Deep Learning based Image Retrieval with Weight Learning and Diffusion Modelling

Yan Zhao

Follow this and additional works at: https://ro.uow.edu.au/theses1
Deep Learning based Image Retrieval with Weight Learning and Diffusion Modelling

Yan Zhao

Supervisor:
A/Prof. Lei Wang
Co-supervisor:
Dr. Luping Zhou

This thesis is presented as required for the conferral of the degree:

Doctor of Philosophy

The University of Wollongong
School of Computing and Information Technology

November 24, 2019
Declaration

I, Yan Zhao, declare that this thesis submitted in fulfilment of the requirements for the conferral of the degree Doctor of Philosophy, from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Yan Zhao
November 24, 2019
To Jing.
Abstract

Image retrieval is to retrieve or search digital images from large databases. The difficulty of image retrieval lies at is the semantic gap, and that the images in a database usually do not have any labels. The research of image retrieval based on deep learning aims to design and propose new ideas, methods, and algorithms to improve its performance.

From the perspective that image retrieval is generally an unsupervised task, this thesis firstly analyses some typical unsupervised deep learning models. Under the unsupervised settings, this work compares these deep learning models with the classical BoW model. This gains valuable insights into unsupervised learning and network training for the subsequent image retrieval studies in this thesis.

After the introduction of deep learning, the computer vision community has realised that pure unsupervised learning is not as competitive as supervised CNN models when applied to image retrieval. Consistent with this knowledge, this thesis focuses on semi-supervised learning and proposes an improved spatial search algorithm, called semi-supervised weight learning (SWL). This method is to solve the problem of manually adjusting the spatial weights of the spatial search algorithm [RSMC15], which is one of the most advanced ConvNet-based image retrieval methods at that time. The experimental results on three benchmark datasets and a newly collected archival photo dataset demonstrate the effectiveness of the proposed weight learning method.

Further, focusing on the unsupervised nature of image retrieval, this thesis proposes a deep learning based method to model diffusion process to improve retrieval performance. The diffusion process, by exploiting potential neighbourhood structure of data, has been proven as an effective mechanism for improving image retrieval. This thesis innovatively treats the diffusion process as a “black box” and models it directly by training deep neural networks. It assimilates the effects of the diffusion process to obtain better image representation, and makes simple Euclidean search work more effectively. Experimental results conducted on multiple benchmark image retrieval datasets validate the advantage of the proposed method.

Finally, this thesis develops a real application of deep learning based image retrieval to the photographic collection of National Archives of Australia (NAA). This thesis introduces the background of NAA image retrieval problem, collects and labels the NAA29k dataset, builds up the retrieval system, and explores some related applications.
In sum, by analysing deep learning models, developing new retrieval algorithms based on deep neural networks, and realising practical image retrieval application, this thesis contributes to adding new knowledge and insights on deep learning based image retrieval.
Acknowledgments

It could not be possible to present this thesis without the support of a number of individuals and organisations. I would like to gratefully acknowledge them below:

Firstly, I would like to extend my sincerely gratitude to my supervisor, A/Prof. Lei Wang. He is experienced and professional in research. He is smart and hard-working. He always can come up with new ideas and discusses with us. He is also kind to help me for problems and always supports me, especially at some tough time. He supported me to attend DICTA2014, and ICML2017 and helped me prepare slides and spotlight for DICTA2016 and BMVC2018. He is also nice and encourages me to collaborate with other researchers. I really enjoy the time discussing ideas and papers with him. And I have been influenced by some of his good research habits and inspired by his perseverance.

I would like to thank my co-supervisor, Dr. Luping Zhou. She is always strict on research and also kind to help me revise the papers. She is brilliant and diligent. I think I am very lucky to be advised by the supervisors of A/Prof. Lei Wang and Dr. Luping Zhou. They are really dedicated for the academics and kind-hearted for life. I really admire their spirits and pursuit of excellence.

During my study, I am lucky to meet my colleagues. I would like to thank Chao Wang, Xinwang Liu, Lingqiao Liu, Jianjia Zhang, Zhimin Gao, Comor Ian, Zhongyan Zhang, Biting Yu, Yu Ding, Kevin Zhang and all other visiting scholars to VILA group. I really appreciate the time working with them. Without their support, I cannot complete my research.

I would like to extend my gratitude to Prof. Dongyun Yi. Thanks for his support so I can get the opportunity to study in Australia. I also like to thank A/Prof. Chenping Hou who introduced me to UOW and VILA group.

Thanks must go to Chinese Scholarship Council, and University of Wollongong for providing me PhD scholarship and tuition fee waiver scholarship. I also thank National Archives of Australia for allowing me to access their archival photo collections. I also want to thank Multi-modal Australian Sciences Imaging and Visualisation Environment (MASSIVE) and National Computational Infrastructure (NCI) to provide computing resources.

I would like to give my gratitude to my dear parents for their support. Special thank to my wife, Jing Cui for her help and understanding. Also, I would like to thank Wollongong, which is a really beautiful and lovely city.
Publications


Contents

Abstract iv

1 Introduction 1
  1.1 Background ........................................ 1
  1.2 Research Questions, Aim and Significance ............. 4
    1.2.1 Research Questions ............................ 4
    1.2.2 Research Aim .................................. 5
    1.2.3 Research Significance ......................... 5
  1.3 Contributions ..................................... 5
  1.4 Organisation of Thesis ............................. 7

2 Literature Review 8
  2.1 Evolution from Early Neural Networks to Deep Learning .... 8
  2.2 Image Retrieval ..................................... 10
  2.3 Image Retrieval based on Deep Learning .................. 11

3 Investigation of Unsupervised Deep Feature Learning 15
  3.1 Introduction ........................................ 15
  3.2 Related Work ....................................... 17
  3.3 Models ............................................. 17
    3.3.1 Bag-of-Words Models ............................ 17
    3.3.2 Deep Learning Models ........................... 18
  3.4 Experimental Study .................................. 21
    3.4.1 HEp-2 Cell Datasets ............................ 21
    3.4.2 Image Preprocessing ............................. 21
    3.4.3 Clustering ...................................... 22
    3.4.4 Results of the BoW Models ....................... 22
    3.4.5 Results of Deep Learning Models ................. 23
    3.4.6 Discussion ..................................... 26
  3.5 Conclusion ......................................... 28

4 Semi-supervised Weight Learning Method 32
  4.1 Introduction ........................................ 32
  4.2 Related Work ....................................... 34
  4.3 Proposed Method: SWL ............................... 35
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.1</td>
<td>Spatial Search Method</td>
<td>35</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Loss Function of the Proposed Method</td>
<td>37</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Triplet Generation</td>
<td>38</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Optimisation Problem for Learning Weights</td>
<td>38</td>
</tr>
<tr>
<td>4.4</td>
<td>Experimental Study</td>
<td>40</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Datasets</td>
<td>40</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Experimental Setting</td>
<td>42</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Experimental Results</td>
<td>42</td>
</tr>
<tr>
<td>4.5</td>
<td>Extension to Kernel SWL</td>
<td>47</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Kernel SWL</td>
<td>49</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Experiment</td>
<td>49</td>
</tr>
<tr>
<td>4.6</td>
<td>Discussion</td>
<td>51</td>
</tr>
<tr>
<td>4.7</td>
<td>Conclusion</td>
<td>52</td>
</tr>
<tr>
<td>5</td>
<td>Modelling Diffusion for Image Retrieval</td>
<td>53</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>53</td>
</tr>
<tr>
<td>5.2</td>
<td>The Proposed Approach</td>
<td>55</td>
</tr>
<tr>
<td>5.2.1</td>
<td>A Mapping View of Diffusion Process</td>
<td>57</td>
</tr>
<tr>
<td>5.2.2</td>
<td>A Deep Metric Learning Approach</td>
<td>58</td>
</tr>
<tr>
<td>5.2.3</td>
<td>A Bootstrapping Framework for Image Retrieval</td>
<td>59</td>
</tr>
<tr>
<td>5.3</td>
<td>Experimental Study</td>
<td>60</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Experimental Setup</td>
<td>60</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Result and Discussion</td>
<td>61</td>
</tr>
<tr>
<td>5.4</td>
<td>Fusion of Diffusion Based on NAA Database</td>
<td>65</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Fusion of Global and Regional Diffusion</td>
<td>68</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Fusion of Diffusion with Visual and Textual Information</td>
<td>70</td>
</tr>
<tr>
<td>5.5</td>
<td>Conclusion</td>
<td>72</td>
</tr>
<tr>
<td>6</td>
<td>NAA Image Retrieval System and Related Applications</td>
<td>74</td>
</tr>
<tr>
<td>6.1</td>
<td>Background of NAA Image Retrieval</td>
<td>74</td>
</tr>
<tr>
<td>6.2</td>
<td>NAA29k Dataset Collection and Annotation</td>
<td>75</td>
</tr>
<tr>
<td>6.2.1</td>
<td>NAA29k Dataset Annotation</td>
<td>76</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Properties of the NAA29k Dataset</td>
<td>77</td>
</tr>
<tr>
<td>6.3</td>
<td>NAA Image Retrieval Web Platform and Software</td>
<td>78</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Image Retrieval Web Platform</td>
<td>78</td>
</tr>
<tr>
<td>6.3.2</td>
<td>NAA Image Retrieval Software</td>
<td>79</td>
</tr>
<tr>
<td>6.4</td>
<td>Discovering and Gathering Region of Interests for NAA Users</td>
<td>82</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Discovering with Object Detection</td>
<td>85</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Discovering with Regional Image Caption Generation</td>
<td>86</td>
</tr>
<tr>
<td>6.4.3</td>
<td>Gathering Regions of Interests with Clustering</td>
<td>91</td>
</tr>
</tbody>
</table>
List of Tables

3.1 ACC for the BoW models and Deep Learning models. . . . . . . . . . 29
3.2 NMI for the BoW models and Deep Learning models. . . . . . . . . . 30
3.3 Comparison of the BoW model and the Deep Learning models. . . . . 31

4.1 Details of the image retrieval datasets. . . . . . . . . . . . . . . . . . 40
4.2 Comparison Spatial Search with other methods on three datasets. . . 43
4.3 Comparison the proposed method on four datasets. . . . . . . . . . . 43
4.4 Result of the proposed method on Holiday and NAA29k. . . . . . . . 47

5.1 Six datasets used in this experiment. . . . . . . . . . . . . . . . . . . 60
5.2 Comparison with two baselines under global image representation. . . 62
5.3 Comparison with the state-of-the-art image retrieval methods. . . . . 64
5.4 Comparison of average time and memory usage in online retrieval. . . 65
5.5 Global diffusion results obtained by $Method_{off}$. . . . . . . . . . . 69
5.6 Regional diffusion results obtained by $Method_{off}$. . . . . . . . . . . 69
5.7 Results of proposed fusion methods. . . . . . . . . . . . . . . . . . . . 69
5.8 Visual and textual information fusion results. . . . . . . . . . . . . . 72
List of Figures

2.1 The evolution of early neural networks to deep learning. ........... 14
3.1 Framework of the BoW model with spatial pyramid matching (SPM). 18
3.2 Denoising Autoencoder. ............................................. 19
3.3 Restricted Boltzmann Machine. ....................................... 19
3.4 Stacked Denoising Autoencoder. ...................................... 19
3.5 Deep Belief Network. .................................................. 19
3.6 Examples of HEp-2 cell images from two datasets. .................. 21
3.7 The effect of patch size and pooling methods for different BoW models. 24
3.8 The effect of patch size and feature number for single-layer networks. 25
4.1 Flowchart of the proposed SWL method. .............................. 36
4.2 Example images in the new dataset NAA29k. ......................... 41
4.3 Effect of hyper-parameters $\gamma$ and $C$ in the proposed learning method. 45
4.4 Results of purely supervised weight learning on the four datasets. ... 46
4.5 Results of purely supervised learning on NAA29k and Holiday. ...... 46
4.6 Example retrieval results on NAA29k dataset. ....................... 48
4.7 Training and test curve of SWL with single kernel. .................. 50
5.1 The proposed framework consists of four components for offline training 56
5.2 The proposed method performance with respect to image insertion. ... 66
6.1 NAA image retrieval system based on BoW framework[WW13] ...... 75
6.2 Early NAA image retrieval system[WW13] .......................... 75
6.3 Some queries and their truly similar images in NAA29k. ............ 77
6.4 NAA image retrieval web platform. .................................. 80
6.5 NAA image retrieval web platform with a retrieval example. ....... 81
6.6 NAA Image Retrieval GUI. ............................................ 82
6.7 NAA Image Retrieval System Flowchart. ............................ 83
6.8 CNN feature extraction flow chart. .................................. 84
6.9 Test results obtained using faster-rcnn on NAA29k. .................. 86
6.10 Region-based caption generation for NAA. .......................... 88
6.11 Region-based caption generation for NAA: an example. .............. 89
6.12 Comparison of retrieval result with ground truth: a sample. ....... 90
6.13 Part of first stage clustering results. .................................. 92
A.1 An example to show the unsatisfactory retrieval result when using
deep features .................................................. 118
Chapter 1

Introduction

1.1 Background

Image retrieval, as one of the computer vision tasks, is to retrieve and search digital images from large databases. It has been largely used in daily life and also serves for other computer vision tasks. After decades of research and development, it has accumulated more than thousands of methods. These methods can be roughly divided into three categories: text-based image retrieval [RHC99], content-based image retrieval [SWS+00b], and semantic-based image retrieval [Bra00]. Text-based image retrieval relies on text annotations. Without manual annotations, this kind of retrieval is not well performed, and such methods disregard the visual content of the image and lose visual information. Traditional content-based image retrieval usually uses low-level visual features such as color features [JV96], edge features [JV96], texture features [MM96], and GIST [OT01]. These low-level visual features are often not directly understood by humans, but also result in the gap between low-level visual features and high-level semantic concepts. Semantic-based image retrieval attempts to narrow this gap by emphasising the association between visual and semantic concepts and constructing a high-level concept [MKS03]. The difficulty and challenge of image retrieval research is how to reduce the semantic gap. Existing image retrieval methods still cannot completely solve this problem.

Deep learning, due to its powerful feature learning ability and hierarchical representation of features, has affected and changed almost every field of computer vision in recent years, so as to influence and change people’s lifestyle. This change and impact come first from the recognition of deep learning concepts at top academic conferences. Since 2012, deep learning has gradually become the mainstream research field for all kinds of top computer vision conferences and machine learning conferences, such as CVPR, ICCV, ECCV, NIPS, ICML, AAAI, and IJCAI. Deep learning has significant advantages in various computer vision tasks, such as some typical applications: image classification [KSH12], image recognition [LGRN09] and object detection [GDDM14]. Benefiting from the excellent performance of deep learning in computer vision applications, deep learning is also rapidly changing and influencing human life. Many well-known companies, such as Google, Facebook, Microsoft, Baidu, Apple, IBM, Adobe, Netflix, NVIDIA, and NEC, as well as some well-known start-ups, such as Face++, iFlytek, and SenseTime Technology, have established deep learning research institutions and developed deep learning technolo-
gies. These companies have commercialised face recognition technology, pedestrian detection technology, driver-less technology, assisted driving technology and video analytics technology, and provided better service and gradually changed people's lifestyles. In addition, deep learning has also affected other fields of application outside of computer vision, such as natural language processing, machine translation and video games, and even art fields such as music and painting.

Deep learning has brought impact to computer vision research, and this influence has also increased with the development of deep learning, especially the performance of convolutional neural networks. In recent years, image classification problem on ImageNet datasets has driven the performance of convolutional neural networks to rise. A series of landmark convolutional neural network models have been proposed, such as Alexnet, VGG, Googlenet, and Resnet, etc. With the model depth going deeper, newly proposed models are refreshing the record of image classification tasks. In general, convolutional neural networks pre-trained on ImageNet datasets can be directly used for image feature extraction and utilised in related applications. Since the convolutional neural network has achieved strong feature representation capability, the network learned based on the ImageNet classification task has a certain invariance and can be utilised in other application fields with similar image datasets. Compared to this simple method of borrowing features, the fine-tuning method combined with new applications can achieve better results, which belongs to the scope of transfer learning. In some specific supervised learning tasks, such as object detection, the convolutional neural network trained with ImageNet can be fine-tuned with new labelled images to achieve even better results. This fine-tuning approach has greatly contributed to the enthusiasm of researchers in various fields. Relying on this simple-tuning transfer learning method, many test datasets in the field of computer vision were continuously conquered from 2013 to 2015. However, at that time, the influence of deep learning on image retrieval research still stays at the stage of borrowing features. Despite this, a number of image retrieval methods based on deep learning have been proposed and improved, and early research of deep learning based image retrieval has been opened.

This thesis summarises the situation and characteristics for deep learning based image retrieval before 2016 as follows:

1) The dataset of the image retrieval problem is different from the classification dataset. It does not contain any data labels. In short, it cannot be trained directly by supervised methods. The method of fine-tuning convolutional neural networks cannot be directly applied to image retrieval tasks. This greatly limits the effect of convolutional neural networks on image retrieval datasets due to the differences between datasets. The progress in the field of image retrieval is slower than the rapid development of other applications in the field of deep learning.
2) The mainstream models of this period are still Alexnet and VGG, etc., which are now considered as relatively shallower models, while the deeper models such as Resnet have just been proposed but have not been widely applied at that time. The feature properties provided by these shallower models have not yet achieved very high performance. Simply using these models to extract features could often be defeated by using typical SIFT-based image retrieval methods (e.g., BoW [FPSV14], Fisher Vector [PD07]). This also slows down the development of deep learning for image retrieval.

3) Due to the first two reasons, many researchers have focused on how to better borrow the characteristics of these convolutional neural networks. Since convolutional neural networks have multiple layers and each of convolutional layers is capable of outputting three-dimensional features. It is an active research topic to study which layer to extract features, how to compress these three-dimensional features, and how to post-process these features. Also being affected by the traditional CBIR image retrieval framework, some researchers propose ideas by considering multi-scale effects and spatial information has also been applied to the extraction and processing of deep learning features. There are many representative methods accumulated in this stage, and the best performance is achieved by the spatial search algorithm [RSMC15]. It has maintained its leading position in multiple datasets for a period of time. By extracting and compressing the convolutional neural network features of different image regions and weighting the distance measure of different regions of the image, spatial search method can effectively consider the multi-scale, different positions and other information in the image, thereby better measuring the image similarity between two images. However, these spatial weights need to be manually adjusted through an exhaustive search based on trial and error. In order to improve this situation, this thesis proposes an automatic semi-supervised weight learning method (SWL) for spatial search algorithms. This method can effectively solve the problem of manually adjusting the weights and achieve better results.

Since 2016, the impact of deep learning on image retrieval research has improved significantly. Research on image retrieval based on deep learning has been further developed, which benefits from three aspects:

1) The proposed and widely used Resnet model significantly improves the effect of convolutional neural networks in image classification. This makes it simple to borrow the model of deep learning to extract features that can surpass the traditional CBIR image retrieval methods. Despite this improved performance from the model rather than algorithms, it still motivates researchers to combine deep learning with the field of image retrieval. Researchers have generally realised the importance of deep learning methods in image retrieval.

2) Fine-tuning with similar datasets can be helpful. By collecting labelled
datasets that are similar to the datasets to be searched on, researchers can make the fine-tuning method feasible. For example, the commonly used Oxford5k dataset contains some well-known architectural images. By collecting a new Landmark dataset with annotation information, Resnet fine-tuned on the Landmark dataset can achieve significantly improved retrieval performance on Oxford5k dataset. Although this method can achieve significant improvements, it has obvious drawbacks. It takes extra time and effort to collect similar datasets, and to label the similar dataset also relies on manual or complex annotation methods. And the generalisation ability of this method is also limited when a new dataset is given. It is often necessary to re-collect additional datasets that are similar to the new dataset and have sufficient annotations.

Considering the above two aspects together, the performance of image retrieval can be greatly improved. For example, the performance on the two benchmark datasets of Oxford5k and Paris6k can be raised from about 80 percentage points to above 90 percentage points [GARL17, RTC16] (out of 100 percentage points).

3) Drawing on some image retrieval post-processing methods in traditional CBIR, retrieval performance is further improved. For example, prior to deep learning, the diffusion process has been proved as an effective mechanism for improving image retrieval by exploiting potential neighbourhood data structures. Recently, by combining with deep learning, the effectiveness of the diffusion process to improve image retrieval has been verified again [ITA+17], and the performance for the two benchmark datasets Oxford5k and Paris6k has almost become saturated. However, this method still has some shortcomings. The diffusion process requires considerable space to store larger neighbourhood maps, spends more online retrieval time, and requires special algorithms in addition to simple Euclidean search. In order to solve these problems, this thesis proposes a method to directly model the diffusion process. The method can assimilate the effect of the diffusion process by training the deep neural network and can obtain better image representation. It only needs to perform simple Euclidean search when retrieving new images and completely avoids the online diffusion process in retrieval. Also, this method is unsupervised and does not require image annotations or external similar, labelled datasets.

1.2 Research Questions, Aim and Significance

1.2.1 Research Questions

Although deep learning has brought tremendous changes in image retrieval, the problem of image retrieval has not been completely solved.

1) Semantic gap. The difficulty of image retrieval is the semantic gap, and the
images do not have any labels. Research on image retrieval based on deep learning aims to design and propose new types of retrieval algorithms based on deep learning. In terms of the semantic gap, deep convolutional neural networks can provide more effective means to extract visual features than traditional CBIR retrieval methods, thus narrowing the gap between visual features and semantic concepts.

2) Unsupervised learning. However, deep learning, especially convolutional neural networks, still have a strong dependence on labelling information. How to construct an unsupervised deep network training method for image retrieval requires researchers to further explore new network training methods and propose better unsupervised training methods.

1.2.2 Research Aim

The overall aims of this thesis include: 1) to design and propose advanced retrieval algorithms based on deep learning, and 2) to explore new network training approach and propose better unsupervised training methods.

1.2.3 Research Significance

Image retrieval is a widely applicable computer vision task. Investigating algorithms with advanced visual features is a key point to improve retrieval performance. Deep learning is a rising research field with strength in learning over-completed hierarchical representations. Its effectiveness has been shown on many computer vision tasks and needs to be further evaluated on image retrieval. By designing advanced deep learning based algorithms for image retrieval, this thesis can contribute to deep learning and improve image retrieval as well. Also, the developed deep learning based algorithms could potentially be useful for other computer vision applications.

1.3 Contributions

The thesis has done the following work in model analysis, algorithm design, and specific applications:

(1) From the perspective that image retrieval is an unsupervised task, this thesis firstly analyses some unsupervised deep learning models for the general unsupervised task. Under unsupervised settings, the study compared the deep learning models with the classical BoW model. This thesis has compared the features of these two typical models through clustering tasks. Experimental results show that the BoW model generally performs better than unsupervised deep learning models around 2013. This provides valuable experience for dealing with the unsupervised issues and network training in subsequent image retrieval studies.
(2) In the early study of this thesis, an improved spatial search algorithm is proposed. As the most advanced ConvNet-based image retrieval method at that time, the spatial search [RSMC15] has shown excellent retrieval performance and is superior to other competitors. A key component of the method is a weighted combination of distances evaluated in different regions of the query image. However, these weights are currently manually adjusted through an exhaustive search based on trial and error. This not only causes a lengthy parameter adjustment process but also makes it difficult to guarantee the optimality of the weights. Moreover, these weights are typically not applicable when the nature of the image dataset changes. In order to improve this situation, this thesis proposes to automatically learn the combination weights by developing a new semi-supervised weight learning (SWL) method. Experimental results on three benchmark datasets and newly collected archival photo datasets demonstrate the effectiveness of the proposed weight learning approach. Furthermore, this thesis extends this SWL method with the kernel instead of the Euclidean distance.

