Summarization Attack via Paraphrasing (Student Abstract)

Jiyao Li and Wei Liu

University of Technology Sydney, 15 Broadway, Ultimo NSW 2007, Australia
Jiyao.Li-1@student.uts.edu.au, Wei.Liu@uts.edu.au

Abstract

Many natural language processing models are perceived to be fragile on adversarial attacks. Recent work on adversarial attack has demonstrated a high success rate on sentiment analysis as well as classification models. However, attacks to 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 paper, we propose a paraphrasing-based attack method to attack summarization models. We first rank the sentences in the document according to their impacts to summarization. Then, we apply paraphrasing procedure to generate adversarial samples. Finally, we test our algorithm on benchmarks datasets against others 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.

Introduction

Adversarial attack has been proven to be effective to invalidate neural networks with small modifications on inputs. In many existing models, gradient-based and Masked Language Model (MLM) are adopted to generate adversarial samples by perturbing tokens and deceiving a model to predict a wrong label. The perturbed inputs are indistinguishable for humans but can cheat the deep learning models to derive wrong predictions. Recent studies on adversarial attack dedicates on misleading the model to generate a wrong label by applying word-level substitution (Yang et al. 2021). Unlike sentences, it is harder to construct adversarial examples for document-level summary models. In this paper, we propose to attack sequence-level abstractive summarization models with paraphrasing candidate generation. The primary challenge is balance perturbation and attack performance. The algorithm will first rank the importance of sentences to search best victims. We then apply baseline paraphrasing model to replace some of the influential sentences. Our contributions are as followed: (i) Abstractive models were utilised to rank the sequences importance of a given text, (ii) Paraphrasing approaches are applied on generating candidates, and (iii) Experiments were conducted on real world datasets to examine our algorithm.

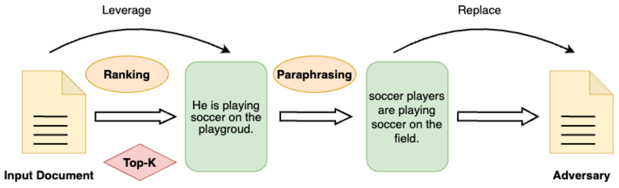
Leverage
Replace
Ranking
Paraphrasing
Top-K
He is playing soccer on the playgroud.
soccer players are playing soccer on the field.
Input Document
Adversary

Figure 1: The brief workflow of our candidate selection with paraphrasing. For each target document, we rank out the sentences in reverse order, and rebuild it by replacing the top-k sentences with sentence produced with paraphrasing model.

Proposed Method

The first section of the method is Sentence Importance Ranking (the completion steps of our proposed method is shown in Algorithm 1). Predominantly, results in summarization works are evaluated by ROUGE score (Lin 2004) which calculate the relationship between generated abstracts and reference abstracts. Inspired by to extremely improve the algorithm's success rate, we perform a ranking procedure in terms of sentence importance. In the procedure, sentences are deleted independently to rebuild the document. The modified documents are then input into summarization models to extract summary. Sentences were ranked out according to the ROUGE score before and after deleting the target sentence. Let $a = [s_1, ..., s_i, ...]$ denote input document, F is the summarization generating model applied in ranking, R denotes ROUGE score function, and D_0, D_i refer to document with or without sentence i respectively. As shown in Equation 1, the sentence importance is defined as:

$$G_i = R(F_s(D_i)) - R(F_s(D_0)) \tag{1}$$

The next step is Sentence Replacement. Previous text attack methods obtain substitutes from the prediction of Masked Language Model. It requires sophisticated selection strategy from predicted words. However, summarization models mine the semantic relationship between sentences and, substituting words has low effect on misleading summary models. Therefore, a hypothesis to build adversary document is to simply remove influential sequences, but lacking bunch of topic sentence would also mislead human readers. Hence in

