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.

Chapter 1

Introduction

1.1 Background

Images as the visual reflections of our daily lives have become widely available due to the improvement of the imaging technology and the rise of the social network, people take photos whenever and wherever they want and share them online. With the enormous users on the social network, images are generated and spread at an explosive pace, which expresses the urgent need for the annotation in order to assist people to understand as well as efficiently retrieve them.

In general, an image can be annotated with multiple labels considering it always bounds with several types of semantic concepts, like objects, attributes, and scene categories. In Figure 1.1, the image describes two people play baseball on a sports field. There are multiple labels which can be annotated to this image. For objects, there are grass, baseball bat, people and filed, and for scene categories, there are batter box, sports field and outdoor from the fine-grained to coarse. By assigning multiple semantic labels to an image, we can transfer the visual information to keywords, which are more convenient to understand and index. The accurate annotation can also spark solutions for various applications, such as retrieval, captioning and semantic segmentation.

Given the significant growth on the image amount every internet minute, it is obvious that manually annotating image contents at the semantic level are not

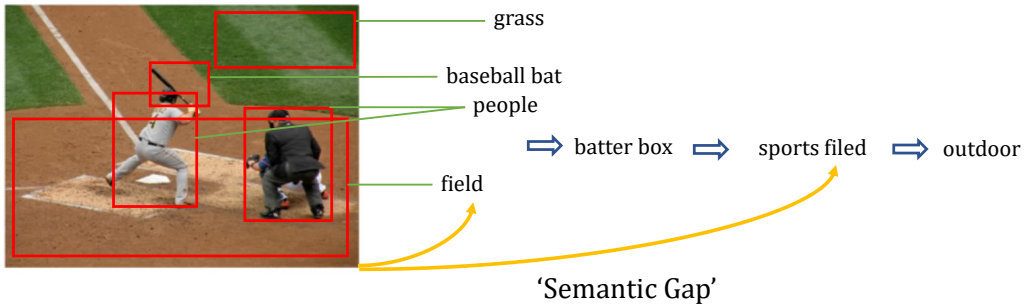


Figure 1.1: An example of the image annotation. There are multiple objects: grass, people and baseball in this image. From the fine-grained to coarse, the image's scene categories are batter box, sports field and outdoor. Bounding boxes are only for the illustration.

applicable. Efficient automatic image annotation methods are required under this circumstance. The key issue behind the annotation is the semantic gap that exists between image visual content and the corresponding multiple labels. Various studies have been conducted on how to bridge this gap from different perspectives.

Generally, there are two main factors in the Generic Image Annotation (GIA). See Figure 1.2 as an illustration. First, the visual relevance, which captures the corresponding relationship between image pixel content and its labels. The second factor is the semantic dependency, which exists within the label set. Semantically dependent labels have high probabilities occur together, for example, the label 'ocean' and 'ship' always appear together in the same image, while the 'ocean' and 'cat' usually never occur together. Moreover, social networks in addition to hosting the images for users, also record the user interactions with images as the vast amount of metadata, which is presented in various forms. As a means to communicate with other users, these metadata can be as informative as the pixel contents to understand images. Utilising the relevance between the image and metadata as well as the label and metadata are critical for the Image Annotation with Metadata Information (IAMI).

Methods for the generic image annotation have been extensively studied for decades, and most mainstream methods have been focused on annotating common visual objects and attributes (Lin, Maire, Belongie, Hays, Perona, Ramanan,

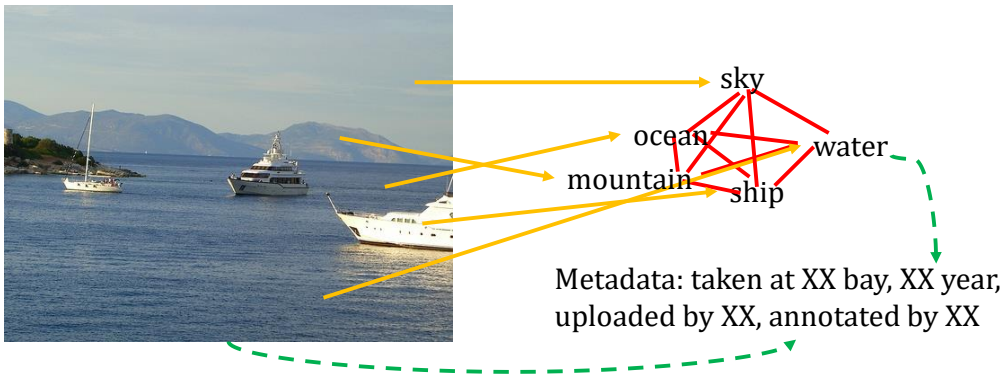


Figure 1.2: The illustration of the main factors in the image annotation.

Dollár & Zitnick 2014, Everingham, Van Gool, Williams, Winn & Zisserman 2010, Chua, Tang, Hong, Li, Luo & Zheng 2009, Huiskes & Lew 2008). Given a set of training images with ground-truth labels, the direct way to uncover the relationship between the image visual content and labels is to train classifiers independently for each label. Hand-crafted image features like SIFT (Lowe 2004), LBP (Ojala, Pietikainen & Maenpaa 2002) and GIST (Oliva & Torralba 2001) combined with classifiers such as random forest (Breiman 2001), SVM (Chang & Lin 2011) and voting (Altman 1992) are widely adopted to fulfill the task. Most recently, deep convolutional neural network (CNN) has shown the advanced ability on various computer vision and pattern recognition tasks, including the image classification (Simonyan & Zisserman 2014, He, Zhang, Ren & Sun 2016) and retrieval (Gordo, Almazán, Revaud & Larlus 2016, Zhao, Huang, Wang & Tan 2015). CNN can obtain high-quality image representations based on the purely supervised learning, and training CNN with a logistic loss function (Guillaumin, Mensink, Verbeek & Schmid 2009) has established a solid baseline for the annotation problem, while multiple works have been focused on modifying the loss function to better fit the problem, such as the pair-wise ranking loss (Gong, Jia, Leung, Toshev & Ioffe 2013). However, by treating each label independently, this line of works only model the relationship between the image visual content and each label but overlook the abundant semantic information carried by the labels themselves.

To bridge the semantic gap, a multi-modal representation is often carried out to learn the image representation as well as the label representation. Canonical correlation analysis (CCA) (Härdle & Simar 2015) and kernel canonical correlation analysis (KCCA) (Rasiwasia, Mahajan, Mahadevan & Aggarwal 2014) based methods project both visual and semantic representations into a latent space to tackle the annotation and retrieval problem. There are also related works leverage the semantic information by capturing the dependencies between label pairs. In general, images with multiple labels have strong correlations among attached labels. Probabilistic graphical models (PGM) such as Markov random field (MRF) (Li 2009) and conditional random field (CRF) (Lafferty, McCallum, Pereira et al. 2001), as well as the widely used recurrent neural network (RNN) (Wang, Yang, Mao, Huang, Huang & Xu 2016), have been proved their efficiencies on capturing high-order label dependencies.

As for the image annotation with metadata information, social networks like Flickr and Instagram record the user interactions with images as the vast amount of metadata, which is presented in various forms. The most common metadata includes the collections, *i.e.* image groups created by users; textual descriptions, *i.e.* tags and captions; as well as user profiles, *i.e.* usernames, locations and friends. As a means to communicate with other users, these metadata can be as informative as the pixel contents (McAuley & Leskovec 2012, Johnson, Ballan & Li 2015) to understand images. There are several lines of works (Ballan, Uricchio, Seidenari & Del Bimbo 2014, Srivastava & Salakhutdinov 2012) conducted by utilising the metadata to assist the annotation, where different types of metadata are studied to be embedded into the annotation framework. In (Johnson et al. 2015), authors propose to nonparametrically use the metadata to generate image neighbours and train an annotation model based on the visual features from the target image and its neighbours.

The goal of this thesis is to study these two types of image annotation problems, namely the generic image annotation and annotation with metadata information. We address the research issues in each problem as below.

1.2 Research Issues

Although the image annotation has achieved significant progress in the past few decades, there are still some unsolved and partially solved issues that can be further explored and addressed.

Deep convolution neural networks (CNN) have demonstrated advanced performance on single-label image classification, and various progress also has been made to apply CNN methods on the image annotation, which requires annotating objects, attributes, scene categories *etc.* in a single shot. Recently, Wang *et al*. (Wang et al. 2016) have shown that the recurrent neural networks (RNN) (Hochreiter & Schmidhuber 1997, Mikolov, Karafiát, Burget, Cernockỳ & Khudanpur 2010) can efficiently capture the high order label dependencies. They unify the CNN and RNN as one framework to exploit the label dependencies at the global level, largely improve the labelling ability. However, predicting small objects and attributes is still challenging for these works due to the limited discrimination of the global visual features.

People sometimes spontaneously assign tags to images when uploading them to the social network, and more tags are attached along with the sharing. Despite these tags provide the semantic illustrations of images to a certain extent, they can be incomplete, imprecise, and biased toward individual perspectives (Li, Uricchio, Ballan, Bertini, Snoek & Bimbo 2016). To reduce the existing tag error for the further retrieval and train the annotation model, some previous works (Zhu, Ngo & Jiang 2012, Liu, Hua, Yang, Wang & Zhang 2009, Li, Snoek & Worring 2009, Li & Snoek 2013, Han, Yang, Ma, Shen, Sebe & Zhou 2014, Zhu, Yan & Ma 2010) are conducted on the tag relevance analysis and refinement. These works explore the tag relevance from different perspectives including the semantic similarities between tags, the visual similarities among image neighbours and the properties of the annotation matrix. The user-provided tag is further refined based on its relevance. The relevant works conducted in (Zhu et al. 2010, Li & Tang 2017) focus on analysing the low-rank property of the annotation matrix. The user-provided annotation matrix and the image feature matrix are used as the model input respectively. However, the image features are not connected to the refined

results in (Zhu et al. 2010), while they are fixed in (Li & Tang 2017). Moreover, images are not isolated subjects; due to the diverse contents and complex styles, some images can be challenging to recognise when neglecting the context. The image metadata such as similar topics and textual descriptions, common friends of users and nearby locations, forms a neighbourhood for each image, which can be used to assist the annotation.

The mainstream methods are proposed for the image annotation of common objects and their attributes and have achieved satisfactory results. However, there are two major drawbacks when they handle the more diverse image annotation. First of all, state-of-the-art methods on the image annotation problem use CNN as the backbone model, which heavily relies on the fine-tuning the pre-trained image representation on the large-scale image dataset ImageNet (Krizhevsky, Sutskever & Hinton 2012). However, the labels of the real-world image can be more obscure and diverse. Even for the same label, the image visual appearance can be very different. Moreover, the domain difference between the training image set and ImageNet can be much larger than current popular datasets. Directly learning the image visual representation from the training set could degrade the performance. Secondly, since the labels of the real-world images can be very diverse, and classic annotation methods tend to treat labels as equally related to the image visual content or rule out ones that are less relevant, which makes their performances unsatisfactory.

1.3 Research Contributions

After researching the above issues, the author has developed corresponding solutions, which are presented in this thesis. These study contributions are showed as blew:

• A Regional Latent Semantic Dependencies (RLSD) model is proposed for the generic image annotation, which effectively captures the latent semantic dependencies at the regional level. The proposed model combines the power of the region based features and the advantages of the RNN based label co-occurrence

models, and achieves the best performance compared to the state-of-the-art models on several benchmark datasets, especially for predicting small objects and visual concepts; (Chapter 3)

• In addition to the RLSD model, an upper bound model (RLSD+ft-RPN) is set up by utilising the object bounding-box coordinates for training. The experimental results show that the RLSD model can approach this upper bound without involving additional bounding-box annotations, which is more realistic in the real world; (Chapter 3)

• For the image annotation on social images, the deep neural network based image annotation, and tag refinement are conducted simultaneously in a weakly-supervised manner. During the training, the user-provided tags are spontaneously refined from 0/1 assignments to the confidence scores, while the trained annotation model can be applied to assign tags to new images; (Chapter 4)

• Different from previously related works that focus on solving the annotation problem at the image level, the annotation and refinement are conducted at both the image level and regional level, which not only enable the visual feature learning to ensure the flexibility and stability but also enhance the ability to model the proposed constraints; (Chapter 4)

• For the annotation on the diverse image set, image visual representation learning is performed based on the metric learning. Different from previous works, the metadata is used to measure similarities among image neighbours. The superior experimental results against hand-crafted features and CNN fine-tuned features indicate the significance of the proposed representation learning methods; (Chapter 5)

• The proposed model annotates all labels via collective knowledge from image neighbours, including the visual representation learning and the semantic embedding, which is crucial for the real-world image annotation, since it focuses on not only the visually related labels but also the semantical labels which are related to the events; (Chapter 5)

• A heritage image collection is collected as the benchmark dataset and the proposed model is grounded on it. Experimental results show that our model

achieves the best performance against the state-of-the-art annotation methods; (Chapter 5)

• Propose to annotate images by exploring their neighbourhoods, which are decided by the metadata similarity. A neighbourhood graph network is established to model the correlations between each image and its neighbours. The learned graph representation is used to assist the annotation of the target image; (Chapter 6)

• To accurately capture the relevant regions in each image neighbour that are beneficial for understanding the target image, a co-attention mechanism is introduced to obtain the node representations in the graph. (Chapter 6)

1.4 Thesis Structure

The thesis is structured as follow:

Chapter 2 introduces the mainstream annotation methods for tackling the annotation problems. Two types of image annotation are studied, namely the generic image annotation and annotation with metadata information. The mainstream annotation methods include the classification-based methods and voting-based methods to leverage the visual content and labels, the multi-modality design for visual and semantic representations, the graphical methods to model the label dependencies and the representative works on annotation with different types of metadata.

Chapter 3 briefly reviews the related work for the generic image annotation and explains the motivation of the proposed method. A convolutional localisation layer is first introduced to capture the image regions that potentially have multiple semantic labels, then the LSTM-based label sequence prediction model is adopted to capture the label correlations inside each region. The max-pooling operation to fuse the regional outputs and the training strategies are carried out last.

Chapter 4 briefly reviews the related work of analysing the user tag relevance regarding the social image content and utilising these tags for the image annotation. Then a weakly-supervised neural network is proposed to learn the image annotation model and refine the user-provided tags simultaneously. The selection

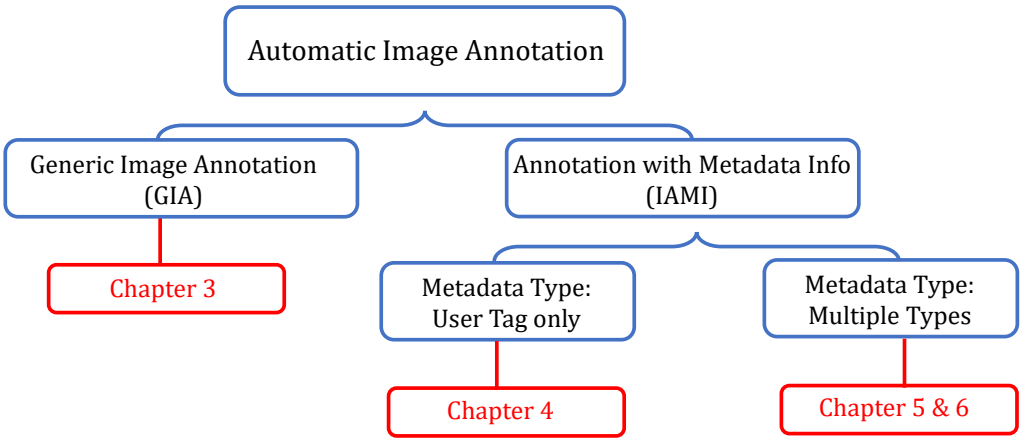


Figure 1.3: The illustration of the thesis structure.

of image regions is first introduced, and the architecture of the regional neural network is established. Then three constraints are proposed, namely the visual consistency, semantic dependency, and user-error sparsity. Finally, the training and prediction details of the tag refinement and image annotation are given.

Chapter 5 introduces the proposed model on using collective knowledge to solve the diverse image annotation problem of a specific domain, which is grounded on the heritage image collection. The metadata similarity measurement and the metric learning for the image visual representation are introduced first, followed by descriptions of the label relevance analysis and the semantic representation. The final annotation model and training details are summarised at last.

Chapter 6 introduces how to represent the image neighbourhood as a graph to assist the image annotation by exploring the image metadata. The first part describes how to locate image neighbours by measuring their metadata similarities, then the architecture of the neighbourhood graph and the node representation by the co-attention mechanism is introduced, followed by the descriptions of the instance-level attention model. The training and prediction details are given at last.

The relationship among Chapter 3, 4, 5 and 6 is illustrated in Figure 1.3.

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

Chapter 2

Related Works

2.1 Summary

In this chapter, we review the mainstream methods proposed for the automatic image annotation problem. As we mentioned before, for the generic image annotation (GIA), there are two major factors: the visual relevance and semantic dependencies. Moreover, we also investigate the methods of exploring the metadata relevance for the annotation with metadata information (IAMI).

The mainstream methods for the generic image annotation are roughly grouped into five parts: classification-based methods, voting-based methods, multi-modality design, graphical inference methods, and deep neural network-based methods. Here we separate deep neural networks as an independent category considering their advanced end-to-end learning ability and superior overall performance. Then we review the annotation methods, which utilise the additional metadata information. It is worth noting that the visual relevance and semantic dependency are still playing important roles in the IACI, the methods we reviewed for the IACI emphasise on how to utilise the metadata to assist the annotation.

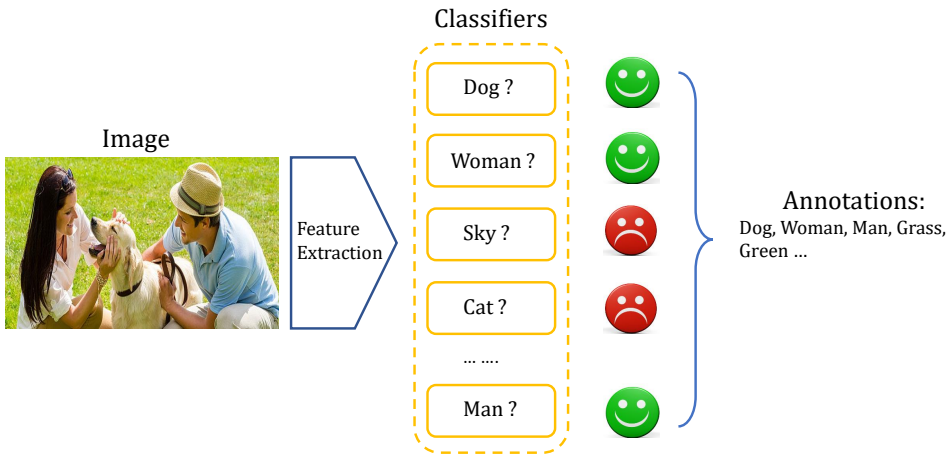


Figure 2.1: The illustration of the classification-based methods. The labels are represented as separate classifiers. When annotating the new image, all the classifiers will be applied, and the fusion of all the classifiers' outputs is the final annotation.

2.2 Generic Image Annotation

2.2.1 Classification-Based

Generic image annotation, as a fundamental computer vision problem, has been studied for decades. Multiple image datasets have been proposed for the research purpose, such as the PASCAL VOC (Everingham et al. 2010), MS-COCO (Lin et al. 2014), Flickr (Huiskes & Lew 2008) and NUS-WIDE (Chua et al. 2009). One way to tackle the problem is to directly model the relationship between the image visual content and associated labels as classifiers. The classification-based methods represent labels as separate classifiers. Each classifier is trained on the image features with the corresponding label while taking other images as negative samples. When annotating the new image, all the classifiers will be applied, and the output from each classifier indicates whether this image contains the corresponding label. We show an illustration of the classification-based methods in Figure 2.1.

Image features can be chose from traditional hand-crafted features (Lowe 2004, Ojala et al. 2002, Oliva & Torralba 2001, Lu & Wang 2015) to the advanced

CNN feature (Simonyan & Zisserman 2014, He et al. 2016), while the classifier can be a wide range of options, including the random forest (Breiman 2001), SVM (Chang & Lin 2011) and neural networks with suitable loss functions (Gong et al. 2013). These methods establish the baselines of the image annotation with satisfactory results, while some works focus on improving the baseline performances by leveraging image regions. In (Xue, Zhang, Zhang, Wu, Fan & Lu 2011), Xue et al. use the multi-label multi-instance method to explore the image features both at the image level and regional level.

More specifically, In (Gao, Fan, Xue & Jain 2006), authors propose a multi-resolution grid framework to extract the image features and adopt a hierarchical boosting algorithm to scale up SVM training in the high dimensional feature space. They use the boosting algorithm to select the most effective SVM classifiers to achieve more accurate predictions. Different from (Gao et al. 2006), in (Chang, Goh, Sychay & Wu 2003), authors use the BPM to soft annotate images for the retrieval task. They adopt the BPM to train classifiers for each label separately, with hand-crafted features as inputs. Although only a small set of images are used for training, the classifiers are applied to all the images to gain the soft assignments. Then the active learning is employed to further improve the initial annotation results.

Instead of using the holistic feature to represent an image, given the multi-label property of the image, the attention mechanisms are applied for the image annotation. In (Huang, Zhang, Li, Mei, He & Zhao 2017, Zhu, Li, Ouyang, Yu & Wang 2017), the attention mechanism is adopted to capture the correlations between the image content and associated labels. Different from these works, we not only apply the attention for the target image regions but also use a co-attention mechanism to guide the attentions between the target image and its neighbours.

2.2.2 Voting-Based

Another feasible way to annotate image is the voting-based approaches. Given a test image, they annotate it by searching its nearest neighbour within the training set. Neighbours are usually allocated in a specific feature space, based on the

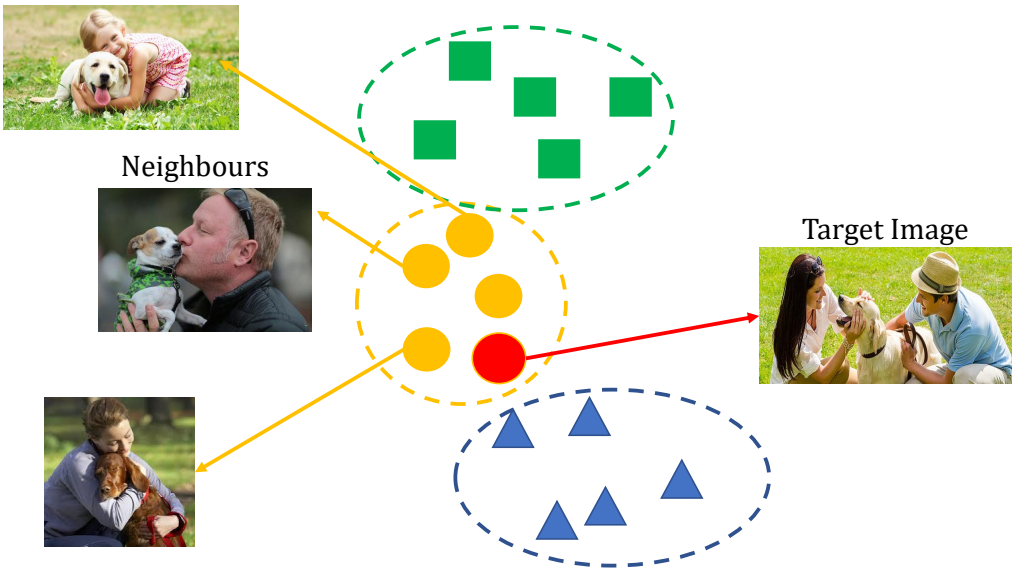


Figure 2.2: The illustration of the voting-based methods. The labels are transferred from the image neighbours within the training set via the neighbour search.

pre-defined distance measurement and feature extraction method. In this case, the selection of the image feature and the definition of the distance are crucial. Moreover, when the training set is large-scale, the selection of the neighbour search method is also needed to be deliberated. We show an illustration of the voting-based methods in Figure 2.2.

In (Makadia, Pavlovic & Kumar 2008), authors summarise the prior works on voting-based annotation methods and propose a baseline model. They utilise different types of hand-crafted features: RGB, HSV and LAB histogram with Gabor and Haar Wavelets. They combine multiple basic distance measurements via the equal joint contribution, *i.e.* all the distance contributions are equal, and the L1-penalised logistic regression *i.e.* lasso, which selectively drops out some redundant distance. As for the search strategy, they adopt the modified k nearest neighbour approach to transfer the labels. Their experiments on three benchmark datasets show that the proposed framework can serve as a general baseline for the annotation problem. In (Wang, Jing, Zhang & Zhang 2008), the authors propose a hybrid voting-based method for the web image annotation. With the internet users

involved, the description of an image becomes more abundant and yet with noise. Authors divide the voting method into four steps. The first step is the content-based retrieval. Given a test image, its neighbours are retrieved and ranked via the visual similarity. Global and regional image features are adopted during the retrieval. The second step is the text-based keyword retrieval. For each image query, a text-based keyword retrieval is applied to rank all the related keywords regarding each training image. The obtained ranked results are merged as the third step, and further refined via the RWR algorithm. A graphical model is constructed during the refinement. Experiment results not only show the framework guarantees the efficiency and effectiveness, but also imply the advantage of voting based methods on the web image dataset. Similar to (Wang et al. 2008), M Batko *et. al.* propose to solve the web image annotation via the combination with the external knowledge to further improve the voting accuracy in (Batko, Botorek, Budikova & Zezula 2013). A hierarchical annotation pipeline is proposed, with the human feedback on the visual relevance of annotations is considered.

2.2.3 Multi-Modality Design

To efficiently exploit the abundant semantic information carried by labels, several approaches (Uricchio, Ballan, Seidenari & Bimbo 2017, Ke, Zhou, Niu & Guo 2017, Gong, Ke, Isard & Lazebnik 2014, Murthy, Maji & Manmatha 2015*a*, Gong, Tao, Maybank, Liu, Kang & Yang 2016) have been proposed to design a multi-modal representation of the image and its labels. The CCA (Härdle & Simar 2015) and KCCA (Rasiwasia et al. 2014) based methods and their variations are widely used in the image annotation and retrieval. In (Hwang & Grauman 2012), Hwang et al. use the KCCA to leverage the importance of textual objects for the image annotation and retrieval. They reveal implied cues about the object importance based on how people naturally annotate images with the text and then translate those cues into a dual-view semantic representation. In (Gong et al. 2014), a third view of the category or concept is added to the CCA to capture the high-level image semantics, which improves the retrieval performance. In (Murthy et al. 2015*a*), Murthy et al. propose to combine the CNN visual representation

with the word embedding by using the CCA, while in (Uricchio et al. 2017), authors propose a label propagation framework based on the KCCA to tackle the annotation problem regardless whether the training set is annotated by experts. In (Gong, Tao, Maybank, Liu, Kang & Yang 2016), authors design a multi-modal curriculum learning (MMCL) strategy to tackle the semi-supervised image annotation problem.

2.2.4 Graphical Inference

Viewing the image annotation as multiple single-label image classification problems only consider the visual relevance between the image content and labels, and the relationships inside the label set are ignored. Voting-based methods involve the semantic dependency but only at a limited level. There are also series of works focus on uncovering semantic information by modelling the semantic dependency within the label set, where the probabilistic graphical models (PGM) (Bradley & Guestrin 2010, Zhang & Zhang 2010, Yang, Huang, Yang, Liu, Shen & Luo 2013, Tan, Shi, van den Hengel, Shen, Gao, Hu & Zhang 2015, Gong, Tao, Yang & Liu 2016) are favoured by the researchers. The most common way is to model the label co-occurrence with the pairwise compatibility probability and adopt the PGM to generate joint label probability during the inference. The graphical models are always combined with the classification method, by exploiting the semantic dependency, these models further improve the classification results. For example, if the label 'ship' exists in an image, then the label 'ocean' is also highly likely appears in the same image, where there is almost no chance the 'tiger' could be in this image too.

For the graphical model based annotation methods, different graph structures can be used. In (Li, Liu, Wang, Bingyuan Liu, Gao, Wu, Song, Ying & Lu 2015), ChowLiu tree is established based on the labels mutual dependency. In (Zhang & Zhang 2010), joint label probability is exploited by a directed acyclic graph, while a chain rule is adopted in (Dembczynski, Waegeman & Hüllermeier 2012). The above methods only construct the graph without the parameter learning together, while in (Wu, Ye, Zhang, Chow & Ho 2015), this drawback is considered and

modified, but it is still time-consuming and not scalable given lots of parameters are used for the graph establishment. In order to address the above issue, in (Tan et al. 2015), M Tian *et. al.* propose a principle way to consider the loss function along with the image features and labels. They formulate the annotation into a max-margin framework and then transform it into a convex optimising problem. The tricky part in the PGM is capturing the label dependency while identifying the independent or irrelevant labels at the same time. Authors propose a clique generating machine with the convex programming model to learn sparse graph structures. The cutting plane algorithm is adopted to active a group of cliques. Another classic graphical method is proposed by in (Li, Zhao & Guo 2014). Authors construct auxiliary labels in the output space to solve the annotation. Two types of space are used in their method. First one is the original label space, another one is the label pair space, which is obtained by a tree-structured graph, maximum spanning is adopted to form the tree. After the model is established, label pairs are considered as new labels to enhance the original label space, and pair-wise training can also be applied. Results show that their method achieves high efficiency and accuracy. In addition, they also conduct experiments with dense graphs, which demonstrate the models scalability.

The graph is an optimal representation of the structured information. In a graph, nodes are connected by edges, which indicate the pair-wise relationships between corresponding nodes. In the Graph Neural Network (GNN) model (Scarselli, Gori, Tsoi, Hagenbuchner & Monfardini 2009), the neighbourhood information is propagated through the graph and the hidden state of each node is updated by the multi-layer perceptrons (MLP), while in (Li, Tarlow, Brockschmidt & Zemel 2015), a recurrent gating mechanism is adopted to update the graph hidden states and extended to output sequences, noted as the Gated Graph Sequence network (GGNN). In (Kipf & Welling 2016), a scalable approach Graph Convolutional Network (GCN) is proposed to learn on the graph-structured data via convolutional operations. In (Wang & Gupta 2018), the video classification is studied by modelling the frames as the spatial-time graph and applying the GCN to infer the video category, the region information and time sequences are used to establish the

graph edges. In (Veličković, Cucurull, Casanova, Romero, Liò & Bengio 2017), the self-attention mechanism is introduced into the GCN to compute the node representation. Each node is embedded by attending over its neighbours. Different from previous works, we ground the image neighbourhood by leveraging their metadata and apply the GCN to infer on the proposed neighbourhood graph for the image annotation problem, and a novel co-attention mechanism is introduced to model the correlations between the target image and its neighbours.

2.2.5 Deep Neural Network-Based

Deep learning has become a popular term in today's research community, which stands for the various machine learning techniques involving deep neural networks to cope with the big data era. Recent advanced progress in the image annotation is made based on the powerful deep convolutional neural networks (CNN), which try to model the high-level abstractions of the visual data by using architectures composed of multiple non-linear transformations. The CNN relies on the pure supervised end-to-end learning mechanism to achieve high classification accuracy on large-scale datasets. Since several approaches are proposed to expand the single-label classification network to the multi-label image annotation problem, the introduction of popular single-label classification models are given first.

The first widely used CNN model for the image classification is proposed in (Krizhevsky et al. 2012), known as AlexNet, it establishes the fundamental CNN architecture and has become the baseline model for various vision applications. The architecture of AlexNet is different from the conventional neural networks; it contains eight existed layers: five convolutional layers and three fully-connected layers, while it also contains several novel layers: ReLu Nonlinearity layer, local response normalisation, and overlapping pooling layer. The AlexNet is the first shows that the CNN can be used to deal with the large-scale image datasets and challenging vision tasks based on the purely supervised end-to-end training. In (Simonyan & Zisserman 2014), authors propose to fine-tune the layer parameters using smaller convolutional filters and expand the convolutional depth of the CNN architecture, which shows a significant performance beyond the prior

model. Two CNN models are established, known as VGGNet-16 and VGGNet-19, which are widely used in different vision tasks, as well as in our proposed frameworks. Their experiments demonstrate that the deeper the network is, the more efficient representation the network can get. Inspired by their discoveries, more efficient network designs have been proposed, including the network in network (NIN) (Lin, Chen & Yan 2013), deep residual learning (He et al. 2016) and densely connected network (Huang, Liu, Van Der Maaten & Weinberger 2017).

