UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

The Attention Mechanism in Vision and Language Analysis

by

Guang Li

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2021

Certificate of Authorship/Originality

I, Guang Li declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy , in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own 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.

Date: Mar 15, 2021

ABSTRACT

The Attention Mechanism in

Vision and Language Analysis

by

Guang Li

In psychology, attention is the cognitive process of concentrating on a particular aspect of information while ignoring other perceivable elements. Human visual/linguistic perceptions can eliminate distracting factors and concentrate on the most relevant components with psychological attention's guidance. In representation learning, an operator imitating the psychological attention mechanism in feature aggregation is also in demand. CNN and RNN are the fundamental frameworks in representation learning, and they have aptitudes for processing structured data. However, the recurrent nature of RNN dilutes the long-term information as the sequence length grows. Moreover, with a fixed kernel size, the convolution has difficulty modeling the long-range relations between pixels. In order to solve the problems above, the attention mechanism is introduced to representation learning. The attention operator treats candidate elements as a set without considering their order or position; therefore, the attention-based models can concentrate on the relevant elements flexibly and free from the bondage of data structure.

This thesis mainly focuses on the attention mechanism for vision and language analysis and researches 1) multimodal attention for image captioning, 2) the positional awareness in attention, 3) local attention for multi-level feature fusion. We begin with the benchmark vision & language task – image captioning, and investigate how to extend the transformer model with the ability to leverage multimodal information simultaneously. Going beyond the attention mechanism exploring content similarity solely, we develop the bilateral attention mechanism, which is equipped with positional

awareness. Comprehensive experiments are conducted on two representative tasks, i.e., semantic segmentation and machine translation, and the encouraging results show that position-awareness is a beneficial supplement for the attention mechanism. Furthermore, We explore if it is feasible to replace the standard convolution with a local attention-based operator based on the attention with positional awareness. Besides, the dynamic local operator demonstrates its adaptiveness in multi-level feature fusion for semantic segmentation. Finally, the thesis is concluded with some future directions on the attention mechanism.

Dissertation directed by Professor Yi Yang
Centre for Artificial Intelligence, School of Software

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Yi Yang, for the guidance, support, and advice he has provided throughout my Ph.D. Study. Prof. Yi gave me enough freedom to pursue the research topics I was interested in and always expressed his patience and encouragement, which is a great fortune in my Ph.D. study. I also want to express my gratitude to my advisors and collaborators: Dr. Yunchao Wei, Dr. Linchao Zhu, Dr. Ping Liu, Dr. Yahong Han, and Dr. Wu Liu. I was fortunate to work with them and engage in intellectual discussions with them.

I would also like to thank my colleagues at the University of Technology Sydney and the ReLER lab. I am very grateful to have worked with so many wonderful people during my Ph.D. study, who have provided so many insightful discussions on my research and various kinds of help in my personal life.

I would also like to thank Data to Decision CRC for supporting my research.

Finally, and most essentially, I am grateful for all the support from my parents, sister, and best friends. They are the source of my strength.

Guang Li

Sydney, Australia, 2021.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In psychology, attention (Wikipedia 2020) is the cognitive process of concentrating on a particular aspect of information while ignoring other perceivable elements. Specifically, human visual attention allows us to focus on the intentioned regions with "high-resolution." Regarding the person's age in Figure 1.1, we tend to focus on their faces. If we are interested in what they are doing, the outfits and snowy background can provide rich context clues. Our visual system can adjust our attention and make inferences accordingly. A similar psychological process happens in the understanding of natural language. Given the companied description of Figure 1.1 as an example. "A female skier is teaching a child in a blue suit to ski on the snow." Our attention tends to spotlight "teaching ski" in the sentence if we interest in the activity. The phases, e.g., "child in the blue suit" or "female skier, " describing the main characters are unconsidered.

In summary, human visual/linguistic perceptions can eliminate distracting factors and concentrate on the most relevant components with psychological attention's guidance. In representation learning, an operator imitating the psychological attention mechanism in feature aggregation is also in demand. Nowadays, the idea of attention (Long, Shelhamer & Darrell 2015, Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin 2017, Wang, Girshick, Gupta & He 2018) is arguably one of the essential concepts in deep learning.

Before introducing the attention mechanism in deep learning, this chapter will

Figure 1.1 : The illustration of human visual attention. Observers favor focusing on the faces when regarding the age. In contrast, visual attention tends to observe the outfits and snowy background when curious about the activity shown in the picture.

briefly review the limitations of CNN/RNN and concisely explain how the attention mechanism breaks their limitations.

In representation learning, the CNN/RNN have aptitudes for processing structured data. The CNN model first proves its capability in the image classification task, which demands to extract representative features from the pixel grids. The convolution operator only has a fixed kernel size. Therefore, the model needs to stack deep layers to enlarge its receptive field.

The recurrent neural networks are effectual in processing sequential data, e.g., sentences. However, the recurrent nature of RNN dilutes the long-term information as the sequence length grows. In the seq2seq (Sutskever, Vinyals & Le 2014) model, the source sequence is compressed into a fixed-length context vector and passed to the decoder for target sequence generation in the translation process. With the

target sequence's growth, the influence of source information fades away gradually.

In order to solve the problems above, the attention mechanism is introduced to representation learning. The attention mechanism treats candidate elements as a set without considering their order or position. Therefore, the attention operator can handle unstructured data efficiently. The attention-based models can concentrate on the relevant elements flexibly and free from the bondage of grids or sequences.

The attention mechanism allows the seq2seq model to freely recapitulate the input sequence instead of compressing all information into a fixed-length vector. In the model equipped with attention, the hidden states generated in the encoding stage can be viewed as additional memory to evade the catastrophic forgetting in the sequence generation. Since the internal memory information is orderless, the permutation-invariant property of attention is essential to randomly access the relevant memory information.

For the long-range dependency problem in CNN models, the attention mechanism (Wang et al. 2018, Huang, Wang, Huang, Huang, Wei & Liu 2019) can enlarge the receptive fields to the whole input features while keeping the computational complexity offerable. The CNN model, augmented by the attention mechanism, achieves remarkable success in computer vision tasks. Meanwhile, the sequence models built total on the attention mechanism, *e.g.*, Transformer (Vaswani et al. 2017), Bert (Devlin, Chang, Lee & Toutanova 2018) et al., also have developed as the fundamental framework in natural language processing. The attention mechanism has shown numerous potential in the two primary tasks of artificial intelligence.

As an essential component of deep learning techniques, the attention mechanism has excellent potentials in analyzing vision and language. This dissertation mainly investigates typical vision tasks, e.g., semantic segmentation, face parsing, and language task, e.g., neural machine translation, and multimodal task, e.g., image

captioning. In chapter 2, a multimodal attention mechanism is devised to extend the transformer leveraging the visual and semantic information simultaneously in image captioning. In Chapter3, I investigate the positional awareness in the self-attention mechanism to enhance its expression ability. This algorithm is verified in semantic segmentation and neural machine translation. Furthermore, a local operator built total on attention mechanism is devised to tackle the multi-level feature fusion in semantic segmentation. The detailed research objectives and contributions are discussed as follows.

1.2 Research Contribution

Entangled Transformer. The typical attention mechanisms are arduous to identify the equivalent visual signals in the image captioning task, especially when predicting highly abstract words. This phenomenon is known as the semantic gap between vision and language. This problem can be overcome by providing semantic attributes that are akin to language. However, when designing elaborate attention mechanisms to integrate visual inputs and semantic attributes, RNN-like variants become unflexible due to their complexities. In Chapter II, we investigate a Transformer-based sequence modeling framework, built only with attention layers and feedforward layers. To bridge the semantic gap, we introduce EnTangled Attention (ETA) that enables the Transformer to exploit semantic and visual information simultaneously. Furthermore, Gated Bilateral Controller (GBC) is proposed to guide the interactions between the multimodal information. We name our model as ETA-Transformer. Remarkably, ETA-Transformer achieves state-of-the-art performance on the benchmark image captioning dataset. The ablation studies validate the improvements of our proposed modules.

Bilateral Attention. Recently self-attention has been extensively explored in various language and vision tasks. Given that self-attention operation is solely content-

based and orderless, there is a surge of methods exploring how to expose position information in the attention model. Existing solutions that extend self-attention with positional awareness usually add absolute/relative position encoding to inputs/hidden embedding. However, the position encoding operates on the same projected query shared by self-attention. In this way, handling of position and content information, which are inherently heterogeneous, is compounded, limiting the effectiveness of incorporating position information for feature aggregation. Motivated by the bilateral filter that combines separate filters to consider photometric similarity and position closeness, in this paper, we disentangle the handling of position from content and separately learn position attention to enforce the positional awareness. Specifically, we propose non-localized position attention with the dynamic convolution as a key ingredient, which will generate position attention aligned to yet independent from the content-based attention. The proposed non-localized attention and the content-based attention are further combined in the bilateral formulation to generate hybrid attention weights used for feature aggregation, which provides a more principled way to enforce consistency of two heterogeneous information instead of heuristic designs. The effectiveness of the proposed method is verified on two representative tasks, i.e., semantic segmentation and machine translation.

Localized Attention for Semantic Segmentation. Convolutions are the paradigms in modern deep neural networks. However, convolving across different positions of an image with fixed kernels makes convolutions content-agnostic and inefficient at modelling dynamic spatial layouts. In image segmentation, the problem is especially prominent since the high-quality reconstruction from low-resolution feature maps demand plenty of adaptiveness. Hence, we investigate the potential of attention mechanisms, which can derive dynamic kernels for spatial aggregation. To equip the permutation-invariant attention operation with position-awareness, we propose the local bilateral attention, which can explore the appearance and

geometry information simultaneously. Compared with standard 3×3 convolution, the BA with the same receptive filed has at least 50% fewer parameters and 40% fewer FLOPS. To estimate more accurate attention affinities, we propose an iterative refinement algorithm, which results in additional parameter reduction. We verify the effectiveness of the proposed operations on DeepLabv3+ and U-Net. Experimental results on two public segmentation datasets show that bilateral attention outperforms the standard convolution in both accuracy and robustness.

1.3 Thesis Organization

As discussed in above, this thesis covers three important aspects of the attention mechanism. In each of the chapter, we will expand the research work of the corresponding topic in the structures as: introduction, related work, proposed method, experiment and conclusion. The topics are organised as follows:

(1) *Chapter 2.* This chapter presents an entangled attention mechanism to enable the transformer framework to leverage multimodal information. The effectiveness of the proposed method is valid on the the image captioning task. This chapter is based on the work (Li, Zhu, Liu & Yang 2019) presented in the ICCV 2019 proceedings.

(2) *Chapter 3.* This chapter elaborates on how to enable the self-attention mechanism with positional awareness. The proposed method demonstrates consistent improvement over three variants for representative tasks of semantic segmentation and machine translation.

(3) *Chapter 4.* This chapter explores the possibility of employing a local attention operator as a replacement for the convolution in semantic segmentation, and the designed operator demonstrates its inclination in the fusion of multi-level features.

(4) *Chapter 5.* A brief summary of the thesis contents are given in the final chapter. Recommendation for future works is given as well.

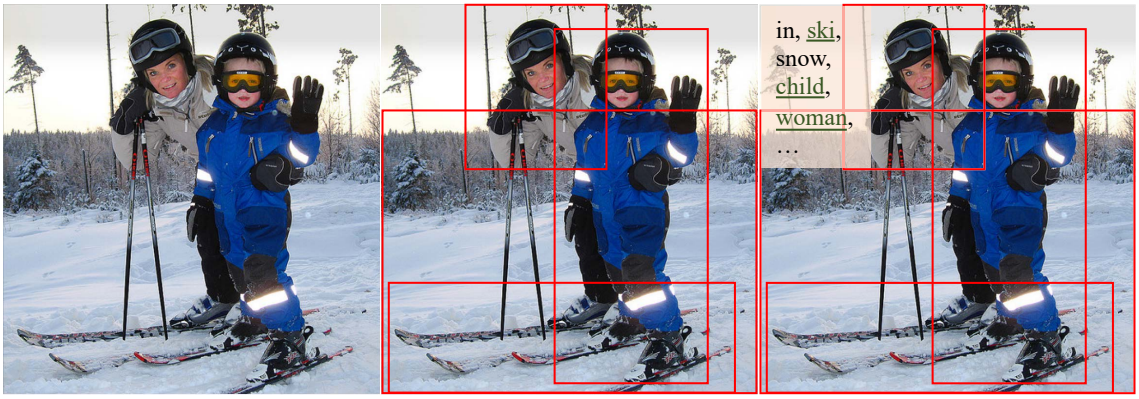
Chapter 2

Entangled Transformer for Image Captioning

2.1 Introduction

Image captioning (Vinyals, Toshev, Bengio & Erhan 2015, Karpathy & Fei-Fei 2015) is one of the essential tasks (Antol, Agrawal, Lu, Mitchell, Batra, Lawrence Zitnick & Parikh 2015, Vinyals et al. 2015, Zhu, Xu, Yang & Hauptmann n.d.) that attempts to break the semantic gap between vision and language. To generate good captions for images, it involves not only the understanding of many concepts, such as objects, actions, scenes, human-objects interactions but also expressing these factors and their relations in a natural language. Recently, the attention mechanism (Xu, Ba, Kiros, Cho, Courville, Salakhudinov, Zemel & Bengio 2015, You, Jin, Wang, Fang & Luo 2016, Fang, Gupta, Iandola, Srivastava, Deng, Dollár, Gao, He, Mitchell, Platt et al. 2015) was introduced to dynamically recap the salient information of the input image for every word.

In previous image captioning works (Xu et al. 2015, You et al. 2016, Fang et al. 2015), the attention mechanism mainly lies in two fields based on the modality of the information they employed: *Visual Attention* and *Semantic Attention*. On the one hand, visual attention exploits the low-level feature maps (Xu et al. 2015) or high-level object ROI-pooled features (Pedersoli, Lucas, Schmid & Verbeek 2017, Anderson, He, Buehler, Teney, Johnson, Gould & Zhang 2018) to identify the most relevant regions for the words. However, due to the semantic gap, not every word in the caption has corresponding visual signals (Lu, Xiong, Parikh & Socher 2018), especially for the tokens associated with abstract concepts and complex



(a). A person is standing in the snow.

(b). Two children are standing in the snow.

(c). A woman and a child are skiing in the snow.

Figure 2.1 : The image captioning results when given different modality information. (a) provides an unsatisfactory caption result only using low-level visual features. When provided with high-level visual information guided from region proposals, (b) can make some improvement, e.g., predict “two children" in the picture. However, it still fails to grab abstract concepts in the image, e.g., “skiing". (c) is the result when utilizing information from complementary modalities: visual and semantic. It is the most accurate result among the three descriptions.

relationships. Figure 2.1 shows an example of this obstacle. On the other hand, researchers develop the semantic attention (You et al. 2016, Fang et al. 2015) which can leverage the high-level semantic information directly. Nevertheless, because of the recurrent nature, RNNs (Elman 1990, Mikolov, Karafiát, Burget, Černockỳ & Khudanpur 2010, Sutskever, Martens & Hinton 2011) have difficulties in memorizing the inputs many steps ago, especially the initial visual input. Consequently, such approaches tend to collapse into high-frequency phrase fragments without regard to the visual cues.

As shown in Figure 2.1(c), the combination of the two complementary attention paradigms can alleviate the harmful impacts of the semantic gap. Therefore, Li *et*

al. (Li & Chen 2018) propose a two-layered LSTM (Hochreiter & Schmidhuber 1997) that the visual and semantic attentions are separately conducted at each layer. Yao *et al.* (Yao, Pan, Li & Mei 2018) employ graph convolutional neural networks to explore the spatial and semantic relationships. They use late fusion to combine two LSTM language models that are independently trained on different modalities. However, due to the inherent recurrent nature and the complex operating mechanism, RNNs fail to explore the two complementary modalities concurrently.

To solve these problems above, we extend the efficient and straightforward Transformer (Vaswani et al. 2017) framework with our proposed Entangled Attention (ETA) and Gated Bilateral Controller (GBC) to explore visual and semantic information simultaneously. The design of ETA is inspired by the studies (Cooper 1974, Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy 1995) about the human visual system, showing the selection of attentive regions in human visual attention can be influenced by a prior linguistic input. To mimic this phenomenon, we use an information injection operation to infuse the input query with the information from the preliminary modality. Then the attention over the target modality can be conducted under the guidance of the preliminary modality. Subsequently, the representations of the target visual and semantic modalities propagate to the next layers under the channel-wise control of GBC.

The advantages of our method are as follows. First, the simplicity of the Transformer (Vaswani et al. 2017) framework relieves us from the limitations of recurrent neural networks. Second, the application of self-attention in the encoder encourages our model to explore the relationships between the detected entities. Our method can efficiently leverage the information in the target modality under the guidance of preliminary modality. Third, the proposed bilateral gating, GBC, can jointly facilitate our module to provide sophisticated control for the propagation of multimodal information. Because of the cohesiveness, our attention module can be readily

applied to the Transformer without violating its parallel nature and modularity.

Our contributions can be summarized as follows:

(1) We devise the EnTangled Attention – a unique attention mechanism which enables the Transformer framework to exploit the visual and semantic information simultaneously.

(2) We propose the Gated Bilateral Controller – a novel bilateral gating mechanism which can provide sophisticated control for the forward propagation of multimodal information as well as their backpropagating gradients.

(3) We comprehensively evaluate our approach on the MSCOCO dataset (Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár & Zitnick 2014), and our method achieves the state-of-the-art performance.

2.2 Related Work

Attention in Visual Captioning. Despite the efforts (Xu et al. 2015, Pedersoli et al. 2017, You et al. 2016, Fang et al. 2015, Lu et al. 2018, Anderson et al. 2018, Wu, Zhu, Jiang & Yang 2018) investigate the attention over monomodal information, many works also try to combine visual and semantic information semoutanouly. Yao *et al.* (Yao, Pan, Li, Qiu & Mei 2017) prove multimodal information can contribute to the image captioning problem and investigate how to employ semantic attributes under the LSTM framework. Li *et al.* (Li & Chen 2018) propose a two-layer visual-semantic LSTM which conducts visual attention and semantic attention at different layers. To explore the relationship between objects and semantic attributes, Yao *et al.* (Yao et al. 2018) apply a graph convolution neural networks in the encoding stage. Tang *et al.* (Tang, Zhang, Wu, Luo & Liu 2019) leverage scene graph to align the relations between vision and language. Conducted only in each modality separately, these methods fail to explore the complementary nature of the visual and semantic

information.

Co-attention in VQA. The widely used co-attention mechanism (Lu, Yang, Batra & Parikh 2016, Fukui, Park, Yang, Rohrbach, Darrell & Rohrbach 2016, Lee, Chen, Hua, Hu & He 2018) in visual question answering (VQA) can explore the visual and semantic information jointly. But the major concern of VQA is to identify the most relevant visual regions based on the question. Hence, the attention mechanism in VQA mainly queries the visual regions with the semantic feature. However, in image captioning, the most salient semantic attributes should also be identified.

