

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

List of Publications	vii
List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Overview	1
1.2 Main Challenges	2
1.3 Contributions and Outlines	3
I Foundations	7
2 Related Works	9
2.1 Traditional Captioning Methods	10
2.1.1 Retrieval-based Methods	10
2.1.2 Template-based Methods	10
2.2 Modern Captioning Methods	11
2.2.1 Neural Network Methods	11
2.2.2 Dense Captioning and Styled Captioning	12
2.2.3 Other Vision-language Tasks	12
3 Problem Statement	15
3.0.4 Training with Cross Entropy Loss	15
3.0.5 Training with Reinforcement Learning	16

II Proposed Methods for Image Captioning	17
4 Language Model: Recall Network	19
4.1 Introduction	20
4.2 LSTM Background	21
4.3 Method	22
4.3.1 Framework Overview	22
4.3.2 Recall the Visual Information Continually in the Decoder	24
4.3.3 Recall with the Depth Dimension LSTM	25
4.3.4 Different Vision Models	27
4.4 Experiments	28
4.4.1 Datasets	29
4.4.2 Evaluation Metrics	30
4.4.3 Training Details	30
4.4.4 Experiments on Full Images	31
4.4.5 Experiments on Dense Captioning	37
4.5 Summary	38
5 Vision Model: Dual GCN	41
5.1 Introduction	41
5.2 GCN Background	42
5.3 Image captioning by two visual relation graphs	43
5.3.1 Problem Formulation	44
5.3.2 Region Visual Graph	44
5.3.3 Grid Visual Graph	45
5.3.4 GCN for Directed Labeled Graph	45
5.3.5 RNN Generator	46
5.4 Experiments	46
5.4.1 Experimental Setup	46
5.4.2 Comparative Models	47
5.4.3 Ablative Analysis	48
5.4.4 Comparison with State-of-the-art Models	49
5.5 Summary	50
6 Training Strategy: Noise Agent	51
6.1 Introduction	51

6.2	RL Background	53
6.3	Boosting Caption Diversity with a Noise Agent	54
6.3.1	Gaussian Noise Agent	54
6.3.2	Noise Critic Training	55
6.3.3	Parallel Noise Decoding	57
6.4	Experiments	57
6.4.1	Experimental Setup	57
6.4.2	Comparative Models	57
6.4.3	Ablative Analysis	58
6.4.4	Comparison with State-of-the-art Models	59
6.4.5	Diversity Evaluation	60
6.4.6	Qualitative analysis	60
6.5	Summary	61

III Extensions 63

7 Creating Modern Chinese Poetry from Images 65

7.1	Introduction	65
7.2	Background	69
7.3	Creating Poetries From Images	70
7.3.1	Problem Formulation	71
7.3.2	Visual Semantic Vector Construction	71
7.3.3	Topic-Aware Poetry Generation	74
7.3.4	Anti-frequency Decoding	77
7.4	Experiments	79
7.4.1	Dataset	79
7.4.2	Implementation Details	79
7.4.3	Evaluation Metrics	80
7.4.4	Comparative Methods	81
7.4.5	Quantitative Evaluation	82
7.4.6	Qualitative Evaluation	86
7.4.7	Effects of Hyperparameter λ	87
7.5	Summary	88

8 Conclusion 89

TABLE OF CONTENTS

8.1	Summary	89
8.2	Future Work	90
	Reference	93

LIST OF FIGURES

FIGURE	Page
1.1 Examples of object recognition, full image captioning and dense captioning. . .	2
1.2 Suboptimal descriptions from automatic generators.	3
3.1 The basic CNN-RNN framework for image captioning.	15
4.1 An illustration of the proposed model. Our model contains two LSTMs at each step. We use the temporal LSTM to transmit the sequential caption states and the depth LSTM to integrate the visual information.	23
4.2 Three different decoder units. (a): The image representation is only processed once with the LSTM. (b): The image representation and word embedding are concatenated as extra inputs for the LSTM. (c): Our Recall Network.	26
4.3 Loss comparison on validation split along the training process.	32
4.4 CIDEr score comparison on validation split along the training process.	33
4.5 Captions generated by the Recall Network. The comparison between the Recall Network and the conventional encoder-decoder model shows that our model generates more accurate and detailed captions. The comparison between the generated results and the human annotated ground truth shows that further work should be carried out for more comprehensive content captions.	36
4.6 Examples generated by the Recall Network in dense captioning. Each caption corresponds to the bounding box in the same color.	38
5.1 Region-level and grid-level visual graph construction.	42
5.2 The proposed CNN-GCN-RNN model. The image is represented by a region level graph and a grid level graph. GCN is utilized to encode neighbourhood information.	43
6.1 (a) Existing RL framework in captioning. (b) Our proposed one with the noise module.	52

6.2	The transition details in the noise-added model.	54
6.3	Generation examples with naive noise and adaptive noise.	56
6.4	CIDEr and BLEU-4 log on validation set while training the noise agent . . .	59
6.5	Generated captions from ResNet baseline, our G-noise model and the Ground Truth (GT).	61
7.1	Three challenges in the image-poetry generation. (a) Semantic inconsistency between the image and the poetry. (b) Topic drift in the subsequent lines. (Relative words are shown in bold.) (c) Frequent occurrence of certain words (shown in bold).	67
7.2	The framework of our proposed CTAM	70
7.3	Comparison on keywords obtained from two different visual models.	74
7.4	The topic-aware LSTM details.	75
7.5	Comparison results. (a) The occurrences of seven frequently appearing characters generated by the model w/ and w/o anti-frequency decoding. (b) The percentage of generated poetries containing keywords.	84
7.6	The number of poetry keywords and caption words.	85
7.7	Visualization of image captions and related keywords.	87
7.8	Visualization of generated results of Sequence model and the CTAM. We manually highlight relative words in poems, and present their generation probabilities in blue bars. Human evaluation results on Semantic consistency (S), Readability (R), Poeticness/Aesthetics (P) and the Average score (A) are shown in orange stripes.	87
7.9	Frequency of character "I" with respect to variant hyperparameter λ	88
8.1	An example with detailed entities. This figure is from RichCaptioning[78] . .	91

LIST OF TABLES

TABLE	Page
4.1 Dataset size summary	29
4.2 Comparison with the baseline methods	32
4.3 Comparison with the Stacked LSTM	33
4.4 Comparison with the state-of-the-art methods on MSCOCO dataset for Offline Evaluation.	34
4.5 Online Evaluation Performance on the MSCOCO Test Server	35
4.6 Dense Captioning Performance on the Visual Genome Dataset	38
5.1 GCN Model Performance comparisons on MSCOCO offline set	48
5.2 Comparisons on the updating times of the Grid GCN.	49
5.3 Performance comparisons with the state-of-the-art methods on MSCOCO offline set	49
5.4 Performance evaluation on MSCOCO online Test Server	50
6.1 Performance comparisons for ablation study on MSCOCO offline set	58
6.2 Performance comparisons with the state-of-the-art methods on MSCOCO offline set	59
6.3 Performance evaluation on MSCOCO online Test Server	60
6.4 Diversity Evaluation	60
7.1 Listing of notations in this chapter	71
7.2 Human Evaluation Results	82
7.3 Ablation study with BLEU-1 (%)	84
7.4 Image captioning performance	86

