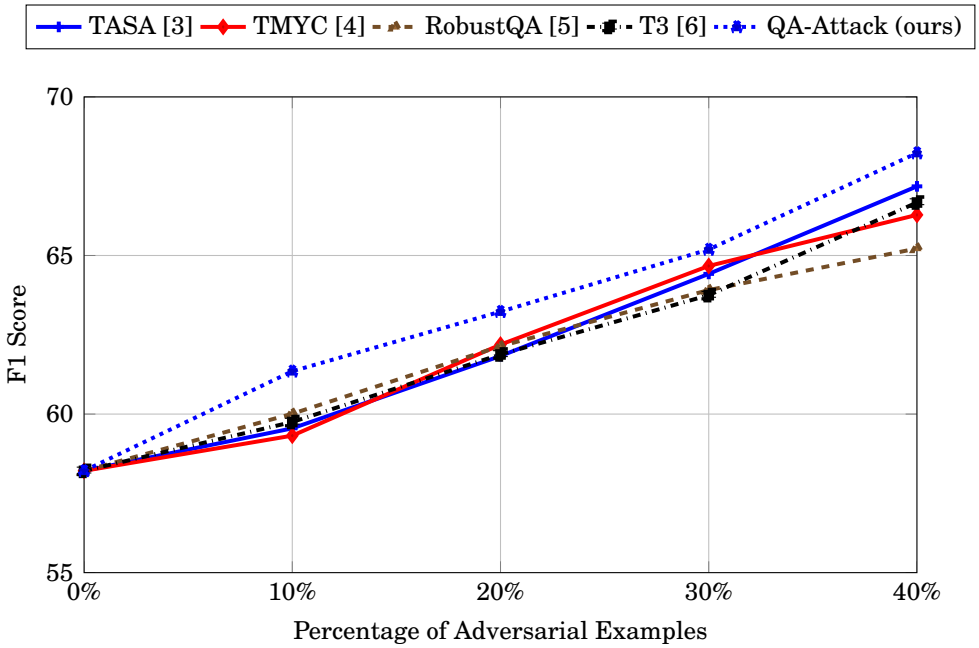

Figure 5.4: F1 scores of attacking T5 models retrained with increasing proportions of adversarial examples generated by baseline methods (TASA [3], TMYC [4], RobustQA [5], T3 [6]) and our QA-Attack.

5.4.8 Attacking Models with Defense Mechanism

Defending NLP models against adversarial attacks is crucial for maintaining the reliability of language processing systems in real-world applications [220]. To further analyze how attacks are performed under defense systems, we deploy two distinct defense mechanisms to investigate our attack performance under defense systems. The first is Frequency-Guided Word Substitutions (FGWS) approach [7], which excels at detecting adversarial examples. The second is Random Masking Training (RanMASK) [8], a technique that enhances model robustness through specialized training procedures. We perform the adversarial attack on T5 on datasets SQuAD 1.1, NarrativeQA and BoolQ, and the outcomes are collected in Table 5.12. The outcomes indicate that QA-Attack demonstrates superior adversarial robustness across multiple benchmark datasets, consistently outperforming existing methods against state-of-the-art defenses.

Table 5.12: Effectiveness of defense mechanisms (FGWS [7] and RanMASK [8]) against QA-Attack: EM scores of T5 model output answers across SQuAD 1.1, NarrativeQA, and BoolQ datasets. Lower scores indicate higher attack success against defenses.

Datasets	Defense	TASA	RobustQA	TMYC	T3	QA-Attack
SQuAD 1.1	FGWS [7]	34.71	39.42	28.51	24.11	**21.03**
	RanMASK [8]	32.17	39.78	44.81	41.09	**30.26**
Narrative QA	FGWS [7]	49.28	44.62	**37.21**	45.17	38.33
	RanMASK [8]	38.41	37.14	41.62	43.81	**34.47**
BoolQ	FGWS [7]	45.71	47.37	38.97	45.33	**38.34**
	RanMASK [8]	41.63	42.88	47.25	42.17	**40.51**

5.4.9 Transferability of Attacks

To evaluate our model's transferability, we test the adversarial samples generated for T5 on three distinct question-answering models: RoBERTa [92], DistilBERT [91], and MultiQA [100]. We also compare the transferability of three baseline methods: TASA, TextFooler, and T3, under identical experimental conditions. As shown in Fig. 5.5, QA-Attack effectively degrades other QA models' performance on both the NarrativeQA and BoolQ datasets. This suggests that the transferring attack performance of our QA-Attack consistently outperforms the baselines.