Inspired by the advanced abilities of CNN models on the single-label image classification, Gong *et al*. (Gong et al. 2013) combine top-k ranking objectives with a CNN architecture to tackle the annotation. By defining a weight function for pair-wised ranking labels, they minimise the loss function so that positive labels are ranked higher than negative ones, while in (Wei, Xia, Huang, Ni, Dong, Zhao & Yan 2014), a regional solution is proposed. Authors proposed to use the image proposals instated of the whole image as the input. The objectness method BING is adopted to generate around one thousand proposals per image; then they apply the normalise cut to cluster the proposals into ten groups based on the overlap between the proposals. Five proposals in each group are selected to represent each group. Then the representative proposals are sent to a shared CNN model, and they max out the outputs for each proposal as the final score. In this way, the region information is considered, which can assist the prediction of the small objects in the image. Most recently, the recurrent neural network (RNN) has been applied to capture the sequential dependencies, which is suitable for the image to language problem including the annotation and captioning. In (Wang et al. 2016), Wang et al. have shown that the RNN can efficiently capture high order label dependencies. They define each image ground-truth as an ordered sequence of labels and use the CNN-RNN as one unified framework to annotate images in an end-to-end fashion.

2.2.6 Comparisons

Most of these classification-based methods focus on modelling the visual relevance, while neglecting the semantic information carried by the labels. Compared

to the classification-based methods, the voting-based methods do not require the training of classifiers. However, they are sensitive to the metric used to allocate the neighbours. Meanwhile, multi-modality methods tend to embed both the visual and semantic information, the hand-crafted embedding methods are often designed. Different from the multi-modality, graphicial models capture the semantic information as the relevance between the visual content and labels. With the advanced performance in representation learning, deep neural network-based methods address the annotation problem by fusing the aforementioned advantages into the neural network.

2.3 Annotation with Metadata

Images are not isolated subjects, due to the diverse contents and complex styles, some images can be challenging to recognise when neglecting the context. In this section, we review the related works for the annotation with metadata information, namely the metadata. Since the majority of images comes from the social network, the user-provided tags become the most commonly existed metadata, and various experiments are conducted to investigate how to utilise these tags to assist the annotation. In addition, other types of metadata are also studied to be embedded into the annotation framework.

2.3.1 User-Tag Relevance Analysis

The most commonly used metadata for the image annotation is a set of user-provided tags, where a multi-modal representation can be learned for the image feature and associated tags. For example, in (Chen, Xu, Tsang & Luo 2012), the image feature is enriched by adding the additional tag feature, which is obtained by the SVM prediction, while in (Pereira, Coviello, Doyle, Rasiwasia, Lanckriet, Levy & Vasconcelos 2014) and (Ballan et al. 2014), the authors design a latent multi-modal space to tackle the annotation problem by the CCA and KCCA. Besides, matrix-based methods are also widely used for the tag refinement, since the user-provided tag, as one type of semantic information, is subject to the low-rank

property. Therefore, the initial annotation matrix can be decomposed into an ideal tagging matrix with a sparse user-error matrix (Zhu et al. 2010, Li & Tang 2017). In (Zhu et al. 2010), they use the pre-defined visual and semantic similarities to assist the process. However, these methods cannot connect the visual features with the annotation results, which makes it unable to perform the annotation. In (Li & Tang 2017), a three-layer network architecture is proposed to bridge the semantic gap. However, since it cannot perform the feature learning and use the matrix as the input, it is less flexible and stable when dealing with the large-scale image sets.

Given the diversity of the tag set in the image training set, tag relevance can guide the annotation process in an effective way. Therefore, we review related works on the tag relevance analysis. In early works (Zhu et al. 2012, Xu, Wang, Hua & Li 2009, Sigurbjörnsson & Van Zwol 2008), the tag relevance is estimated based on the semantic similarities between tag pairs. In (Sun & Bhowmick 2010), Sun et al. propose two distance metrics to quantify the tag relevance of the image visual content, which is measuring the tag visual relevance at a global level. In (Li, Snoek & Worring 2009), Li et al. use the hand-crafted visual feature similarity to find each image's neighbours and employ the KNN method to vote the tag relevance to the image content, which is measuring the image-specific tag relevance. In (Liu et al. 2009), the tag relevance is initially leveraged by the kernel density estimation; then the random walk is employed to refine the relevance based on the visual and semantic information. In (Guillaumin et al. 2009, Li & Snoek 2013), nearest neighbour voting is used to estimate the tag visual relevance and annotate images.

2.3.2 Other Metadata Types

In addition to user-provided tags, there are investigations conducted on other types of metadata. The user information ranges from the simplest user identities, annotation preferences to the user reliabilities. Under the assumption that tags chosen by the majority of the users to label the visually similar images are more likely relevant to the image visual contents, in (Li, Snoek & Worring 2009), authors utilise

user identities to ensure that learning examples come from distinct users. A similar idea is reported in (Kennedy, Slaney & Weinberger 2009), which finds visually similar image pairs with matching tags from different users. (Ginsca, Popescu, Ionescu, Armagan & Kanellos 2014) proposes to improve the image retrieval accuracy by ranking the images uploaded by users with good credibility first. The credibility of a user is defined by counting matches between the user-provided tags and machine labels predicted by visual concept detectors. In (Sawant, Datta, Li & Wang 2010), personal annotation preference is considered in the form of tag statistics computed from images that a user has uploaded in the past. These past images are used in (Liu, Li, Tang, Jiang & Lu 2014) to learn a user-specific embedding space. Timestamps are exploited for the time-sensitive image annotation and retrieval (Kim & Xing 2013), where the connection between image occurrence and various temporal factors is modelled. In (McParlane, Whiting & Jose 2013), time-constrained tag co-occurrence statistics are considered to refine the outputs of visual classifiers for the annotation. GPS and EXIF are used in (Joshi, Luo, Yu, Lei & Gallagher 2011, Li, Crandall & Huttenlocher 2009) to annotate the landmark images, while in (Stone, Zickler & Darrell 2008) friendships are contributed to the label recommendation. In (McAuley & Leskovec 2012, Long, Collins, Swears & Hoogs 2018) multi-type of textual features and network linkage information are used to construct a CRF-based inference model for the image annotation. Johnson *et al.* propose to allocate image neighbours by nonparametrically exploring the metadata in (Johnson et al. 2015), and incorporate features from the target image and its neighbours to annotate.

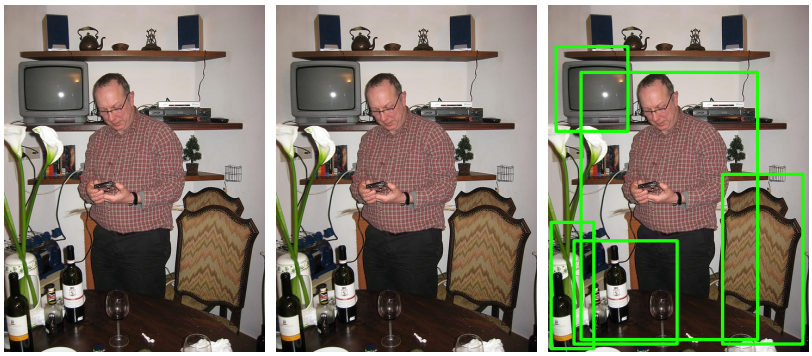
Chapter 3

Regional Latent Semantic Dependencies

3.1 Introduction

As we addressed before, for the generic image annotation (noted as the 'image annotation' for the expression simplicity in this chapter), the key issue is bridging the semantic gap (Han et al. 2014) existing between the image visual content and multiple semantic labels. There exists not only image-label visual relevance but also abundant semantic dependencies within the label set. With the availability of the large-scale images and the enrichment of annotations, image annotation has drawn lots of attentions (Guillaumin et al. 2009, Guo & Gu 2011). Inspired by the advanced performance of the deep neural network, especially convolutional neural networks (CNN) (He et al. 2016, Krizhevsky et al. 2012), various efforts have been made to apply the neural network on this problem (Wang et al. 2016, Wei et al. 2014).

The most straightforward approach is to treat the image annotation as several separate single-label classification problems, and train the independent classifier for each label with a cross-entropy (Guillaumin et al. 2009) or ranking loss (such as WARP (Gong et al. 2013)). Wei *et al.* (Wei et al. 2014) provide a regional solution allow predicting labels independently at the regional level. However,



Ground truth:
person; chair; dining table;
bottle; wine glass; vase; TV

Multi-CNN: person; chair; TV
CNN+LSTM: person; chair;
dining table; TV

RLSD:
person; chair; dining table;
bottle; **wine glass**; **vase**; TV

Figure 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' *etc*.

it is difficult for them to model the label dependencies between different labels. Intuitively, images with multiple labels usually contain strong correlations among the labels, for example, 'ocean' and 'ship' usually appear in the same image, while 'ocean' and 'cat' normally never occur together. To conveniently explore the label dependencies, the probabilistic graphical models (PGM) are usually employed in the previous works (Li et al. 2014, Tan et al. 2015, Tang, Li, Qi & Chua 2010).

Most recently, Wang *et al.* (Wang et al. 2016) have shown that the recurrent neural networks (RNN) (Hochreiter & Schmidhuber 1997, Mikolov et al. 2010) can efficiently capture the high order label dependencies. They unify the CNN and RNN as one framework to exploit the label dependencies at the global level, largely improving the labelling ability. However, predicting small objects and attributes is still challenging for these works due to the limited discrimination of the global visual features.

In this work, our main contribution is that we propose a Regional Latent Semantic Dependencies (RLSD) model for the image annotation, which effectively

captures the latent semantic dependencies at the regional level. The proposed model combines the power of the region based features and the advantages of the RNN based label co-occurrence models, and achieves the best performance compared to the state-of-the-art annotation models on several benchmark datasets, especially for predicting small objects and visual concepts. Figure 3.1 shows an example output of our proposed RLSD model compared with the baselines models. We can see that the Multi-CNN and CNN+LSTM both fail to predict the 'bottle', 'vase' and 'wine glass' in the image due to their small size, while our model efficiently predicts them along with other large objects.

The framework of the proposed model is shown in Figure 3.2. An input image is first processed through a CNN to extract convolutional features, which are further sent to an RPN-like (Regional Proposal Network) localisation layer. Different from the conventional RPN in the object detection framework (such as faster R-CNN (Ren, He, Girshick & Sun 2015)) which tries to predict the proposals with a single object inside, our localisation layer is designed to localise the regions in an image that may contain multiple semantically dependent labels. These regions are encoded with a fully-connected neural network and further sent to an RNN, which captures the latent semantic dependencies at the regional level. The RNN unit sequentially outputs a multi-class prediction, based on the outputs of the localisation layer and the outputs of previous recurrent neurons. Finally, a max-pooling operation is carried out to fuse all the regional outputs as the final prediction.

In addition, we set up an upper bound model (RLSD+ft-RPN) by utilising the object bounding-box coordinates for training. Our experimental results show that our model can approach this upper bound without involving additional bounding-box annotations, which is more realistic in the real world.

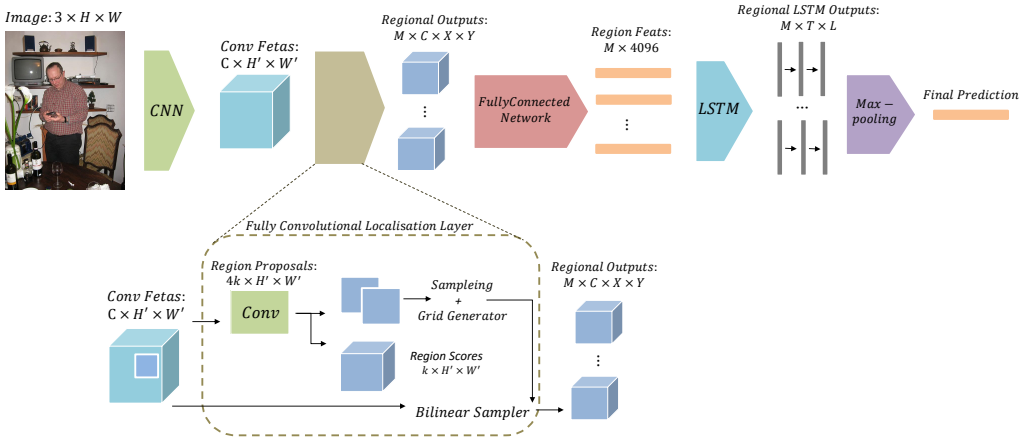


Figure 3.2: Our proposed Regional Latent Semantic Dependencies model. An input image $3 \times H \times W$ is first processed through a CNN to extract convolutional features, which are further sent to an RPN-like fully convolutional localisation layer. The localisation layer localises the M regions in 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.

3.2 The RLSD Model

3.2.1 Framework Overview

The critical characteristic of our proposed model is that it can capture the regional semantic label dependencies. The novelty lies in the fact that this is achieved by a localisation architecture, followed by a few LSTMs (Long-Short Term Memory). The purpose of the localisation layer is to localise the regions that contain multiple highly-dependent labels, while the LSTMs are employed to characterise the latent semantic label dependencies sequentially. A max-pooling operation is carried out to fuse all the regional outputs finally. Figure 3.2 shows the entire network of our proposed model.

In the following part, the localisation layer is first introduced in Section 3.2.2,

and the LSTM based label sequence prediction model is described in Section 3.2.3. The final max-pooling operation and the loss function is summarised in Section 3.2.4. The model initialisation and some training details are given in Section 3.2.5.

3.2.2 Localising Multi-label Regions

To explore the image at the regional level, we need to generate the regions that potentially contain the multiple objects and visual concepts. Hence, the first component of our proposed model is to localise these regions. The conventional object proposal algorithms, such as Selective Search (Uijlings, van de Sande, Gevers & Smeulders 2013), Objectness (Alexe, Deselaers & Ferrari 2012), BING (Cheng, Zhang, Lin & Torr 2014) and MCG (Pont-Tuset, Arbelaez, Barron, Marqués & Malik 2017) are ruled out since these methods only focus on predicting single object proposals, which means that a proposed region normally only contains one single object. Instead, Johnson *et al*. (Johnson, Karpathy & Fei-Fei 2016) propose a fully convolutional neural network, extended from the Region Proposal Network (RPN) (Ren et al. 2015), to localise regions that can be described by a sentence, instead of a single label. Therefore, the proposed regions in (Johnson et al. 2016) normally enjoy bigger label density, as well as the label complexity. Inspired by their work, we develop our approach of generating region proposals that are tailored for the image annotation.

Convolutional Features as Input

Since the convolutional layers of the CNN still preserve the spatial information of an image, which is necessary for us to explore the semantic dependencies at the regional level, we use them to extract image features. Specifically, we use the VGGNet (Simonyan & Zisserman 2014) convolutional layer configuration, which consists of 13 convolutional layers with 3×3 kernel size and five max-pooling layers with 2×2 kernel size. The output of the last convolutional layer is used as image features. Given an input image with size $3 \times H \times W$, the convolutional

features will be $C \times H' \times W'$, where $H' = \lfloor \frac{H}{16} \rfloor$, $W' = \lfloor \frac{W}{16} \rfloor$, and $C = 512$, same as in the VGGNet setting. The convolutional features are further sent to the localisation layer to generate the region proposals that we are interested.

Fully Convolutional Localisation Layer

The input of the localisation layer is the convolutional features extracted from the last step, while outputs are numbers of spatial regions of interest with a fixed-sized representation for each region.

Anchors and Regression By referring to (Johnson et al. 2016, Ren et al. 2015), we predict region proposals by regressing offsets from a set of generated anchors. Specifically, each point inside the convolutional feature map is projected back into the original image ($H \times W$) and further used as a center to generate k different aspect ratio anchor boxes. Each anchor box is sent into a fully convolutional network to produce the predicted box scalars and a confidence score. The fully convolutional network consists of 256 convolutional filters with 3×3 kernel size, a ReLU layer and a final convolutional layer with $(4 + 1) \times k$ filters, where four stands for the number of the box scalars, and one stands for the confidence score. We set $k = 12$ in our proposed models. As for the bounding-box regression from the anchors to the region proposals, we refer to (Ren et al. 2015) for the parameterisation. We apply the log-space scaling transformations on the anchor box, which means given an anchor box's parameters (a_x,a_y,a_w,a_h), where (a_x,a_y) is the center of the anchor box, and a_w, a_h stands for the width and height of the anchor box respectively, we generate the region coordinates $b = (b_x,b_y,b_w,b_h)$ by following volumes:

$$b_x = a_x + t_x a_w \qquad b_y = a_y + t_y a_h \tag{3.1}$$

$$b_w = a_w exp(t_w) \qquad b_h = a_h exp(t_h) \tag{3.2}$$

The scalars t_x,t_y,t_w,t_h are predicted by our model. Smooth L_1 norm is employed as the loss function to regress the region location. This loss is only used when we finetune the localisation layer in the upper-bound model RLSD+ft-RPN.

Given the ground-truth coordinates $g = (g_x,g_y,g_w,g_h)$, the loss function is defined as:

$$L(b,g) = \sum_{i \in x,y,w,h} Smooth_{L_1}(b_i,v_i) \tag{3.3}$$

where

$$Smooth_{L_1} = \begin{cases} 0.5x^2 & if|x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{3.4}$$

Box Sampling and Bilinear Interpolation A sampling mechanism is employed here to subsample the generated region proposals since it is expensive to send all the proposals to the further LSTM-based label generation step. By referring to (Johnson et al. 2016, Ren et al. 2015), a $M = 256$ size minibatch is sampled. The regions with the top $M/2$ highest confidence score are considered as the positive samples, and the lowest $M/2$ regions are negative. In RLSD+ft-RPN, A box with IoU (Intersection of Union) which is larger than 0.7 ratio of the ground-truth region is considered as the positive sample and less than 0.3 as the negative. We also restricted that at most of the half of the boxes in one minibatch are positive samples and the other half are negative. During the test stage, non-maximum suppression is used to select the top M highest ranked proposals. In practical, the entire image is also added to the proposals, since some labels are related to the whole image.

To ensure that the region proposal features can be accepted by the fully-connected layer and gradients can be back propagated to both the input features and box coordinates, the bilinear interpolation is used to replace the ROI pooling layer in (Ren et al. 2015). We refer to the bilinear sampling operation in (Johnson et al. 2016), which results in $M \times C \times X \times Y$ feature maps for the top M region proposals, where $C = 512$ is the VGGNet convolutional feature map size, and X, Y are the bilinear sampling grid size. In our case, we set $X = Y = 7$, referring to (Jaderberg, Simonyan, Zisserman et al. 2015, Johnson et al. 2016).

Encoded by a Fully-Connected Network

After the regional features as $M \times C \times X \times Y$ are obtained, they are sent to a fully-connected network, which is formed by two 4096-d fully-connected layers and regularised by dropout. Features from each region are flattened into a vector and passed through this fully-connected network. Thus, each proposal region is encoded as a feature vector v with 4096 dimensions. All the regions' fully-connected features form a minibatch $V = [v_1, v_2, ...v_i, ...v_M]$ with the size of $M \times 4096$, where i indicates the i^{th} region proposal.

Figure 3.3 shows some examples of the comparison results between our localisation layer proposed regions and the MCG (Pont-Tuset et al. 2017) produced object proposals. The bounding boxes generated by our model are normally bigger, and some of them contain multiple objects. Therefore, our model not only can explore the sufficient label dependencies but also can outperform the current methods in predicting small objects and visual concepts. To show the effectiveness of the localisation layer, we set a baseline model that uses the MCG (Pont-Tuset et al. 2017) to replace our multi-label region localisation layer for the further annotation. More details can be found in the Section 3.3.

3.2.3 An LSTM-based Multi-Label Generator

To capture the latent semantic dependencies existed in those regions, we employ LSTMs to generate the sequence of label probability distributions in every single region.

The LSTM is a memory cell which encodes the knowledge at every time step for what inputs that have been observed up to this step. Figure 3.4 shows the basic structure of LSTM. We follow the model used in (Hochreiter & Schmidhuber 1997). Letting σ be the sigmoid nonlinearity, the LSTM updates for time step t

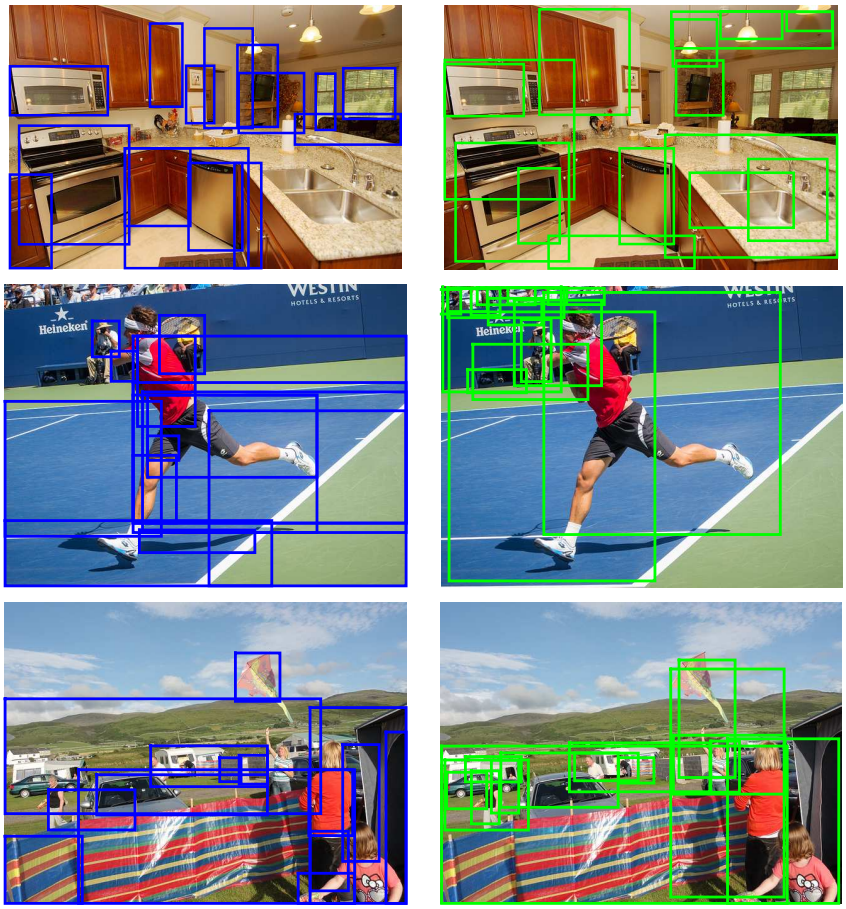


Figure 3.3: The comparison results between the Top-15 regions generated by MCG (left) and our localisation layer (right). Some of our generated regions contain multiple objects, for example, the generated regions contain the object of 'oven/microwave/kitchenwares', 'person/tennis racket,' and 'person/kite/car' altogether.

given inputs x_t, h_{t-1}, c_{t-1} are:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{3.5}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{3.6}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{3.7}$$

$$g_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{3.8}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{3.9}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3.10}$$

$$p_t = \text{softmax}(h_t) \tag{3.11}$$

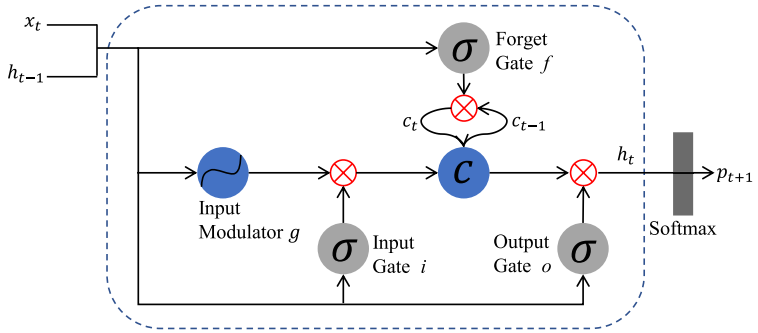


Figure 3.4: The structure of LSTM.

Here, i_t, f_t, c_t, o_t are the input, forget, memory, output state of the LSTM. The various W matrices are trained parameters and $\odot$ represents the product with a gate value. h_t is the hidden state at time step t and is fed to a Softmax, which will produce a probability distribution p_t over all labels.

Given a region feature vector v, we set $x_0 = W_{ev}v$, where W_{ev} is the learnable region features embedding weights. Following the equations (3.5) to (3.11), it gives us an initial hidden state h_0 which can be used in the next time step. From $t = 1$ to $t = T$, we set $x_t = W_{es}S_t$ and the hidden state h_{t-1} is given by the previous step, where W_{es} is the learnable label embedding weights. T is the numbers of the label in a region and S_t is the input label at time step t. In practical, a label is represented as a one-hot vector, where one indicates the label exists, and 0 elsewhere. In our RLSD model, since only the global multi-label ground-truth are provided and no regional ground-truth can be used in the training stage (as well as in the testing), we call the S_t as a latent label, which can be obtained by the following equation:

$$S_t^i = \mathbb{1}\{p_{t-1}^i = \max(p_{t-1})\} \tag{3.12}$$

where $\mathbb{1}$ is an indicator function and S_t is a one-hot vector that index $i = 1$, and 0 elsewhere. i is the index of the maximum value of the probability distribution p_{t-1} over all the labels, which is computed by the LSTM feed-forward process at the previous $t - 1$ time step. After all the labels in a region are predicted, an 'END' label is added to finalise the prediction.

Giving all the M region features in a minibatch (all the regions in a minibatch come from the same image) into the LSTM model, we gather the prediction p_{tm} at each time step t, on each region m, to form a tensor with the shape $M \times T \times L$, where L is the label size of a dataset. If a region label length is less than T, we will pad 0.

3.2.4 Max-pooling and Loss Function

To suppress the possibly noisy prediction on some region proposals or at certain time steps, a cross region and time max-pooling is carried out to fuse the outputs into one integrative prediction. Compared to other pooling methods such as average-pooling, the max-pooling is more suitable to eliminate the possible prediction noise. Suppose p_{tm} is the output prediction of region m at time step t and $p_{tm}^j(j = 1,..., L)$ is the j^{th} component of p_{tm}. The max-pooling in the fusion layer can be formulated as:

$$p^j = \max(p_{11}^j,...,p_{1M}^j,p_{21}^j,...,p_{2M}^j,...,p_{T1}^j,...,p_{TM}^j) \tag{3.13}$$

where p^j can be considered as the predicted value for the j^{th} category of the given image.

The max-pooling fusion is a crucial step for the proposed RLSD model to be robust to the noise. The output of the fusion layer is fed into a multi-way softmax layer with the squared loss as the cost function, which is defined as:

$$J = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}(p_i^j - \hat{y}_i^j)^2 \tag{3.14}$$

where $\hat{y}_i = y_i/||y_i||_1$ is the ground-truth probability vector of i^{th} image and p_i is the predictive probability vector of i^{th} image. N is the number of images.

Figure 3.5 shows the illustration of the proposed RLSD model for the test image. The potential multi-label regions of the test image are generated by the 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' *etc.*

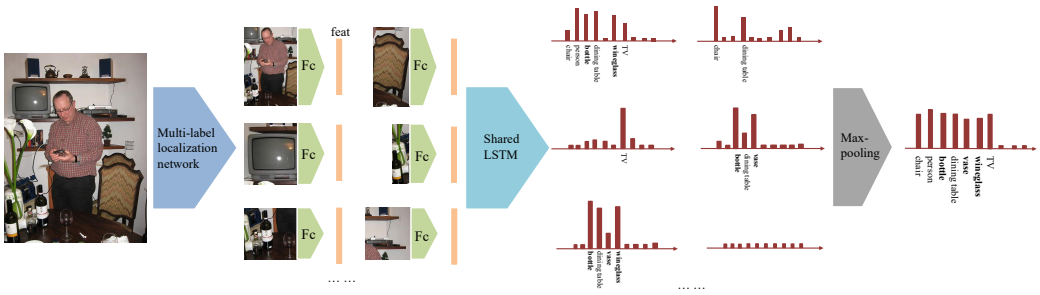


Figure 3.5: An illustration of proposed RLSD model for the test image. The potential multi-label regions of the test image are generated by localisation layer and further used to extract features and input to shared LSTM. As we can see, the small-sized objects like 'wine glass,' 'bottle' and 'vase' *etc*. can be included in the regions due to our multi-label localisation network. The test is also performed in an end-to-end fashion.

can be included in the regions due to our multi-label localisation network. The test is also performed in an end-to-end fashion.

3.2.5 Initialisation and Pre-Training

Our model is capable of training end-to-end from scratch, but a proper initialisation and pre-training mechanism is essential to achieve a promising performance.

Localisation Layer Pre-Training The localisation layer is pre-trained on the Visual Genome region caption dataset (Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein & Fei-Fei 2017). Different from other object detection datasets, each region in the image of this dataset normally contains several objects and visual concepts, which is very suitable for our multi-label region localisation task.

LSTM Pre-training In the training stage, the LSTM is first pre-trained on the global image without region proposals, where every time step has the global image label as ground-truth to compute loss. This global model is also implemented and compared in the experimental part, known as CNN-LSTM. Then the pre-trained LSTM is used as the initialisation of the regional LSTM in our proposed RLSD

model. We find this initialisation process is important for the model to converge fast.

3.3 Experiments

In this section, we present our experimental results and analysis to show the effectiveness of our proposed RLSD model for multi-label image classification problem. We evaluate the proposed model on four benchmark datasets: MS-COCO (Lin et al. 2014), NUS-WIDE (Chua et al. 2009), PASCAL VOC 2012 and 2007 (Everingham et al. 2010). By comparing with several state-of-the-art models and baseline models, we show that our proposed RLSD model achieves the best performance. We further analyse the precision-recall along with the bounding-box size to show our model is especially good at predicting small objects.

3.3.1 Implementation Details

As mentioned in the Section 3.2.2, the VGGNet (Simonyan & Zisserman 2014) is used to initialise the convolutional layers for the localisation network. The embedding size of the image region feature (dimension of W_{ev}) and the label embedding size is 64 (dimension of W_{es}). We only use the one-layer LSTM, while the LSTM memory cell size is 512. Stochastic gradient descent (SGD) is used for optimisation, and we employ learning rate 10^{-5}, momentum 0.9 and drop rate 0.5.

Data preprocessing In proposed RLSD model, the ground truth bounding boxes are not used. The localisation layer in RLSD model is pre-trained on the Visual Genome dataset to give a proper initialisation for fast converging since each region in the image of this dataset normally contains several objects and visual concepts. The localisation layer is not fixed during training; RLSD finetunes the whole network in an end-to-end fashion. Moreover, for MS-COCO (Lin et al. 2014) and VOC 2007/2012 datasets (Everingham et al. 2010, Lin et al. 2014), the ground-truth of bounding-boxes for the single objects are provided. Therefore, we can

process the bounding-box coordinates to finetune the localisation layer in our model. By using this additional information, the fine-tuned model can extract the regions more accurately. Therefore, it can be viewed as an upper bound model to verify the RLSD's generality. We call this upper bound model RLSD+ft-RPN (RLSD with finetuned RPN).

The data pre-processing procedures are: given an image with several single-object bounding-boxes, we first compute their centres. Since the relevant objects usually appear close to each other, Euclidean distance based hierarchical clustering is applied to these bounding-box centres to generate several clusters. Bounding-boxes which belong to the same cluster is merged as one novel region. These novel regions are further used as the ground-truth to finetune the localisation layer. Please note that only the bounding-box coordinates are used in this pre-processing, there are no label annotations involved.