Model Structures. The recurrent nature of RNN dilutes the long-term information at every time step (Sukhbaatar, Weston, Fergus et al. 2015). To get rid of the catastrophic forgetting in long-term memory, Gu *et al.* (Gu, Wang, Cai & Chen 2017) introduce temporal CNN to impose the experienced semantic information at every step of the generation procedure. Additionally, to overcome the inherently recurrent nature of the RNNs, Gehring *et al.* (Gehring, Auli, Grangier, Yarats & Dauphin 2017) propose to use Convolutional Neural Networks (CNN) to model the sequence-to-sequence problem. Afterward, Aneja *et al.* (Aneja, Deshpande & Schwing 2018) adapt this model to image captioning. Different from the local convolution operation, whose receptive field is determined by the kernel size and layer depth, the self-attention can access the information globally. Besides, there are only a few attempts (Chen, Li, Zhang & Huang 2018, Zhou, Zhou, Corso, Socher & Xiong 2018, Sharma, Ding, Goodman & Soricut 2018) to employ the Transformer in visual captioning.

2.3 Preliminary

To overcome the inherent recurrence in RNN model, the Transformer reformulate the calculation of the hidden state in Eq. 2.1. Thus, the hidden state of current time step $\mathbf{h}_t$ only depends on the feature embeddings of the input image and history words, rather than the previous hidden state $\mathbf{h}_{t-1}$. This formulation enables the

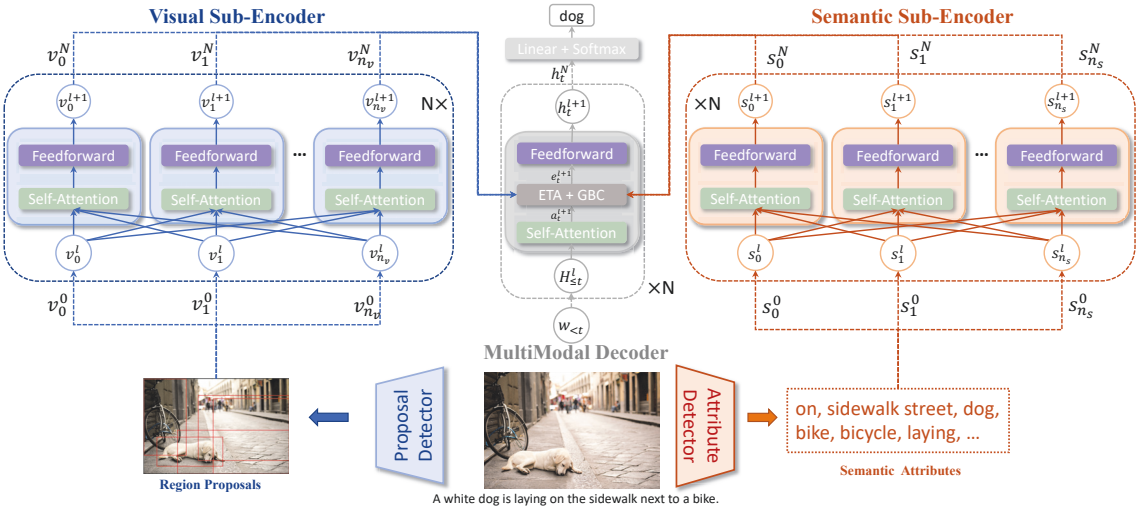


Figure 2.2 : The overall architecture of ETA-Transformer. Our model consists of three components: the visual sub-encoder, the semantic sub-encoder, and the multimodal decoder. The generation procedure has three steps: (1) detecting region proposals and semantic attributes; (2) encoding the visual and semantic features separately; (3) decoding word by word to obtain the final caption. Notice that the Residual Connections, Layer Normalizations, and Embedding Layers are omitted.

Transformer model to execute in parallel.

$$\mathbf{h}_t = \mathbf{TransformerDecoder}(I; \mathbf{w}_1, \ldots, \mathbf{w}_{t-1}) \tag{2.1}$$

To handle the variable-length inputs, such as image regions and word sequence, Transformer employs attention to convert the unfixed number of inputs to a unified representation. Moreover, positional encoding (Vaswani et al. 2017) is employed both in the encoder and decoder to inject sequential information.

There are two particular attention mechanisms in the Transformer model. Here we start with the *scaled dot-product attention* (Vaswani et al. 2017), in which the inner product is applied to calculate the attention weights. Given a query $\mathbf{q}_i$ from all m queries, a set of keys $\mathbf{k}_t \in \mathbb{R}^d$ and values $\mathbf{v}_t \in \mathbb{R}^d$ where $t = 1, \ldots, n$, the scaled dot-product attention outputs a weighted sum of values $\mathbf{v}_t$, where the weights are

determined by the dot-products of query $\mathbf{q}_i$ and keys $\mathbf{k}_t$. In order to implement the dot product operation by highly optimized matrix multiplication code, the queries, keys, and values are packed together into matrices $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_m)$, $\mathbf{K} = (\mathbf{k}_1, \ldots, \mathbf{k}_n)$, and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$. In practice,

$$\mathbf{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \tag{2.2}$$

where d is the width of the input feature vectors.

To extend the capacity of exploring subspaces, Transformer employs the *multi-head attention* (Vaswani et al. 2017) which consists of h parallel scaled dot-product attentions named *head*. The inputs including queries, keys, and values are projected into h subspaces, and the attention performs in the subspaces seperately:

$$\begin{aligned} \mathbf{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \mathbf{Concat}(\mathbf{H_1}, \ldots, \mathbf{H_h})\mathbf{W}^O, \\ \mathbf{H}_i &= \mathbf{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \end{aligned} \tag{2.3}$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{\frac{d}{h} \times d}$ are the independent head projection matrices, $i = 1, 2, \ldots, h$ and $\mathbf{W}_i^O \in \mathbb{R}^{d \times d}$ denotes the linear transformation. Note that the bias terms in linear layers are omitted for the sake of concise expression, and the subsequent descriptions follow the same principle.

2.4 Methodology

In this section, we devise our ETA-Transformer model. As shown in Figure 2.2, the overall architecture follows the encoder-decoder paradigm. First, a dual-way encoder maps the original inputs into highly abstract representations and then the decoder incorporates the multimodal information simultaneously to generate the caption word by word.

2.4.1 Dual-Way Encoder

In most cases, CNNs like VGG (Simonyan & Zisserman 2015) or ResNet (He, Zhang, Ren & Sun 2016) are first considered for encoding the visual information,

while the transformer encoder is originally designed for sequence modeling. However, we argue that a transformer encoder with sophisticated design can better explore the inter- and intra- relationships between the visual entities and semantic attributes. Specifically, we devise a dual-way encoder that consists of two sub-encoders. Each sub-encoder is self-attentive and of the same structure, i.e., a stack of N identical blocks.

Take the output of the l-th ($0 \leq\ l < N$) block $\mathbf{O}^l \in R^{d \times n}$ as an example. They are first fed into the multi-head self-attention module in the $(l+1)$-th block:

$$\mathbf{M}^{l+1} = \mathbf{MultiHead}(\mathbf{O}^l, \mathbf{O}^l, \mathbf{O}^l), \tag{2.4}$$

where $\mathbf{M}^{l+1}$ is the hidden state calculated by multi-head attention. The query, key and value matrices have the same shape. Notice that the $\mathbf{O}^0$ is the output of the embedding layer.

The subsequent sub-layer is a position-wise feed-forward network (FFN) which consists of two linear transformations with a ReLU activation in between:

$$\begin{aligned} \mathbf{FFN}(\mathbf{x}) &= \mathbf{W}_2 \cdot \mathbf{ReLU}(\mathbf{W}_1 \cdot \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2, \\ O^{l+1} &= [\mathbf{FFN}(\mathbf{M}^{l+1}_{\cdot,1}); \ldots; \mathbf{FFN}(\mathbf{M}^{l+1}_{\cdot,n})], \end{aligned} \tag{2.5}$$

where $\mathbf{W}_2 \in\ \mathbb{R}^{d \times d_m}$, $\mathbf{W}_1 \in \mathbb{R}^{d_m \times d}$, $\mathbf{O}^{l+1} \in R^{d \times n}$ are the outputs of the $(l+1)$-th block, and $\mathbf{M}^{l+1}_{\cdot,i}$ represents column i of matrix $\mathbf{M}$, thus the i-th feature vector. The two equivalent expressions are used interchangeably in the subsequent description. Same to (Vaswani et al. 2017), the residual connection and layer normalization are used after the forementioned sub-layers, and we omit them for a concise explanation.

The structure described above can be used for encoding both the visual and semantic features. Before feeding into the sub-encoder, the n_v visual features are mapped into $\mathbf{V}^0 \in \mathbb{R}^{d \times n_v}$ by a linear transformation, and the n_s one-hot semantic attributes are projected into $\mathbf{S}^0 \in \mathbb{R}^{d \times n_s}$ by an embedding layer. Furthermore, we

share the word embeddings between the semantic encoder and the decoder so that our model can utilize the target information directly.

2.4.2 Multimodal Decoder

In addition to the basic block of the encoder, the decoder block inserts an ETA module and a GBC module between the self-attention sub-layer and the feed-forward sub-layer, which empowers the decoder block to perform attention over the visual outputs $\mathbf{V}^N$ and semantic outputs $\mathbf{S}^N$ of the dual-way encoder simultaneously. Similar to the encoder, the decoder consists of N identical blocks, and we employ residual connections around each of the sub-layers, followed by layer normalization.

Suppose the decoder is generating the t-th word in the target sentence. We denote $\mathbf{w}_t \in \mathbb{R}^{d\times 1}$ as the vector representation of the t-th word, which is the sum of word embedding and positional encoding. Therefore, the input matrix representation for time step t is:

$$\mathbf{W}_{<t} = [\mathbf{w}_0; \ldots; \mathbf{w}_{t-1}], \tag{2.6}$$

where $\mathbf{W}_{<t} \in \mathbb{R}^{d\times t}$ and $\mathbf{w}_0$ is the feature vector of the token representing the start of sentence.

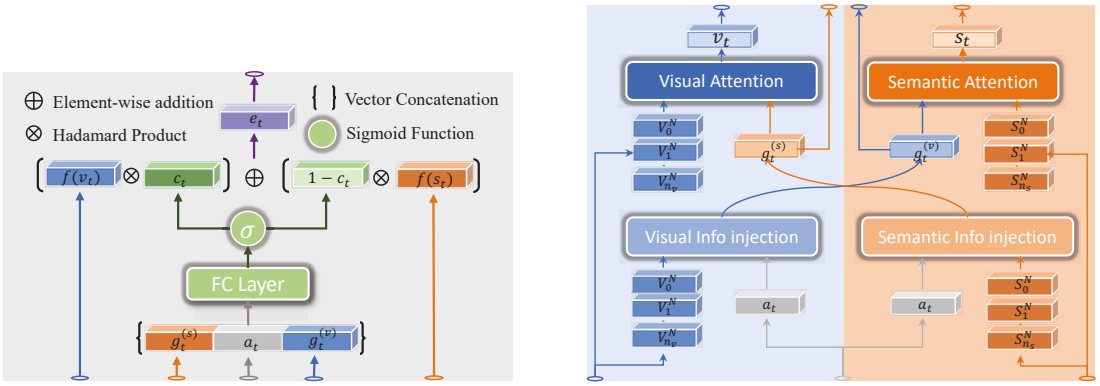
For the $(l+1)$-th block, the inputs $\mathbf{H}^l_{\leq\ t} \in \mathbb{R}^{d\times t} = (\mathbf{h}^l_1, \ldots, \mathbf{h}^l_t)$ are fed into a multi-head self-attention sub-layer, notice that $\mathbf{h}^0_t$ corresponds to $\mathbf{w}_{t-1}$:

$$\mathbf{A}^{l+1}_{\cdot,t} = \mathbf{MultiHead}(\mathbf{H}^l_{\cdot,t}, \mathbf{H}^l_{<t}, \mathbf{H}^l_{<t}), \tag{2.7}$$

where $\mathbf{H}^l_{\cdot,t} \in \mathbb{R}^{d\times 1}$, $\mathbf{A}^l_{\cdot,t} \in \mathbb{R}^{d\times 1}$, and $\mathbf{h}^0_t = \mathbf{w}_{t-1}$. Notice that $\mathbf{W}_{<t}$ is the inputs for the first layer. Subsequently, the self-attention output $\mathbf{a}^{l+1}_t$ is passed into the ETA to incorporate with visual and semantic features:

$$\mathbf{E}^{l+1}_{\cdot,t} = \mathbf{ETA}(\mathbf{A}^{l+1}_{\cdot,t}, \mathbf{V}^N, \mathbf{S}^N), \tag{2.8}$$

where $\mathbf{E}^{l+1}_{\cdot,t} \in \mathbb{R}^{d\times 1}$ contains the visual and semantic information which is elaborately



(a) Gated Bilateral Controller (b) EnTangled Attention

Figure 2.3 : The multimodal representations are first fed into ETA to conduct EnTangled Attention, then to GBC to obtain the final representation.

integrated according to the importance of modalities in channel level. After the process of **FFN**, we obtain the output $\mathbf{h}_t^{l+1} = \mathbf{FFN}(\mathbf{e}_t^{l+1})$ of current layer.

Finally, the output of layer N is fed into the classifier over vocabulary to predict next word. Notice that the procedure described above illustrates the incremental generation in inference. Because all the input tokens are known in the training stage, the attention is implemented with highly optimized matrix multiplication.

2.4.3 EnTangled Attention

Most of the previous attempts trying to integrate multimodal information for image captioning only perform attention over the multiple modalities separately and then fuse the independent attention representations. Therefore, they fail to leverage the complementary nature of visual and semantic information in attention operations. Differently, as shown in Figure 2.3 (b), we implement the attention in an entangled manner so that it can be affected by the preliminary modality while performing attention over the target one.

Here we take the visual pathway in ETA as an illustration. To mimic the

attention mechanism of the human vision system, we need a function which can inject the information of preliminary modality $\mathbf{S}^N$ into the self-attention output $\mathbf{a}_t$ (see Eq. 2.7) so that the generated representation $\mathbf{g}_t^{(s)} \in \mathbb{R}^{d\times 1}$ (the superscript (s) is donated for the sign of modality) can provide proper guidance for the attention in target modality. In order to handle the variable number of semantic attributes, we choose multi-head attention as the preliminary information injection function:

$$\mathbf{g}_t^{(s)} = \mathbf{MultiHead}(\mathbf{a}_t, \mathbf{S}^N, \mathbf{S}^N). \tag{2.9}$$

Next, we use the semantic guidance $\mathbf{g}_t^s$ to perform multi-head attention over the target modality $\mathbf{V}^N$:

$$\mathbf{v}_t = \mathbf{MultiHead}(\mathbf{g}_t^{(s)}, \mathbf{V}^N, \mathbf{V}^N), \tag{2.10}$$

where $\mathbf{v}_t \in \mathbb{R}^{d\times 1}$ is the final representation generated with the guidance of semantic modality. And in a similar manner but reversed order, we could obtain the semantic representation $\mathbf{s}_t \in \mathbb{R}^{d\times 1}$. Notice that all the attention layers in ETA are followed with residual connection and layer normalization which are omitted for concise expression.

2.4.4 Gated Bilateral Controller

In this section, we present the *Gated Bilateral Controller* (GBC) specially designed for the integration of the generated representations $\mathbf{s}_t$ and $\mathbf{v}_t$. The gating mechanisms controlling the path through which information flows to the subsequent layers are widely used in the famous sequence models like LSTM (Hochreiter & Schmidhuber 1997), GRU (Cho, Gulcehre, Bahdanau, Schwenk & Bengio 2014), and ConvS2S (Gehring et al. 2017). Such multiplicative gates are adept at dealing with gradient explosion and vanishing, which enable the information to propogate unimpededly through long timesteps or deep layers. As illustrated in Figure 2.3 (a), the context gate $\mathbf{c}_t$ in GBC is determined by the current self-attention output $\mathbf{a}_t$,

the visual guidance $\mathbf{g}^{(v)}$ and the semantic guidance $\mathbf{g}^{(s)}$:

$$\mathbf{c}_t = \sigma\left(\mathbf{W}_c \cdot [\mathbf{g}_t^{(s)}, \mathbf{g}_t^{(v)}, \mathbf{a}_t]\right), \tag{2.11}$$

where $\mathbf{c}_t \in \mathbb{R}^{d\times 1}$, $\mathbf{W}_c \in \mathbb{R}^{d\times 3d}$ and $\sigma(\cdot)$ denotes the sigmoid function.

Different from the previous gating mechanism managing only one pathway, we extend it with a bilateral scheme. The gate value $\mathbf{c}_t$ controls the flow of visual guidance $\mathbf{v}_t$ while the complement part $(1-\mathbf{c}_t)$ governs the propagation of semantic information $\mathbf{s}_t$:

$$\mathbf{e}_t = f(\mathbf{v}_t) \odot \mathbf{c}_t + f(\mathbf{s}_t) \odot (1-\mathbf{c}_t), \tag{2.12}$$

where $\odot$ represents the hadamard product, $f(\cdot)$ can be an activation function or identity function, and $\mathbf{e}_t \in \mathbb{R}^{d\times 1}$ denotes the output of ETA.

The Effect of f Function. In LSTM or GRU, the left part of the Hadamard product is always activated with function f which can be Sigmoid, Tanh or ReLU (Le, Jaitly & Hinton 2015), *etc.* Whereas, we do not apply any activation over v_t and s_t which are merely the outputs of the linear transformation in multi-head attention. Compared with the saturate activations mentioned above, the identity function $id(x) = x$ allows gradients to propagate through the linear part without downscaling. Here, following the analysis in (Dauphin, Fan, Auli & Grangier 2017), we take the left part of the Eq. 2.12 as an example, whose gradient is:

$$\nabla\left[f(\mathbf{x}) \odot \ \mathbf{c}_t\right] = f'(\mathbf{x})\nabla\ \mathbf{x} \odot \ \mathbf{c}_t. \tag{2.13}$$

As shown in the Eq. 2.13, the $f'(x)$ can act as a scale factor of the gradients. Additionally, $tanh'(\cdot) \in (0, 1]$, $\sigma'(\cdot) \in (0, 0.25]$, while $id'(\cdot) = 1$. Thus, the saturate activations will downscale the gradient and make gradient vanishing even worse with the stacking of layers. Although the non-saturate activation ReLU has similar property with identity function, here we argue the activated gate $\mathbf{c}_t$ has equipped the module with non-linearity (Dauphin & Grangier 2016). For the principle of

simplicity, we do not apply any activations over $\mathbf{v}_t$ and $\mathbf{s}_t$. By comparing the effect of f function experimentally in Section 2.5.4, we find the activations deteriorate the performance greatly while the identity function achieves the best.