5.4.10 Parts of Speech Preference

To further understand the candidate words' distribution of our word-level attack, we examine its attacking preference in terms of Parts of Speech (POS), highlighting vulnerable areas within the input context. We use the Stanford POS tagger [221] to label each attacked word, categorizing them as *noun*, *verb*, *adjective (Adj.)*, *adverb (Adv.)*, and *others (e.g., pronoun, preposition, conjunction)*. Table 5.13 illustrates the POS preference of our QA-Attack compared to baseline methods in the base setting. For "informative queries" on SQuAD dataset, most attacking methods predominantly target *nouns*, while TASA shows a slight preference for *adverbs*. In the case of "boolean queries" on BoolQ

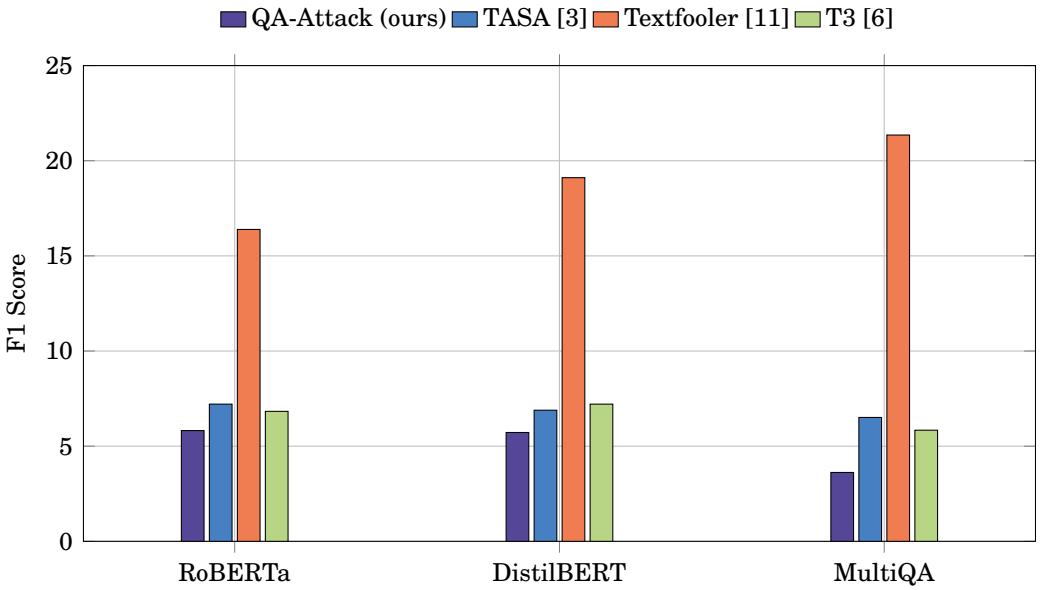

Figure 5.5: F1 scores for transfer attacks on three other QA models using adversarial samples generated for UnifiedQA. A lower value indicates better performance.

dataset, all methods frequently focus on *adjectives* and *adverbs*. Notably, our QA-Attack demonstrates a higher preference for the "others" category. Given that these parts of speech (pronouns, prepositions, and conjunctions) carry limited semantic content, we suggest that altering them may not significantly affect the linguistic or semantic aspects of prediction. However, such modifications could disrupt sequential dependencies, potentially compromising the contextual understanding of QA models and misleading their answers.

5.4.11 Robustness versus the Scale of Pre-trained Models

From the attacking results in Table 5.5 discussed in Sec 5.4.4, we recognize the limitation of our QA-Attack on $\text{BERT}_{\text{base}}$, with $L = 12$ and $H = 768$, which does not sufficiently support robust experimental outcomes. To address this issue and gain more comprehensive insights, we conducted experiments with four different sizes of BERT [28] models[3]:

[3]Different sizes of BERT models can be obtained from `https://github.com/google-research/bert/`

Table 5.13: POS preference with respect to choices of victim words among attacking methods. (TASA is incompatible with Boolean queries.)

Datasets	Methods	Noun	Verb	Adj.	Adv.	Others
SQuAD 1.1	TASA [3]	*35%*	12%	13%	**36%**	4%
	TMYC [4]	**47%**	*21%*	11%	5%	17%
	RobustQA [5]	**34%**	13%	*22%*	16%	15%
	TextFooler [11]	**44%**	13%	*23%*	8%	12%
	T3 [6]	**60%**	*17%*	6%	7%	10%
	QA-Attack (ours)	*34%*	9%	18%	3%	**36%**
BoolQ	TASA [3]	N/A	N/A	N/A	N/A	N/A
	TMYC [4]	14%	*19%*	12%	**35%**	20%
	RobustQA [5]	*19%*	14%	**27%**	23%	17%
	TextFooler [11]	**41%**	15%	*27%*	7%	10%
	T3 [6]	**42%**	*13%*	20%	16%	9%
	QA-Attack (ours)	10%	19%	*25%*	18%	**28%**

Table 5.14: A comparative analysis to attacking various sizes of BERT model on SQuAD 1.1 dataset. A lower value indicates better attack performance.