3.3.2 Evaluation Metrics

We refer to the evaluation metrics used in (Gong et al. 2013, Wang et al. 2016) to compute precision and recall for the predicted labels. For each test image, we predict k highest ranked labels and compare against the image ground-truth. The precision is the number of the correctly annotated labels divided by the number of predicted labels; the recall is the number of correctly annotated labels divided by the number of ground-truth labels. We compute *overall precision & recall (op & or)* and *per-class precision & recall (cp & cr)* based on the formulas blew. We also compute the *mean average precision* (mAP) for the comparisons.

$$op = \frac{\sum_{i=1}^{c} N_i^c}{\sum_{i=1}^{c} N_i^p}, \; or = \frac{\sum_{i=1}^{c} N_i^c}{\sum_{i=1}^{c} N_i^g}, \tag{3.15}$$

$$cp = \frac{1}{c}\sum_{i=1}^{c} \frac{N_i^c}{N_i^p}, \; cr = \frac{1}{c}\sum_{i=1}^{c} \frac{N_i^c}{N_i^g}, \tag{3.16}$$

where c is the number of total labels, N_i^c is the number of images that are accurately labeled for i_{th} label,N_i^p is the number of the predicted images for i_{th} label, and N_i^g is the number of ground-truth images for i_{th} label.

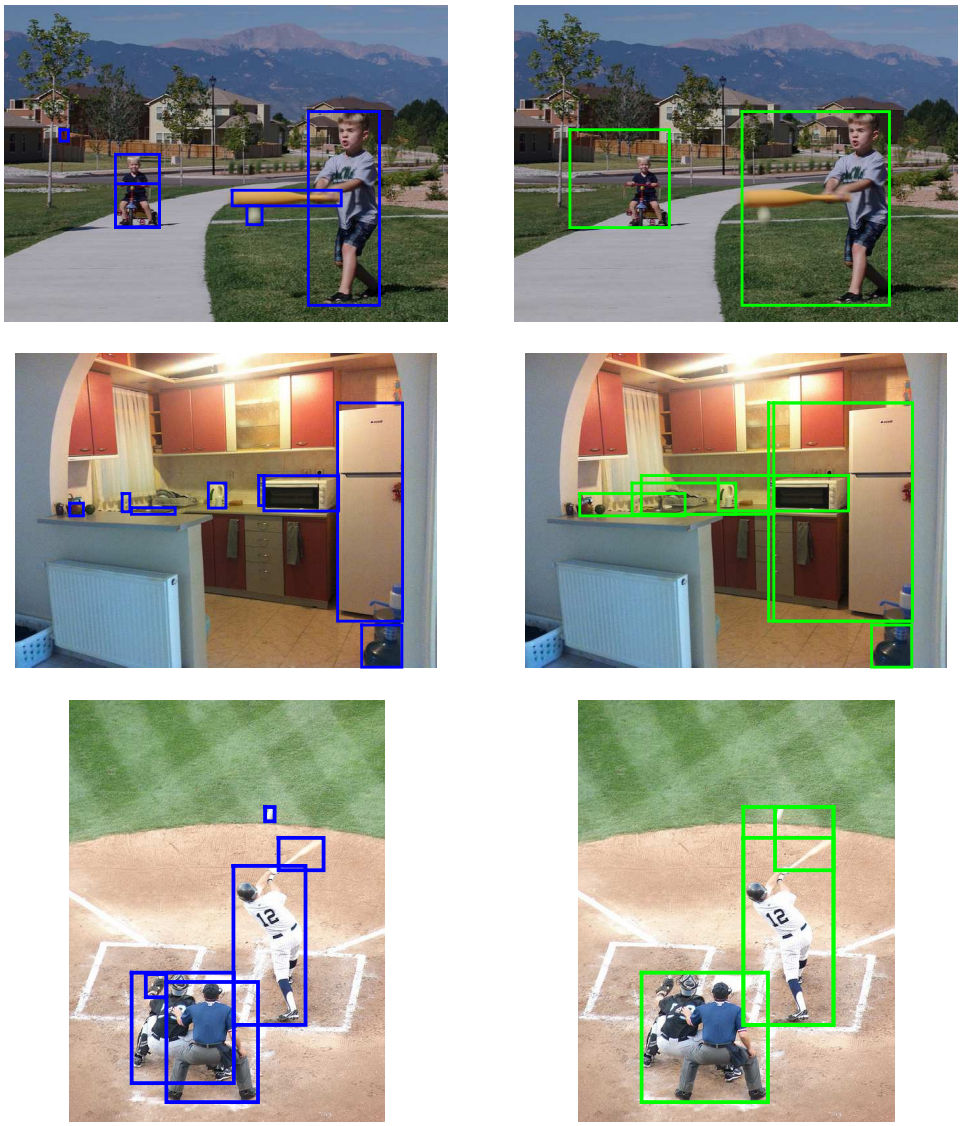


Figure 3.6: The comparison results between the object bounding boxes (left) and generated multi-label regions (right). Our generated regions contain multiple objects, for example, the generated regions contain the object of 'person/baseball/baseball bat,' 'cup/oven' and 'oven/fridge' altogether.

3.3.3 Baseline Models

To evaluate the effectiveness of our proposed RLSD model, we implement four baseline models for comparisons. The experimental results of these baseline mod-

els indicate the significance of our model's components.

Multi-CNN This is a standard CNN model without considering any label-dependencies. We use the VGGNet (Simonyan & Zisserman 2014) pre-trained on the ImageNet for parameters initialisation, and finetune the model on the benchmark datasets. The dimension of the last fully-connected layer is changed to the class number of each dataset, and the element-wise logistic loss is applied. The learning rates of the last two fully-connected layers are initialised as 0.001 and 0.05 respectively, the rest of the layers are fixed. We run around 45 epochs in total until the model leads to convergence and decrease the learning rate to one-tenth every 15 epochs, the weight decay is 0.0005 and momentum is 0.9.

CNN+LSTM This is a model with the same configurations as our proposed RLSD model, except that it only considers the semantic dependencies at the global level. We use the last fully-connected layer of the pre-trained VGGNet (Simonyan & Zisserman 2014) as the global image representation. Then we feed the image feature into the LSTM to train an annotation model. This baseline is similar to the CNN-RNN model proposed in (Simonyan & Zisserman 2014), except the image feature is only fed into the LSTM once at the first time step. The dimensions of the label embedding and LSTM layer are the same as our proposed model: 64 and 512 respectively.

MCG-CNN+LSTM This is a baseline model, which is used to verify the effectiveness of our localisation of multi-label regions. In this model, we use the same configurations as our proposed model in Section 3.2 except that the region proposal network is replaced by a bottom-up object proposal tool: the Multiscale Combinatorial Grouping (MCG) (Pont-Tuset et al. 2017). The pre-extracted top-256 object proposals (based on the object proposal confidence score) for each image are sent into our region-based LSTM for training and testing. As shown in Figure 3.3, our region-based localisation layer can generate the proposals with multiple labels.

	C-P	C-R	C-F1	O-P	O-R	O-F1	mAP	mAP@10
WARP (Gong et al. 2013)	59.3	52.5	55.7	59.8	61.4	60.7	-	49.2
CNN-RNN (Wang et al. 2016)	66.0	55.6	60.4	69.2	**66.4**	**67.8**	-	61.2
Multi-CNN	54.8	51.4	53.1	56.7	58.6	57.6	60.4	57.8
CNN+LSTM	62.1	51.2	56.1	68.1	56.6	61.8	61.8	59.2
EdgeBox-CNN+LSTM	63.8	52.6	57.7	62.8	58.9	60.8	62.5	61.8
MCG-CNN+LSTM	64.2	53.1	58.1	61.3	59.3	61.3	64.4	62.4
RLSD	**67.7**	56.4	61.5	**70.5**	59.9	64.8	67.4	65.9
RLSD+ft-RPN	67.6	**57.2**	**62.0**	70.1	63.4	66.5	**68.2**	**66.7**

Table 3.1: Comparisons on the MS-COCO dataset for $k = 3$. mAP@10 measures are additionally computed for comparison.

EdgeBox-CNN+LSTM This is an equivalent baseline model of the MCG-based model, where the bottom-up proposal method EdgeBox (Zitnick & Dollár 2014) is used to replace the region proposal network. The configurations of this model are same as MCG-CNN+LSTM baseline. Top-256 generated object proposals are sent to the regional based LSTM for training and test.

3.3.4 Results on the MS-COCO

MS-COCO dataset (Lin et al. 2014) is a large scale benchmark dataset for several vision tasks. There are in total 123,287 images for training and validation with 80 object concepts annotated. We use all the annotated object labels in an image as the multi-label ground-truth, and employ its training set as the training data and validation set as the test data. After removing the images without annotation, we have 82,081 images for training and 40,137 images for testing. We obtain the semantic dependencies of these labels by computing their co-occurrence rates and form them as a matrix. We have found that there is strong dependencies in its label set, e.g. 'keyboard' and 'computer' always appear together.

Table 3.1 shows the comparisons between proposed methods, including the WARP (Gong et al. 2013) model and CNN-RNN (Wang et al. 2016) model, along with the baselines models on the MSCOCO dataset. The WARP model uses

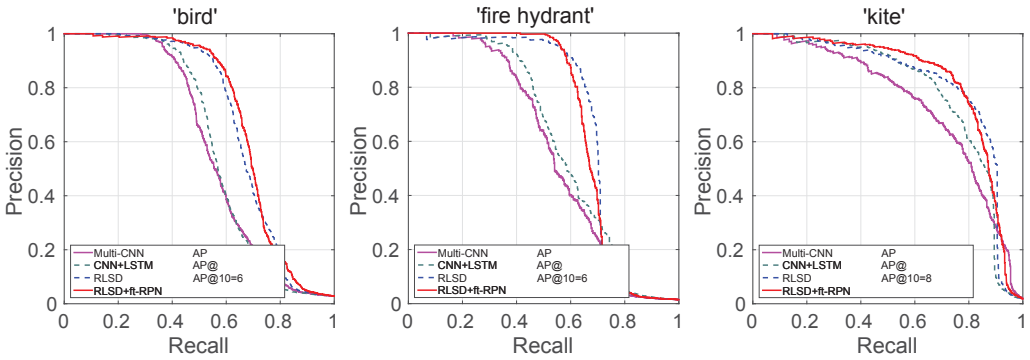


Figure 3.7: Precision-Recall curves for the 'bird,' 'fire hydrant' and 'kite' classes in the MS-COCO dataset, for our RLSD models and the baseline models. The average precision @ ten is also given in the figure.

the ranking based loss to optimise the positive label rank higher than the negative ones, while the CNN-RNN model explores the semantic dependencies on the whole image. Our proposed RLSD model uses the pre-trained region proposal network and fixes the localisation layer during the training, while the upper bound mode RLSD+ft-RPN finetunes the localisation layer with the region-based bounding-boxes as the guidance. The region bounding-boxes generation details have been discussed ins Section 3.3.1. From Table 3.4, we can see that our RLSD model outperforms the WARP on all evaluation matrixes and outperforms the CNN-RNN on the evaluation metrics of per-class precision & recall, and overall precision. This proves the regional semantic dependencies are more advanced than only considering the semantic dependencies at the global level. Although we have a lower overall recall, because our model may predict less than k labels for an image (a threshold $t = 0.5$ is set on the prediction score to decide whether the label should be output). We have higher mAP and mAP@10 measures than the state-of-the-art methods. The comparisons between Multi-CNN and CNN+LSTM baselines further proves that the regional semantic dependencies are more advanced than only considering the semantic dependencies at the global level. The MCG/EdgeBox-CNN+LSTM use the bottom-up methods to generate the region proposals. However, the superior performance against these baselines indicates that since the bottom-up proposals mainly refer to the single

object, the semantic dependencies inside these regions are limited to explore. We show some example proposals from MCG and the proposed model in Figure 3.3.

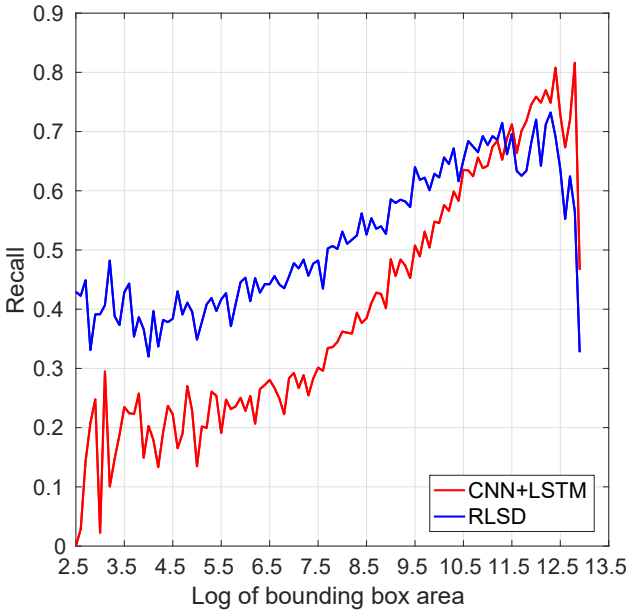


Figure 3.8: The relationship between recall and bounding-box area on the MS-COCO dataset.

Figure 3.7 shows the precision-recall curves of three selected classes (bird, fire hydrant, and kite) in the MS-COCO dataset, these three kinds of objects are visually small in the images of this dataset. The PR curves illustrate that our RLSD model significantly surpasses the baseline models and is especially good at predicting these small objects. Some qualitative results are shown in Figure 3.9. Small sized objects are noted with bold. As we can see, small objects like 'sports ball', 'bird', and 'bottle' *etc*. are only predicted by our RLSD model. Figure 3.8 analyses the recall along with the object ground-truth bounding-box size on the MS-COCO dataset, which shows that our model achieves much higher recall than the CNN+LSTM model on those labels that corresponding bounding-boxes are smaller. This means that our RLSD model is more sensitive to 'small' objects and attributes. Both recall curves start to fall when the object is so large that it almost fills the whole image, the similar case is also observed in (Wang et al. 2016).

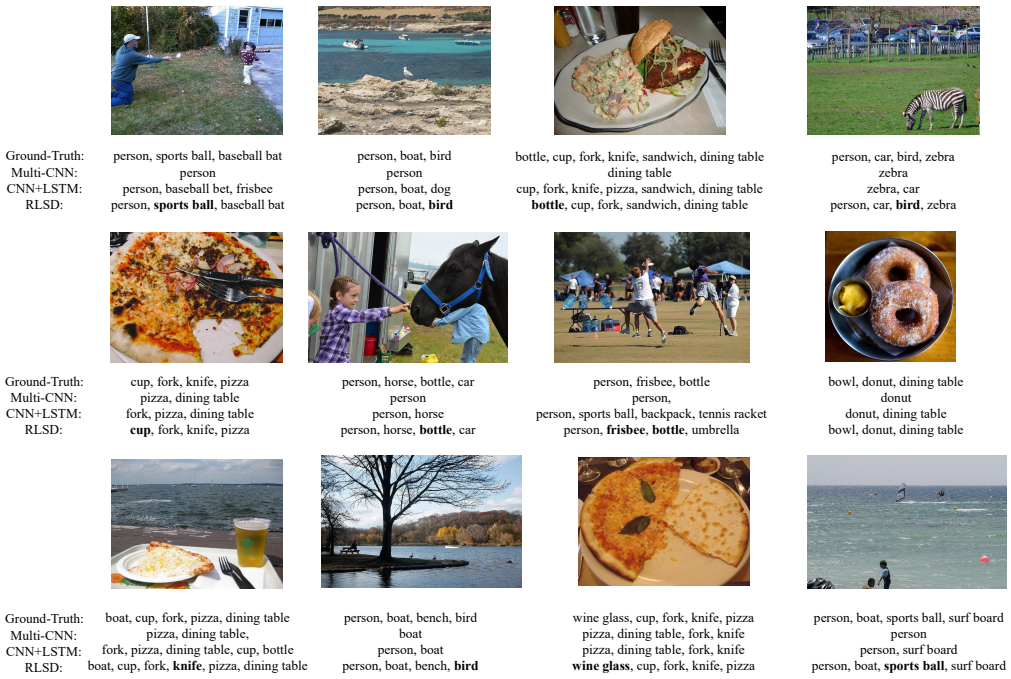


Figure 3.9: Some example annotation results from the MS-COCO dataset.

3.3.5 Results on the NUS-WIDE

NUS-WIDE (Chua et al. 2009) is a web image dataset associated with user tags. The whole dataset contains 269,648 images and 1,000 tags. These images are further manually labelled into 81 concepts. After removing the no-annotated images, the training set and test set contains 125,449 and 83,898 images respectively.

Table 3.2 shows the comparison results on the NUS-WIDE dataset on 81 concepts. We compare the proposed RLSD model with previous mentioned methods including metric learning (Li, Lin, Rui, Rui & Tao 2015), multi-edge graph (Liu, Yan, Rui & Zhang 2010), K nearest neighbour (Chua et al. 2009), softmax prediction, WARP method (Gong et al. 2013) and the state-of-the-art method CNN-RNN (Wang et al. 2016). Same as (Wang et al. 2016), we do not finetune the CNN image representation. Our proposed RLSD model outperforms these methods in a large margin. Specifically, we achieve 4% higher precision than the CNN-RNN. The superior performance on the NUS-WIDE dataset indicates that using LSTM to explore the dependencies within the multi-label regions can largely improve

the annotation accuracy. Since the bounding-boxes are not provided in the NUS-WIDE dataset, there is no upper bound for this dataset.

	C-P	C-R	C-F1	O-P	O-R	O-F1	mAP
Metric Learning (Li, Lin, Rui, Rui & Tao 2015)	-	-	-	-	-	21.3	-
Multi-edge graph (Liu et al. 2010)	-	-	-	35.0	37.0	36.0	-
KNN (Chua et al. 2009)	32.6	19.3	24.3	42.9	53.4	47.6	-
softmax (Wang et al. 2016)	31.7	31.2	31.4	47.8	59.5	53.0	-
WARP (Gong et al. 2013)	31.7	35.6	33.5	48.6	60.5	53.9	-
CNN-RNN (Wang et al. 2016)	40.5	30.4	34.7	49.9	61.7	55.2	-
Multi-CNN	40.2	42.8	41.5	53.4	66.4	59.2	47.1
CNN+LSTM	42.2	45.8	43.9	53.4	66.5	59.3	49.9
EdgeBox-CNN+LSTM	41.8	43.7	42.7	53.6	65.7	59.0	51.2
MCG-CNN+LSTM	44.3	48.1	46.1	54.1	67.2	59.9	52.4
RLSD	**44.4**	**49.6**	**46.9**	**54.4**	**67.6**	**60.3**	**54.1**

Table 3.2: Comparisons on the NUS-WIDE dataset on 81 concepts for $k = 3$.

3.3.6 Results on the PASCAL VOC Datasets

PASCAL VOC 2012 and 2007 (Everingham et al. 2010) are two benchmark datasets for image annotation, segmentation, and detection tasks. In VOC 2012 dataset, there are 11,540 images for training and validation, and 10,991 images for the test. In VOC 2007 dataset, there are 9,963 images in total, of which 5,011 and 4,952 images for training/validation and testing respectively. Twenty common objects are annotated in these two datasets.

The results of VOC 2012 are shown in Table3.3. We compare proposed models with baseline models and several state-of-the-art methods, including the Edge-Box (Zitnick & Dollár 2014), PRE-1000C* (Oquab et al. 2014) and HCP-VGG (Wei et al. 2014). EdgeBox (Zitnick & Dollár 2014) generates single object proposals based on the hand-crafted edge features, while HCP-VGG uses single object proposals as inputs and finetunes the CNN pre-trained on single label images. As we can see, the proposed RLSD and RLSD+ft-RPN achieve the best performance against all these models on the VOC 2012 dataset. The improvements

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
EdgeBox(Zitnick & Dollár 2014)	97.3	84.2	80.8	85.3	60.8	89.9	86.8	89.3	75.4	77.8	75.1	83.0	87.5	90.1	95.0	57.8	79.2	73.4	94.5	80.7	82.2
PRE-1000C*(Oquab, Bottou, Laptev, Sivic et al. 2014)	93.5	78.4	87.7	80.9	57.3	85.0	81.6	89.4	66.9	73.8	62.0	89.5	83.2	87.6	95.8	61.4	79.0	54.3	88.0	78.3	78.7
HCP-VGG(Wei et al. 2014)	**99.1**	92.8	97.4	94.4	79.9	93.6	89.8	98.2	78.2	94.9	79.8	97.8	97.0	93.8	96.4	74.3	**94.7**	71.9	96.7	88.6	90.5
Multi-CNN	97.4	86.4	92.7	90.3	62.8	89.1	80.2	95.1	72.4	82.9	78.9	93.2	93.1	90.6	94.2	53.1	86.5	64.5	94.0	79.8	83.8
CNN+LSTM	97.8	85.3	92.9	91.6	64.3	89.9	82.3	95.5	74.7	82.5	81.0	94.6	93.1	90.5	95.0	57.5	86.2	67.5	95.6	82.4	85.0
EdgeBox-CNN+LSTM	98.0	85.7	93.1	91.7	64.4	90.5	82.4	95.6	75.5	84.2	91.3	94.8	92.9	91.0	95.2	57.9	87.2	69.0	95.5	82.5	85.4
MCG-CNN+LSTM	98.1	90.9	94.0	90.1	69.2	92.5	88.1	95.9	55.6	89.5	65.9	95.8	95.8	93.2	92.7	59.1	88.5	72.4	96.7	86.3	85.5
RLSD	98.9	94.4	97.4	94.8	77.8	94.9	91.6	98.1	82.2	92.9	84.1	97.7	97.1	96.3	97.5	74.6	92.4	79.2	**97.8**	89.9	91.5
RLSD+ft-RPN	99.0	**95.6**	**97.9**	**95.2**	**82.5**	**95.5**	**94.9**	**98.5**	**86.1**	**95.4**	**84.4**	**98.4**	**98.1**	**97.1**	**98.8**	**81.5**	**94.7**	**83.4**	97.7	**92.2**	**93.3**

Table 3.3: Comparisons of annotation results on the VOC 2012 dataset.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
INRIA (Harzallah, Jurie & Schmid 2009)	77.2	69.3	56.2	66.6	45.5	68.1	83.4	53.6	58.3	51.1	62.2	45.2	78.4	69.7	86.1	52.4	54.4	54.3	75.8	62.1	63.5
FV (Perronnin, Sánchez & Mensink 2010)	75.7	64.8	52.8	70.6	30.0	64.1	77.5	55.5	55.6	41.8	56.3	41.7	76.3	64.4	82.7	28.3	39.7	56.6	79.7	51.5	58.3
PRE-1000C*(Oquab et al. 2014)	88.5	81.5	87.9	82.0	47.5	75.5	90.1	87.2	61.6	75.7	67.3	85.5	83.5	80.0	95.6	60.8	76.8	58.0	90.4	77.9	77.7
HCP-Alex (Wei et al. 2014)	**95.4**	**90.7**	**92.9**	88.9	53.9	81.9	**91.8**	**92.6**	60.3	**79.3**	73.0	90.8	89.2	**86.4**	92.5	**66.9**	**86.4**	**65.6**	94.4	**80.4**	**82.7**
Multi-Alex	95.4	90.7	92.9	89.9	53.9	81.9	91.8	92.6	60.3	79.3	73.0	90.8	89.2	86.4	92.5	66.9	86.4	65.6	94.4	80.4	82.7
Alex+LSTM	92.4	82.0	89.0	83.7	38.0	73.9	87.8	89.4	65.9	58.7	73.5	87.5	86.0	80.0	92.8	41.8	43.7	60.7	93.5	70.0	74.5
EdgeBox-Alex+LSTM	94.4	84.0	90.7	86.8	37.4	77.3	88.5	90.8	66.8	61.5	76.5	88.2	86.8	81.7	93.5	42.7	68.3	63.1	94.7	69.8	77.2
MCG-Alex+LSTM	93.2	83.3	90.0	86.1	42.7	76.6	88.1	90.1	66.8	64.3	74.8	88.8	87.1	81.8	93.2	45.9	62.8	62.2	94.0	71.6	77.2
RLSD-Alex	95.7	87.7	92.1	**90.0**	49.2	80.7	89.5	91.9	69.4	69.2	**78.5**	90.6	89.9	84.9	**94.4**	56.3	80.0	66.7	**95.5**	75.3	81.4
RLSD+ft-RPN-Alex	95.2	87.4	91.8	**90.0**	**54.2**	**82.7**	89.0	91.7	**68.6**	77.2	77.6	**91.2**	**91.6**	86.1	93.8	58.2	82.9	65.3	94.9	76.8	82.3
HCP-VGG (Wei et al. 2014)	**98.6**	**97.1**	**98.0**	**95.6**	**75.3**	**94.7**	**95.8**	**97.3**	73.1	**90.2**	**80.0**	**97.3**	**96.1**	**94.9**	96.3	**78.3**	**94.7**	**76.2**	**97.9**	**91.5**	**90.9**
CNN-RNN (Wang et al. 2016)	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
Multi-VGG	95.8	88.7	93.2	91.1	51.5	82.1	89.5	91.8	68.3	80.2	75.7	91.4	92.6	88.7	92.8	61.2	82.9	68.0	96.2	78.0	83.0
VGG+LSTM	96.8	89.7	93.3	90.4	54.6	85.6	89.0	92.3	68.9	80.9	76.8	90.5	93.6	89.5	93.0	60.3	84.3	67.4	96.2	76.7	83.5
EdgeBox-VGG+LSTM	97.0	90.2	93.3	92.2	57.0	84.0	90.1	93.1	71.1	80.5	80.6	92.3	92.8	87.5	94.8	61.5	85.1	69.3	96.1	79.0	84.4
MCG-VGG+LSTM	94.3	90.4	91.4	91.4	64.9	90.3	92.9	94.3	71.9	83.7	55.6	92.5	94.5	90.0	96.8	60.1	90.5	71.3	96.3	84.7	84.9
RLSD-VGG	95.3	92.4	91.2	92.1	71.9	91.1	93.3	94.8	**74.9**	86.1	70.4	93.3	95.6	89.7	**98.0**	66.8	89.4	75.7	96.6	85.9	87.3
RLSD+ft-RPN-VGG	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5

Table 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.

against CNN+LSTM again proves that exploring the semantic dependencies inside the multi-label regional proposals is more effective than at the global level.

Table 3.4 shows the comparisons between our proposed models and state-of-the-art methods on the VOC 2007 dataset, including the INRIA (Harzallah et al. 2009), FV (Perronnin et al. 2010), PRE-1000C*(Oquab et al. 2014), HCP-Alex/VGG (Wei et al. 2014), and CNN-RNN (Wang et al. 2016), as well the baseline models. Our models outperform the baseline models and most state-of-the-art methods, especially the improvement against CNN-RNN model, which validates our contributions. However, we observe that both RLSD and RLSD+ft-RPN are hard to catch up with the HCP-Alex/VGG, which uses a single label region as inputs to finetune the CNN trained on the single images. The reason is that VOC 2007 dataset contains more than 50% single label or two-label images and is relatively small scale compared to other multi-label datasets, such as MS-COCO, NUS-WIDE and VOC 2012, which cause the dependencies between label pairs are relatively weak and even weaker for our generated regional proposals. This situation is not in favour of the LSTM based model to capture the semantic dependencies; similar results are also observed on the CNN-RNN model, which explores the semantic dependencies at the global level. However, by introducing the regional information, we manage to strength the semantic dependencies to outperform CNN-RNN model. Compared to the VOC 2007 dataset, VOC 2012 dataset is twice scaled with more multi-label images, which provides more opportunities for our model to capture the semantic dependencies. We show the statistics of training/validation sets in VOC 2007 and VOC 2012 datasets in Figure 3.10. We also visualise the semantic dependencies of these two datasets by computing the label co-occurrence rates and form them as a matrix in Figure 3.11. As we can see, the VOC 2012 dataset possesses stronger dependencies compared to the VOC 2007.

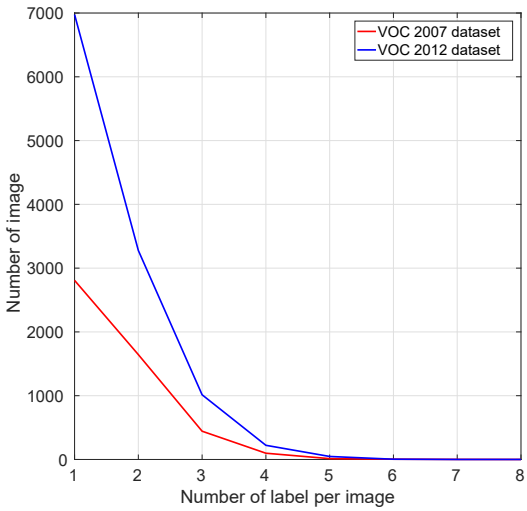


Figure 3.10: The statistics of training/validation set of VOC datasets.

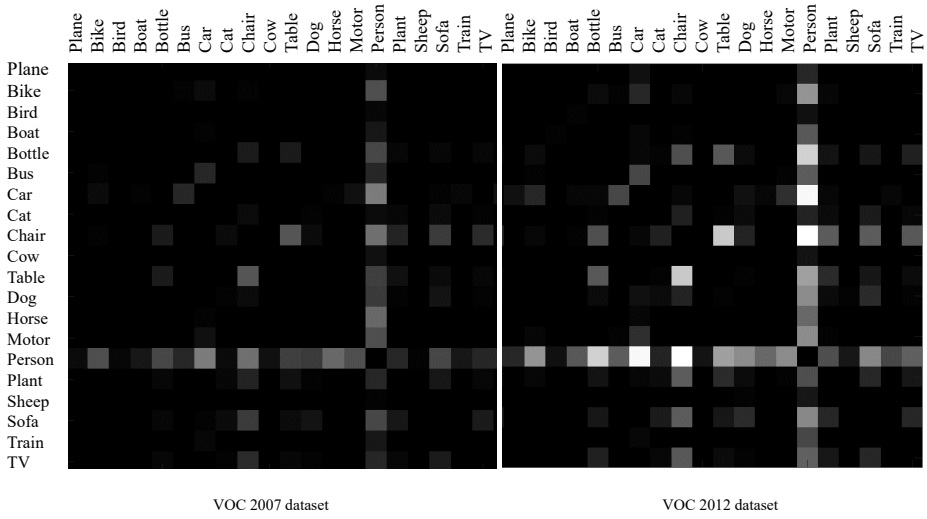


Figure 3.11: The visualisation of semantic dependencies for the VOC 2007 and VOC 2012 datasets. The brighter the block is, the higher dependency the corresponding label pair has.

3.4 Summary

Generic image annotation is an important and realistic problem in the multimedia area because it is not only more challenging than the single-label image classification but also closer to the real-world applications. In this work, we propose a Re-

gional Latent Semantic Dependencies (RLSD) model to address this problem. As its name suggests, this proposed model can capture the latent label dependencies at the regional level. There are two main advantages in the proposed model; one is the fully-convolutional localisation network for capturing the regions potentially have the multiple dependent labels, another is the shared LSTM for analysing the latent semantic dependencies within the regions. Experimental results on several benchmark datasets show that the proposed RLSD model consistently achieves the superior overall performance to the state-of-the-art approaches, especially for predicting small objects and visual concepts occurred in the image.

Chapter 4

Weakly-Supervised Annotation & Tag Refinement

4.1 Introduction

To better understand and efficiently retrieve the images on the social network, it is essential to develop an efficient annotation model. Traditional methods on the image annotation focus on using expert-labelled images as the training data to uncover the relationships between the image visual content and tags (Guillaumin et al. 2009, Makadia, Pavlovic & Kumar 2010, Ballan et al. 2014). In recent years, the deep neural network (Krizhevsky et al. 2012, Simonyan & Zisserman 2014) has achieved superior performance on the image feature learning and been widely used in the image classification and related vision tasks (Wei et al. 2014, Wang et al. 2016). However, these deep models are purely based on supervised learning and require a significant amount of expert-labelled training samples to obtain satisfactory results. The labelling process for the training data can be intensive and prohibitive, and it is unrealistic for the human to continuously label the large-scale images that pop into the social networks every day, to obtain the high-quality training data.

