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