

Faculty of Engineering and Information Technology
University of Technology Sydney

Towards Automatic Construction of Diverse, High-quality Image Dataset

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

by

Yazhou Yao

September 2018

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

Production Note:
Signature removed prior to publication.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisors Prof. Jian Zhang for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

I also would like to appreciate my co-supervisor Prof. Zhenming Tang and Fumin Shen for providing me with continuous support throughout my Ph.D study and research. Without their professional guidance and persistent help, this thesis would not have been possible.

I am very grateful to Senior Researcher Xian-sheng Hua at Alibaba Research, and I have benefited greatly from every high-quality group meeting. Their cutting-edge research perspectives and solid theoretical foundation make me open-minded and inspire me to work harder.

I thank my fellow labmates in Global Big Data Technologies Center: Yucheng Wang, shangrong Huang for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last three years.

Last but not the least, I would like to thank my family: my wife and my parents, for their unconditional support, both financially and emotionally throughout the whole Ph.D studying.

Yazhou Yao

March 2018 @ UTS

Contents

Certificate	i
Acknowledgment	iii
List of Figures	ix
List of Tables	xiii
List of Publications	xv
Abstract	xvii
Chapter 1 Introduction	1
1.1 Background	1
1.2 Research Issues	4
1.3 Research Contributions	6
1.4 Thesis Structure	7
Chapter 2 Literature Review and Foundation	10
2.1 Polysemy-oriented Methods	11
2.2 Diversity-oriented Methods	13
2.3 Accuracy-oriented Methods	15
2.3.1 Manual Annotation Methods	15
2.3.2 Active Learning Methods	17
2.3.3 Automatic Methods	17
2.4 Privileged Information	18
Chapter 3 Polysemy	20
3.1 Introduction	20
3.2 Framework and Methods	22

CONTENTS

3.2.1	Discovering Possible Semantic Senses	22
3.2.2	Merging and Pruning Semantic Senses	23
3.2.3	Distinguishing Visual Senses	29
3.3	Experiments	30
3.3.1	Classifying Sense-specific Images	30
3.3.2	Re-ranking Search Results	34
3.4	Conclusions	37
Chapter 4	Diversity	42
4.1	Introduction	42
4.2	Domain robust dataset construction	43
4.2.1	Query Expanding	44
4.2.2	Noisy Expansions Filtering	45
4.2.3	Noisy Images Filtering	47
4.3	Experiments	54
4.3.1	Image Dataset DRID-20 Construction	54
4.3.2	Comparison of Classification Ability, Cross-dataset Generalization Ability, and Dataset Diversity	55
4.3.3	Comparison of Object Detection Ability	61
4.4	Conclusions	63
Chapter 5	Accuracy	69
5.1	Introduction	69
5.2	Framework and Methods	71
5.2.1	Multiple Textual Metadata Discovering	71
5.2.2	Noisy Textual Metadata Filtering	72
5.2.3	Noisy Images Filtering	75
5.3	Experiments	81
5.3.1	Image Dataset Construction	81
5.3.2	Comparison of Image Classification Ability and Cross-dataset Generalization Ability	83
5.3.3	Comparison of Object Detection Ability	86

5.3.4	Parameter Sensitivity Analysis	88
5.3.5	Platform Introduction	89
5.4	Conclusions	89
Chapter 6 Privileged Information		94
6.1	Introduction	94
6.2	Framework and Methods	96
6.2.1	Formulation	96
6.2.2	Optimization	98
6.3	Experiments	101
6.3.1	Image categorization	101
6.3.2	Image sub-categorization	105
6.3.3	Parameter Sensitivity Analysis	108
6.3.4	Time Complexity Analysis	110
6.4	Conclusions	110
Chapter 7 Conclusions and Future Work		113
7.1	Conclusions	113
7.2	Future Work	114
Appendix A Solutions		116
A.1	The detailed solutions to (5.1)	116
Bibliography		119

List of Figures

1.1	Visual polysemy. For example, the query “mouse” returns multiple visual senses on the first page of results. The retrieved web images suffer from the low precision of any particular visual sense.	4
1.2	Most discriminative images for “airplane” from four different datasets. Each dataset has their preference for image selection.	5
1.3	The profile of work in this thesis.	9
3.1	Illustration of the process for obtaining selected semantic senses.	22
3.2	A snapshot of the retrieved images for visual consistency and non-consistency semantic senses.	26
3.3	A snapshot of the retrieved images for selected semantic senses. Due to the error index of image search engine, even we retrieve the sense-specific images, some instance-level noise may also be included. The noisy images are marked with red bounding boxes.	28
3.4	The detailed performance comparison over 5 categories on the MIT-ISD dataset.	35
3.5	Examples of multiple visual senses discovered by our proposed approach. For example, our approach automatically discovers and distinguishes four senses for “ <i>Note</i> ”: notes, galaxy note, note tablet and music note. For “ <i>Bass</i> ”, it discovers multiple visual senses of: bass fish, bass guitar and Mr./Mrs. Bass <i>etc.</i>	38