People sometimes spontaneously assign tags to images when uploading them, and more tags are attached along with the sharing. Despite these tags provide

Image Annotation:

sky; blue; cloud; tree;
building; house; city

User-Provided Tags:
pink; gerber; gerbera; daisies;
green; gift; athankyou; project365;
nikond70s; 365days

Tag Refinement:
color; daisy; flower; petal; green; pink

Figure 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.

the semantic illustrations of images to a certain extent, they can be incomplete, imprecise, and biased toward individual perspectives (Li et al. 2016). See Figure 4.1 as an example. To reduce the existing tag error for the further retrieval and train the annotation model, some previous works (Zhu et al. 2012, Liu et al. 2009, Li, Snoek & Worring 2009, Li & Snoek 2013, Han et al. 2014, Zhu et al. 2010) are conducted on the tag relevance analysis and refinement. These works explore the tag relevance from different perspectives including the semantic similarities between tags, the visual similarities among image neighbours and the properties of the annotation matrix. The user-provided tag is further refined based on its relevance. The most relevant works are conducted in (Zhu et al. 2010, Li & Tang 2017), which focus on analysing the low-rank property of the annotation matrix. The user-provided annotation matrix and the image feature matrix are used as the model input respectively. However, the image features are not connected to the refined results in (Zhu et al. 2010), while they are fixed in (Li & Tang 2017). Moreover, these methods focus on solving the problem at the image level, while the associated tags of one image may correspond to the different regions, instead of the whole image.

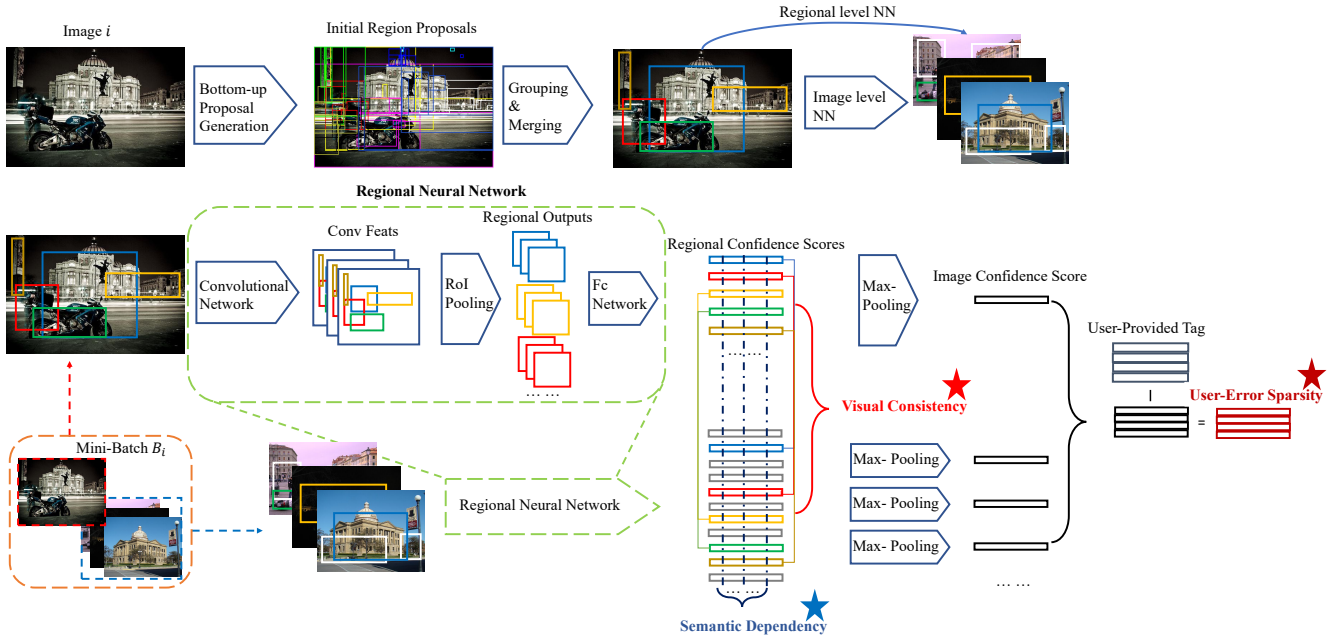


Figure 4.2: The proposed framework. Each image $i \in I$ is first extracted with multiple region proposals; then representative regions are selected by the grouping and merging. The Nearest Neighbor (NN) approach is performed at both the image level and regional level to locate the region pairs for the 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.

To address the above issues, in this work, we propose to learn an image annotation model and refine the user-provided tags simultaneously in a weakly-supervised manner. The whole framework is shown in Figure 4.2. A region based Deep Convolutional Neural Network (DCNN) is adopted as the feature learning and backbone annotation model. Different from the regular DCNN that is trained on the supervised information, which is usually annotated by experts, we consider the user-provided tags as the weakly-supervised information to assist the training. Since a single image is often associated with multiple tags, some tags may refer to the regions instead of the whole image. See Figure 4.3 as an example. In order to discover those regions with semantical meanings, we first group the potential regions generated by state-of-the-art proposal method (Krähenbühl & Koltun 2014), and select the representative ones as the regional inputs by merging each group. The input image and its region coordinates are then passed through the Region of Interest (RoI) pooling network to extract visual features and are further encoded with a fully-connected network for predicting tags with confidences. A max-pooling operation is carried out to fuse the regional outputs as the final prediction for the image.

At the training stage, we propose to learn the regional visual representation and the relationship between the visual content and tags by exploring the visual consistency among the regional neighbours and the semantic dependencies between the tag pairs. That is, image regions with similar visual appearance are usually annotated with the same tags, and semantic dependent tags always jointly appear in the same image. To efficiently model these two constraints, the pre-defined visual and tag similarity are adopted at the regional level. Moreover, considering users share the general knowledge about the semantic annotation and each image usually is assigned with very few tags compared to the entire tag set. Therefore, the errors of the user-provided tags on a batch of images are sparse. By setting these constraints at both the image level and regional level, we can train the neural network to conduct tag refinement and learn the annotation model at the same time. In summary, the main contributions of our model are as follows:

1) We propose to obtain the deep neural network based image annotation

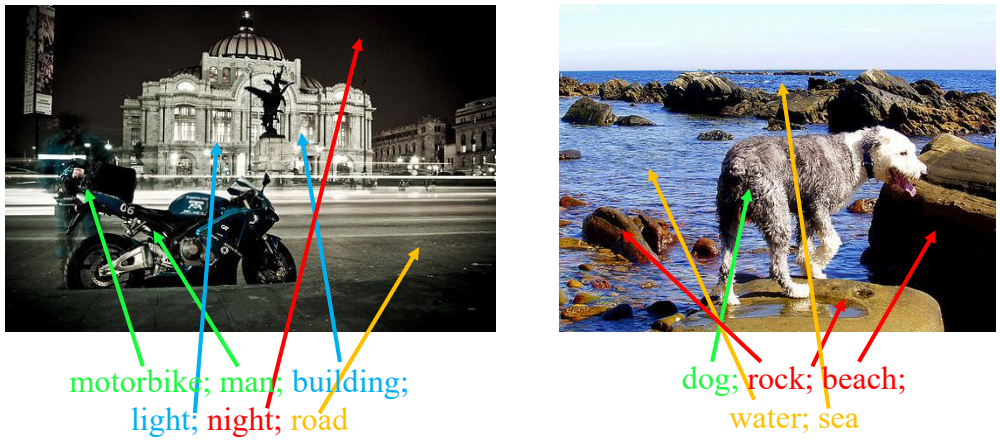


Figure 4.3: The examples of images assigned to multiple tags. As we can see, different tags such as 'motorbike,' 'building,' and 'dog' *etc.* correspond to the different regions of the image, instead of the whole image, which is the motivation that we explore the image regional information for both tasks.

model and conduct the tag refinement simultaneously in a weakly-supervised manner. During the training, the user-provided tags are spontaneously refined from 0/1 assignments to the confidence scores, while the trained annotation model can be applied to assign tags to new images;

2) Different from previously related works that focus on solving the problem at the image level, we propose to conduct the annotation and refinement process at both the image level and regional level, which not only enable the visual feature learning to ensure the flexibility and stability but also enhance the ability to model the proposed constraints;

3) By utilising three proposed constraints, our model achieves state-of-the-art performance for both the image annotation and tag refinement experiments on two benchmark datasets.

4.2 The Proposed Model

The key characteristic of our model is that we formulate the constraints of the regional deep neural network from two aspects. One is the internal relationship of the dataset, which is reflected as the visual consistency among region neighbours and the semantic dependencies between tag pairs. The other one is the general knowledge that the errors of the user-provided tags are sparse. To appropriately introduce these constraints into the neural network, we adopt the regional architecture to receive the selected regions as network inputs. The entire framework is shown in Figure 4.2.

In the following part, the selection of image regions is first introduced in Section 4.2.1, and the architecture of the regional neural network is described in Section 4.2.2. Then we introduce the proposed three constraints, namely the visual consistency in Section 4.2.3, semantic dependency in Section 4.2.4 and user-error sparsity in Section 4.2.5 respectively. Finally, the training and prediction details of the tag refinement and image annotation are given in Section 4.2.6.

4.2.1 Representative Region Selection

As we can see in Figure 4.3, the multiple tags assigned to the social image may correspond to the certain regions instead of the whole image. Therefore, directly use the image level feature can undermine the annotation performance. To explore the image at the regional level, we need to generate regions that contain semantical meaning.

Since we design the model to handle the diverse social images, and the bounding boxes of semantical regions are not available, we adopt the bottom-up proposal method to generate region candidates. Considering the semantical regions can exist in both the foreground and background, we choose the Geodesic Object Proposals (GOP) (Krähenbühl & Koltun 2014) as the proposal method to cover the whole image. Let I be the image set, T be the set of possible initial tags provided by the users, where $|I| = N$ and $|T| = C$, corresponding to the image and tag set size respectively. For each image i, we generate $R_i = \{r_1, r_2, ..., r_{n_i}\}$ region

proposals. Since one image only has a certain number of regions corresponding to the tags and the overlap ratios of these proposals are significantly high, we conduct the grouping method to select representative proposals as final region inputs for the neural network.

The motivation of using the proposal selection method based on the clustering are two-fold: first, the proposal numbers can be huge, it would be time consuming and inefficient to process all the region proposals; secondly, there are considerable overlaps among these generated proposals, clustering can group these regions to avoid repeat computation. Given image i with the region proposal set R_i, where $|R_i| = n_i$, we construct a $n_i \times n_i$ affinity matrix w, where each element w_{jk} stands for the Intersection over Union (IoU) score of the proposal pair (r_j, r_k):

$$w_{jk} = \frac{|r_j \cap r_k|}{|r_j \cup r_k|} \tag{4.1}$$

where $(j, k) \in [1, n_i]$ and $|\cdot|$ indicates the number of pixels. Then we apply the normalised cut (Shi & Malik 2000) to group the proposal set into $m - 1$ clusters. We remove proposals with the high height-width ratio or small areas, and for each cluster, we merge top q proposals as one representative region. See Figure 4.4 for the examples of the region selection. Since some tags may refer to the whole image scene, we also include the whole image as a special region. After the selection process, we now have m representative regions for each image i.

4.2.2 Regional Neural Network

The deep neural network is adopted as the feature learning and annotation backbone model, and since the image regions are used as the network inputs to uncover the relationship between the image visual content and tags, inspired by (Ren et al. 2015, Zhang, Wu, Shen, Zhang & Lu 2018), we utilise a regional network architecture as the framework.

Since the convolutional layer of the neural network preserves the spatial information of the input image, which is essential for exploring the regional image information, given image $i \in I$ with m regions, we pass them through the convolutional layers to extract features. Specifically, we adopt the VGG-16 network

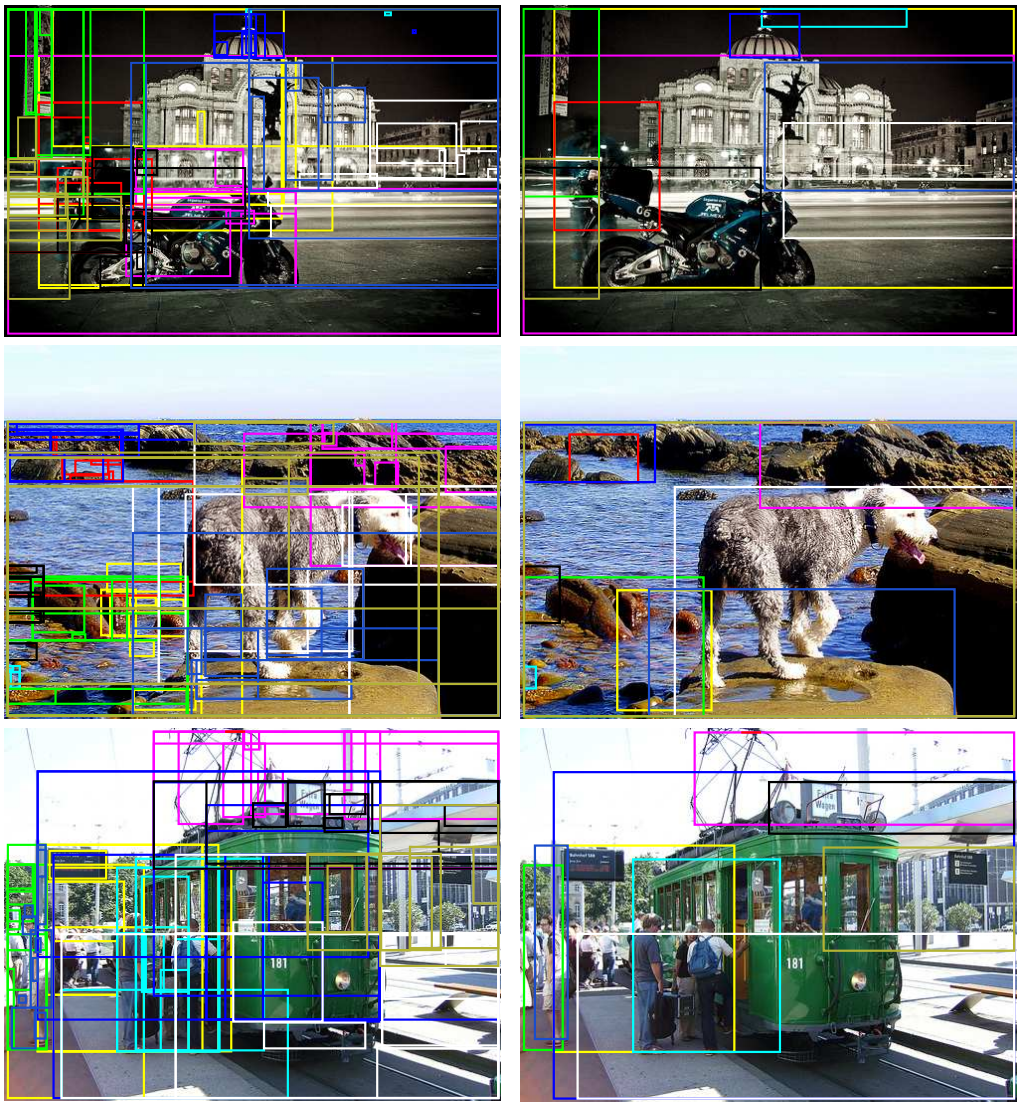


Figure 4.4: The representative region selection. We group proposals generated by GOP (left column) and merge the top proposals in each group into one representative region (right column).

configuration. The output of the last convolutional layer is used as the image feature, noted as $\Theta(i)$. For each region r_j, where $j \in [1, m]$, the coordinates of region boxes are also projected on the convolutional feature map $\Theta(i)$. For each region from the input list, we take the projection section of $\Theta(i)$ and scales it to the pre-defined size $X \times Y$ by adopting the RoI pooling operation, in our case, we

set $X = Y = 7$. By utilising the RoI pooling operation, we can reuse the feature map from the convolutional network to significantly speed up both the training and prediction time.

After the regional features obtained, we feed them into a fully-connected (Fc) network which is composed of three fully-connected layers. Features from each region are encoded by the first two layers as a $4096 - d$ feature vector. All the regional feature vectors from image i form a batch V_i with the size of $m \times 4096$. The batch V_i is then encoded by the last Fc layer to generate the confidence scores for each region. Let p_{ij} be the confidence vector of j_{th} region with size C, and p_{ij}^k is the k_{th} component of p_{ij}, where $k \in [1, C]$. We carry out the tag-wise max-pooling operation to fuse the regional outputs as the final confidence vector p_i of image i:

$$p_i^k = max(p_{i1}^k, \quad p_{i2}^k, \quad ..., \quad p_{im}^k) \tag{4.2}$$

After we establish the network architecture of our model, we start to train the neural network by considering the visual consistency L_{vis} with semantic dependency L_{sem} and user-error sparsity L_{err} altogether; we give the final form of our network constraints:

$$L_{final} = L_{vis} + \lambda_1 L_{sem} + \lambda_2 L_{err} \tag{4.3}$$

where λ_1 and λ_2 are set to balance the different constraints. In the following subsections, we first introduce the input batch selection for the network and the visual consistency constraint. The semantic dependency is presented next, and followed by the user-error sparsity constraint. Finally, we summarise how the proposed model performs the tag refinement and predict tags for new images. Implementation details will also be given in this section.

4.2.3 Input Batch Selection & Visual Consistency

As the first constraint component, we consider the image visual information. Visually similar images tend to be annotated by the same tag. As we addressed before, social images are usually annotated with multiple tags, and different tag

may correspond to the different regions instead of the whole image. Therefore, we evaluate the image visual similarity at the regional level and formulate the input batches based on the region visual similarity to activate the visual consistency constraint.

Input Batch Selection After the representative region selection 4.2.1, we now have m regions for each image $i \in I$. To make the most of the visual similarity between image regions, we need to carefully select input batches for the regional neural network. Comprehensive studies in the image annotation (Li & Tang 2017, Li et al. 2016) and retrieval task (Gao, Xue, Zhou, Pang & Tian 2016) indicate that distances among low-level image features, including the colour, texture, and edge *etc.* can reflect the visual similarity between image pairs. Therefore, for each image i with m regions, we first extract the low-level features of each image and its regions as $\Phi(i)$ and $\Phi(r_{in})$, where $n \in [1, m]$. The visual similarity between image and region pairs is defined as the Euclidean distance between low-level features.

To utilise the visual consistency constraint, we propose to select visually similar and dissimilar region pairs to conduct the metric learning, which will be detailed later. In order to efficiently achieve this objective, we conduct a twofold selection process for input batches. First, we build the KD-tree at the image level based on the distances between image feature pairs. For each image i, we query k nearest neighbours as the positive candidate set $Z_{\{i,pos\}}$ and k farthest images as the negative candidate set $Z_{\{i,neg\}}$. Then we measure the visual similarity between regions among image i and its candidate sets $Z_{\{i,pos\}}$ and $Z_{\{i,neg\}}$. The regional level KD-trees are built for positive and negative sets respectively based on the regional features. Since image i and each image in the candidate sets Z_i have m regions, the Nearest Neighbor approach (NN) at the regional level is performed to vote the final positive and negative set. Take the positive set as an example, for each region r_{in}; the NN is conducted within the region set with size $m \times k$, which is comprised of the regions from the positive image candidate set. The NN results are projected back to the image candidates, and we select top $k'(k' < k)$ images as

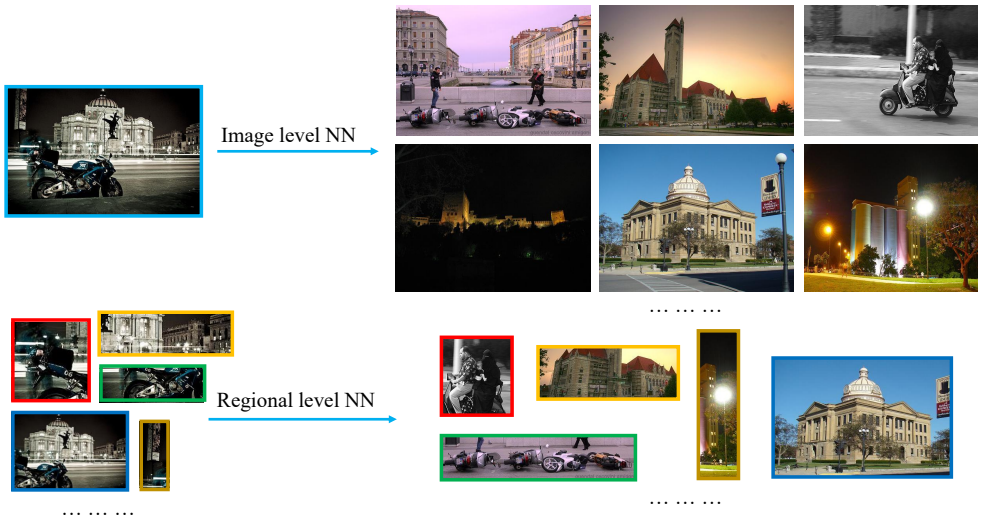


Figure 4.5: An illustration of the selection process for the positive region pairs. Nearest neighbour is performed in twofold. First, the image candidates are selected at the image level; then we measure the visual similarity between regions among image and its candidates at the regional level.

the final positive set $Z'_{\{i,pos\}}$. The equivalent process is performed for the negative set, while the k' images are selected as $Z'_{\{i,neg\}}$. The reason we select the final sets by two steps is to efficiently locate the images with most similar and dissimilar regions since the input batches of the neural network are comprised of images. We demonstrate the selection process for the positive region pairs in Figure 4.5. The input generation process is described in Algorithm.4.1.

Visual Consistency After generating the input batches of the neural network, now we introduce the first constraint of our model. Based on the observation that visually similar images intend to be annotated with the similar tags (*i.e.*, tag distributions should be close), we define the visual consistency constraint as follows.

Each mini-batch B_j is comprised of image i_j with k' positive and k' negative images. During the input generation process, we reserve the indexes of matched positive and negative region pairs. That is, for each region $r_{jn}, n \in [1, m]$, there are w_{pos} and w_{neg} matched region pairs among k' image pairs. The visual similar-

Algorithm 4.1 Generate Input Batches for the Network Training.

Input:

I : Image set; S : Batch size

$\Phi(\cdot)$: Feature extraction for images and regions;

k : Initial candidate set size;

k' : Final positive and negative sets size;

m : Number of regions per image;

Output:

B_j : Input mini-batch contains image i_j;

B : Input batch per iteration;

for Each $j \in [1, S]$ **do**
  Image level NN approach for i_j based on $\Phi(i_j)$;
  Build condidiate sets $Z_{\{i_j,pos\}}$, $Z_{\{i_j,neg\}}$ with size k;
  Build region condidiate sets $Z_{R_{\{i_j,pos\}}}$, $Z_{R_{\{i_j,neg\}}}$ with size $m \times k$;
  for Each $n \in [1, m]$ **do**
    Regional level NN vote for r_{jn} within Z_R;
  end for
  Project NN votes into Z;
  Select k' images as Z' to form j_{th} mini-batch B_j;
end for
Concatenate all the S mini-batches as the input batch B.

ity score between positive region pair is defined as:

$$\gamma_{jn,l} = exp(-(\|\Phi(r_{jn}) - \Phi(r_l)\|^2)/\sigma) \tag{4.4}$$

where $l \in [1, w_{pos}]$, σ is the medium value of γ, $\Phi(\cdot)$ is the regional feature. Then the visual consistency constraint L_{vis} inside a mini-batch can be carried out as follows:

$$d_p = \sqrt{\{p(r_{jn}) - p(r_l)\}^2} \tag{4.5}$$

$$L_{vis} = \min_{P_{B_j}} \frac{1}{m} \sum_{n=1}^{m} \{ \frac{1}{2w_{pos}} \sum_{l=1}^{w_{pos}} \gamma_{jn,l} \cdot d_p^2 + \frac{1}{2w_{neg}} \sum_{l=1}^{w_{neg}} max(0, \epsilon - d_p)^2 \} \tag{4.6}$$

where $p(r)$ stands for the tag confidence prediction for region r, $|p(r)| = C$, d_p measures the Euclidean distance between two confidence vectors. P_{B_j} is the corresponding confidence for the whole mini-batch B_j. ϵ is the margin. The tag confidence prediction can be viewed as the semantic representation of the region content, as we can see, the visual consistency constrains the visually similar regions to have closer tag confidence prediction, while pushing the dissimilar pairs beyond the margin.

4.2.4 Semantic Dependency

Besides the visual consistency among image regions, we also consider the semantic dependencies between the tag pairs. It is natural that social tags are not assigned separately, semantically similar tags often appear together in similar images, as well as the regions. Based on this knowledge, we first estimate the tag pair similarity.

We consider the tag pair similarity from two aspects: the context and knowledge base. Given two tags t_i and t_j, the context ($dist_{ctx}$) is defined as the Google distance (Cilibrasi & Vitányi 2007) of two tags in the given set (we use the image instead of the web page), while the knowledge base ($dist_{kb}$) is the WordNet similarity (Miller 1995) based on the information content of the least common subsumer and input synsets. The motivation of doing so is to eliminate the possible bias brought by the training set without causing the domain drift, namely transfer the general external knowledge to the given set. Therefore, a weighted sum of two aspects is used as the final measurement. That is:

$$dist_{ctx}(t_i, t_j) = \frac{max(logf(t_i),logf(t_j)) - logf(t_i,t_j)}{logN - min(logf(t_i),logf(t_j))} \tag{4.7}$$

$$dist(t_i, t_j) = dist_{ctx}(t_i, t_j) + \alpha dist_{kb}(t_i, t_j) \tag{4.8}$$

$$\xi_{t_{ij}} = exp(-dist(t_i, t_j)^2/\sigma) \tag{4.9}$$

where $f(t_i)$ is the frequency of tag t_i in dataset D, $f(t_i,t_j)$ is the co-occurrence of tag pair (t_i, t_j), α is set to balance two metrics, σ is the medium value of ξ. We define the semantic dependency constraint L_{sem} at the regional level, which is carried out as:

$$L_{sem} = \min_{P_{B_R}} \frac{1}{C^2} \sum_{i=1}^{C} \sum_{j=1}^{C} \xi_{t_{ij}} \|p'_{t_i} - p'_{t_j}\|^2 \tag{4.10}$$

where p'_{t_i} stands for the confidence scores of the corresponding region batch annotated by t_i, $|p'_{t_i}| = m \cdot S \cdot (2k' + 1)$, which can be viewed as the region distribution of the corresponding tag. B_R stands for the outputs of the input batch at the regional level, where $|B_R| = (m \cdot S \cdot (2k' + 1)) \times C$. As we can see, the semantic dependency constrains the semantically similar tag pair has the close region distributions. To further speed up the training process, by referring to (Zhu et al. 2010), we use the matrix form of this constraint, let Q be the diagonal matrix, then the semantic dependency constraint can be written as:

$$Q_{ii} = \sum_{i \neq j} \xi_{t_{ij}} \quad i,j \in [1, C] \tag{4.11}$$

$$L_{sem} = \min_{P_{B_R}} \frac{1}{C^2} Tr[P_{B_R}^T (Q - \xi) P_{B_R}] \tag{4.12}$$

4.2.5 User-Error Sparsity

Considering people share the general knowledge about the semantic annotation and each image is usually assigned with few tags compared to the entire tag set. Therefore, the errors of user-provided tags on a batch of input images are sparse. Let $G_B(|G_B| = (S \cdot (2k' + 1)) \times C)$ be the user-provided annotation matrix of the

input batch, which has the same dimensions as the batch of confidence scores P_B at the image level. Each row vector in G_B represents the user annotation for each image. The difference matrix between G_B and P_B is the user-error matrix. Thus, the sparsity constraint L_{err} is defined as follows:

$$L_{err} = \min_{P_B} \|G_B - P_B\|_1 \tag{4.13}$$

as we can see, the error sparsity constrains the user-provided error on a batch of input images sparse by minimising the l_1 norm of the corresponding difference matrix between the ideal and user annotation.

4.2.6 Training and Prediction

The VGG-16 configuration is used for initialising the convolutional network component in our model. The training process is two-staged: we first train the whole regional network on the user-provided tags for the fast convergent and get a good initialisation for the weakly-supervised training process. Then we train the model based on the proposed constraints. We train new layers in all models for thirty epochs, with learning rate of 0.001 from the start and decrease it to one-tenth every ten epochs. Stochastic gradient descent (Bottou 2010) is used to optimise the models. We extract three types of low-level image features: 73-d edge direction histogram, 59-d LBP and 256-d SIFT bag of words. After we obtain these features, l_2 normalisation is applied to concatenate all the features together, and they are further encoded into a 280-d feature vector by PCA and whitening. The low-level features are used to measure the visual similarity at both the image level and regional level, which is further used to guide the input selection and visual consistency constraint as addressed in Section 4.2.3. The grid-search strategy is adopted to tune the hyperparameters including $\lambda_1, \lambda_2, \epsilon$, and α by referring to the previous works (Zhu et al. 2010, Li & Tang 2017). Moreover, to verify the proposed constraints, we visualise them in the next section.

	Img number N	Tag set size C	Expert Label number	User tags per img
Mirflickr	25,000	444	14	2.7
NUS-WIDE	201,302	3010	81	6.8

Table 4.1: The statistics of the Mirflcikr 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.

4.3 Experiments

In this section, we present our experimental results and analyse the effectiveness of our proposed model. Our model is evaluated on two benchmark datasets: Mirflickr (Huiskes & Lew 2008) and NUS-WIDE (Chua et al. 2009). By comparing with the baselines and state-of-the-art models, we show that the proposed model achieves the best performance.

4.3.1 Data Preprocessing

Images and user-provided tags of the Mirflickr and NUS-WIDE dataset are obtained from the Flickr website. The tags are free-form and need to be unified to conduct adequate research. Besides, for a tag to be meaningful, it needs to be assigned to a certain number of images. Therefore, we carry out the preprocessing as follows to obtain training and test sets: first, we lemmatise all the tags to their dictionary forms and remove the ones that do not appear in the WordNet, then we exclude the tags that do not meet the occurrence threshold (0.1% of total image number). We evaluate all the models on tags which are manually corrected (noted as the expert label in this section). We choose 5,000 and 50,000 training images for Mirflickr and NUS-WIDE respectively. The experiments are repeated five times, and average results are reported. The statistics of the two datasets are shown in Table 4.1.

4.3.2 Evaluation Metrics

Several metrics are employed to evaluate the performance of the proposed model and state-of-the-art methods. The results of the image annotation and the tag refinement are both reported. We refer to the previous works (Li et al. 2016, Li & Tang 2017) to compute the average precision (AP). For each image, a good model should rank relevant tags before the irrelevant ones. Moreover, for a given tag query, relevant images should be returned first before the irrelevant ones. Therefore, we use the mean image average precision (miAP) and mean average precision (mAP) to measure the model performance. miAP is computed by averaging the APs on all the images, while the mAP is computed by averaging the APs on all the given tags. Similar to AP, the global and average performance of AUC is measured. MicroAUC is computed by concatenating all the tag probability vectors together and average the AUC, while MacroAUC is calculated by averaging the mean AUC of each given tag.

4.3.3 Baselines and Compared Methods

We give the descriptions of the baselines first and then list all the compared state-of-the-art methods of the image annotation and tag refinement:

RandomGuess (Li et al. 2016) This is a baseline for the image annotation. Given a new image, RandomGuess assigns tags by randomly selecting from the tag set. We run RandomGuess eighty times, and evaluate it by averaging the predicted scores.

UserTag (Li et al. 2016) This is a baseline for the tag refinement. All the user-provided tags are reserved, and the performance is evaluated based on them.

KNN (Makadia et al. 2010) This is a baseline for the image annotation. KNN model measures the tag relevance respect to the given image by retrieving the k nearest neighbours from the image set. Then the tags are assigned based on their

occurrence rates among the neighbours. The image feature used in this model is the 4096-d vector extracted by VGG-16.

Multi-CNN This baseline has the same configurations as VGG-16, except the dimension of the last layer is modified to $|T| = C$. For the image annotation, we train the neural network using the aforementioned training data with logistic loss and test it on the rest of the set. For the tag refinement, the model is trained on the whole dataset to better evaluate the refinement performance for large-scale datasets.