2.5 Experiments

	Proposal	Semantic	Cross-Entropy Loss						Sequence-Level Optimization					
			B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
SCST (Rennie, Marcheret, Mroueh, Ross & Goel n.d.)	✗	✗	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
LSTM-A (Yao et al. 2017)	✗	✓	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
VS-LSTM (Li & Chen 2018)	✓	✓	76.3	34.3	26.9	-	110.2	-	78.9	36.3	27.3	-	120.8	-
Up-Down (Anderson et al. 2018)	✓	✗	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM$_{fuse}$ (Yao et al. 2018)	✓	✓	77.4	37.1	28.1	57.2	117.1	21.1	80.9	38.3	28.6	58.5	**128.7**	22.1
ETA	✓	✓	77.3	37.1	28.2	57.1	117.9	21.4	81.5	39.3	28.8	58.9	126.6	**22.7**
ETA$_{fuse}$	✓	✓	**77.6**	**37.8**	**28.4**	**57.4**	**119.3**	**21.6**	**81.5**	**39.9**	**28.9**	**59.0**	127.6	22.6

Table 2.1 : MSCOCO Offline Evaluation. The ETA denotes the ETA-Transformer. ✓ indicates the corresponding features (region proposals or semantic attributes) are applied, and ✗ means otherwise. All values are reported as percentage (%).

2.5.1 Datasets and Evaluation

We use the MSCOCO 2014 captions dataset (Lin et al. 2014) to evaluate our proposed captioning model. In offline testing, we use the 'Karpathy' splits (Karpathy & Fei-Fei 2015) that have been used extensively for reporting results in previous works. This split contains 113,287 training images with five captions each, and 5K images respectively for validation and testing. Our MSCOCO test server submission is trained on the Karpathy's training split, and chosen on the Karpathy's test split.

Data processing We follow standard practice and perform only minimal text pre-processing, converting all sentences to lower case, tokenizing on white space, and keeping words that occur at least five times, resulting in a model vocabulary of 9,487 words. To evaluate caption quality, we use the standard automatic evaluation metrics,

	B@1	B@4	M	R	C	S
VS-LSTM$_s$	74.3	33.3	26.5	-	105.1	-
VS-LSTM$_v$	75.1	33.5	26.5	-	105.8	-
VS-LSTM	76.3	34.3	26.9	-	110.2	-
GCN-LSTM$_s$	77.3	36.8	27.9	57.0	116.3	20.9
GCN-LSTM$_v$	77.2	36.5	27.8	56.8	115.6	20.8
GCN-LSTM$_{fuse}$	**77.4**	**37.1**	28.1	**57.2**	117.1	21.1
Transformer$_s$	71.1	29.0	25.3	52.8	96.2	18.2
Transformer$_v$	75.9	34.0	27.5	56.1	112.2	21.0
ETA	77.3	**37.1**	**28.2**	57.1	**117.9**	**21.4**
ETA^{oracle}	97.0	76.7	47.9	84.2	204.2	34.7

Table 2.2 : The results on single modality. The ETA denotes the ETA-Transformer. Subscript indicates that the visual modality or semantic modality is applied.

namely SPICE (Anderson, Fernando, Johnson & Gould 2016), CIDEr-D (Vedantam, Lawrence Zitnick & Parikh 2015), METEOR (Denkowski & Lavie 2014), ROUGE-L (Lin 2004) and BLEU (Papineni, Roukos, Ward & Zhu 2002).

2.5.2 Implementation Details

Visual & Semantic Features. For visual features, we use the region proposals as the visual representations. To select the salient regions, we follow the settings in Up-Down (Anderson et al. 2018). When comparing with some previous methods (Karpathy & Fei-Fei 2015, Aneja et al. 2018), we also encode the full-sized input image with the final convolutional layer of VGG-16 (Simonyan & Zisserman 2015) and use adaptive pooling to resize the outputs into a fixed size of 7x7. For semantic features, we follows the settings of Fang *et. al* (Fang et al. 2015) to detect semantic attributes. The backbone of attribute detector is fine-tuned from VGG16 equipped with a noisy-OR version of multiple instance loss. We only keep the top-1000 frequent

words as labels. And in the training stage, we use the detected semantic attributes rather than the ground truth.

Model Settings & Training. We follow the same hyper-parameter settings in (Vaswani et al. 2017). We use $N = 6$ identical layers in both encoder and decoder. The output dimension of the word embedding layers is 512, and the input visual features are also mapped into 512 with a linear projection. The inner-layer of the feed-forward network has dimentionality $d_m = 2048$. And $h = 8$ parallel attention layers are employed in multi-head attention. Besides, we also share the word embedding between semantic sub-encoder and the decoder in order to leverage the target word representation directly. In training stage, we use the same learning rate schedule as (Vaswani et al. 2017). The input batch size is 75 image-sentence pairs and the warm-up step is 20000. We use the Adam optimizer (Karpathy & Fei-Fei 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$.

Model	B@1		B@2		B@3		B@4		M		R-L		C-D	
-	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.3	56.3	70.7	114.7	116.0
LSTM-A	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
VS-LSTM	78.8	94.6	62.8	87.5	47.9	77.3	35.9	66.3	27.0	35.3	56.5	70.3	116.6	119.5
Up-Down	80.2	**95.2**	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
GCN-LSTM	-	-	**65.5**	**89.3**	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	**125.3**	**126.5**
ETA	**81.2**	95.0	**65.5**	89.0	**50.9**	**80.4**	**38.9**	**70.2**	**28.6**	**38.0**	**58.6**	**73.9**	122.1	124.4

Table 2.3 : MSCOCO Online Evaluation. The ETA denotes the ETA-Transformer. cX means evaluation on X captions. All values are reported as percentage (%).

2.5.3 Comparision with State-of-the-Art Methods

Offline Evaluation. Table 2.1 shows the performance of our model and state-of-the-art approaches in recent two years. Note that the comparative methods are all based on LSTM and its variants, which is the dominant framework in

	B@1	B@4	M	R	C	S
LSTM (Karpathy & Fei-Fei 2015)	71.3	30.3	24.7	52.5	91.2	17.2
Convolution (Aneja et al. 2018)	71.1	28.7	24.4	52.2	91.2	17.5
Transformer$_s$	71.1	29.0	25.3	52.8	96.2	18.2
- Encoder	70.3	28.5	24.8	52.0	93.1	17.7
Transformer$_v$	71.0	30.2	24.9	52.6	93.8	18.0
- Encoder	70.2	28.2	24.2	51.6	91.8	17.2
ETA	**72.2**	**31.9**	**25.7**	**53.4**	**99.2**	**18.6**

Table 2.4 : Comparison with different model structures. And “-Encoder" implies the Encoder is removed from the model. All results are reported in token-level training.

image captioning. All the baselines adapt ResNet-101 as the backbone network of visual representation. The self-critical sequence-level training strategy devised in SCST (Rennie et al. n.d.) is applied by Up-Down (Anderson et al. 2018), GCN-LSTM (Yao et al. 2018) and ETA-Transformer for optimizing the CIDEr-D score, while VS-LSTM (Li & Chen 2018) employs an improved version of SCST. LSTM-A (Yao et al. 2017) investigates how to utilize the predicted semantic attributes efficiently. We use them as the LSTM baselines. Up-Down (Anderson et al. 2018) presents a two-layer LSTM to conduct attention over bottom-up and top-down visual features separately. VS-LSTM (Li & Chen 2018) use a similar design but replace the low-level visual features with semantic attributes. Restricted by the complexity of LSTM, the models have difficulties in stacking deep layers. Benefits from the scalability of the Transformer and the cohesiveness of our proposed modules, the multimodal attention can be conducted at different levels of abstraction. In our experiments, we employ N=6 multimodal attentions in the decoding stage. Thus, our method outperforms them with a large margin. Aiming at modeling the relations of objects, GCN-LSTM (Yao et al. 2018) introduced graph convolutional neural network to encode the detected entities. To make a fair comparison, we also provide

the late-fused performance of two models with different initialization. The result shows that our model achieves superior performance on the cross-entropy training. And in sequence level training, our model produces higher performance in five out of six metrics, especially the BLEU@4(39.9%) and SPICE(22.7%).

To provide a more detailed comparison, we also report the results on single modality. ETA-Transformer and VS-LSTM use weak semantic labels generated from the ground truth captions (see (Fang et al. 2015) for more details), but the GCN-LSTM employs a fully-supervised model trained on the region-level annotations of Visual Genome (Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma et al. 2017). Therefore, the GCN-LSTM$_s$ has superior performance to the Transformer$_s$ and VS-LSTM$_s$. However, as shown in Table 2.2, our model provides the most significant improvements when combining the two modalities. This comparison further proves the effectiveness of our proposed modules in leveraging the complementary information. We also report the performance of our model under an Oracle setting (see the ETA^{oracle}), where the semantic attributes tokenized from the ground truth captions are provided during test time. This can be viewed as the upper bound of our method when we have a perfect attribute detector.

Online Evaluation. We ensembled three models trained on sequence-level criterion with different initalization, and submitted our results to the online testing server. Table 2.3 includes the top-5 methods which have been officially published, and it shows that the ETA-Transformer is among the top-2 performance over all the metrics. In particular, the B@3, B@4, METEOR, and ROUGE-L are superior on both c5 and c40 testing sets. The submission results named *ETA-Transformer* have been public on the leaderboard [*].

[*] https://competitions.codalab.org/competitions/3221

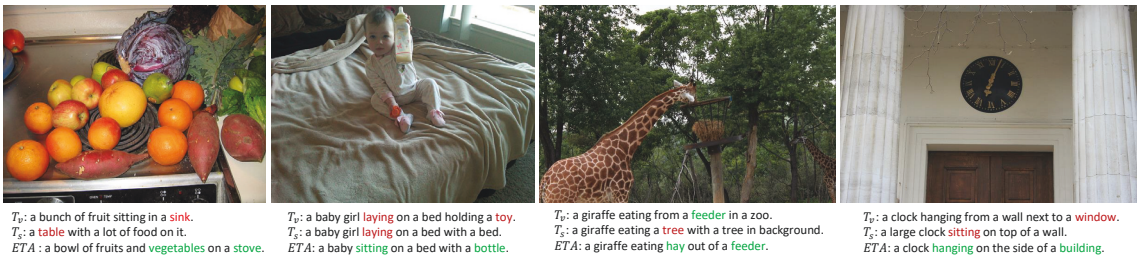


Figure 2.4 : Qualitative examples of different methods. Compared with Transformer$_v$ (T$_v$) and Transformer$_s$ (T$_s$), the ETA-Transformer (ETA) generates more descriptive and more accurate captions.

2.5.4 Ablation Study

Comparison with Different Frameworks

In the ablation study, we first compare the Transformer with the other two classical sequence model LSTM (Xu et al. 2015) and ConvS2S (Aneja et al. 2018, Gehring et al. 2017). The two models are all equipped with visual attention mechanism. Following the feature extraction settings in (Aneja et al. 2018), we use 7x7 feature maps of the fifth convolution layer in VGG-16 as our visual representations. The performance on Table 2.4 shows that the standard Transformer is comparable with LSTM and ConvS2S model in the image captioning problem.

Further, to validate the previous declaration that the self-attention can benefit the feature representation from modeling the relationships of input entities, we report the results of the encoder-removed transformer on both modalities. As shown in Table 2.4, the performance has dropped significantly over all the metrics.

Comparison with Strong Baselines

In this section, we provide other two simplified versions of our proposed modules, and the late-fusion of Transformer$_s$ (T$_s$) & Transformer$_v$ (T$_v$), as strong baselines.

	B@1	B@4	M	R	C	S
Transformer$_v$	80.6	38.3	28.5	58.3	124.1	22.3
Transformer$_s$	76.6	32.6	25.5	54.4	102.8	19.1
T$_v$ & T$_{s\ fuse}$	79.6	37.5	27.6	57.6	118.7	19.8
Parallel	80.9	38.7	28.8	58.7	124.9	22.4
Stacked$_v$	80.7	39.1	28.6	58.6	125.0	22.4
Stacked$_s$	80.8	38.8	28.6	58.5	124.5	22.5
ETA	**81.5**	**39.3**	**28.8**	**58.9**	**126.6**	**22.7**

Table 2.5 : Ablation experiments. ETA is denotes the ETA-Transformer. And all results are trained on sequence-level criterion.

In the first one, we remove the GBC module and extract one pathway of ETA as the first version. We refer this version as Stacked Attention (SA) because it has two stacked multi-head attentions. In the second one, we remove the preliminary information injection blocks in ETA and simply use GBC to integrate the outputs of encoder (S^N and V^N) directly. This version is named as Parallel Attention (PA). In T$_s$ & T$_{v\ fuse}$, we train the two standard transformer model separately and late-fused the results of them.

The late fusion of monomodal models can only have limited gains, sometimes, even severe degeneration. As shown in Table 2.5, the performance of T$_s$ & T$_{v\ fuse}$ is worse even compared with T$_v$. This is mainly caused by the inferior single model T$_s$. Differently, the ETA-Transformer, which integrates the multimodal information at the feature level, obtains significant and stable improvement in performance.

In Table 2.5, compared results of the multimodal versions with Transformer$_s$ or Transformer$_v$, we can find that visual and semantic modalities are complementary. The integration of visual and semantic information can contribute to better performance despite that the semantic representations are considerably worse than

	B@1	B@4	M	R	C	S
Sigmoid	74.5	32.1	26.3	54.8	104.9	19.4
Tanh	74.8	32.0	26.2	54.8	104.1	19.6
ReLU	**76.3**	36.1	27.9	56.2	114.0	20.8
Linear	**76.3**	**36.3**	**28.1**	**56.5**	**115.2**	**21.0**

Table 2.6 : The effect of activation functions in GBC. All results are reported in token-level training.

the visual representations. Notwithstanding the huge performance gap between Transformer$_v$ and Transformer$_s$, SA$_s$ and SA$_v$ (see the Stacked$_s$ and Stacked$_v$ in Table 2.5 have near performance on all the metrics. These experimental results show the EnTangled Attention can benefit from fusing the visual and semantic information with an ordered manner. Besides, the widely used skip connection, which equally combines the preliminary and target representations without any adaptive trade-off, sustains the impact of the preliminary modality. Thus the performance of semantic information is enhanced.

Without using the EnTangled Attention mechanism, the parallel attention only employs the gated bilateral controller to combine the encoded visual and semantic representations adaptively. And PA gains comparable and slightly better performance than SA$_s$ and SA$_v$. Furthermore, The ETA can be viewed as the combination of PA and SA, which incorporates the advantages of both. Shown in Table 2.5, the ETA achieves superior performances against the two strong baselines in all the metrics noteworthily.

The Effect of Activation in GBC

As shown in Table 2.6, the saturated activation functions like Sigmoid and Tanh deteriorate the performance of GBC significantly, while the identity function and the

non-saturated activation ReLU do not suffer from this degeneration. The identity function only outperforms ReLU slightly. Following the analysis in 2.4.4, bacause $tanh'(\cdot) \in (0, 1]$ has a larger range compared with $\sigma'(\cdot) \in (0, 0.25]$, Tanh should outperform Sigmoid. We think that the saturated area, where the gradients are close to zero, occupies most of the feasible domain in saturated activation functions —— consequently, Tanh still suffers serious deterioration as Sigmoid.

Further, we compare the design principle of the gating mechanism between RNN and Transformer. For RNN, the supervision information is provided for every time step. Thus the gating mechanism should be able to restrict gradient explosion in the backpropagation through time. Differently, the supervision only provided in the last layer of the Transformer Decoder, where the gradient vanishing becomes the dominant problem. Therefore, the identity function should be considered first when stacking deep layers.

2.6 Conclusion

In this work, we devise an effective multimodal sequence modeling framework for image captioning. By introducing the EnTangled Attention and Gated Bilateral Controller, the Transformer model is extended to exploit complementary information of visual regions and semantic attributes simultaneously. Moreover, comprehensive comparisons with state-of-the-art methods and adequate ablation studies demonstrate the effectiveness of our framework.

Chapter 3

Bilateral Attention: Rethinking the Positional Awareness in Self-Attention

3.1 Introduction

Recently, the self-attention mechanism that is originated from the Transformer model (Vaswani et al. 2017) has been extensively explored in both language and vision society, e.g., machine translation (Devlin et al. 2018), semantic segmentation (Huang et al. 2019, Zhao, Zhang, Liu, Shi, Change Loy, Lin & Jia 2018, Peng, Zhang, Yu, Luo & Sun 2017, Fu, Liu, Tian, Li, Bao, Fang & Lu 2019, Zhu, Xu, Bai, Huang & Bai 2019), object detection (Hu, Gu, Zhang, Dai & Wei 2018, Carion, Massa, Synnaeve, Usunier, Kirillov & Zagoruyko 2020), video classification (Wang et al. 2018), and etc. Self-attention aims to compute the response of each query element by a weighted combination of projected keys. However, the compatibility function used in self-attention is solely measured on the feature representations of query-key pairs, which only accounts for content information and ignores position information that is crucial for learning structured data, e.g., images and sequences. Given this situation, there is a surge of methods exploring how to expose position information to the model.

The most popular solution is to simply apply the absolute position encoding to the Transformer model (Vaswani et al. 2017) along with self-attention. Particularly, it encodes the absolute position into a embedding vector and adds them to the inputs to expose information on how a token at one position attend to tokens at other positions. As an alternative to the absolute position encoding, Shaw et al. (Shaw,

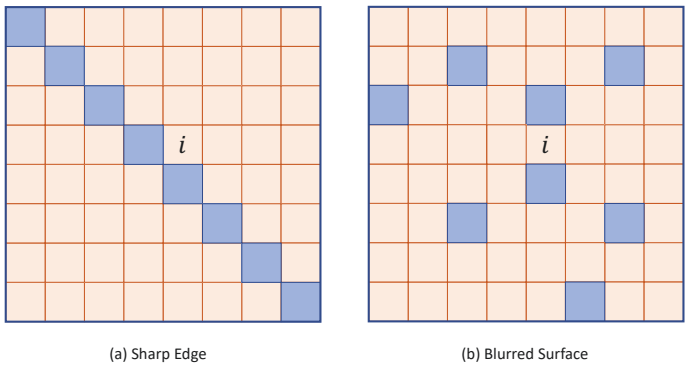


Figure 3.1 : The permutation-invariant property will hurt the capability of self-attention in modeling structured data. e.g., there are eight blue rectangles in both of the images above. The self-attention will generate the same output for the position i, ignoring the local positional relationships are different for a sharp edge and a blurred surface.

Uszkoreit & Vaswani 2018*a*) propose to extend self-attention in Transformer by incorporating learned relative position encoding. Although effectiveness has been validated, relative position encoding actually operates on the same projected query that is also shared for self-attention computation. Considering that heterogeneous property of content and position information, this compounded operation may limit the capacity of the incorporated position information.

With a spirit to effectively incorporate both content and position information, we have noted that the bilateral filter (Tomasi & Manduchi 1998), which combines heterogeneous filters that measure photometric similarity and position closeness of pixels in a bilateral formulation for image smoothing. Motivated by the effectiveness of bilateral mechanism, in this paper we disentangle the handling of position from content and separately learn a position attention. More specifically, we propose a non-localized position attention with dynamic convolution (Wu, Fan, Baevski, Dauphin & Auli 2019*a*) as the key ingredient, which will generate position attention that is aligned to yet still independent from the content-based attention help fully excite the

potential of incorporating position information. The proposed non-localized position attention and the content-based attention are then combined in a bilateral formulation and forms the Bilateral Attention. With the bilateral formulation, heterogeneous information from content and position are combined in a more principled way rather than heuristic designs.

