Exploring deep learning for intelligent image retrieval

Chen, W.

Chapter 1

Introduction

1.1 Background

The use of images to convey information is playing a crucial role in our daily life in diverse areas such as medicine, journalism, advertising, design, education, and entertainment. The rapid growth of images has been facilitated by numerous technologies ranging from fast Internet to digital cameras and smartphones to inexpensive solid state storage. For example, in 2020, the number of smartphone users worldwide surpasses 3.6 billion and is forecast to further grow by several hundred million in the next few years. It is straightforward to capture and store the visual imagery. However, searching or retrieving the image is difficult because computers do not understand the relationship between the image pixel representations and high level human concepts. Image retrieval has become a major research area worldwide. Given a query image that describes the user's needs, image retrieval is the process of searching for semantically matched or similar images in a large image dataset by analyzing their visual content, for example, a zoologist would look for the photograph of a particular animal, and so on.

To enable accurate and efficient retrieval in massive image collections, compact and rich image feature representations are at the core of successful image retrieval. In the past two decades, remarkable progress has been made, focusing more on using primitive colour features, texture features, and shape features [1, 2, 3]. These low-level features heavily depend on hand-engineered feature descriptors such as Scale-Invariant Feature Transform (SIFT) [4], Speeded Up Robust Features (SURF) [5], and Histogram of Oriented Gradients (HOG) [6]), middle-level representations such as bag of words (BoW) [7] and Fisher Vector (FV) [8, 9]. However, the low-level image features do not have sufficient discriminatory power to effectively bridge the semantic gap, which characterizes the difference between the high-level concepts of humans and the low-level features typically derived from images [10]. In recent years, deep learning, particularly the influential ImageNet [11] and the Deep Convolutional Neural Network (DCNN) AlexNet [12] enables learning powerful feature representations with multiple levels of abstraction directly from data. Deep learning techniques have attracted enormous attention and have brought about considerable breakthroughs in various computer vision tasks including image classification [12, 13, 14], object detection [15, 16, 17, 18], semantic segmentation [19, 20, 21], and image generation [22, 23, 24]. This has also directly impacted the field of image retrieval [25, 26, 27, 28, 29, 30, 31, 32]. DCNNs can learn high-level semantic-aware features to perform retrieval tasks. For instance, current literature may directly use a pre-trained deep model to extract features for input images. Likewise, visual features can be captured from the sub-patches of images to satisfy instance-level image retrieval. Different layers (*e.g.* convolutional layers and fully-connected layers) can capture different levels of features: global features (by fully-connected layers) and local features (by convolutional layers). Model-level and layer-level feature fusion schemes are further studied to incorporate multi-scale deep features by using

methods such as concatenation and weighted averaging. Despite these strategies to improve the performance of image retrieval, there exist many other problems to be addressed, which is the core of this thesis.

1.2 Research Questions and Contributions

Semantic information that helps us understand the world usually comes from different sensory modalities. In the real world, we can express the same concept using different ways: for example by using writing text or by taking a picture. People can search for images of interest by submitting any media content at hand (*e.g.* a word, a phrase, or a sentence) as the query item. Cross-modal retrieval, as a natural search method becomes increasingly important to augment image retrieval. For cross-modal retrieval, although raw data from two modalities (*e.g.* an image and a text) have similar semantic concepts, feature vectors extracted from these data are distributed in different spaces so that their semantics are not well associated. Therefore, these vectors are not directly comparable, leading to inconsistent distributions. This is what is referred to as a heterogeneity gap. The challenge of performing cross-modal retrieval lies in how to measure the semantic similarity, which is defined as a metric to quantify the likeness of two items. In the context of cross-modal retrieval, these items (image and text data) are from different modalities. In terms of this challenge, we come to our first **Research Question 1** (**RQ 1**): **How to learn a commonly shared embedding space for visual modality and textual modality to reduce the heterogeneity gap?** Focusing on the **RQ 1**, we explore cross-modal retrieval by using information entropy to measure the domain uncertainty of the shared space for visual and textual modalities. Furthermore, we combine information theory and adversarial learning into an end-to-end framework. This work is the first to explore information theory in reducing the heterogeneity gap for cross-modal retrieval, a method that is beneficial for constructing a shared space for learning commonalities between cross-modal features. In addition, we introduce a regularization term based on Kullback-Leibler (KL) divergence [33] with temperature scaling to address the issue of data imbalance. Our contributions are based on the following publications [34, 35]:

> **Chen, W.**, Pu, N., Liu, Y., Bakker, E. M., and Lew, M. S., "Domain Uncertainty Based On Information Theory for Cross-Modal Hash Retrieval." IEEE International Conference on Multimedia and Expo, 2019, pp. 43-48.
>
> **Chen, W.**, Liu, Y., Bakker, E. M., and Lew, M. S., "Integrating Information Theory and Adversarial Learning for Cross-modal Retrieval." Pattern Recognition, 2021, 117, pp. 107983.