(3) In the later study of this thesis, a diffusion process modelling method is proposed. Diffusion process, by exploiting the underlying neighbourhood structure of data, has been shown as an effective mechanism to improve image retrieval. However, the diffusion process stores large neighbourhood maps, takes more online retrieval time, and requires special algorithms in addition to simple Euclidean search. In order to solve these problems, this thesis proposes to regard the diffusion process as a “black box” and directly model it by training deep neural networks to obtain better image representation, assimilate the effect of the diffusion process, and realise a simple Euclidean search. This thesis firstly proposes a kernel mapping interpretation of the diffusion process, and then represents the modelling as a deep metric learning problem. The proposed method is unsupervised because it requires neither image annotations nor external datasets, and completely avoids the online diffusion process in retrieval. Experiments validate its effectiveness and study its attractive features, such as the new image insertion.

After understanding the effectiveness of the diffusion process on image retrieval, how to design a better diffusion process become an important research direction. In recent years, from the perspective of data fusion, methods for integrating multiple diffusion processes have been proposed and studied. Also, application to the archival photo collection of National Archives of Australia (NAA) has always been the driving force behind the problem of deep learning image retrieval in this thesis. It has a good application scenario and practical significance. Another major content is the exploration of a diffusion-fusion framework based on the NAA database. Around the NAA database, this thesis explores the fusion process based on the diffusion process from several different perspectives in order to better realise the effects of
the diffusion process: 1) simple global and regional diffusion fusion; 2) diffusion and fusion of visual and textual information.

(4) Finally, at the application level, an application of image retrieval is developed. This thesis introduces the background of NAA image retrieval problem, the collection and labelling of NAA29k dataset, the retrieval system construction, and some related applications.

1.4 Organisation of Thesis

The thesis is organised as follows:

The first chapter introduces the background, research question and contribution of this thesis.

The second chapter introduces the current status of relevant research.

The third chapter studies several unsupervised deep learning models and compares the performance of the BoW model and the deep learning models in unsupervised learning tasks. Experimental results show that the BoW model studied in this chapter generally performs better than the deep learning models studied in this chapter. Also, this chapter illustrates that the BoW model and the deep learning models have complementary properties.

The fourth chapter develops an automatic learning method for combining weights in the spatial search algorithm. It proposes to develop optimisation problems under the distance metric learning framework and use the gallery images to generate unsupervised constraints to help learning. In addition, an efficient solver algorithm is developed and evaluated on the three benchmark datasets in image retrieval and the NAA datasets. The experimental results show the effectiveness and advantages of the method.

The fifth chapter proposes a method for modelling highly non-linear diffusion processes. By using the powerful modelling capabilities of deep neural networks, a clear and better representation of features is generated for image retrieval. This method preserves the positive effects of the diffusion process while reducing the computational cost and retrieval complexity of online diffusion. An interesting unsupervised learning framework for guiding image retrieval is proposed. By using the underlying structural information of the images in the database, the sophisticated diffusion process can be converted into a simpler and better Euclidean search.

The sixth chapter focuses on the practical application of image retrieval. This chapter introduced NAA with its background, the collection and labelling of NAA29k dataset, the retrieval system construction and related applications.

The last chapter summaries the work conducted in this thesis, discusses the limitations of the current work, and puts forward future research ideas.
Chapter 2

Literature Review

This chapter reviews the literature related to the work in the thesis from three parts: 1) deep learning model evolution; 2) image retrieval; and 3) deep learning based image retrieval.

2.1 Evolution from Early Neural Networks to Deep Learning

Deep learning, as a way to implement deep neural networks, has recently gained wide-ranging influence. Because it can perform significantly better than traditional models and algorithms, it has demonstrated excellent applications in various machine learning tasks such as image recognition, speech recognition and natural language processing, among others. One key to the success of deep learning is that it has found a way to effectively train deep neural networks. In the 1980s, the Back Propagation Algorithm (BP) [RHW88] was first introduced for training Multilayer Perceptron (MLP) [Ros58], which has a good performance on a shallow network. It was then used to train deep networks, but after a long exploration, it still failed. In 1991, researchers clearly noticed the problem of disappearing or bursting gradients in training. This is regarded the basic issue of deep learning today [Sch14]. In 2006, Hinton proposed a novel way to initialise parameters to deal with the basic issue to optimise deep neural networks by using layered pre-training of multi-layer perceptrons [HOT06]. Researchers then proposed a number of deep learning models, such as Deep Belief Networks (DBN) [HOT06], Deep Autoencoders and Deep Boltzman Machine (DBM) [SH09]. Models emerging from the evolution of early neural networks (NN) to deep learning (DL) can be analysed from three perspectives. They are: 1) shallow or deep; 2) supervised or unsupervised; 3) direct encoding or probabilistic. This thesis summarises the entire evolution process and illustrates it in Fig.2.1.

1) From shallow to deep. In the early 1980s, the models were almost shallow, including linear regression, Perceptron [Ros58], linear autoencoder, and Hopfield network [Hop82]. At that time, they were all deterministic models although they could be given a probabilistic explanation. The probabilistic models appeared until Boltzmann Machine [AHS85] was proposed. With the development and application of neural networks, hierarchical models are studied. A typical example is Multilayer
Perceptron [RRK+90], which is a basic feedforward neural network that faces the problem of disappearing or bursting gradients [Sch14]. In fact, the emergence of deep learning reflects the fact that more levels and enough complexity are needed to build better models.

2) From supervised models to unsupervised models. Some studies have shifted to learning unsupervised models that can be used to better train supervised models [HOT06, SH09]. Since Hinton proposed a layered pre-training scheme for feedforward neural networks in 2006, this has been a sound and versatile method. Many models such as DBN and DBM use this scheme to pre-train the feedforward neural network and fine-tune the parameters through the labels. However, as the number of data increases, especially for tagged data, it seems unnecessary to use an unsupervised model to pre-train the supervised model. For example, for the convolutional neural network CNN [LB95], there is no significant difference between the results using pre-training and individual random initialisation when a sufficient number of labelled data are available. This raises the question about the need to use an unsupervised model to pre-train the supervised model.

3) From direct encoding to probabilistic models. For the shallow models, linear autoencoders and Hopfield networks [Hop82] are both deterministic unsupervised models. For deep models, deep autoencoders are stacked by multiple linear autoencoders. It usually utilises some techniques, including sparse coding, manifold learning, denoising and shrinking, to build variants like automatic encoder [AB14], denoising autoencoder [SOC+12], and compression autoencoder [RVM+11]. Direct encoding models can also be transferred to probabilistic models. For example, Hopfield networks (direct encoding) can be transferred to Boltzmann Machine (probabilistic models) by adding random units instead of treating them as deterministic units. This thesis analyses these two paradigms of direct encoding and probabilistic modelling. On one hand, these two paradigms have different perspectives to build the network. For example, Boltzmann machines pursue logical and meticulous solutions by using random approximations, while automatic encoders seek fast and usable solutions by using deterministic approximations. On the other hand, these two paradigms have some shared models, and hence have strong relationships. One representative shared model is the Sigmoid Belief Network [SJJ96] which was first proposed as a deterministic structure, and then was transferred to probabilistic models by using random units. What’s more, based on the Sigmoid belief network, Hinton proposes a deep belief network [SH09] by stacking RBMs. Also, there is evidence of strong relationships between the two paradigms. For example, the Gaussian RBM equals the denoising autoencoder if estimating it by score matching technique [Vin11]. Other strong relationships can also be found in the development of the two paradigms, like the Predictive Sparse Decomposition [KRL10] is a model
which combines the ideas of the two paradigms, as well as the Gaussian process latent variable model [Law03].

2.2 Image Retrieval

Image retrieval has been an active research area since 1970. From the perspective of different methods that are bases on, image retrieval can be separated into text-based image retrieval (TBIR) [RHC99], content-based image retrieval (CBIR) [SWS+00b], and Semantic-based Image Retrieval (SBIR) [Bra00]. In the early days of development, TBIR used manually annotated images, keywords or descriptions to retrieve. It is limited by the difficulty of manually annotating large image databases and the inaccuracy of comments with similar content [AEEB15]. CBIR uses the visual features of the image instead of annotations to overcome the above limitations, so image representation is the key to CBIR. CBIR has been proposed with different visual features, such as color features [JV96], edge features [JV96], texture features [MM96], GIST [OT01] to name just a few. Typical methods include the Bag-of-Words (BoW) model [YJHN07, Liu13] and the Fisher Vectors-based model [PLSP10, JDS11]. The BoW model [FPSV14] mainly includes the following stages: feature extraction, dictionary learning, feature coding, and feature pooling. The BoW model is also widely used in other areas of computer vision in addition to image retrieval. The advantage of CBIR is the use of rich visual features, but there is still a semantic gap between low-level visual features and high-level semantic concepts. In order to narrow this gap, many attempts have been made based on conceptual methods. For example, using the object ontology construction concept [MKS03] and semantic signature [WLT11] has formed a new method called semantic-based image retrieval (SBIR). SBIR pays more attention to concepts and their association with visual features than the visual features themselves. Other methods, such as automatic image annotation [SNP07], image tagging [GWCT10], and title generation [KPO+13] are also closely relevant to image retrieval, and they can also be used to reduce the semantic gap. At the same time, from the perspective of image contents, image retrieval can be divided into instance level (or object level) image retrieval and scene level image retrieval. For object-level image retrieval, its main challenge comes from the variations of the same object, which is usually caused by geometric transformations, illumination changes, and the diversity of objects in the same category. These challenges are also involved in other object-level image applications, such as object recognition [Dau93] and object detection [VJ01]. However, scene-level image retrieval faces more difficulties. Scene-level images usually contain multiple objects. The mutual relationship, relative position and overall semantic level of the objects bring new challenges to image retrieval.
2.3 Image Retrieval based on Deep Learning

The literature closely related to the thesis consists of three parts. The first part of the literature is mainly before 2016, which focuses on the feature extraction design for deep features. The second part appears after 2016 which mainly includes the methods based on fine-tuning network. And the last part focuses on the methods involving diffusion processes.

The first attempt to conduct image retrieval with deep learning was to train the deep autoencoder [KH11] in an unsupervised manner. The deep autoencoder is pre-trained by Deep Belief Network (DBN) [HOT06] and tested on the 80 million small image dataset [TFF08]. After that, deep learning, especially the Convolutional Neural Network (CNN) [KSH12], has made significant progress in image classification [KSH12]. Researchers found that the CNN model has considerable generality. The CNN model learned from an image classification task could be effectively applied to other recognition tasks [OBLS14, DJV14+, ZF14]. Gradually, this generality has driven CNN’s impact on various areas of computer vision. The first study using CNN for image retrieval was conducted in [RASC14a], but the performance was not satisfactory. With additional spatial information, CNN’s performance becomes comparable to traditional SIFT-based BoW and VLAD encoding. After that, multi-scale orderless collection (MOP) [GWGL14] uses a sliding window to aggregate multi-scale image blocks instead of extracting CNN features from the whole image, and the retrieval performance is improved. However, this method has difficulty in selecting the window size. At the same time, researchers have begun working on how to extract CNN features. Firstly, the second last fully connected layer is utilised to extract features, which was originally designed to optimise the image classification performance [BSCL14]. Soon after, extracting features from higher convolutional layers became more popular because it retains more object information, which was helpful for image retrieval. Since the convolutional layer output is a 3D array, how to map it to a multidimensional feature vector has attracted interest. By borrowing ideas from the traditional CBIR framework, some pooling methods, such as max-pooling [BL15], R-MAC [TSJ15], and spatial and channel-aware weighting pooling methods [KMO16] were correspondingly proposed. Another approach that is closer to the traditional CBIR is to treat each local descriptor in the 3D array (along its depth dimension) as “super SIFT”, and then encode these local descriptors using the established VLAD or Fisher vector model and merge the coding coefficients on the feature map [GWGL14, NYD15]. Different from the traditional CBIR, a spatial search algorithm based on deep features [RSMC15] is proposed. The key component of this method is to evaluate, in a weighted combination way, the distances based on the different regions of an image. Excellent results were obtained in benchmark re-
CHAPTER 2. LITERATURE REVIEW

trieval datasets. The above literature shows that CNN can effectively better bridge
the semantic gap between low-level visual features and semantic concepts.

Another attempt is to fine-tune CNN on new datasets by trimming the CNN
to learn better representations [GARL17, RTC16, GARL16, AGT+18, RIT+18,
NAS+17, TAZS19]. Although the retrieval performance is good, it is time-consuming
to collect and label new datasets. This is because, in order to train the network,
additional datasets need to be collected related to a given image retrieval task, and
careful data selection or cleaning procedures need to be implemented. Two recent
papers [GARL17] and [RTC16] have basically employed the fine-tuning strategy for
image retrieval of landmark buildings. These two works both depend on some aux-
iliary datasets. [GARL17] uses the clean version of the labelled dataset provided in
the work of [BSCL14] and [GARL17] uses the dataset obtained from the Internet
through the landmark keyword search. Triplet networks containing three CNN tow-
ers are fine-tuned by mining positive and negative image pairs. By fine-tuning the
CNN in this way, the image retrieval performance on the target dataset is largely
improved. The current fine-tuning strategy has succeeded in giving CNN the inform-
ation that it needs to learn. The collection of target-related image sets is helpful
to better understand the image changes, and triplet mining is helpful to explore
the invariance of target objects. However, this strategy relies heavily on data col-
lection. First, collecting target-related datasets is not an efficient solution and can
be time-consuming, especially when dealing with multiple different retrieval tasks.
Secondly, the keyword search engine module is used in the above dataset collection
strategy, which requires a clear description of the query image or given concepts.
This description is simple for a single concept datasets, like Oxford5k and Paris6k.
However, the real difficulty lies in the fact that in the real image retrieval appli-
cation, the images are often more complex and not well indexed. It will be more
difficult to clearly construct the detailed concepts, and then collecting the related
training dataset becomes a huge and complicated task, which is not in line with the
actual application conditions.

At the same time, some post-processing methods in the traditional CBIR can be
used to further improve retrieval. For example, prior to deep learning based image
retrieval, the diffusion process has proven to be an effective mechanism for impro-
ing image retrieval by utilising potential neighbourhood data structures. Although
there are many ways of doing this, they can be organised around three key factors:
the initialisation method, the definition of the transfer matrix, and the diffusion
method. Unlike the typical image retrieval which directly measures image similarity
by pairwise affinity, the diffusion process propagates pairwise affinity on the neigh-
bourhood graphs constructed with these images. By diffusion, the improved affinity
is used as a ranking score for image retrieval. In practice, even with significant
changes such as lighting, proportions, or cluttered background, the diffusion process still facilitates image retrieval to link two similar images. In addition, [BZW+19] studied the issue of optimal fusion of multiple similarity measures in image retrieval based on diffusion process. The proposed regularised set diffusion proves once again the effectiveness of the diffusion process for image retrieval in this case.

The above three parts of literature review gives a fundamental understanding of deep learning model evolution, development of image retrieval, and recent research field of deep learning based image retrieval. It helps the thesis to focus on some important research issues. It helps the thesis analyse some limitations of some current methods dealing with the issues. It also inspires the thesis to propose better deep learning based image retrieval methods in the following chapters.
Figure 2.1: Models emerging from the evolution of early neural networks (NN) to deep learning (DL) can be analysed from three perspectives: 1) shallow or deep; 2) supervised or unsupervised; 3) direct encoding or probabilistic. To well demonstrate the evolution, relationship (green dashed) between two models is annotated with a simple sentence. Note: CNN is not considered in this graph. The basic methods are marked in blue. The variants are marked in purple. The type of methods is marked by green (supervised or unsupervised).
Chapter 3

Experimental Investigation of Unsupervised Deep Feature Learning

From the perspective that image retrieval is an unsupervised task, this thesis firstly analyses some unsupervised deep learning models for general unsupervised tasks. Under the unsupervised setting, the thesis compares the deep learning models with the classical BoW model. The thesis investigates the characteristics of these two typical models by the clustering task. Experimental results show that the BoW model generally performs better than these early unsupervised deep learning models. Also, this thesis illustrates the complementary properties of the BoW model and the unsupervised deep learning models.

3.1 Introduction

At the early study of deep learning, unsupervised deep feature learning is a general but difficult task to the research community. Investigation of unsupervised deep feature learning is important and will be useful for the following research of deep learning based image retrieval in this thesis. At that early time, different to nowadays, the performance of unsupervised deep models is not satisfactory since they are relatively shallow, and training strategies are relatively simple. It is easy to ask whether unsupervised deep feature learning can surpass the classical methods, such as the typical BoW model. To answer this question, this thesis analyses unsupervised deep learning models by using the clustering of HEp-2 cell images as a case study.

Automatic identification of HEp-2 cell images has recently received increasing research attention. In pathology, the antinuclear antibodies (ANAs) test is a common methodology to diagnose systemic autoimmune diseases, such as Systemic Lupus Erythematosus, Rheumatoid Arthritis, and Sjogren’s syndrome. One of the standard methodologies for autoimmune diseases diagnosing [MS10] is Indirect immunofluorescence (IIF) on Human Epithelial-2 (HEp-2) cells, since it is effective for ANAs test. The staining patterns in IIF images can indicate the presence of ANAs in the patient serum and some specific autoimmune diseases are related to the type of the staining patterns. Manual evaluation of IIF images via visually analysing the distribution of staining patterns within the images leads to subjective results, which has low repeatability [FPSV13]. Automatic identification of staining patterns can
greatly address these limitations and has attracted an increasing amount of interest. Feature representation plays a crucial role in achieving a good identification performance. Many methods have been proposed for automatic identification of HEp-2 staining patterns in recent years, especially since 2012 when the first HEp-2 cells classification competition was held by International Conference on Pattern Recognition (ICPR). These methods can be categorised into two categories. One category is to construct the methods by utilising the hand-crafted features. One of the popular frameworks is the bag-of-words (BoW) model [FPSV14]. The other category is to learn the hierarchical representation by training deep neural networks. Although these two categories of methods have achieved promising performance with different perspectives, their success can both be attributed to the effective feature representation.

Supervised feature learning for automatic cell identification depends on labelled images. However, these labels are often expensive to obtain. They are commonly manually annotated by specialists with intensive labour and high cost of time.

This thesis therefore focuses on unsupervised feature learning. It compares the quality of the features of these two typical models through clustering. The thesis investigates the two categories of methods. For the first category, this thesis considers the typical BoW model. For the second category, it adopts single-layer networks such as Denoising Autoencoders [VLL+10] and Restricted Boltzmann Machines [Smo87] and further moves to multi-layers neural networks including Stacked Denoising Autoencoders (SDAE) [SOC+12] and Deep Belief Network (DBN) [HOT06]. With these learnt features, this thesis employs the widely used k-means for clustering with two measurements. The contributions of this chapter are summarised as follows: 1) It evaluates the performance of the BoW models and the deep learning models in unsupervised learning tasks on HEp-2 cell image dataset. Generally, the BoW model obtains better performance than these early deep learning models. Its success can be attributed to the local hand-crafted features and the generation of a discriminative codebook. For these deep learning models, their performance seems to be constrained by the inadequate training samples. 2) The thesis finds that the BoW and the deep learning models have complementary properties when performing a clustering task. They respond differently at the following four aspects, namely: dataset scale, dataset resolution, patch size, and feature number. And the results shows that the BoW model and the deep learning models are good at extracting local features and global features, respectively.


CHAPTER 3. INVESTIGATION OF UNSUPERVISED DEEP FEATURE LEARNING

3.2 Related Work

This thesis focuses on the BoW model and the deep learning models. It briefly reviews the relevant work. One of the popular frameworks is the bag-of-words (BoW) model [FPSV14]. The stages related to features are feature extraction, dictionary learning, feature coding, and feature pooling. Methods based on this model attempt to enhance one or more of the four stages. Wiliem et al. [WSW+14] adopt the BoW models with scale-invariant feature transform (SIFT) and Discrete Cosine Transformation descriptors. To use the Spatial Pyramid Matching (SPM) to form regional histograms, Kong et al. [KLC+14] adopt a discriminative dictionary learning strategy and then sparsely encode the cell images. Shen et al. [SLWY14] integrate the intensity order pooling based feature into the BoW model to achieve rotation invariance, and SPM is also utilised to incorporate spatial information of cell images. In 2006, Hinton [HOT06] proposed a method called greedy layer-wise pre-training which overcomes the difficulty of training a deep neural network. The success of this approach opens a new research area called deep learning and starts a so-called second neural network renaissance. Conventionally, deep neural networks contain two families. One family is deep feed-forward neural networks (FNN), and it mainly contains CNN and Deep Autoencoders. The second family is recurrent neural networks, which consists of Boltzmann Machines and their restricted variants. Among these neural networks, they are all unsupervised methods except CNN. This thesis starts with single-layer neural networks, Restricted Boltzmann Machine and Autoencoders and move to multi-layer neural networks, Deep Belief Network, and Stacked Denoising Autoencoders.

3.3 Models

3.3.1 Bag-of-Words Models

A typical BoW model consists of four basic modules: 1) Local feature extraction: a large number of local features within images are extracted from small overlapping patches of input images to form feature descriptors. The most popularly employed descriptor is scale-invariant feature transform (SIFT) [Low04] which can reflect the information of local intensity; 2) Codebook generation: a codebook is built from the descriptors of training images to represent the shared visual features among all the local features. Each element in the codebook is a “visual word”. A commonly utilised method to generate codebook is $k$-means; 3) Local feature coding: each local feature of an image is encoded to form a coding vector using the generated codebook. Recently, the commonly used soft-assignment coding [LWL11] assigns
weights of local features to some of the visual words in the codebook. And an image is represented by a collection of coding vectors; 4) Feature pooling: all the coding vectors of an image are pooled to compute an overall coding vector to represent the image. Commonly used pooling methods are sum-pooling (or average-pooling) and max-pooling.

The classical BoW model represents the input as an orderless collection of features, discarding the spatial layout of the features. It is less discriminative because it cannot capture the shape or location information of the input. Spatial pyramid matching (SPM) is proposed by Lazebnik et al. [LSP06] to alleviate this limitation for the BoW model. SPM partitions the input image into an increasingly finer level of spatial subregions, and each subregion is represented as a histogram pooled over all the coding vectors within the region. The final representation is the concatenation of all the subregions’ histogram. The framework of the BoW model with SPM is shown in Fig. 3.1.

3.3.2 Deep Learning Models

3.3.2.1 Denoising Auto-Encoders

The encoder is a function \( f \) that maps an observed unit \( x \in \mathcal{R}^m \) to hidden unit \( h \in \mathcal{R}^n \), where \( m \) and \( n \) are the dimensions of observed unit and hidden units respectively. It has the form \( h = f(x) = s_f(Wx + b_h) \), where \( s_f \) is a nonlinear activation function, typically a logistic \( \text{sigmoid}(z) = 1/(1 + e^{-z}) \). The decoder function \( s_g \) maps hidden representation \( h \) back to a reconstruction \( \tilde{x} = g(h) = s_g(W^\top h + c) \), where \( s_g \) is the decoder’s activation function. For parameters, \( W \) is a real valued weight matrix associated with the edges between units \( x \) and \( h \). \( b \) and \( c \) are bias vectors associated
with the observed unit and the hidden unit, respectively. Auto-encoder training aims to find parameters \( \{ W, b, c \} \) that minimise the reconstruction error on a training set of examples. For standard networks, the backpropagation algorithm is applied to compute the gradient of the error function with respect to the parameters. Then, the parameters are updated using stochastic gradient descent. A successful alternative form of autoencoder is obtained through the technique of denoising auto-encoders (DAE) by Vincent et al. [VLL+10]. It simply corrupts an input \( \mathbf{x} \) before feeding it into the autoencoder, and it aims to reconstruct a clean version via training. The structure of DAE is shown in Fig. 3.2.