Overall, the main contribution of our work can be concluded in three-fold:

(1) We disentangle position operation from those for content and learn a separate position attention to enforce positional awareness to self-attention. A non-localized position attention is presented to generate position attention that is aligned to yet independent from the content-based attention.

(2) We exploit a bilateral formulation to effectively combine content and position information, which provides a more principled way to incorporate heterogeneous information.

(3) The proposed method demonstrates consistent improvement over its counterparts for representative tasks of semantic segmentation and machine translation, demonstrating the effectiveness and generality.

3.2 Related Work

Self-attention and Variants. Self-attention (Vaswani et al. 2017) aims to compute the response for a query with a weighted summation of transformed keys based on compatibility of query-key pairs. Motivated by its success on sequential problems, several variants have been designed to adapt self-attention to vision tasks (Hu et al. 2018, Wang et al. 2018, Carion et al. 2020, Huang et al. 2019). For example, Wang et al. (Wang et al. 2018) propose a non-local operation that computes response of a position by attending to all the other positions across space, time or spacetime. To releases the computational burden of non-local operation, Huang et al. (Huang

et al. 2019) propose a criss-cross attention that only attends to locations on a criss-cross path of the query location, which is shown to be effective. Self-attention and its variants compute compatibility function solely based on the appearance of query-key features, ignoring position information. In this paper, we propose a bilateral attention to enforces both content and position consistency. Self-attention is also highly related to graph networks (Atwood & Towsley 2016, Niepert, Ahmed & Kutzkov 2016, Gilmer, Schoenholz, Riley, Vinyals & Dahl 2017)

Position Encoding and Variants. To enforce position-awareness with self-attention, the most intuitive solution is adding absolute position encoding to the inputs (Vaswani et al. 2017). As an alternative, Shaw et al. (Shaw et al. 2018*a*) extend self-attention by adding a learned relative position encoding to hidden representations. Recently, several variants have have been proposed to extend relative position encoding to vision tasks. One solution (Parmar, Ramachandran, Vaswani, Bello, Levskaya & Shlens 2019, Bello, Zoph, Vaswani, Shlens & Le 2019) is to factorize it into two dimensions along the weight and height direction, with relative position encoding learned for each dimension. Without directly learning relative position encoding as network parameters, Hu et al. (Hu, Zhang, Xie & Lin 2019*a*) model them with an additional small network based on relative positions of pixel pairs and show better results than the directly learning fashion for image recognition. Despite its effectiveness, relative position encoding operates on the same projected query shared for computing self-attention. In this way, the handling of content and position information is thus compounded, which may hinder the capacity of incorporating position information for representation learning. In this paper, we disentangle position operation from self-attention and separately learn a non-localized position attention to operate position information independently but still uniformly with content-based attention.

Dynamic Convolutions. Dynamic convolutions (Wu et al. 2019*a*) was proposed

recently for machine translation. In an alternative to self-attention, a lightweight depth-wise convolution is employed to aggregate keys for the query. The set of convolution parameters is shared across different query element, which drastically reduces the computational burden of self-attention. On top of the lightweight convolutions, a dynamic convolutions is designed by predicting a different convolution kernel at every time-step based on the input at current time-step only. Dynamic convolutions enjoys the effectiveness of dynamically generated weights at different positions as well as the efficiency of using only a small network to model the weights. In this paper, we extend dynamic convolution as our non-localized position attention to provide aligned yet independent position attention with the content-based attention to enable bilateral attention.

Bilateral Filter. Bilateral filter (Tomasi & Manduchi 1998) is an image smoothing technique which computes a weighted mean of similar and nearby pixel values in an image. The basic idea is that two pixels in the image convey relationships of photometric similarity and position closeness, and either one cannot fully depict the compatibility of pixel pairs. To consider both factors, two filters that are separately in charge of the intensity and position are combined as a bilateral filter to improve filtering performances. Motivated by the effectiveness of bilateral formulation to combine heterogeneous information, in this paper we propose Bilateral Attention to incorporate the position information with self-attention in a more principled way.

3.3 Preliminary: Bilateral Filter

Bilateral filter (Tomasi & Manduchi 1998) is a technique for image smoothing which replaces the intensity of each pixel with a nonlinear combination of intensity values from nearby pixels. Traditional filters compute a weighted average of pixel values in the neighborhood, in which the weights decrease with distance from the neighborhood center. Differently, the weights for bilateral filter depend not only on

Euclidean distance, but also on the photometric differences, which helps to preserve sharp edges after filtering.

Filters that only consider spatial closeness of pixels and use weighs fall off with distance is called domain filter. While filters that measures photometric similarity and use weights decay with dissimilarity is denoted as range filter. Either using one of these two filters alone is not fully sufficient and hence they are combined to form the bilateral filter. Taking a grayscale image I as an example, bilateral filter with normalization factor $\mathcal{C}$ can be written as:

$$\begin{aligned} I_i^{'} &= \frac{1}{\mathcal{C}} \sum_{j \in \Omega} f_r(\|I_i - I_j\|) g_s(\|i - j\|) I_j \\ \mathcal{C} &= \sum_{j \in \Omega} f_r(\|I_i - I_j\|) g_s(\|i - j\|), \end{aligned} \tag{3.1}$$

where $I^{'}$ denote the filtered image, i is the coordinates of the current pixel to be filtered, Ω is the window centered at i, and hence $j \in \Omega$ is another pixel. h_r and p_s represent the kernel for range filter and domain filter, respectively. When Gaussian distribution is exploited to model the distances in the range and in the domain, the hybrid kernel of bilateral filter will be:

$$\begin{aligned} w(i,j) &= f_r(\|I_i - I_j\|) g_s(\|i - j\|) \\ &= \exp(-\frac{\|i - j\|^2}{2\sigma_s^2} - \frac{\|I_i - I_j\|^2}{2\sigma_r^2}) \end{aligned} \tag{3.2}$$

where σ_s^2 and σ_r^2 are smoothing parameters.

3.4 Bilateral Attention

3.4.1 General Formulation

The attention mechanism, which computes the response of the current element by attending to other pixels in its neighbourhood, can also be view as a filtering process. Generally, an attention operation for information aggregation can be written as:

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x}_i)} \sum_{j \in \Omega_i} \alpha(\mathbf{x}_i, \mathbf{x}_j, i, j) \beta(\mathbf{x}_j), \tag{3.3}$$

where $\mathbf{x}_i$ is the original feature representation at position i and $\mathbf{y}_i$ is the newly aggregated feature at position i. $\alpha(\mathbf{x}_i, \mathbf{x}_j, i, j)$, in analogy to a filter kernel, can also be regarded as a kernel function that produces the weight for each projected feature $\beta(\mathbf{x}_j)$ in the neighbourhood Ω_i of feature $\mathbf{x}_i$. $\mathcal{C}(\mathbf{x}_i)$ is the normalization factor which amounts to $\sum_{j \in \Omega_i} \alpha(\mathbf{x}_i, \mathbf{x}_j, i, j)$.

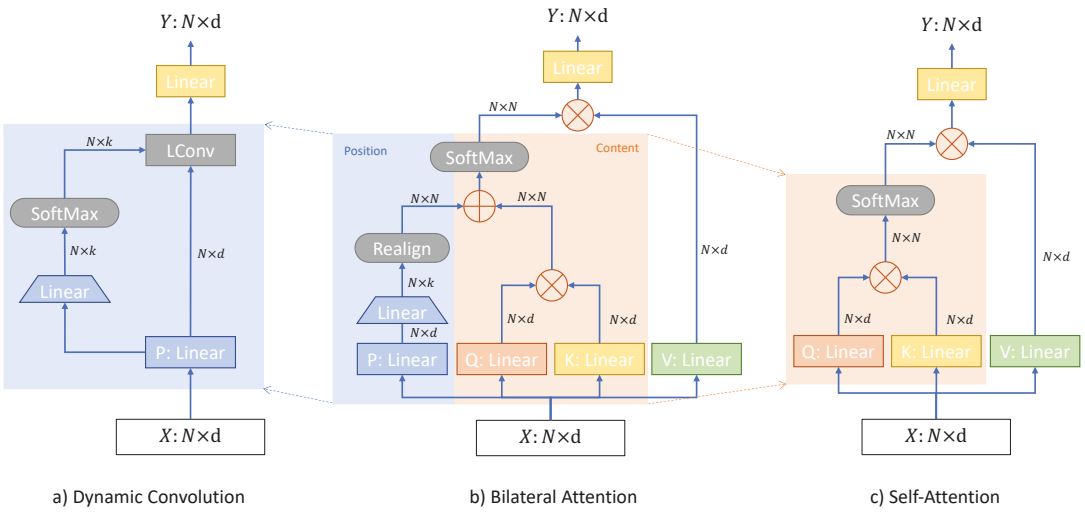


Figure 3.2 : The detailed comparison of a) dynamic convolution, b). bilateral attention, c). self-attention. Our proposed bilateral attention can be viewed as a combination of the self-attention and non-localized position attention, which is adapted from dynamic convolution by replacing the light-weight convolution (LConv) operation with a logit-realignment operation. The logits generated by two components are joined together bilaterally.

Motivated by recent works (Shaw et al. 2018*a*, Parmar et al. 2019, Bello et al. 2019, Hu et al. 2019*a*) that take additional efforts to model position relationship to augment self-attention, we aim to directly encode both the content similarity and position closeness in the kernel function $\alpha(\mathbf{x}_i, \mathbf{x}_j, i, j)$ to generate a set of more representative attention weights for feature aggregation. In analogy to the bilateral filter that considers two filters that are separately in charge of the photometric similarity and spatial closeness, we also have access to two different attention operations, i.e.,

content-based attention and position-based attention, which can be written as:

$$\begin{aligned}\mathbf{y}_i^c &= \frac{1}{\mathcal{C}_c(\mathbf{x}_i)} \sum_{j \in \Omega_i^c} \alpha_c(\mathbf{x}_i, \mathbf{x}_j, i, j)\beta(\mathbf{x}_j) \\ \mathbf{y}_i^p &= \frac{1}{\mathcal{C}_p(\mathbf{x}_i)} \sum_{j \in \Omega_i^g} \alpha_p(\mathbf{x}_i, \mathbf{x}_j, i, j)\beta(\mathbf{x}_j),\end{aligned} \tag{3.4}$$

where α_c is the content-based kernel function that mainly accounts for content similarities and α_p is the kernel functions that mainly in charge of position closeness. As bilateral filter (Tomasi & Manduchi 1998) that considers both position-related and content-related filters to enforce joint consistency, we also combine above two attentions and build up the proposed Bilateral Attention. Following Equation. 3.6, the kernel function for bilateral attention can be initially written as:

$$\alpha(\mathbf{x}_i, \mathbf{x}_j, i, j) = \alpha_c(\mathbf{x}_i, \mathbf{x}_j, i, j)\alpha_p(\mathbf{x}_i, \mathbf{x}_j, i, j). \tag{3.5}$$

For α_c, we follow existing content-based attention mechanisms that operate only on feature representations and simplify $\alpha_c(\mathbf{x}_i, \mathbf{x}_j, i, j)$ as $\alpha_c(\mathbf{x}_i, \mathbf{x}_j)$. While for α_p, the most intuitive way for formulation is to use distances in the image space, which will be solely dependent on position i and j. With only distances considered, a fixed geometry prior will be exposed across different locations, ignoring the context. However, we argue that the distribution of distances for the key element $\mathbf{x}_i$ is desired to be adaptive to the context, especially for vision tasks. For example, the distance distribution for pixels in a round football may prefer to be symmetric in a local window, while for pixels in a curved road, such a symmetric distance distribution may not be suitable anymore. To this end, we formulate the position attention in a way that is sensitive to local context. Thus, $\alpha_p(\mathbf{x}_i, \mathbf{x}_j, i, j)$ will be simplified as: $\alpha_p(\mathbf{x}_i, i, j)$. With the purified content-based and position-based kernel functions, the kernel function for our bilateral attention operation will be updated as:

$$\alpha(\mathbf{x}_i, \mathbf{x}_j, i, j) = \alpha_c(\mathbf{x}_i, \mathbf{x}_j)\alpha_p(\mathbf{x}_i, i, j). \tag{3.6}$$

Following existing attention models that exploits Gaussian function to model weight generation (Vaswani et al. 2017, Wang et al. 2018), we can further present kernel functions for bilateral attention as:

$$\begin{aligned} \alpha_c(\mathbf{x}_i, \mathbf{x}_j) &= \exp(\delta_c(\mathbf{x}_i, \mathbf{x}_j)) \\ \alpha_p(\mathbf{x}_i, i, j) &= \exp(\delta_p(\mathbf{x}_i, i, j)), \end{aligned} \tag{3.7}$$

where $\delta_c \in R$ and $\delta_p \in R$ are distance functions based on content and position, respectively. In this case, the overall normalization factor will become: $\mathcal{C} = \sum_{j \in \Omega_i} \exp(\delta_c(\mathbf{x}_i, \mathbf{x}_j) + \delta_p(\mathbf{x}_i, i, j))$. With the softmax function, we can wrap bilateral attention in a more compact mode:

$$\mathbf{y}_i = \sum_{j \in \Omega_i} \mathbf{softmax}\left[\delta_c(\mathbf{x}_i, \mathbf{x}_j) + \delta_p(\mathbf{x}_i, i, j)\right] \beta(\mathbf{x}_j), \tag{3.8}$$

Up to now we have finished building up most of the basis for the proposed bilateral attention operation. However, in practical, we found that logits' values of $\delta_c(\mathbf{x}_i, \mathbf{x}_j)$ and $\delta_p(\mathbf{x}_i, i, j)$ sometimes can resident in very distinct ranges, making the direct summation biased to values with large magnitude. Besides, logits with large magnitude values will also tend to saturate the softmax function, which should be avoided. To counteract this effect, we scale the output logits of function δ_c and δ_p with two smoothing factors: σ_c and σ_p, respectively. The role of smoothing factor is similar to the temperature set for in self-attention (Vaswani et al. 2017), which has shown to be effective. With smoothing factors, our bilateral attention will accordingly be updated:

$$\mathbf{y}_i = \sum_{j \in \Omega_i} \mathbf{softmax}\left[\frac{\delta_c(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_c} + \frac{\delta_p(\mathbf{x}_i, i, j)}{\sigma_p}\right] \beta(\mathbf{x}_j). \tag{3.9}$$

σ_c and σ_p can be chosen as pre-defined hyper-parameters. Beyond using fixed σ_s and σ_p for magnitude scaling, we also propose an adaptive way to balance the magnitude of outputs from two distance functions, which is demonstrated to be more effective in our experiments.

3.4.2 Formulation of Distance Functions

In terms of the distance function $\delta_c(\mathbf{x}_i, \mathbf{x}_j)$ for content-based attention, we follow the generally adopted dot-product approach (Vaswani et al. 2017, Wang et al. 2018) to model the content-wise similarity:

$$\delta_c(\mathbf{x}_i, \mathbf{x}_j) = \theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j), \tag{3.10}$$

where θ and ϕ are projection functions that encode input $\mathbf{x}$ into a new embedding space. Usually they are implemented as $\theta(\mathbf{x}) = W_\theta \mathbf{x}$ and $\phi(\mathbf{x}) = W_\phi \mathbf{x}$, each of which parameterized by W_θ and W_ϕ, respectively. With distance function δ_c, various content-based attention mechanism can be instantialized in terms of different neighbourhood Ω_i. For example, Ω_i could be a 1-D window for sequences or a 2-D window for images, which corresponds to self-attention and non-local operation, respectively. Details on the instantiations together with the position attention are presented in Section. 3.4.4.

When it comes to the distance function $\delta_p(\mathbf{x}_i, i, j)$ in the position attention, we adopt dynamic convolutions (Wu et al. 2019*a*) as a key ingredient. Dynamic convolutions model the kernel weights of a depth-wise convolution as a function of input at the current time step, with dynamic weights derived at different positions. Similarly, we model the position attention weights with an additional small network followed by the softmax normalization, based on the local query element $\mathbf{x}_i$:

$$\mathbf{W}_{\Omega_i^p} = f(\mathbf{x}_i), \tag{3.11}$$

where $\mathbf{W}_{\Omega_i^p}$ is the logits to be fed into the softmax function, Ω_i^p denotes a pre-defined local window with length k (in 1-D situation). As shown in the blue part of Figure. 3.2 (b), position-related compatibility function f is simply implemented as a two-layer neural networks, a linear embedding layer first projects features into the embedding space, followed by a linear mapping layer outputting k logits for each of N features.

Benefited from the localized nature of dynamic convolution (Wu et al. 2019*a*), the position-related compatibility function f can be applied to arbitrary-sized sequences/image, just by sliding the function across different query elements. Dynamic attention logits will be generated subsequently for each query element $\mathbf{x}_i$. However, the efficiency of dynamic convolution incurs another problem: for one specific query element $\mathbf{x}_i$, the logits $\mathbf{W}_{\Omega_i^p}$ is only valid within the defined local window Ω_i^p, which may not align with the content-based attention. For example, if content-based attention is instantiated as self-attention which attends to every other positions in a sequence, the result from the compatibility function f in the position attention that only occupy a local window will cause problematic combination of the content-based and position-based attention. To mitigate this discrepancy, we further complete the distance function δ_p with a padding operation, with additional values padded for positions that are outside of the defined window Ω_i^p while are considered in the content-based attention. We thus propose the *non-localized position attention* which is modeled with a piecewise distance function δ_p to cater for "outlier" positions and enable alignment of attentions from the position and content domain. Considering the 1-D situation, the distance function δ_p for our non-localized position attention can be written as:

$$\delta_p(\mathbf{x}_i, i, j; k) = \begin{cases} f(\mathbf{x}_i)_{\lceil \frac{k-1}{2} \rceil + 1 - (i-j)}, & |i-j| \leq \lceil \frac{k+1}{2} \rceil \\ v_i, & |i-j| > \lceil \frac{k+1}{2} \rceil \end{cases} \tag{3.12}$$

where $f : R^d \mapsto R^k$, and $|i-j|$ denotes the relative distance between position i and j, k is the length of window Ω_i^p defined in f, v_i is the padded logits that is constant across the 'outlier' positions. $f(x)_s$ indicates the value in $f(x)$ at index s. As shown in Figure. 3.2, we wrap up all the padding operations for N query elements and realign it as a $N \times N$ maps, which is to be ready to fused with the $N \times N$ outputs from the self-attention operation. Besides, it is also intuitive to extend δ_p to 2-D situation with a window size of $k \times k$ for compatibility function $f : R^d \mapsto R^{k \times k}$.

The relative distance $|i - j|$ will be measured across the width and height direction, respectively.