Versions	BERT tiny	BERT mini	BERT medium	BERT large
Size	L = 2 H = 128	L = 4 H = 256	L = 4 H = 256	L = 24 H = 1024
EM ↓	11.82	13.26	13.31	14.25
F1 ↓	5.67	6.35	6.42	7.24
SIM ↓	6.23	7.12	7.43	8.38

$\text{BERT}_{\text{tiny}}$, $\text{BERT}_{\text{mini}}$, $\text{BERT}_{\text{medium}}$, and $\text{BERT}_{\text{large}}$. Our findings, detailed in Table 5.14, demonstrate a positive correlation between model size and experimental robustness. The effectiveness of adversarial attacks decreases as the complexity and capacity of the BERT model increase, suggesting that deeper architectures provide better protection against adversarial perturbations.

5.5 Summary and Discussion

In this chapter, we presented QA-Attack, a novel approach that employs Hybrid Ranking Fusion (HRF) to conduct effective attacks by identifying and modifying critical tokens

in input text. By combining attention-based and removal-based ranking strategies, QA-Attack successfully disrupts model predictions while preserving semantic and linguistic coherence. Comprehensive experiments demonstrate that our method surpasses existing attack techniques in terms of attack success, fluency, and computational efficiency across multiple datasets, validating its effectiveness in compromising the robustness of state-of-the-art QA models. While QA-Attack reveals vulnerabilities in question-answering systems, these findings also provide valuable insights for enhancing model robustness. Future work will focus on developing defense mechanisms to mitigate these vulnerabilities. Additionally, we plan to extend QA-Attack to address more complex scenarios, including multiple-choice questions and multi-hop reasoning [222], ensuring our method remains effective for evaluating and improving QA system robustness in an evolving landscape of adversarial threats.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This chapter concludes with a summary of our key findings in Section 6.1, followed by proposed directions for future research in Section 6.2.

6.1 Conclusion

This thesis investigates the vulnerabilities of Natural Language Processing (NLP) and Computer Vision models, specifically focusing on adversarial attacks against abstractive summarization, image captioning, and question answering systems. Our research reveals critical insights into model weaknesses and introduces novel methodologies that both expose and enhance the robustness of contemporary deep learning systems. The key contributions are as follows:

1. First, we introduce an innovative paraphrasing-based attack framework for abstractive summarization models. Our approach addresses the challenge of attacking long-form text inputs by incorporating a sentence importance ranking mechanism based on ROUGE score differentials. By identifying and paraphrasing the most

influential sentences, our method generates adversarial examples that effectively deceive summarization models while maintaining semantic coherence. Experimental results demonstrate superior attack success rates compared to existing methods across multiple datasets.

2. Second, we propose a novel attention-based adversarial attack targeting image captioning systems. The method leverages an attention score to identify salient image regions and employs differential evolution to optimize perturbations, all within a black-box setting that requires no access to model gradients or architecture details. Through comprehensive evaluation, we show that our approach achieves state-of-the-art attack success rates while ensuring perturbations remain imperceptible to human observers.

3. Finally, we develop a hybrid word-level adversarial framework for question-answering systems that handle both boolean and informative queries. Our method combines attention-based and removal-based ranking strategies to identify vulnerable words, which are then replaced with contextually appropriate synonyms generated by a masked language model. This approach preserves linguistic fluency and semantic meaning while successfully misleading modern QA models, including T5 and BERT, achieving high attack success rates with minimal semantic distortion.

This research advances our understanding of adversarial vulnerabilities in textual and visual models, revealing their susceptibility to subtle, imperceptible attacks. Through innovative strategies that balance attack effectiveness with semantic and perceptual coherence, this work establishes new benchmarks for evaluating and enhancing AI system robustness. The methodological contributions offer practical insights for improving the security and reliability of AI applications in real-world contexts, such as automated services and accessibility tools.

6.2 Future Work

Building upon the findings presented in this thesis, several promising directions for future research emerge. These directions seek to expand the scope of adversarial attacks, deepen our understanding of model vulnerabilities, and enhance the robustness of AI systems across Natural Language Processing and Computer Vision tasks. By investigating emerging AI paradigms such as large language models, exploring attack transferability across diverse tasks, and developing innovative defense mechanisms, future studies can address both theoretical challenges and practical applications. Furthermore, leveraging insights from adversarial attacks to refine model architectures and training methodologies presents significant opportunities for advancing the reliability and security of AI systems in real-world deployments. The key areas for future investigation are outlined below:

- Adversarial Attacks on Large Language Models: Future research could extend adversarial attack strategies to large language models like GPT [56], T5 [16], and PaLM [223], focusing on their unique architectural vulnerabilities. This includes investigating how attacks on LLMs differ from traditional NLP models due to their scale, few-shot learning capabilities, and extensive pre-training. Exploring attacks in conversational, generative, and multi-turn dialogue contexts could reveal novel insights into model robustness. Additionally, testing adversarial strategies on LLM-based APIs deployed in real-world applications would bridge the gap between theoretical findings and practical implications.