Att-CNN Since the proposed model uses the image regions as the network input; we also set up a baseline using the attention mechanism (You, Jin, Wang, Fang & Luo 2016) to attend on the image regions instead of the bottom-up proposal method. The VGG-16 is used to extract the convolutional features with size $14 \times 14 \times 512$; then the features are sent into a two-part sub-net to capture tag-wise attention. The first part receives the features with two convolutional layers followed by a channel-wise SoftMax operation to generate tag-wise attention map with size $14 \times 14 \times C$, while the second part feeds the features into a convolutional layer to produce the confidence map with size $14 \times 14 \times C$. Element-wise multiplication followed by a sum-pooling is carried out on two maps for the image level prediction. This baseline is trained with the same configurations as the Multi-CNN baseline.

Regional-CNN This is a baseline for the initialisation of the proposed model, which has the same network architecture but trained on the user-provided tags without the proposed constraints.

Multi-CNN+L_{vse} (Zhang, Wu, Zhang, Shen & Lu 2018) This is a baseline model, which is the regular Multi-CNN trained with the proposed constraints.

Compared Methods For the image annotation, we compare with several state-of-the-art methods, including TagProp (Guillaumin et al. 2009), CCA (Murthy,

Maji & Manmatha 2015*b*), TagFeature (Chen et al. 2012), TagExample (Li & Snoek 2013) and WDNL (Li & Tang 2017). For the tag refinement, we compare with TagCooccur (Sigurbjörnsson & Van Zwol 2008), TagVote (Li, Snoek & Worring 2009), and RPCA (Zhu et al. 2010). For a fair comparison, the compared models and baselines are using the pre-trained VGG-16. We implement an equivalent model of WDNL, named NMF, by using the fixed VGG-16 feature matrix as the input with a low rank matrix decomposition.

4.3.4 Results on Image Annotation

For the image annotation task, we train our model on the training data as described in the preprocessing. We set the initial image candidate size $k = 100$, one batch B is formed by $S = 4$ mini-batches, where each mini-batch is comprised of one image i with $k' = 8$ positive and negative image pairs. Each image contains $m = 10$ representative regions.

Table 4.2 and Table 4.3 indicate that the proposed model, noted as Regional-CNN+L_{vse}, achieves the best performance on all the evaluation metrics of both datasets. As we can see, all methods outperform the baseline RandomGuess, which proves that learning with the user-provided tags is useful for the image annotation task. KNN, multi-CNN, Att-CNN and Regional-CNN are four baselines that utilise the image visual information to conduct the annotation. KNN uses the visual similarities at the image level to retrieve the possible tags from image neighbours, while Multi-CNN learns the classifier for each tag independently. Compared to them, Att-CNN and Regional-CNN capture the image visual information at the regional level, which surpasses the baselines performed at the image level, since the multiple tags assigned to one image may correspond to the regions instead of the whole image. It is worth noting that Regional-CNN, which uses the bottom-up proposed method, achieves better results against the attention mechanism. This is because that attention map is learned from the image level supervised information, it is hard to capture the attention when social images are assigned with diverse user-provided tags. On the contrary, the bottom-up proposal method, that independently relies on the image content, is not affected by the im-

Method	miAP	mAP	MicroAUC	MacroAUC
RandomGuess (Li et al. 2016)	0.072	0.072	0.501	0.498
KNN (Makadia et al. 2010)	0.243	0.499	0.785	0.926
Multi-CNN	0.404	0.556	0.865	0.925
Att-CNN	0.411	0.562	0.853	0.926
Regional-CNN	0.426	0.571	0.869	0.926
Multi-CNN+L_{vse} (Zhang, Wu, Zhang, Shen & Lu 2018)	0.449	0.591	0.893	0.941
CCA (Murthy et al. 2015*b*)	0.318	0.325	0.753	0.837
NMF (Li & Tang 2017)	0.422	0.412	0.882	0.865
TagProp (Guillaumin et al. 2009)	0.386	0.518	0.822	0.907
TagFeature (Li & Snoek 2013)	0.313	0.414	0.786	0.892
TagExample (Li & Snoek 2013)	0.324	0.537	0.728	0.915
Regional-CNN+L_{vse}	**0.501**	**0.627**	**0.906**	**0.948**

Table 4.2: The image annotation results on the Mirflickr dataset.

Method	miAP	mAP	MicroAUC	MacroAUC
RandomGuess (Li et al. 2016)	0.023	0.023	0.500	0.504
KNN (Makadia et al. 2010)	0.388	0.357	0.916	0.941
Multi-CNN	0.405	0.369	0.922	0.934
Att-CNN	0.431	0.376	0.924	0.936
Regional-CNN	0.433	0.384	0.926	0.939
Multi-CNN+L_{vse} (Zhang, Wu, Zhang, Shen & Lu 2018)	0.412	0.398	0.922	0.942
CCA (Murthy et al. 2015*b*)	0.363	0.364	0.865	0.928
NMF (Li & Tang 2017)	0.383	0.369	0.910	0.923
TagProp (Guillaumin et al. 2009)	0.359	0.373	0.921	0.930
TagFeature (Li & Snoek 2013)	0.240	0.302	0.831	0.906
TagExample (Li & Snoek 2013)	0.356	0.335	0.919	0.924
Regional-CNN+L_{vse}	**0.440**	**0.423**	**0.932**	**0.950**

Table 4.3: The image annotation results on the NUS-WIDE dataset.

precision brought by user-provided tags, which is the reason why we choose the regional network as our backbone architecture. The proposed model outperforms these four baselines in a large margin, which indicates the significance of proposed constraints. The better results against Multi-CNN+L_{vse} (Zhang, Wu, Zhang, Shen & Lu 2018) proves the effectiveness of exploring the image information at the regional level.

As for the compared methods, CCA (Murthy et al. 2015*b*) learns the latent space for both the image visual feature and tag semantic feature, and models their correlation, while Tagfeature (Chen et al. 2012) first learns the classifiers for each tag, then recycles the classifiers' predictions as the semantic representations and concatenates them with the image visual feature to retrain the classifiers. Since we also utilise both the image visual information and tag semantic information, the better results against these methods show that eliminate the tag noise is necessary for the social image annotation. TagProp (Guillaumin et al. 2009) is a trained nearest neighbour approach, which models the tag distribution by the metric learning. Different from Tagprop, the metric learning is conducted to constrain the semantic representation of the region pairs in our model. TagExample (Li & Snoek 2013) is proposed to progressively select the relevant positive and negative samples for training the annotation model. Compared to this method, our model trains the annotation network with proposed constraints, which implicitly modifies the training samples. NMF (Li & Tang 2017) uses the low-rank matrix decomposition with the user-error sparsity to eliminate the tag noise. However, the fixed feature matrix is used as the input to conduct the matrix decomposition, which can consume large computation resources when facing the large-scale dataset. Instead of the fixed inputs, we enable the feature learning by exploring the dataset intrarelationships, including the visual consistency among image regions and semantic dependency between tag pairs at the regional level, which achieves the significant better results. Some qualitative results are shown in Figure 4.6.

Moreover, we compare the average precision between the proposed model and baselines. The results are shown in Figure 4.7. As we can see, the majority of values are above $y = x$, which proves the effectiveness of our model on the whole tag set.

4.3.5 Results on Tag Refinement

For the tag refinement, we train our model on the entire dataset to show the effectiveness of our model when refining the large-scale datasets. We set parameters same as the annotation experiment.

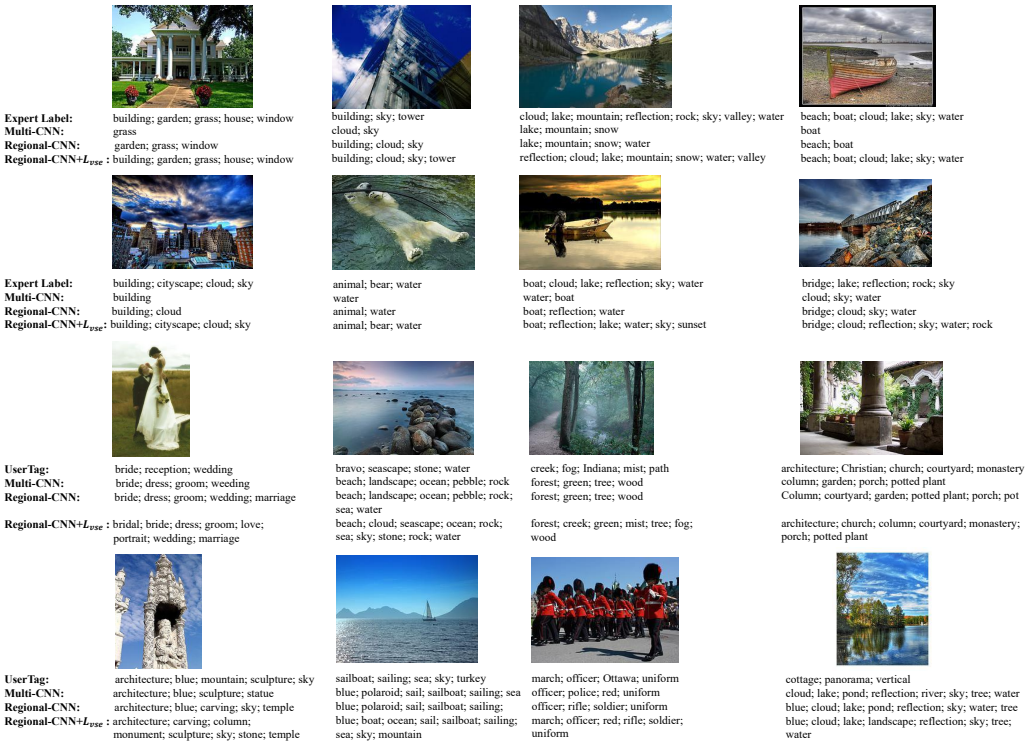


Figure 4.6: The example results of the image annotation and tag refinement. The first two rows are the image annotation examples, while the last two rows are the tag refinement results.

Method	miAP	mAP	MicroAUC	MacroAUC
UserTag (Li et al. 2016)	0.100	0.263	0.544	0.642
Multi-CNN	0.426	0.597	0.872	0.937
Att-CNN	0.448	0.606	0.885	0.939
Regional-CNN	0.467	0.612	0.891	0.940
Multi-CNN+L_{vse} (Zhang, Wu, Zhang, Shen & Lu 2018)	0.476	0.633	0.907	0.952
TagCoocur (Sigurbjörnsson & Van Zwol 2008)	0.159	0.260	0.587	0.699
TagVote (Li, Snoek & Worring 2009)	0.201	0.323	0.594	0.708
RPCA (Zhu et al. 2010)	0.384	0.541	0.840	0.914
Regional-CNN+L_{vse}	**0.494**	**0.662**	**0.913**	**0.960**

Table 4.4: The tag refinement results on the Mirflickr dataset.

Table 4.4 and Table 4.5 indicate that the proposed model outperforms the baselines and state-of-the-art models on both datasets. As we can see from two tables,

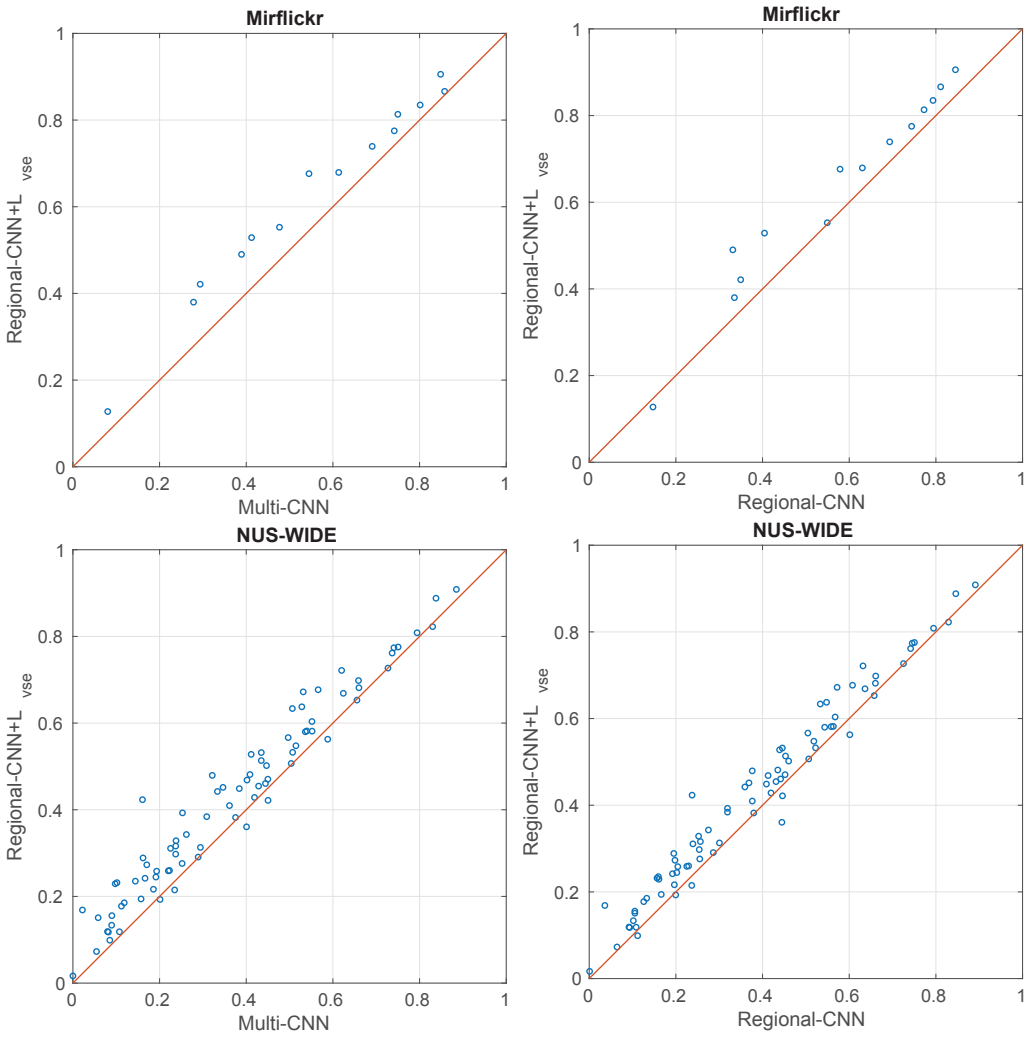


Figure 4.7: The AP comparisons of the proposed model Regional-CNN+L_{vse} against baselines Multi-CNN and Regional-CNN of the image annotation on both datasets.

the baseline UserTag directly uses user-provided tags without any refinement, while different refinement methods have shown different degrees of improvements against it. TagCoocur (Sigurbjörnsson & Van Zwol 2008) only uses the tag co-occurrence rate and frequency to rank the tags respective to the image, no visual information involved, while the TagVote (Li, Snoek & Worring 2009) also considers the visual similarities among images to refine the ranking results. However, by feeding the selected region pairs into the regional neural network and trained

Method	miAP	mAP	MicroAUC	MacroAUC
UserTag (Li et al. 2016)	0.187	0.338	0.656	0.783
Multi-CNN	0.424	0.416	0.921	0.935
Att-CNN	0.438	0.426	0.931	0.949
Regional-CNN	0.439	0.430	0.931	0.950
Multi-CNN+L_{vse} (Zhang, Wu, Zhang, Shen & Lu 2018)	0.431	0.446	0.927	0.950
TagCoocur (Sigurbjörnsson & Van Zwol 2008)	0.277	0.298	0.650	0.818
TagVote (Li, Snoek & Worring 2009)	0.311	0.368	0.905	0.864
RPCA (Zhu et al. 2010)	0.404	0.426	0.918	0.872
Regional-CNN+L_{vse}	**0.457**	**0.470**	**0.941**	**0.958**

Table 4.5: The tag refinement results on the NUS-WIDE dataset.

with proposed constraints, our model can perform the feature learning while conducting the refinement, which outperforms these methods. The better results against the baselines Multi-CNN, Att-CNN, and Regional-CNN, again proves the effectiveness of the proposed constraints, and the superior results against Multi-CNN+L_{vse} (Zhang, Wu, Zhang, Shen & Lu 2018) also indicates the significance of exploring the image regional information. Moreover, we achieve better results compared to the most relevant refinement method RPCA. As illustrated in (Li et al. 2016), since the RPCA optimises the whole tagging matrix, it could not be easily applied to the large-scale datasets due to its high demand in both CPU time and memory. On the contrary, we formulate three constraints at both the image level and regional level, which activates the feature learning and also make our model flexible and stable to deal with the large-scale datasets. We give some examples in Figure 4.6 to show the refinement results. As we can see, our proposed model removes the inaccurate user-provided tags and adds relevant tags to images. Similarly, we also compare the AP against baselines in Figure 4.8 to show the significance of our model on the whole tag set.

4.3.6 Ablation Analysis

To further investigate the individual contribution of each constraint, we conduct a comprehensive ablation study. We combine the visual consistency and seman-

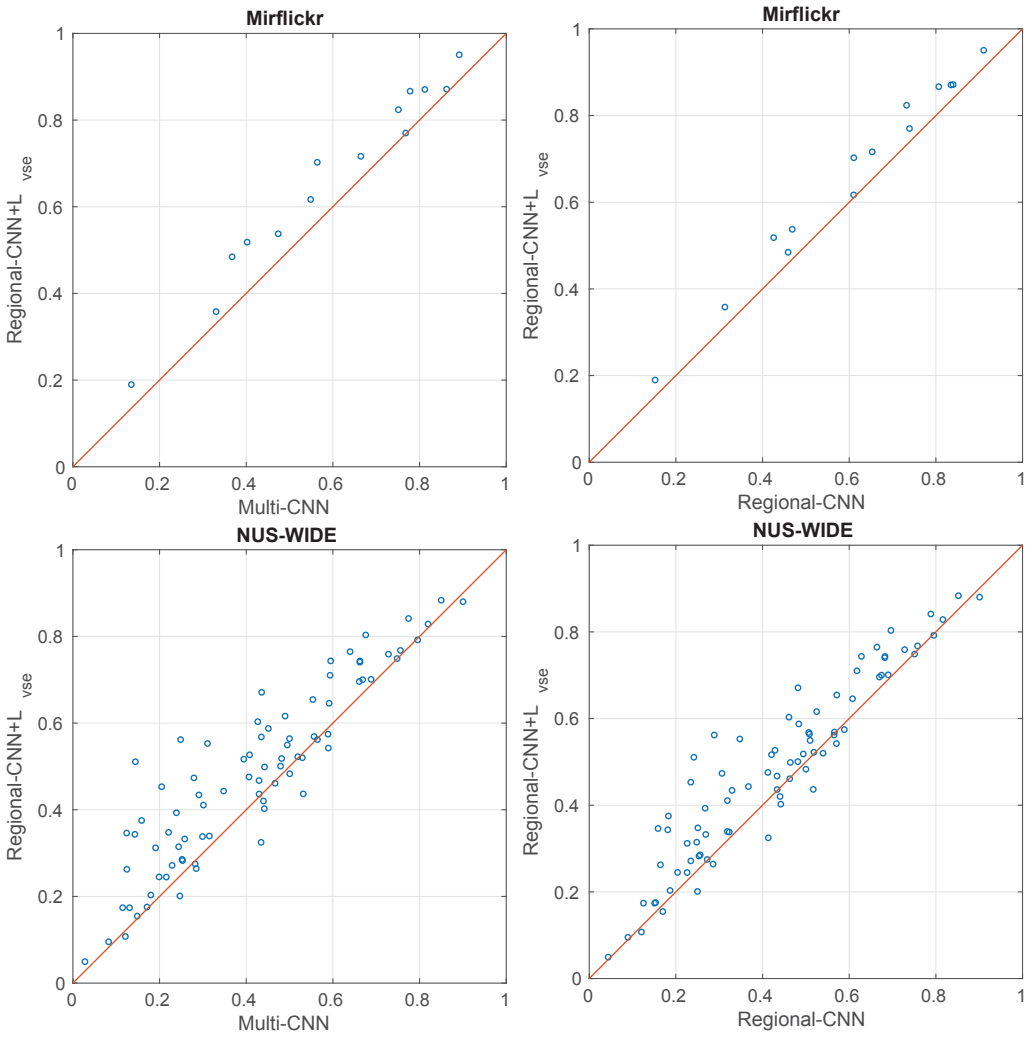


Figure 4.8: The AP ratios of the proposed model Regional-CNN+L_{vse} against baselines Multi-CNN and Regional-CNN of the tag refinement on both datasets.

tic dependency with user-error sparsity respectively, noted as Regional-CNN+L_{ve} and Regional-CNN+L_{se}, and compare them with the Regional-CNN and the final model with three proposed constraints Regional-CNN+L_{vse}.

The ablation results of the image annotation are shown in Table 4.6 and Table 4.7, while results of the tag refinement are shown in Table 4.8 and Table 4.9. As we can see, the Regional-CNN+L_{ve} and Regional-CNN+L_{se} show improvements against baseline Regional-CNN on both tasks, which proves the effectiveness of

Method	miAP	mAP	MicroAUC	MacroAUC
Regional-CNN	0.426	0.571	0.869	0.926
Regional-CNN+L_{ve}	0.462	0.617	0.894	0.944
Regional-CNN+L_{se}	0.453	0.609	0.887	0.942
Regional-CNN+L_{vse}	**0.501**	**0.627**	**0.906**	**0.948**

Table 4.6: The ablation results of the image annotation on the Mirflickr dataset.

Method	miAP	mAP	MicroAUC	MacroAUC
Regional-CNN	0.433	0.384	0.926	0.939
Regional-CNN+L_{ve}	0.435	0.421	**0.932**	0.948
Regional-CNN+L_{se}	0.432	0.417	0.929	0.948
Regional-CNN+L_{vse}	**0.440**	**0.423**	**0.932**	**0.950**

Table 4.7: The ablation results of the image annotation on the NUS-WIDE dataset.

Method	miAP	mAP	MicroAUC	MacroAUC
Regional-CNN	0.467	0.612	0.891	0.940
Regional-CNN+L_{ve}	0.483	0.654	0.910	0.956
Regional-CNN+L_{se}	0.472	0.645	0.905	0.958
Regional-CNN+L_{vse}	**0.494**	**0.662**	**0.913**	**0.960**

Table 4.8: The ablation results of the tag refinement on the Mirflickr dataset.

Method	miAP	mAP	MicroAUC	MacroAUC
Regional-CNN	0.439	0.430	0.931	0.950
Regional-CNN+L_{ve}	0.456	0.467	**0.941**	0.957
Regional-CNN+L_{se}	0.453	0.454	0.939	0.956
Regional-CNN+L_{vse}	**0.457**	**0.470**	**0.941**	**0.958**

Table 4.9: The ablation results of the tag refinement on the NUS-WIDE dataset.

the sole constraint. Since the different constraints are designed from the different aspects, combining them can mutually remedy each other, and further improves the overall performance.

Metro	Underground	Fog	Mist
Aircraft	Airplane	Bug	Insect
Auto	Car	Ocean	Sea
Chocolate	Dessert	Bride	Wedding
Art	Sculpture	Railroad	Train

Table 4.10: The high-dependent tag pairs.

4.3.7 Visualisation of the Constraints

To verify the proposed constraints, we visualise them in this section, the visual consistency is shown first, followed by the semantic dependency, and finally, we show the user-error sparsity.

Visual Consistency We first select three pairs of visually similar regions, which are indicated by the red and blue boxes in Figure 4.9. Then we visualise their tag probability distributions. The images are sampled from the Mirflickr dataset, where the tag set size $C = 444$. Each row under the region pair is the corresponding visualised tag distributions. The dimension of each row is $2 \times C$. The brighter the row element is, the higher the confidence score it has. As we can see, the tag distributions of the visually similar regions are close.

Semantic Dependency We show top ten high-dependent tag pairs in Table 4.10. Moreover, we visualise three tag pairs' region distributions on randomly selected 100 regions in Figure 4.10. Each row stands for the corresponding region distributions for each tag pair with the size 2×100. The region distribution is shown as the probability of the image annotated by the corresponding tag. Same as the visual consistency, the brighter the element is, the higher probability it has. As we can see, the high semantic dependent tag pairs have similar annotation distributions.

User-Error Sparsity We compute the user-error matrix on one batch of input images from the Mirflickr dataset and visualise it in Figure 4.11. The brighter the matrix element is, the higher the error score it has. As we can see, it verifies

Figure 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.

the assumption of the sparsity of the user-error matrix, which is used as one of the proposed constraint in the weakly-supervised neural network. This is also corresponding to the similar assumptions drawn from classic related works that people share the general knowledge about the semantic annotation and each image is usually assigned with few tags compared to the entire tag set. Therefore, the errors of user-provided tags on a batch of input images are sparse.

Figure 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.

Figure 4.11: The visualisation of the user-error matrix with the size 68×444.

4.4 Summary

The social image annotation and tag refinement have been the research focus since the image sharing networks become popular. In this work, we propose to solve these problems in a weakly-supervised manner. By adopting a regional deep neural network as the backbone model architecture, we propose to simultaneously refine the training tags and learn the annotation model. The visual consistency, semantic dependency, and user-error sparsity are introduced as the network con-

straints, which explore the image and tag relationships at both the image level and regional level. The advanced experimental results against state-of-the-art methods on two benchmark datasets demonstrate the significance of the proposed model.

Chapter 5

Annotation via Collective Knowledge

5.1 Introduction

The mainstream annotation methods are mainly proposed for the image annotation of common objects and their attributes and have achieved satisfactory results. However, there are two major drawbacks when they handle the more diverse image annotation. First of all, state-of-the-art methods on the image annotation problem use CNN as the backbone model, which heavily relies on the fine-tuning the pre-trained image representation on the large-scale image dataset ImageNet (Krizhevsky et al. 2012). However, the labels of the real-world image can be more obscure and diverse. Take Figure 5.1 as an example. The left image is sampled from the Flickr dataset (Huiskes & Lew 2008). These labels indicate the objects and their attributes, which are tightly related to the image visual content, while the right image is selected from the heritage image collection we collected online, it contains not only the easily recognised labels 'crowd' and 'city street', but also labels that can be inferred from context or related knowledge, such as 'anniversary' and 'festival'. Even for the same label, the image visual appearance can be very different. For example in Figure 5.2, all images are annotated with the label 'theatre.' However, the visual appearance of these images is quite diverse.

(a)
blue, cloud, landscape, sunset, sky

(b)
procession, crowd, city street, band, festival, anniversary

Figure 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.

(a) (b) (c) (d)

Figure 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.

Moreover, the domain difference between the training image set and ImageNet can be much larger than current datasets. Directly learning the image visual representation from the training set could degrade the performance. Secondly, since the labels of the real-world images can be very diverse, and classic annotation methods tend to treat labels as equally related to the image visual content or rule out ones that are less relevant, which makes their performances unsatisfactory.

To address above issues, we propose to uncover relationships between the image and its neighbours and utilise them to conduct the image representation learn-

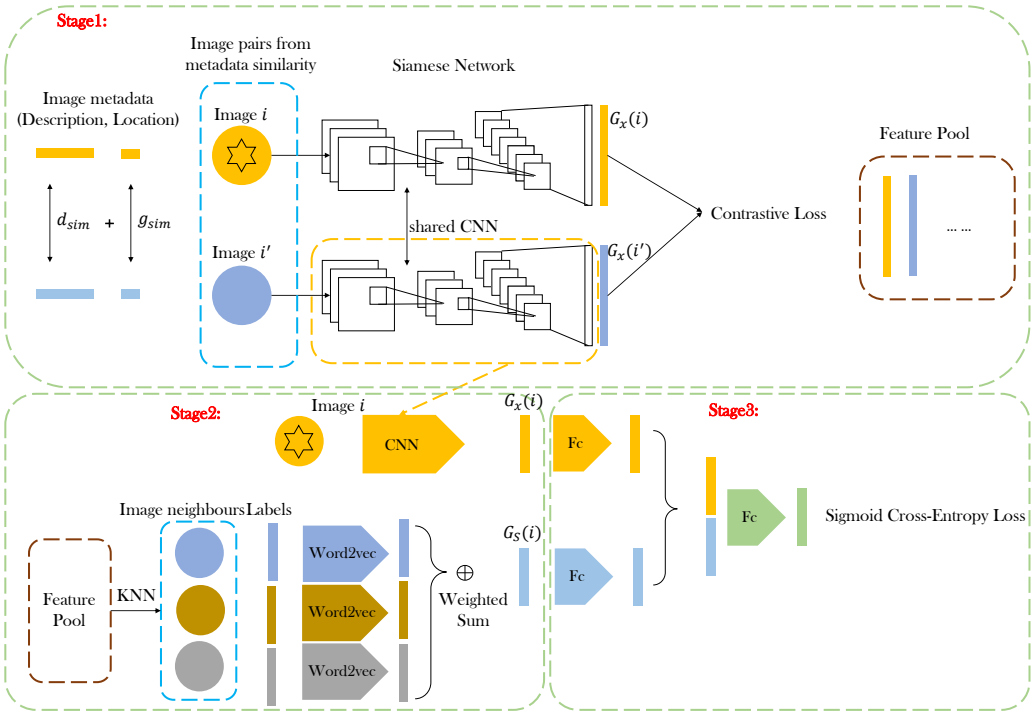


Figure 5.3: The framework of the proposed model. At the training stage, we first generate image pairs (i, i') for the visual representation learning based on the metadata similarity, in our case, the combination of the description similarity d_{sim} and location similarity g_{sim}. Image pairs are passed 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.

ing and train an annotation model, namely annotation via collective knowledge. We ground our proposed model on the heritage image collection, given its visual and label diversity. The whole framework is shown in Figure 5.3. Specifically, we define neighbourhood relationships between images based on their metadata, which are the descriptions of the background of images and locations. Images with similar contexts have the similar visual appearance, which can help us to eliminate the visual ambiguous. Therefore, we conduct the metric learning among

image neighbours to learn effective visual representations. To handle the diversity of the label set, we collect label candidates from image neighbours and embed their semantic information with the visual representation to help infer all possible labels. The deep neural network is adopted in the representation learning and the annotation model. We collect a heritage image collection from the library online open data as the benchmark dataset to verify the effectiveness of the proposed model. In summary, the main contributions of our model are as follows:

1) We propose to learn the image visual representation based on the metric learning. Different from previous works, we use metadata to measure similarities among image neighbours. The superior experimental results against hand-crafted features and CNN finetuned features indicate the significance of our proposed representation learning methods.

2) Different from previous works, our model annotates all labels via collective knowledge from image neighbours, including the visual representation learning and the semantic embedding, which is crucial for the real-world image annotation, since it focuses on not only the visually related labels but also the semantical labels which are related to the events.

3) We collect a heritage image collection as the benchmark dataset and ground our proposed model on it. Experimental results show that our model achieves the best performance against the state-of-the-art annotation methods.

5.2 Proposed Model

5.2.1 Model Overview

The key characteristic of our model is that we use collective knowledge to solve the image annotation problem, which is grounded on the heritage image collection. The collective knowledge is reflected from two perspectives. One is the image visual representation. We uncover relationships between the image and its neighbours by measuring similarities among their metadata and conduct the metric learning to obtain the effective visual representation. The other is that we

generate the semantic representation for the image by summarising contributions from its neighbours' labels. The visual and semantic representations are embedded together and passed through a neural network to finalise the annotation model. The entire model is shown in Figure 5.3.

In the following part, the metadata similarity measurement and the metric learning for the image visual representation are first introduced in Section 5.2.2, and the label relevance analysis and the semantic representation are described in Section 5.2.3. The final annotation model and training details are summarised in Section 5.2.4 and Section 5.2.5 respectively.

5.2.2 Visual Representation Learning

Considering the visual diversity of the training set, and labels are related to the image visual content in varying degrees, directly finetuning the visual representation of the deep neural network towards the training ground-truth can compromise the representational capacity. However, metadata as the reflection of the image content can be used to eliminate the image visual ambiguous and locate image neighbours. In our case, the heritage image collection is attached with the abundant information noted by photographers and librarians when cataloguing them. These metadata are presented as the descriptions of historical events when the images were taken, and the locations of where these events happened, which reflect the visual content in a semantical way. See Figure 5.4 for an example. Inspired by (Johnson et al. 2015), we define image neighbours for the heritage collection based on metadata and conduct the metric learning to obtain the effective visual representation.

