Exploring deep learning for intelligent image retrieval

Chen, W.

Chapter 9

Conclusions

Finding information in digital datasets and libraries is one of the grand challenges of our generation. Finding images or image retrieval is a major sub-problem and has recently had significant advances due to the groundbreaking developments in deep visual learning. In this thesis, we have explored and designed algorithms for retrieval tasks via deep learning methods, including unimodal image retrieval and cross-modal retrieval.

In Chapter 2, we presented a comprehensive review on deep learning for image retrieval. We introduced the popular backbone deep network architectures, that is widely used for extracting retrieval feature representations, and summarized three aspects of the challenge for deep image retrieval, including (1) reducing the semantic gap, (2) improving retrieval scalability, and (3) balancing retrieval accuracy and efficiency. Based on these main challenges, we presented methodologies for retrieval, including feature extraction, feature fusion, and feature enhancement methods. These methods can be employed in off-the-shelf convolutional neural networks. Also, they can be applied when deep networks are fine-tuned on the new target retrieval datasets. We analyzed supervised and unsupervised fine-tuning methods for the updating of network parameters. For these methodologies, we compared their performance on four retrieval benchmarks. This chapter aims to give a global view of intelligent image retrieval.

Finding an image of interest may require searching through thousands, millions, or even billions of images. Therefore, searching efficiently is as critical as searching accurately. To enable accurate and efficient retrieval of massive image collections, learning compact and rich feature representations is critical. In Chapter 3, we focus on cross-modal hash retrieval because hash code learning has high efficiency in computation and storage. We proposed an information entropy loss function based on Shannon information theory to reduce the heterogeneity gap, and thereby build a better common space to align the visual and textual modalities. We regularized real-valued features and the binary hash codes using the proposed information entropy loss. As demonstrated in Chapter 3, the challenge of performing cross-modal retrieval lies in how to measure the semantic similarity between data from different modalities. For this purpose, in Chapter 4, we proposed to integrate information theory and adversarial learning to learning the cross-modal features. Combining information theory and adversarial learning is beneficial in discovering the distribution differences between modalities to minimize the heterogeneity gap and enable more accurate retrieval. To guarantee the semantic similarity between data from visual and textual modalities, we adopted a bi-directional ranking loss function and a cross-modal feature projection method. Moreover, we adopted the Kullback–Leibler divergence to address the data imbalance issue which exists in the cross-modal datasets where each image is described by five sentences. The proposed method is evaluated by thorough experimental results on four well-known datasets using four deep models.

In Chapters 3 and 4, the proposed methods were trained and evaluated on fixed datasets. In Chapter 5, we explored fine-grained image retrieval in the context of incremental learning where deep networks are trained by using new data only. The new data is added at once or sequentially into the existing old data, where we employed the knowledge distillation method, which is computed based on the output probabilities from the final classifier, and the maximum mean discrepancy loss, which is based on the retrieval feature representations from the intermediate layer. The proposed method was compared with the state-of-the-art methods and to show its efficacy. We also applied the proposed method for incremental image classification tasks.

In Chapter 6, we further explored methods for incremental fine-grained image retrieval. Previously, in Chapter 5, we only used the penultimate model as the teacher model to regularize the current student model which learns on the new task. As incremental learning proceeds, especially when new data are added sequentially, knowledge distillation based on the stream of models will be memory-consuming and make the learning complex. We proposed a feature estimation method to estimate representative features from the models trained on earlier old tasks so that saving this model stream is unnecessary. Quantitative and qualitative experiments on two common benchmarks demonstrate that the proposed approach is effective for achieving optimal performance on both the old and new tasks when new incoming data are added at once or sequentially.

In Chapter 7, we explored fine-grained image retrieval in a lifelong manner. In contrast to Chapter 6 and Chapter 5, the images in the newly added data are semantically different from the ones in the already trained data. These semantic drifts make minimizing the forgetting ratio on previous tasks more difficult. In addition, we considered improving the generalization ability of the trained networks on the new tasks. To this end, we proposed a dual knowledge distillation framework that includes two professional teachers and a self-motivated student. To further alleviate the forgetting issue, we used the stored running statistics of the BatchNorm layers of the frozen teacher to generate several representative images. We evaluated the proposed framework on three benchmarks, where the scenarios of two-task sequence and three-task sequence are considered.

In Chapter 8, we presented four popular multimodal applications, including cross-modal retrieval, image captioning, image generation, and visual question answering. We introduced recent new ideas and trends of these applications from the viewpoint of structure for multimodal feature extraction and the strategies for multimodal feature learning. These novel ideas are important for better multimodal content understanding and can be further used to improve performance in intelligent image retrieval.

9.1 Limitations and Possible Solutions

Although our research has reached its aims, there still exist some limitations for our initial explorations for intelligent image retrieval.

