

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.

Production Note:

Signature: 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

Certificate	ii
Abstract	iii
Acknowledgments	v
List of Figures	ix
List of Tables	xii
1 Introduction	1
1.1 Background	1
1.2 Research Contribution	4
1.3 Thesis Organization	6
2 Entangled Transformer for Image Captioning	7
2.1 Introduction	7
2.2 Related Work	10
2.3 Preliminary	11
2.4 Methodology	13
2.4.1 Dual-Way Encoder	13
2.4.2 Multimodal Decoder	15
2.4.3 EnTangled Attention	16
2.4.4 Gated Bilateral Controller	17
2.5 Experiments	19

2.5.1 Datasets and Evaluation	19
2.5.2 Implementation Details	20
2.5.3 Comparison with State-of-the-Art Methods	21
2.5.4 Ablation Study	24
2.6 Conclusion	27

3 Bilateral Attention: Rethinking the Positional Awareness

in Self-Attention	28
3.1 Introduction	28
3.2 Related Work	30
3.3 Preliminary: Bilateral Filter	32
3.4 Bilateral Attention	33
3.4.1 General Formulation	33
3.4.2 Formulation of Distance Functions	37
3.4.3 Distance Function in 2-D Mode	39
3.4.4 Instantiations	40
3.5 Experiments for 2-D Bilateral Attention	42
3.5.1 Comparison Results	44
3.5.2 Comparison to Relative Position Encoding	44
3.5.3 Controlled Experiments	45
3.5.4 Visualization of Attention Maps	48
3.6 Experiments For 1-D Bilateral Attention	49
3.6.1 Comparison to the State-of-the-art Methods	49
3.6.2 Controlled Experiments	50
3.7 Conclusion	52

4 Localized Bilateral Attention with Iterative Refinement	
for Image Segmentation	53
4.1 Introduction	53
4.2 Background	55
4.2.1 The General Formulation	55
4.2.2 Content Attention	58
4.2.3 Geometric Attention	58
4.3 Localized Bilateral Attention	60
4.3.1 Bilateral Combination	60
4.3.2 Iterative Refinement	61
4.3.3 Computational Complexity	62
4.4 Experiments	63
4.4.1 Experiments on DeepLabv3+	64
4.4.2 Experiments on U-Net	67
4.4.3 Detailed Configuration for U-Net	70
4.5 Conclusion	71
5 Future Directions	72
Bibliography	74

List of Figures

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.	2
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.	8
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.	12
2.3	The multimodal representations are first fed into ETA to conduct EnTangled Attention, then to GBC to obtain the final representation.	16

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.	24
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.	29
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.	34
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.	39
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.	47

- 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 55
- 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. 57
- 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. 69

List of Tables

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 (%).	19
2.2	The results on single modality. The ETA denotes the ETA-Transformer. Subscript indicates that the visual modality or semantic modality is applied.	20
2.3	MSCOCO Online Evaluation. The ETA denotes the ETA-Transformer. cX means evaluation on X captions. All values are reported as percentage (%).	21
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.	22
2.5	Ablation experiments. ETA is denotes the ETA-Transformer. And all results are trained on sequence-level criterion.	25
2.6	The effect of activation functions in GBC. All results are reported in token-level training.	26
3.1	Comparison results on Cityscapes and ADE20K, and multiple-scale is applied for testing.	44
3.2	Comparison of bilateral attention with relative position encoding (RPE) on Cityscapes.	45

3.3	Ablation experiments on Cityscapes.	46
3.4	The controlled comparison of different kernel size of the bilateral criss-cross attention.	48
3.5	Translation quality evaluation (BLEU scores).	50
3.6	Ablation experiments on IWSLT'14	51
3.7	The effect of the kernel size in geometric attention	51
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.	64
4.2	Ablation study on the Cityscapes dataset of the proposed bilateral attention module with mean IOU.	66
4.3	Face parsing results on the HELEN dataset with class-wise F1-score and overall accuracy.	67
4.4	Face parsing results on the LFW-PL dataset with class-wise F1-score and overall accuracy.	69
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 Resnet-18. "up- i " represents the i -th upsampling-fusion blocks in the decoder.	71