### 3.3.2.2 Restricted Boltzmann Machines

![Figure 3.2: Denoising Autoencoder.](image1)

![Figure 3.3: Restricted Boltzmann Machine.](image2)

![Figure 3.4: Stacked Denoising Autoencoder.](image3)

![Figure 3.5: Deep Belief Network.](image4)

A Restricted Boltzmann Machine (RBM) [Smo87] is a Markov Random Field (MRF) associated with a bipartite undirected graph, shown in Fig. 3.3. It consists of \( m \) observed units \( \mathbf{X} = (X_1, ..., X_m) \) to represent observable data and \( n \) hidden units \( \mathbf{H} = (H_1, ..., H_n) \) to describe dependencies between observed variables. In binary RBMs, the random variables \( (\mathbf{X}, \mathbf{H}) \) take values \( (x, h) \in \{0, 1\}^{m+n} \) and the
joint probability distribution under the model is given by the Gibbs distribution
\[ p(x, h) = \frac{1}{Z} e^{-E(x, h)} \]
with the energy function
\[ E(x, h) = - \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} h_i x_j - \sum_{j=1}^{m} b_j x_j - \sum_{i=1}^{n} c_i h_i \]  
(3.1)
and the partition function is
\[ Z = \sum_{x, h} e^{-E(x, h)} \]
where \( w_{ij} \) is a real valued weight associated with the edge between units \( x_j \) and \( h_i \). \( b_j \) and \( c_i \) are real valued bias terms associated with the \( j \)th observed and the \( i \)th hidden variable, respectively. RBM can be trained using contrastive divergence approximation [MSCdF10]. Sparse RBM adds an additional penalty term that encourages the units to maintain a low average activation.

### 3.3.2.3 Stacked Denoising Auto-Encoders

Stacked Denoising Auto-Encoders (SDAE) [SOC+12], shown in Fig. 3.4, consists of denoising autoencoders stacked on top of each other, and is trained in a greedy layer-wise manner. When there are \( L-1 \) intermediate layers of hidden units in both of the encoder and decoder, the encoder becomes \( h = f(x) = f_{L-1} \circ f_{L-2} \circ ... \circ f_1(x) \). And the decoder becomes \( x = g(h) = g_1 \circ g_2 \circ ... \circ g_{L-1}(h) \), where \( f_l \) and \( g_l \) are the encoding and decoding nonlinear modules at the \( l \)-th layer. \( f_l \) and \( g_l \) are defined by
\[ f_l(\tilde{x}_l) = \phi_l \left( W_l^T \tilde{x}_{l-1} + b_l \right) \]
and
\[ g_l(\tilde{x}_l) = \varphi_l \left( U_l \tilde{x}_{l+1} + c_l \right) \]
where \( W_l, U_l, b_l \) and \( c_l \) are the parameters of the \( l \)th hidden module, and \( \phi_l \) and \( \varphi_l \) are component-wise non-linearities used in the encoder and decoder, respectively. The parameters of a deep autoencoder can be optimised by minimizing the difference between the original input \( x_n \) and the reconstructed input \( \tilde{x}_n \) for all \( N \) training samples. The difference may be measured by any suitable distance metric such as a squared Euclidean distance or a cross-entropy loss in the case of binary inputs.

### 3.3.2.4 Deep Belief Network

The deep belief network (DBN), shown in Fig. 3.5, was proposed by Hinton et al. [HOT06]. It is a hybrid model that has both directed and undirected edges. The top two hidden layers \( h^{L-1} \) and \( h^L \) are connected to each other (without any intra-layer edges) by undirected edges, while all subsequent pairs of layers below are connected by directed downward edges. Hence, the top two layers generatively model the prior distribution of \( h^{L-1} \), and all the other hidden layers below model the conditional distribution that generates the states of the units in the subsequent layer below.

The joint distribution \( P(h^L, h^{L-1}) \) of the top two layers is a RBM. DBN can be trained using a greedy layer-wise algorithm. For each layer, it can train a RBM to obtain the output of a lower layer as the input of a higher layer.
3.4 Experimental Study

3.4.1 HEp-2 Cell Datasets

3.4.1.1 ICPR 2014 Cell Dataset

The dataset of ICPR 2014 HEp-2 cell contains 13,596 training cell images, and test cell images are not published. All the cell images are cropped from 83 specimen images, and they have been manually segmented and annotated. Each cell image belongs to one of the six classes of staining patterns: centromere, golgi, homogeneous, nucleolar, nuclear membrane, and speckled (see examples in Fig. 3.6).

3.4.1.2 ICPR 2012 Cell Dataset

The ICPR 2012 cell dataset is smaller than ICPR 2014 Cell dataset, which consists of 1455 cell images extracted from 28 specimen images. The dataset has been partitioned into a training set (721 images) and test set (734 images). Each cell image also belongs to one of the six types of staining patterns: centromere, cytoplasmatic, homogeneous, nucleolar, fine speckled, and coarse speckled. Note that two of the six staining patterns are different from those of ICPR 2014 dataset (see examples in Fig. 3.6). In this experiment, the training set and the test set are merged for unsupervised feature learning.

![Examples of HEp-2 cell images from ICPR 2014 dataset (top row) and ICPR 2012 dataset (bottom row).](image)

Figure 3.6: Examples of HEp-2 cell images from ICPR 2014 dataset (top row) and ICPR 2012 dataset (bottom row).

3.4.2 Image Preprocessing

In order to enhance the contrast, each cell image is normalised by first subtracting the minimum intensity value of the image and then being divided by the difference
between the maximum intensity and the minimum intensity of the image. Furthermore, each image is resized to a fixed size of 64 × 64 or 32 × 32. And this experiment therefore obtains four datasets: dataset 2014_64, dataset 2014_32, dataset 2012_64 and dataset 2012_32.

3.4.3 Clustering

3.4.3.1 K-means

Feature vectors learnt by the BoW model or the deep learning models are clustered by \(k\)-means. In the following experiments, to enhance the reliability of the clustering results, feature vectors are clustered by 100 times under the same setting of plain \(k\)-means [VF08]. Also, this experiment builds a baseline by using the vectorised raw image as input, which corresponds to the case where no feature learning is conducted.

3.4.3.2 Evaluation Measures for Clustering

Two commonly used measurements, the accuracy (ACC) and the normalised mutual information metric (NMI), are utilised to evaluate the clustering performance [XLG03].

3.4.4 Results of the BoW Models

For the BoW model, this experiment utilises the classical 4-stage BoW pipeline: local feature extraction, codebook generation, local feature coding, and feature pooling. Popular SPM is optionally adopted to incorporate spatial information for the final image representation. More specifically, SIFT features are extracted from patches of each cell image. A codebook is generated by \(k\)-means. Local soft-assignment coding (LSC) is employed to encode the low-level SIFT descriptors. For the SPM, the partition of three levels is carried out at 1 × 1, 2 × 2 and 1 × 3 grids.

The pipeline of the BoW model has several hyper-parameters: 1) the patch size for feature extraction; 2) codebook size in codebook generation; 3) pooling method and 4) the optional use of SPM. The experiment first evaluates the effect of these hyper-parameters in dataset 2014_64. Patch size is chosen as the range of 9 × 9, 11 × 11, 13 × 13, and 15 × 15. Codebook size is selected from 1000, 2000, and 4000. And the pooling method can be max-pooling or sum-pooling. Other hyper-parameters are set empirically. For instance, this experiment sets the patch stride as 2 to balance the performance and the computation cost.

Fig. 3.7.(a) and Fig. 3.7.(b) clearly show the effect of path size. Among the chosen patch size, the best one is 9 × 9 in terms of both ACC and NMI. And the
performance will decrease if the patch size is increased. This can be understood from two aspects. One aspect is that the relatively small local receptive field can provide enough information on statistics for the locally hand-crafted descriptors. The other aspect is that a hand-crafted descriptor, like SIFT, does not seem to be good at extracting global features.

The use of SPM slightly harms the ACC by comparing the left part (with SPM) and the right part (without SPM) of Fig. 3.7(a). In fact, SPM is originally designed to build a hierarchical representation to improve the discriminative ability of features. However, it leads to the dimensionality curse by multiplying the number of features, especially when the codebook size is large and the data scale is limited (for instance, $8 \times 4000$ features, for dataset 2014_64 with 13596 images).

Also, Fig. 3.7(c) and Fig. 3.7(d) show the same phenomenon as that reported by [LW14], which illustrates that the max-pooling method can get much better performance than using sum-pooling (or average-pooling). This result can be explained as that average-pooling wipes off the diversity at local while max-pooling increases this diversity by only retaining the largest value.

The experiments also test the BoW models for the other three datasets. The results are shown in Table 3.1 and Table 3.2 (left part). Similarly, it can be found the negative effect of using SPM. Especially, for the dataset 2012_32 when using 4000 as the codebook size, the largest loss of ACC can be 0.055. And using a larger size of codebook can slightly increase the performance without SPM. Also, when reducing the image size from $64 \times 64$ pixels to $32 \times 32$ pixels, the performance of the BoW will substantially decrease, which illustrates that the BoW model needs to use images with sufficient resolutions.

### 3.4.5 Results of Deep Learning Models

Deep learning models can be trained from patches or directly from full images. To make a comparison with the BoW model, this experiment firstly trains the single-layer networks (RBM and DAE) using the patches. Then this experiment moves to the case of using full images by training a multi-layer networks (DBN and SDAE).

#### 3.4.5.1 Single-layer networks trained with patches

To train single-layer networks with patches, this experiment adopts the framework in [CNL11], which regards RBM and DAE as self-learned feature detectors. To get the feature vector of a whole image, the first step is to extract the $K$-dimensional feature vector from each patch using RBM and DAE. After that, using a max-pooling method over 4 quadrants of images is used to form the final $4K$-dimensional feature vector.
CHAPTER 3. INVESTIGATION OF UNSUPERVISED DEEP FEATURE LEARNING

(a) ACC of the BoW model with different patch size.

(b) NMI of the BoW model with different patch size.

(c) ACC of the BoW model with different pooling method.

(d) NMI of the BoW model with different pooling method.

Figure 3.7: The effect of patch size and pooling methods for different BoW models. ACC is the accuracy of the clustering and NMI, called normalised mutual information. The BoW models are evaluated by different size of codebook and optional use of SPM. For instance, 2000-NoSPM means the BoW model with a codebook of 2000 and without utilising SPM. For (a) and (b), this experiment uses max-pooling; For (c) and (d), this experiment fixes the patch size as 11.
CHAPTER 3. INVESTIGATION OF UNSUPERVISED DEEP FEATURE LEARNING

Figure 3.8: The effect of patch size and feature number for single-layer networks, compared with the BoW models. Patch size is selected from $9 \times 9$, $11 \times 11$, $13 \times 13$, and $15 \times 15$; Feature number is selected from 500, 1000, 2000, and 4000. Single-layer networks achieved better performance by learning more features. For (a) and (b), this experiment fixes feature number as 2000; For (c) and (d), this experiment fixes the patch size as 11.
To make a comparison with the BoW model, this experiment also evaluates the effect of two hyper-parameters: 1) the patch size; 2) the number of features or the number of the hidden units of the network. Patch size is also selected from $9 \times 9$, $11 \times 11$, $13 \times 13$, and $15 \times 15$ and $K$ is selected from 500, 1000, and 2000. Other hyper-parameters related to the training of networks are chosen using cross-validation. For instance, learning rate ($0.1$, $0.01$, and $0.001$), momentum ($0.1$, $0.3$, $0.5$, $0.7$, and $0.9$) and sparsity target ($0.1$, $0.01$, and $0.001$) are cross-validated.

The effect of patch size and the number of features are shown in Fig. 3.8.(c). Compared with the BoW model, which prefers to smaller patch size, single-layer networks does not have such an obvious trend. However, as is shown in Fig. 3.8.(d), the dimension of features, for which the BoW model is not very sensitive, has an obvious impact on the single-layer network. In other words, single-layer networks achieved better performance by learning more features.

### 3.4.5.2 Multi-layer Networks trained with Full Images

This experiment has trained multi-layer networks directly from full images instead of using the patches on all four datasets. Both SDAE and DBN are built from three to five layers and the number of units in each layer is fixed as 1000. Compared with the BoW model, the networks have a larger number of uncertain hyperparameters to set, and the training of network is known to be difficult [HOT06]. When utilising images directly (with a larger input size than image patch), the training becomes more difficult since the number of parameters is significantly increased.

The results of multi-layer networks trained with full images are shown in Table 3.1 and Table 3.2 (right part). Multi-layer networks obtain the best result ($0.571$ of ACC and $0.404$ of NMI) in dataset 2014, and the ACC value is even higher than that obtained by the BoW model. SDAE performs well on the relatively lower resolution dataset. For small-scale datasets, namely, dataset 2012, 2012, and 2012, the deep learning models achieve the same or even slightly worse results compared with the baseline, obtained by k-means clustering. This result shows the failure of training the deep learning models when the data samples are inadequate.

### 3.4.6 Discussion

This experiment evaluates clustering performance for the BoW model and deep learning models from four aspects, namely: dataset scale, dataset resolution, patch size, and feature number. Table 3.3 summaries the comparison of these two models in terms of the sensitivity or dependence of the above four aspects. This experiment finds that the BoW model and the deep learning models have some different or complementary properties in terms of these issues. For instance, to achieve a good
CHAPTER 3. INVESTIGATION OF UNSUPERVISED DEEP FEATURE LEARNING

performance, the deep learning models depend on a larger-scale dataset and prefer lower-resolution images while the BoW model prefer higher-resolution images. Also, the BoW model is sensitive to the patch size and is relatively stable with the feature number. Differently, the deep learning models are sensitive to feature number and are relatively stable for patch size when the single-layer networks are utilised.

Essentially, these complementary properties also reflect the complementary mechanisms of the BoW model and the deep learning models in handling features. This thesis tries to give an explanation based on the understanding of these two kinds of models. At the first place, it should be noticed that these two kinds of models utilises different features. For the first aspect on the scale of the dataset, when the scale is small, it will be hard for deep models to learn good features since the number of training data is limited. However, this seems have less impact on the BoW model since it utilises SIFT feature. For the second aspect on image resolution, when using images with higher resolution, the BoW model performance will be better since it improves the image patch quality, and hence get better performance. While for deep models, it will result in higher model complexity and in turn training difficulty. For the third aspect on patch size, as mentioned above, SIFT feature may prefer higher-quality patches while deep models extract global features which will less be affected by patch size. For the last aspect on feature number, higher feature dimension will also increase the deep models’ complexity and this may also incur the training difficulty. For the datasets used in this experiment, the inadequacy of data samples is the main reason why the deep learning models investigated in this experiment performs worse than the BoW model.

What’s more, the codebook generation procedure, by merging the similar features into clusters, can also be helpful for the BoW model to increase its discriminative ability. Contrarily, the deep learning models studied in this experiment do not have such procedure. These models are primitively designed as the generative models with the goal of reconstructing the original data. Hence, although these deep learning models can learn global features, there is still no guarantee for them to learn highly discriminative features. This should be another reason why the investigated deep learning models cannot perform better than the BoW model in most of the experiments. Certainly, the current study only utilises the cell images. Whether the identified characteristics could be observed in other image sceneries will be an interesting piece of work for future research.

This chapter has investigated unsupervised deep feature learning. Feature learning technique is quite important to various computer vision tasks as well as image retrieval since image retrieval performance relies on feature representation. Particularly, for image retrieval, unsupervised deep feature learning could bring great benefit. This is true since that there are no labels for image retrieval tasks. The
investigation of deep feature learning in this chapter serves the thesis as a basis and helps the subsequent chapters to develop new methods for image retrieval.

3.5 Conclusion

This chapter investigated the unsupervised feature leaning through the application of clustering the HEp-2 cell image. It compares two kinds of typical models: the BoW model and the deep learning models. The results show that the BoW model obtains better clustering performance on most datasets than the investigated deep learning models. This success can be attributed to the fitness of utilising the local hand-crafted features and the discriminative codebook generation in the BoW model. This chapter evaluates the BoW model and the deep learning models in four aspects and finds that they have complementarity properties. It seems that the BoW model and the deep learning models studied in this chapter are good at extracting local features and global features, respectively. This provides valuable experience for unsupervised learning tasks and network training in the subsequent image retrieval tasks to be conducted in this thesis.
Table 3.1: ACC for the BoW models and Deep Learning models. Using a plain $k$-means, this experiment can evaluate the features leant by these models. ACC is the accuracy of the clustering performance. The BoW models are evaluated by different size of codebook: 1000, 2000 and 4000 with optional use of SPM. Deep learning models, SDAE and DBN, are evaluated with 3, 4 and 5 layers with a fixed number of hidden units as 1000. Also, the baseline is built by using the vectorised raw image as input. The dataset, for instance, 2014_64 means the ICPR 2014 HEp-2 Cell dataset with each image resized as $64 \times 64$. The best performance is marked as bold. And the ± means the variations for multiple experiments.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Baseline</th>
<th>BoW with SPM</th>
<th>BoW without SPM</th>
<th>Deep Learning Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SDAE_3 SDAE_4 SDAE_5</td>
</tr>
<tr>
<td>2014_64</td>
<td></td>
<td>0.495</td>
<td>0.623 0.609 0.613</td>
<td>0.626 0.629 <strong>0.640</strong></td>
<td>0.529 0.52 0.536</td>
</tr>
<tr>
<td></td>
<td></td>
<td>±0.027</td>
<td>±0.023 ±0.029 ±0.04</td>
<td>±0.022 ±0.024 ±<strong>0.031</strong></td>
<td>±0.028 ±0.033 ±0.04</td>
</tr>
<tr>
<td>2014_32</td>
<td></td>
<td>0.494</td>
<td>0.514 0.519 0.509</td>
<td>0.52 0.527 0.511</td>
<td>0.564 0.563 <strong>0.571</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>±0.024</td>
<td>±0.03 ±0.037 ±0.025</td>
<td>±0.042 ±0.043 ±0.031</td>
<td>±0.046 ±0.049 ±<strong>0.049</strong></td>
</tr>
<tr>
<td>2012_64</td>
<td></td>
<td>0.413</td>
<td>0.599 0.593 0.592</td>
<td>0.595 0.604 <strong>0.622</strong></td>
<td>0.425 0.425 0.423</td>
</tr>
<tr>
<td></td>
<td></td>
<td>±0.012</td>
<td>±0.047 ±0.06 ±0.076</td>
<td>±0.033 ±0.039 ±<strong>0.048</strong></td>
<td>±0.028 ±0.026 ±0.026</td>
</tr>
<tr>
<td>2012_32</td>
<td></td>
<td>0.407</td>
<td>0.511 0.517 0.495</td>
<td>0.532 <strong>0.545</strong> 0.540</td>
<td>0.421 0.398 0.388</td>
</tr>
<tr>
<td></td>
<td></td>
<td>±0.016</td>
<td>±0.045 ±0.047 ±0.055</td>
<td>±0.045 ±<strong>0.05</strong> ±0.046</td>
<td>±0.031 ±0.027 ±0.025</td>
</tr>
</tbody>
</table>
Table 3.2: NMI for the BoW models and Deep Learning models. Using a plain $k$-means, this experiment can evaluate the features learnt by these models. NMI is another metric for measuring clustering performance, called normalized mutual information. The BoW models are evaluated by different size of codebook: 1000, 2000 and 4000 with optional use of SPM. Deep learning models, SDAE and DBN, are evaluated with 3, 4 and 5 layers with a fixed number of hidden units as 1000. Also, the baseline is built by using the vectorised raw image as input. The dataset, for instance, 2014\_64 means the ICPR 2014 HEp-2 Cell dataset with each image resized as $64 \times 64$. The best performance is marked as bold. And the $\pm$ means the variations for multiple experiments.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Baseline</th>
<th>BoW with SPM</th>
<th>BoW without SPM</th>
<th>SDAE_3</th>
<th>SDAE_4</th>
<th>SDAE_5</th>
<th>DBN_3</th>
<th>DBN_4</th>
<th>DBN_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014_64</td>
<td>0.37</td>
<td>0.489</td>
<td>0.525</td>
<td>0.477</td>
<td>0.513</td>
<td>0.378</td>
<td>0.393</td>
<td>0.379</td>
<td>0.305</td>
<td>0.359</td>
</tr>
<tr>
<td></td>
<td>±0.011</td>
<td>±0.019</td>
<td>±0.013</td>
<td>±0.027</td>
<td>±0.012</td>
<td>±0.008</td>
<td>±0.013</td>
<td>±0.012</td>
<td>±0.012</td>
<td>±0.009</td>
</tr>
<tr>
<td>2014_32</td>
<td>0.37</td>
<td>0.421</td>
<td>0.425</td>
<td>0.413</td>
<td>0.404</td>
<td>0.403</td>
<td>0.402</td>
<td>0.35</td>
<td>0.357</td>
<td>0.357</td>
</tr>
<tr>
<td></td>
<td>±0.012</td>
<td>±0.025</td>
<td>±0.025</td>
<td>±0.025</td>
<td>±0.022</td>
<td>±0.023</td>
<td>±0.022</td>
<td>±0.003</td>
<td>±0.002</td>
<td>±0.003</td>
</tr>
<tr>
<td>2012_64</td>
<td>0.296</td>
<td>0.523</td>
<td>0.498</td>
<td>0.509</td>
<td>0.527</td>
<td>0.273</td>
<td>0.281</td>
<td>0.284</td>
<td>0.294</td>
<td>0.299</td>
</tr>
<tr>
<td></td>
<td>±0.007</td>
<td>±0.042</td>
<td>±0.045</td>
<td>±0.043</td>
<td>±0.038</td>
<td>±0.014</td>
<td>±0.013</td>
<td>±0.01</td>
<td>±0.004</td>
<td>±0.004</td>
</tr>
<tr>
<td>2012_32</td>
<td>0.295</td>
<td>0.374</td>
<td>0.402</td>
<td>0.426</td>
<td>0.416</td>
<td>0.291</td>
<td>0.275</td>
<td>0.274</td>
<td>0.296</td>
<td>0.297</td>
</tr>
<tr>
<td></td>
<td>±0.006</td>
<td>±0.027</td>
<td>±0.028</td>
<td>±0.038</td>
<td>±0.013</td>
<td>±0.01</td>
<td>±0.007</td>
<td>±0.003</td>
<td>±0.004</td>
<td>±0.003</td>
</tr>
</tbody>
</table>
Table 3.3: Comparison of the BoW model and the deep learning models. They have complementary properties on these four aspects: dataset scale, image resolution, patch size, and feature number.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>BoW Models</th>
<th>Deep Learning Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset scale</td>
<td>Less dependent</td>
<td>Dependent</td>
</tr>
<tr>
<td>Image resolution</td>
<td>Positively-related</td>
<td>Negatively-related</td>
</tr>
<tr>
<td>Patch size</td>
<td>Sensitive</td>
<td>Less sensitive</td>
</tr>
<tr>
<td>Feature number</td>
<td>Less sensitive</td>
<td>Sensitive</td>
</tr>
</tbody>
</table>
Chapter 4

Semi-supervised Weight Learning for the Spatial Search Method

As the state-of-the-art ConvNet-based image retrieval method, spatial search [RSMC15] has shown excellent retrieval performance and outperformed other competitors. A key component of this method is a weighted combination of distances evaluated at different regions of a query image. However, these weights are currently manually tuned, by a trial-and-error based exhaustive search. This not only incurs a lengthy parameter tuning process but is also hard to guarantee the optimality of the tuned weights. Moreover, these weights may not be generally applied when the nature of image dataset changes. To improve this situation, this thesis proposes to automatically learn the combination weights based on retrieval ground truth. Specifically, this thesis develops a method, called semi-supervised weight learning (SWL), based on the framework of distance metric learning. In addition to generating triplet constraints with retrieval ground truth, this thesis leverages unlabelled images to generate numerous unsupervised constraints to stabilise the learning process and improve learning efficiency. By linking with the latest primal solver of linear support vector machines, an efficient algorithm is put forward to solve the resulting large-scale optimisation problem. Experimental results on three benchmark datasets and a newly collected archival photo dataset to demonstrate the effectiveness of the proposed weight learning approach. It achieves comparable or better retrieval performance than the manual tuning approach, especially on the new archival photo dataset. Then, this thesis extends the SWL method with kernel instead of Euclidean distance.