The existing fine-grained datasets for image retrieval have limited categories to support a retrieval system. As noted, the wide popularity of mobile devices make large

image collections available. However, the deep models are only trained and validated on these limited image categories. The retrieval ability of deep models is limited to the existing categories, but cannot be extended to some of new incoming categories. Unlike the continuous learning process of human beings, deep neural networks suffer from a catastrophic forgetting problem [36], a phenomenon that occurs when a network is trained successively on a series of new data and the learning of this data degrades the performance on previous data. To satisfy domain context constraints, the deep models are required to learn on a series of categories sequentially. Thus we come to our second **RQ 2: What kind of knowledge is more beneficial, for a deep model to learn fine-grained categories incrementally, in order to reduce catastrophic forgetting of previous data?** We extend fine-grained image retrieval in the context of incremental learning and explore this question in the following publication [37]:

> **Chen, W.**, Liu, Y., Wang, W., Tuytelaars, T., Bakker, E. M., and Lew, M. S., "On the Exploration of Incremental Learning for Fine-grained Image Retrieval." The British Machine Vision Conference (BMVC), 2020, pp. 1-10.

To the best of our knowledge, this is the first work to study this problem. Furthermore, incremental learning has been employed to make deep models learn on old data and newly added data successively. However, as incremental learning proceeds, each training session produces a specific model. Saving this stream of models is memory-consuming. Thus, we come to our third **RQ 3: How to utilize the model stream in incremental learning to transfer previously learned information to the deep model trained on the new data?** We investigate this question by proposing a feature estimation method for incremental fine-grained image retrieval, which is based on the following publication [38]:

> **Chen, W.**, Liu, Y., Pu, N., Wang, W., Liu L., and Lew, M.S., "Feature Estimations based Correlation Distillation for Incremental Image Retrieval." IEEE Transactions on Multimedia, 2021.

In general, deep models are trained and validated on fixed or stationary datasets. Exploring incremental learning by adding new fine-grained categories is still far from realizing the model's continuous retrieval ability because the images in old categories and new incoming categories are semantically similar. However, the images added to the fixed datasets may have different semantic contents, for instance when a model trained on a vehicle dataset is transferred to learn a new dataset which includes different breeds of flowers. For the context of incremental learning, the noted semantic shifts between different datasets (*e.g.* flower and car images) make the problem of minimizing the forgetting ratio more difficult when training deep models in a lifelong manner. In this case, we consider the fourth **RQ 4: How to perform lifelong image retrieval on different training datasets by reducing**

the impact of semantic shifts? We investigate this question by proposing a dual knowledge distillation framework, which is based on the submitted manuscript:

> **Chen, W.**, Pu, N., Liu, Y. , Lao, M., Wang, W., Bakker, E. M., Liu L., Tuytelaars, T., and Lew, M.S., "Lifelong Image Retrieval via Dual Knowledge Distillation." submitted to Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI) (*under review*), 2021.

1.3 Thesis Overview

The contents of this thesis is based on the articles where I have been the primary author. To organize this thesis for a better understanding, in **Chapter 2** we first present a comprehensive review of intelligent image retrieval via deep learning. It includes the main challenges of intelligent image retrieval, the categorization for retrieval methodologies, the popular convolutional neural networks for image retrieval, *etc.* This chapter aims at giving a global view for intelligent image retrieval, and is based on the submitted manuscript:

> **Chen, W.**, Liu, Y., Wang, W., Bakker, E. M., Georgiou T., Fieguth P., Liu L., and Lew, M. S., "Deep Image Retrieval: A Survey." submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (*Major revision*), 2021.

In Chapter 3 and Chapter 4, we target **RQ1** and present related work on cross-modal retrieval. We focus on cross-modal hash retrieval in **Chapter 3** and propose using information theory for measuring the domain uncertainty of binary codes in the shared feature space for image and text. In **Chapter 4**, we further combine information theory in reducing the heterogeneity gap for cross-modal retrieval. This method is beneficial for constructing a shared space for further learning commonalities between cross-modal features.