An ideal annotation model should be flexible to handle the different forms of metadata, to this end, we measure the similarity between metadata nonparametrically to define image neighbours. Let I be the heritage image collection, T be the set of label candidates, and $D = \{(i,t)|i \in I, t \subseteq T\}$ indicates images associated with a set of labels. Metadata of the heritage image collection can be various, and based on the benchmark dataset we use, we consider the descriptions of historical events and the locations.

(a)
Description: St Patrick's Day procession
Location: Hyde Park, College Street, Sydney

(b)
Description: Royal Easter Show
Location: Sydney Showground

(c)
Description: Women's winter fashion
Location: Snow's Department Store, Pitt Street, Sydney

Figure 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.

The description of a historical event is presented as a free-form phrase. Given the image i, we tokenise the description d_i into words, which results in a vocabulary V of word candidates. Therefore, for each image $i \in I$, a subset $v_i \subseteq V$ represents the description. We use the Jaccard similarity to compare two descriptions, that is, given two images i and i':

$$d_{sim} = |v_i \cap v_{i'}|/|v_i \cup v_{i'}| \tag{5.1}$$

Since each image i only possesses one geographical location g_i, we simply use the indicator function to represent the location similarity:

$$g_{sim} = \mathbf{1}(g_i, g_{i'}) : \begin{cases} 1 & \text{if} \quad g_i = g_{i'} \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

After we have the similarities of the different type of metadata, we now give the final form of the metadata similarity, where λ is used to balance the two metrics:

$$meta_{sim} = d_{sim} + \lambda g_{sim} \tag{5.3}$$

Based on the metadata similarity, we apply the nearest neighbour approach to generate image pairs for the metric learning. For each image i, we select top n_p similar images to form positive pairs, while negative pairs are composed of top n_n dissimilar images. We pass image pairs through the siamese network to conduct the metric learning for the visual representation. The siamese network

consists of two identical neural networks; each receives one of input image pairs, the last layers of networks are fed to a contrastive loss function (Chopra, Hadsell & LeCun 2005), which captures the similarity between two images. That is, given the image pair i and i', with the label y indicates the similarity, 1 for the positive pair, while 0 for the negative one. The networks are trained by minimising the loss L:

$$D_W = \sqrt{\{G_X(i) - G_X(i')\}^2} \tag{5.4}$$

$$L = y\frac{1}{2}D_W^2 + (1-y)\frac{1}{2}\{max(0, m - D_W)\}^2 \tag{5.5}$$

where G_X is the output of the neural network that stands for the visual representation of the input image, m is the margin. D_W measures the Euclidean distance between two image representations $G_X(i)$ and $G_X(i')$. The contrastive loss L constrains the image pairs with the similar metadata to have the closer visual representations while pushing the dissimilar pairs beyond the margin m. Since the metadata is the reflection of the image content, by training based on the metadata similarity, we can eliminate the image visual ambiguous and locate image neighbours, and by training with image pairs, we avoid finetuning the visual representation directly based on the diverse ground-truth. The experimental results, which are presented in Section 5.3, prove the effectiveness of the metric learning for the visual representation.

5.2.3 Label Relevance Analysis & Semantic Summarisation

Motivation Since we consider the annotation for the individual image via collective knowledge from its neighbours, it is important to evaluate each label's relevance to the image content. We choose evaluation metrics from two aspects: the semantic and visual information, namely the semantic field (Zhu et al. 2012) and the neighbour voting, which is orthogonal to each other.

Since the labels are relevant to the image visual content in various degree, some of them are easier predicted based on the combination of semantical information than the individual image. We generate a semantic representation of the image to help infer all possible labels by summarising label information from its

neighbours. This is intuitive that when librarians manually annotate the heritage image collection, they not only observe the individual image but also look into the archived collection to find connections between images, and consider whether to transfer the labels.

Label Relevance Analysis Given the image i, semantic field evaluates the label $\tau \in t$ based on average semantic similarities between τ and other labels that are associated with the image i. We consider the semantic similarity from two aspects: the context of the current image collection and the general knowledge base. Given two labels τ_j and τ_k, the context similarity refers to the semantic similarity that is obtained based on the context information of the current image collection. We adopt the normalised Google distance (Cilibrasi & Vitányi 2007), where context statistics, including the label frequence and label pair co-occurrence, are sampled from the image collection instead of the webpage. For the general knowledge base, we use the Euclidean distance between word2vec vectors that are pre-trained from the Google News corpus (Mikolov, Sutskever, Chen, Corrado & Dean 2013). We combine two metrics to balance the general and domain-specific knowledge for the given collection, that is:

$$s_{ctx}(\tau_j, \tau_k) = \frac{max(logf(\tau_j), logf(\tau_k)) - logf(\tau_j, \tau_k)}{logN - min(logf(\tau_j), logf(\tau_k))} \tag{5.6}$$

$$s(\tau_j, \tau_k) = s_{ctx}(\tau_j, \tau_k) + \alpha s_{w2v}(\tau_j, \tau_k) \tag{5.7}$$

$$\xi_{\tau_{j,k}} = exp(-s(\tau_j, \tau_k)^2/\sigma) \tag{5.8}$$

where $f(\tau)$ indicates the frequence of label τ in the image collection, $f(\tau_j,\tau_k)$ is the co-occurrence of label τ_j and τ_k, α is the weight, σ is the medium value of ξ, and N is the total number of images in the collection. Then the semantic filed similarity score of label τ_j for image i can be computed by:

$$r_{ss}(\tau_j|i) = \frac{1}{|t|-1} \sum_{k=1,k\neq j}^{|t|-1} \xi_{\tau_{j,k}} \tag{5.9}$$

$$r(\tau_j|i) = \text{CombSUM}(r_{ss}(\tau_j|i), r(\tau_j|i)) \tag{5.10}$$

The neighbour voting measures the label relevance of τ_j with respect to the image i by examining the visually similar images. Based on the visual represen-

tation we obtained from the last step, the metric retrieves multiple nearest neighbours, and the number of images annotated with label τ_j is used as the similarity score $r_{nn}(\tau_j|i)$. Finally, two relevance scores are normalised and merged by the CombSUM as $r(\tau_j|i)$. By performing the label relevance analysis, we convert the ground-truth of the training set from the hard assignment 0-1, where 0-1 denotes the absence-present of a label, to a confidence vector, where each dimension indicates the label relevance (0 occupies the absent label). Considering the diversity of the label set, this step can benefit the summarisation of collective knowledge by leveraging the label relevance.

Semantic Summarisation Given image i, we retrieve m nearest neighbours based on the visual representation to form a candidate set Z_i, each image $i_j \in Z_i$ is associated with a relevance vector r_{i_j}, where $|r_{i_j}| = |T|$, we reserve the KD-tree structure for training set after we perform the nearest neighbour approach, which will be used during the label prediction for new images. The summarised label information of i is indicated as:

$$\varphi_{i,i_j} = exp(-\|G_X(i) - G_X(i_j)\|^2/\sigma) \tag{5.11}$$

$$G_S(i) = \sum_{j=1}^{m} \varphi_{i,i_j}(r_{i_j} \cdot H) \tag{5.12}$$

where σ is the medium value of φ. φ_{i,i_j} is the similarity of the image pair i and i_j, H is the matrix of the pre-trained word2vec, and each row of H stands for the word2vec of a label. $G_S(i)$ gathers all the label information from image neighbours based on the image similarity and the label relevance of each neighbour. The reason that we adopt the weighted sum is to better leverage the importance of the annotation carried by the image neighbour with respect to the query image. Therefore, we can summarise the semantic information in a more accurate way.

5.2.4 Annotation via Collective Knowledge

For each image $i \in I$, we now have a visual representation $G_X(i)$ by the metric learning and a semantic representation $G_S(i)$ by summarising from its neighbours. We compute h_1 and h_2 dimensional hidden states for each image's visual

and semantic representations respectively, by a fully connected layer followed by a ReLU nonlinearity transform ψ, which are parameterised by (w_X, b_X), and (w_S, b_S). At this point, we concatenate hidden states of visual and semantic representation, and feed into a third fully connected layer parametered by (w_P, b_P) to obtain the probabilities of labels, that is:

$$\vartheta_X = \psi(w_X \cdot G_X(i) + b_X) \tag{5.13}$$

$$\vartheta_S = \psi(w_S \cdot G_S(i) + b_S) \tag{5.14}$$

$$P(t|i) = w_P \cdot [\vartheta_X; \vartheta_S] + b_P \tag{5.15}$$

where ϑ_X and ϑ_X indicate the hidden states of visual and semantic representations respectively. The neural network can be trained by minimising the sigmoid cross-entropy loss towards the image ground-truth.

5.2.5 Training and Prediction

We use the VGGNet-16 as our backbone network for the metric learning, the output of the fc7 layer with ReLU transform is used as the visual representation. The training process is three-stage: first, we generate positive and negative image pairs for the visual representation learning based on similarities between image metadata. Then, we measure the label relevance of the image content by exploring the semantic and visual information from image neighbours, and summarise the semantic representation for each image based on collective knowledge. And finally we fuse the visual and semantic representations and feed it into fully connected layers. The whole training process is shown in Algorithm 5.1.

As for the label prediction during the test, when a new image k arrives, we first extract its visual representation by $G_X(k)$, since we build the KD-tree during training, we can easily query its neighbours from the training set, and summarise the semantic representation $G_S(k)$. Then we feed two representations into the annotation model to compute hidden states and predict labels.

Algorithm 5.1 Training stages of the annotation model.

Input: $D, \forall i \in I, t \subseteq T, v_i \subseteq V$

1: Measure the metadata similarity $meta_{sim}$ of image pairs based on eq. 5.3;
2: Generate positive and negative training samples for the metric learning by minimising eq. 5.5, and obtain the visual representation $G_X(i)$;
3: Evaluate each label's relevance of image i based on eq. 5.10;
4: Summarise the semantic representation $G_S(i)$ based on eq. 5.12;
5: Compute the hidden states of representations and concatenate them to pass through fully connected layers for the label generation based on eq. 5.15;
6: Training the network with the sigmoid cross-entropy loss.

5.3 Experiments

In this section, we present the experimental details and results. Our model is evaluated on the heritage image collection we collected online. By comparing with baseline models and state-of-the-art methods, we show that our model achieves the best performance, and comprehensive ablation studies indicate the significance of each component in our model.

5.3.1 Data Preprocessing & Evaluation Metrics

We collect the raw heritage image collection from the library online open data. The dataset contains 37,931 valid images with textual information, including label, description, and location, etc., which are annotated by librarians. To conduct the effective annotation, we lemmatise all the textual words to their dictionary forms, then we exclude labels that blew the occurrence threshold (0.2% of $|I|$) to avoid the insufficient sampling. Finally, images without labels are removed. The preprocessing results in 31,815 images with a size of the label set $|T| = 257$; each image is attached with metadata: a description and a geographic location. We use 21,210 images for the training and 10,605 images for the test, the metadata is only used with training images.

For overall and per-label evaluation metrics, we use the average precision

(AP). As an effective annotation model, the relevant label should be ranked higher than irrelevant ones with respect to the image, and the same applies to the image with respect to a label query. Therefore, we compute both the mean image average precision (imAP), which is averaging APs over all images and the mean average precision (mAP), which is averaging APs over all labels. Moreover, to conduct the quantitative evaluation, we predict up to three highest ranked labels above the threshold for each image to compare against the ground-truth. Overall and per-label precision/recall/F1 score noted as $(O_P, O_R, O_{F1}, C_P, C_R, C_{F1})$ are reported.

5.3.2 Implementation Details

As mentioned in Section 5.2.5, the output of the last two fully connected layers in the VGGNet-16 is used for the visual representation. For the neighbour voting in the label relevance analysis and the semantic summarisation, we retrieve $m = 50$ neighbours. The pre-trained word2vec for each label is queried from the Google News corpus, which is a 300-d vector. And we set the hidden state dimension $h_1 = 512$ and $h_2 = 256$ for the visual and semantic representation respectively. As for the hyperparameters including λ and α, by referring to the previous works (Johnson et al. 2015, Li et al. 2016), we adopt the grid-search to tune them. Our experimental result $\lambda = 0.85$ indicates that, in the given heritage image collection, we rely on the description similarity d_{sim} more, which is reasonable considering multiple historical events can happen in the same location. Similarly, α is used to balance the domain-specific and general knowledge for the given collection. The experimental result $\alpha = 0.76$ illustrates that we value domain-specific knowledge s_{ctx} more when measuring the semantic similarity, and the general knowledge base s_{w2v} is jointly considered to ensure the similarity metric's generalisation. And we set $n_p = 1$ and $n_n = 2$ for the metric learning[1]. We train all models including the siamese network and fully-connected layers for 30 epochs, with the SGD optimisation and the learning rate decreases from 0.001 to 1/10 every ten

[1]We varied the values of n_p and n_n during the experiments and no significant differences were observed.

epochs. During the test, a new image is extracted with the visual and semantic representation and passed through the annotation model to get the prediction.

5.3.3 Baselines and Compared Methods

To evaluate the effectiveness of our proposed model, we implement four baselines for comparisons, here we give the descriptions of these baselines and compared state-of-the-art annotation methods:

RandomGuess All the annotation methods should achieve better results against RandomGuess (Li et al. 2016), which randomly selects a subset of labels. The random prediction is run 50 times, and average evaluation scores are reported.

Multi-CNN This is a standard CNN model without involving any additional information (Zhang, Zhang, Lu, Shen, Curr, Phua, Neville & Edmonds 2016). The VGGNet-16 is used as the backbone model and trained on the image ground-truth with the sigmoid cross-entropy loss. The model is trained for 30 epochs with the SGD optimisation and the learning rate decreases from 0.001 to 1/10 every ten epochs.

KNN This is a simple and widely used annotation baseline model (Makadia et al. 2008). KNN annotates the image by retrieving the m nearest neighbours, and labels are assigned based on the occurrence among neighbours. We set $m = 50$, and use the 4096-d visual representation from Multi-CNN last two fully-connected layers.

CNN+LSTM This is an equivalent model proposed in (Wang et al. 2016), which annotates images by leveraging the image visual-label relationship and label pair dependencies simultaneously at the image level. Tags associated with the image are ranked as an ordered sequence based on the occurrence rate. The image is first sent to the CNN for the visual representation extraction, then the visual representation and labels are recurrently encoded and fed to the LSTM to infer the next

label. The embedded dimensions for the visual representation and label are 512 and 256 respectively.

Compared Methods We compare our proposed model with several state-of-the-art annotation methods, including CCA (Murthy et al. 2015*b*), TagProp (Guillaumin et al. 2009), HCP (Wei et al. 2014) TagFeature (Chen et al. 2012), and TagExample (Li & Snoek 2013). For fair comparisons, the image representation used in these model are the 4096-d vector from Multi-CNN last two fully-connected layers. Moreover, since these methods are not originally designed to take the metadata into consideration, to further show the effectiveness of our proposed model, we implement five comparison experiments that utilise the metadata, including KNN*, CCA*, TagProp*, TagFeature* and TagExample*, where * stands for the image features in these models are obtained after the proposed visual representation learning.

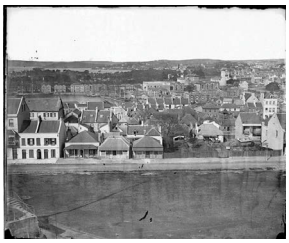
5.3.4 Results on Image Annotation

Table 5.1 shows that our proposed model achieves the best performance on all evaluation metrics against all baselines and state-of-the-art methods. As expected, all methods outperform the baseline RandomGuess in a large margin, which proves the learning from the image ground-truth is necessary for the effective annotation. Multi-CNN baseline directly models the relationship between the image visual content and associated labels, since our model uses metadata for the visual representation learning and summarise the semantic representation from image neighbours, the performance of our model surpasses this baseline. KNN baseline only uses the image visual representation to retrieve relevant neighbours to vote the annotation; therefore, with the more accurate visual representation obtained from the metric learning, our model also outperforms this baseline. CNN+LSTM is the state-of-the-art model for the image annotation, which models the image-label and the label sequence correlation by utilising a unified framework of CNN and LSTM. However, it is hard for LSTM to directly model the diverse label correlation as an ordered sequence.

Table 5.1: Results of the Image Annotation.

Method	imAP	mAP	O_P	O_R	O_{F1}	C_P	C_R	C_{F1}
RandomGuess (Li et al. 2016)	0.008	0.009	0.008	0.011	0.009	0.008	0.000	0.000
Multi-CNN (Zhang et al. 2016)	0.411	0.346	0.541	0.315	0.398	0.515	0.200	0.288
KNN (Makadia et al. 2008)	0.330	0.282	0.399	0.360	0.378	0.425	0.143	0.214
CNN+LSTM (Wang et al. 2016)	0.410	0.342	0.539	0.316	0.399	0.517	0.314	0.391
CCA (Murthy et al. 2015*b*)	0.424	0.367	0.498	0.384	0.434	0.509	0.286	0.366
TagProp (Guillaumin et al. 2009)	0.430	0.366	0.555	0.329	0.413	0.520	0.257	0.344
HCP (Wei et al. 2014)	0.449	0.390	0.536	0.394	0.454	0.519	0.308	0.387
TagFeature (Chen et al. 2012)	0.417	0.357	0.520	0.297	0.378	0.509	0.229	0.315
TagExample (Li & Snoek 2013)	0.425	0.368	0.525	0.353	0.422	0.517	0.314	0.391
KNN*	0.391	0.331	0.475	0.350	0.403	0.464	0.286	0.354
CCA*	0.443	0.386	0.500	0.410	0.450	0.513	0.310	0.387
TagProp*	0.475	0.408	0.541	0.403	0.462	0.529	0.257	0.346
TagFeature*	0.425	0.359	0.499	0.378	0.430	0.506	0.305	0.381
TagExample*	0.441	0.376	0.511	0.389	0.442	0.513	0.342	0.410
Our Model	**0.511**	**0.445**	**0.607**	**0.416**	**0.494**	**0.575**	**0.371**	**0.451**

As for the compared methods, CCA (Murthy et al. 2015*b*) learns a latent space to embed the image and label representation together to enhance the representation, and the similar idea is also used in TagFeature (Chen et al. 2012), where the predictions from label classifiers are concatenated with the image representation to retrain the annotation model. We also learn the hidden states of both image and semantic representation for the annotation, however, since we obtain the semantic representation from image neighbours and we have better image visual representation based on the metric learning, our model outperforms both methods. TagProp (Guillaumin et al. 2009) is a trained nearest neighbour approach, which directly maximises the probability of the label distribution in training set by the integration of the metric learning. Different from this method, our metric learning for the visual representation is performed under the guidance of image metadata instead of modelling the label distribution, considering the diversity of labels. The superior performance against TagProp verifies this assumption. Instead of operating at the image level, HCP (Wei et al. 2014) proposes to apply the bottom-up proposal method to generate image regions, captures image regions with associated labels and fuse them. However, considering the various range of the image visual content in the training set, the bottom-up proposal methods like BING (Cheng



(a)
Ground-Truth: historic building; house; panoramic view; suburb; city view
KNN: historic building; city street; city view
Multi-CNN: historic building; panoramic view
Our Model: historic building; house; city street; panoramic view; city view, suburb

(b)
Ground-Truth: festival; procession; official event; bridge; float procession; band
KNN: festival; anniversary; procession; bridge
Multi-CNN: festival; procession; official event; bridge;
Our Model: festival; procession; official event; bridge; float procession

(c)
Ground-Truth: historic building; children; teacher; bank roof
KNN: historic building;
Multi-CNN: historic building;
Our Model: historic building; children; teacher

(d)
Ground-Truth: crowd; evening clothes; theater foyer
KNN: crowd; theater; association; audience
Multi-CNN: crowd
Our Model: crowd; evening clothes; theater foyer

(e)
Ground-Truth: streetscape; coronation; decoration; city street
KNN: streetscape; commercial establishment; flag
Multi-CNN: crowd; city street; procession; decoration
Our Model: city street; streetscape; decoration

(f)
Ground-Truth: group people; uniform; football team
KNN: group people; football team; children
Multi-CNN: group people; football team
Our Model: group people; uniform; football team

Figure 5.5: Some example annotation results on the heritage image collection.

et al. 2014) fail to capture valid image regions in most cases. TagExample (Li & Snoek 2013) explores both positive and negative training samples by analysing the label relevance with respect to the image content. Compared to this method,

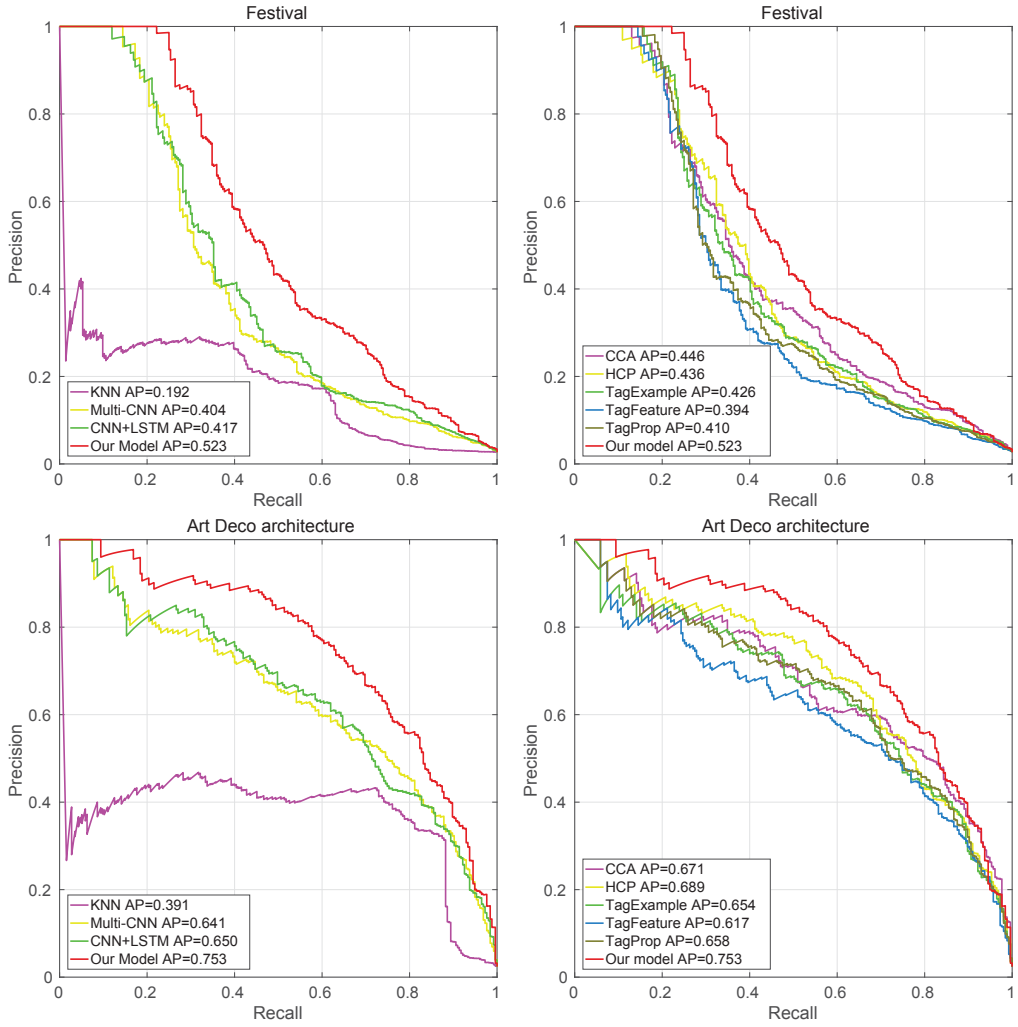


Figure 5.6: (a) The first row is the PR-curves of the label 'festival' compared with baselines and state-of-the-art methods. (b) The second row is the PR-curves of the label 'Art Deco architecture' compared with baselines and state-of-the-art methods. The average precision is also given in the figure. Better view in color.

we use the metadata similarity for the image representation learning to avoid any visual ambiguous caused by the label diversity and apply label relevance analysis later to help summarise the semantic representation from neighbours. The advanced performance indicates the effectiveness of our proposed model. These compared methods are not originally designed to take the metadata into consider-

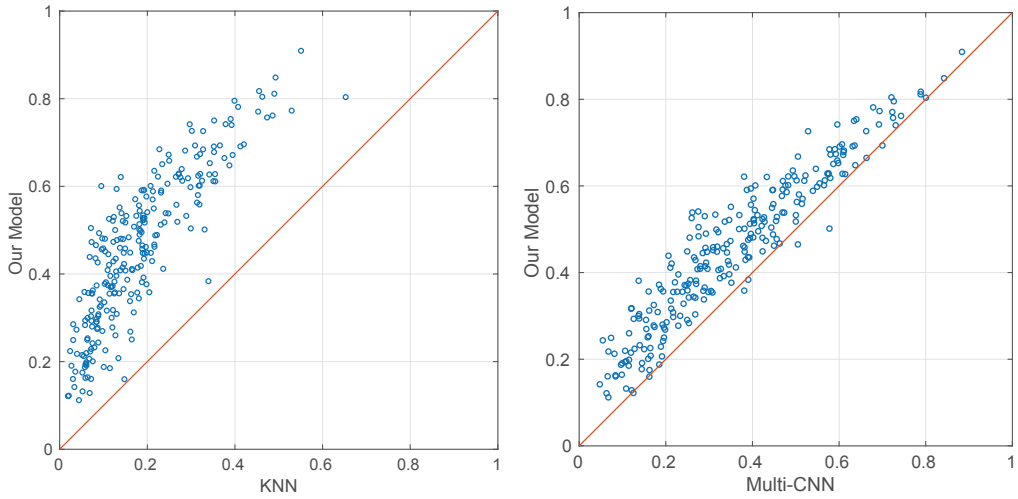


Figure 5.7: The comparisions of the average precision (AP) values between the proposed model and baselines on all labels. The left one is the AP values of our model against KNN baseline, while the right one is our model against Multi-CNN baseline. Better view in color.

ation, which is one of the advantages of our proposed model when we tackle the diverse image annotation problem on the heritage image collection. To include the metadata in these models, we also report the results of the *-models in Table 5.1. As we can see, the *-models outperform their original models in most cases, which validate the effectiveness of the learned image visual representation. Moreover, our model achieves the best performance against *-models again shows the significance of the annotation via collective knowledge. Annotation examples of the proposed model are shown in Figure 5.5.

In Figure 5.6, we show the precision-recall curves of our model against baselines and compared methods on the label 'festival' and 'Art Deco architecture.' As we can see, our model achieves the best performance compared against baselines and state-of-the-art methods. The label 'festival' and 'Art Deco architecture' are both semantical and related to the image visual content. By combining the visual and semantic representation, we show large improvements compared to baselines. We also compare the average precisions (AP) between the proposed model and KNN/Multi-CNN baselines. The results are shown in Figure 5.7, where the x-ray

Figure 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.

Table 5.2: Results of the Ablation Study.

Method	imAP	mAP	O_P	O_R	O_{F1}	C_P	C_R	C_{F1}
Multi-CNN	0.411	0.346	0.541	0.315	0.398	0.515	0.200	0.288
MeL+Fc	0.455	0.403	0.508	0.410	0.454	0.514	0.286	0.367
Our Model	**0.511**	**0.445**	**0.607**	**0.416**	**0.494**	**0.575**	**0.371**	**0.451**

stands for the baseline AP value, and y-ray is the corresponding proposed model's AP value. As we can see, the majority of the values are above $y = x$, which proves the effectiveness of our model on the whole label set.

5.3.5 Ablation Study

We conduct the comprehensive ablation study to further investigate the individual contribution of each component in the proposed model. In this section, we first analyse the effectiveness of the metric learning on the visual representation. Then we investigate the contribution of the semantic summarisation for the annotation.

Visual Representation

To conduct a quantitative analysis, we extract the image visual representation after the metric learning and fed into the fully-connected network to train an annotation model. The network is trained with the sigmoid cross-entropy loss, same as Multi-CNN baseline. We note this model as MeL+Fc and show the results in Table 5.2. As we can see, MeL+Fc shows large improvement compared to baseline Multi-CNN, which proves the metric learning based on the metadata similarity is necessary to obtain the accurate visual representation.

In Figure 5.8 we show some examples of training pairs for the metric learning based on the metadata similarity. As we can see, it is reasonable to generate positive and negative samples for the visual representation learning based on the metadata similarity.

Semantic Summarisation

As we mentioned in Section 5.2.3, after we obtain the visual representation, we retrieve nearest neighbours and utilise them to generate the semantic representation. In Table 5.2, we observe the improvement from the comparison between MeL+Fc and our final model, which proves that summarise semantic information from neighbours can help boost the annotation performance. We show some examples of label suggestions retrieved from image neighbours in Figure 5.9. As we can see, these suggestions can assist the annotation model to infer all labels.

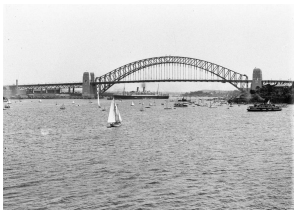
5.4 Summary

Images are the reflections of the real world. Building an automatic annotation model is a crucial step to understand these images as well as efficiently retrieve them. Different from the traditional image annotation problem, which focuses on annotating the common visual objects and attributes, in this work, we focus on the image annotation with diverse visual appearances and labels. We ground the annotation task on the heritage collection given its visual and label diversity,

(a)

Ground-Truth:
crowd; festival; city street; procession; band; anniversary

Suggestions from neighbours:
crowd; procession; band; festival; spectator; city street; association; anniversary;



(b)

Ground-Truth:
official event; harbor; bridge

Suggestions from neighbours:
harbor; bridge; official event; bridge construction; ship

(c)

Ground-Truth:
evening clothes; table setting; dance

Suggestions from neighbours:
evening clothes; table setting; dance; group people; convention;

(d)

Ground-Truth:
theatrical costume; actor; theatrical production

Suggestions from neighbours:
portrait; theatrical costume; actor; theatrical production

Figure 5.9: Some examples of label suggestions retrieved from image neighbours. We summarise these suggestions as the semantic representation.

and propose to tackle this problem via collective knowledge. We uncover image neighbours by measuring the metadata similarity and adopt the metric learning to obtain the effective visual representation. Moreover, we generate semantic representations from image neighbours to assist the annotation model to infer all labels. Advanced experimental results against baselines and state-of-the-art methods show the significance of the proposed model.

Chapter 6

Annotation via Graph Co-Attention Networks

6.1 Introduction

With the rise of the social network, people like to capture vivid moments and share them on the internet. These images are generated and spread at an explosive pace, which yields the urgent need for an efficient annotation method to assist users to understand and retrieve images. Significant progresses have been made on the image annotation by uncovering relationships between image pixel contents and labels, such as the classification (Breiman 2001, Chang & Lin 2011), clustering (Guillaumin et al. 2009, Yu & Grauman 2014), and graph inference (Li et al. 2014, Tan et al. 2015). Most recently, deep neural networks (He et al. 2016, Simonyan & Zisserman 2014) have demonstrated advanced abilities of the image feature learning, which inspire various network-based models (Wang et al. 2016, Wei et al. 2014). These models treat the individual image as an independent object and focus on solving the annotation without the context information. However, due to the diverse contents and complex styles, some images are still difficult to annotate on their own.