- Transferability Across Modalities and Tasks: Building on current findings, future work could investigate the transferability of adversarial examples across various architectures and tasks. For NLP, this could include testing attacks on multi-modal systems, translation models, and domain-specific applications in legal and medical

contexts. In Computer Vision, transfer attacks could target advanced tasks like video captioning [224], 3D vision [225], and generative models such as GANs [226]. Understanding how adversarial examples generalize across different architectures, datasets, and training conditions would provide valuable insights into AI system robustness.

- Potential Defense Systems: The findings from this thesis can inform the development of innovative defense mechanisms. This includes designing adversarial training approaches specifically tailored to vulnerabilities identified in summarization, image captioning, and QA models. Hybrid defense systems that combine adversarial detection with robust optimization techniques could provide comprehensive protection. Furthermore, incorporating explainable AI techniques and uncertainty quantification could enhance model resistance to adversarial perturbations. Collaboration with industry partners to deploy and validate these defenses in operational settings would demonstrate their practical effectiveness.

- Improvements to CV and NLP Models Based on Attack Results: Insights from these adversarial attacks could guide improvements in model architectures and training methodologies. For NLP, this involves developing models with enhanced contextual understanding and redundant decision-making pathways. In Computer Vision, incorporating adaptive attention mechanisms could help models resist adversarial perturbations. Cross-task learning approaches, where models are trained to identify and counter adversarial patterns across multiple tasks, could provide breakthrough advances. Additionally, attack insights could inspire novel pretraining objectives and regularization techniques that simultaneously enhance performance and robustness.

- Societal Threat Quantification: While this thesis evaluates attacks using semantic similarity and NLP metrics, these technical measures inadequately capture

real-world consequences. Future work should develop frameworks mapping attack characteristics to domain-specific risks-including market volatility in finance, treatment errors in healthcare, and electoral influence in politics. Targeted demonstration experiments could strengthen these frameworks by simulating real-world failures, such as compromised content moderation systems or autonomous vehicles misinterpreting adversarial captions. By establishing quantitative relationships between attack metrics and societal harm indicators, researchers could better prioritize defenses for high-stakes applications and guide deployment decisions in sensitive domains.

BIBLIOGRAPHY

[1] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization, 2015.

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

[3] Yu Cao, Dianqi Li, Meng Fang, Tianyi Zhou, Jun Gao, Yibing Zhan, and Dacheng Tao. TASA: Deceiving question answering models by twin answer sentences attack. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11975–11992, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[4] Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401, 2019.

[5] Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 976–986, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[6] Boxin Wang, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li. T3: Tree-autoencoder constrained adversarial text generation for targeted attack. *arXiv preprint arXiv:1912.10375*, pages 6134–6150, November 2020.

[7] Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. Frequency-guided word substitutions for detecting textual adversarial examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 171–186, Online, April 2021. Association for Computational Linguistics.

[8] Jiehang Zeng, Xiaoqing Zheng, Jianhan Xu, Linyang Li, Liping Yuan, and Xuanjing Huang. Certified robustness to text adversarial attacks by randomized [mask]. *Computational Linguistics*, 49(2):395–427, jun 2023.

[9] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[10] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *International Conference on Learning Representations (ICLR)*, 2018.

[11] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment, 2019.

[12] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, 2016.

[13] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, 2017.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, pages 7871–7880, July 2020.

[15] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR, 2020.

[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[17] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, 2020.

[18] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, 2019.

[19] Jianbo Cheng, Yi Zhao, Yun Han, Shuguang Zhou, and Tao Zhang. Adversarial attack and defense strategies for deep text classification models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7210–7225, 2020.

[20] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.

[21] Chien-Chou Lin, Chih-Hung Kuo, and Hsin-Te Chiang. Cnn-based classification for point cloud object with bearing angle image. *IEEE Sensors Journal*, 22(1):1003–1011, 2022.

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[26] Xin Zhang, An Yang, Sujian Li, and Yizhong Wang. A survey on machine reading comprehension: Tasks, evaluation metrics, and benchmark datasets. *Applied Sciences*, 10(21):7640, 2020.

[27] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[29] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 33:1877–1901, 2020.

[30] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[31] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardnern, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*, pages 2153–2162, November 2019.

[32] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, 2018.

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[34] Animesh Chakraborty, Tamal Alam, Somesh Dey, Arunava Chattopadhyay, and Dipankar Mukhopadhyay. Adversarial attacks and defenses: A survey. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(2):126–139, 2018.