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; $\alpha$ is the maximum number of sentences to perturb; Function to get adversarial sample$F_{adv}$.
Output : Adversarial samples $adv$
/* Sentence Importance Ranking */
$sentence\_rank \leftarrow [\,]$
for $s_i$ in $a$ do
    Build document without sentence $s_i$
    Calculate ROUGE score $R_i$ without $si$
    $sentence\_rank.append(R_0 - R_i)$  ▷ Calculate the ROUGE difference
end for
$size \leftarrow len(D)$
$length \leftarrow 0$
for $a$ in $D$ do
    $length += len(a)$
end for
$ave\_length \leftarrow length/size$
$threshold \leftarrow min(ave\_length * K, alpha)$
$ranked \leftarrow sentence\_rank[: threshold]$
/* Generating adversarial example */
for $s_i$ in $ranked$ do
    $s_{adv} \leftarrow F_{adv}(s_i)$  ▷ Generating adversarial samples
    $a_{adv} \leftarrow [s_1, ..., s_{adv}, ..., s_n]$
end for
return $adv$
```

our approach, we tested top-K from 10% to 30% of document length to balance the attack cost and semantics of document. The chosen paraphrasing model uses hierarchical sketches to build semantic preserved sentences with various vocabulary and grammar. It is known that paraphrasing approach maintain the grammatical and semantic feature of each sentence with combination sentence structure.

Experiment Results and Analysis

To evaluate the algorithm, several baselines deep learning models were imported for comparison of sentence importance rank and candidates generating. For the target summarization model, Pegasus (Zhang et al. 2019) from Google were chosen. In sentence importance ranking section, Pegasus, textrank (Mihalcea and Tarau 2004), and tf-idf summarization models were chosen. We chose textrank and tf-idf for ranking procedure because their summary are directly relied on sentence importance. In order to examine the performance of paraphrasing attack, translation and deleting methods are introduced to obtain candidates in the experiment. In translation scheme, transformers encoder-decoder model were picked to translate input sentence to German and back to English text. As for the deleting method, the top-k candidates were removed from input. The algorithm was investigate on XSum dataset (Zhang et al. 2019) pre-trained in Google Pegasus model. 1000 samples were randomly selected from train split of the dataset, the top-k were finally set to 20% sentences with 5 sentences as upper bound. We provide our code for reproducibility of the experiments [1].

Attack method	Rank method	sim	R_{diff}
Translation	tf-idf	68.2	10.3
	Texrank	**76.5**	11.3
	Pegasus	75.8	14.3
Deleting	tf-idf	N/A	13.2
	Texrank	N/A	12.9
	Pegasus	N/A	17.8
Paraphrasing	tf-idf	67.8	13.3
	Texrank	71.5	12.0
	Pegasus	72.8	**18.4**

Table 1: Comparisons of Attack Results.

The success of each attack is measured by the decrease of ROUGE F1 score on the summary produced on modified document on Pegasus, denoted as R_{diff}. The semantic similarity (sim) comparing word vectors between original and adversarial sentences were considered in experiments. As shown in Table 1, the ROUGE score number decreased by 18.4 on Pegasus ranking under paraphrasing scenario which was the best. It can be observed that semantic similarity in translation is higher than paraphrasing attack, however the R_{diff} is lower than both deleting and paraphrasing. According to the experiment results, the proposed method reached superior performance on summarization model attacks.

Conclusion and Future Work

In this research, we studied adversarial attack on summarization models, and proposed Pegasus ranking and paraphrasing adversary generation strategy. Experiments on two different datasets and baseline algorithms demonstrate the effeteness on generating adversarial samples for abstractive summarization models of our approach. In the future, more tests are needed on summarization models to optimize the Pegasus ranking efficiency.

References

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Mihalcea, R.; and Tarau, P. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. Barcelona, Spain: Association for Computational Linguistics.

Yang, X.; Liu, W.; Tao, D.; and Liu, W. 2021. BESA: BERT-based Simulated Annealing for Adversarial Text Attacks. In *IJCAI*, 3293–3299.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. J. 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv:1912.08777.

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