Social networks like the Flickr, Instagram and Facebook record the user interactions with images as the vast amount of the metadata, which is presented in

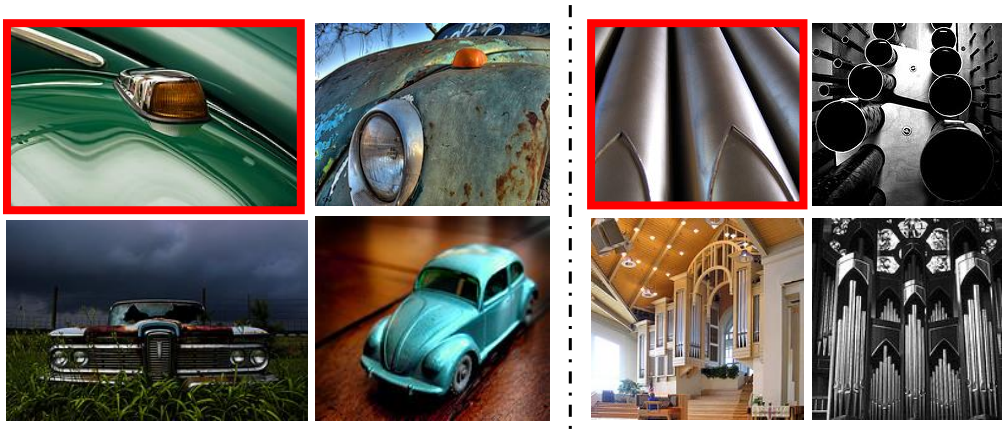


Figure 6.1: For the target images with the red boxes, they are hard to recognise on their own. However, in the context of the neighbours with the similar metadata, such as 'vehicle, vintage, Beetle' and 'church, instrument, art', it is more clear that the target images are a car and a pipe organ. Based on this motivation, we propose a neighbourhood graph as *'neighbourhood watch'* to assist the image annotation.

various forms. The most common metadata includes the collections, *i.e.* image groups created by users; textual descriptions, *i.e.* tags and captions; as well as user profiles, *i.e.* user-names, locations and friends. As a means to communicate with other users, these metadata can be as informative as visual pixels (Johnson et al. 2015, McAuley & Leskovec 2012) to understand images. See Fig. 6.1 as an example, images are hard to be recognised and annotated without seeing its metadata related images. There are several lines of works (Ballan et al. 2014, Srivastava & Salakhutdinov 2012) conducted by utilising the metadata to assist the annotation, where different types of metadata are studied to be embedded into the annotation framework. In (Johnson et al. 2015), authors propose to non-parametrically use the metadata to generate image neighbours and train an annotation model based on the visual features from the target image and its neighbours. However, the features from image neighbours are independently embedded from each other, and only global features are considered.

In this chapter, we address the image annotation problem by routinely checking its neighbours in a graph, which is constructed by the equipped meta infor-

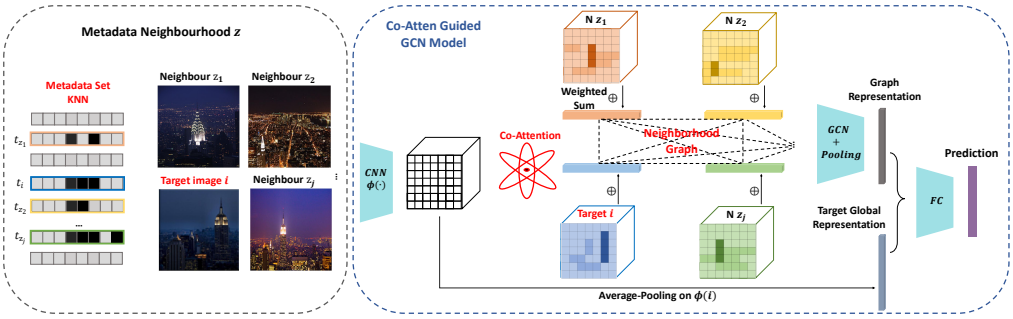


Figure 6.2: The framework of the proposed model. The neighbourhood z of the target image i is decided by measuring their metadata similarities. Then we establish the neighbourhood graph using image representations as nodes and correlations as edges. To accurately harvest visual clues from its neighbours, 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.

mation of the image, which is different from the previous chapter 3 generic image annotation (training image samples are annotated only with expert labels), chapter 4 (directly learning from noisy user tags during training), while chapter 5 introduces a solution for the diverse image annotation problem by utlising the metadata, the proposed model in this chapter focuses on the graph based approach and target at contemporary images, and since the metadata is only used to ground image neighbourhood, therefore, unlike the chapter 5, the metadata dictionary in this chapter does not need to be consistent between training and test sets.

The whole framework is shown in Fig. 6.2. Since the metadata explicitly or implicitly indicates the connections between images, for example, semantically similar textual descriptions like tags and captions usually associated with images that have similar visual appearances (Guillaumin et al. 2009), friends who share same interests have the high probabilities of following images with similar topics (Stone et al. 2008), and landmark photos are always taken at the fixed locations (Li, Crandall & Huttenlocher 2009). We locate the image neighbours by measuring the similarities among their metadata. Then we establish a graph network to model the correlations between the target image and its neighbours. The whole

neighbourhood is represented as a graph, where each node is the corresponding image feature. The graph edges indicate the correlations between node pairs.

Considering the diverse visual appearances of neighbours, different attention should be paid according to its content. Therefore, we introduce a co-attention mechanism to obtain the node representation. That is, we obtain the co-attention maps by successively switching the attention between the target image and its neighbours. Given the graph structure, we can perform reasoning on the graph and infer the representation by applying the graph convolutional operations. The graph representation is finally concatenated with the target global feature to predict the label confidence, which is intuitive since we want to capture the connections among image neighbours to assist the annotation of the target image. We name our proposed model as **M**et**a**data **n**eighbourhood **g**raph c**o**-attention **net**work (MangoNet).

In summary, the main contributions of our method are as follows:

1) We propose to annotate images by exploring their neighbourhoods, which are allocated by the metadata similarity. A neighbourhood graph network is established to model the correlations between each image and its neighbours. The learned graph representation is used to assist the annotation of the target image.

2) To accurately capture the relevant regions in each image neighbour that are beneficial for understanding the target image, we introduce a graph co-attention mechanism to obtain the node representations in the graph.

3) We evaluate our proposed method on three benchmark datasets. Our model achieves state-of-the-art performances on all of them.

6.2 The MangoNet

The key characteristic of our proposed Metadata Neighbourhood Graph Co-Attention Network (MangoNet) is to represent the image neighbourhood as a graph to assist the image annotation. The neighbourhood graph is established via the metadata. A co-attention mechanism is introduced to guide the visual attention between the target image and its neighbours to obtain node representations. The whole frame-

work is shown in Fig. 6.2.

In the following sections, we first describe how to locate image neighbours by measuring their metadata similarities; then we introduce the architecture of the neighbourhood graph and the node representation updating process by unifying the instance-level and co-attention mechanisms. The training and implementation details are given at last.

6.2.1 Neighbourhood Graph Co-Attention

Graph Construction

The motivation behind the neighbourhood graph is that a graph structure, where its edges indicate the correlations between image nodes, can competently represent the image neighbourhood. By reasoning on the graph, we aggregate node features as the graph representation to assist the target image annotation. Metadata, as a means of bridges between images, it connects images with each other. We first locate image neighbours by measuring their metadata similarities.

Formally, let I be a set of images, V be the set of manual labels $|V| = C$, and $D = \{(i, v)|i \in I, v \subseteq V\}$ be the image dataset, where each image is associated with a subset of labels. Let T be the vocabulary of the metadata carried by I. Given the vocabulary T, each image i is associated with a subset of $t_i \subseteq T$ metadata. We use the Jaccard metric to measure the similarity between the metadata pair, that is, given two images i,$j \in I$ with $t_i, t_j \subseteq T$, $\varphi(i, j) = |t_i \cap t_j|/|t_i \cup t_j|$, where $\varphi(i, j) \in [0, 1]$. Based on the metadata similarity, the nearest neighbour approach can be applied to locate the neighbours.

Since most of the metadata are generated based on the user behaviours (Johnson et al. 2015), metadata vocabularies can be very large, and the metadata associated with each image can be imprecise at the certain level (Li et al. 2016). To accurately and efficiently locate the neighbours, we conduct the hierarchy search strategy or the semantic search strategy in the light of circumstances. Specifically, for each image in the large-scale dataset with large metadata vocabulary size like the NUS-WIDE (Chua et al. 2009), we perform the neighbour search based on the

user tags within a sub-image set, which is consist of the images from the collections with similar topics. For the dataset with relatively clean semantic metadata such as the MS-COCO (Lin et al. 2014) with human-labelled captions, we extract metadata representations of each image as the weighted sum of the semantic representations of visual attributes. We set the search parameters as m neighbours for each image and obtain the candidate neighbourhood $z = \{z_1, ..., z_m\}$ for each image i. The image i and its neighbours z are fully connected with each other to form a neighbourhood graph.

Graph Co-Attention Mechanism

Node Attention Since we intend to obtain visual clues from image neighbours to assist the annotation of the target image, for each neighbour, different attention should be paid according to its content. We adopt a co-attention mechanism to generate attention maps for the target image and its neighbours, *i.e.* 'mind the neighbours'. The co-attention mechanism can be viewed as to learn image visual correlations to contribute to the node representation.

Considering we only have the image-level supervision information *i.e.* annotated ground-truth, we propose to generate the instance-level attention maps regarding individual image first, then combine them as the co-attention maps we desired, which is consistent with our motivation that visual clues from the neighbour are harvested from common semantic classes with the target image, and the target image is revisited with the clues from its neighbourhood. The framework of the instance-level attention module and the proposed co-attention module are shown in Fig. 6.3 and Fig. 6.4 respectively.

More specifically, multiple labels associated with an image are always semantically related to different image regions. By referring to (Zhu et al. 2017), we adopt an instance-level attention module to capture this multi-label property for each image. That is, learning attention weights for each label (instance) respectively from the image ground-truth, noted as the InsAtten module. We employ the ResNet-101 (He et al. 2016) as the backbone CNN model to extract the outputs of the convolutional layer as visual features $\phi \in N \times d$, where $N = H \times W$. Given

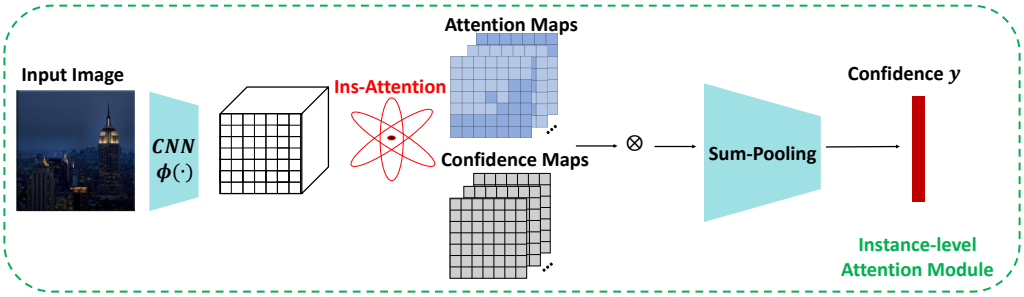


Figure 6.3: The instance-level attention module we used to capture the regional semantic correspondences between image content and associated labels.

the image i with feature $\phi(i)$ and associated label subset v, the attention weights for each label are generated as follows:

$$x = q(\phi(i), \eta), \quad x \in \mathcal{R}^{H \times W \times C} \tag{6.1}$$

$$\beta = \text{softmax}(x), \tag{6.2}$$

where the attention is estimated by $q(\cdot)$ with the parameters η as three convolutional layers with 512 kernels of 1×1, 512 kernels of 3×3 and C kernels of 1×1. And β is the normalised attention weights with $\beta \in \mathcal{R}^{H \times W \times C}$, of which the third dimension stands for the size of the whole label set V. Each attention map $\beta^k \in \mathcal{R}^{H \times W \times 1}$, where $k \in [1, C]$, is used to weighted sum the image feature for k_{th} label as:

$$\tilde{i^k} = \sum_l \beta^k \odot \phi(i), \tag{6.3}$$

where $l \in [1, N]$ indexes the spatial position. The weighted image feature $\tilde{i^k}$ represents the regions related to the k_{th} label. Then the weighted feature can be fed to the FC layer to generate the confidence for each label. To efficiently learn the attention weights, by referring to (Zhu et al. 2017), we reformulate the FC layer as applying the label-specific linear classifier at every spatial location of the image feature $\phi(i)$ and then aggregating the label confidences based on β. That is, we forward $\phi(i)$ to a convolutional layer with C kernels of 1×1 to generate confidence maps $E \in \mathcal{R}^{H \times W \times C}$, then E and β are element-wise multiplied and

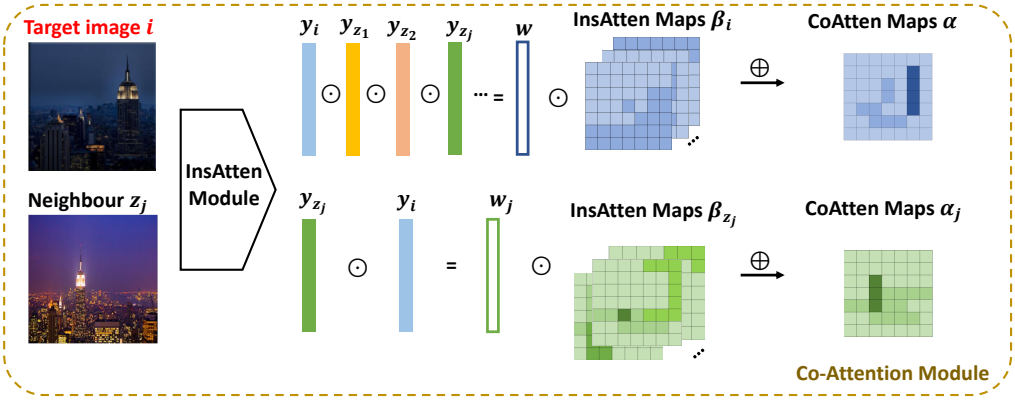


Figure 6.4: The co-attention attention module. The co-attention maps are the semantic overlap of instance-level attention maps between the target image i and its neighbourhood z.

sum-pooled to obtain the confidence vector $y \in \mathcal{R}^C$, which can be trained with the image ground-truth v.

Since the co-attention mechnisam is introduced to capture the correlations between the target image and its neighbourhood, we formulate the operation of the co-attention mechanism as the weighted sum of the instance-level attention maps. That is, the target image i and its neighbours $z(|z| = m)$ firstly pass through the backbone and InsAtten module to obtain the attention maps β_i and β_z, and confidence vectors y_i and y_z, where $y \in R^C$,$\beta \in R^{H\times W\times C}$. Then the co-attention for the neighbour z_j can be computed as:

$$w_j = y_i \odot y_{z_j} \tag{6.4}$$

$$\alpha_j = softmax(\sum_{k=1}^{C} w_j^k \odot \beta_{z_j}^k) \tag{6.5}$$

$$\hat{z}_j = \sum_l \alpha_j \odot \phi(z_j) \tag{6.6}$$

where $\alpha_j \in R^{H\times W\times 1}$, and $\odot$ stands for the element-wise multiply with broadcasting, w_j represents the semantic overlaps between the target and its neighbour z_j. The weighted sum $\hat{z}_j$ is the weighted feature for j_{th} neighbour. We omit $\phi(\cdot)$ in the weighted feature for the expression simplicity. Similarly, for the target image

i, the co-attention is computed as:

$$w = y_i \odot \{y_{z_1} ... \odot y_{z_j} ... \odot y_{z_m}\} \tag{6.7}$$

$$\alpha = softmax(\sum_{k=1}^{C} w^k \odot \beta_i^k) \tag{6.8}$$

$$\hat{i} = \sum_{l} \alpha \odot \phi(i) \tag{6.9}$$

where $\hat{i}$ is the weighted target image feature. The weighted image (and its neighbours) features will be used as node representations in the proposed neighbourhood graph.

Graph Reasoning with Attented Features After we obtain the attended representation for image i and its m neighbours as $\hat{z} = \{\hat{z_0}, \hat{z_1}, ..., \hat{z_m}\}$ [1]. The correlation between every two images can be represented as:

$$s(\hat{z_k}, \hat{z_l}) = \psi(\hat{z_k})^T \psi(\hat{z_l}), \quad \forall k, l \in [0, m] \tag{6.10}$$

where the $\psi(\cdot)$ is modelled as FC layer with the hidden state size 512. We apply the softmax function on each row of the correlation matrix, which normalises the sum of all the edge values connected to each node to be one. The normalised matrix S is taken as the adjacency matrix for the proposed graph.

We adopt the graph convolutional network (GCN) (Kipf & Welling 2016) to reason on the graph. Based on the definition of neighbourhood relations, the GCN can compute the response of each node and pass messages inside the graph. The outputs of the GCN are updated node features, which will be aggregated for further use. More specifically, one layer of the graph convolution is represented as:

$$Z' = S\hat{Z}W \tag{6.11}$$

where S is the introduced adjacency matrix with $(m+1) \times (m+1)$ dimensions, $\hat{Z}$ is the features of image nodes with $(m+1) \times d$ dimensions, W is the learnable weight matrix with $d \times d$ dimensions. The output Z' of one graph convolutional

[1] The weighted feature $\hat{i}$ of the target image i is noted as $\hat{z_0}$ for the unified graph representation.

layer is $(m+1) \times d$ dimensions, which is followed by a ReLU activation. For the fast convergence, we use a residual unit to update the node features *i.e.* $Z' = Z' + \hat{Z}$. The updated node features are fed to an average pooling layer to obtain a $1 \times d$ representation. Moreover, we also perform an average pooling on the target image feature to obtain a $1 \times d$ global representation. Two representations are concatenated together and sent to the Fully-Connected (FC) layer to predict the final confidence vector $y_{neb} \in \mathcal{R}^C$, where C is the number of classes. See Fig. 6.2 for the illustration.

6.2.2 Implementation Details

We employ the pre-trained ResNet-101 (He et al. 2016) as the initial backbone CNN model $\phi(\cdot)$ to extract the convolutional features. We adopt the stage-wise training strategy: first we finetune the pretrained backbone model on each dataset, then we train the InsAtten module by referring to (Zhu et al. 2017), and finally we train the CoAtten-GCN module. In practice, the size of the instance-level attention maps are initially trained with the size of 14×14 and then max-pooled as 7×7. The cross-entropy loss function and the stochastic gradient descent (Bottou 2010) are used for the optimisation. We train the models with the batch size 64 and the learning rate of 0.001 from the start and vary the sizes of the image neighbourhood as $m = 3/7/15$. By referring to (Johnson et al. 2015), the nearest neighbour search for the training and test metadata are performed separately.

6.3 Experiments

We present the experimental results in this section and analyse the effectiveness of the proposed model. Our model is evaluated on three benchmark datasets: NUS-WIDE (Chua et al. 2009), Mirflickr (Huiskes & Lew 2008) and MS-COCO (Lin et al. 2014). We compare our model with several baselines and state-of-the-art methods. An ablation study is then performed to evaluate the contribution of each component of our model. We finally visualise some of the attention map examples to show their effectiveness.

6.3.1 Datasets

Both the NUS-WIDE (Chua et al. 2009) and Mirflickr (Huiskes & Lew 2008) contain a large number of images collected from the Flickr website, a commonly used image-sharing social network. Each image in the dataset is manually annotated for the presence of the pre-defined label set. By referring to (Johnson et al. 2015, McAuley & Leskovec 2012), we query the metadata of images via the Flickr API. To ensure the dataset scale, we tokenise and lemmatise the most common metadata *i.e.* user tags and image collection descriptions for our experiments. We remove the duplicates in the processed metadata sets, and the image records without valid URLs or ground-truth labels. It is worth noting that the metadata information is only used for locating the image neighbourhood, it does not affect the dataset scale or involve in the model training. We select the optimal sizes of the metadata sets based on the grid search. After the preprocess, we use $201{,}302$ images for the NUS-WIDE with 81 labels, $3{,}010$ user tags and 704 collection topics; $12{,}682$ images for the Mirflickr with 14 labels, 450 user tags. We then select $150{,}000$ and $51{,}302$ images for training and test respectively on the NUS-WIDE; $5{,}200$ and $7{,}482$ images for training and test respectively on the Mirflickr. For the MS-COCO (Lin et al. 2014), the descriptions of each image take the form of a set of captions. We tokenise the captions and extract most common 256 visual attributes as the metadata for this dataset. For the semantic representations of the visual attributes, we query the pre-trained word2vec. We use the official train/val split, which is $82{,}783$ for training and $40{,}504$ for test.

6.3.2 Evaluation Metrics

We employ several metrics to evaluate the performance of the proposed models and compared methods. By referring to previous works (Johnson et al. 2015, Li et al. 2016, Zhu et al. 2017), we compute the average precision (AP), it ranks the retrieved results based on the relevance regarding the query. For each label, relevant images should be ranked higher than the irrelevant ones, noted as the mAP_C. To take into consideration of the label imbalance problem on the dataset

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
Meta+Logistic (Johnson et al. 2015)	0.527	0.667	0.679	0.393	0.475	0.768	0.450	0.567
KNN (Makadia et al. 2010)	0.462	0.669	-	-	-	-	-	-
Multi-CNN (He et al. 2016)	0.580	0.789	0.693	0.408	0.490	0.810	0.565	0.666
CNN_Voting (Johnson et al. 2015)	0.599	0.799	0.674	0.423	0.497	0.795	0.597	0.682
TagProp (Guillaumin et al. 2009)	0.541	0.742	0.674	0.408	0.496	0.784	0.572	0.661
Link-CRF (McAuley & Leskovec 2012)	0.542	0.770	0.698	0.347	0.437	0.805	0.548	0.652
SRN (Zhu et al. 2017)	0.600	0.806	0.704	0.415	0.496	0.811	0.587	0.682
NCNN (Johnson et al. 2015)	0.598	0.803	0.725	0.390	0.478	0.800	0.598	0.685
MangoNet-m3	0.617	0.806	0.715	0.419	**0.507**	0.802	0.599	0.686
MangoNet-m7	0.626	**0.808**	0.720	0.415	0.505	0.804	0.599	**0.687**
MangoNet-m15	**0.628**	**0.808**	0.739	0.410	0.501	0.806	0.599	**0.687**

Table 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.

splits, we also compute the overall mAP by treating each label assignment as an independent label, noted as the mAP_O. Moreover, to conduct the quantitative evaluation, we predict up to three ranked labels above the confidence threshold 0.5 for each image to compare against the ground-truth. Mean scores of per label and overall precision, recall and F1 score noted as $C_P, C_R, C_{F_1}/O_P, O_R, O_{F_1}$ are reported.

6.3.3 Overall Performance

We compare the proposed model with several popular and state-of-the-art annotation models, which involve utilising the image metadata or the attention mechanism. Since the different metadata and dataset splits are studied in these models, for fair comparisons, we re-implement some of them (Guillaumin et al. 2009, Johnson et al. 2015, McAuley & Leskovec 2012, Makadia et al. 2010, Zhu et al. 2017) by using the metadata vocabularies and dataset splits we processed, and the hand-crafted features from the original models are replaced with the average pooled 2048-d convolutional features from the backbone. For fair comparisons, the compared models share the same finetuned backbone on each dataset. We give

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
Meta+Logistic (Johnson et al. 2015)	0.571	0.719	0.771	0.334	0.429	0.767	0.494	0.601
KNN (Makadia et al. 2010)	0.745	0.839	-	-	-	-	-	-
Multi-CNN (He et al. 2016)	0.816	0.915	0.889	0.638	0.696	0.857	0.800	0.827
CNN_Voting (Johnson et al. 2015)	0.825	0.916	0.902	0.630	0.685	0.860	0.798	0.828
TagProp (Guillaumin et al. 2009)	0.818	0.856	0.776	0.625	0.657	0.864	0.594	0.704
Link-CRF (McAuley & Leskovec 2012)	0.800	0.902	0.883	0.601	0.671	0.853	0.766	0.807
SRN (Zhu et al. 2017)	0.831	**0.925**	0.839	0.711	0.760	0.853	0.827	0.840
NCNN (Johnson et al. 2015)	0.840	0.918	0.840	0.724	0.765	0.849	0.826	0.837
MangoNet-m3	0.849	0.924	0.870	0.713	0.769	0.865	0.822	0.843
MangoNet-m7	**0.852**	0.924	0.866	0.718	**0.772**	0.865	0.826	**0.845**
MangoNet-m15	0.851	**0.925**	0.881	0.705	0.761	0.867	0.823	0.844

Table 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.

the details of the main compared models as follows:

- *Meta+Logistic* (Johnson et al. 2015): This model is to investigate the annotation capability by only using the metadata. Each image is represented by the binary vector with the metadata vocabulary size $|T|$-dimension, and trained with the logistic loss to generate the label confidence.

- *CNN_Voting* (Johnson et al. 2015): This model is to investigate the contribution of the image metadata neighbourhood for the voting-based methods. Different from the KNN, which allocates the visually similar neighbourhood in the training set, this model uses the metadata neighbourhood directly generated from the test set (same as the proposed MangoNet). Then the label confidence of the test image is set to be a weighted sum of its Multi-CNN prediction and the mean of the Multi-CNN predictions of its neighbours.

- *SRN* (Zhu et al. 2017): This is a state-of-the-art attention-based model, which utilises the instance-level attention as the spatial and semantic regularisation to strengthen the CNN framework. We report the results on the MS-COCO from the original model since we also use the same official dataset splits. For the NUS-WIDE and Mirflickr, the dataset splits are different after the preprocess, therefore, we train and test this model using the same splits as the proposed model

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
Meta+Logistic (Johnson et al. 2015)	0.703	0.779	0.851	0.556	0.643	0.882	0.575	0.696
KNN (Makadia et al. 2010)	0.699	0.766	-	-	-	-	-	-
Multi-CNN (He et al. 2016)	0.738	0.812	0.827	0.563	0.646	0.848	0.601	0.703
CNN_Voting (Johnson et al. 2015)	0.750	0.818	0.816	0.558	0.635	0.839	0.582	0.687
TagProp (Guillaumin et al. 2009)	0.720	0.814	0.815	0.570	0.641	0.832	0.607	0.703
Link-CRF (McAuley & Leskovec 2012)	0.718	0.787	0.823	0.548	0.642	0.831	0.594	0.693
SRN (Zhu et al. 2017)	0.771	0.839	0.852	0.588	0.674	0.874	0.625	0.729
NCNN (Johnson et al. 2015)	0.760	0.833	0.838	0.579	0.669	0.871	0.608	0.716
MangoNet-m3	0.775	0.843	0.871	0.579	0.676	0.895	0.619	0.732
MangoNet-m7	0.778	0.845	0.881	0.577	0.676	0.902	0.618	0.733
MangoNet-m15	**0.779**	**0.846**	0.876	0.584	**0.680**	0.898	0.622	**0.735**

Table 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.

for these two datasets.

- *NCNN* (Johnson et al. 2015): This model employs the metadata neighbourhood to assist the target annotation. The image neighbours are embedded with the hidden state size 512 from image global features and then max-pooled. We use the same processed metadata neighbourhood in this model to investigate the importance of the co-attention GCN module.

Tab. 6.1, 6.2 and 6.3 show that our proposed models, noted as the ***MangoNet***, achieve the best performances on the overall evaluation metrics mAP_C, mAP_O and O_{F_1} on all three datasets. The ***Meta+Logistic*** (Johnson et al. 2015) represents each image with a binary vector indicating the presence of the metadata and trains the classifiers with the logistic loss regarding each label. Most vision-based methods outperform this model, which proves that learning from the image visual content is crucial for the annotation. The ***Multi-CNN*** (He et al. 2016) is a standard multi-label annotation model trained with the logistic loss, which serves as a baseline. The ***CNN_Voting*** (Johnson et al. 2015) utilises the same metadata neighbourhood we processed and averages the label confidences from the ***Multi-CNN*** on the test set neighbours. It outperforms the classic ***KNN*** (Makadia et al. 2010), which indicates that the metadata can be useful for eliminating the visual ambigu-

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
NGCN	0.612	**0.807**	0.727	0.381	0.470	0.808	0.595	0.685
MangoNet w/o GF	0.499	0.691	0.656	0.328	0.412	0.739	0.492	0.591
InsAtten	0.579	0.801	0.700	0.383	0.468	0.822	0.562	0.668
MangoNet	**0.617**	0.806	0.715	0.419	**0.507**	0.802	0.599	**0.686**

Table 6.4: Ablation studies of our model on the NUS-WIDE.

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
NGCN	0.843	0.923	0.885	0.667	0.731	0.866	0.813	0.838
MangoNet w/o GF	0.768	0.876	0.807	0.613	0.679	0.821	0.737	0.777
InsAtten	0.820	0.916	0.884	0.665	0.723	0.857	0.806	0.831
MangoNet	**0.849**	**0.924**	0.870	0.713	**0.769**	0.865	0.822	**0.843**

Table 6.5: Ablation studies of our model on the Mirflickr.

ous and locating the neighbours that contribute to the annotation. The superior performance against these models shows the significance of the proposed graph structure for exploring the neighbourhood feature.

Instead of treating image neighbours equally, the ***TagProp*** (Guillaumin et al. 2009) is a trained nearest neighbour method, where each image neighbour is re-weighted by the discriminative metric learning. Compared to this model, our proposed ***MangoNet*** represents relationships between the target image and its neighbours as a graph. By reasoning on the graph, we can aggregate the node features as the neighbourhood representation to assist the annotation, which outperforms this model by a large margin. The state-of-the-art method ***SRN*** (Zhu et al. 2017) employs the instance-level attention mechanism as the spatial and semantic regularisation to boost the annotation on the individual image, while in our model, we not only consider the target regional attention but also value the visual clues harvested by the proposed graph model from the neighbourhood. The graphical solution ***Link-CRF*** (McAuley & Leskovec 2012) defines the image relations via metadata, and models them as the CRF. We have the similar motivation to use the neighbourhood defined by the metadata. However, we not only look into

Method	mAP_C	mAP_O	C_P	C_R	C_{F_1}	O_P	O_R	O_{F_1}
NGCN	0.765	0.834	0.842	0.568	0.663	0.880	0.610	0.721
MangoNet w/o GF	0.692	0.774	0.811	0.504	0.599	0.854	0.547	0.667
InsAtten	0.741	0.816	0.831	0.565	0.649	0.851	0.603	0.706
MangoNet	**0.775**	**0.843**	0.871	0.579	**0.676**	0.895	0.619	**0.732**

Table 6.6: Ablation studies of our model on the MS-COCO.

the neighbourhood but also propose to capture the visual clues from each neighbour by a co-attention mechanism, which achieves better results. The most related method ***NCNN*** (Johnson et al. 2015) also proposes to utilise the neighbour features to assist the annotation. However, in this model, these neighbours are embedded separately from the target image, and only the holistic features are considered. Different from the ***NCNN***, our ***MangoNet*** establishes a graph structure to represent the neighbourhood and employ a co-attention mechanism to guide the message passing within the graph.

Moreover, we also investigate the influences of the different sizes of the neighbourhood, we report the results of the $m = 3/7/15$ in the tables. As we can see, in general, with the larger neighbourhood, the visual ambiguous can be further eliminated, and the proposed model achieves the better results. To indicate the significance of the proposed components on the whole label set, we show the comparisons of AP values against the NCNN (no co-attention neighbourhood graph is adapted in this model) in Fig. 6.5, where the x-ray stands for the NCNN value on each label, and y-ray is the corresponding value of our MangoNet. As we can see, the majority of the values are above the $y = x$, which proves the effectiveness of the proposed graph model on the whole tag set.

6.3.4 Ablation Study

We conduct the ablation analysis on three datasets to further investigate the individual contributions of proposed components in a tiered manner. We compare the following ablation models:

- *NGCN*: This is a plain graph model without the proposed co-attention mech-

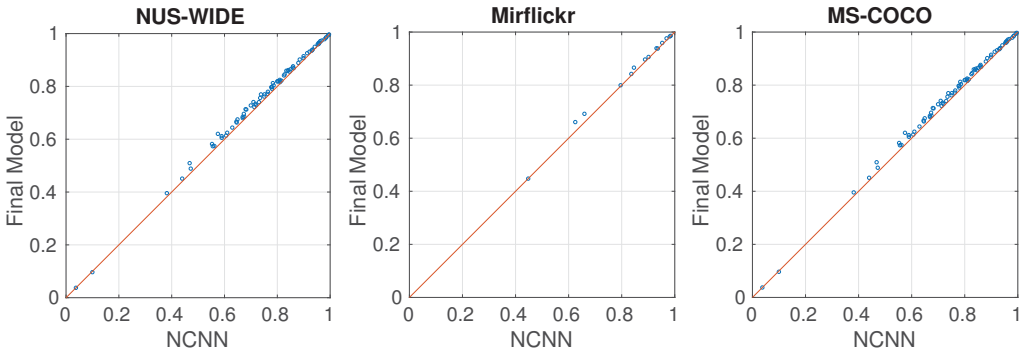


Figure 6.5: The AP comparisons of the proposed MangoNet against the NCNN model on the NUS-WIDE, Mirflickr and MS-COCO.

anism. To implement this model, we replace the node representations in our proposed GCN model as the holistic features, *i.e.* average-pooled convolutional features.