4.1 Introduction

Deep convolutional neural networks (ConvNet) based image retrieval has become an attractive research area in recent years. Researchers in this area have made great efforts in designing effective image representations with the features extracted from pre-trained ConvNet. The existing methods can roughly be categorised into two groups. The first group methods focus on generating global features from different layers [BSCL14] [NYD15] or with different pooling strategies [BL15] [KMO16]. The second group focuses on local features. Methods with multi-region features followed by different aggregating schemes [GWGL14] [MB15] [TSJ15] or matching
strategies [RSMC15] have been proposed. In general, the region-based methods, which are encoded with more local appearance information of an image, usually outperform the global-based ones.

One typical region-based method is the spatial search developed in [RASC14b] [RSMC15]. Instead of comparing images as a whole via aggregated local features, this method compares individual regions cropped from a query and a gallery image to better handle background clutter and spatial variance of objects. Specifically, for each region of a query, its best match among the regions of a gallery image is identified based on the distance of local features. After that, a combination of these distances is deemed the final distance of the two images for retrieval. Although storing and cross-matching all the regions incurs a relatively high computational cost, the performance of spatial search has greatly surpassed other competitors. As one of the state-of-the-art methods, spatial search has recently evolved from adopting an unweighted average combination [RASC14b] to utilising a weighted combination [RSMC15].

A major issue with existing spatial search method lies at that it adopts a trial-and-error based approach to exhaustively search for the weights for regions of various sizes and locations. Although promising performance is reported in [RSMC15] via this approach, it incurs a lengthy parameter tuning process and requires prior knowledge and experience. Furthermore, this tuning process may have to be repeatedly performed for new retrieval tasks, considering that the tuned weights do not necessarily generalise to image datasets of a different nature.

Note that in order to evaluate the goodness of various sets of weights to select the best one, some retrieval ground truth is required by the trial-and-error approach. To address the aforementioned issue, this thesis proposes to automatically learn the combination weights by taking advantage of the access to retrieval ground truth. The goal is to learn a set of weights based on which the resulting distance between a query and a gallery image can best agree with their semantic similarity. Considering that retrieval is a ranking problem and that it is convenient to define the relationship in the way of “image \( A \) is more similar than image \( B \) to a query \( q \)”, this thesis builds work upon the framework of distance metric learning, with a set of triplet constraints. A triplet contains three elements: query, positive sample and negative sample, like \((q, A, B)\). And triplet constraints are utilised to learn better distance metrics making similar images closer to the query image than the dissimilar images. Nevertheless, the amount of retrieval ground truth (i.e., the number of queries for which true similar and not-similar gallery images are labelled) is usually limited, because collecting ground truth is time-consuming and labour-intensive. This could make the learning process overfit training data and fail to generalise to unseen queries. To mitigate this issue, this thesis performs weight learning in a
semi-supervised manner to incorporate the unlabelled (gallery) images in a dataset. Specifically, besides the triplet constraints generated with retrieval ground truth, this thesis generates unsupervised triplet constraints by randomly selecting gallery images as queries and determining the relationship \((q,A,B)\) with the spatial search method employing an (default) unweighted combination. In doing so, this thesis obtains a large number of unsupervised triplet constraints to use as a sort of regularisation to stabilise the learning process. The proposed semi-supervised weight learning (SWL) method boils down to a quadratic programming problem with a large number of linear constraints. Identifying its close relationship to the primal problem of linear support vector machines (SVM), this thesis adapts its latest solver developed in [NHWH14], which is based on augmented Lagrange multipliers (ALM) method, to this problem. This brings a highly efficient algorithm to handle a large number of triplet constraints in the optimisation.

The contributions of this chapter are summarised as follows. 1) This thesis develops a method to automatically learn the combination weights in spatial search [RASC14b] [RSMC15], a state-of-the-art ConvNet-based image retrieval method; 2) This thesis formulates the optimisation problem under the framework of distance metric learning and exploit gallery images to generate unsupervised constraints to help learning. In addition, an efficient solver is developed for the resulting optimisation program; 3) The proposed method is evaluated on three benchmark datasets in image retrieval and a new archival photo dataset collected by this thesis. The experimental result shows the effectiveness and advantage of this automatic weight learning approach.

4.2 Related Work

This thesis reviews ConvNet-based image retrieval from two perspectives: global features and regional features.

The early work, Neural Codes [BSCL14], studies the global feature with the fully connected layer of ConvNet, followed by principal component analysis (PCA) dimension reduction. Later on, research focus shifts from the features of fully connected layers to the feature maps at convolutional layers, and various pooling strategies are also investigated to create better image representations. Typical methods in this line include sum-pooled convolutional features (SPoC) [BL15] and features derived after cross-dimensional weighting and pooling (CroW) [KMO16]. Beyond manipulating features extracted from pre-trained models, fine-tuning ConvNet for better global features are also investigated, by utilising extra labelled data [BSCL14, YMK11] or additional information [WSL+14, AGT+15, WWH+14, RTC16]. Methods based on global features without fine-tuning are generally less competitive than those based
Research works based on regional or local features try to explore the ConvNet features as local descriptors, with different aggregating schemes [NYD15, GWGL14, MB15, TSJ15] or matching strategies [RSMC15]. One of the aggregation based methods are multi-scale orderless pooling (MOP) [GWGL14]. It has been introduced to achieve more invariant features by removing spatial information with VLAD [JPD+12]. Another work in this line [NYD15] exploits convolutional layer features instead of fully connected layer features. Work in [MB15] adopts max-pooling on features of detected regions along each channel and claims that it is more orderless than MOP since concatenating features by VLAD still preserves some spatial information. Different from [MB15], R-MAC [TSJ15] considers a set of regions in multiple scales and aggregates them into a final representation with max-pooling. The common property of aggregation based methods is to build invariant and compact features by reducing spatial information.

Spatial search method [RSMC15] works with regional features. It firstly extracts features from regions of various sizes and locations across each image. After that, it measures the distance of two images by cross-matching their regional features, instead of aggregating them. Although incurring extra computation than global feature based methods, spatial search [RSMC15] demonstrates superior retrieval performance on multiple benchmarks datasets. For this method, choosing appropriate weights to combine the distances evaluated at various regions of a query is important. Current spatial search method selects the weights by conducting an exhaustive search with the help of prior knowledge. Since it assumes the availability of retrieval ground truth, this thesis can directly learn these weights through optimising a retrieval-performance-related criterion with the ground truth.

### 4.3 Proposed Method: SWL

This section describes the proposed semi-supervised weight learning (SWL) method. A flowchart is provided in Fig. 4.1 to illustrate this method.

#### 4.3.1 Spatial Search Method

Given an image, spatial search firstly crops the largest square patch centred in the image, and further crops multiple square patches from this centred patch at various scales and locations. Let $a$ be the length of the centred patch and $L$ be the number of different scales. The length of different square patches is expressed as

$$a_l = \frac{2a}{l+1}, \quad l = 1, ..., L. \quad (4.1)$$
Figure 4.1: Flowchart of the proposed SWL method. It begins with an initial retrieval by using the spatial search with an unweighted average combination. Based on the retrieval ground truth and this initial retrieval result, supervised and unsupervised triplet constraints are generated and fed into the solver to learn the optimal combination weights. After that, the learned weights are integrated into the spatial search method for any future retrieval. Note that the weight learning only needs to be performed once for a given image dataset.
At each scale, $l^2$ square patches are cropped at the centre

$$
\left( \frac{a_l}{2} + (i-1)b, \frac{a_l}{2} + (j-1)b \right), \; i, j = 1, 2, \cdots, l
$$

(4.2)

where $b = \frac{a-a_l}{l-1}$ ($l \neq 1$). Also, two extra patches are added by rotating the previous largest square patch by $90^\circ$ and $-90^\circ$, respectively. In total, there are $m = 2 + \sum_{l=1}^{L} l^2$ square patches. Let \( \{f_k\}_{k=1}^{m} \) denote a set of ConvNet features extracted from the $m$ patches.

Given a query image $I^q$ and a gallery image $I^g$ and their feature sets \( \{f^q_k\} \) and \( \{f^g_k\} \). The distance between a patch $I^q_k$ extracted from the query and the whole gallery image $I^g$ is defined as

$$
d^*(I^q_k, I^g) = \min_{1 \leq j \leq m} d(f^q_k, f^g_j), \tag{4.3}
$$

where $d(\cdot, \cdot)$ is a predefined distance measure and a Euclidean distance is used in [RSMC15]. The final distance between the two images is then defined as a weighted sum of $d^*(I^q_k, I^g)$, with the weights denoted by $w = [w_1, w_2, \ldots, w_m]^\top$.

$$
D(I^q, I^g) = \sum_{k=1}^{m} w_k \cdot d^*(I^q_k, I^g) \tag{4.4}
$$

### 4.3.2 Loss Function of the Proposed Method

Following the framework of distance metric learning, given a triplet of query image $I^q$, positive (similar) image $I^+$ and negative (not similar) image $I^-$, this thesis uses the inequality $D(I^q, I^-) - D(I^q, I^+) \geq 1$ to reflect their relationship. In this way, this thesis can define a hinge loss function as

$$
\ell(I^q, I^+, I^-) = (1 - D(I^q, I^-) + D(I^q, I^+))^p \tag{4.5}
$$

where the operator $(x)^+_p$ stands for $\max(x, 0)$ and $p$ is the power with $p = 1, 2$. Combining with Eq.(4.4), Eq.(4.5) can be rewritten as

$$
\ell(I^q, I^+, I^-) = (1 - w^\top x)^p_+ \tag{4.6}
$$

where $x$ is a column vector of $[x_1, x_2, \ldots, x_m]^\top$, $x_k$ is

$$
x_k = d^*(I^q_k, I^-) - d^*(I^q_k, I^+), \; k = 1, 2, \ldots, m,
$$

and $I^q_k$ is the $k$th patch cropped from image $I^q$. Given a triplet, $x$ can be pre-computed and it only needs to be calculated once. Also, once computed, it remains
CHAPTER 4. SEMI-SUPERVISED WEIGHT LEARNING METHOD

a constant for the following optimisation problem.

4.3.3 Triplet Generation

Generating supervised triplets. Given a query image \( q \) with retrieval groundtruth, this thesis firstly forms a set \( \Omega_p \) that contains all the positive images with respect to this query. To collect hard negative images, this thesis performs the spatial search method with an unweighted average combination (i.e., \( w = [\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}]^\top \)). An initial ranking list of all the gallery images is then obtained. This thesis uses the top-ranked negative images to form another set \( \Omega_n \). Without loss of generality, this thesis keeps \(|\Omega_p| + |\Omega_n| = K\), where \(|\cdot|\) denotes the cardinality of a set. In this way, this thesis can pick one image \( A \) from \( \Omega_p \) and another image \( B \) from \( \Omega_n \) to form a triplet \((q, A, B)\). This is repeated by exhaustively pairing all the positive and negative images in the two sets.

Generating unsupervised triplets. The procedure is largely similar to the above. The main difference is that this thesis now randomly select a gallery image as a query \( q \) (therefore no retrieval groundtruth). By using the same spatial search method, an initial ranking list of all the gallery images is obtained, and this thesis forms a set \( \Omega \) with the top \( K \) images. Noting that there is no retrieval groundtruth for this query, this thesis exhaustively selects each pair of images \((A, B)\) from \( \Omega \) to form an unsupervised triplet \((q, A, B)\). The one ranked higher (lower) is used as the positive (negative) image in this triplet. As previously mentioned, generating a sufficient number of unsupervised triplets is important for avoiding overfitting and achieving effective learning.

4.3.4 Optimisation Problem for Learning Weights

By minimising the triplet-based loss function and the \( L_2 \) norm of the weight vector, the following optimisation problem is designed for weight learning.

\[
\min_w \frac{1}{2} w^\top w + \sum_{i=1}^n C_i \left(1 - w^\top x_i\right)_+^p,
\]

where \( n \) is the total number of triplets. Also, this thesis emphasises the loss from supervised triplets and unsupervised triplets differently by defining the regularisation parameter \( C_i \) as

\[
C_i = \begin{cases} 
C & \text{if } x_i \in \text{supervised triplets}; \\
C/\gamma & \text{if } x_i \in \text{unsupervised triplets}.
\end{cases}
\]

The optimisation problem shares a similar form with the primal SVM problem,
with the difference that it does not contain the class label variable \( y \) and the bias parameter \( b \) as SVM. Many efficient solvers have been developed for large-scale primal SVM problem in the literature [NHWH14, CHL08, LWK07]. One of the state-of-the-art solvers [NHWH14] is designed by using the augmented Lagrange multipliers (ALM) method [GR12]. Based on that solver, this thesis develops an efficient algorithm for the optimisation problem as follows. Firstly, an auxiliary variable \( e_i = 1 - w^\top x_i \) is introduced, and Eq.(4.8) becomes

\[
\min_{w,e} \frac{1}{2} w^\top w + \sum_{i=1}^{n} C_i (e_i)_+^p.
\]  

(4.10)

Eq.(4.10) can be written as an unconstrained optimisation problem via its Lagrangian function

\[
\min_{w,e} \frac{1}{2} w^\top w + \sum_{i=1}^{n} C_i (e_i)_+^p + \lambda^\top (X^\top w - 1 + e)
\]  

(4.11)

where \( X_{m\times n} = [x_1, x_2, ..., x_n], 1_{n\times 1} = [1, 1, ..., 1]^\top \) and \( e_{n\times 1} = [e_1, e_2, ..., e_n]^\top \). With the ALM method, another term \( \frac{\mu}{2} \| X^\top w - 1 + e \|^2 \) is appended, and the optimisation becomes

\[
\min_{w,e,\lambda} \frac{1}{2} w^\top w + \sum_{i=1}^{n} C_i (e_i)_+^p + \frac{\mu}{2} \| X^\top w - 1 + e + \mu^{-1} \lambda \|^2,
\]  

(4.12)

where \( \mu \) is constant within each iteration and will be gradually increased with iterations according to a given sequence.

For solving Eq.(4.12), \( e \) and \( w \) are updated alternately. When \( w \) is fixed, it can be decomposed into \( n \) independent single-variable minimisation problems

\[
e_i = \arg \min_{e_i} \frac{C_i}{\mu} (e_i)_+^p + \frac{1}{2} (e_i - t_i)^2
\]  

(4.13)

where \( t_i = 1 - w^\top x_i - \frac{\lambda_i}{\mu} \) and \( \lambda_i \) is the \( i \)th element of \( \lambda \). The optimal \( e_i \) can be analytically obtained for \( p = 1, 2 \). When \( e \) is fixed, \( w \) is updated by solving an \( L_2 \)-norm regularised least square regression problem for one step only,

\[
\min_w \mu^{-1} w^\top w + \| X^\top w - z \|^2,
\]  

(4.14)

where \( z = 1 - e - \mu^{-1} \lambda \). In addition, at each iteration \( \lambda \) is updated by

\[
\lambda^{(j)} = \lambda^{(j-1)} + \mu^{(j)} (X^\top w - 1 + e)
\]  

(4.15)

where \( \lambda^{(j)} \) is \( \lambda \) at the \( j \)-th iteration.

This thesis summarises the algorithm in Algorithm 1. Its complexity is \( O(Tnm) \), where \( T \) is the maximum number of iterations, \( n \) is the number of triplet constraints
and $m$ is the dimensionality of variable $w$. In Algorithm 1, $\epsilon$ is a small value to control termination. More details such as convergence analysis can be found in [NHWH14].

**Algorithm 1** Solution to the proposed SWL method

1. **Input:** $T, X, C, \left\{ \mu_j \right\}_{j=1}^{T}$
2. Initialise $w = 1, \lambda = 1, J = 1$;
3. **repeat**
4. Update $e$ by solving Eq.(4.13);
5. Update $w$ by solving Eq.(4.14);
6. Update $\lambda$ with Eq.(4.15);
7. $j = j + 1$;
8. **until** $\| \nabla \text{obj}(w) \| \leq \epsilon$ or $\| \nabla w \| \leq \epsilon$ or $j > T$.

Algorithm 1 is tested with a large number of triplet constraints and it is of high efficiency. For example, with the MATLAB code running on a computer with one Intel Core i5-3550 CPU @ 3.30GHz, it only takes around two minutes with 3 GB memory to solve the proposed optimisation problem in 200 iterations, for 5 million constraints with 32 weights. The speed could be further accelerated by optimised implementation.

### 4.4 Experimental Study

#### 4.4.1 Datasets

This thesis conducts experiments on four datasets including three benchmark datasets for image retrieval, **Oxford5k** [PCI+07a], **Paris6k** [HLC08] and **Holiday** [JDS08], and a new dataset **NAA29k** collected by this thesis. Table 4.1 shows their details.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total</th>
<th>Gallery</th>
<th>Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>image</td>
<td>images</td>
<td>Full</td>
</tr>
<tr>
<td>Oxford5k</td>
<td>5063</td>
<td>5008</td>
<td>55</td>
</tr>
<tr>
<td>Paris6k</td>
<td>6392</td>
<td>6338</td>
<td>55</td>
</tr>
<tr>
<td>Holiday</td>
<td>1491</td>
<td>941</td>
<td>500</td>
</tr>
<tr>
<td>NAA29k</td>
<td>28912</td>
<td>28312</td>
<td>600</td>
</tr>
</tbody>
</table>

The dataset called ‘NAA29k’ is part of the database of National Archives of Australia (NAA). Images of this database are historically valuable and naturally
Figure 4.2: Example images in the new dataset NAA29k.
collected around Australia, illustrating various contexts. The database contains images of landscapes, portraits, objects, buildings, aerial photographs, group shots, factories, signs and posters. They are in wide styles: black and white, sepia-tone, and some colour images. The total number of images in the database is around 350,000 and this thesis randomly selects 28,912 images to form a smaller dataset. From these 29k images, this thesis randomly picks 1000 as queries. To generate retrieval ground truth, this thesis first performs retrieval with global CovNet features and only keep the top 500 retrieved results for each query. Then this thesis manually labels similar images from these retrieved images. By removing those ‘easiest’ 400 queries, only 600 queries are retained to form the retrieval ground truth for NAA29k. Containing diverse and naturally collected images, this dataset is much more challenging than other benchmark image retrieval datasets in this experiment. Fig. 4.2 shows some example images in the NAA29k dataset.

### 4.4.2 Experimental Setting

For all experiments, this thesis adopts ImageNet-Vgg16 pre-trained model in [SZ14] when not indicated otherwise. The number of spatial scales $L$ is set as 4. For each image, this thesis keeps the aspect ratio and resize it to ensure the length of its longer side to be 576. This thesis crops 32 patches for each image by following [RSMC15]. For each patch, this thesis uses MatConvNet [VL15] to extract feature of $Conv5_2$ layer, followed by 2-by-2 grid max-pooling. After extracting features for all patches, this thesis follows the pipeline in [RSMC15] to post-process the features. In specific, all features are applied $L_2$ normalisation, PCA whitening, dimensionality reduction to 512 dimensions, and $L_2$ re-normalisation. The weights to learn are initialised as uniform weights. For triplets generation, $K$ is set as 40. For the solver, $\mu$ is set as 1.05, $T$ is set as 200 and $\epsilon$ is $10^{-8}$. The hyper-parameters $C$ and $\gamma$ are empirically set as $2^3$ and $2^8$. In each experiment, this thesis randomly halves the queries as training and test, as shown in Table 4.1. Each experiment is repeated 20 times and the averaged mean-Average Precision (mAP) is reported. All results are obtained on test queries by default.

### 4.4.3 Experimental Results

This thesis reports results with two methods. One is spatial search with an un-weighted average combination [RASC14b] (i.e., uniform weights) and the other method is spatial search with manually tuned weights [RSMC15]. The first method is viewed as a baseline for comparison. For the second method, this thesis uses the tuned weight provided in [RSMC15]. Since NAA29k is a new dataset, this thesis adopts the combination weight of Holidays in [RSMC15] for it, considering both
datasets contain natural scene images. In addition, to demonstrate the superiority of the spatial search method to other comparable methods in the literature, results from other recent related work are also included as reference. These results are originally reported in Neural Codes [BSCL14], Spoc [BL15], Crow [KMO16], MOP [GWGL14], Exploiting [NYD15], Object Level [MB15], R-MAC [TSJ15] and Spatial Search [RSMC15].

Table 4.2: Comparison on three datasets. The results are originally reported in recent work on full queries, please see the full set in the Table 4.1. As shown, spatial search outperforms other competitors.

<table>
<thead>
<tr>
<th>mAP</th>
<th>Oxford5k</th>
<th>Paris6k</th>
<th>Holiday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Pooling</td>
<td>0.464</td>
<td>0.312</td>
<td>0.638</td>
</tr>
<tr>
<td>Sum Pooling</td>
<td>0.606</td>
<td>0.468</td>
<td>0.809</td>
</tr>
<tr>
<td>Neural Codes [BSCL14]</td>
<td>0.435</td>
<td>-</td>
<td>0.749</td>
</tr>
<tr>
<td>Spoc [BL15]</td>
<td>0.589</td>
<td>-</td>
<td>0.802</td>
</tr>
<tr>
<td>Crow [KMO16]</td>
<td>0.796</td>
<td>0.682</td>
<td>0.849</td>
</tr>
<tr>
<td>MOP [GWGL14]</td>
<td>-</td>
<td>-</td>
<td>0.784</td>
</tr>
<tr>
<td>Exploiting [NYD15]</td>
<td>0.649</td>
<td>0.694</td>
<td>0.838</td>
</tr>
<tr>
<td>Object Level [MB15]</td>
<td>0.607</td>
<td>0.662</td>
<td>0.885</td>
</tr>
<tr>
<td>R-MAC [TSJ15]</td>
<td>0.668</td>
<td>0.830</td>
<td>-</td>
</tr>
<tr>
<td>Spatial Search [RSMC15]</td>
<td><strong>0.844</strong></td>
<td><strong>0.853</strong></td>
<td><strong>0.897</strong></td>
</tr>
</tbody>
</table>

Table 4.3: Comparison on four datasets. The results are conducted in this thesis on test queries to compare spatial search methods before and after the proposed weight learning. The proposed method achieves comparable or better performance than the baseline and the spatial search. The baseline is the spatial search method with an unweighted average combination.