[35] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. *Machine Learning and Knowledge Discovery in Databases*, pages 387–402, 2013.

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[37] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1467,Äì1474, Madison, WI, USA, 2012. Omnipress.

[38] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3520,Äì3532, Red Hook, NY, USA, 2017. Curran Associates Inc.

[39] Luis Muñoz González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, page 27,Äì38, New York, NY, USA, 2017. Association for Computing Machinery.

[40] Chaofei Yang, Qing Wu, Hai Li, and Yiran Chen. Generative poisoning attack method against neural networks. *arXiv preprint arXiv:1703.01340*, 2017.

[41] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[42] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2(3):4, 2017.

[43] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 1528,Äì1540, New York, NY, USA, 2016. Association for Computing Machinery.

[44] Tom B. Brown, Dandelion Mane, Aurko Roy, Martin Abadi, and Justin Gilmer. Adversarial patch, 2018.

[45] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples, 2018.

[46] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics, July 2002.

[47] Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos, and Constantine D. Spyropoulos. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 160,Äì167, New York, NY, USA, 2000. Association for Computing Machinery.

[48] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[49] Harry Zhang. The optimality of naive bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 562–567, 2004.

[50] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML)*, pages 137–142. Springer, 1998.

[51] Tong Zhang and Francis J Oles. Text categorization based on regularized linear classification methods. *Information Retrieval*, 4(1):5–31, 2001.

[52] Robin M. Schmidt. Recurrent neural networks (rnns): A gentle introduction and overview, 2019.

[53] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[54] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

[55] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1480–1489, 2016.

[56] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Technical Report*, 2018.

[57] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[58] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. In *Proceedings 2019 Network and Distributed System Security Symposium*, NDSS 2019. Internet Society, 2019.

[59] Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online, July 2020. Association for Computational Linguistics.

[60] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 6193–6202, Online, November 2020. Association for Computational Linguistics.

[61] Paul Michel, Xian Li, and Graham Neubig. On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3103–3114, 2019.

[62] Kang Gu, Ehsanul Kabir, Neha Ramsurrun, Soroush Vosoughi, and Shagufta Mehnaz. Towards sentence level inference attack against pre-trained language models. *Proceedings on Privacy Enhancing Technologies*, 2023.

[63] Xinyang Zhou, Wei Wang, Javid Ebrahimi, Zhifang Deng, and Dejing Dou. Learning to generate adversarial examples for text classification using syntax-guided adversarial networks. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1134–1145, 2019.

[64] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

[65] Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020.

[66] Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, 2007.

[67] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27:3104–3112, 2014.

[68] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[70] Yinhan Liu, Jiatao Wang, Hanzi Li, et al. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[71] Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3601–3608, Apr. 2020.

[72] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2018.

[73] Zhiting Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

[74] Wei Chen et al. Cross-lingual transfer attacks on machine translation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2485–2497, 2021.

[75] Hui Lin and Vincent Ng. Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9815–9822, 2019.

[76] Dragomir R Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938, 2004.

[77] Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 93–98, 2016.

[78] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[79] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforcement learning approach to abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2018.

[80] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[81] Ziqiang Cao, Wenjie Wang, Sujian Li, and Furu Li. Faithful to the original: Fact aware neural abstractive summarization. *arXiv preprint arXiv:1805.05245*, 2018.

[82] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.03760*, 2020.

[83] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguis-*

tics, pages 1085–1097, Florence, Italy, July 2019. Association for Computational Linguistics.

[84] Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020.

[85] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1875–1885, 2018.

[86] Aly M Kassem and Sherif Saad. Finding a needle in the adversarial haystack: A targeted paraphrasing approach for uncovering edge cases with minimal distribution distortion. *arXiv preprint arXiv:2401.11373*, 2024.

[87] T Ter-Hovhannisyan, H Aleksanyan, and K Avetisyan. Adversarial attacks on language models: Wordpiece filtration and chatgpt synonyms. *Journal of Mathematical Sciences*, 285(2):210–220, 2024.

[88] Marco Antonio Calijorne Soares and Fernando Silva Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, 32(6):635–646, 2020.

[89] Gulsum Yigit and Mehmet Fatih Amasyali. From text to multimodal: A comprehensive survey of adversarial example generation in question answering systems. *Knowledge and Information Systems*, 66:7165–7204, 2024.

[90] Zhen Wang. Modern question answering datasets and benchmarks: A survey. *arXiv preprint arXiv:2206.15030*, 2022.

[91] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, Vancouver, Canada, 2019. NeurIPS.

[92] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China.

[93] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[94] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, November 2020. Association for Computational Linguistics.

[95] Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online, June 2021. Association for Computational Linguistics.