As illustrated in Figure. 3.2, the realign operation bridges the gap between the localized dynamic convolution as position attention and the global content-based attention, making it possible for the proposed bilateral attention to effectively benefit from both the content and position information. Various potential padded values v_i also injects flexibility to our bilateral attention. For example, if the padded value v_i is set to $-\infty$, with the softmax operation, the bilateral attention for padded positions will be 0, which amounts to the situations that the 'outlier' positions are not considered completely. The padded value can also be set statistically, e.g., as the average or minimum values of outputs from the compatibility function $f(\mathbf{x}_i)$, and it is also possible to ask $f(\mathbf{x}_i)$ to directly output one more logit as the padding value. Detailed experimental results and analysis in terms of the padded values will be presented in Section 3.4.4.

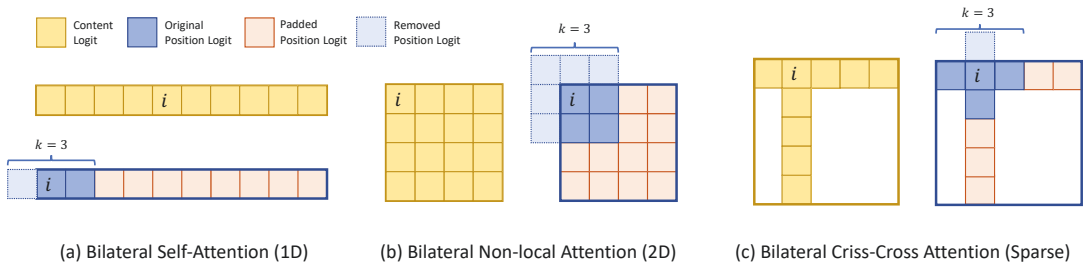


Figure 3.3 : The generated content logits and position logits for position i. For current position i, each of the small yellow rectangle denotes content logit in position j and each of the small blue or pink rectangle denotes position logits.

3.4.3 Distance Function in 2-D Mode

In the paper we have described the 1-D distance function to compute the position attention, e.g., for sequential models. Here we further illustrate its 2-D version.

For the 2-D bilateral non-local attention (Bi-NL), kernel size k is expanded as (k_0, k_1) as well as the position $i = (i_0, i_1)$, $j = (j_0, j_1)$, $i - j = (i_0 - j_0, i_1 - j_1)$. Similarly, the distance $|i - j| \leq \lceil \frac{k+1}{2} \rceil$ requires $|i_0 - j_0| \leq \lceil \frac{k_0+1}{2} \rceil$ and $|i_1 - j_1| \leq \lceil \frac{k_1+1}{2} \rceil$. $f(\mathbf{x}_i)_{\lceil \frac{k-1}{2} \rceil + 1 - (i-j)}$ is the entry located at $\lceil \frac{k-1}{2} \rceil + 1 - (i_0 - j_0)$-th row and $\lceil \frac{k-1}{2} \rceil + 1 - (i_1 - j_1)$-th column of matrix $f(\mathbf{x}_i)$. Besides, the function $f : \mathbb{R}^d \mapsto \mathbb{R}^{k_0 \times k_1}$ will generate $k_0 \times k_1$ logits, covering each position in the window centered at i.

The coordinate expanding in 2-D criss-cross attention (Bi-CC) follows similar policy, while the difference lies in distance measure $|i - j|$. In Bi-CC, only the elements in a criss-cross path are considered. Therefore, $|i - j| \leq \lceil \frac{k+1}{2} \rceil$ denotes $|i_0 - j_0| \leq \lceil \frac{k_0+1}{2} \rceil$ when $i_1 = j_1$ or $|i_1 - j_1| \leq \lceil \frac{k_1+1}{2} \rceil$ when $i_0 = j_0$. Moreover, in order to avoid the double counting of $j = (i_0, i_1)$, we only consider $k_0 + k_1 - 1$ elements in the position logit generation. Consequently, the function $f : \mathbb{R}^d \mapsto \mathbb{R}^{k_0 + k_1 - 1}$.

3.4.4 Instantiations

As mentioned in Section. 3.4.2, different definitions of neighbourhood Ω_i for information aggregation will give rise to distinct type of content-based attention mechanisms. For illustration, we instantiate bilateral attention with three typical content-based attentions, in conjunction with the corresponding non-localized position attention.

Bilateral Multi-head Self-attention. The first content-based attention we consider is the self-attention in the Transformer model for machine translation (Vaswani et al. 2017), with neighbourhood Ω_i defined as all the positions in the sequence. Since multiple attention heads is always jointly used with self-attention, we extend the compatibility function $f(\mathbf{x}_i)$ (Equation 3.12) for position-based attention generation from $f : \mathbb{R}^d \mapsto \mathbb{R}^k$ to $f : \mathbb{R}^d \mapsto \mathbb{R}^{m \cdot k}$, with k logits generated for each of the m attention head. After padding, the logits for self-attention and non-localized attention will align with each other across multiple heads to enable proper processing.

It is also notable that masked self-attention is used in the transformer decoder to avoid attending to locations that are later than the current time-step. Accordingly, the window size k for non-localized position attention used in the decoder needs to be modified to $\lceil \frac{k+1}{2} \rceil$ to keep consistency. Using Equation. 3.9, the bilateral multi-head attention can be derived by combining the kernel function for multi-head self-attention and the corresponding non-localized position attention.

Figure. 3.3 (a) illustrates the process to generate logits for the proposed non-localized position attention (lower row) given the logits for self-attention (upper-row) for 1-D sequence. When we talk about logits in terms of attention, it usually denotes values that are yet to be passed to the softmax function for attention weight generation, i.e., output of distance function δ_c for content-based attention and δ_p for non-localized position attention here. As observed in Figure. 3.3 (a), to generate the non-localized position logits that are aligned to those of the self-attention (yellow squares), the compatibility function $f(\mathbf{x}_i)$ which is modeled as a small network will be first employed to generates the position logits (deep blue squares) within a window of length k. For positions that are not accessible by function $f(\mathbf{x}_i)$ but are considered in the self-attention, logits will be padded accordingly (pink squares). If the position i of the current query element happens to be around the border of the sequence, position logits that exceed the neighbourhood range will be removed for compatibility (light blue squares).

Bilateral Non-local Attention. The second content-based attention being considered is the non-local attention proposed in (Wang et al. 2018) for vision tasks. In this situation, it matches to the case when the neighbourhood Ω_i in the bilateral attention spans over the whole feature maps. Accordingly, the 2-D version of distance function δ_p should be employed, i.e., $f : \mathbb{R}^d \mapsto \mathbb{R}^{k \times k}$. Figure 3.3 (b) illustrates an example case of logits prediction for the non-local attention and the non-localized position attentions for 2-D images, and for each dimension similar operations can be

observed as in the Bilateral Multi-head Self-attention.

Bilateral Criss-cross Attention. Another content-attention we take into consideration is the criss-cross attention proposed in (Huang et al. 2019) for semantic segmentation. For efficient computation, it adopts a criss-cross path for the current input, and hence the neighbourhood Ω_i for feature aggregation will be the criss-cross path centered at element $\mathbf{x}_i$. Correspondingly, the compatibility function $f(\mathbf{x}_i$ will become $f : R^d \mapsto R^{2k-1}$. Similarly, Figure. 3.3 (c) also illustrates the process to generate logits of the criss-cross attention and non-localized position attention.

3.5 Experiments for 2-D Bilateral Attention

To demonstrate the effectiveness of the proposed bilateral attention, we first exploit the task of semantic segmentation as the testbed. The dense property of semantic segmentation task and its desire to the content and position infomration makes it a suitable task to diagnostic our 2-D bilateral attention. Two popular attention models, i.e., non-local (NL) network (Wang et al. 2018) and criss-cross (CC) network (Huang et al. 2019), are selected as the frameworks to embody the proposed bilateral attention. By replacing the corresponding self-attention layer and the criss-cross attention layer with the bilater attention, we can obtain the bilateral non-local network (Bi-NL) and the bilateral criss-cross (Bi-CC) network.

Experimental Setup We use the open sourced library: MMSegmentation* for model implementation. Besides, ImageNet-pretrained Resnet-50 and Resnet-101 build up model backbones. For other network configurations, we follow those in the previous work (Chen, Papandreou, Schroff & Adam 2017) and the output stride is set to 8.

Following prior works (Chen, Papandreou, Schroff & Adam 2017), we employ a

*https://github.com/open-mmlab/mmsegmentation

poly learning rate policy with the initial learning rate as 1e-2, which is multiplied by $(1 - \frac{iter}{max_iter})^{power}$ with *power* set to 0.9. We set maximum iteration number to 80K for Cityscapes, and 160K for experiments on the ADE20K. Momentum and weight decay are set to 0.9 and 0.0001, respectively. For Cityscapes, the training images are first augmented by randomly scaling (from 0.75 to 2.0), from which high-resolution patches (769×769) are randomly cropped. Similar scaling strategy is applied to the training images in ADE20K, with randomly cropped patches in the size of 512×512 patches. For evaluation, the single-scale testing strategy is applied unless specifically denoted.

We set the default local window size of the neighbourhood Ω_i^p of the compatibility function f in the non-localized position attention to 31. For the smoothing factors σ_c and σ_p to balance the magnitude of content and position logits in the bilateral attention, it could be chosen as $\sqrt{d}$ where d is the number of feature channel. We also consider another way which directly normalizes the logits to $(0 \sim 1)$ with their mean and std value. we set the default optio to be an impleicit operation of normalization operation is set as default choice.

Datasets Comprehensive experiments have been carried on two challenging datasets for semantic segmentation: Cityscapes (Cordts, Omran, Ramos, Rehfeld, Enzweiler, Benenson, Franke, Roth & Schiele 2016) and ADE20K (Zhou, Zhao, Puig, Fidler, Barriuso & Torralba 2017). Cityscapes (Cordts et al. 2016) contains 5,000 images and the size of each image is in 2048×1024. It provides high quality pixel-level annotations of 19 semantic classes. There are 2,979, 500 and 1,525 images for training, validation and testing, respectively. ADE20K (Zhou, Zhao, Puig, Fidler, Barriuso & Torralba 2017) is a very challenging scene parsing benchmark which contains dense labels of 150 stuff/object categories. There are separately 20K, 2K and 3K images set for training, validation and testing. In terms of evaluation metrics, both *mean of class-wise intersection over union* (mIOU) and *pixel-wise accuracy*(PixAcc) are

adopted for evaluation. All the results are reported on the validation sets.

3.5.1 Comparison Results

Table 3.1 : Comparison results on Cityscapes and ADE20K, and multiple-scale is applied for testing.

Dataset	Model	Attention Type	mIOU
Cityscapes	NL (Wang et al. 2018)	Content	80.85
	Bi-NL	Position + Content	81.13
	CC (Huang et al. 2019)	Content	81.30
	Bi-CC	Position + Content	81.45
ADE20K	NL (Wang et al. 2018)	Content	44.83
	Bi-NL	Position + Content	45.23
	CC (Huang et al. 2019)	Content	45.22
	Bi-CC	Position + Content	45.54

Table 3.1 shows the comparison results of the proposed bilateral attention with two popular content-based attention models, i.e., non-local network (NL) and criss-cross network (CC), on the two datasets of Cityscapes and ADE20K, respectively. It can be observed that with the proposed bilateral attention , i.e., models Bi-NL and Bi-CC, leads to consistent improvements over the two strong baselines across the two datasets. This demonstrates the necessity of extending the self-attention with positional awareness, as well as the efficiency of the bilateral attention.

3.5.2 Comparison to Relative Position Encoding

To further validate the effectiveness of the proposed non-localized position attention within the bilateral attention, we conduct additional experiments on Cityscapes to compare with the relative positional encoding (RPE) (Bello et al. 2019), a popular method to inject position information into self-attention. In particular, We utilize the

Table 3.2 : Comparison of bilateral attention with relative position encoding (RPE) on Cityscapes.

Model	Resnet-50		Resnet-101	
	mIOU	PixAcc	mIOU	PixAcc
NL	78.98	86.55	79.40	86.55
NL+RPE (Bello et al. 2019)	79.31	87.02	79.47	87.18
Bi-NL	79.80	87.20	79.89	87.45
CC	79.10	86.11	79.45	86.98
CC+RPE (Shaw et al. 2018*a*)	79.64	87.14	79.81	86.95
Bi-CC	80.17	87.43	80.45	87.79

self-attention block with decomposed relative positional encoding in (Bello et al. 2019) for comparison. As shown in Table 3.2, both positional methods can improve over the baseline, while the non-localized attention bring higher gains than RPE (Bello et al. 2019). Notably for with the CC framework, the bilateral combination of positional information can result in at least 1% increment in term of mIOU for both backbones.

3.5.3 Controlled Experiments

We conduct comprehensive ablation studies to diagnostic the proposed bilateral attention in 2-D mode. All the experiments are conducted on Cityscapes based on Resnet-50, with single-scale testing mode.

Padding and Smoothing The padding operation in the non-localized positional attention and the smoothing strategies in the bilateral attention has significant effects to the results. As shown in Table 3.3, when padding with $-\infty$, the bilateral attention will degenerate to a local operator with the kernel size of k. Even with a large kernel ($k = 31$), the performance still drops dramatically, indicating the necessary to consider the "outlier" positions. When padding with 0, the position logits act as

Table 3.3 : Ablation experiments on Cityscapes.

Model	Padding $v(\mathbf{X}_i)$	$\sigma_c = \sigma_p = \sqrt{d}$		$\mathcal{N}(\delta_p(\mathbf{x}_i))$	
		mIOU	PixAcc	mIOU	PixAcc
Bi-CC	$-\infty$	74.98	83.81	-	-
	0	79.66	87.05	79.91	86.96
	$f(\mathbf{x}_i)_{k+1}$	79.97	87.28	80.17	87.43
Bi-NL	0	77.87	85.86	-	-
	$f(\mathbf{x}_i)_{k+1}$	79.24	86.63	79.8	87.20

inductive bias for the content logits within the kernel size while leaving the logits out of the kernel size as it is. Note in this situation the performance for Bi-CC model increases by 0.56% while the performance for Bi-NL decreases by more than 1%. When the smoothing factor $\sigma_c = \sigma_p$ is $\sqrt{d}$ as in the self-attention, we observe consistent increments for both Bi-CC and Bi-NL by changing the zero-padding with a learned value of $f(\mathbf{x}_i)_{k+1}$. When we use the z-score normalization, we also observe similar improvements enabled by padding with a learned value.

Local Window Size. Table 3.4 shows the effects of using different size for the local window Ω_i^p in the non-localized position attention with bilateral criss-cross attention model. It can be observed that the bilateral attention outperforms the original criss-cross attention when the local window size is larger than 15. We conjecture that, with a small window size, the positional logits tend to restrict the information in a local range. When the window size is larger than 31, the performance tends to saturation. Based on these observation, we set the window size as 31 in our experiments considering the computational efficiency.

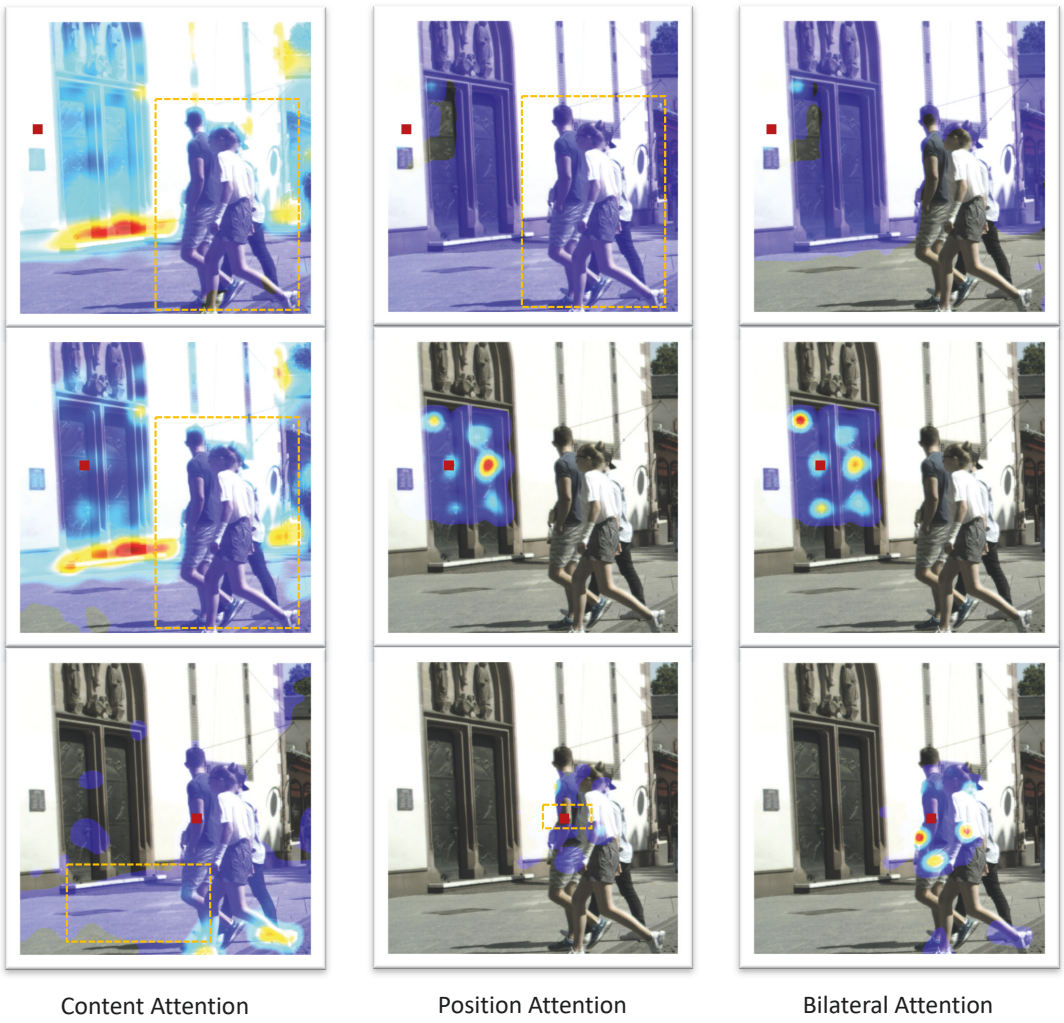


Figure 3.4 : The detailed comparison of a) dynamic convolution, b). bilateral attention, c). self-attention. Our proposed bilateral attention can be viewed as a combination of the self-attention and non-localized position attention, which is adapted from dynamic convolution by replacing the light-weight convolution (LConv) operation with a logit-realignment operation. The logits generated by two components are joined together bilaterally.

Table 3.4 : The controlled comparison of different kernel size of the bilateral criss-cross attention.