LIST OF FIGURES

3.6	The detailed performance comparison over 30 categories on the CMU-Poly-30 dataset.	39
4.1	Domain robust image dataset construction framework. The input is text query that we would like to build a image dataset for. The outputs are a set of selected images corresponding to the given query.	44
4.2	Image classification ability of CIFAR-10, STL-10 and DRID-20 on PASCAL VOC 2007 dataset: (a) airplane, (b) bird, (c) cat, (d) dog, (e) horse, (f) car/automobile and (g) average. . .	66
4.3	Image classification ability of Optimol, Harvesting, ImageNet, AutoSet and DRID-20 on PASCAL VOC 2007 dataset.	67
4.4	Cross-dataset generalization ability of classifiers learned from CIFAR-10, STL-10, DRID-20 and then tested on: (a) CIFAR-10, (b) STL-10, (c) DRID-20, (d) Average.	67
4.5	(a) Comparison of the lossless JPG file sizes of average images for five different categories in DRID-20, ImageNet and STL-10. (b) Example images from DRID-20, ImageNet, STL-10 and average images for each category indicated by (a).	67
4.6	Cross-dataset generalization ability of classifiers learned from Optimol, Harvesting, ImageNet, AutoSet, DRID-20 and then tested on: (a) Optimol, (b) Harvesting, (c) ImageNet, (d) AutoSet, (e) DRID-20, (f) Average.	68
5.1	Illustration of the process for obtaining multiple textual metadata. The input is a textual query that we would like to find multiple textual metadata for. The output is a set of selected textual metadata which will be used for raw image dataset construction.	70

5.2	Illustration of the process for obtaining selected images. The input is a set of selected textual metadata. Artificial images, inter-class noisy images, and intra-class noisy images are marked with red, green and blue bounding boxes separately. The output is a group of selected images in which the images corresponding to different textual metadata.	72
5.3	A snapshot of the retrieved images for visual non-salient and less relevant textual metadata.	73
5.4	The image classification accuracy (%) comparison over 14 and 6 categories on the PASCAL VOC 2007 dataset.	91
5.5	The cross-dataset generalization ability of various datasets by using a varying number of training images, and tested on (a) ImageNet, (b) Optimol, (c) Harvesting, (d) DRID-20, (e) Ours, (f) Average.	92
6.1	Examples of textual tags (privileged information) for images on image sharing website “Flickr”. Both of useful and noisy tags are included.	95
6.2	Sub-categorization accuracy (%) of the different methods (a) using a varying number of training images for per subcategory, and (b) using a varying number of testing images for per subcategory.	107
6.3	The parameter sensitiveness of C_1 , η , N_p and N_n in terms of image categorization accuracy.	109
6.4	The detailed performance comparison over 20 categories on the PASCAL VOC 2007 dataset.	111
6.5	The detailed performance comparison over 10 categories on the (a) STL-10 dataset, (b) CIFAR-10 dataset.	112

List of Tables

2.1	The publicly available automatic datasets.	18
3.1	The average performance comparison of classification accuracy on the CMU-Poly-30 and MIT-ISD dataset.	34
3.2	Web images for polysemy terms were annotated manually. For each term, the number of annotated images, the semantic senses, the visual senses and their distributions are provided, with core semantic senses marked in boldface.	40
3.3	Area Under Curve (AUC) of all senses for “bass” and “mouse”.	41
4.1	Object detection results (A.P.) (%) on PASCAL VOC 2007 (TEST).	65
5.1	The average accuracy (%) comparison over 14 and 6 common categories on the PASCAL VOC 2007 dataset.	85
5.2	The average recall and precision for ten categories corresponding to different S_i	88
5.3	The average accuracy of inter-class noisy images filtering for ten categories corresponding to different δ	88
5.4	Object detection results (A.P.) (%) on PASCAL VOC 2007 dataset (Test).	93
6.1	The average performance comparison on the PASCAL VOC 2007, STL-10 and CIFAR-10 dataset.	105

LIST OF TABLES

6.2 The detailed number of subcategories used for image sub-
categorization in this experiment. 106

List of Publications

Papers Published

- **Yazhou Yao**, Fumin Shen, Jian Zhang, Li Liu, Zhenmin Tang, and Ling Shao (2018), Extracting Privileged Information for Enhancing Classifier Learning, *in* 'Proceedings of IEEE Transactions on Image Processing (**TIP**)', doi=10.1109/TIP.2018.2869721.
- **Yazhou Yao**, Fumin Shen, Jian Zhang, Li Liu, Zhenmin Tang, and Ling Shao (2018), Discovering and Distinguishing Multiple Visual Senses for Web Learning, *in* 'Proceedings of IEEE Transactions on Multimedia (**TMM**)', doi=10.1109/TMM.2018.2847248.
- **Yazhou Yao**, Jian Zhang, Fumin Shen, Xian-Sheng Hua, Wankou Yang and Zhenmin Tang (2018), Extracting Privileged Information from Untagged Corpora for Classifier Learning, *in* 'Proceedings of the International Joint Conference on Artificial Intelligence (**IJCAI 18**)', pp, 1085-1091.
- **Yazhou Yao**, Jian Zhang, Fumin Shen, Wankou Yang, Pu Huang and Zhenmin Tang (2018), Discovering and Distinguishing Multiple Visual Senses for Polysemous Words. *in* 'Proceedings of the AAAI Conference on Artificial Intelligence (**AAAI 18**)', pp, 523-530.
- **Yazhou Yao**, Jian Zhang, Fumin Shen, Xian-Sheng Hua, Jingsong Xu and Zhenmin Tang (2017), Exploiting Web Images for Dataset

Construction: A Domain Robust Approach. *in* 'Proceedings of IEEE Transactions on Multimedia (TMM)', pp. 1771-1784.