<table>
<thead>
<tr>
<th></th>
<th>Oxford5k</th>
<th>Paris6k</th>
<th>Holiday</th>
<th>NAA29k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.823</td>
<td>0.787</td>
<td>0.900</td>
<td>0.629</td>
</tr>
<tr>
<td>Spatial Search [RSMC15]</td>
<td>0.839</td>
<td><strong>0.803</strong></td>
<td>0.898</td>
<td>0.623</td>
</tr>
<tr>
<td>Proposed SWL ($p = 1$)</td>
<td>0.841</td>
<td>0.792</td>
<td><strong>0.904</strong></td>
<td>0.664</td>
</tr>
<tr>
<td>Proposed SWL ($p = 2$)</td>
<td><strong>0.843</strong></td>
<td>0.801</td>
<td><strong>0.904</strong></td>
<td><strong>0.665</strong></td>
</tr>
</tbody>
</table>

Comparison with spatial search. This thesis builds two settings for comparison. One setting is to use all queries (full setting) and the other is to only use test queries (test setting). The full setting helps the thesis compare methods under unsupervised manner and the test setting allows the methods to perform semi-supervised learning.
Firstly, this thesis compares the spatial search with some other competitive methods on As shown in Table 4.2 (on the full set of queries), spatial search outperforms other competitors on three benchmark datasets, demonstrating its superiority. Also, as seen in Table 4.3 (on the test set of queries), the proposed SWL method achieves comparable or better performance than the spatial search in [RSMC15], at the test set of queries (Note that to be fair, the training set of queries are excluded because they have been used for training). The improvement is most significant at the new dataset NAA29k, where SWL outperforms spatial search [RSMC15] by four percentage points. This verifies the effectiveness of the proposed weight learning method and its advantage. Also, this thesis reports the results of SWL with different $p$-norm. It achieves the best performance with $p = 1$ on Holiday and with $p = 2$ on Oxford5k and NAA29k, although the difference is not significant. Fig. 4.6 shows some retrieval results on NAA29k dataset. The proposed method performs the best among the three methods.

Effect of hyper-parameters. The proposed method has two important hyper-parameters $C$ and $\gamma$. $C$ balances the term of weight norm and the hinge loss term, while $\gamma$ balances the supervised and unsupervised triplet constraints. Fig. 4.3 shows the effect of $C$ and $\gamma$ on all the datasets for both training and test. For each dataset, training and test graphs have similar profiles and almost peak at the same locations. These curves indicate that SWL enjoys a good generalisation ability and the learned weights perform well for new queries. Also, as expected, this thesis found that $\gamma$ should be assigned an appropriate value, neither too large or too small. If using a too large value, it will take off the regularisation ability and if using a too small value, it will suppress the retrieval ground truth information.

Purely supervised learning. This thesis has also extent the semi-supervised learning to a extreme case: purely supervised learning by removing the unsupervised triplet constraints in the formulation, frequently leads to over-fitting. As shown in Fig. 4.4, the curve drops sharply if emphasising the supervised triplet constraints too much by setting $C$ to an overlarge value. This, in turn, indicates that the unsupervised triplet constraints generated from unlabelled images do play an important role in maintaining good learning performance.

With different pre-trained model. This thesis also explores the effects of various pre-trained models. It conducts experiments with another pre-trained model, called Place-Vgg19 model [WGHQ15] on Holiday and NAA29k datasets which are scene related. This model is pre-trained on scene-level datasets of Place205 [ZLX +14]. Table 4.4 reports the results. This thesis finds that adopting appropriate features is important for image retrieval. With Place-Vgg19 features, all three methods in comparison achieve better retrieval performance on NAA29k. Particularly, the proposed SWL achieves 0.674, which is one percentage point higher than the result
Figure 4.3: Effect of hyper-parameters $\gamma$ and $C$ in the proposed learning method. Each plot shows the mAP of the proposed semi-supervised weight learning method on the training or test query sets (power $p = 1$, and ImageNet-Vgg16 features are used). Both $\gamma$ and $C$ are in the logarithm of 2. The plane presents the baseline, which is the spatial search method with an unweighted average combination.
CHAPTER 4. SEMI-SUPERVISED WEIGHT LEARNING METHOD

Figure 4.4: Results of purely supervised weight learning on the four datasets with ImageNet-Vgg16 features. It is conducted by removing the unsupervised triplet constraints in the optimisation problem. The baseline is the spatial search method with an unweighted average combination. \( C \) is in the logarithm of 2.

Figure 4.5: Results of purely supervised learning on NAA29k and Holiday with Place-Vgg19 features. It is conducted by removing the unsupervised triplet constraints in the optimisation problem. The baseline is the spatial search method with an unweighted average combination. \( C \) is in the logarithm of 2.
Table 4.4: Result on Holiday and NAA29k with different pre-trained models and different initial rank lists. The baselines of ImageNet-Vgg16 and Place-Vgg19 are obtained by spatial search with unweighted weights, while the baseline of Place-Vgg19* is obtained by using global feature based retrieval.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Holiday</th>
<th>NAA29k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ImageNet-Vgg16</td>
<td>Place-Vgg19</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.900</td>
<td>0.889</td>
</tr>
<tr>
<td>Spatial Search [RSMC15]</td>
<td>0.898</td>
<td><strong>0.892</strong></td>
</tr>
<tr>
<td>SWL (proposed)</td>
<td><strong>0.904</strong></td>
<td><strong>0.892</strong></td>
</tr>
</tbody>
</table>

using ImageNet-Vgg16 features (0.664). This observation is consistent with the fact that the new dataset NAA29k contains many scene-level images. Also, this thesis conducts purely supervised learning. Again, this thesis finds that purely supervised learning leads to over-fitting, as shown in Fig. 4.5a and Fig. 4.5b.

With different initial rank list. To extend the flexibility and efficiency of the proposed method, this thesis conducts experiments by generating initial rank lists with some simple methods rather than the default approach. By default, a uniformly weighted spatial search is applied to create the initial list. One drawback of this way could be that it will take too much time to perform retrieval on large datasets. A possible option to quickly generate an initial ranking list is to directly use the global features for retrieval, while this might affect the learning performance of the proposed SWL since the overall quality of triplet constraints could decrease. To investigate this option, this thesis conducts experiments and report result in the last column of Table 4.4. It shows that SWL can still achieve competitive performance with an initial ranking list obtained with global features (0.889 vs. 0.892 on Holiday, and 0.671 vs. 0.674 on NAA29k).

4.5 Extension to Kernel SWL

The original SWL method uses Euclidean distance when calculating the distance between two images. An extension is to apply kernel instead of Euclidean distance. Following this idea, this following part proposes a kernel-based SWL.
Figure 4.6: Example retrieval results on NAA29k dataset. Four queries are selected and shown at the top of each block. In each block, there are three columns of results: the left is for the baseline, the middle is for the spatial search method in [RSMC15], and the right is for the proposed SWL. Correctly retrieved images are marked with green rectangles while the wrong ones are marked in red. The proposed method shows the overall best performance.
4.5.1 Kernel SWL

Given a gallery image $I^g$ and a query image $I^q$, $I^q_k$ is the $k$th patch cropped from $I^q$. Considering using the RBF kernel as an example, the similarity between the patch $I^q_k$ and the gallery image $I^g$ can be defined as:

$$s^*(I^q_k, I^g) = e^{-\frac{(d^*(I^q_k, I^g))^2}{2\sigma^2}}$$  \hspace{1cm} (4.16)

Then the similarity between the two images is then defined as a weighted sum of $S^*(I^q_k, I^g)$, with the weights denoted by $w = [w_1, w_2, ..., w_m]^\top$. Recall that $m$ denotes the total number of patches per image.

$$S(I^q, I^g) = \sum_{k=1}^{m} w_k \cdot s^*(I^q_k, I^g)$$  \hspace{1cm} (4.17)

Given a triplet of query image $I^q$, positive (similar) image $I^+$ and negative (not similar) image $I^-$, this thesis redefines a hinge loss function as:

$$\ell(I^q, I^+, I^-) = (1 - S(I^q, I^+) + S(I^q, I^-))^p$$  \hspace{1cm} (4.18)

where the operator $(x)_+$ stands for max($x, 0$) and $p$ is the power with $p = 1, 2$. Combining with Eq.(4.17), Eq.(4.18) can be rewritten as

$$\ell(I^q, I^+, I^-) = (1 - w^\top x)_+^p$$  \hspace{1cm} (4.19)

where $x$ is a column vector of $[x_1, x_2, ..., x_m]^\top$, and $x_k$ denotes

$$x_k = s^*(I^q_k, I^+) - s^*(I^q_k, I^-), \ k = 1, 2, ..., m$$  \hspace{1cm} (4.20)

4.5.2 Experiment

This thesis first applies a single RBF kernel with SWL on Oxford5k dataset and NAA29k dataset. $\sigma_0$ is calculated as the average value of the Euclidean distances of all patch pairs. Here, $\sigma_0 = 1.3517$ for Oxford5k dataset. Different $\sigma$ is selected from $\{0.3, 0.4, ..., 1.4\}$ times of $\sigma_0$. The regularisation parameter $C$ and $\gamma$ used in Eq.(4.9) are fixed as $2^3$ and $2^7$ respectively.

Results for a single RBF kernel on Oxford5k are shown in Fig. 4.7(a) and Fig. 4.7(b). Results the for the single RBF kernel on NAA29k are shown in Fig. 4.7(c) and Fig. 4.7(d). From the results, it can be seen that SWL with a single RBF kernel consistently achieves better performance than SWL with Euclidean distance on both Oxford5k dataset and NAA29k dataset.

This experimental result suggest that the SWL method could be extended with
Figure 4.7: Training and test curve of SWL with single kernel on Oxford5k and NAA29k.
various machine learning techniques. Future work can focus on more extensions: 
1) SWL using multiple kernel learning (MKL); 2) SWL with non-negative weight 
constraints; 3) SWL with sparsity constraints. It is expected that these variants 
could further enhance the retrieval performance of the SWL method.

4.6 Discussion

In very recent years, deep learning has affected every corner of the field of com-
puter vision. Firstly, supervised CNN trained from large-scale image databases 
has constantly refreshed the records of image classification and image detection 
tasks with deeper networks and higher computational ability. Secondly, de-
derived from the strongly supervised CNNs with the benefit from the so-called pre-
trained models, Deep Metric Learning (DML) has expanded the power of CNN 
from image classification and detection to many other computer vision applica-
tions. Recently, DML has been associated with Face verification [LHT17], Per-
sion re-identification [DLZ+18, SZL+15, HLZJ15], Fine-grained visual categoriza-
tion [QJZL15, CZLB16], Text-image embedding [WLL16], Image set classifica-
tion [LWD+15], Image Retrieval [LMP+16], and Image Matching task [HLJ+15]. 
Among these applications, image retrieval, typically regarded as a fully unsup-
ervised learning task, is a promising research direction to extend deep learning from 
supervised field to unsupervised field. Yet, image retrieval has its own difficulty 
since the gallery is not well-categorised and no label information is provided.

However, the existing DML methods have successfully avoided this difficulty by 
using some “fake” retrieval datasets instead of the real retrieval datasets. These 
fake retrieval datasets are well-categorised datasets and originally utilised for im-
age classification tasks. For example, Mir-flicker 25k [HL08], Cifar10 [Kri09], and 
NUS-Wide [CTH+09] are all well-labelled datasets. With class labels, the triplets 
generation procedure in these DML is easy and is often proceeded by a) randomly 
selecting from the intra-class for positive images and b) randomly selecting from the 
inter-class for the negatives. To accelerate the learning procedure, hard-negative 
sampling is utilised to generate high-quality triplets. Nevertheless, for real retrieval 
tasks, the existing triplets generation procedure will lose efficacy since there are no 
labels available in the gallery. It is important to study how to generate triplets 
for real image retrieval datasets. The proposed SWL performs in a semi-supervised 
manner to incorporate the unlabelled (gallery) images in a dataset. In specific, 
besides the triplet constraints generated with retrieval ground truth, this thesis 
generates unsupervised triplet constraints by randomly selecting gallery images as 
queries and determining its positive and negative images with an initial retrieved 
list. The obtained large number of unsupervised triplet constraints perform as a
sort of regularisation to stabilise the learning process. Future work can focus on developing more advanced methods for triplets generation.

4.7 Conclusion

This thesis proposes a semi-supervised weight learning method to automatically learn the combination weights in the spatial search method for image retrieval. The proposed method achieves comparable or better retrieval performance on four datasets than the existing spatial search that manually tunes these weights. Although this work currently focuses on learning weights by using off-the-shelf ConvNet features, the distance metric learning based weight learning framework in this thesis can be readily extended to fine-tune the ConvNet features and learn the combination weights simultaneously. The proposed SWL can also be extended to kernel SWL.
Chapter 5

Modelling Diffusion Process by Deep Neural Networks for Image Retrieval

By considering the underlying neighbourhood structure of images, diffusion process can better evaluate image similarity and has proven highly effective in improving image retrieval. Nevertheless, the diffusion process stores a large neighbourhood graph, costs more online retrieval time, and requires special algorithms other than simple Euclidean search. To address these issues, this thesis proposes to treat diffusion process as a “black box” and directly model it by training deep neural networks, so as to obtain better image representation that assimilates the effect of diffusion process and works with Euclidean search. This thesis firstly puts forward a kernel mapping interpretation to diffusion process, and then formulates the modelling as a deep metric learning problem. The proposed approach is unsupervised in the sense that it needs neither image labels nor external datasets, and completely avoids online diffusion process in retrieval. More interestingly, this thesis finds that this approach could even achieve better retrieval than the original diffusion process, instead of merely approximating it. Experiments verify its effectiveness and investigate its appealing characteristics such as the generalisation to new image insertion.

5.1 Introduction

Content-based image retrieval aims to retrieve from an image database the images that can meet the requirement set by a user, and the typical scenario may be to find the images visually similar to a query of example. As an important topic in computer vision, image retrieval has received intensive research and gained significant progress during the past two decades [SWS+00a, DJLW08, ZYT18]. In particular, the recent deep learning techniques greatly boost the performance of image retrieval. With the powerful deep feature representations, a simple Euclidean distance based search has been able to achieve excellent retrieval performance.

Diffusion process, by exploiting the underlying neighbourhood structure of data, has been shown as an effective mechanism to improve image retrieval [DB13, ZWG+03]. Through propagating affinity information on this structure, diffusion process can more accurately evaluate the similarity between images, showing ro-
bustness to background clutter, partial occlusion, and the variation on scale or illumination. Its effectiveness has been shown not only via traditional SIFT features [YKL09, DB13], but also with the recent deep features [ITA+17]. The latter has exhibited the state-of-the-art performance on benchmark datasets. A particular attractive property of diffusion process is that it improves image retrieval in an unsupervised manner. A more detailed introduction on diffusion process can be found in Section 5.2.1.

Nevertheless, for image retrieval, diffusion process is more sophisticated than a Euclidean search. It needs to store a large neighbourhood graph whose size increases linearly (or even quadratically) with the size of the image database. For a given query, diffusion needs to be performed in an online manner to evaluate the similarity of the query to the images in a database. These not only consume a large amount of memory, but also delay the response of retrieval. In short, although diffusion process brings better image similarity, to benefit from it has to pay the price on computational cost, real-time performance, and search complexity.

This thesis aims to improve the above situation to make diffusion-based image retrieval more efficient and practical. Above all, this thesis interprets diffusion process as performing a kernel-induced implicit mapping on the input feature representation of each image. It produces a more advanced feature representation upon which the simple Euclidean distance becomes effective in evaluating the similarity of images. This interpretation motivates the thesis to treat diffusion process as a “black box,” and instead of precisely modelling the underlying physical process of diffusion, this thesis explicitly learns such a mapping from the result of diffusion process. In doing so, this thesis will be able to avoid performing online diffusion but retain its positive effect, and enjoy the nice properties of Euclidean search such as simplicity, low computational cost, and the access to many data structures and algorithms.

The recent deep neural networks, characterised by the well-proven capacity in modelling complex functional mappings, provide an instrumental tool. To realise the above idea, this thesis proposes to formulate the modelling of diffusion process as a deep metric learning problem. Given an image database, this thesis first extracts feature representations for all images with a pre-trained deep network. Diffusion process is then performed off-line, once only, to evaluate the similarities among these images. According to these similarities, image triplets are generated to fine-tune the above deep network, making it learn the implicit kernel-induced mapping and therefore “assimilate” the effect of diffusion process. This fine-tuned network is then used to re-extract feature representations of all images in the database. Once a query (could be out of the database) is given, its representation will be extracted with the same fine-tuned network, and all retrieval in the sequel will purely be
performed with Euclidean distance on this new feature representation. Note that the proposed approach does not require any image label information or external datasets for training, and is therefore unsupervised.

In recent literature, several pieces of work have been aware of the aforementioned issue and made efforts to resolve it. The following two are particularly relevant to this work. The authors in [IAT+18] perform an off-line low-rank spectral decomposition of the affinity matrix in diffusion process, which helps to realise online diffusion with Euclidean and dot product search. Another work [ITAC18] shares an even similar spirit as ours\(^a\) but has a different focus. It utilises the change of \(k\)-nearest neighbourhoods before and after diffusion process to mine hard training examples for deep metric learning. In contrast, this thesis focuses more explicitly on modelling the diffusion process by interpreting it as a kernel mapping. Thanks to this different perspective, this thesis has several interesting findings on the effectiveness of this direct modelling (e.g., it could even achieve better performance than the original diffusion-based retrieval), its database-specific characteristic, and its generalisation and robustness with respect to the insertion of new images to a given database.

The contributions are summarised as follows:

1) By taking advantage of the powerful modelling capability of deep neural networks, this thesis proposes to model the highly non-linear diffusion process to generate explicit, better feature representation for image retrieval. It retains the positive effect of diffusion process but avoids online diffusion, significantly reducing computational cost and search complexity.

2) This thesis indicates an interesting unsupervised learning framework to bootstrap image retrieval, which exploits the underlying structure information of images in a database and converts it to better feature representations for Euclidean search. Moreover, better retrieval could be attained when this bootstrapping process is conducted with more iterations.

3) Experimental study shows multiple appealing properties of the proposed approach. In particular, it could even outperform diffusion process for image retrieval, although the original goal is merely to simulate the effect of diffusion process. This is significant and inspiring, and better justifies the value of the proposed approach.

\[5.2\] The Proposed Approach

As pointed out in Section 5.1, although diffusion process can effectively improve image retrieval, it also brings a number of issues. In detail, this thesis identifies the following ones: i) large memory cost to store the neighbourhood graph; ii) 

\(^a\)The author of the thesis would like to clarify that the work in this thesis has been independently developed since 2017.
Figure 5.1: The proposed framework consists of four components for offline training: 1) original features extracted for the images in a database with a convolutional neural network (CNN); 2) constructing the neighbourhood graph with the extracted features and performing diffusion with the graph to obtain image similarities; 3) image triplet generation based on the rankings obtained with the image similarities; and 4) training a deep metric CNN network with the generated triplets. Specifically, the three coloured CNNs at the center of the figure will be trained with stochastic gradient descent. The yellow arrows connecting them mean that the weights are shared across the three CNNs. After the training process, the features for all images in the database are re-extracted with the newly trained CNN network. For online retrieval, when a query is submitted, extract the features of this query with the same trained CNN network and simply perform a Euclidean search over the database.
prolonged retrieval time; iii) special treatment to handle a query out of a given image database; iv) having to perform query-specific diffusion, and v) having to update the neighbourhood graph when new images are inserted into a database. All of these issues, more or less, significantly affect image retrieval in practice. The idea is to view diffusion process as performing an unknown, implicit, highly non-linear mapping from the input feature space to another feature space in which a Euclidean-based measure can align well with the image similarities obtained by diffusion process. The framework of the proposed method is shown in Fig. 5.1.

5.2.1 A Mapping View of Diffusion Process

Let \( \{x_1, x_2, \cdots, x_n\} \) denote a set of data points (e.g., images) in a vector space \( \mathcal{X} \). Diffusion process usually starts from computing an \( n \times n \) pairwise affinity matrix \( A \). A weighted undirected graph is then constructed, with each node corresponding to a data point and each edge corresponding to the pairwise affinity of two linked points. This thesis now shows that diffusion process can be interpreted as performing an implicit, non-linear kernel mapping.

As surveyed in [DB13], although many variants of diffusion process have been developed in the literature, they can be well categorised and summarised according to three factors, i.e., the initialisation matrix \( W_0 \), the transition matrix \( T \), and the update scheme. The update scheme in [YKL09] gives the best image retrieval performance, and it is expressed as

\[
W_{t+1} = TW_tT^\top, \tag{5.1}
\]

where \( T \) denotes matrix transpose. Note that most of the methods surveyed in [DB13] set the initialisation \( W_0 \) as the affinity matrix \( A \). In this case, it is not difficult to obtain that

\[
W_{t+1} = T^{t+1}W_0(T^{t+1})^\top = T^{t+1}A(T^{t+1})^\top, \tag{5.2}
\]

where the superscript of \( T \) denotes the order of power. A common way to compute the entries of \( A \) uses a Gaussian RBF kernel, making \( A \) positive definite (PD). Immediately, this makes \( W_{t+1} \) a PD matrix and satisfy the Mercer’s condition [Bur98]. So, this thesis can interpret \( W_{t+1} \) as a kernel matrix. In particular, the kernel function between points \( x_i \) and \( x_j \) can be written as

\[
\kappa(x_i, x_j | A) \triangleq W_{t+1}(i, j) = T^{t+1}(i, :)A(T^{t+1}(j, :))^\top, \tag{5.3}
\]

where \( T^{t+1}(i, :) \) denotes the \( i \)th row of the matrix \( T^{t+1} \). As seen, this is a “context-
aware” kernel and its value depends on the whole matrix $A$ due to the diffusion process. This well shows the characteristic of diffusion process. Therefore, since i) the output of diffusion process, $W_{t+1}$, can be interpreted as a kernel matrix obtained via the kernel $\kappa$ and ii) each kernel induces an implicit non-linear mapping from an input space to another feature space, this thesis can indeed interpret diffusion process as performing an implicit, non-linear kernel mapping.

Meanwhile, it is worth noting that this thesis uses this mapping view to primarily illustrate the idea behind the proposed method, that is, showing what the deep neural network is essentially modelling. In practice, the proposed method requires neither the positive definiteness of $W_{t+1}$ nor the existence of a kernel function like $\kappa(x_i, x_j|A)$. What this thesis needs will just be the ranking information of images, from which this thesis can generate image triplets to train the network. This requirement allows the proposed method to readily work with any diffusion process employed in the literature of image retrieval.

### 5.2.2 A Deep Metric Learning Approach

The challenge of learning the aforementioned implicit non-linear mapping via deep neural networks lies at how to train the deep neural networks. This thesis can certainly train the network to produce feature representations such that their inner products best approximate the obtained affinity values in $W_{t+1}$. Nevertheless, considering that this thesis is dealing with image retrieval and many diffusion processes used in the literature output ranking scores instead of affinity values, this thesis formulates this idea as a common deep metric learning problem, that is, a deep neural network is trained with a set of triplets of the images chosen from a database. Each triplet is composed of one anchor image, one closer image, and one farther image.\(^b\)

Being closer or farther from an anchor image is defined according to the ranking scores produced by diffusion process. During training, this thesis enforces that for a given query, its distance to the closer image should be smaller than the distance to the farther one by a margin. In the following part, three key issues on training the deep triplet network are elaborated, including 1) network structure, 2) triplet generation, and 3) triplet loss function.

