Doctor of Philosophy

CRICOS Code: 058666A

September 2019

Course Code:C02047

Generating Descriptive and Accurate Image Captions with Neural Networks

Lingxiang Wu

School of Electrical and Data Engineering
Faculty of Engg. & IT
University of Technology Sydney
NSW2007, Australia

Generating Descriptive and Accurate Image Captions with Neural Networks

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

by

Lingxiang Wu

to

School of Electrical and Data Engineering
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW2007, Australia

September 2019

Certificate of original authorship

I, Lingxiang Wu 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 reference 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: 2019/09/01

Abstract

Image captioning is to automatically describe an image with a sentence, which is a topic connecting computer vision and natural language processing. Research on image captioning has great impact to help visually impaired people understand their surroundings, and it has potential benefits for the sentence-level photo organization. Early work typically tackled this task by retrieval methods or template methods. Modern methods were mainly based on a combination of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). However, generating accurate and descriptive captions remains a challenging task. Accurate captions refer to sentences consistent with the visual content, and descriptive captions refer to those with diverse descriptions rather than plain common sentences. Generally, the vision model is required to encode the context comprehensively and the language model is required to express the visual representation into a readable sentence consistently. Additionally, the training strategy also affects the performance.

In this thesis, we develop methods and technics to generate descriptive and accurate image captions from three aspects with neural networks. First, we consider how to express the visual representation consistently in the language model. We propose a Recall Network that can selectively import the visual information using GridLSTM units in the RNN. This design efficiently prevents the RNN from deviating from the visual representation while gradually generating each word. Second, we explore a comprehensive visual representation in the vision model based on Graph Neural Networks (GCN). A grid-level visual graph is introduced to work collaboratively with a region-level graph, and GCN are applied to aggregate visual neighborhood information in the graphs. Finally, we design a new training strategy with Reinforcement Learning (RL) technics to boost captions' diversity. Unlike previous CNN-RNN frameworks, our framework contains an additive noise module which can manipulate the transition hidden states in the RNN. We train the noise module by our proposed noise-critic training algorithm.

Then we extend the caption generation into Modern Chinese Poetry creation from images. We identify three challenges in this task: (a) semantic inconsistency between images and poems, (b) topic drift problems, and (c) frequent occurrence of certain words. Regarding the challenges, we develop a Constrained Topic-Aware Model. Particularly, we construct a visual semantic vector via image captions. A topic-aware generator is developed based on the Recall Network. An Anti-Frequency Decoding scheme is introduced to constrain the high-frequency characters.

Overall, we propose three novel modules for image captioning and an image-to-poetry framework. This thesis can facilitate the connection between computer vision and natural language, and it extends the generation into specific domains.

Acknowledgments

First and foremost, I must thank my supervisor Min Xu for her continuous support and guidance throughout my PhD. Min provided me much freedom on what I want to do, and thus made my PhD life pretty enjoyable. She is foresighted and professional, and she always provided helpful suggestions when I was confused. I am grateful that Min kindly supported me when I met difficulties in research or in personal life.

I am also thankful to Wenguan Wang and Shengsheng Qian for their generous guidance, encouragement and collaboration. It is a great opportunity to collaborate with them in two projects. I enjoyed very much discussing research with them, and I learned various aspects from them such like insightful thinking and technical writing.

There are many others I would like to express thanks during my two interns. Great thanks to Jinqiao Wang and Guibo Zhu who were my mentors in National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. They introduced me in the research community and helped me a lot in my first research attempt. Also, a big thanks to Jianwei Cui, who is my mentor in Xiaomi Corporation. I was fortunate enough to work with Jianwei and had the freedom to choose my tendentious task.

My life in UTS would not be such enjoyable without my friends and colleagues in Sydney: Tianrong Rao, Haimin Zhang, Zhongqin Wang, Lei Sang, Ruiheng Zhang, Xiaoxu Li etc. Great thanks to Lei Sang for his generous ideas and professional collaboration. Thanks to Stuart Perry, Qiang Wu and Sean He for being my candidate assessment panels. Thanks for Dadong Wang for being my co-supervisor. Thanks to Jian Zhang for making their server available to me when we are short of GPUs at the beginning of my PhD. Thanks to Chandranath Adak (UTS) for providing this thesis template. Thanks to NVIDIA who provided a GPU for our research.

Finally, I owe many thanks to my boyfriend Xinyu Tang for his constant support and free help. He is an anonymous author in most of my papers. This thesis is dedicated to my parents and boyfriend for your love and support all these years.

LIST OF PUBLICATIONS

RELATED TO THE THESIS :

1. **Lingxiang Wu**, Min Xu, Jinqiao Wang, Stuart Perry, *Recall What You See Continually Using GridLSTM in Image Captioning[J]*, IEEE Transactions on Multimedia (TMM) 2019. (***Published***).

2. **Lingxiang Wu**, Min Xu, Shengsheng Qian, Jianwei Cui , *Image to Modern Chinese Poetry Creation via A Constrained Topic-Aware Model[J]*, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). (***Under review***).

3. **Lingxiang Wu**, Min Xu, Wenguan Wang, Lei Sang, *Diverse Image Captioning with Trainable Noise and Dual Level Visual Graphs[J]*,IEEE Transactions on Image Processing (TIP). (***Under review***).

OTHERS :

1. **Lingxiang Wu**, Jinqiao Wang, Guibo Zhu, Min Xu, Hanqing Lu, *Person re-identification via rich color-gradient feature[C]*, IEEE International Conference on Multimedia and Expo (ICME) 2016. (***Published***).

2. **Lingxiang Wu**, Min Xu, Guibo Zhu, Jinqiao Wang, Tianrong Rao, *Appearance features in Encoding Color Space for visual surveillance[J]*, Neurocomputing 308: 21-30 (2018). (***Published***).

3. Ruiheng Zhang, **Lingxiang Wu**, Yukun Yang, Wanneng Wu, Yueqiang Chen, Min Xu, *Multi-camera Multi-player Tracking with Deep Player Identification in Sports Video[J]*, Pattern Recognition. (***Under review***).

Table of Contents

II Proposed Methods for Image Captioning 17

List of Figures

Figure **Page**

List of Tables

Table Page