First, in Chapter 4, we explored integrating information theory and adversarial learning for cross-modal retrieval, in which the information entropy loss was computed only based on image modality and text modality. Therefore, the feature vectors extracted from these two modalities are projected into a common feature space but the associations and alignments between cross-modal features are neglected. However, retrieval performance depends on the matching of each image-text feature pair. For some large-scale datasets, each category may include a large number of image-text pairs. Thus, it is valuable to make the information entropy loss specific for each category so that the discrepancy between two modalities can be reduced more granularly.

Second, we explored image retrieval in the context of incremental learning in Chapters 5, 6, and 7, by focusing on the representations extracted from the teacher-student structure to distill correlations. Thus, both old tasks and new tasks are trained on the same representations. However, regularizing directly on the representations may be overly restrictive for the learning on the new tasks. We find the accuracy of new tasks on the CUB-Birds dataset is still lower than the upper bound of joint training (see Table 6.1). For this limitation, instead of regularizing the representations, it may be promising to project them into a sub-space using an auto-encoder or a variational auto-encoder. Afterward, informative parts of the representations for the old tasks are captured and kept unchanged, while others that are not meaningful for the old tasks allow the learning for new tasks.

Third, we proposed a feature estimation method in Chapter 6 to minimize the forgetting ratio in previous tasks. In fact, effectively estimating representations for all previous models depends on the parameter inheritance of model initialization at the start of each incremental step. However, estimated features from the penultimate model to the first one are not accurate enough due to the accumulative estimation errors. We resolved this limitation by aligning estimated features with descending importance and demonstrated its effectiveness experimentally. Nevertheless, distilling knowledge on the stream of models is worth further investigation theoretically. Sequence modeling via the recurrent network [369] may be a direction that deserves to be explored.

9.2 Future Research Directions

In terms of future work, there are several directions into which we can extend our research work:

1. Unsupervised intelligent image retrieval. We have explored intelligent image retrieval in a supervised manner. However, supervisory information such as class labels are time-consuming and labor-intensive to collect. Therefore, it is valuable to investigate unsupervised image retrieval. For example, the proposed Shannon information loss functions in Chapter 3 and Chapter 4 are label-free and can be used in some unsupervised learning scenarios. It may be more difficult for lifelong image retrieval in an unsupervised manner that uses the teacher-student framework. Without the supervisory information to regularize the training of the student network, the student network may suffer from more severe forgetting on the previous tasks. One possible solution is to employ the Variational AutoEncoder (VAE), which can be used for unsupervised learning, in lifelong representations learning.

2. Multimodal retrieval. In an information era, people can search for the item of interest by using different kinds of queries which makes the field of multimodal retrieval an area that richly deserves to be explored. One of the challenges for multimodal retrieval is to align features from different modalities in a shared latent space. We have examined the application of combining Shannon information entropy with adversarial learning for cross-modal retrieval. We find that Shannon information entropy can be used for multimodal feature learning by estimating the modality uncertainty. It will be promising to explore Shannon entropy further when applied to other kinds of cross-modal feature learning similar to image-text retrieval, such as video-text, audio-video, and audio-text matching, which aims at learning modality-invariant representations.

3. Zero-shot learning for image retrieval. The popularity of media platforms and the rapid development of novel techniques makes it very convenient for people to share their images, and as a result, the number of images on the Internet has increased exponentially where there often exist "unseen" images or categories. However, most datasets are static and offer a limited amount of objects and categories for feature learning. Thus, the retrieval algorithms or systems may suffer from the scarcity of the appropriate training data for these unseen images. Therefore, there is a need to extend conventional image retrieval methods to a zero-shot learning scenario where we can retrieve both seen and unseen categories from the system. Furthermore, combined with unsupervised methods, zero-shot learning algorithms can significantly improve the flexibility and generalization of image retrieval systems.

4. Incremental learning for image retrieval. Content-based image retrieval can be divided into category-level image retrieval and instance-level image retrieval. In our work, we have paid attention to explore category-level image retrieval in the context of lifelong learning. To avoid forgetting ratios on the already trained tasks, more techniques may be necessary, such as hierarchical learning. Since images used in the incremental fine-grained image retrieval share subtle inter-class variations and

larger intra-class variations, it is valuable to learn hierarchical domain knowledge. Furthermore, examining instance-level image retrieval in incremental learning is also promising.

5. Deploy image retrieval for practical applications. Existing image retrieval technologies are trained and evaluated on standard benchmarks, and various metric learning methods are explored for retrieval on fine-grained datasets. However, these technologies are still far from real-world applications such as face search, fashion search, person re-identification, shopping recommendation systems, or medical image retrieval. In these practical applications, the purpose of image retrieval may not just be retrieving images for general content on standard benchmarks, but also for more refined information. It is challenging to deploy image retrieval for specific scenarios. For example, as a specific instance search topic, person re-identification systems may encounter images with low-resolution or with inferior quality due to inadequate illumination. Existing techniques such as attention mechanisms and region proposal networks can be adopted to guarantee performance. In addition, it is valuable to explore multi-modal retrieval in practical applications. This means that image retrieval can also be combined with other auxiliary modalities such as words, phrases, and sentences to meet the different retrieval expectations of users.