Kernel Size	Resnet-50		Resnet-101	
	mIOU	PixAcc	mIOU	PixAcc
15	79.38	86.33	79.41	86.70
31	80.17	87.43	80.45	87.79
63	79.77	87.23	80.23	87.85
127	80.10	87.29	80.36	87.23

3.5.4 Visualization of Attention Maps

To demonstrate the effectiveness of the proposed bilateral attention, we visualize the attention maps using the Bilateral Non-local (Bi-NL) model for a sample image[†] from the Cityscapes dataset. Bi-NL uses ResNet-50 as backbone. In Figure 3.4, each row shows the content attention, position attention and the proposed bilateral attention sequentially for a unique query position, which is marked as red dot on the image. For better visualization results, we truck each attention map by ignoring attention values smaller than 0.02. As observed from the first row of Figure 3.4, for a query position located on the wall, either the content or position attention tends to attend to the whole image, which is not desired. While our bilateral attention effectively aggregates the two heterogeneous information and generates reasonable attention map that more concentrates to the wall and gate, avoiding the attention on irrelevant ground and pedestrians, as denoted in the yellow dashed box. Similar situations can be observed in the other two rows with different query positions selected. The bilateral attention map shows consistently shows reasonable results compared to content or position attention solely, especially for the areas in the yellow dashed boxes.

[†] `val/frankfurt/frankfurt_000001_011835_leftImg8bit.png`

3.6 Experiments For 1-D Bilateral Attention

To validate the effectiveness of bilateral attention in 1-D situation, We choose the neural machine translation task as another testbed.

Experiment Setup. We implement our algorithm based on the open-sourced library FariSeq[‡]. The evaluation experiment are conducted on three mainstream datasets: WMT English to German (En-De), WMT English to French (En-Fr) and IWSLT German to English (De-En). We follow steps mentioned in (Wu et al. 2019*a*) for pre-processing. And we apply byte-pair encoding (BPE) (Sennrich, Haddow & Birch 2016) to the sentences and generate joint vocabulary shared across source and target vocabulary.

Evaluation Protocol. For all experiments, we measure case-sensitive tokenized BLEU scores with multi-bleu[§]. Following (Vaswani et al. 2017), we apply compound splitting for WMT En-De. For all datasets, we used beam search with beam with 5. And, we tuned a length penalty as well as the number of checkpoints to average on the validation set.

3.6.1 Comparison to the State-of-the-art Methods

We compare our Bi-SA with the state-of-the-art methods in machine translation, by replacing all the self-attention layers in Transformer model with Bi-SA. Table 3.5 shows that our method outperforms current state-of-the-art methods on all three datasets. Specifically, our bilateral self-attention defeats the relative position encoding method (Shaw et al. 2018*a*) with a large margin. Moreover, we emphasis that our Bi-SA also outperforms the Transformer and dynamic convolution, which can be treated as our basis.

[‡] https://github.com/pytorch/fairseq

[§] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

Table 3.5 : Translation quality evaluation (BLEU scores).

Model	WMT' 14		IWSLT'14
	En-De	En-Fr	De-En
Transformer (Vaswani et al. 2017)	28.4	41.0	34.4
Transfomer + RPE (Shaw et al. 2018*a*)	29.2	41.5	-
Dynamic Conv (Wu et al. 2019*a*)	29.7	43.2	35.2
Scaling NMT (Ott, Edunov, Grangier & Auli 2018)	29.7	43.2	-
Locality (Fonollosa, Casas & Costa-jussà 2019)	29.7	43.3	35.7
Large Kernel (Lioutas & Guo 2020)	29.6	43.2	35.5
Bi-SA (Ours)	**29.85**	**43.42**	**35.93**

3.6.2 Controlled Experiments

We conduct comprehensive ablation studies to illustrate the behaviors of the proposed bilateral attention in 1-D mode. All experiments are conducted on IWSLT'14 De-En dataset. And the kernel size for encoder and decoder are 21 and 11, respectively.

Padding. Here we evaluate the padding and smoothing strategies in the non-localized positional attention. As shown in Table 3.6, when padding with the minimum of $f(\mathbf{x}_i) \in \mathbb{R}^k$, the performance degenerates dramatically compared with the SA baseline. This is may caused by the inaccurate estimation of the minimum, because that some of the logits will be ignored in the realignment process. When padding with $-\infty$, the non-localized positional attention degenerates as a local operator, the same to the self-attention. Unlike the performance corruption in the segmentation task, the performance in NMT still increases a little compared with the transformer baseline. This reflects that the kernel size is large enough for natural language. We also try to pad with $f(\mathbf{x}_i)_{k+1}$ and 0. And zero-padding achieves the best performance. As the experiments in segmentation, we also try to apply normalization operation on the realigned positional logits, but the experiments

Table 3.6 : Ablation experiments on IWSLT'14

Model	Padding $v(\mathbf{X}_i)$	BLEU
Bi-SA	$\min(f(\mathbf{x}_i))$	31.28
	$-\infty$	35.43
	$f(\mathbf{x}_i)_{k+1}$	35.73
	0	**35.93**
SA (Ott et al. 2018, Vaswani et al. 2017)	-	35.09
SA-RPE (Shaw et al. 2018*a*)	-	35.49

Table 3.7 : The effect of the kernel size in geometric attention

Kernel Size		BLEU
Encoder	Decoder	
05	21	35.36
11	21	35.50
11	06	35.60
21	11	35.93
31	16	35.68
41	21	35.69

shows normal training process, while in testing, the normalization prevent the model from generating *EOS* (end of sentence) token. This phenomenon requires further investigation. Therefore, in the translation experiments, we use the root of embedding dimension $\sqrt{d}$ as the smoothing factor for logits.

Local Window Size. Similar to the segmentation task, the local window size of the sequence model is also very important. As shown in Table 3.7, the performance reaches the best when the kernel sizes are 21 and 11 respectively.

3.7 Conclusion

In this paper, we aim to design a new attention scheme, taking both content similarities and position closeness into account. Accordingly, we propose bilateral attention, which first disentangles position-related attention from content-related one and then combines them via bilateral formulation to enforce jointed consistency. Our bilateral attention is generic and can help augment previous content-based attention approaches to achieve better performance. We investigate the effects of our bilateral attention on several popular attention approaches, including 1D-based self attention and 2D-based no-local and criss-cross attention. Extensive experiments for NLP and Vision tasks well demonstrate the effectiveness of our bilateral attention.

Chapter 4

Localized Bilateral Attention with Iterative Refinement for Image Segmentation

4.1 Introduction

As one of the most fundamental tasks in computer vision, semantic segmentation produces considerable impacts on various real-world applications, such as autonomous driving (Fritsch, Kuehnl & Geiger 2013), medical image analysis (Ronneberger, Fischer & Brox 2015), and image editing (Gu, Bao, Yang, Chen, Wen & Yuan 2019). Recently, the application of fully convolutional networks (Chen, Papandreou, Kokkinos, Murphy & Yuille 2017, Yu & Koltun 2016) under the encoder-decoder architecture (Chen, Zhu, Papandreou, Schroff & Adam 2018, Ronneberger et al. 2015) brings great leap-forwards to semantic segmentation.

In modern segmentation frameworks, the feature aggregation across different levels (Chen, Zhu, Papandreou, Schroff & Adam 2018, Ronneberger et al. 2015) is widely used in the decoding stage. The high-level features are upsampled and then fused with the low-level features, which contains rich spatial information, by a convolution layer. Two challenges accompany this process. First, the content-agnostic upsampling operations tend to over smooth the feature maps and consequently blur the edges in the prediction. Second, the downsampling/upsampling operations in the model cause feature misalignment in the fusion of different level of features.

The conventional convolution has limitations in handling the above two issues. First, the convolution is content-agnostic. A fixed kernel is usually applied to all the pixels irrespective of their content, i.e., no matter the pixel is in a smooth surface or

a sharp edge. Second, the spatial aggregation in the convolution is only a simple summation operation, without any balancing strategy among the candidate features in the receptive field. This further sets barriers in the selection of the most aligned features from the supporting region.

Alternatively, we investigate the potential of another group of atomic operations – attention mechanisms to improve the decoding efficiency for segmentation. Most existing segmentation methods take advantage from the attention mechanism to model the long-range interactions in the encoding stage, while the usage of attention mechanism in the decoder stage has not been fully explored. In this paper, we propose locally bilateral attention (BA) to help excite the performance of the decoder. Specifically, the locally bilateral attention is modeled as the coupling of a content attention and a geometric attention, which will adaptively attend to local features that share both content and geometry-wise similarity to the target feature. In this way, it will be more effective to model the varying spatial layouts and relieves the feature misalignment in the decoding process. To improve the accuracy of the estimated attention affinities, we further propose an iterative refinement scheme to complete the design of the BA module. Beyond the effectiveness, the proposed LBA module also enjoys huge efficiency compared to either of the convention non-local attention and the convolution operation. For example, the computational complexity is reduced from $o(N^2)$ to $o(N)$ for the processing of a feature map with totally N pixels. While compared to a convolution layer with kernel size $N \times N$, the BA also shows surpassing efficiency. These features make the BA especially suitable for the processing of the decoder, which usually operates on high-resolution feature maps.

Overall, our contributions can be summarized as follows. First, a general locally bilateral attention, which considers the appearance and geometry information simultaneously, is proposed. Second, we propose an iterative refinement procedure to further purify the attention weights. Moreover, detailed experiments on benchmark

segmentation datasets and two classical backbones show that bilateral attention outperforms the standard convolution in both accuracy and robustness.

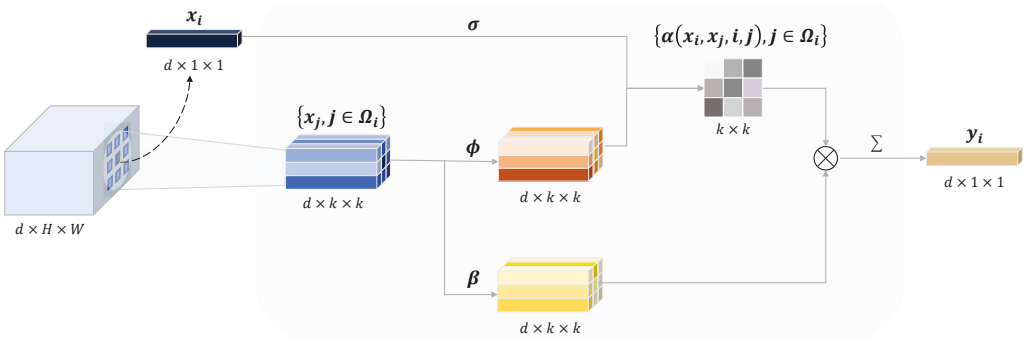


Figure 4.1 : Illustration of the attention mechanism. The feature maps are shown as the shape of their tensors, e.g., $d \times H \times W$ for d channels. $\otimes$ denotes matrix multiplication. Given an input feature map X, the attention mechanism transforms each of its feature vector x_i to y_i by dynamically aggregating the contents from a $k \times k$ neighborhood of x_i via an attention weight map. Specifically, the target feature x_i and its neighborhood features $x_j; \forall j \in \Omega_i$ are first mapped to the common space via function γ and ϕ (γ and ϕ are usually implemented as linear transformation), respectively, to generate the query-key feature pairs. Meanwhile, the neighborhood features are further transformed to values via function β. Considering the appearance similarity of the query-key pairs, the attention model generates an attention map, which are then used as pixel-wise weighting scalar to aggregate the transformed values and output the final feature y_i.

4.2 Background

4.2.1 The General Formulation

The attention operation can be viewed as a filtering process, which is applied to an input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times d_{in}}$ to produce a response $\mathbf{Y} \in \mathbb{R}^{H \times W \times d_{out}}$. An attention function, which is given a query element $\mathbf{x}_i, i \in [1, \ldots, HW]$ and a set of

key elements $\mathbf{x}_j$ in supporting set Ω_i, can dynamically aggregate the key content according to attention weight $\alpha(\mathbf{x}_i, \mathbf{x}_j, i, j) \in \mathbb{R}$ that measures the compatibility of query-key pair. This process can be formulated as follows:

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x}_i)} \sum_{j \in \Omega_i} \alpha(\mathbf{x}_i, \mathbf{x}_j, i, j) \odot \beta(\mathbf{x}_j) \tag{4.1}$$

where $\odot$ is the Hadamard product, and $\alpha(\mathbf{x}_i, \mathbf{x}_j, i, j)$ produces a scalar to measure the *content similarity* or *geometric relation* between the centre $\mathbf{x}_i$ and a nearby point $\mathbf{x}_j$. $\beta : R^{d_{in}} \mapsto R^{d_{out}}$ is a transformation, $\mathcal{C}(\mathbf{x}_i)$ acts like a normalization factor.

To allow the model to attend to the representations from different subspaces and different positions, the feature space is divided into M parallel subspaces, and M attention modules are conducted independently in each of the subspace.

$$\mathbf{y}_i = W_M[\dots, \frac{1}{\mathcal{C}_m(\mathbf{x}_i)} \sum_{j \in \Omega_i} \alpha_m(\mathbf{x}_i, \mathbf{x}_j) \odot \beta_m(\mathbf{x}_j), \dots], \tag{4.2}$$

Here m indexes over the total M attention heads, $[\dots]$ denotes the concatenation of vectors, $W_H \in \mathbb{R}^{d_{out} \times d_{in}}$ is the learnable weights. In the following section, we mainly discuss the attention in one subspace, and head index m is omited for the sake of concise expression.

In the upsampling process of low-resolution feature maps, the distribution or texture over a small supporting region is much more important than the "non-local information". Therefore, we only consider a region with kernel size k around the centre position i instead of all the N positions. Due to the fact that $k \ll N$, the proposed attention is more computational efficient and memory-friendly than the non-local operation, which are crucial for the processing of large resolution inputs. Even compared with convolution operation which is also a local operation, the local attention still has advantages on the number of parameters and FLOPs as well.

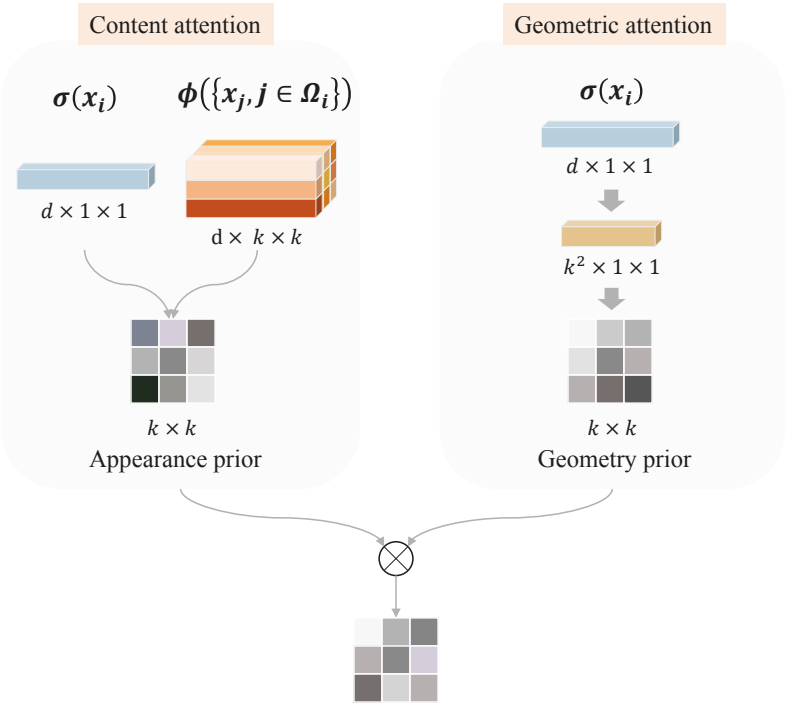


Figure 4.2 : A bilateral attention module. The bilateral attention operation includes two parts: the content attention part and the geometric attention part. The content attention generates appearance-based attention priors by measuring the similarity of the appearance between the target feature and its surrounding features. We further incorporate a geometric attention to generate geometry-based attention priors based on the embedding of features' positions. Finally, the two independently generated attention priors are combined to form the final bilateral attention weight.

4.2.2 Content Attention

In the content attention, only the information in the feature space is taken into consideration. Consequently, the affinity function $\alpha(\mathbf{x}_i, \mathbf{x}_j, i, j)$ degenerates as $\alpha(\mathbf{x}_i, \mathbf{x}_j)$. Inspired by the classical image filter, e.g.: bilateral filter, the gaussian distribution is a good choice to model the distribution of affinity.

$$\alpha_c(\mathbf{x}_i, \mathbf{x}_j) = \exp(\sigma(\mathbf{x}_i)^T \phi(\mathbf{x}_j)). \tag{4.3}$$

Here ϕ and σ project the input features into the common space, in which the inner-product can be applied to measure the content affinity. Noticing that $\mathcal{C}(\mathbf{x}_i) = \sum_{j \in \Omega_\mathbf{i}} \exp(\sigma(\mathbf{x}_i)^T \phi(\mathbf{x}_j))$ and $\alpha(\mathbf{x}_i, \mathbf{x}_j)/\mathcal{C}(\mathbf{x}_i)$ is the *Softmax* function. There are various options for the projection function σ and ϕ, e.g., a multi-layer perceptron. In our implementation, the linear transformation is applied:

$$\alpha_c(\mathbf{x}_i, \mathbf{x}_j) = \exp(\mathbf{x}_i^T W_\sigma^T W_\phi \mathbf{x}_j), \tag{4.4}$$

Besides, σ and ϕ can the same function, or even an identical mapping, in the context of self-attention where $\mathbf{x}_i$ and $\mathbf{x}_j$ share a common feature space. There are several works (Hu, Zhang, Xie & Lin 2019*b*, Ramachandran, Parmar, Vaswani, Bello, Levskaya & Shlens 2019) exploring to replace the convolution in CNN models with content-based attention.

4.2.3 Geometric Attention

The standard convolution operation can be disentangled as $k \times k$ position-wise linear transformations followed by spatial summation of the outputs. Following similar formulation but being equipped with more dynamics, the geometric attention adaptively assigns attention weights for each of the linear transformed feature at

position j according to the geometric closeness measured by $\gamma : R^{d_{in}} \mapsto R$:

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x}_i)} \sum_{j \in \Omega_i} \alpha_g(\mathbf{x}_i, j) \odot \beta(x_j) = \frac{1}{\mathcal{C}(\mathbf{x}_i)} \sum_{j \in \Omega_i} e^{\gamma(\mathbf{x}_i, j)} \odot \beta(x_j). \tag{4.5}$$

In this formulation, $k \times k$ linear transformation is reduced to 1×1 for the sake of efficiency. Gaussian distribution is applied, and to make the operation content-aware, we use $\gamma(\mathbf{x}_i, j)$ instead of $\gamma(i, j)$.

There are various implementations of the function γ and they all have strong relations to the positional embedding. In the 2D relative position embeddings (Shaw, Uszkoreit & Vaswani 2018*b*, Ramachandran et al. 2019), the relative distance between $i = (i_{row}, i_{col})$ and $j = (j_{row}, j_{col})$ is factorized across dimensions. The row and column offset embeddings are concatenated to form the embedding $\mathbf{r}_{i-j} = [\mathbf{r}_{row}, \mathbf{r}_{col}]$. Here we use $i - j$ denotes the coordinate shift between position i and j. Thus, $\gamma = \mathbf{x}_i^T \mathbf{r}_{i-j}$. And this formulaiton results in $2k$ embeddings to represent the $k \times k$ positions.