[96] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, page 4208,Äì4215. International Joint Conferences on Artificial Intelligence Organization, July 2018.

[97] Sara Rosenthal, Mihaela Bornea, and Avirup Sil. Are multilingual bert models robust? a case study on adversarial attacks for multilingual question answering. *arXiv preprint arXiv:2104.07646*, 2021.

[98] Jiyao Li and Wei Liu. Summarization attack via paraphrasing (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16250–16251, 2023.

[99] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[100] Alon Talmor and Jonathan Berant. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy, July 2019. Association for Computational Linguistics.

[101] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, and Diego Andina. Deep learning for computer vision: A brief review. *Intell. Neuroscience*, 2018, January 2018.

[102] Xingshuai Dong and Massimiliano L Cappuccio. Applications of computer vision in autonomous vehicles: Methods, challenges and future directions. *arXiv preprint arXiv:2311.09093*, 2023.

[103] Marco Leo, Pierluigi Carcagni, Pier Luigi Mazzeo, Paolo Spagnolo, Dario Cazzato, and Cosimo Distante. Analysis of facial information for healthcare applications: A survey on computer vision-based approaches. *Information*, 11(3), 2020.

[104] Ansam A. Abdulhussein, Hasanien Kariem Kuba, and Alaa Neamah Azeez Alanssari. Computer vision to improve security surveillance through the identification of digital patterns. In *2020 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, pages 1–5, 2020.

[105] Zahid Mahmood, Tauseef Ali, Shahid Khattak, Laiq Hasan, and Samee U Khan. Automatic player detection and identification for sports entertainment applications. *Pattern Analysis and Applications*, 18:971–982, 2015.

[106] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[107] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[108] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[109] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[110] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[111] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[112] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[113] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum, 2017.

[114] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[115] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[116] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[117] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 10–17, 2018.

[118] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without

training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec)*, pages 15–26, 2017.

[119] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani Srivastava. Genattack: Practical black-box attacks with gradient-free optimization. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, pages 1111–1119, 2019.

[120] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*, pages 506–519, 2017.

[121] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[122] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[123] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. *European Conference on Computer Vision (ECCV)*, pages 21–37, 2016.

[124] Hao Liu, Jianan Li, Yang Song, Linjun Wang, and Qinghua Hu. Adversarial attacks against object detection models. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4056–4064, 2018.

[125] Qianyu Zhao, Quanlong Zou, and Shengping Wang. Targeted adversarial attacks for deep object detection models. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1558–1569, 2019.

[126] Yao Duan, Jing Li, Qiang Zhou, and Wei Wang. Adversarial transfer attacks for object detection models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 110–117, 2020.

[127] Xinlei Cheng, Xiaolong Wang, and Kaiming Lin. Query-efficient black-box adversarial attacks for object detection models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[128] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[129] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[130] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659, 2016.

[131] Jianwei Yang, Jiasen Lu, Dhruv Batra, and Devi Parikh. Object relationship detection with sentence regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12212–12221, 2019.

[132] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7008–7024, 2017.

[133] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[134] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Duane Boning, Cho-Jui Hsieh, and Xiaolin Hu. Show-and-fool: Crafting adversarial examples for neural image captioning. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5499–5507, 2018.

[135] Zhenqin Yin, Yue Zhuo, and Zhiqiang Ge. Transfer adversarial attacks across industrial intelligent systems. *Reliability Engineering & System Safety*, 237:109299, 2023.

[136] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

[137] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017.

[138] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Mubarak Shah, and Ajmal Mian. Language model agnostic gray-box adversarial attack on image captioning. *IEEE Transactions on Information Forensics and Security*, 18:626–638, 2022.

[139] Xinghao Yang, Weifeng Liu, Shengli Zhang, Wei Liu, and Dacheng Tao. Targeted attention attack on deep learning models in road sign recognition. *IEEE Internet of Things Journal*, 8(6):4980–4990, 2020.

[140] Zhizhou Yin, Fei Wang, Wei Liu, and Sanjay Chawla. Sparse feature attacks in adversarial learning. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1164–1177, 2018.

[141] Aneesh Sreevallabh Chivukula and Wei Liu. Adversarial learning games with deep learning models. In *2017 international joint conference on neural networks (IJCNN)*, pages 2758–2767. IEEE, 2017.

[142] Jiaqi Li, Cihang Xie, Mingkui Tan, and Boqing Gong. Improving the transferability of adversarial examples with advanced diversity-ensemble method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9002–9012, 2022.

[143] Zongze Wang, Jiahao Zhang, and Xiaoyan Zhu. Improving transferability of adversarial examples with virtual step and auxiliary gradients. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1651–1657, 2022.

