

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 incited 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

Certificate	i
Acknowledgment	ii
List of Figures	vii
List of Tables	xiv
List of Publications	xvi
Abstract	xviii
Chapter 1 Introduction	1
1.1 Background	1
1.2 Research Issues	5
1.3 Research Contributions	6
1.4 Thesis Structure	8
Chapter 2 Related Works	10
2.1 Summary	10
2.2 Generic Image Annotation	11
2.2.1 Classification-Based	11
2.2.2 Voting-Based	12
2.2.3 Multi-Modality Design	14
2.2.4 Graphical Inference	15
2.2.5 Deep Neural Network-Based	17
2.2.6 Comparisons	18
2.3 Annotation with Metadata	19
2.3.1 User-Tag Relevance Analysis	19

2.3.2	Other Metadata Types	20
Chapter 3	Regional Latent Semantic Dependencies	22
3.1	Introduction	22
3.2	The RLSD Model	25
3.2.1	Framework Overview	25
3.2.2	Localising Multi-label Regions	26
3.2.3	An LSTM-based Multi-Label Generator	29
3.2.4	Max-pooling and Loss Function	32
3.2.5	Initialisation and Pre-Training	33
3.3	Experiments	34
3.3.1	Implementation Details	34
3.3.2	Evaluation Metrics	35
3.3.3	Baseline Models	36
3.3.4	Results on the MS-COCO	38
3.3.5	Results on the NUS-WIDE	41
3.3.6	Results on the PASCAL VOC Datasets	42
3.4	Summary	46
Chapter 4	Weakly-Supervised Annotation & Tag Refinement	48
4.1	Introduction	48
4.2	The Proposed Model	53
4.2.1	Representative Region Selection	53
4.2.2	Regional Neural Network	54
4.2.3	Input Batch Selection & Visual Consistency	56
4.2.4	Semantic Dependency	60
4.2.5	User-Error Sparsity	61
4.2.6	Training and Prediction	62
4.3	Experiments	63
4.3.1	Data Preprocessing	63
4.3.2	Evaluation Metrics	64
4.3.3	Baselines and Compared Methods	64

4.3.4	Results on Image Annotation	66
4.3.5	Results on Tag Refinement	68
4.3.6	Ablation Analysis	71
4.3.7	Visualisation of the Constraints	74
4.4	Summary	76
Chapter 5	Annotation via Collective Knowledge	78
5.1	Introduction	78
5.2	Proposed Model	81
5.2.1	Model Overview	81
5.2.2	Visual Representation Learning	82
5.2.3	Label Relevance Analysis & Semantic Summarisation	84
5.2.4	Annotation via Collective Knowledge	86
5.2.5	Training and Prediction	87
5.3	Experiments	88
5.3.1	Data Preprocessing & Evaluation Metrics	88
5.3.2	Implementation Details	89
5.3.3	Baselines and Compared Methods	90
5.3.4	Results on Image Annotation	91
5.3.5	Ablation Study	96
5.4	Summary	97
Chapter 6	Annotation via Graph Co-Attention Networks	99
6.1	Introduction	99
6.2	The MangoNet	102
6.2.1	Neighbourhood Graph Co-Attention	103
6.2.2	Implementation Details	108
6.3	Experiments	108
6.3.1	Datasets	109
6.3.2	Evaluation Metrics	109
6.3.3	Overall Performance	110
6.3.4	Ablation Study	114

6.3.5	Visualisation of Attention	116
6.4	Summary	116
Chapter 7	Conclusions and Future Work	118
7.1	Conclusions	118
7.2	Future Work	120
Bibliography	121

List of Figures

1.1	An example of the image annotation. There are multiple objects: grass, people and baseball in this image. From the fine-grained to coarse, the image's scene categories are batter box, sports field and outdoor. Bounding boxes are only for the illustration.	2
1.2	The illustration of the main factors in the image annotation.	3
1.3	The illustration of the thesis structure.	9
2.1	The illustration of the classification-based methods. The labels are represented as separate classifiers. When annotating the new image, all the classifiers will be applied, and the fusion of all the classifiers' outputs is the final annotation.	11
2.2	The illustration of the voting-based methods. The labels are transferred from the image neighbours within the training set via the neighbour search.	13
3.1	Example results of multi-label prediction from different models. The left is the ground-truth, and the middle column shows the results from baseline models, Multi-CNN and CNN+LSTM. The right column displays the outputs of our proposed RLSD model, including predicted multiple labels and selected region proposals. Compare to the baseline methods, our model produces much richer predictions and is especially good at predicting small objects, such as 'bottle,' 'wine glass' and 'vase' <i>etc.</i>	23

