UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Exploring Visual Attention Mechanism for Scene Understanding in Image Captioning

by Zongjian Zhang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of A/Prof. Qiang Wu,
A/Prof. Yang Wang,
Prof. Fang Chen

University of Technology Sydney
Faculty of Engineering and Information Technology

September 2020

Certificate of Original Authorship

I, **Zongjian Zhang** declare that this thesis, is submitted in fulfilment of the requirements for the award of **Doctor of Philosophy**, in the **School of Electrical and Data Engineering/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: 28 Sep 2020

ABSTRACT

Exploring Visual Attention Mechanism for Scene Understanding in Image Captioning

by

Zongjian Zhang

Supervisor: A/Prof. Qiang Wu

Co-supervisor: Prof. Fang CHEN, A/Prof. Yang WANG

Scene understanding is a high-level computer vision research task that requires multiple fundamental vision tasks on different visual elements. Image captioning is a typical scene understanding task that understands salient visual contents in a real-world scene of an image and then automatically describes its understandings via a natural language sentence. This thesis concentrates on exploring visual attention mechanism for the scene understanding in image captioning, aiming to achieve a comprehensive multimodal and transparent scene understanding ability. Specifically, four problems are studied to enhance the visual attention mechanism to fine-grained spatial level, to a comprehensive semantic level, and to a high-level ability of attending to visual relationships for interaction words.

Firstly, this thesis proposes a fine-grained and semantic-guided visual region attention model based on a novel Fully Convolutional Network (FCN)-Long Short Term Memory (LSTM) framework. It can attend to both object and "stuff" regions at a fine-grained grid-wise resolution and only focuses on the principal object information in each grid cell. In addition, grid-wise semantic labels are introduced to provide semantic guidance to ensure that related visual regions in different grid cells are correlated to each other. Moreover, an additional semantic context can be summarized from these textual semantic labels.

Secondly, this thesis proposes a novel high-resolution FCN encoder is used with

a residual attention structure to achieve a high-resolution attention supporting four high fine-grained resolutions (i.e., 27×27, 40×40, 60×60, 80×80). A size-invariant "attention correctness" metric is further proposed for evaluating the attention accuracy under different resolutions. Based on the COCO-Stuff dataset, pixel-level evaluations are conducted on both object and "stuff" regions to analyse the performances of high-resolution fine-grained visual region attention.

Thirdly, this thesis concentrates on visual relationship attention exploring visual relationships between object regions under spatial constraints based on the parallel attention mechanism. The goal is to fully explore the relationships/interactions between visual regions for achieving an accurate description of interaction words in the caption. Moreover, it is trained implicitly through an unsupervised approach without using any explicit visual relationship annotations.

Last but not the least, this thesis takes a further step to achieve an adaptive attention module that can perform the role of both visual region attention or visual relationship attention adaptively according to the needs of language decoder. The dynamic linguistic context of the language decoder is fully leveraged for exploring and attending to related visual relationships for interactive words.

Acknowledgements

First of all, I would like to express my deepest gratitude to my principal supervisor, A/Prof. Qiang Wu. Without his guidance, this thesis would not have been accomplished with high quality, and I would not have achieved qualified PhD research outcomes. With abundant knowledge and the insightful and rigorous way of thinking about research problems, he patiently guided me through many research challenges in this valuable journey of PhD study. Besides this degree itself, I've learned the methodology of exporing and solving a new problem, which I believe is more important for my future work and life. Moreover, he is also a great friend that cares about my personal life and offers helps when I face problems. I feel very lucky to finish my PhD under his supervision and can't imagine a better supervisor. Deeply appreciate it!

I am also profoundly grateful to my co-supervisors Prof. Fang Chen and A/Prof. Yang Wang for hosting me as a visiting student in Data 61 (formerly as NICTA) and providing great supports. Thanks to these, I can further extend my knowledge and techniques about data analytics, machine learning, and industrial projects. I believe these are the most valuable experiences for my future work career.

I would also like to express my sincere gratitude to my mentor Dr. Zelin Li. I've gained lots of valuable suggestions and experiences of PhD research and research projects in Data 61. Moreover, he gave me lots of help that are very important to my personal life, which significantly makes my life in Australia easier.

Last but not least, I also want to thank my family. Thank you, my beloved wife, for accompanying me and fully supporting me when I am pursuing this PhD degree. Thank you, my selfless parents, for helping us taking care of our two lovely kids, so that we can work with dedication. Thank you, my kids, for trusting us and experiencing a harsh life transfer with us. You all are my motivations for doing good

life research, including this PhD.

Many thanks to UTS and Data 61 for providing scholarship and research platform to me, so I can accomplish myself with a PhD degree.

Zongjian Zhang

Sydney, Australia, 2020.

List of Publications

Journal Papers

J-1. **Z. Zhang**, Q. Wu, Y. Wang and F. Chen, "High-Quality Image Captioning With Fine-Grained and Semantic-Guided Visual Attention," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1681-1693, 2019.

J-2. **Z. Zhang**, Q. Wu, Y. Wang and F. Chen, "Exploring Region Relationships Implicitly: Image Captioning with Visual Relationship Attention," *Image and Vision Computing*, 2021.

J-3. **Z. Zhang**, Q. Wu, Y. Wang and F. Chen, "Adaptively Exploring Visual Relationship through Linguistic Context," *IEEE Transactions on Multimedia*, Under review.

Conference Papers

C-1. **Z. Zhang**, Q. Wu, Y. Wang and F. Chen, "Fine-Grained and Semantic-Guided Visual Attention for Image Captioning," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1709-1717, Mar. 12-15, 2018.

C-2. **Z. Zhang**, Q. Wu, Y. Wang and F. Chen, "Size-Invariant Attention Accuracy Metric for Image Captioning with High-Resolution Residual Attention," *Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1-8, Dec. 10-13, 2018.

C-3. **Z. Zhang**, Q. Wu, Y. Wang and F. Chen, "Visual Relationship Attention for Image Captioning," *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, Dec. 14-19, 2019.

Contents

List of Figures