**Network structure.** As illustrated in Fig. 5.1, the deep triplet network consists of three CNNs, with all layers shared. They accept the anchor, closer, and farther images as the input, respectively. This thesis adopts the residual network architectures [HZRS16] for the CNN due to its outstanding performance demonstrated in the recent literature. The triplet loss function is applied to the features\(^b\)

\(^b\)Note that different from existing deep metric learning methods, this thesis does not access any labelled data. Therefore, this thesis uses “closer” and “farther” (instead of positive and negative) to be more precise.
output by the three CNNs. The weights of these three CNNs will be learned with
the stochastic gradient descent technique.

**Triplets generation.** The training of deep triplet network relies on the generation of high-quality image triplets. Generating triplets by randomly sampling from the images in a database can hardly provide useful information to benefit the training. This thesis takes a locally constrained triplet generation method. Specifically, given an anchor image $I_a$, its $k$-nearest neighbouring images are identified based on the ranking scores obtained by the diffusion process, and they are collectively denoted by a set $N_k(I_a)$. Two images are then randomly sampled from the set $N_k(I_a)$. According to their ranking positions with respect to $I_a$, this thesis regards them as the closer image $I_c$ and the farther image $I_f$, respectively, to form a triple $(I_a, I_c, I_f)$. This thesis denotes all the generated triplets collectively by a set $S$.

**Triplet loss function.** This thesis uses the following triplet loss which has a soft margin

$$
L = \sum_{(I_a, I_c, I_f) \in S} \left[ d(I_a, I_c) - d(I_a, I_f) + \frac{|r_f - r_c|}{k} m_0 \right]_+, \quad (5.4)
$$

where $r_c$ and $r_f$ denote the ranking positions of $I_c$ and $I_f$ with respect to $I_a$, $[x]_+$ denotes $\max(x, 0)$, and $d(I, J)$ is the Euclidean distance between images $I$ and $J$ based on the features output by the three CNNs. $k$ is the size of neighbourhood used in the triplet generation step and $m_0$ is a constant as the basic margin. The coefficient $\frac{|r_f - r_c|}{k}$ is a slight modification of the commonly used triplet loss function. In doing so, the magnitude of this soft margin can therefore adapt to the ranking difference between the closer and farther samples, and this thesis finds that this is helpful for the network to learn.

### 5.2.3 A Bootstrapping Framework for Image Retrieval

Built upon the above deep metric learning approach to modelling the diffusion process, this thesis proposes an unsupervised bootstrapping framework for image retrieval as follows. It could iterate between performing diffusion process and learning better feature representations to maximise the improvement on retrieval performance.

1. Given an image database, extract feature representations of (part or all of) the images, denoted by $\{x_1, x_2, \cdots, x_n\}$, in the database with a pre-trained deep neural network;

2. Construct the affinity matrix $A$ with the extracted features. Perform diffusion process with $A$ (or any of its variants) to obtain the scores on image similarity;
5.3 Experimental Study

5.3.1 Experimental Setup

**Dataset.** The proposed approach is tested on six benchmark datasets, including Oxford5k [PCI+07b], Pairs6k [PCI+08], INSTRE [WJ15], Sculpture [AZ11], Oxford105k, and Pairs106k. The last two are obtained by adding 100k distractor images collected from Flicker. For each dataset, there is no overlapping between the images in the database and query images. To ensure an objective evaluation, this thesis only uses the images in the database to do diffusion and train the network, and reserves the query images exclusively to test retrieval performance. This is consistent with the protocol commonly adopted in the literature. In addition, this thesis evaluates INSTRE by following the recent work [ITA+17] and uses standard evaluation protocol for all the other datasets. Mean average precision (mAP) is used to measure retrieval performance in all experiments. These datasets are summarised in Table 5.1.

**Network training.** The CNN of ResNet101 pre-trained with ImageNet [HZRS16] is referred to as “ResNet101-ImageNet” and used in Tasks 1 and

---

**Table 5.1:** Six datasets are used in this experiment. This thesis only uses the images in each database to train the proposed method, and reserves the corresponding query images exclusively to test retrieval performance.

<table>
<thead>
<tr>
<th>Benchmark datasets</th>
<th>Total No. of images</th>
<th>No. of images in the database</th>
<th>No. of query images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxford5k</td>
<td>5063</td>
<td>5008</td>
<td>55</td>
</tr>
<tr>
<td>Paris6k</td>
<td>6392</td>
<td>6337</td>
<td>55</td>
</tr>
<tr>
<td>Oxford105k</td>
<td>105134</td>
<td>105079</td>
<td>55</td>
</tr>
<tr>
<td>Paris106k</td>
<td>106463</td>
<td>106408</td>
<td>55</td>
</tr>
<tr>
<td>INSTRE</td>
<td>28543</td>
<td>27293</td>
<td>1250</td>
</tr>
<tr>
<td>Sculpture</td>
<td>3170</td>
<td>3100</td>
<td>70</td>
</tr>
</tbody>
</table>
2 defined in the last paragraph of Section 5.3.1. During training, all images are resized with the longer side having 600 pixels. Stochastic gradient descent technique is used. The learning rate is initialised as 0.01 and gradually attenuates during the training process. The coefficients for weight decay and the momentum are 0.0001 and 0.9. The batch size is 40, and the training process usually takes 1000 epochs to converge. The margin $m_0$ in Eq.(5.4) is empirically set as 0.1, and the number of nearest neighbours, $k$, to generate triplets is set as 300. Note that to clearly show the basic performance of the proposed method, this thesis only trains the deep metric network with the diffusion process once in all experiments, except the part particularly investigating the case of multiple iterations.

**Retrieval setting and baseline.** R-MAC [TSJ15] feature representation is used to describe each image. The LCDP [YKL09] update scheme (in Eq.(5.1)) is adapted to perform diffusion. This thesis compares the proposed method with the following two baselines: 1) R-MAC+E and 2) R-MAC+D, where “E” and “D” mean that Euclidean distance search and diffusion process are used for retrieval, respectively. By the comparison, this thesis wants to verify whether the proposed approach can effectively achieve or even improve over the retrieval performance obtained via diffusion process. To ensure fair comparison, this thesis implements the baselines and the proposed methods under the same experimental setting.

**Experimental tasks.** There are five tasks: 1) Compare the proposed method with R-MAC+E and R-MAC+D, where each image is represented by a global representation (i.e., R-MAC); 2) Compare it with the state-of-the-art retrieval methods where each image is represented by a set of regional representations; 3) Compare it with various recent image retrieval methods to give a whole picture; 4) Compare it with diffusion-based image retrieval in terms of time and memory cost in online retrieval; and 5) Investigating important properties of the proposed method, such as its generalisation and robustness to image insertion and the help of multiple iteration training. The results are reported in order in the next section.

### 5.3.2 Result and Discussion

**Task 1.** Table 5.2 compares the proposed method with R-MAC+E and R-MAC+D under the global image representation. As shown, by conducting diffusion process, R-MAC+D consistently achieves higher retrieval performance (5 to 16 percentage points) than R-MAC+E that uses Euclidean distance based search. By assimilating the effect of diffusion process via deep metric learning, the proposed method, still using Euclidean search, not only achieves comparable performance as R-MAC+D on Oxford5k, Oxford105k, and INSTRE, but also outperforms it on Paris6k, Paris106k, and Sculpture. In particular, the proposed method brings up to 14 percentage points
CHAPTER 5. MODELLING DIFFUSION FOR IMAGE RETRIEVAL

Table 5.2: Comparison with two baseline methods under a global image representation, where “E” denotes the Euclidean distance based search while “D” denotes diffusion process. The “ResNet101-ImageNet” network is used in this experiment.

<table>
<thead>
<tr>
<th>Method (mAP)</th>
<th>Oxford5k</th>
<th>Paris6k</th>
<th>Oxford105k</th>
<th>Paris106k</th>
<th>INSTRE</th>
<th>Sculpture</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-MAC+E (global)</td>
<td>58.5</td>
<td>73.3</td>
<td>57.3</td>
<td>67.4</td>
<td>37.7</td>
<td>51.0</td>
</tr>
<tr>
<td>R-MAC+D (global)</td>
<td>63.1</td>
<td>83.5</td>
<td>62.0</td>
<td>77.0</td>
<td>53.0</td>
<td>61.6</td>
</tr>
<tr>
<td>Proposed (global)</td>
<td><strong>63.2</strong></td>
<td><strong>89.6</strong></td>
<td><strong>62.4</strong></td>
<td><strong>82.6</strong></td>
<td><strong>54.5</strong></td>
<td><strong>75.5</strong></td>
</tr>
</tbody>
</table>

Table 5.2: Comparison with two baseline methods under a global image representation, where “E” denotes the Euclidean distance based search while “D” denotes diffusion process. The “ResNet101-ImageNet” network is used in this experiment.

of mAP increase on Sculpture (75.5 (ours) vs 61.6). This result is significant. It shows that the proposed method is indeed effective in assimilating the effect of diffusion process and converting it to enhanced feature representation. Furthermore, it is surprising to observe these large improvements on three datasets. This thesis attributes such improvement to the wholly manner of the proposed method in approximating diffusion process. That is, the proposed deep metric learning is performed upon a large number of image triplets generated from the whole database. This provides the network with a “global” view about the similarity of these images, and therefore may help the network to produce overall better feature representations. As for R-MAC+D, a specific diffusion process is performed for a given query, and this process is initialised by or dependent on this query. Such a “local” view may limit its overall retrieval accuracy. This interesting issue will be further explored in future work.

Task 2. Region-based image retrieval methods have recently shown excellent performance by representing an image as a set of regional features. The image similarity is usually evaluated by summarising the similarity across image regions. In this case, diffusion process is performed on the graph constructed upon these regional deep features. In this task, this thesis focuses on Paris6k to compare with two state-of-the-art methods of this kind: the cross-region matching method [RASC14b] and the regional diffusion method [ITA+17] (re-implemented by the thesis with the network “ResNet101-ImageNet”). They obtain the mAP of 84.4 and 91.8, respectively. The better performance of the regional diffusion method is due to its use of diffusion process to evaluate the similarity of image regions. The proposed method achieves an mAP of **93.8**, which further improves the regional diffusion method by two percentage points. This again indicates the effectiveness and advantage of the proposed method in approximating diffusion process.

Task 3. To give a whole picture about the performance of the proposed method, Table 5.3 compares it with the image retrieval methods developed in the recent literature. To be consistent with the state-of-the-art methods, this thesis uses the CNN
structure provided in [GARL17], which fine-tunes ResNet101 with an additional landmark dataset (called “ResNet101-Landmarks” in this thesis), for the proposed deep metric learning approach. This thesis categorises all the retrieval methods in comparison into two groups: 1) the methods only applying Euclidean distance based search with a global feature representation, as shown in the upper part of the table; and 2) the methods applying diffusion process or post-processing steps (such as query expansion, matching, and verification), which appear in the lower part of the table. The first group of methods enjoys higher computational efficiency in retrieval, while the second group of methods generally achieves higher retrieval performance. For the proposed method, which only conducts Euclidean search to retrieve images, it can well outperform most of the methods in the first group and achieve quite competitive performance to those in the second group that have more sophisticated online retrieval mechanisms. This result shows that by assimilating the effect of diffusion process with new features, the proposed method can enjoy both high computational efficiency and high retrieval accuracy for online retrieval. In addition, it is worth noting that the results in Tables 5.2 and 5.3 are not directly comparable. The proposed method in Table 5.2 is implemented based on the “ResNet101-ImageNet” network, while in Table 5.3 it is implemented based on the network “ResNet101-Landmarks” to be consistent with the state-of-the-art methods.

**Task 4.** To show computational efficiency, this thesis compares the time and memory cost of the proposed method with diffusion-based image retrieval in performing online retrieval on three datasets. The experiment is conducted with Matlab2017a on a desktop computer of Intel@core i7-4720 2.60GHz CPU and the result is reported in Table 5.4 for retrieval with the global and regional representations, respectively. As expected, the proposed method is consistently faster and can shorten online retrieval up to 10 times for a single query. Furthermore, due to the use of Euclidean distance, the proposed method can readily be sped up by utilising off-the-shelf data structure and algorithms. Also, because it does not need to store the neighbourhood graph, it incurs no extra memory usage in this aspect.

**Task 5.**

1) **Image insertion.** One drawback of diffusion-based image retrieval lies at that it needs to update its neighbourhood graph when new images are inserted into a database. This experiment investigates the robustness and the generalisation capability of the proposed method in this situation. Now, this thesis only uses part of the images \(n_0\) in a database to build the graph, conduct diffusion, and generate triplets for training. However, when a query is submitted, the retrieval will be performed on the whole database. This simulates the case that all the remaining images (i.e., other than these \(n_0\) ones) are newly inserted after the proposed method is trained. As previous, the proposed method uses the learned feature representations, respectively, to perform retrieval. The result is plotted in Fig. 5.2. The horizontal
Table 5.3: Comparison with the state-of-the-art image retrieval methods. The result shows that the proposed method effectively assimilates the effect of diffusion process to generate better feature representations, upon which it achieves very competitive retrieval performance with simple Euclidean distance. [ITA+17]* and [ITA+17]** are the results reported by [ITA+17] as the re-implementation of [RTC16] and [GARL17] with ResNet101 fine-tuned on an external landmark dataset. Top three values per column are in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dim.</th>
<th>Oxford5k</th>
<th>Paris6k</th>
<th>Oxford105k</th>
<th>Paris106k</th>
<th>INSTRE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global image representation with Euclidean search</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[JZ14]</td>
<td>128</td>
<td>43.3</td>
<td>-</td>
<td>35.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[BSCL14]</td>
<td>128</td>
<td>55.7</td>
<td>-</td>
<td>52.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[YHNYD15]</td>
<td>128</td>
<td>59.3</td>
<td>59.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[BL15]</td>
<td>256</td>
<td>53.1</td>
<td>-</td>
<td>50.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[TSJ15]</td>
<td>512</td>
<td>66.9</td>
<td>83.0</td>
<td>61.6</td>
<td>75.7</td>
<td>-</td>
</tr>
<tr>
<td>[KMO16]</td>
<td>512</td>
<td>68.2</td>
<td>79.7</td>
<td>63.3</td>
<td>71</td>
<td>-</td>
</tr>
<tr>
<td>[ITA+17]*</td>
<td>512</td>
<td>77.7</td>
<td>84.1</td>
<td>70.1</td>
<td>76.8</td>
<td>47.7</td>
</tr>
<tr>
<td>[ITAC18]</td>
<td>512</td>
<td>78.2</td>
<td>85.1</td>
<td>72.6</td>
<td>78</td>
<td>57.7</td>
</tr>
<tr>
<td>[RTC16]</td>
<td>512</td>
<td>79.7</td>
<td>83.8</td>
<td>73.9</td>
<td>76.4</td>
<td>-</td>
</tr>
<tr>
<td>[JZ14]</td>
<td>1024</td>
<td>56.0</td>
<td>-</td>
<td>50.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[TSJ15]</td>
<td>2048</td>
<td>69.4</td>
<td>85.2</td>
<td>63.7</td>
<td>77.8</td>
<td>-</td>
</tr>
<tr>
<td>[GARL17]</td>
<td>2048</td>
<td>86.1</td>
<td>94.5</td>
<td>82.8</td>
<td>90.6</td>
<td>-</td>
</tr>
<tr>
<td>[ITA+17]*</td>
<td>2048</td>
<td>83.9</td>
<td>93.8</td>
<td>80.8</td>
<td>89.9</td>
<td>62.6</td>
</tr>
<tr>
<td>[AGT+18]</td>
<td>4096</td>
<td>71.6</td>
<td>79.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Global image representation + diffusion / query expansion / matching / verification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[KMO16]</td>
<td>-</td>
<td>72.2</td>
<td>85.5</td>
<td>67.8</td>
<td>79.7</td>
<td>-</td>
</tr>
<tr>
<td>[SLBW14]</td>
<td>-</td>
<td>75.2</td>
<td>74.1</td>
<td>72.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[QGB+11]</td>
<td>-</td>
<td>81.4</td>
<td>80.3</td>
<td>76.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[CMPM11]</td>
<td>-</td>
<td>82.7</td>
<td>80.5</td>
<td>76.7</td>
<td>71.0</td>
<td>-</td>
</tr>
<tr>
<td>[DJL+13]</td>
<td>-</td>
<td>84.3</td>
<td>83.4</td>
<td>80.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[MPCM14]</td>
<td>-</td>
<td>84.9</td>
<td>82.4</td>
<td>79.5</td>
<td>77.3</td>
<td>-</td>
</tr>
<tr>
<td>[TAJ16]</td>
<td>-</td>
<td>86.9</td>
<td>85.1</td>
<td>85.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[TJ14]</td>
<td>-</td>
<td>89.4</td>
<td>82.8</td>
<td>84.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[TSJ15]</td>
<td>512</td>
<td>77.3</td>
<td>86.5</td>
<td>73.2</td>
<td>79.8</td>
<td>-</td>
</tr>
<tr>
<td>[ARS+16]</td>
<td>512</td>
<td>79.0</td>
<td>85.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[RTC16]</td>
<td>512</td>
<td>84.5</td>
<td>86.4</td>
<td>80.4</td>
<td>79.7</td>
<td>-</td>
</tr>
<tr>
<td>[ITA+17]*</td>
<td>512</td>
<td>85.4</td>
<td>88.4</td>
<td>79.7</td>
<td>83.5</td>
<td>57.3</td>
</tr>
<tr>
<td>[TSJ15]</td>
<td>2048</td>
<td>78.9</td>
<td>89.7</td>
<td>75.5</td>
<td>85.3</td>
<td>-</td>
</tr>
<tr>
<td>[ITA+17]</td>
<td>2048</td>
<td>87.1</td>
<td><strong>96.5</strong></td>
<td>87.4</td>
<td><strong>95.4</strong></td>
<td><strong>80.5</strong></td>
</tr>
<tr>
<td>[GARL17]</td>
<td>2048</td>
<td>90.6</td>
<td>96.0</td>
<td><strong>89.4</strong></td>
<td>93.2</td>
<td>-</td>
</tr>
<tr>
<td>[ITA+17]*</td>
<td>2048</td>
<td><strong>89.6</strong></td>
<td>95.3</td>
<td><strong>88.3</strong></td>
<td>92.7</td>
<td>70.5</td>
</tr>
<tr>
<td>[IAT+18]</td>
<td>2048</td>
<td><strong>87.5</strong></td>
<td><strong>96.4</strong></td>
<td><strong>87.9</strong></td>
<td><strong>95.3</strong></td>
<td><strong>80.5</strong></td>
</tr>
</tbody>
</table>

The proposed global image representation (by modelling diffusion process) + Euclidean search

| Proposed    | 2048 | 85.4 | **96.3** | 85.1 | **94.7** | **71.7** |

axis is the ratio of images taken from a database used for performing diffusion and training the proposed method, while the vertical axis shows the mAP value. The three dotted lines indicate the baseline performance for Oxford105k, Paris6k, and
Table 5.4: Comparison of average time / memory usage (Second / GB) in online retrieval. The dimensions of image feature representation are 2048. Time cost is averaged over all of the queries.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Global feature representation</th>
<th>Regional feature representation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oxford5k</td>
<td>INSTRE</td>
</tr>
<tr>
<td>Diff. based</td>
<td>0.020/0.01</td>
<td>0.100/0.03</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.002/N.A.</td>
<td>0.011/N.A.</td>
</tr>
</tbody>
</table>

INSTRE, respectively, when diffusion process is not used. The three solid lines show the corresponding performance of the proposed method. As seen, its performance steadily improves with the increasing ratio and quickly approaches the level when all images (i.e., ratio = 1.0) in a database are used for diffusion process and training the proposed method. For Paris6k, an improvement over the baseline can be observed even when the ratio is as low as 0.1. As for Oxford105k and INSTRE, clear improvement can be obtained once the ratio exceeds 0.3. These results show that the proposed method generalises well with respect to the insertion of new images. 2) Iterative training by re-applying diffusion. The proposed bootstrapping framework for image retrieval supports iterative training. This thesis can alternate between learning new feature representation and performing diffusion process with this learned representation. Tested on INSTRE, the proposed method does obtain better retrieval by one or two extra iterations, and the mAP result is 71.7 at the first iteration, 74.2 at the second, and 74.5 at the third, all higher than the baseline of 70.8. This result initially shows the effectiveness of the proposed bootstrapping framework and its potential will be further explored in the future work.

5.4 Fusion of Diffusion Based on NAA Database

NAA database has been a driving force for deep learning image retrieval research in this thesis, and it has a good application and practical significance. After understanding the effectiveness of the diffusion process on image retrieval, how to design a better diffusion method becomes an important research direction. In recent years, from the perspective of data fusion, methods for integrating multiple diffusion processes have been proposed and studied.

The main content of this section is to explore the fusion of diffusion framework from two perspectives: 1) fusion of global and regional diffusion; 2) fusion of visual and textual diffusion. The two fusion methods are based on a recent work [ITA+17]. This work has provided an off-the-shelf method, marked as Method_{off}. In specific, Method_{off} studies a new approach to build the affinity matrix and estimate the
Different ratios of images are used for diffusion and training

Figure 5.2: Retrieval performance of the proposed method when different ratios of images are taken from a database for conducting diffusion process and training the proposed method. This experiment investigates its generalisation capability with respect to new image insertions. The three dotted lines indicate the baselines when diffusion process is not used. The solid lines show the corresponding performance of the proposed method. The “ResNet101-Landmarks” network is used in this experiment.
Given a dataset of \( n \) images with features \( \chi := \{ \mathbf{x}_1, ..., \mathbf{x}_n \} \subset \mathbb{R}^d \), where \( d \) is the dimension of features. The similarity between two data points \( \mathbf{x} \) and \( \mathbf{z} \) is defined as

\[
s(\mathbf{x}|\mathbf{z}) = \begin{cases} 
    s(\mathbf{x}, \mathbf{z}) & \text{if } \mathbf{x} \in \text{NN}_k(\mathbf{z}); \\
    0, & \text{otherwise}
\end{cases}
\]  
(5.5)

in which \( \text{NN}_k(\mathbf{z}) \) denotes the set of the \( k \) nearest neighbours of \( \mathbf{z} \) in \( \chi \). Also, \( s(\mathbf{x}, \mathbf{z}) \) can be calculated by inner product of \( \mathbf{x} \) and \( \mathbf{z} \). Then

\[
s_k(\mathbf{x}, \mathbf{z}) = \min\{s_k(\mathbf{x}|\mathbf{z}), s_k(\mathbf{z}|\mathbf{x})\}
\]  
(5.6)

An affinity matrix \( \mathbf{A} \in \mathbb{R}^{n \times n} \) is constructed by

\[
\mathbf{A}_{ij} = s_k(\mathbf{x}_i, \mathbf{z}_j), \forall (i, j) \in [n]^2, [n] = 1, 2, ..., n.
\]  
(5.7)

The affinity matrix is normalised by \( \mathbf{S} := \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2} \), in which \( \mathbf{D} \) is degree matrix of \( \mathbf{A} \). Then diffusion process or random walk is performed by solving

\[
\mathbf{f}^* = (1 - \alpha) \mathcal{L}_\alpha^{-1} \mathbf{y}
\]  
(5.8)

in which \( \mathcal{L}_\alpha = \mathbf{I}_n - \alpha \mathbf{S} \), and \( \alpha \) is a jump probability in random walk which is smaller than 1 and larger than 0. Based on the similarity measurement in Eq.(5.6), given a query image and its associated data point \( q, y_i \), the \( i \)th element of its query score \( \mathbf{y} \), is estimated by

\[
y_i = s_k(\mathbf{x}_i, \mathbf{q}), \forall i \in [n]
\]  
(5.9)

Finally, \( \mathbf{f}^* \) is obtained as a vector of ranking scores (\( \mathbf{RS} \)).