We can also extend the dynamic convolution (Wu, Fan, Baevski, Dauphin & Auli 2019*b*), which is originally designed for sequence modelling, into the two-dimension mode. A linear transformation ($W_\gamma \in R^{d_{in} \times k \times k}$) is applied to predict $k \times k$ position-wise attention weights. If we omit the effect of bias term and view the fully connected layer W_γ as a embedding layer, and slice the j-th row vector $W_\gamma[j, \cdot]$ as the embedding of position j. we obtain $\gamma(\mathbf{x}_i, j) = W_\gamma[j, \cdot]^T \mathbf{x}_i$. Interestingly, if we extend the receptive field into the whole image with a fixed input size of $H \times W$, the 2D dynamic convolution will result in the collection operation of the point-wise spatial attention (Zhao et al. 2018). In summary, we choose the 2D dynamic convolution in our implementation due to the brief concept and efficient implementation. And we predict $k \times k$ attention weights for every subspace.

4.3 Localized Bilateral Attention

Compared with the convolution that is content-agnostic, the content attention has advantages in adaptively aggregating the correlated information according to the similarity in feature space. However, due to the permutation-invariant property, its expressivity in vision tasks still limits by the lack of positional information. Therefore, there is a huge necessity to combine *geometric attention* and *content attention*, together casting as the *bilateral attention*.

4.3.1 Bilateral Combination

As to the fusion of the proposed attentions, the straightforward idea is to concatenate their outputs (Liu, Jiang, Huang & Yang 2019). However, the attention in late fusion fails to make the two attentions affect each other. The appropriate solution is to enforce both geometric and content affinity simultaneously. The combined one, thus bilateral attention, can be described as follows:

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x}_i)} \sum_{j \in \Omega_i} \left(\alpha_c(\mathbf{x}_i, \mathbf{x}_j) \cdot \alpha_g(\mathbf{x}_i, j)\right) \odot \beta(x_j) \tag{4.6}$$

In this way, the unilateral attention weights can modulate the strength of the other one. Due to the fact that the gaussian distribution is widely used in previous methods. The formulation of α can be finally instantiated as:

$$\begin{aligned} \alpha(\mathbf{x}_i, \mathbf{x}_j, i, j) &= \exp(\mathbf{x}_i^T W_\sigma^T W_\phi \mathbf{x}_j) \cdot \exp(W_\gamma[j, \cdot]\mathbf{x}_i) \\ &= \exp\left(\mathbf{x}_i^T W_\sigma^T W_\phi \mathbf{x}_j + W_\gamma[j, \cdot]\mathbf{x}_i\right) \end{aligned} \tag{4.7}$$

with the normalization:

$$\mathcal{C}(\mathbf{x}_i) = \sum_{l \in \Omega_i} \exp\left(\mathbf{x}_i^T W_\sigma^T W_\phi \mathbf{x}_l + W_\gamma[l, \cdot]\mathbf{x}_i\right) \tag{4.8}$$

If we use the projected $\mathbf{v}_i = W_\sigma \mathbf{x}_i$ instead of $\mathbf{x}_i$ to calculate the geometric attention, we obtain:

$$\begin{aligned} \alpha(\mathbf{x}_i, \mathbf{x}_j, i, j) &= \exp\left((W_\phi \mathbf{x}_j)\mathbf{v}_i + W_\gamma[j, \cdot]\mathbf{v}_i\right) \\ &= \exp((W_\phi \mathbf{x}_j + W_\gamma[j, \cdot])\mathbf{v}_i) \end{aligned} \tag{4.9}$$

This (Eq. 4.9) provides another explanation of why the position embeddings () in the Transformer model can be added to the input features directly before inputting into the attention layers, instead of being simply explained as a feature fusion.

4.3.2 Iterative Refinement

The performance of attention-based methods heavily relies on the estimation of coupling coefficients between the key-query pairs. To improve the accuracy of the coefficients, we further consider to refine the initial attention weights based on an iterative manner. In the multi-head attention, the original feature vector is factorized as M heads, and if we treat each of the feature vector $\mathbf{x}_i \in R^{d_{in}/M}$, which consists of d_{in}/M neurons, as a capsule, we can naturally analogy the attention procedure with the routing by agreement algorithm in capsule networks. The essential difference is that our routing procedure happens locally and spatially. Besides, we consider the information aggregation from pixel j to pixel i rather than the diffusion manner in capsule networks. We show the adapted routing algorithm 1 below, and we only perform this procedure in content attention.

Here we consider the iterative refinement procedure for position i and its supporting region Ω_j. Before starting the iterative procedure, the inputs $\mathbf{x}_j$ is transformed by the linear weight matrix W. and then the initial attention logit c_j is calculated. Following the procedure in vanilla attention, we obtain the updated version of feature $\mathbf{x}_i$. To calculate the Δc_j, we first applied *squash activation* $\mathbf{u}_i = \frac{\|\mathbf{x}_i\|^2}{1+\|\mathbf{x}_i\|^2}\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$ over the updated $\mathbf{x}$, and recalculate the inner product between $\mathbf{u}_i$ and $\mathbf{x}_j$. This refinement

Algorithm 1 Iterative refinement algorithm

procedure ITERATIVE REFINEMENT($\mathbf{x}_i$)
  $\mathbf{x}_j \leftarrow W_\phi \mathbf{x}_j; \forall j \in \Omega_i$ ▷ note that $\mathbf{x}_i \in \{\mathbf{x}_j; \forall j \in \Omega_i\}$, and $W_\phi = W_\sigma = W_\gamma$
  $c_j \leftarrow {\mathbf{x}_i}^T \mathbf{x}_j$
  for $t \in [1, 2, ..., T]$ **do**
    $a_j \leftarrow \frac{\exp(c_j)}{\sum_{l \in \Omega_i} \exp(c_l)}; \forall j \in \Omega_i$ ▷ *softmax* normalization
    $\mathbf{x}_i \leftarrow \sum_{j \in \Omega_i} a_j \mathbf{x}_j$ ▷ spatial feature aggregation
    $\mathbf{u}_i \leftarrow squash(\mathbf{x}_i)$ ▷ activate the target feature vector
    $\Delta c_j \leftarrow \mathbf{u}_i^T \mathbf{x}_j$
    $c_j \leftarrow c_j + \Delta c_j$ ▷ update the *content similarity*
  return $\{c_j; \forall j \in \Omega_i\}$

term is treated as if it was a log-likelihood and is added to the initial c_j step by step. After T iterations, we finally obtain the refined attention affinities in content space.

4.3.3 Computational Complexity

In this section, we mainly compare the complexity of the proposed attention modules and the convolution operation based on two metrics: parameter size and FLOPs. Table 4.1 shows the detailed comparison results.

In terms of the number of parameters, we mainly consider the weight terms of linear transformations and omit the bias terms. The complexity of parameter sizes of different operations are listed in the second column (#Params) in Table 4.1. We separately calculate the parameters of the content attention part and the geometric attention part in our proposed BA module for detailed investigation. As observed, the convolution whose parameter size increases along with the quadratic of the kernel size (k). For the content attention part, it is independent from the kernel size; for the geometric attention part, the number of heads (M) is generally much smaller than

the input/output feature dimensions (d_i/d_o). Overall, it can be concluded that the proposed BA module outperforms the convolution in the parameter size as long as the kernel size is larger than the value of 2, which is usual in implementations. When taking into the iterative refinement algorithm (IR) into consideration, the parameter size can be further reduced by 50% compared to the vanilla content attention. This is due to the fact that the IR algorithm enables the parameter-sharing of different heads in the content attention.

Table 4.1 also shows the complexity of FLOPs for different operations in the 3th∼6th columns. For clear illustration, we decompose the attention/convolution operation as linear transformation and spatial summation, respectively. The spatial summation is further divided into the attention weight generation part ('Affinity' in the table) and the weighted summation part ('Summation' in the table). As observed, the convolution has large FLOPs in the linear transformation part, while the attention operations have more FLOPs in the spatial summation part. Considering the usual values for the parameter k, d_i and d_o, the FLOPs of the linear transformation part will dominate the computation of the overall FLOPs of each operation. In this situation, the attention modules also enjoys advantages in the FLOPs compared to the convolution operation, when the kernel size is larger than 2. Besides, another important discovery is that the IR algorithm leads to a further reduction of the FLOPs of the vanilla content attention, contributed by the decreasing of the FLOPs in the linear transformation part. It is notable that the efficiency of the IR algorithm comes at expenses, e.g., additional memory requirement for intermediate variables, which may hinder its application to large-sized feature maps.

4.4 Experiments

To verify the effectiveness of the proposed bilateral attention, we experiment with two classical encoder-decoder architectures: deeplabv3+ (Chen, Zhu, Papan-

Table 4.1 : Complexity analysis for the convolution operation and proposed attention operations. "IR" represents Iterative Refinement algorithm, d_i and d_o represents the input/output dimension of the feature, respectively. M denotes the number of heads in attention. Our analysis is based on the fact that $k \ll d_i$ and $k \ll d_o$, and some inessential terms are omitted.

Methods	#Params	#FLOPs			
		Linear Transformation	Spatial Summation		
			Affinity	Summation	Total
Convolution	$O(k^2 \cdot d_i \cdot d_o)$	$O(k^2 \cdot d_i d_o)$	-	$O(d_o \cdot k^2)$	$O(d_o \cdot k^2)$
Content Attention	$O(4 \cdot d_i \cdot d_o)$	$O(4 \cdot d_i d_o)$	$O(2d_o \cdot k^2)$	$O(d_o \cdot 2k^2)$	$O(4d_o \cdot k^2)$
Geometric Attention	$O(k^2 \cdot d_i \cdot M)$	$O(d_i d_o)$	$O(M \cdot 2d_o \cdot k^2)$	$O(d_o \cdot 2k^2)$	$O(2(M+1)d_o \cdot k^2)$
Content Attention + IR	$O(2 \cdot d_i \cdot d_o)$	$O(2 \cdot d_i d_o)$	$O(3(2d_o \cdot k^2))$	$O(3d_o \cdot 2k^2)$	$O(12d_o \cdot k^2)$

dreou, Schroff & Adam 2018) and U-Net (Ronneberger et al. 2015), by replacing the convention convolution layers in the upsampling blocks of their decoder parts. Deeplabv3+ contains only one upsampling-fusion block in its decoder, which is ideal to diagnose the effects of a single bilateral attention module. We start from the experiments with Deeplabv3+ on the challenging Cityscapes (Cordts et al. 2016) dataset, which contains elaborate annotations for large-resolution images. With the initial observations of single-layer BA, we continue the experiment with U-Net–which contains over four consecutive upsampling-fusion blocks – to verify the effects of multi-layer BA. Experiments with U-Net are conducted on another two fine-grained segmentation datasets for face-parsing: HELEN (Smith, Zhang, Brandt, Lin & Yang 2013) and LFW-FP (Kae, Sohn, Lee & Learned-Miller 2013).

4.4.1 Experiments on DeepLabv3+

The Cityscapes dataset is a large-scale urban-scene dataset, which contains 5K images and high-quality pixel-level annotations for 19 semantic labels. The image set is split into 2,975, 500 and 1,525 for training, validation and test, respectively.

Following the settings in previous works, we employ the SGD optimizer with the momentum of 0.9. The learning rate is updated by the polynomial learning rate policy (Chen, Zhu, Papandreou, Schroff & Adam 2018) and initialized as 1e-2. Typical data augmentation techniques for image segmentation are used, including random horizontally flipping, random scaling in the range of [0.5,2], and random cropping as 769×769. In all the ablation experiments, we train the models for 40K iterations with a batch size of 8 images. In the testing stage, the whole-image single-scale inference strategy is applied. Mean IOU is reported as the evaluation metric.

Our method is implemented based on Pytorch toolbox (Paszke, Gross, Massa, Lerer, Bradbury, Chanan, Killeen, Lin, Gimelshein, Antiga et al. 2019). The ImageNet (Krizhevsky, Sutskever & Hinton 2012) pre-trained ResNet-101 (He et al. 2016) is adopted as the backbone for deeplabv3+ model. We remove the last two downsampling layers and employ dilated convolutions in the subsequent convolutional layers following previous works (Chen, Zhu, Papandreou, Schroff & Adam 2018), where the output stride becomes 8. Different with the original deeplabv3+, we only keep one 3x3 convolution block in the upsampling-fusion blocks of the decoder. This aims to investigate the effect of a single bilateral attention module.

We conduct detailed ablation study via removing or altering each component independently on the Cityscapes dataset. Note that both the training and testing procedures of each ablation experiment are kept exactly the same for a fair comparison. The experiment results are shown in Table 4.2. Several conclusions could be drawn:

The importance of bilateral combination. The Comparison between a and b shows the relation in content space is more important than the geometric affinities for semantic segmentation. Although the content attention and the geometric attention operation cannot compete with the deeplabv3+ (j) when they are considered solely, their joint effects, i.e., the bilateral attention operation successfully surpasses the

Table 4.2 : Ablation study on the Cityscapes dataset of the proposed bilateral attention module with mean IOU.

ID	Content	Geometric	#Heads	#Iterations	Share	#Params (K)	#FLOPS (G)	Mean IOU (%)
a	√		8	1		265.22	2.45	78.8
b		√	8	1		217.56	2.01	77.3
c	√	√	1	1		267.55	2.48	79.08
d	√	√	4	1		274.54	2.54	79.16
e	√	√	8	1		283.86	2.63	79.72
f	√	√	16	1		302.51	2.80	79.41
g	√		8	3	√	1.3184	1.22	79.08
h	√	√	8	1	√	150.27	1.39	79.21
i	√	√	8	3	√	150.27	1.71	79.90
j	DeepLabv3 Plus (Baseline)					590.08	5.44	79.14

deeplabv3+ on the Mean IOU, even with much lower computational complexity (in parameters and FLOPs) (comparing *a*, *b*, *e* vs *j*). As observed, the bilateral attention reduces the number of parameter by 50%, and the FLOPS by more than 40%.

The importance of iterative refinement. With more accurate attention affinities, the iterative refinement scheme is helpful in further boosting the segmentation precision, no matter which component it is grounded on (comparing *a* vs *g*, *h* vs *i*). In the iteration process, the feature of query, key and value should come from the same feature space to maintain a stable feature update. Without the refinement procedure, simply sharing the parameters of σ, ϕ and γ will result in collapse of the performance (comparing *h* and *i*). The refinement procedure can eliminate the negative effect of weight-sharing. The weight-sharing reduce half of the parameters in linear transformation as well as the computation load of FLOPS. Therefore, although the iteration procedure brings in extra processing steps, the overall MACs still reduce thanks to the weight-sharing design.

The effects of number of heads. The number of attention heads is a key parameter that influences the effects of the proposed bilateral attention operation. Increasing of heads in a certain range can continuously provide improvement, while too many of heads will hurt the performances (comparing $c \sim f$). This maybe the reason that the number of neurons in a subspace will reduce as the number of heads increases, and too few neurons will defeat the representation of the feature in subspace. Besides, the complexity of content attention does not increase with the number of heads, while the geometric attention operation does. This leads to the increment of parameters with the number of heads.

Table 4.3 : Face parsing results on the HELEN dataset with class-wise F1-score and overall accuracy.

	Skin	Nose	Upper-lip	Inner-mouth	Lower-lip	Brows	Eyes	Mouth	Overall
Smith *et al.* (Smith et al. 2013)	88.2	92.2	65.1	71.3	70.0	72.2	78.5	85.7	80.4
Liu *et al.* (Liu, Shi, Liang & Yang 2017)	92.1	93.0	74.3	89.1	81.7	77.0	86.8	89.1	88.6
Wei *et al.* (Wei, Sun, Wang, Lai & Liu 2017)	91.5	93.7	-	-	-	78.6	84.7	91.5	90.2
Lin *et al.* (Lin, Yang, Chen, Zeng, Wen & Yuan 2019)	94.5	**95.6**	79.6	86.7	**89.8**	**83.1**	**89.6**	**95.0**	**92.4**
U-Net	94.8	94.3	80.0	87.9	87.9	82.7	89.1	94.0	91.5
U-Net-BA	**94.9**	94.4	**80.8**	**88.0**	88.3	82.9	89.3	94.3	91.8
U-Net (w/o alignment)	94.4	93.7	80.2	89.3	90.2	81.5	88.0	94.6	91.0
U-Net-BA (w/o alignment)	94.7	93.7	82.4	88.8	90.7	82.1	89.9	95.4	91.6

4.4.2 Experiments on U-Net

In this experiment we validate the proposed bilateral operation with U-Net on two face parsing datasets: HELEN (Smith et al. 2013) and LFW-PL (Kae et al. 2013). The HELEN dataset contains 2,330 face images. Each image is annotated with 11 labels: "background", "facial skin", "left/right brow", "left/right eye", "nose", "upper/lower lip", "inner mouth" and "hair". We adopt the same dataset division setting as in (Lin et al. 2019) that uses 2,000, 230 and 100 images for training, validation and testing, respectively. The LFW-PL dataset contains 2,972 face images, which are

manually annotated with three labels: “skin”, “hair” and “background”. We use 1,500, 500, and 927 images are split for training, validation and testing following previous works (Kae et al. 2013, Zhou, Liu & He 2017).

For both of the Helen and LFW-PL dataset, the same network configurations are adopted. Specifically, Following the previous settings (Lin et al. 2019), the ImageNet pre-trained ResNet-18 is adopted as the backbone model for U-Net, whose last two down-sampling layers are removed. Corresponding to the four residual blocks in ResNet-18, the U-Net contains four consecutive upsampling-fusion modules. To train the U-Net with the proposed bilateral attention (U-Net-BA), we replace the 3x3 convolutions in the decoder of U-Net with the proposed bilateral attention module. The iterative refinement scheme is only applied to the first BA layer. Please refer the supplementary materials for detailed network configuration.

We use Adam as the optimizer, weight decay of 1e-5 and batch size of 12. Poly learning rate policy is used with an initial learning rate of 3e-4. All the models are trained for 100 epoch, and the learning rate warming-up is applied to the training of U-Net-BA. Similar to the previous works, face alignment is implemented as a pre-processing step on the Helen dataset. The face alignment places a strong prior to the layout of the image, which in advance alleviates the effects of some challenging effects on the performances, *e.g.*, large pose, rotation. Although it helped improve the performances, the robustness of a model cannot be fully revealed. To demonstrate the robustness of the proposed method, we also train another U-Net and U-Net-BA model on HELEN dataset without the pre-processing of face alignment, respectively. F1-score is adopted as the quantitative evaluation metric as previous works.

The comparison results on HELEN dataset are shown in Table 4.3. Each column in the table column shows the F1-score percentage corresponding to a specific face label, respectively. The last column uses the union of all the inner facial components

Figure 4.3 : Qualitative comparison on Helen dataset. The first and the second row shows parsing results on face images with and without the pre-processing of face alignment, respectively. For each image pair, the left and side shows the result of the U-Net-BA model and the original U-Net, respectively.