[144] Fangzhou Suya, Ananya Bhattad, Yao-Yuan Li, Prateek Khosla, Liam Fowl, Micah Goldblum, John P. Dickerson, and Tom Goldstein. Visual attention enhances adversarial robustness. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[145] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2739, 2019.

[146] Hongge Wu, Yisen Xia, Chao Wang, and Deng Cai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1161–1169, 2020.

[147] Rainer Storn and Kenneth Price. Differential evolution,Äîa simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.

[148] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260, 2017.

[149] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, and Qi Ju. Improving image captioning with conditional generative adversarial nets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8142–8150, Jul. 2019.

[150] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[151] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Show-and-fool: Crafting adversarial examples for neural image captioning. *arXiv preprint arXiv:1712.02051*, 2, 2017.

[152] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. Exact adversarial attack to image captioning via structured output learning with latent variables, 2019.

[153] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

[154] Xinghao Yang, Weifeng Liu, Dacheng Tao, and Wei Liu. Besa: Bert-based simulated annealing for adversarial text attacks. In *IJCAI*, pages 3293–3299, 2021.

[155] Yuxin Huang, Zhengtao Yu, Junjun Guo, Zhiqiang Yu, and Yantuan Xian. Legal public opinion news abstractive summarization by incorporating topic information. *International Journal of Machine Learning and Cybernetics*, 11:2039–2050, 2020.

[156] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388, 2021.

[157] Mahsa Momenzad, Babak Majidi, and Mohammad Eshghi. Deep summarization of academic textbooks for adaptive gamified virtual learning environments. In *2018 2nd National and 1st International Digital Games Research Conference: Trends, Technologies, and Applications (DGRC)*, pages 88–94. IEEE, 2018.

[158] Mohammed Elsaid Moussa, Ensaf Hussein Mohamed, and Mohamed Hassan Haggag. A survey on opinion summarization techniques for social media. *Future Computing and Informatics Journal*, 3(1):82–109, 2018.

[159] Praveen Kumar Katwe, Aditya Khamparia, Deepak Gupta, and Ashit Kumar Dutta. Methodical systematic review of abstractive summarization and natural language processing models for biomedical health informatics: Approaches, metrics and challenges. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.

[160] Haoran Zhang and Ying Li. Adversarial attacks against deep learning-based abstractive text summarization systems. In *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM)*, pages 767–776. IEEE, 2020.

[161] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[162] Ion Androutsopoulos and Prodromos Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187, 2010.

[163] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96, 2016.

[164] Colin Raffel, Noam Shazeer, Adam Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[165] Tao Xie, Yue Xu, and Xiaoguang Zhang. '. *Journal of Financial Markets*, 55:100–118, 2021.

[166] Yanshan Wang, Sunghwan Sohn, Sijia Liu, Feichen Shen, and Liwei Wang. Medical text summarization using large language models: A systematic review. *Journal of Biomedical Informatics*, 138:104–117, 2023.

[167] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. Assessing the risks of 'infodemics' in response to covid-19 epidemics. *Nature Human Behaviour*, 4(12):1285–1293, 2020.

[168] Gordon Pennycook and David G Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019.

[169] James N Druckman. Evaluating framing effects. *Journal of Economic Psychology*, 22(1):91–101, 2001.

[170] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[171] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don‘t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, oct-nov 2018. Association for Computational Linguistics.

[172] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693,Äì1701, Cambridge, MA, USA, 2015. MIT Press.

[173] Mahdi Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 76–81. Association for Computational Linguistics, 2018.

[174] Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084. Association for Computational Linguistics, 2019.

[175] Jaehyung Kim, Jisun Kim, and Alexander M. Rush. Abstractive summarization of reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2519–2531. Association for Computational Linguistics, 2019.

[176] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In Fei Liu and Thamar Solorio, editors, *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[177] Yunpeng Chen, Jiliang Liu, and Xiangyu Zhang. Detecting and defending against adversarial examples in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8926–8935, 2023.

[178] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2267–2281, 2019.

[179] Tianyu Zhang, Xuelong Xu, and Shuai Wang. Adversarial robustness of vision-language models: A comprehensive survey. *ACM Computing Surveys*, 55(4):1–35, 2022.

[180] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.

[181] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016.

[182] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

[183] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):539–559, 2023.

[184] Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, Anan Liu, Zhendong Mao, Zhenyu Chen, and Xingyu Gao. Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):43–51, 2022.

[185] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV*, pages 740–755. Springer, 2014.

[186] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *The Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[187] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections, 2022.

[188] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. Expansionnet v2: Block static expansion in fast end to end training for image captioning, 2022.

[189] Mohammad Nuruzzaman and Omar Khadeer Hussain. A survey on chatbot implementation in customer service industry through deep neural networks. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, pages 54–61, 2018.

[190] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021.