3.2	Our proposed Regional Latent Semantic Dependencies model. An input image $3 \times H \times W$ is first processed through a CNN to extract convolutional features, which are further sent to an RPN-like fully convolutional localisation layer. The localisation layer localises the M regions in an image that potentially contain multiple highly-dependent labels. These regions are encoded with a fully-connected neural network and sent to the regional LSTM to produce T timestep probability distributions over dataset labels L , which results in a shape $M \times T \times L$ tensor. Finally, a max-pooling operation is carried out to fuse all the regional outputs as the final prediction.	25
3.3	The comparison results between the Top-15 regions generated by MCG (left) and our localisation layer (right). Some of our generated regions contain multiple objects, for example, the generated regions contain the object of ‘oven/microwave/kitchenwares’, ‘person/tennis racket,’ and ‘person/kite/car’ altogether.	30
3.4	The structure of LSTM.	31
3.5	An illustration of proposed RLSD model for the test image. The potential multi-label regions of the test image are generated by localisation layer and further used to extract features and input to shared LSTM. As we can see, the small-sized objects like ‘wine glass,’ ‘bottle’ and ‘vase’ <i>etc.</i> can be included in the regions due to our multi-label localisation network. The test is also performed in an end-to-end fashion.	33
3.6	The comparison results between the object bounding boxes (left) and generated multi-label regions (right). Our generated regions contain multiple objects, for example, the generated regions contain the object of ‘person/baseball/baseball bat,’ ‘cup/oven’ and ‘oven/fridge’ altogether.	36

3.7	Precision-Recall curves for the ‘bird,’ ‘fire hydrant’ and ‘kite’ classes in the MS-COCO dataset, for our RLSD models and the baseline models. The average precision @ ten is also given in the figure.	39
3.8	The relationship between recall and bounding-box area on the MS-COCO dataset.	40
3.9	Some example annotation results from the MS-COCO dataset. . .	41
3.10	The statistics of training/validation set of VOC datasets.	46
3.11	The visualisation of semantic dependencies for the VOC 2007 and VOC 2012 datasets. The brighter the block is, the higher dependency the corresponding label pair has.	46
4.1	An example of the social image annotation and the tag refinement. As we can see, the image annotation assigns tags that describe the visual content to the image, while the tag refinement removes the inaccurate tags from the user-provided set, and add relevant ones to the image.	49

4.2	The proposed framework. Each image $i \in I$ is first extracted with multiple region proposals; then representative regions are selected by the grouping and merging. The Nearest Neighbor (NN) approach is performed at both the image level and regional level to locate the region pairs for the network input. Each input mini-batch for the regional neural network B_i is comprised of the image i and its image neighbours, while the corresponding region pairs are marked. After the regional level confidence scores are obtained, the visual consistency and semantic dependency constraints are applied. Moreover, with the max-pooling operation, the regional scores are fused as the final image confidence score, where the user-error sparsity constraint is performed. The annotation model is trained in an end-to-end fashion, while the tag refinement is conducted during the training. When the new image comes, the representative regions are first extracted, then the image and the regions are sent to the regional neural network to generate the tag prediction.	50
4.3	The examples of images assigned to multiple tags. As we can see, different tags such as ‘motorbike,’ ‘building,’ and ‘dog’ <i>etc.</i> correspond to the different regions of the image, instead of the whole image, which is the motivation that we explore the image regional information for both tasks.	52
4.4	The representative region selection. We group proposals generated by GOP (left column) and merge the top proposals in each group into one representative region (right column).	55
4.5	An illustration of the selection process for the positive region pairs. Nearest neighbour is performed in twofold. First, the image candidates are selected at the image level; then we measure the visual similarity between regions among image and its candidates at the regional level.	58

4.6	The example results of the image annotation and tag refinement. The first two rows are the image annotation examples, while the last two rows are the tag refinement results.	69
4.7	The AP comparisons of the proposed model Regional-CNN+ L_{use} against baselines Multi-CNN and Regional-CNN of the image annotation on both datasets.	70
4.8	The AP ratios of the proposed model Regional-CNN+ L_{use} against baselines Multi-CNN and Regional-CNN of the tag refinement on both datasets.	72
4.9	The visualisation of the visual consistency on three visually similar region pairs. Each row under two region pairs stands for the tag probability distributions with the size $2 \times C$, where $C = 444$ for the Mirflickr dataset.	75
4.10	The visualisation of three semantic dependent tag pairs' region distributions. As we can see, the high semantic dependent tag pairs have similar annotation distributions. Each row stands for the region distributions of the tag pair with the size 2×100	76
4.11	The visualisation of the user-error matrix with the size 68×444 . .	76
5.1	Examples of the training image from Flickr dataset and the heritage image collection. As we can see, the labels of the heritage image collection are more diverse and semantical.	79
5.2	The visual ambiguous of the heritage image collection. All images are annotated with the label 'theatre.' However, the visual appearance of them is quite different. Figure (a) is the exterior of a theatre, Figure (b) is the interior of a theatre, Figure (c) is the hall of a theatre, while Figure (d) is a group people taking a photo outside a theatre.	79