Table 4.4 : Face parsing results on the LFW-PL dataset with class-wise F1-score and overall accuracy.

Methods	Skin	Hair	Background	Accuracy
Zhou *et al.* (Zhou, Liu & He 2017)	94.10	85.16	96.46	95.28
Liu *et al.* (Liu et al. 2017)	97.55	83.43	94.37	95.46
Lin *et al.* (Lin et al. 2019)	95.77	88.31	98.26	96.71
U-Net	95.83	89.36	98.43	96.95
U-Net-BA	95.90	89.71	98.46	97.02

(eyes/brows/nose/mouth) as labels. It can be observed that the proposed method surpasses the base model of U-Net and also outperforms most of the previous methods in terms of the overall performances, which demonstrates the effectiveness of the bilateral attention. Recently, Lin (Lin et al. 2019) achieve the best performances on HELEN dataset. They have leveraged the Mask-RCNN (He, Gkioxari, Dollár & Girshick 2017) branch, which benefits the segmentation results with the detection of inner facial components. In spite of this, our U-Net-BA model still achieve the top accuracy in some fine-grained classes, i.e., "upper-lip", "inner-mouth" and "skin". Comparing the performance of U-Net and U-Net-BA on the unaligned images,

we can observe that the U-Net suffers from severe degeneration in performances while the performance of proposed U-Net-BA only slightly decreased. This further demonstrates the robustness of the proposed bilateral attention compared to the convention convolution operations.

Comparison results with the U-Net on LFW-PL dataset are shown in Table 4.4. The F1-score percentages corresponding to skin, hair and background are reported in each column, respectively, with the overall accuracy in the last column. As observed, U-Net-BA outperforms the baseline model of U-Net on all the metrics, and also achieves the best performance on the LFW-PL dataset compared to other state-of-the-art methods.

4.4.3 Detailed Configuration for U-Net

We adapt the open-source implementation of U-Net with the backbone of Resnet-18 and replace all the 3x3 convolution layers in the upsampling-fusion block of the decoding stage with our proposed bilateral attention (BA) modules. Detailed configurations of the BA module in the U-Net (U-Net-BA) can be found in Table 4.5. Take the first upsampling-fusion block (dubbed as up-1) for example, the output of res-5 is first upsampled by two times, concatenated with the output from res-4, and then fused by a 3×3 BA module. The BA in up-1 has eight heads, with the iterative refinement (IR) algorithm applied.

Table 4.5 : The detailed configuration for the decoder in the U-Net-BA. "res-i" denotes the output of the i-th stage in the Resent-18. "up-i" represents the i-th upsampling-fusion blocks in the decoder.

Module ID	Input Features (#channels)		Output Features (#channels)	#Head	Kernel Size	IR
	Low-level	High-level				
up-1	res-4 (256)	res-5 (512)	up-1 (512)	8	3×3	$\surd$
up-2	res-3 (128)	up-1 (512)	up-2 (256)	8	3×3	$\times$
up-3	res-2 (64)	up-2 (256)	up-3 (256)	4	3×3	$\times$
up-4	res-1 (64)	up-3 (256)	up-4 (128)	4	3×3	$\times$

4.5 Conclusion

The general encoder-decoder architecture encompasses two important problems: over-smooth and feature misalignment. To tackle these challenges, in this work we propose a general local bilateral attention module as an alternative to the convention convolution layers in the decoder part. An iterative refinement scheme is further proposed to generate more accurate attention affinities. The bilateral attention jointly leverages the appearance and geometric information to infer the attention weights, which are then used to complete the transformation of the input feature map. Competitive experimental results on two datasets with two popular encoder-decoder architectures demonstrate the efficacy of the bilateral attention, with general improved performances and drastically decreased computational complexity. In the future, we expect to explore the application of bilateral attention in other vision tasks such as image generation model or super-resolution.

Chapter 5

Future Directions

This dissertation focuses on the attention mechanism. The first part proposes an entangled attention mechanism to enable the transformer framework to explore multimodal information. According to the bilateral formulation, the second part elaborates on how to enable the self-attention mechanism with positional awareness. The third part explores the possibility of employing a local attention operator as a replacement for the convolution in semantic segmentation, and the designed operator demonstrates its inclination in the fusion of multi-level features. There are a few research directions that the author hopes to take in the future to continue the research projects presented in this dissertation.

Interpretability. Interpretability of deep model is a critical problem in various applications, e.g., medical image, self-driving, dialogue system, et al. Since attention layers explicitly weight the importance of input features by model the similarity between query and keys, it is believed that attention could be interpreted to identify the elements that models found useful. Most of the papers on the attention mechanism illustrate their interpretability by visualization the attention maps. However, qualitative verification is far from revealing the mystery behind the attention mechanism and its explainability. More investigations are required to explore the correlation or causality between attention weights and model predictions.

Attention-based framework for Computer Vision. Convolutions are a fundamental building block of modern computer vision systems. Still, the attention mechanism has advantages in two aspects. First, the attention-based operator enjoys

adaptiveness in aggregation local context, as shown in Chapter IV. Moreover, nonlocal operations based on attention are proven to be effective at capturing long-range dependency. Inspired by the numerous success that the transformer has made in NLP, a purely attention-based model may revolutionize the computer vision frameworks.

Efficient Attention Mechanism. Current state-of-the-art attention-based models are computation-intensive; the quadratic time and space complexity of self-attention prevent it from applying to many real-time applications. This disadvantage is especially notable when applying the attention mechanism to computer vision. Therefore, the acceleration of the attention mechanism tends to convey it into broader applications. The design of efficient attention is meaningful.

Bibliography

Anderson, P., Fernando, B., Johnson, M. & Gould, S. (2016), Spice: Semantic propositional image caption evaluation, *in* 'ECCV', Springer.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. & Zhang, L. (2018), Bottom-up and top-down attention for image captioning and visual question answering, *in* 'ICCV'.

Aneja, J., Deshpande, A. & Schwing, A. G. (2018), Convolutional image captioning, *in* 'CVPR'.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C. & Parikh, D. (2015), Vqa: Visual question answering, *in* 'ICCV'.

Atwood, J. & Towsley, D. (2016), Diffusion-convolutional neural networks, *in* 'NeurIPS', pp. 1993–2001.

Bello, I., Zoph, B., Vaswani, A., Shlens, J. & Le, Q. V. (2019), Attention augmented convolutional networks, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 3286–3295.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. & Zagoruyko, S. (2020), 'End-to-end object detection with transformers', *arXiv preprint arXiv:2005.12872* .

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2017), 'Deeplab: Semantic image segmentation with deep convolutional nets, atrous

convolution, and fully connected crfs', *IEEE transactions on pattern analysis and machine intelligence* **40**(4), 834–848.

Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. (2017), 'Rethinking atrous convolution for semantic image segmentation', *arXiv preprint arXiv:1706.05587* .

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. (2018), Encoder-decoder with atrous separable convolution for semantic image segmentation, *in* 'Proceedings of the European conference on computer vision (ECCV)', pp. 801–818.

Chen, M., Li, Y., Zhang, Z. & Huang, S. (2018), Tvt: Two-view transformer network for video captioning, *in* 'ACML'.

Cho, K., Gulcehre, B. v. M. C., Bahdanau, D., Schwenk, F. B. H. & Bengio, Y. (2014), 'Learning phrase representations using rnn encoder–decoder for statistical machine translation'.

Cooper, R. M. (1974), 'The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing.', *Cognitive Psychology* .

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. & Schiele, B. (2016), The cityscapes dataset for semantic urban scene understanding, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3213–3223.

Dauphin, Y. N., Fan, A., Auli, M. & Grangier, D. (2017), Language modeling with gated convolutional networks, *in* 'ICML'.

Dauphin, Y. N. & Grangier, D. (2016), Predicting distributions with linearizing belief networks, *in* 'ICLR'.

Denkowski, M. & Lavie, A. (2014), Meteor universal: Language specific translation evaluation for any target language, *in* 'SMT-W', pp. 376–380.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018), 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805* .

Elman, J. L. (1990), 'Finding structure in time', *Cognitive science* **14**(2), 179–211.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C. et al. (2015), From captions to visual concepts and back, *in* 'CVPR'.

Fonollosa, J. A., Casas, N. & Costa-jussà, M. R. (2019), 'Joint source-target self attention with locality constraints', *arXiv preprint arXiv:1905.06596* .

Fritsch, J., Kuehnl, T. & Geiger, A. (2013), A new performance measure and evaluation benchmark for road detection algorithms, *in* '16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)', IEEE, pp. 1693–1700.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z. & Lu, H. (2019), Dual attention network for scene segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3146–3154.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T. & Rohrbach, M. (2016), 'Multimodal compact bilinear pooling for visual question answering and visual grounding', *arXiv preprint arXiv:1606.01847* .

Gehring, J., Auli, M., Grangier, D., Yarats, D. & Dauphin, Y. N. (2017), Convolutional sequence to sequence learning, *in* 'ICML'.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. (2017), Neural message passing for quantum chemistry, *in* 'ICML', pp. 1263–1272.

Gu, J., Wang, G., Cai, J. & Chen, T. (2017), An empirical study of language cnn for image captioning, *in* 'ICCV'.

Gu, S., Bao, J., Yang, H., Chen, D., Wen, F. & Yuan, L. (2019), Mask-guided portrait editing with conditional gans, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3436–3445.

He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2017), Mask r-cnn, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 2961–2969.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'CVPR'.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* .

Hu, H., Gu, J., Zhang, Z., Dai, J. & Wei, Y. (2018), Relation networks for object detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3588–3597.

Hu, H., Zhang, Z., Xie, Z. & Lin, S. (2019*a*), Local relation networks for image recognition, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 3464–3473.

Hu, H., Zhang, Z., Xie, Z. & Lin, S. (2019*b*), Local relation networks for image recognition, *in* 'The IEEE International Conference on Computer Vision (ICCV)'.

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y. & Liu, W. (2019), Ccnet: Criss-cross attention for semantic segmentation, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 603–612.

Kae, A., Sohn, K., Lee, H. & Learned-Miller, E. (2013), Augmenting crfs with boltzmann machine shape priors for image labeling, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2019–2026.

Karpathy, A. & Fei-Fei, L. (2015), Deep visual-semantic alignments for generating image descriptions, *in* 'CVPR'.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A. et al. (2017), 'Visual genome: Connecting language and vision using crowdsourced dense image annotations', *IJCV* .

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'Advances in neural information processing systems', pp. 1097–1105.

Le, Q. V., Jaitly, N. & Hinton, G. E. (2015), 'A simple way to initialize recurrent networks of rectified linear units', *CoRR* **abs/1504.00941**.

Lee, K.-H., Chen, X., Hua, G., Hu, H. & He, X. (2018), Stacked cross attention for image-text matching, *in* 'ECCV'.

Li, G., Zhu, L., Liu, P. & Yang, Y. (2019), Entangled transformer for image captioning, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 8928–8937.

Li, N. & Chen, Z. (2018), Image cationing with visual-semantic lstm, *in* 'IJCAI-18'.

Lin, C.-Y. (2004), 'Rouge: A package for automatic evaluation of summaries', *Text Summarization Branches Out* .

Lin, J., Yang, H., Chen, D., Zeng, M., Wen, F. & Yuan, L. (2019), Face parsing with roi tanh-warping, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5654–5663.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014), Microsoft coco: Common objects in context, *in* 'ECCV', Springer.

Lioutas, V. & Guo, Y. (2020), 'Time-aware large kernel convolutions', *ICML* .

Liu, H., Jiang, B., Huang, W. & Yang, C. (2019), 'One-stage inpainting with bilateral attention and pyramid filling block', *arXiv preprint arXiv:1912.08642* .

Liu, S., Shi, J., Liang, J. & Yang, M.-H. (2017), 'Face parsing via recurrent propagation', *arXiv preprint arXiv:1708.01936* .

Long, J., Shelhamer, E. & Darrell, T. (2015), Fully convolutional networks for semantic segmentation, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3431–3440.

Lu, J., Xiong, C., Parikh, D. & Socher, R. (2018), Knowing when to look: Adaptive attention via a visual sentinel for image captioning, *in* 'ICCV'.

Lu, J., Yang, J., Batra, D. & Parikh, D. (2016), Hierarchical question-image co-attention for visual question answering, *in* 'Neurips'.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J. & Khudanpur, S. (2010), Recurrent neural network based language model, *in* 'INTERSPEECH'.

Niepert, M., Ahmed, M. & Kutzkov, K. (2016), Learning convolutional neural networks for graphs, *in* 'ICML', pp. 2014–2023.

Ott, M., Edunov, S., Grangier, D. & Auli, M. (2018), Scaling neural machine translation, *in* 'Proceedings of the Third Conference on Machine Translation: Research Papers', pp. 1–9.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002), Bleu: a method for automatic evaluation of machine translation, *in* 'ACL', Association for Computational Linguistics.

Parmar, N., Ramachandran, P., Vaswani, A., Bello, I., Levskaya, A. & Shlens, J. (2019), Stand-alone self-attention in vision models, *in* H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox & R. Garnett, eds, 'Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada', pp. 68–80.
URL: *http://papers.nips.cc/paper/8302-stand-alone-self-attention-in-vision-models*

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. (2019), Pytorch: An imperative style, high-performance deep learning library, *in* 'Advances in Neural Information Processing Systems', pp. 8024–8035.

Pedersoli, M., Lucas, T., Schmid, C. & Verbeek, J. (2017), Areas of attention for image captioning, *in* 'ICCV'.

Peng, C., Zhang, X., Yu, G., Luo, G. & Sun, J. (2017), Large kernel matters—improve semantic segmentation by global convolutional network, *in* 'CVPR', pp. 1743–1751.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A. & Shlens, J. (2019), 'Stand-alone self-attention in vision models', *arXiv preprint arXiv:1906.05909* .

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J. & Goel, V. (n.d.), Self-critical sequence training for image captioning, *in* 'CVPR', pp. 7008–7024.

Ronneberger, O., Fischer, P. & Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* 'International Conference on Medical image computing and computer-assisted intervention', Springer, pp. 234–241.

Sennrich, R., Haddow, B. & Birch, A. (2016), Neural machine translation of rare words with subword units, *in* 'Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 1715–1725.

Sharma, P., Ding, N., Goodman, S. & Soricut, R. (2018), Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, *in* 'ACL'.

Shaw, P., Uszkoreit, J. & Vaswani, A. (2018*a*), Self-attention with relative position representations, *in* M. A. Walker, H. Ji & A. Stent, eds, 'Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)', Association for Computational Linguistics, pp. 464–468.
URL: *https://doi.org/10.18653/v1/n18-2074*

Shaw, P., Uszkoreit, J. & Vaswani, A. (2018*b*), 'Self-attention with relative position representations', *arXiv preprint arXiv:1803.02155* .

Simonyan, K. & Zisserman, A. (2015), 'Very deep convolutional networks for large-scale image recognition', *ICLR* .

Smith, B. M., Zhang, L., Brandt, J., Lin, Z. & Yang, J. (2013), Exemplar-based face parsing, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3484–3491.

Sukhbaatar, S., Weston, J., Fergus, R. et al. (2015), End-to-end memory networks, *in* 'Neurips'.

Sutskever, I., Martens, J. & Hinton, G. E. (2011), Generating text with recurrent neural networks, *in* 'ICML-11', pp. 1017–1024.

Sutskever, I., Vinyals, O. & Le, Q. V. (2014), Sequence to sequence learning with neural networks, *in* 'Advances in neural information processing systems', pp. 3104–3112.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. C. (1995), 'Integration of visual and linguistic information in spoken language comprehension', *Science* .

Tang, K., Zhang, H., Wu, B., Luo, W. & Liu, W. (2019), Learning to compose dynamic tree structures for visual contexts, *in* 'CVPR'.

Tomasi, C. & Manduchi, R. (1998), Bilateral filtering for gray and color images, *in* 'Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)', IEEE, pp. 839–846.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), Attention is all you need, *in* 'Neurips'.

Vedantam, R., Lawrence Zitnick, C. & Parikh, D. (2015), Cider: Consensus-based image description evaluation, *in* 'CVPR', pp. 4566–4575.

Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. (2015), Show and tell: A neural image caption generator, *in* 'CVPR'.

Wang, X., Girshick, R., Gupta, A. & He, K. (2018), Non-local neural networks, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 7794–7803.

Wei, Z., Sun, Y., Wang, J., Lai, H. & Liu, S. (2017), Learning adaptive receptive fields for deep image parsing network, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2434–2442.

Wikipedia (2020), 'Attention — Wikipedia, the free encyclopedia'. [Online; accessed 06-December-2020].
URL: *https://en.wikipedia.org/wiki/Attention*

Wu, F., Fan, A., Baevski, A., Dauphin, Y. N. & Auli, M. (2019*a*), Pay less attention with lightweight and dynamic convolutions, *in* '7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019', OpenReview.net.
URL: *https://openreview.net/forum?id=SkVhlh09tX*

Wu, F., Fan, A., Baevski, A., Dauphin, Y. N. & Auli, M. (2019*b*), 'Pay less attention with lightweight and dynamic convolutions', *arXiv preprint arXiv:1901.10430* .

Wu, Y., Zhu, L., Jiang, L. & Yang, Y. (2018), Decoupled novel object captioner, *in* 'ACM MM', ACM.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. & Bengio, Y. (2015), Show, attend and tell: Neural image caption generation with visual attention, *in* 'ICML'.

Yao, T., Pan, Y., Li, Y. & Mei, T. (2018), Exploring visual relationship for image captioning, *in* 'ECCV'.

Yao, T., Pan, Y., Li, Y., Qiu, Z. & Mei, T. (2017), Boosting image captioning with attributes, *in* 'ICCV'.

You, Q., Jin, H., Wang, Z., Fang, C. & Luo, J. (2016), Image captioning with semantic attention, *in* 'CVPR'.

Yu, F. & Koltun, V. (2016), Multi-scale context aggregation by dilated convolutions, *in* Y. Bengio & Y. LeCun, eds, '4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings'.
URL: *http://arxiv.org/abs/1511.07122*

Zhao, H., Zhang, Y., Liu, S., Shi, J., Change Loy, C., Lin, D. & Jia, J. (2018), Psanet: Point-wise spatial attention network for scene parsing, *in* 'Proceedings of the European Conference on Computer Vision (ECCV)', pp. 267–283.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A. & Torralba, A. (2017), Scene parsing through ade20k dataset, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 633–641.

Zhou, L., Liu, Z. & He, X. (2017), 'Face parsing via a fully-convolutional continuous crf neural network', *arXiv preprint arXiv:1708.03736* .

Zhou, L., Zhou, Y., Corso, J. J., Socher, R. & Xiong, C. (2018), End-to-end dense video captioning with masked transformer, *in* 'CVPR'.

Zhu, L., Xu, Z., Yang, Y. & Hauptmann, A. G. (n.d.), 'Uncovering the temporal context for video question answering', *IJCV* .

Zhu, Z., Xu, M., Bai, S., Huang, T. & Bai, X. (2019), Asymmetric non-local neural networks for semantic segmentation, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 593–602.