The above method is based on the global diffusion mechanism. Following the same way utilised in [ITA+17], the method can be applied to the regional diffusion mechanism. For regional diffusion, \( y_i \) is estimated by a set of query data points \( \mathbf{Q} \), which are the \( \mathbf{Q} \) are k-nearest neighbour image regions of the given query image.

\[
y_i = \sum_{\mathbf{q} \in \mathbf{Q}} s_k(\mathbf{x}_i, \mathbf{q}), \forall i \in [n]
\]  
(5.10)

If looking into these equations, the key input of Method_{off} consists of two parts. One is the \( k \) nearest neighbours information \( \text{NN}_k \) and the other information is \( s \) for all data points. Noting a knngraph, denoted as \( \mathbf{KG} \), is just two parts of \( \text{NN}_k \) and \( s \), where \( \text{NN}_k \) can be inferred if given \( s \) and \( k \). To simplify the definition in the following section, recalling that \( \mathbf{KG}^* = (\text{NN}_k^*, s^*) \) for any marker \( * \). In short, the input of Method_{off} is \( \mathbf{KG} \) and the output is \( \mathbf{RS} \).
5.4.1 Fusion of Global and Regional Diffusion

Proposed Method

This section designs an experimental study to explore the fusion of global diffusion and regional diffusion based on Method off. Let’s define knngraph built by local features as KG\textsubscript{global} and knngraph built by regional features as KG\textsubscript{regional}. Also, define RS\textsubscript{global} and RS\textsubscript{regional} as the ranking scores obtained by Method off if given KG\textsubscript{global} and KG\textsubscript{regional}. The fusion can be applied in two ways. One way is simple by directly adding two ranking scores RS\textsubscript{global} and RS\textsubscript{regional} after each diffusion. The obtained ranking score is RS\textsubscript{after} = RS\textsubscript{global} + RS\textsubscript{regional}.

The other way is more complicated by fusing the two knngraphs KG\textsubscript{global} and KG\textsubscript{regional} before diffusion, by building a new knngraph KG\textsubscript{before}. Given image I and its associated data point x, and regional image rI and its associated data point rx. Denoting rx ∈ x if the regional image rI is cropped from the image I. The reduced knngraph KG\textsubscript{reduced} is built by s\textsubscript{reduced}(x, z) = \sum_{rx∈x, rz∈z} s\textsubscript{regional}(rx, rz).

Secondly, fusing two knngraph by obtaining s\textsubscript{before}(x, z) = s\textsubscript{reduced}(x, z) + s\textsubscript{global}(x, z). Then feeding KG\textsubscript{before} into Method off to obtain RS\textsubscript{before}.

Experiments

For the experimental study, it is noticed that there are two hyper-parameters k (utilised in Eq.(5.5)) and kq (utilised in Eq.(5.9)) in the work [ITA+17] which may affect the performance. One is the number of nearest neighbours k for constructing the knngraph and the other is the number of nearest neighbours kq to keep for a query. For this experiment, k is varied within 10, 20, 30, 50, and 200, and kq is varied within 1, 2, 3, 5, and 10. The NAA29k dataset is utilised for the experiments.

Results

Firstly, the results of global diffusion and regional diffusion are obtained and they are shown in Table 5.5 and Table 5.6. Compared with the baseline (without applying diffusion), the Method off can achieve 6 to 10 percentage points increase with a different hyper-parameters choice on both global and regional diffusion. This verifies that the Method off also works well on the NAA29k dataset. Also, regional diffusion is much better than global diffusion since regional diffusion can better retrieve some smaller objects that global diffusion cannot handle. Secondly, the two proposed fusion of diffusion methods are applied and the results are shown in Table 5.7. Although the proposed methods are simple and easy to implement based on the off-the-shelf diffusion methods, the performance for the proposed fusion methods can still exceed that of the regional diffusion method by around two percentage points. This indicates that global diffusion and regional diffusion can be complementary and it is worth to exploring a fusion of diffusion.
Table 5.5: Global diffusion results obtained by $Method_{off}$ with different hyperparameters of nearest neighbourhood $k$ and $k_q$. and “w/o” denotes “without”.

<table>
<thead>
<tr>
<th>k = 10</th>
<th>k = 20</th>
<th>k = 30</th>
<th>k = 50</th>
<th>k = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td>61.07</td>
<td>64.48</td>
<td>66.98</td>
<td>68.20</td>
<td>69.08</td>
</tr>
<tr>
<td>64.07</td>
<td>66.09</td>
<td>68.19</td>
<td>68.39</td>
<td>68.72</td>
</tr>
<tr>
<td>64.39</td>
<td>66.43</td>
<td>67.88</td>
<td>67.92</td>
<td>68.49</td>
</tr>
<tr>
<td>60.67</td>
<td>62.16</td>
<td>63.39</td>
<td>64.03</td>
<td>65.50</td>
</tr>
</tbody>
</table>

Table 5.6: Regional diffusion results obtained by $Method_{off}$ with different hyper-parameters of nearest neighbourhood $k$ and $k_q$. and “w/o” denotes “without”.

<table>
<thead>
<tr>
<th>k = 10</th>
<th>k = 20</th>
<th>k = 30</th>
<th>k = 50</th>
<th>k = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td>73.31</td>
<td>76.21</td>
<td>78.54</td>
<td>78.67</td>
<td>78.69</td>
</tr>
<tr>
<td>76.05</td>
<td>76.79</td>
<td>77.74</td>
<td>78.43</td>
<td>78.69</td>
</tr>
<tr>
<td>74.16</td>
<td>74.68</td>
<td>75.41</td>
<td>76.26</td>
<td>76.19</td>
</tr>
<tr>
<td>74.45</td>
<td>74.05</td>
<td>74.38</td>
<td>73.99</td>
<td>73.84</td>
</tr>
</tbody>
</table>

Table 5.7: Results of proposed fusion methods. Two proposed approaches, fusion before diffusion and fusion after diffusion, perform better than the global diffusion and regional diffusion method. Different $k_{q_{global}}$ and $k_{q_{regional}}$ are utilised. $k$ is fixed as 50.

<table>
<thead>
<tr>
<th>$k_{q_{global}}$</th>
<th>$k_{q_{regional}}$</th>
<th>global</th>
<th>regional</th>
<th>fusion after diffusion</th>
<th>fusion before diffusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>69.08</td>
<td>78.54</td>
<td>79.39</td>
<td>79.67</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>69.08</td>
<td>77.74</td>
<td>79.33</td>
<td>79.58</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>68.72</td>
<td>78.54</td>
<td>80.15</td>
<td><strong>80.53</strong></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>68.72</td>
<td>77.74</td>
<td>80.14</td>
<td>80.12</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>68.72</td>
<td>75.41</td>
<td>80.13</td>
<td>80.01</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>68.49</td>
<td>77.74</td>
<td>79.18</td>
<td>79.04</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>68.49</td>
<td>75.41</td>
<td>79.18</td>
<td>79.01</td>
</tr>
</tbody>
</table>
5.4.2 Fusion of Diffusion with Visual and Textual Information

This section explores fusion of diffusion with visual and textual information. For image retrieval applications, images in the database are often provided with textual information. For example, each image in NAA dataset is associated with metadata consisting of a few sentences. These sentences usually incorporate some high-level semantic information. Hence, it provides a chance to explore the fusion of visual and textual information. This section proposes a method to fuse visual and textual information for diffusion, which can achieve better results than using single information.

Proposed Method

Given a dataset of \( n \) images with visual features \( \chi^v := \{x^v_1, ..., x^v_n\} \subset \mathbb{R}^{d_v} \), where \( d_v \) is the dimension of visual features, and their associated textual feature \( \chi^m := \{x^m_1, ..., x^m_n\} \subset \mathbb{R}^{d_m} \), where \( d_m \) is the dimension of textual features. This thesis defines the similarity between two data points \( x = \{x^v, x^t\} \) and \( z = \{z^v, z^t\} \) with both visual and textual local constrains. Let

\[
s_{k_v, k_t}(x|z) = \begin{cases} 
  s(x^v, z^v), & \text{if } x^v \in \text{NN}_{k_v}^v(z^v) \text{ and } x^t \in \text{NN}_{k_t}^t(z^t); \\
  0, & \text{otherwise}
\end{cases} 
\]  

(5.11)

be the similarity of \( x \) with respect to \( z \), in which \( \text{NN}_{k_v}^v(z^v) \) denotes the set of the \( k_v \) nearest neighbours of \( z^v \) in \( \chi^v \) and \( \text{NN}_{k_t}^t(z^t) \) is \( k_t \) nearest neighbours of \( z^t \) in \( \chi^t \). Also, \( s(x^v, z^v) \) can be calculated by inner product of \( x^v \) and \( z^v \). Then

\[
s_{k_v, k_t}(x, z) = \min\{s_{k_v, k_t}(x|z), s_{k_v, k_t}(z|x)\} 
\]  

(5.12)

An affinity matrix \( A \in \mathbb{R}^{n \times n} \) constructed by

\[
A_{ij} = s_{k_v, k_t}(x, z), \quad \forall (i, j) \in [n]^2, \quad [n] = 1, 2, ..., n.
\]  

(5.13)

This thesis normalises the affinity matrix by \( S := D^{-1/2} AD^{-1/2} \), in which \( D \) is degree matrix of \( A \). Then diffusion process or random walk is performed by solving

\[
f^* = (1 - \alpha)L_\alpha^{-1}y
\]  

(5.14)

in which \( L_\alpha = I_n - \alpha S \), and \( \alpha \) is a jump probability in random walk which is smaller than 1 and larger than 0. Based on the similarity measurement in Eq.(5.12), given a query image and its associated data point \( q, y_i \), the \( i \)th element of its query score
y_i is estimated by

\[ y_i = s_{k_v,k_t}(\mathbf{x}_i, \mathbf{q}), \forall i \in [n] \] (5.15)

Finally, \( f^* \) is obtained as a vector of new scores for ranking.

For regional diffusion, \( y_i \) is estimated by a set of query data points \( Q \), which are the \( Q \) are k-nearest neighbour image regions of the given query image.

\[ y_i = \sum_{q \in Q} s_{k_v,k_m}(\mathbf{x}_i, \mathbf{q}), \forall i \in [n] \] (5.16)

**Experiments**

The experiments are conducted on NAA29k dataset. NAA29k provides two test settings with 100 queries or 1k queries. To compare with the proposed method, several baselines are also introduced. Firstly, textual and visual information are separately utilised. Given textual information, TF-IDF (term frequency-inverse document frequency) [Joa96] is applied. Given visual information, spatial search [RSMC15] and regional diffusion [ITA+17] are applied. Secondly, textual and visual information are both utilised. Re-ranking method is applied by integrating textual and visual information. In specific, given a query image and \( n \) gallery images in datasets, the TF-IDF method provides a textual ranking list (well-sorted) \( R^t = \{R^t_1, R^t_2, ..., R^t_{k_t}\} \), where each element \( R^t_i \) is an image from the dataset, and \( k_t \) is the number of top ranked images (set as 500). And regional diffusion method provides a visual ranking list (well-sorted) \( R^v = \{R^v_1, R^v_2, ..., R^v_{k_v}\} \), where each element \( R^v_i \) is also an image from the dataset, and \( k_v \) is the number of top ranked images (set as 500). The \( R^v \) is shorten by removing its elements, which are not appeared in the list of \( R^t \) while keeping the relative order of the elements retained in the list \( R^v \) unchanged. Then, a re-ranking list is obtained. This re-ranking method fuses both visual and textual information by regarding textual information as qualitative factor, and regarding visual information as quantitative factor.

For visual features, this thesis adopts VGG16 model pre-trained on ImageNet, and uses MatConvNet [VL15] to extract the feature of Conv5_2 layer followed by 2-by-2 grid max-pooling. After extracting all the regional features, this thesis follows the pipeline proposed in [RSMC15] as post-processing step. All features are L2 normalised, PCA-whitened with reduced dimension as 512, and re-L2 normalised. This thesis also defines the affinity function using a monomial kernel as \( s(\mathbf{x}, \mathbf{z}) = \max(\mathbf{x}^T \mathbf{z}, 0)^3 \). The diffusion parameter \( \alpha \) is 0.99. The default values of parameters utilised in Eq.(5.12) are 200 for \( k_v \) and 500 for \( k_t \).

**Results**

The results of the performance are summarised in Table 5.8. For methods that only use textual information, the performance of TF-IDF is 50.47 percentage
Table 5.8: Visual and textual information fusion results.

<table>
<thead>
<tr>
<th>Method</th>
<th>100 queries</th>
<th>1k queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual: TF-IDF [Joa96]</td>
<td>50.47</td>
<td>48.55</td>
</tr>
<tr>
<td>Visual: Spatial Search [RSMC15]</td>
<td>70.77</td>
<td>59.93</td>
</tr>
<tr>
<td>Visual: Regional Diffusion [ITA+17]</td>
<td>78.19</td>
<td>68.67</td>
</tr>
<tr>
<td>Textual and Visual: Re-ranking</td>
<td>82.66</td>
<td>74.36</td>
</tr>
<tr>
<td>Textual and Visual: (Proposed)</td>
<td><strong>84.04</strong></td>
<td><strong>76.85</strong></td>
</tr>
</tbody>
</table>

points for 100 queries and 48.55 percentage points for 1k queries, which is about 10 to 20 percentage points lower than the methods applied with visual information. For visual-only methods, the regional diffusion method performs better than the spatial search on both 100 queries and 1k query settings. A re-ranking method, by integrating textual and visual information, can further enhance the regional diffusion method by an increase of nearly 5 percentage points. It indicates that visual and textual information are complementary. What’s more, the proposed method achieves the best performance with 84.04 percentage points for 100 queries setting and 76.85 percentage points for 1k queries setting, which exceeds the re-ranking method about 2 percentage points. This result verifies that the proposed method has designed a better way to fuse the visual and textual information within diffusion process than simply fusing the visual and textual rankings.

5.5 Conclusion

Utilising the modelling capability of deep neural networks, this thesis assimilates the effect of diffusion process into new feature representation, achieving similar or better retrieval with simple Euclidean search. Also, it gives an unsupervised framework to bootstrap image retrieval by exploiting the manifold structure of the images in a database. It effectively improves retrieval without the aid of additional labels or external datasets. Experimental study on benchmark datasets demonstrates its effectiveness and advantages. This thesis also explores the fusion with diffusion in two different perspectives.

This thesis takes a database-specific approach by assuming access to a database in advance. How to generalise it to unseen databases will be an interesting issue to explore in future work. It is believed that training it with a sufficiently large and generic database will enhance its generalisation capability to some extent. Also, due to its database-specific characteristic, the feature representation learned by the proposed approach on one database may not be effectively applied to another database of a significantly different nature. This issue will also be addressed in future work. Integrating the proposed approach with domain adaptation and transfer learning
techniques could be a potential solution.
Chapter 6

NAA Image Retrieval System and Related Applications

This chapter mainly introduces a practical application scenario of image retrieval, the National Archives of Australia Project (NAA). The thesis introduces the background of NAA image retrieval problem, NAA29k dataset collection and labelling, retrieval system construction, and the related applications.

6.1 Background of NAA Image Retrieval

The NAA image retrieval project has been accompanied by a background project of this Ph.D. program. NAA staff collected many historical images for recording the history of Australia. They would like to have a retrieval system that helps staff and public users explore these images by looking for what they are interested in (such as important events of the US president’s visit to Australia). Currently, the image retrieval system of NAA is based on a TBIR system. As a result, it is difficult to find similar images through image content, style, event, or scene. The task of this project is to build a CBIR system for NAA. This retrieval system is expected to effectively serve NAA staff and be further open to the public for a better understanding of the history of Australia.

The initial work of the project was completed from 2011 to 2012. A BoW-based CBIR image retrieval system is basically built, as shown in Fig. 6.1. Several subsequent processing modules are then included to enhance this traditional CBIR framework, forming an early NAA image retrieval system, as shown in Fig. 6.2. The system achieved good results at that time. After 2013, the thesis started to take over this project. On the basis of this early CBIR system, the thesis gradually began to apply deep learning to NAA image retrieval. During the research work, the thesis identified the following issues.

1) Measurement issue. It is hard to measure the quality of the retrieval results since there is no ground truth for NAA. The debugging and improvement of the algorithm usually requires visual checking, and the subjectivity is strong; 2) Effective deep features. With the advent of deep learning, the thesis finds that a simple way to utilise deep features (i.e., no need of the subsequent processing modules) can usually achieve better results than the BoW based design. For example, the retrieval performance based on the fc6-layer feature of VGG16 is much better than
the BoW-based retrieval performance through visual checking; 3) Deep feature’s limitation. The thesis can still find some unsatisfactory retrieval results even if using deep features. For example, in the retrieval result for a given query image, some irrelevant images could appear between two relevant images, as demonstrated in Figure A.1. At the same time, pre-trained CNN is still inadequate. It still lacks sufficient feature invariance with respect to illumination change, large viewing angle change, and complicated non-rigid change. Particularly, it does not work well for scenes, events, and people-related images in the NAA collections.

These issues need to be addressed by developing better features and metrics in order to build up an effective retrieval system.

6.2 NAA29k Dataset Collection and Annotation

Building a good measurement of retrieval performance relies on the collection and labelling of NAA datasets. Its difficulty lies in how to judge the concept of similarity
and the intensive labour associated with it. The similarity is a subjective matter. Just as one thousand readers have one thousand Hamlets, it is also subjective to judge whether two images are similar or not. Usually, image retrieval applications apply a manual annotation method to build a test set to evaluate the performance of retrieval method. The NAA dataset covers more than 340,000 unlabelled images. As the research progresses, this thesis increasingly needs an evaluation dataset to index and test.

From the perspective of developing the NAA project, this thesis needs to compare and evaluate the performance of various algorithms. However, the entire NAA database contains 340k images, which for this thesis is too large to be exhaustively manually labelled. Therefore, this thesis selects a part of images (29k) for labelling, and a total of 600 query images are finally selected. The thesis will release this NAA29k dataset for peer researchers.

6.2.1 NAA29k Dataset Annotation

In this section, the thesis introduces the information about the NAA29k dataset. Here are a few key data annotation steps:

1) Feature extraction and retrieval. 1137 images are randomly selected from the whole collection as initial query images. The feature of the last layer of VGG16 is extracted for the whole 340k images and l2-normalised. For each image in the 1137 initial queries, the whole 340k images are ranked by using Euclidean distance. And only the 1000 top ranked images are kept to find the truly similar images by manually visual checking.

2) Manually visual checking. A human observer needs to visually check the 1000 top ranked images and judge whether it is similar to the query or not. Those dissimilar images, judged by the observer, will be removed from the retrieval list. Most of the query images can finally have 7 to 19 truly similar images. These images are regarded as ground truth for the given query. Note, that this step is highly labour-intensive, and this thesis therefore only selected about 1137 query images.

3) Removing easy queries. Using the ground truth, the average precision (AP) value of each query image is calculated. Part of the query images are further removed since the images similar to them can be easily retrieved, that is, their AP values are very close to 1.00. Only 600 queries are kept after this step.

4) Obtaining 29k dataset. Putting the 600 query images and their associated top 1000 ranked similar images together, a dataset of 28,913 images is finally obtained, called NAA29k dataset. And it has the ground truth for evaluating the performance of retrieval algorithm is labelled in step 2.
6.2.2 Properties of the NAA29k Dataset

The NAA29k dataset is different from most existing retrieval benchmark datasets. It is naturally formed and contains a variety of objects, scenes, locations, and activities from various periods in Australia, rather than being manually collected based on certain specific concepts. For example, the Oxford5k and Paris6k datasets are building-related datasets; the INRIA Holiday dataset is collected during the INRIA researchers’ travelling, and the INSTRE dataset is a collection of 200 object instance-related images and extended by web-based retrieval. The NAA29k dataset covers a wide range of concepts including buildings, cars, portraits, flowers, crowds, events, celebrations, sports, birds, animals, urban landscapes, books, factories, trains, beaches, sea, table, office, and plane, etc. The NAA29k is very diverse and unbalanced compared to other manually acquired image datasets. This makes the retrieval task more challenging. Also, the NAA29k dataset does not pro-
vide bounding box based annotation. This is because most of the images in the NAA29k are related to scenes, places, events, or characters, making it very difficult to determinate bounding boxes clearly. This is different from instance-level image datasets.

6.3 NAA Image Retrieval Web Platform and Software

Designing an image retrieval web platform and developing image retrieval software is a great approach to converting scientific research into real products. In particular, the thesis builds an image retrieval web platform and develops the image retrieval software for NAA, which could also be potentially used for other datasets. The image retrieval web platform is mainly used to display the results of image retrieval algorithms and serve public users to retrieve images. The NAA image retrieval software is mainly used by professional archival researchers and managers to process image information and obtain image retrieval results. The author’s own contribution lies on the design and development of the web platform. Also, the author makes efforts in the establishment of the version 1.0 software with core retrieval functions embedded. The author would like to thank Chao Wang, Ian Cormor, and Kevin Zhang, who have given nice suggestions and made efforts in updating the design of the NAA image retrieval web platform. At the same time, the author would like to thank Song Liu to extent the software to version 2.0 by adding various user-friendly functions, improving the software efficiency, and optimising the software interface.

6.3.1 Image Retrieval Web Platform

Image retrieval web platform is an important application tool for researchers who study image retrieval. It cannot only be regarded as a tool to visualise the retrieval result but also inspire researchers to design new algorithms. From some wrongly retrieved images, the shortcomings of the methods utilised behind the ranking result can be identified. At the same time, the image retrieval web platform can also serve public users to retrieve whatever images they are interested in. Surfing through the image retrieval web platform can help public users understand the whole dataset and get more information about the history of Australia.

The NAA image retrieval web platform, designed in this thesis, is shown in Fig. 6.4 and Fig. 6.5. It consists of two types of web pages. One is the main page, shown in Fig. 6.4 and another is the page to display retrieval result, shown in Fig. 6.5. In the main page, it provides a login function to authorise users to visit the retrieval system. All images in the main page are randomly selected from the
NAA dataset and can be different for different login sessions. By clicking one of the images (regarding it as a query) in the main page, it can jump to another page, which shows similar images to this query image as well as the metadata of this query image. Users can find more similar images by clicking the left and right arrow buttons in this page.

The current version of NAA image retrieval web platform focuses on retrieving images within the NAA dataset. To ensure retrieval efficiency, the server directly stores the static ranking result for every image in NAA and does not need repeatedly to perform feature extraction or similarity calculations. Meanwhile, this system can be readily extended to support users to upload any new image as query to conduct retrieval. To reach the goal of retrieving with any query image, current NAA image retrieval web platform can be updated with a more capable server with sufficient computational resource to make the online retrieval fast enough.

### 6.3.2 NAA Image Retrieval Software

Although NAA image retrieval web platform could be satisfactory for public users, NAA staffs expect more professional software to help them manage the NAA collections. To build a good NAA image retrieval software, this thesis has made efforts in a series of related research and knowledge reserves, such as CNN feature extraction, retrieval system demonstration, deep learning algorithm study, and image annotation. The following part describes the version 1.0 NAA image retrieval software, which mainly implements the core algorithm functions. Optimisation of the interface of the software will be carried out by professional software developers.

The NAA image retrieval software implements automatic retrieval progress based on textual information and visual information. The software interface is shown in Fig.6.6. The flowchart and basic modules of the NAA image retrieval software are described below, and shown in Fig. 6.7.