[191] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[192] Jorge Martinez-Gil. A survey on legal question–answering systems. *Computer Science Review*, 48:100552, 2023.

[193] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, page 555,Äì565, New York, NY, USA, 2017. Association for Computing Machinery.

[194] Huoyuan Dong, Jialiang Dong, Shuai Yuan, and Zhitao Guan. Adversarial attack and defense on natural language processing in deep learning: A survey and perspective. In *International conference on machine learning for cyber security*, pages 409–424. Springer, 2022.

[195] Jigna J Hathaliya, Sudeep Tanwar, and Priyanka Sharma. Adversarial learning techniques for security and privacy preservation: A comprehensive review. *Security and Privacy*, 5(3):e209, 2022.

[196] Hui Sun, Tianqing Zhu, Zhiqiu Zhang, Dawei Jin, Ping Xiong, and Wanlei Zhou. Adversarial attacks against deep generative models on data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3367–3388, 2021.

[197] Karan Singhal, Shekoofeh Azizi, Tien Tu, and et al. Large language models encode clinical knowledge. *Nature*, 2023.

[198] Ning Zheng, Bingbing Tan, Huan Zhang, Xitong Wu, and Zhiyuan Yu. Does gpt understand statutory reasoning? In *Proceedings of ICAIL*, 2023.

[199] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States, July 2022. Association for Computational Linguistics.

[200] Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online, November 2020. Association for Computational Linguistics.

[201] Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. Penalizing confident predictions on largely perturbed inputs does not improve out-of-distribution generalization in question answering. In *Proceedings of the Workshop on Knowledge Augmented Methods for NLP (KnowledgeNLP) at AAAI 2023*, 2023.

[202] Tassilo Klein and Moin Nabi. Learning to answer by learning to ask: Getting the best of gpt-2 and bert worlds. *arXiv preprint arXiv:1911.02365*, 2019.

[203] Pietro Bongini, Federico Becattini, and Alberto Del Bimbo. Is gpt-3 all you need for visual question answering in cultural heritage? In *European Conference on Computer Vision*, pages 268–281. Springer, 2022.

[204] Fares Antaki, Daniel Milad, Mark A Chia, Charles-Édouard Giguère, Samir Touma, Jonathan El-Khoury, Pearse A Keane, and Renaud Duval. Capabilities of gpt-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. *British Journal of Ophthalmology*, 108(10):1371–1378, 2024.

[205] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[206] Priyabrata Karmakar, Shyh Wei Teng, and Guojun Lu. Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition. *Intelligent Systems with Applications*, 23:200406, 2024.

[207] Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291,Äì4308, October 2021.

[208] Weimin Lyu, Songzhu Zheng, Lu Pang, Haibin Ling, and Chao Chen. Attention-enhancing backdoor attacks against BERT-based models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10672–10690, Singapore, December 2023. Association for Computational Linguistics.

[209] Mingze Ni, Ce Wang, Tianqing Zhu, Shui Yu, and Wei Liu. Attacking neural machine translations via hybrid attention learning. *Mach. Learn.*, 111(11):3977,Äì4002, November 2022.

[210] Tong Xiao and Jingbo Zhu. Introduction to transformers: an nlp perspective, 2023.

[211] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR)*, 2013.

[212] Zhendong Dong and Qiang Dong. Hownet-a hybrid language and knowledge resource. In *International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003*, pages 820–824. IEEE, 2003.

[213] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[214] Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

[215] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[216] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long and Short Papers), pages 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[217] D. Naber. *A Rule-Based Style and Grammar Checker*. GRIN Verlag, 2003.

[218] Katharina Kann, Sascha Rothe, and Katja Filippova. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[219] George A. Miller. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[220] Shreya Goyal, Sumanth Doddapaneni, Mitesh M Khapra, and Balaraman Ravindran. A survey of adversarial defenses and robustness in nlp. *ACM Computing Surveys*, 55(14s):1–39, 2023.

[221] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, page 173,Äì180, USA, 2003. Association for Computational Linguistics.

[222] Jianxing Yu, Wei Liu, Shuang Qiu, Qinliang Su, Kai Wang, Xiaojun Quan, and Jian Yin. Low-resource generation of multi-hop reasoning questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6729–6739, Online, July 2020. Association for Computational Linguistics.

[223] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Se-

bastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1), March 2024.

[224] Moloud Abdar, Meenakshi Kollati, Swaraja Kuraparthi, Farhad Pourpanah, Daniel McDuff, Mohammad Ghavamzadeh, Shuicheng Yan, Abduallah Mohamed, Abbas Khosravi, Erik Cambria, et al. A review of deep learning for video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[225] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.

[226] Abdul Jabbar, Xi Li, and Bourahla Omar. A survey on generative adversarial networks: Variants, applications, and training. *ACM Comput. Surv.*, 54(8), October 2021.