5.3	The framework of the proposed model. At the training stage, we first generate image pairs (i, i') for the visual representation learning based on the metadata similarity, in our case, the combination of the description similarity d_{sim} and location similarity g_{sim} . Image pairs are passed through the siamese network and trained with the contrastive loss. Then we retrieve image neighbours and summarise the semantic representation by adopting the weighted sum based on the image pair similarity and label relevance. Both the visual representation $G_X(i)$ and the semantic representation $G_S(i)$ are fed into the fully-connected network to compute hidden states and further trained with the sigmoid cross-entropy loss.	80
5.4	Examples of the description and the location of the heritage image collection. As we can see, these metadata can help understand the image at a semantical level.	83
5.5	Some example annotation results on the heritage image collection.	93
5.6	(a) The first row is the PR-curves of the label ‘festival’ compared with baselines and state-of-the-art methods. (b) The second row is the PR-curves of the label ‘Art Deco architecture’ compared with baselines and state-of-the-art methods. The average precision is also given in the figure. Better view in color.	94
5.7	The comparisons of the average precision (AP) values between the proposed model and baselines on all labels. The left one is the AP values of our model against KNN baseline, while the right one is our model against Multi-CNN baseline. Better view in color. . .	95
5.8	The examples of training pairs for the metric learning based on the metadata similarity. The first column is the query image with the blue box, the second column is its positive pair with the green box, and the third and fourth columns are negative ones with brown boxes.	96

5.9	Some examples of label suggestions retrieved from image neighbours. We summarise these suggestions as the semantic representation.	98
6.1	For the target images with the red boxes, they are hard to recognise on their own. However, in the context of the neighbours with the similar metadata, such as ‘vehicle, vintage, Beetle’ and ‘church, instrument, art’, it is more clear that the target images are a car and a pipe organ. Based on this motivation, we propose a neighbourhood graph as ‘ <i>neighbourhood watch</i> ’ to assist the image annotation.	100
6.2	The framework of the proposed model. The neighbourhood z of the target image i is decided by measuring their metadata similarities. Then we establish the neighbourhood graph using image representations as nodes and correlations as edges. To accurately harvest visual clues from its neighbours, we introduce a co-attention mechanism to guide the Graph Convolutional Network (GCN) and obtain the graph representation, which is then concatenated with the target global feature to generate the label confidence.	101
6.3	The instance-level attention module we used to capture the regional semantic correspondences between image content and associated labels.	105
6.4	The co-attention attention module. The co-attention maps are the semantic overlap of instance-level attention maps between the target image i and its neighbourhood z	106
6.5	The AP comparisons of the proposed MangoNet against the NCNN model on the NUS-WIDE, Mirflickr and MS-COCO.	115
6.6	The visualisations of the co-attention maps of the targets and their neighbours. The first column is the target and its attention map, the rest columns are the neighbours, where the neighbourhood size $m = 3$	117

List of Tables

3.1	Comparisons on the MS-COCO dataset for $k = 3$. mAP@10 measures are additionally computed for comparison.	38
3.2	Comparisons on the NUS-WIDE dataset on 81 concepts for $k = 3$	42
3.3	Comparisons of annotation results on the VOC 2012 dataset.	43
3.4	Comparisons of annotation results (AP in %) on the VOC 2007 dataset. Both AlexNet and VGGNet have been used as the backbone model. The best results of AlexNet and VGGNet based models are noted in bold respectively.	44
4.1	The statistics of the Mirflickr and NUS-WIDE dataset after the preprocessing, including the total image amount, the size of the user-provided tag set and expert label set, and the number of user-provided tags per image.	63
4.2	The image annotation results on the Mirflickr dataset.	67
4.3	The image annotation results on the NUS-WIDE dataset.	67
4.4	The tag refinement results on the Mirflickr dataset.	69
4.5	The tag refinement results on the NUS-WIDE dataset.	71
4.6	The ablation results of the image annotation on the Mirflickr dataset.	73
4.7	The ablation results of the image annotation on the NUS-WIDE dataset.	73
4.8	The ablation results of the tag refinement on the Mirflickr dataset.	73
4.9	The ablation results of the tag refinement on the NUS-WIDE dataset.	73
4.10	The high-dependent tag pairs.	74

5.1	Results of the Image Annotation.	92
5.2	Results of the Ablation Study.	96
6.1	Image annotation results compared with other state-of-the-art methods and our MangoNet on the NUS-WIDE, where m indicates the neighbourhood size used in our model.	110
6.2	Image annotation results compared with other state-of-the-art methods and our MangoNet on the Mirflickr, where m indicates the neighbourhood size used in our model.	111
6.3	Image annotation results compared with other state-of-the-art methods and our MangoNet on the MS-COCO, where m indicates the neighbourhood size used in our model.	112
6.4	Ablation studies of our model on the NUS-WIDE.	113
6.5	Ablation studies of our model on the Mirflickr.	113
6.6	Ablation studies of our model on the MS-COCO.	114

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.