The software consists of three modules: 1) The textual information processing module, that is, the TFIDF module, which indexes each image by textual information; 2) CNN feature extraction, namely FeaExtra module, which extracts all image features and supports AlexNet, VGG, and other deep neural networks. For the specific process, see Fig.6.8; 3) Re-indexing module that re-indexes each image with visual information and textual information.

The main contribution of this software lies in the following parts. Firstly, a self-developed feature extraction module is built with 3,000 lines of C++ code. This module is implemented by designing underlying functional modules, defining the convolutional neural network, realising different layer, and completing the network forward propagation function. Compared to current feature extraction progress
Figure 6.4: NAA image retrieval web platform. It provides a login function to authorise users to visit the retrieval system. All images in the main page are randomly selected from the NAA dataset and can be different for different login sessions. Users can explore the NAA collections by clicking any image, being regarded as a query, and it will jump to another page showing similar images to this query.
**Figure 6.5:** NAA image retrieval web platform with a retrieval example. The top left image is regarded as a query image, as well as the metadata of this query images on the top right. The similar images are shown on the bottom. Users can find more similar images by clicking the left and right arrow buttons in this page. Users can further explore the NAA collections by clicking any similar image appeared in this page, and it will jump to a new page by regarding the clicked image as new query.
which only needs 10-line codes. This kind of module still has its significance. It is an independent module with high autonomy, and does not rely on any third-party modules or platforms such as Caffe, Torch, Tensorflow, Matconvnet, PyTorch, and so on. This is very important for commercial use that has a different nature from academic research. Secondly, this FeaExtract module can suit various CNN models, which provides the potentiality for a wide range of application scenarios. A user manual for NAA image retrieval software is provided in Appendix.

6.4 Discovering and Gathering Region of Interests for NAA Users

The current version of NAA software is functionally insufficient for users. It can only provide a ranking list for a query image (the retrieval is based on the visual feature of the whole image, as well as in the database side) by a certain retrieval method. However, users may be interested in various parts of an image only. For example,
users may be interested in a building in an image, other objects like mountains, rivers, and trees. The user may want to find the images having the similar buildings but not the mountain, river, or trees. This part of image (i.e., building) can be called regions of interests. This application scenario makes the discovery of regions of interests important. The discovered regions of interests can be fed into the existing retrieval loop as a new query to obtain a ranking list of similar images. Many current retrieval systems simply leave this task to users by providing a tool to manually crop a regions of interests. However, this section tries to provide an alternative way by using automatic method to discover the regions of interests for NAA users. This will enhance the efficiency of the retrieval system and make it more user-friendly. The following part will introduce some approaches to realise this idea, especially based on the off-the-shelf methods. What’s more, a user may be not only interested in a single region of a given image but also would like to comprehend all the interesting regions in the whole dataset. This urges the thesis to provide a method to gather the regions of interests.

Discovering the region of interests is close to three computer vision tasks, object detection, caption generation, and clustering. Given one image, object detection is a task to find an object consisted in the given image and predict the bounding box as well as the object category; caption generation cannot only predict the bounding box but also generate one sentence to describe the context appeared in the predicted regions of the given image. Opposite to focusing on one single image, images clustering, especially images patches clustering, can gather similar images together and effectively organise the whole image dataset.

Benefited from many off-the-shelf methods on object detection, caption genera-
Figure 6.8: CNN feature extraction flow chart. It adopts AlexNet struture with five convolutional layers and two fc layers. In each layer, it contains various procedures like Im2Col procedure, Conv procedure, ReLu procedure, Pool procedure, Norm procedures, and Fc procedure. Since this software is established for commercial purpose, all of the network struture, layers, as well as these procedures, are re-designed and re-implemented from raw C++ codes without using any package from existing deep learning platforms.
tion, and images clustering, it is possible to provide the regions of interests for NAA users. This function could be added into a future version of NAA software as an alternative approach to the cropping tool. Also, it may provide a full view for the NAA users to gather the regions of interests.

6.4.1 Discovering with Object Detection

The first attempt is to detect objects in the NAA29k. Given one image, object detection can predict many bounding boxes as well as their object categories. The thesis uses the pre-trained faster-rcnn model [RHGS17], which is applied with AlexNet trained on VOC2012 datasets. Part of the detection results are shown in Fig. 6.9. For each image, several yellow bounding boxes are marked with some predicted categories. These bounding boxes can be treated as the potential regions of interests and provided to the users. By visually checking, it can be found that these regions are satisfactory and they almost contain all the important objects appeared in the image. For example, the third left image of the first row in Fig. 6.9 has correctly found all five persons in the image. Then these bounding boxes (regions of interests) can be provided to the user for their next round of retrieval.

However, if looking into the detection performance in terms of category correctness, it is not very good. The faster-rcnn model (trained on the 20 categories in VOC2012 datasets) can detect various objects, like train, person, boat, and car. Many of them are correct if the objects belong to these 20 categories, like people or boats. Some of them will become incorrect because the VOC2012 detection database only contains a limited object categories. Some detected objects are out of these 20 categories, such as buildings or landmarks.

The thesis notices that the object detection could be important and helpful for image retrieval tasks. Many datasets, like Oxford5k and Paris6k datasets, provide bounding boxes (regarded as manual object detection) for query images to reduce the noise from the image background, and the retrieval performance will be improved when using the part of query image in the bounding boxes rather than the entire query image. However, NAA dataset is a hybrid of instance and scene images and there are no bounding boxes provided. At the same time, if a well-designed detection algorithm is applied on the database side, the image retrieval performance will also be improved because the adverse impact of the background and other noisy image areas is largely removed. Detection techniques are important as a filtering module in a retrieval system when the query is an instance or object. Although recent detection methods such as fast-rcnn [Gir15] and faster-rcnn [RHGS17] have been well developed, the state-of-the-art retrieval methods still largely use bounding boxes rather than automatically detected image regions. For some datasets, bounding
Figure 6.9: Test results obtained using faster-rcnn on NAA29k. For each image, several yellow bounding boxes are marked with some predicted categories. Some regions prediction are satisfactory. For example, the third left image of the first row has correctly found all five persons in the image. However, some are not satisfactory. For example, the regions prediction for images in the second row are all incorrect.

boxes may be are satisfactory and can be utilised to predict the region of interest. However the category predication is still difficult especially for NAA29k dataset whose visual content is quite diverse. Hence, there is still a gap to push the detection embedded retrieval system into practical applications.

6.4.2 Discovering with Regional Image Caption Generation

Image caption generation is a computer vision task, which focuses on understanding the images. It tries to link the visual part of image context to the language part of a caption. Similar to object detection, image caption generation can also predict bounding boxes. The only difference lies in the prediction of the region. Object detection predicts the category of the region while image caption generation can predict one sentence to describe the region. Regional image caption generation is an extension of image caption generation, and it can provide richer semantic information. Hence, it is a more challenging problem and requires the algorithm to identify the relationship between two objects.
As a state-of-the-art regional caption generation method, DenseCap [JKF16], can predict very rich bounding boxes and their associated captions by providing both visual and caption features for a given image. It utilises a fine-tuned (trained on a dataset [JKF16]) Vgg-16 model. This thesis has conducted an experiment with DenseCap method on the NAA29k dataset. Part of the results is shown in Fig. 6.10. There are four cases. For each case, it contains one image and their associated captions. Each image is marked with several bounding boxes in various colours. And the caption with the same colour is associated to the bounding box. The thesis finds that although the regional image caption generation can give a relatively rich and correct regional description, there are still some difficulties in high-level semantic understanding. For example, in Fig. 6.11, the algorithm can identify two different people in the front, people’s arms, people wearing gray clothes, and people in the background. However, it cannot correctly understand what the two people are doing. In fact, the two aboriginal Australian people are wrestling in the image, which is different from the general wrestling scene. More interestingly, the algorithm incorrectly interprets the foot of an up-side-down person as a hand holding a Frisbee. By tracing the training set used by DenseCap [JKF16] and checking the original training data utilised in DenseCap [JKF16], the thesis has found that a large amount of labelled training data are related to playing Frisbee and it does not contain sufficient wrestling training data. This biased training samples issue may be one of the reasons leading to the wrong prediction. To further show the difficult, the thesis shows the retrieval results for the wrestling query image by NAA software in the bottom of Fig. 6.12. It also shows the ground truth of the same query image in top of Fig. 6.12. It has found that the ground truth images has covered a variety of different wrestling scenes consisting of images with different backgrounds. However, the NAA software can only find one correct similar image, and other incorrect images have either consisted of similar background or crowded people. These results indicate that understanding scene images is still a difficult problem for retrieval.

In short, regional image caption generation can provide a rich prediction of region of interests. However, the caption generated may be not precise or even incorrect. These predicted bounding boxes can be helpful for users’ next round retrieval but the caption information is still not satisfactory for the users to understand the images. Currently predicted captions are too loose or simple when compared to the NAA metadata associated with the images in the NAA collections. Understanding scene-level images is still a difficult computer vision task and linking the predicted caption with the metadata information remains a challenge problem.
**Figure 6.10:** Region-based caption generation for NAA. Here are four cases. For each case, it contains one image and their associated captions. Each image is marked with several bounding boxes in various colours. And the caption with the same colour is associated to the bounding box. Beyond the last case in right-bottom, captions generated for other three cases are satisfactory. In the last case, it fails to predict that two men are wrestling.
Figure 6.11: Region-based caption generation for NAA: an example. Here is an incorrect predicted case for caption generation. It should be that two men are wrestling rather than a man holding frisbee.
(a) Ground truth of a query image of wrestling. Five truly similar images are shown.

(b) Retrieval results obtained by NAA software. Only one similar image is correctly retrieved, others are all incorrect.

Figure 6.12: Comparison of retrieval result with ground truth: a wrestling image sample. The top of the figure is the ground truth of a query image of wrestling. And the bottom of the figure gives the retrieval results obtained by NAA software. The NAA software can only find one correct similar image, and other incorrect images have either consisted of similar background or crowded people. These results indicate that understanding scene images is still a difficult problem for retrieval.
6.4.3 Gathering Regions of Interests with Clustering

Discovering the region of interests is important for understanding a given image. However, to fully understand whole dataset, gathering these regions of interests can be important. This sections introduces an approach to gathering the regions of interests with clustering. In specific, the above mentioned two methods can provide many meaningful proposals or regions of interests. With these regions of interests, a hierarchical clustering method can effectively organise them and reveal useful information.

The following part of the section elaborates the experiment. Firstly, given the input image from NAA29k, the faster rcnn [RHGS17] pre-trained on PASCAL-VOC2007 [EGW+10] is utilised to predict detection boxes. It can predict the $R \times (4K)$ bounding boxes, where $K$ includes the background as the object class used by the pre-trained CNN model. For example, $K$ is 21 for the PASCAL-VOC2007 dataset. $R$ is the number of regions of interests and varies across to different input images. The thesis sorts these bounding boxes according to the predicted scores and only keeps the $K_1$ top-ranked bounding boxes. Here, $K_1$ is set to 50. For other hyper-parameters, the thesis sets the non-maximum suppression NMS threshold to 0.3, which is the default value used by faster-rcnn [PYF]. Secondly, DenseCap [JKF16] can also predict the bounding box when predicting the dense captions. The thesis ignores these caption information, only retains the $K_2$ top-ranked bounding boxes as candidate detection boxes. $K_2$ is also set to 50 and the NMS threshold is set to 0.3. The DenseCap model [DES] is pre-trained on the Visual Genome dataset [KZG+17]. Then, the thesis combines these proposals or regions of interests together. It applies a 2nd NMS with a threshold of 0.75 for all the regions of interests since these regions are generated by two different methods. Finally, the thesis obtains about one million detection boxes for the NAA29k image dataset, corresponding to about one million image patches.

Multi-level clustering is applied to these regions of interests with two stages. In the first stage of clustering, the feature extraction is applied for these one million image patches. For each image in NAA29k, the thesis uses the Vgg16-ImageNet model to extract the conv5.2 layer feature. Based on the bounding box information, a $c$-dimensional feature vector is assigned to each image patch by applying a max-pooling. Here, $c$ is the channel value in the conv5.2 layer of Vgg16, equals to 512. Then k-means is applied to these one million image patches with the software package of vlfeat [VF08]. The number of clusters $k_1$ is set to 100. The thesis randomly selects some of the image patches in the first major category and shows them in Fig. 6.13. It can be found that the clustering granularity of the first stage is still relatively large, and the image patches are relatively mixed.
Figure 6.13: Part of first stage clustering results. The patches, shown in this figure, are randomly selected from the first major category. It can be found that the clustering granularity of the first stage is still relatively large, and the image patches are relatively mixed.
In the second stage of clustering, all the one million image patches are resized into small-sized patches (64 by 64) and features are re-extracted by Vgg16-ImageNet. The feature vector of image patch is obtained by applying max-pooling and L2 normalisation on the feature map of the conv5_2 layer. For image patches belonging to the same large category in the first stage of clustering, the thesis clusters them again by k-means with the number of clusters as 50. Finally, the thesis obtains 100 major clusters and 5000 sub-clusters. Fig. 6.14 and Fig. 6.15 show some of the images in two adjacent sub-clusters. It can be found that most image patches in the 20th sub-clusters are a single person wearing formal suits, while the images in the 21st sub-cluster are mostly multi-person with formal wear. It verifies that image patches after the second stage clustering are much more well-ordered than the images patches after the first stage clustering.

By gathering these regions of interests with clustering, users can get a better picture about the NAA dataset and can quickly go through all the regions of interests rather than image by image. It will both save user’s time and provide a more efficient way to look into the dataset.

In short, by discovering and gathering regions of interests, the NAA system is expected to provide an alternative way to attract user’s attention and an efficient way to browse the full dataset. It can facilitate the users to explore their own interests and save their time.

6.5 Summary

This chapter mainly introduces a real application scenario of image retrieval. The thesis has introduced the background of NAA image retrieval problem, the collection and labelling of NAA29k dataset, system construction, and related applications. The NAA project has been the background project of this thesis. Conducting research on a project based on real application scenarios is an effective to stimulate new ideas and new algorithms. At the same time, it also helps to understand the research field more completely and deeply, to transform ideas to algorithms, and to implement them into software systems that can be widely used.
Figure 6.14: Second stage clustering results, sub-cluster 20. The patches, shown in this figure, are randomly selected from the sub-cluster 20. Most patches in this sub-cluster are single person wearing formal suits.
Second stage clustering results, sub-cluster 21. The patches, shown in this figure, are randomly selected from the sub-cluster 21. Most patches in this sub-cluster are multi-person with formal wear.
Chapter 7

Conclusion

7.1 Summary

The thesis has done the following work in model analysis and algorithm design:

(1) Firstly, several unsupervised deep learning models were analysed. Under unsupervised setting, the study compared the deep learning models with the classical BoW model. Typical methods include the BoW model (based on hand-crafted features) and several typical deep learning modes other than convolutional neural networks. Labels used in supervised feature learning requires a large amount of intensive labour and consumes a lot of time. They are usually manually annotated by experts and are expensive. This thesis therefore focuses on unsupervised feature learning in this part. This thesis compares the characteristics of these two typical models by the clustering algorithm. Experimental results show that the BoW model usually performs better than the deep learning model when there are only a limited number of training samples.

(2) Secondly, an improved spatial retrieval algorithm is proposed. As the most advanced ConvNet-based image retrieval method, the spatial search [RSMC15] has shown excellent retrieval performance and is superior to other competitors. A key component of the method is a weighted combination of the estimated distances of different regions of the retrieved image. However, these weights are currently manually adjusted through an exhaustive search based on trial and error. This not only causes a lengthy parameter adjustment process but also makes it difficult to guarantee the optimality of the tuning weight. Moreover, these weights may not be suitable any more when the nature of the image dataset changes. In order to improve this situation, this thesis proposes to automatically learn the combination weight based on the retrieval of the basic facts. Specifically, this thesis develops a method called semi-supervised weight learning (SWL), based on the framework of distance metric learning. In addition to using the basic facts to generate triplet constraints, this thesis also uses unlabelled images to generate a large number of unsupervised constraints to stabilise the learning process and improve learning efficiency. By combining with the latest linear support vector machine solver, an effective algorithm is proposed to solve the large-scale optimisation problem. The experimental results of three benchmark datasets and the newly collected archival photo dataset demonstrate the validity of the proposed weight learning approach. It achieves comparable or better retrieval performance than manual adjustment methods, especially on new
archival photo datasets. Then, this thesis extends the SWL method with the kernel instead of the Euclidean distance.

(3) Thirdly, this thesis proposes a diffusion process modelling method based on deep learning. The diffusion process, by exploiting potential neighbourhood data structures, has proven to be an effective mechanism for improving image retrieval. Recently, the diffusion process of image retrieval using the deep image feature representation [ITA+17, BZW+19] is a good example to show the effectiveness of the diffusion process for image retrieval. However, the diffusion process stores large neighbourhood maps, takes more online retrieval time, and requires special algorithms in addition to simple Euclidean search. In order to solve these problems, this thesis proposes to treat the diffusion process as a “black box” and model it directly by training deep neural networks to obtain better image representation, assimilate the effects of the diffusion process, and make Euclidean Search effective. This thesis first proposes the kernel mapping interpretation of the kernel diffusion process and then represents the modelling as a deep quantity learning problem. The proposed method is unsupervised because it requires neither image tags nor external datasets, and completely avoids the online diffusion process in retrieval. At the same time, after understanding the effectiveness of the diffusion process on image retrieval, how to design a better diffusion process method becomes an important research direction. In recent years, from the perspective of data fusion, methods for integrating multiple diffusion processes have been proposed and studied. The NAA project has always been the driving force behind the problem of image retrieval in this thesis. It has a good application scenario and practical significance. Another major content is the exploration of a diffusion-fusion framework based on the NAA database. Around the NAA database, this thesis explores the fusion framework based on the diffusion process from two different perspectives in order to better achieve the effects of the diffusion process: 1) global and regional diffusion fusion; 2) diffusion fusion of visual and textual information.

(4) Finally, at the application aspect, a real application scenario of image retrieval is introduced, called the National Archives of Australia (NAA) Image Retrieval Project. This thesis introduces the background of NAA image retrieval problem, the collection and labelling of NAA29k dataset, system construction, and the related applications. Conducting research on a project based on real application scenarios is an effective to stimulate new ideas and new algorithms. At the same time, it also helps to understand the research field more completely and deeply, to transform ideas to algorithms, and to implement them into software systems that can be widely used.
7.2 Future Work

In future work, there are several directions worthy to explore:

(1) Unsupervised learning requires further study and analysis of the latest deep models. The research interest of deep learning evolves from early unsupervised methods to later supervised methods with CNN as the mainstream, and then moves to the recent unsupervised generative adversarial networks (GAN), showing an alternate upward trend. Unsupervised learning has once again been paid attention. It reflects the limitations of CNN networks, and it also puts forward higher requirements for deep learning. This thesis does not study the related field of GAN. Future work can conduct more research on GAN.

(2) This thesis proposes a semi-supervised weight learning method for automatically learning combination weights for image retrieval in the spatial search method. While the work in this thesis currently focuses on learning weights by using off-the-shelf ConvNet features, the weight learning framework based on distance metric learning in this thesis can be easily extended to a joint way of fine-tuning ConvNet and learning spatial weights at the same time.

(3) Using the modelling capabilities of deep neural networks, this thesis converts the effects of the diffusion process into new feature representations, achieving similar or better retrieval through simple Euclidean search. This thesis uses a database-specific approach by assuming the access to the database. How to promote it to an unseen database will be an interesting question in future work. Training it with a large enough and versatile database will enhance its generalisation ability to some extent. Moreover, due to its database-specific characteristics, the feature representations of the proposed method learned on one database may not be effectively applied to another database with significantly different properties. This issue will also be addressing in future work. Combining the proposed approach with transfer learning techniques may be a potential solution. At the same time, the fusion framework of the diffusion process weights needs to be further tested and validated on more datasets.

(4) The NAA retrieval system needs to be further improved. The next step is to integrate the latest retrieval algorithms into the system and simultaneously carry out the related applications.

In addition, by looking at the whole picture of image retrieval, this thesis presumes that the following items could possibly indicate the next big things for this research field. Firstly, unsupervised learning methods may become more popular in the next few years, pushing the image retrieval techniques to be more effective and efficient on practical applications. Secondly, image retrieval techniques on mobile devices and platforms will be largely investigated with the pervasive use of smart phones and advent of 5G technique. Lastly, image retrieval techniques on auto-
driving devices and robots will be highly needed because place-based localisation, mapping and navigation plays a key role for these applications.
Bibliography


[JDS08] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In


Appendix A

NAA Image Retrieval Software User Manual

To run the software, simply: 1) Double-click the “Retrieve” application file. 2) Enter or select a text data folder, an image folder, and an output folder. 3) Click start and run.

There is no need to install the NAA File Photo Retrieval System. Just extract the .zip file and double click. EXE file to retrieve it. All other dependent files are provided, including some dlls and parameters. Depending on the file, the software uses OpenCV (version 2.4.8) and some “dll” of openblas to read the image and perform a matrix calculation. They are all open source and free. In addition, for image feature extraction, this thesis uses ImageNet trained CNN parameters, which is also free for commercial properties. They are stored in a folder called “parameter”.

The manual also provides a “sample” folder to help the users become familiar with the software. Before the users try a lot of image retrieval. This thesis suggests that first testing it on this example. All the users should provide is listed: 1) A folder contains all text data, such as the folder “./example/in_text”; 2) A folder contains all images, such as the folder “./example/in_image”; The folder where the results are saved, such as the “./example/out” folder.

Note on input and output input: 1) Image and text data should be well matched according to their barcode. For example, given a text data “801004”. In the text data folder, if the barcode is “801004”, there should be “801004.jpg” in an image folder and vice versa. Otherwise, if text data is missing, the system will stop and report the error. 2) For image folders and text data folders, it does not support subfolders, which means the users should put all the images in one image folder and all the text data in another text data folder.

The output folder will contain the “final_result” folder and “imageindex.txt” file. You should only care about the final result. See the details in the “./example/out” example. 1) “imageindex.txt” is an image index used to find the final result, which is to index all image barcodes using numbers (from 0 to total image file number - 1); 2) “final_result” folder, which contains the retrieval results. For each “txt” file, it contains the top 1000 most similar in text data and visual information. In each file, it contains two lines. The first line is the query image index. The second line contains the first 1000 image indexes. They are separated by a “:”. To view image barcodes, use the index in “imageindex.txt”. 3) Some log files, “log.txt”,
“log_text.txt”, “log_image.txt” and “log_reindex.txt” will be created automatically to record the running process of each dll. If the system crashes, they will be helpful to find errors.

The sample folder contains 1284 images and text data. To retrieve them, it takes about 5 minutes on a PC with Intel Core i5-2500 CPU @ 3.30GHz. The thesis also tests it on larger data sets of 10,843 image and text data. It takes about an hour. When using this software to retrieve large data sets, such as 340k. It takes a few days to complete all the steps, depending on the PC. To see the flow when running the software, the users can check the file number in the temporary file folder of the output directory. When the users run the application, two additional folders called “metadata_result” and “images_features” will be created, which will be created and automatically deleted if the system has been completed. The most time-consuming step is the second step: feature extraction step. In general, it can process approximately 15K images per hour.
Figure A.1: An example to show the unsatisfactory retrieval result when using deep features. As seen, for the given query, some irrelevant images (i.e., without showing the same textured floor) mix with relevant images.