In **Chapter 5**, we target **RQ2** and explore image retrieval in the context of incremental learning. We focus on expanding the continuous retrieval ability of deep networks, and explore using feature correlations between deep features as knowledge to transfer from the teacher model to the student model. This research is limited to fine-grained datasets.

In **Chapter 6**, we target **RQ3** and explore how to utilize the model stream in incremental fine-grained image retrieval. We propose a feature estimation method to transfer more previous knowledge to further minimize the forgetting ratio on previous old data. The benefit of the feature estimation method is that the streams of deep models trained for the previous tasks are unnecessarily saved. We demonstrate the effectiveness of the proposed method on fine-grained datasets.

In **Chapter 7**, we target **RQ4** and explore lifelong image retrieval on different datasets. Compared to the proposed methods in Chapter 5 and Chapter 6, lifelong

image retrieval is important to numerous practical retrieval applications. To train a deep model on different datasets sequentially, the semantic shifts between the training data (*e.g.* from flower images to vehicle images) will make the problem of minimizing the forgetting ratio difficult. Moreover, prior works for image retrieval mainly improve the generalization ability of the deep model, but did not consider reducing forgetting simultaneously. To address these challenges, we propose a dual knowledge distillation framework and utilize the stored statistics in the BatchNorm layers of a frozen teacher model, which can minimize the forgetting ratio on the old tasks and simultaneously improve generalization on the new task.

In **Chapter 8**, we present recent new ideas and trends for multimodal content understanding. These methods can be used for intelligent image retrieval to seek performance improvement. This chapter is based on the following publication [39]:

> **Chen, W.**, Wang, W., Liu, L., and Lew, M.S., "New Ideas and Trends in Deep Multimodal Content Understanding: A Review." Neurocomputing, 2020, pp. 195-215.

Finally, in **Chapter 9** we conclude this thesis and reflect on future research directions. In addition, this thesis has been inspired by the insights and experiences from the related works in the following publications during my PhD studies:

- Liu, Y., **Chen, W.**, Liu, L., and Lew, M. S., "SwapGAN: A Multi-stage Generative Approach for Person-to-Person Fashion Style Transfer." IEEE Transactions on Multimedia, 2019, 21(9), pp. 2209-2222.
- Pu, N., **Chen, W.**, Liu, Y., Bakker, E. M., and Lew, M. S. "Dual Gaussian-based Variational Subspace Disentanglement for Visible-Infrared Person Re-Identification." ACM International Conference on Multimedia, 2020, pp. 2149-2158.
- Pu, N., **Chen, W.**, Liu, Y., Bakker, E. M., and Lew, M. S. "Lifelong Person Re-Identification via Adaptive Knowledge Accumulation." IEEE International Conference on Computer Vision and Pattern Recognition, 2021.
- Liu, Y., Guo, Y., **Chen, W.**, and Lew, M. S., "An Extensive Study of Cycle-Consistent Generative Networks for Image-to-Image Translation." IEEE International Conference on Pattern Recognition, 2018, pp. 219-224.
- Georgiou T, Liu Y, **Chen W**, Lew M. S. "A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision." International Journal of Multimedia Information Retrieval, 2020, 9(3), pp. 135-170.
- Lao, M., Guo, Y., Pu, N., **Chen, W.**, Liu, Y., and Lew, M. S. "Multi-stage hybrid embedding fusion network for visual question answering." Neurocomputing, 2021, 423, pp. 541-550.

- Lao, M., Guo, Y., Liu, Y., **Chen, W.**, Pu, N., and Lew, M. S. "From Superficial to Deep: Language Bias driven Curriculum Learning for Visual Question Answering." ACM International Conference on Multimedia, 2021.
- Georgiou, T., Schmitt, S., Bäck, T., **Chen, W.**, and Lew, M. S. "Norm Loss: An efficient yet effective regularization method for deep neural networks." IEEE International Conference on Pattern Recognition, 2020, pp. 8812-8818.
- Georgiou, T., Schmitt, S., Bäck, T., Pu, N., **Chen, W.**, and Lew, M. S. (2021). "Comparison of deep learning and hand crafted features for mining simulation data." IEEE International Conference on Pattern Recognition, 2020, pp. 1-8.