Faculty of Engineering and Information Technology
University of Technology Sydney

Vision to Keywords: Automatic Image Annotation by Filling the Semantic Gap

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Junjie Zhang

July 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Junjie 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 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 removed prior to publication.

Signature of Candidate: _____

Date: 08/07/2019

Acknowledgments

First of all, I would like to express the deepest appreciation to my principal supervisor Prof. Jian Zhang for his continuous support of my Ph.D. study; he has been a tremendous mentor for me. I have benefited a lot from the weekly discussions around the research problems, the valuable insights on the paper writing and the generous support for attending conferences. His patience, enthusiasm and professional guidance have helped me to grow as a research scientist as well as the writing of this thesis.

I would also like to thank my co-supervisor Prof. Qiang Wu and the collaborators: Prof. Chunhua Shen, Dr. Qi Wu and Prof. Jianfeng Lu, for not only their insightful guidance and comments on revising my paper drafts but also for the thought-provoking questions, which have incented me to widen my research from various perspectives.

I am grateful to my fellow labmates in the Global Big Data Technologies Centre: Xiaoshui Huang, Yifan Zuo, Lina Li, Huaxi Huang, Muming Zhao and Zongjian Zhang for the stimulating discussions on the research problems and the sleepless nights we worked together before deadlines, and all the fun we have had for the last three years.

Last but not least, my heartfelt gratitude to my parents, it is their selfless supports and confidence in me that help me to overcome the difficulties I have encountered during my Ph.D. study.

Junjie Zhang
July 2019 @ UTS

Contents

List of Figures

List of Tables

List of Publications

Papers Published

- **Junjie Zhang**, Qi Wu, Jian Zhang, Chunhua Shen, and Jianfeng Lu (2019), Mind Your Neighbours: Image Annotation with Metadata Neighbourhood Graph Co-Attention Networks. *in* IEEE International Conference on Computer Vision and Pattern Recognition (CVPR19).

- Huaxi Huang, **Junjie Zhang**, Jian Zhang, Qiang Wu, Jinsong Xu (2019), Compare More Nuanced: Pairwise Alignment Bilinear Network For Few-shot Fine-grained Learning. *in* IEEE International Conference on Multimedia and Expo (ICME19).

- **Junjie Zhang**, Qi Wu, Jian Zhang, Chunhua Shen, Jianfeng Lu and Qiang Wu (2019), Heritage image annotation via collective knowledge. *in* Pattern Recognition (PR), vol. 93, pp. 204-214.

- **Junjie Zhang**, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu (2018), Multi-label Image Classification with Regional Latent Semantic Dependencies. *in* IEEE Transactions on Multimedia (T-MM), vol. 20, no. 10, pp. 2801-2813.

- **Junjie Zhang**, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel (2018), Goal-Oriented Visual Question Generation via Intermediate Rewards. *in* The European Conference on Computer Vision (ECCV18), pp. 189-204.

- **Junjie Zhang**, Qi Wu, Jian Zhang, Chunhua Shen, and Jianfeng Lu (2018), Kill Two Birds with One Stone: Weakly-Supervised Neural Network for Image Annotation and Tag Refinement. *in* Proceedings of the AAAI Conference on Artificial Intelligence (AAAI18), pp. 7550-7557.

- **Junjie Zhang**, Jian Zhang, Qi Wu, Qiang Wu, Jinsong Xu, Jianfeng Lu, Robin Phua, Kate Curr, and Zhenmin Tang (2017), Historical Image Annotation by Exploring the Tag Relevance. *in* The 4th Asian Conference on Pattern Recognition (ACPR17), published online.

- **Junjie Zhang**, Jian Zhang, Jianfeng Lu, Chunhua Shen, Kate Curr, Robin Phua, Richard Neville, and Elise Edmonds (2016), SLNSW-UTS: A Historical Image Dataset for Image Multi-Labeling and Retrieval. *in* Proceedings of the Digital Image Computing: Techniques and Applications (DICTA16), pp. 1-6.

Abstract

Nowadays, images are generated at an explosive pace, which yields the urgent need for an efficient annotation method to assist people to understand them. By assigning multiple labels to an image, we can transfer the visual information to keywords, which are more convenient to index.

The key issue behind this topic is bridging the semantic gap existing between the image visual content and multiple semantic labels. Given a training set of images with manually annotated expert labels, there are two main factors to establish an efficient annotation model, namely the visual relevance between the image and labels, and the semantic dependency between the label pairs. Moreover, images often carry abundant metadata, these metadata can be as informative as the pixel contents and exploring the relevance between the image and metadata also plays an important role in the image annotation. This thesis summarises the works that have conducted on utilising these factors for the image annotation.

In Chapter 3, a Regional Latent Semantic Dependencies model is introduced for the generic image annotation, which effectively captures the latent semantic dependencies at the regional level. In Chapter 4, the weakly supervised annotation and tag refinement for social images are studied. User-provided tags are the most common metadata on the social network, even though they reveal the semantic meaning of the image visual content, it is well-known that they can be incomplete and imprecise to a certain extent. We propose to learn an image annotation model and refine the user-provided tags simultaneously in a weakly-supervised manner. In Chapter 5, the representation learning and image annotation for the diverse image set are studied. We uncover relationships between the image and

its neighbours by measuring similarities among their metadata and conduct the metric learning to obtain the representations of image contents; we also generate semantic representations for images given the collective semantic information from their neighbours. In Chapter 6, the image annotation problem is addressed by routinely checking its neighbours in a graph, which is constructed by the equipped meta information of the image. We propose a graph network to model the correlations between each target image and its neighbours. To accurately capture the visual clues from the neighbourhood, a co-attention mechanism is introduced to embed the target image and its neighbours as graph nodes. Chapter 7 concludes the thesis and outlines the scope of future work.