- **Yazhou Yao**, Xiansheng Hua, Fumin Shen, Jian Zhang and Zhenmin Tang (2016), A Domain Robust Approach for Image Dataset Construction. *in* 'Proceedings of the ACM International Conference on Multimedia (ACM MM 16)', pp, 212-216.
- **Yazhou Yao**, Jian Zhang, Fumin Shen, Xian-Sheng Hua, Jingsong Xu and Zhenmin Tang (2016), Automatic Image Dataset Construction with Multiple Textual Metadata. *in* 'Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 16)', pp, 1-6.
- **Yazhou Yao**, Jian Zhang, Fumin Shen, Xian-Sheng Hua, Jingsong Xu and Zhenmin Tang (2017), A New Web-supervised Method for Image Dataset Construction. *in* 'Proceedings of Neurocomputing', pp. 23-31.
- **Yazhou Yao**, Wankou Yang, Pu Huang, Qiong Wang, Yunfei Cai and Zhenmin Tang (2018), Exploiting Textual and Visual Features for Image Categorization. *in* 'Proceedings of Pattern Recognition Letters', doi=<https://doi.org/10.1016/j.patrec.2018.05.028>.
- **Yazhou Yao**, Jian Zhang, Xian-Sheng Hua, Fumin Shen and Zhenmin Tang (2016), Extracting Visual Knowledge from the Internet: Making Sense of Image Data. *in* 'Proceedings of the International Conference on Multimedia Modelling', pp, 862-873.

Papers Under Review

- **Yazhou Yao**, Jian Zhang, Fumin Shen, Dongxiang Zhang, Zhenmin Tang, and Heng Tao Shen (2017), Towards Automatic Construction of Diverse, High-quality Image Dataset, IEEE Transactions on Knowledge and Data Engineering (TKDE) (**major revision**).

Abstract

The availability of labeled image datasets has been shown critical for high-level image understanding, which continuously drives the progress of feature designing and models developing. However, the process of manual labeling is both time-consuming and labor-intensive. To reduce the cost of manual annotation, there has been increased research interest in automatically constructing image datasets by exploiting web images. Datasets constructed by existing methods tend to suffer from the disadvantage of low accuracy and low diversity. These datasets tend to have a weak domain adaptation ability, which is known as the “dataset bias problem”.

This research aims at automatically collect accurate and diverse images for given queries from the Web, and construct a domain robust image dataset. Thus, within this thesis, various methods are developed and presented to address the following research challenges. The first is the retrieved web images are usually noisy, how to remove noise and construct a relatively high accuracy dataset. The second is the collected web images are often associated with low diversity, how to address the dataset bias problem and construct a domain robust dataset.

In Chapter 3, a framework is presented to address the problem of polysemy in the process of constructing a high accuracy dataset. Visual polysemy means that a word has several semantic (text) senses that are visually (image) distinct. Solving polysemy can help to choose appropriate visual senses for sense-specific images collection, thereby improving the accuracy of the collected images. Unlike previous methods which leveraged the human-

developed knowledge such as Wikipedia or dictionaries to handle polysemy, we propose to automate the process of discovering and distinguishing multiple visual senses from untagged corpora to solve the problem of polysemy.

In Chapter 4, a domain robust framework is presented for image dataset construction. To address the dataset bias problem, our framework mainly consists of three stages. Specifically, we first obtain the candidate query expansions by searching in the Google Books Ngram Corpus. Then, by treating word-word (semantic) and visual-visual distance (visual) as features from two different views, we formulate noisy query expansions pruning as a multi-view learning problem. Finally, by treating each selected query expansion as a “bag” and the images therein as “instances”, we formulate image selection and noise removal as a multi-instance learning problem. In this way, images from different distributions can be kept while noise is filtered out.

Chapter 5 details a method for noisy images removing and accurate images selecting. The accuracy of selected images is limited by two issues: the noisy query expansions which are not filtered out and the error index of image search engine. To deal with the noisy query expansions, we divide them into two types and propose to remove noise from visual consistency and relevancy respectively. To handle noise induced by error index, we classify the noisy images into three categories and filter out noise by different mechanisms separately.

Chapter 6 proposes an approach for enhancing classifier learning by using the collected web images. Different from previous works, our approach, while improving the accuracy and robustness of the classifier, greatly reduces the time and labor dependence. Specifically, we proposed a new instance-level MIL model to select a subset of training images from each selected privileged information and simultaneously learn the optimal classifiers based on the selected images.

Chapter 7 concludes the thesis and outlines the scope of future work.