- MangoNet w/o GF: This is the co-attention guided GCN part of the proposed model without the target global feature concatenation.

- InsAtten: In this model, we train the instance-level attention module independently.

- MangoNet is the proposed full model with the neighbourhood size $m = 3$, same as all other ablation models.

The results are shown in Tab. 6.4, 6.5 and 6.6. As we can see, by introducing the co-attention mechanism into the graph model, our ***MangoNet*** achieves better results against the ***NGCN***, which proves the effectiveness of the proposed attention mechanism in capturing the correlations within the neighbourhood. Without using the global feature concatenation in the GCN module, as we can see, ***MangoNet w/o GF*** has lower performance, since the image neighbours do not guarantee to contain all the labels associated with the target image. The instance-level only attention model ***InsAtten*** performs lower than others, but based on the co-attention maps generated from the InsAtten module, our full model ***MangoNet*** achieves the best performance.

6.3.5 Visualisation of Attention

To verify the proposed attention mechanisms, we visualise the attention maps from the co-attention module. The brighter colour (yellow) indicates the higher attention weights. We show the examples of the co-attention maps of the given images in Fig. 6.6. The first column of every two rows is the target image and its co-attention map, while the rest columns are the neighbours and their corresponding attention maps. As we can see, the co-attention maps capture the correlations between the target image and its neighbours in both simple and complex scenes. For example, small subjects like 'surfboard' and 'ball' are well-captured in example 2 and 4, while in the complex scenes such as example 3, the 'people' and 'car' are also well-captured in the neighbours. The co-attention mechanism is employed to find the most relevant visual features to eliminate the recognition uncertainties, then the neighbourhood graph receives these features and communicates within the neighbourhood, which improves the ability of the annotation model to recognise the target image.

6.4 Summary

Images are connected to each other via the abundant metadata. Fully making use of these connections can assist the image annotation. In this chapter, we explore the image neighbours by measuring their metadata similarities and propose a graph network to model the correlations between the target image and its neighbours. A co-attention mechanism is introduced to leverage the visual attention within the neighbourhood. Experimental results on three benchmark datasets show that the proposed model achieves the best performances against compared methods. Since the textual metadata is mainly used in our experiments, we will explore other metadata types in the future work. Moreover, how to ground the image neighbourhood more efficiently and dynamicly is also worth further investigations.

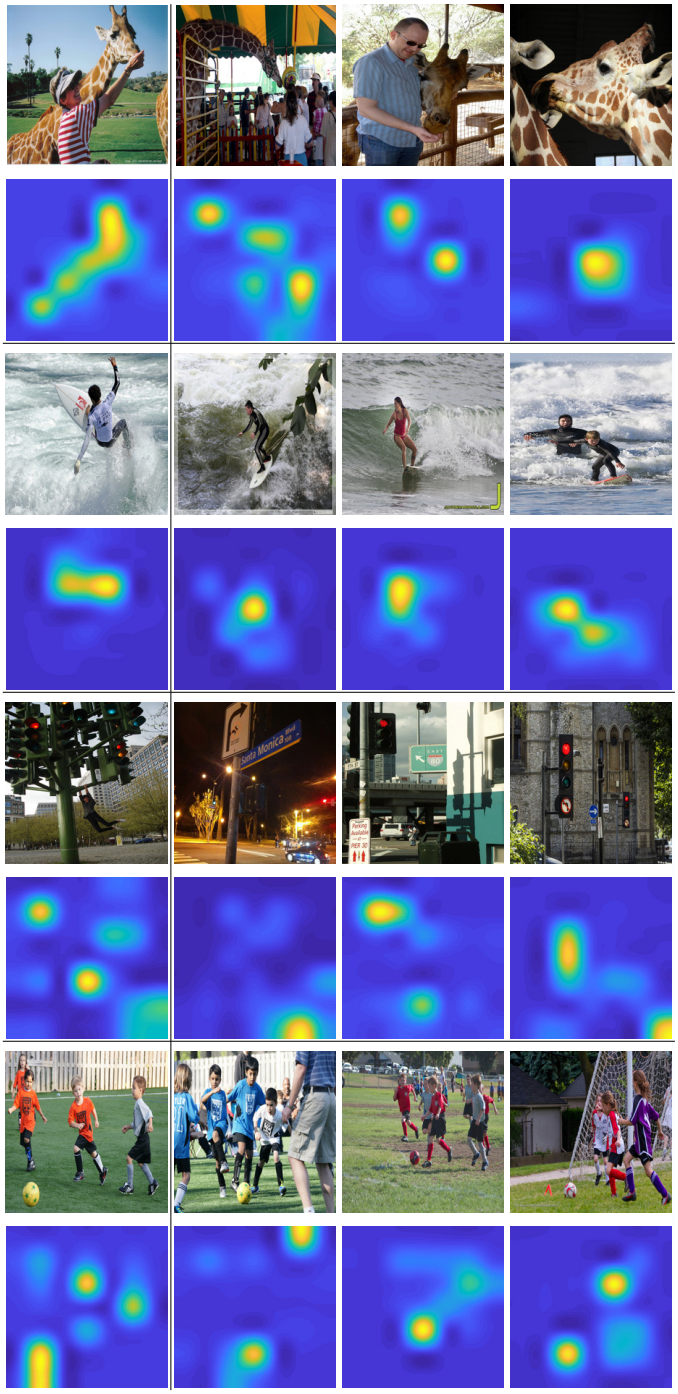


Figure 6.6: The visualisations of the co-attention maps of the targets and their neighbours. The first column is the target and its attention map, the rest columns are the neighbours, where the neighbourhood size $m = 3$.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis presents several methods to address two types of image annotation problems, namely the Generic Image Annotation (GIA), which receives the images with manually annotated labels as training samples; and the Annotation with Metadata Information (IAMI), which in addition to the images and labels, also considering the metadata carried by the social networks or specific domain sets.

Chapter 3 proposes a regional latent semantic dependency model for the generic image annotation, which explores the label correlations at the regional level. Since predicting small visual concepts are challenging due to the limited discrimination of the global image feature, by utilising the regional information, the proposed model achieves the state-of-the-art overall performance and more sensitive to the 'small' objects. There are two main characteristics in the proposed model; one is the fully-convolutional localisation network for capturing the regions potentially have the multiple dependent labels, another is the shared LSTM for analysing the latent semantic dependencies within the regions. The chapter 3 is supported by the journal publication at IEEE T-MM (Zhang, Wu, Shen, Zhang & Lu 2018).

Chapter 4 tackles the annotation with social network metadata problem. More specifically, the social image annotation and tag refinement. Different from the conventional works, which train the annotation model from the expert labels, the

proposed model treat relatively noisy user tags from the social network as weakly-supervised information and directly learns from them. A regional neural network is employed as the backbone model, while two tasks are tackled simultaneously via three proposed constraints, namely the visual consistency among image neighbours, the semantic dependencies between tag pairs, and the user-error sparsity. The chapter 4 is supported by the conference publication at AAAI18 (Zhang, Wu, Zhang, Shen & Lu 2018).

Chapter 5 targets at the diverse image annotation problem, which also utilises the metadata to assist the annotation. Different from the Chapter 4, the metadata in this chapter is collected from the specific image domain. We ground the task on the heritage image collection, given its visual and semantic diversity. The proposed model focuses on the representation learning via the collective knowledge, including the visual representation, which is conducted among the image neighbours defined by the metadata similarity, and the semantic representation, which is summarised from the neighbours based on the label relevance analysis. Two representations are concatenated together as the final feature to infer all labels. This chapter is supported by the conference publications at DICTA16 (Zhang et al. 2016) and ACPR17 (Zhang, Zhang, Wu, Wu, Xu, Lu, Phua, Curr & Tang 2017), and the journal publication at Pattern Recognition (Zhang, Wu, Zhang, Shen, Lu & Wu 2019).

Chapter 6 proposes a graph-based solution to the annotation with context information. Multi-type of metadata are used to construct an image neighbourhood via their metadata similarities, where each node is an image neighbour, and graph edges indicate the visual similarities. To accurately capture the correlations between the target image and its neighbours, a co-attention mechanism is introduced to embed the node features. The graph convolutional operations are carried out to infer the graph representation. Moreover, an instance-level attention model is employed to capture the relationships between the image content and corresponding labels, which further regularises the annotation results. This chapter is supported by the conference publication at CVPR19 (Zhang, Wu, Zhang, Shen & Lu 2019).

7.2 Future Work

Most recently, the adversarial learning (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville & Bengio 2014) has been introduced into the vision-language problems, such as the image captioning (Melnyk, Sercu, Dognin, Ross & Mroueh 2018, Shetty, Rohrbach, Hendricks, Fritz & Schiele 2017) and visual dialogue generation (Wu, Wang, Shen, Reid & van den Hengel 2017). Such techniques have not been extensively studied in annotation problems. How to connect the image visual content and multiple semantic labels via the generative and discriminative models is worth exploring. Moreover, some studies have been conducted on analysing the human annotation preferences (Wu, Chen, Sun, Liu, Ghanem & Lyu 2018), which aim at predicting more diverse and distinct labels like a human subject. For future works, some insights are expected to address such problems.

Bibliography

Alexe, B., Deselaers, T. & Ferrari, V. (2012), 'Measuring the objectness of image windows', *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(11), 2189–2202.

Altman, N. S. (1992), 'An introduction to kernel and nearest-neighbor nonparametric regression', *The American Statistician* **46**(3), 175–185.

Ballan, L., Uricchio, T., Seidenari, L. & Del Bimbo, A. (2014), A cross-media model for automatic image annotation, *in* 'Proc. ACM Int. Conf. Multimedia Retrieval.', ACM, p. 73.

Batko, M., Botorek, J., Budikova, P. & Zezula, P. (2013), Content-based annotation and classification framework: a general multi-purpose approach, *in* 'Proceedings of the 17th International Database Engineering & Applications Symposium', ACM, pp. 58–67.

Bottou, L. (2010), Large-scale machine learning with stochastic gradient descent, *in* 'Proceedings of COMPSTAT'2010', Springer, pp. 177–186.

Bradley, J. K. & Guestrin, C. (2010), Learning tree conditional random fields, *in* 'Proc. Int. Conf. Mach. Learn.', pp. 127–134.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Chang, C.-C. & Lin, C.-J. (2011), 'Libsvm: a library for support vector machines', *ACM Trans. Intell. Syst. Technol.* **2**(3), 27.

Chang, E., Goh, K., Sychay, G. & Wu, G. (2003), 'Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines', *IEEE Transactions on Circuits and Systems for Video Technology* **13**(1), 26–38.

Chen, L., Xu, D., Tsang, I. W. & Luo, J. (2012), 'Tag-based image retrieval improved by augmented features and group-based refinement', *IEEE Trans. Multimedia* **14**(4), 1057–1067.

Cheng, M.-M., Zhang, Z., Lin, W.-Y. & Torr, P. (2014), Bing: Binarized normed gradients for objectness estimation at 300fps, *in* 'Proc. IEEE Conf. Comp. Vis. Patt. Recogn.', pp. 3286–3293.

Chopra, S., Hadsell, R. & LeCun, Y. (2005), Learning a similarity metric discriminatively, with application to face verification, *in* 'Proc. IEEE Conf. Comp. Vis. Patt. Recogn.', pp. 539–546.

Chua, T., Tang, J., Hong, R., Li, H., Luo, Z. & Zheng, Y. (2009), NUS-WIDE: a real-world web image database from national university of singapore, *in* 'Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009'.

Cilibrasi, R. & Vitányi, P. M. B. (2007), 'The google similarity distance', *IEEE Trans. Knowl. Data Eng.* **19**(3), 370–383.

Dembczynski, K., Waegeman, W. & Hüllermeier, E. (2012), An analysis of chaining in multi-label classification., *in* 'ECAI', pp. 294–299.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J. & Zisserman, A. (2010), 'The pascal visual object classes (voc) challenge', *Int. J. Comput. Vision* **88**(2), 303–338.

Gao, Y., Fan, J., Xue, X. & Jain, R. (2006), Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers, *in* 'Proceedings of the 14th ACM international conference on Multimedia', ACM, pp. 901–910.

Gao, Z., Xue, J., Zhou, W., Pang, S. & Tian, Q. (2016), 'Democratic diffusion aggregation for image retrieval', *IEEE Trans. Multimedia* **18**(8), 1661–1674.

Ginsca, A. L., Popescu, A., Ionescu, B., Armagan, A. & Kanellos, I. (2014), Toward an estimation of user tagging credibility for social image retrieval, *in* 'Proceedings of the 22nd ACM international conference on Multimedia', ACM, pp. 1021–1024.

Gong, C., Tao, D., Maybank, S. J., Liu, W., Kang, G. & Yang, J. (2016), 'Multi-modal curriculum learning for semi-supervised image classification', *IEEE Trans. Image Process.* **25**(7), 3249–3260.

Gong, C., Tao, D., Yang, J. & Liu, W. (2016), Teaching-to-learn and learning-to-teach for multi-label propagation, *in* 'Proc. Conf. AAAI', pp. 1610–1616.

Gong, Y., Jia, Y., Leung, T., Toshev, A. & Ioffe, S. (2013), 'Deep convolutional ranking for multilabel image annotation', *arXiv preprint arXiv:1312.4894* .

Gong, Y., Ke, Q., Isard, M. & Lazebnik, S. (2014), 'A multi-view embedding space for modeling internet images, tags, and their semantics', *Int. J. Comput. Vision* **106**(2), 210–233.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014), Generative adversarial nets, *in* 'Proc. Advances in Neural Inf. Process. Syst.', pp. 2672–2680.

Gordo, A., Almazán, J., Revaud, J. & Larlus, D. (2016), Deep image retrieval: Learning global representations for image search, *in* 'Proc. Eur. Conf. Comp. Vis.', Springer, pp. 241–257.

Guillaumin, M., Mensink, T., Verbeek, J. & Schmid, C. (2009), Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, *in* 'Proc. IEEE Int. Conf. Comp. Vis.', pp. 309–316.

Guo, Y. & Gu, S. (2011), Multi-label classification using conditional dependency networks, *in* 'Proc. Int. Joint Conf. Artificial Intell.', Vol. 22, p. 1300.

Han, Y., Yang, Y., Ma, Z., Shen, H., Sebe, N. & Zhou, X. (2014), 'Image attribute adaptation', *IEEE Trans. Multimedia* **16**(4), 1115–1126.

Härdle, W. K. & Simar, L. (2015), Canonical correlation analysis, *in* 'Applied Multivariate Statistical Analysis', Springer, pp. 443–454.

Harzallah, H., Jurie, F. & Schmid, C. (2009), Combining efficient object localization and image classification, *in* 'Proc. IEEE Int. Conf. Comp. Vis.', IEEE, pp. 237–244.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proc. IEEE Conf. Comp. Vis. Patt. Recogn.', pp. 770–778.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural Computation* **9**(8), 1735–1780.

Huang, F., Zhang, X., Li, Z., Mei, T., He, Y. & Zhao, Z. (2017), Learning social image embedding with deep multimodal attention networks, *in* 'Proceedings of the on Thematic Workshops of ACM Multimedia 2017', ACM, pp. 460–468.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017), Densely connected convolutional networks., *in* 'Proc. IEEE Conf. Comp. Vis. Patt. Recogn.', Vol. 1, p. 3.

Huiskes, M. J. & Lew, M. S. (2008), The MIR flickr retrieval evaluation, *in* 'Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2008, Vancouver, British Columbia, Canada, October 30-31, 2008', pp. 39–43.

Hwang, S. J. & Grauman, K. (2012), 'Learning the relative importance of objects from tagged images for retrieval and cross-modal search', *Int. J. Comput. Vision* **100**(2), 134–153.

Jaderberg, M., Simonyan, K., Zisserman, A. et al. (2015), Spatial transformer networks, *in* 'Proc. Advances in Neural Inf. Process. Syst.', pp. 2017–2025.

Johnson, J., Ballan, L. & Li, F. (2015), Love thy neighbors: Image annotation by exploiting image metadata, *in* 'Proc. IEEE Int. Conf. Comp. Vis.', pp. 4624–4632.

Johnson, J., Karpathy, A. & Fei-Fei, L. (2016), Densecap: Fully convolutional localization networks for dense captioning, *in* 'Proc. IEEE Conf. Comp. Vis. Patt. Recogn.'.

Joshi, D., Luo, J., Yu, J., Lei, P. & Gallagher, A. (2011), Using geotags to derive rich tag-clouds for image annotation, *in* 'Social media modeling and computing', Springer, pp. 239–256.

Ke, X., Zhou, M., Niu, Y. & Guo, W. (2017), 'Data equilibrium based automatic image annotation by fusing deep model and semantic propagation', *Pattern Recogn.* **71**, 60–77.

Kennedy, L., Slaney, M. & Weinberger, K. (2009), Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases, *in* 'Proceedings of the 1st workshop on Web-scale multimedia corpus', ACM, pp. 17–24.

Kim, G. & Xing, E. P. (2013), Time-sensitive web image ranking and retrieval via dynamic multi-task regression, *in* 'Proceedings of the sixth ACM international conference on Web search and data mining', ACM, pp. 163–172.

Kipf, T. N. & Welling, M. (2016), 'Semi-supervised classification with graph convolutional networks', *Proc. Int. Conf. Learn. Representations* .

Krähenbühl, P. & Koltun, V. (2014), Geodesic object proposals, *in* 'Proc. Eur. Conf. Comp. Vis.', Springer, pp. 725–739.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S. & Fei-Fei, L. (2017), Visual genome: Connecting language and vision using crowdsourced dense image annotations, *in* 'Int. J. Comput. Vision', Vol. 123, pp. 32–73.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'Proc. Advances in Neural Inf. Process. Syst.', pp. 1097–1105.

Lafferty, J., McCallum, A., Pereira, F. et al. (2001), Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *in* 'Proc. Int. Conf. Mach. Learn.', Vol. 1, pp. 282–289.

Li, J., Lin, X., Rui, X., Rui, Y. & Tao, D. (2015), 'A distributed approach toward discriminative distance metric learning', *IEEE Trans. Neural Netw. & Learn. Syst.* **26**(9), 2111–2122.

Li, S. Z. (2009), *Markov random field modeling in image analysis*, Springer Science & Business Media.

Li, X. & Snoek, C. G. (2013), Classifying tag relevance with relevant positive and negative examples, *in* 'Proc. ACM Int. Conf. Multimedia.', ACM, pp. 485–488.

Li, X., Snoek, C. G. & Worring, M. (2009), 'Learning social tag relevance by neighbor voting', *IEEE Trans. Multimedia* **11**(7), 1310–1322.

Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C. G. & Bimbo, A. D. (2016), 'Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval', *ACM Computing Surveys* **49**(1), 14.

Li, X., Zhao, F. & Guo, Y. (2014), 'Multi-label image classification with a probabilistic label enhancement model', *Proc. Uncertainty in Artificial Intell* .

Li, Y., Crandall, D. J. & Huttenlocher, D. P. (2009), Landmark classification in large-scale image collections, *in* 'Proc. IEEE Int. Conf. Comp. Vis.', IEEE, pp. 1957–1964.

Li, Y., Liu, J., Wang, Y., Bingyuan Liu, J., Gao, Y., Wu, H., Song, H., Ying, P. & Lu, H. (2015), Hybrid learning framework for large-scale web image

annotation and localization, *in* 'CLEF 2015 Evaluation Labs and Workshop, Online Working Notes'.

Li, Y., Tarlow, D., Brockschmidt, M. & Zemel, R. (2015), 'Gated graph sequence neural networks', *arXiv preprint arXiv:1511.05493* .

Li, Z. & Tang, J. (2017), Weakly-supervised deep nonnegative low-rank model for social image tag refinement and assignment., *in* 'Proc. Conf. AAAI', pp. 4154–4160.

Lin, M., Chen, Q. & Yan, S. (2013), 'Network in network', *arXiv preprint arXiv:1312.4400* .

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014), Microsoft coco: Common objects in context, *in* 'Proc. Eur. Conf. Comp. Vis.', pp. 740–755.

Liu, D., Hua, X.-S., Yang, L., Wang, M. & Zhang, H.-J. (2009), Tag ranking, *in* 'Proc. Int. Conf. World Wide Web.', ACM, pp. 351–360.

Liu, D., Yan, S., Rui, Y. & Zhang, H.-J. (2010), Unified tag analysis with multi-edge graph, *in* 'Proc. ACM Conf. on Multimedia', ACM, pp. 25–34.

Liu, J., Li, Z., Tang, J., Jiang, Y. & Lu, H. (2014), 'Personalized geo-specific tag recommendation for photos on social websites', *IEEE Transactions on Multimedia* **16**(3), 588–600.

Long, C., Collins, R., Swears, E. & Hoogs, A. (2018), 'Deep neural networks in fully connected crf for image labeling with social network metadata', *arXiv preprint arXiv:1801.09108* .

Lowe, D. G. (2004), 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vision* **60**(2), 91–110.

Lu, Z. & Wang, L. (2015), 'Learning descriptive visual representation for image classification and annotation', *Pattern Recogn.* **48**(2), 498–508.

Makadia, A., Pavlovic, V. & Kumar, S. (2008), A new baseline for image annotation, *in* 'Proc. Eur. Conf. Comp. Vis.', Springer, pp. 316–329.

Makadia, A., Pavlovic, V. & Kumar, S. (2010), 'Baselines for image annotation', *Int. J. Comput. Vision* **90**(1), 88–105.

McAuley, J. & Leskovec, J. (2012), Image labeling on a network: using social-network metadata for image classification, *in* 'Proc. Eur. Conf. Comp. Vis.', Springer, pp. 828–841.

McParlane, P., Whiting, S. & Jose, J. (2013), Improving automatic image tagging using temporal tag co-occurrence, *in* 'International Conference on Multimedia Modeling', Springer, pp. 251–262.

Melnyk, I., Sercu, T., Dognin, P. L., Ross, J. & Mroueh, Y. (2018), 'Improved image captioning with adversarial semantic alignment', *arXiv preprint arXiv:1805.00063* .

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J. & Khudanpur, S. (2010), Recurrent neural network based language model., *in* 'Interspeech', Vol. 2, p. 3.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* 'Proc. Advances in Neural Inf. Process. Syst.', pp. 3111–3119.

Miller, G. A. (1995), 'Wordnet: a lexical database for english', *Communications of the ACM* **38**(11), 39–41.

Murthy, V. N., Maji, S. & Manmatha, R. (2015*a*), Automatic image annotation using deep learning representations, *in* 'Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015', pp. 603–606.

Murthy, V. N., Maji, S. & Manmatha, R. (2015*b*), Automatic image annotation using deep learning representations, *in* 'Proc. ACM Int. Conf. Multimedia Retrieval.', ACM, pp. 603–606.

Ojala, T., Pietikainen, M. & Maenpaa, T. (2002), 'Multiresolution gray-scale and rotation invariant texture classification with local binary patterns', *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987.

Oliva, A. & Torralba, A. (2001), 'Modeling the shape of the scene: A holistic representation of the spatial envelope', *Int. J. Comput. Vision* **42**(3), 145–175.

Oquab, M., Bottou, L., Laptev, I., Sivic, J. et al. (2014), Weakly supervised object recognition with convolutional neural networks, *in* 'Proc. Advances in Neural Inf. Process. Syst.'.

Pereira, J. C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G. R., Levy, R. & Vasconcelos, N. (2014), 'On the role of correlation and abstraction in cross-modal multimedia retrieval', *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(3), 521–535.

Perronnin, F., Sánchez, J. & Mensink, T. (2010), Improving the fisher kernel for large-scale image classification, *in* 'Proc. Eur. Conf. Comp. Vis.', Springer, pp. 143–156.

Pont-Tuset, J., Arbelaez, P., Barron, J. T., Marqués, F. & Malik, J. (2017), Multiscale combinatorial grouping for image segmentation and object proposal generation, *in* 'IEEE Trans. Pattern Anal. Mach. Intell.', Vol. 39, pp. 128–140.

Rasiwasia, N., Mahajan, D., Mahadevan, V. & Aggarwal, G. (2014), Cluster canonical correlation analysis, *in* 'Artificial Intelligence and Statistics', pp. 823–831.

Ren, S., He, K., Girshick, R. & Sun, J. (2015), Faster r-cnn: Towards real-time object detection with region proposal networks, *in* 'Proc. Advances in Neural Inf. Process. Syst.', pp. 91–99.

Sawant, N., Datta, R., Li, J. & Wang, J. Z. (2010), Quest for relevant tags using local interaction networks and visual content, *in* 'Proc. ACM Int. Conf. Multimedia Retrieval.', ACM, pp. 231–240.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. (2009), 'The graph neural network model', *IEEE Transactions on Neural Networks* **20**(1), 61–80.

Shetty, R., Rohrbach, M., Hendricks, L. A., Fritz, M. & Schiele, B. (2017), Speaking the same language: Matching machine to human captions by adversarial training, *in* 'Proc. IEEE Int. Conf. Comp. Vis.'.

Shi, J. & Malik, J. (2000), 'Normalized cuts and image segmentation', *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(8), 888–905.

Sigurbjörnsson, B. & Van Zwol, R. (2008), Flickr tag recommendation based on collective knowledge, *in* 'Proc. Int. Conf. World Wide Web.', ACM, pp. 327–336.

Simonyan, K. & Zisserman, A. (2014), 'Very deep convolutional networks for large-scale image recognition', *arXiv preprint arXiv:1409.1556* .

Srivastava, N. & Salakhutdinov, R. R. (2012), Multimodal learning with deep boltzmann machines, *in* 'Proc. Advances in Neural Inf. Process. Syst.', pp. 2222–2230.

Stone, Z., Zickler, T. & Darrell, T. (2008), Autotagging facebook: Social network context improves photo annotation, *in* 'Computer Vision and Pattern Recognition Workshops', IEEE, pp. 1–8.

Sun, A. & Bhowmick, S. S. (2010), Quantifying tag representativeness of visual content of social images, *in* 'Proc. ACM Int. Conf. Multimedia.', pp. 471–480.

Tan, M., Shi, Q., van den Hengel, A., Shen, C., Gao, J., Hu, F. & Zhang, Z. (2015), Learning graph structure for multi-label image classification via clique generation, *in* 'Proc. IEEE Conf. Comp. Vis. Patt. Recogn.', pp. 4100–4109.

Tang, J., Li, H., Qi, G.-J. & Chua, T.-S. (2010), 'Image annotation by graph-based inference with integrated multiple/single instance representations', *IEEE Trans. Multimedia* **12**(2), 131–141.

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T. & Smeulders, A. W. M. (2013), 'Selective search for object recognition', *Int. J. Comput. Vision* **104**(2), 154–171.

Uricchio, T., Ballan, L., Seidenari, L. & Bimbo, A. D. (2017), 'Automatic image annotation via label transfer in the semantic space', *Pattern Recogn.* **71**, 144–157.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. & Bengio, Y. (2017), 'Graph attention networks', *arXiv preprint arXiv:1710.10903* .

Wang, C., Jing, F., Zhang, L. & Zhang, H.-J. (2008), 'Scalable search-based image annotation', *Multimedia Systems* **14**(4), 205–220.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C. & Xu, W. (2016), Cnn-rnn: A unified framework for multi-label image classification, *in* 'Proc. IEEE Conf. Comp. Vis. Patt. Recogn.', pp. 2285–2294.

Wang, X. & Gupta, A. (2018), 'Videos as space-time region graphs', *Proc. Eur. Conf. Comp. Vis.* .

Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y. & Yan, S. (2014), 'Cnn: Single-label to multi-label', *arXiv preprint arXiv:1406.5726* .

Wu, B., Chen, W., Sun, P., Liu, W., Ghanem, B. & Lyu, S. (2018), 'Tagging like humans: Diverse and distinct image annotation', *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* **3**.

Wu, Q., Wang, P., Shen, C., Reid, I. & van den Hengel, A. (2017), 'Are you talking to me? reasoned visual dialog generation through adversarial learning', *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* .

Wu, Q., Ye, Y., Zhang, H., Chow, T. W. & Ho, S.-S. (2015), 'Ml-tree: a tree-structure-based approach to multilabel learning', *IEEE transactions on neural networks and learning systems* **26**(3), 430–443.

Xu, H., Wang, J., Hua, X.-S. & Li, S. (2009), Tag refinement by regularized lda, *in* 'Proc. ACM Int. Conf. Multimedia.', ACM, pp. 573–576.

Xue, X., Zhang, W., Zhang, J., Wu, B., Fan, J. & Lu, Y. (2011), Correlative multi-label multi-instance image annotation, *in* 'Proc. IEEE Int. Conf. Comp. Vis.', IEEE, pp. 651–658.

Yang, Y., Huang, Z., Yang, Y., Liu, J., Shen, H. T. & Luo, J. (2013), 'Local image tagging via graph regularized joint group sparsity', *Pattern Recogn.* **46**(5), 1358–1368.

You, Q., Jin, H., Wang, Z., Fang, C. & Luo, J. (2016), Image captioning with semantic attention, *in* 'Proc. IEEE Conf. Comp. Vis. Patt. Recogn.', pp. 4651–4659.

Yu, A. & Grauman, K. (2014), Predicting useful neighborhoods for lazy local learning, *in* 'Advances in Neural Information Processing Systems', pp. 1916–1924.

Zhang, J., Wu, Q., Shen, C., Zhang, J. & Lu, J. (2018), 'Multi-label image classification with regional latent semantic dependencies', *IEEE Trans. Multimedia* .

Zhang, J., Wu, Q., Zhang, J., Shen, C. & Lu, J. (2018), 'Kill two birds with one stone: Weakly-supervised neural network for image annotation and tag refinement', *Proc. Conf. AAAI* .

Zhang, J., Wu, Q., Zhang, J., Shen, C. & Lu, J. (2019), 'Mind your neighbours: Image annotation with metadata neighbourhood graph co-attention networks', *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* .

Zhang, J., Wu, Q., Zhang, J., Shen, C., Lu, J. & Wu, Q. (2019), 'Heritage image annotation via collective knowledge', *Pattern Recognition* **93**, 204–214.

Zhang, J., Zhang, J., Lu, J., Shen, C., Curr, K., Phua, R., Neville, R. & Edmonds, E. (2016), Slnsw-uts: A historical image dataset for image multi-labeling and retrieval, *in* 'Proc. Int. Conf. Digital Image Computing: Techniques and Applications', pp. 1–6.

Zhang, J., Zhang, J., Wu, Q., Wu, Q., Xu, J., Lu, J., Phua, R., Curr, K. & Tang, Z. (2017), 'Historical image annotation by exploring the tag relevance'.

Zhang, M.-L. & Zhang, K. (2010), Multi-label learning by exploiting label dependency, *in* 'Proc. ACM Int. Conf. Knowledge discovery & data mining', ACM, pp. 999–1008.

Zhao, F., Huang, Y., Wang, L. & Tan, T. (2015), Deep semantic ranking based hashing for multi-label image retrieval, *in* 'Proc. IEEE Conf. Comp. Vis. Patt. Recogn.', IEEE, pp. 1556–1564.

Zhu, F., Li, H., Ouyang, W., Yu, N. & Wang, X. (2017), 'Learning spatial regularization with image-level supervisions for multi-label image classification', *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* .

Zhu, G., Yan, S. & Ma, Y. (2010), Image tag refinement towards low-rank, content-tag prior and error sparsity, *in* 'Proc. ACM Int. Conf. Multimedia.', ACM, pp. 461–470.

Zhu, S., Ngo, C.-W. & Jiang, Y.-G. (2012), 'Sampling and ontologically pooling web images for visual concept learning', *IEEE Trans. Multimedia* **14**(4), 1068–1078.

Zitnick, C. L. & Dollár, P. (2014), Edge boxes: Locating object proposals from edges, *in* 'Proc. Eur. Conf. Comp. Vis.', Springer, pp. 391–405.