Exploring deep learning for intelligent image retrieval

Chen, W.

English Summary

We are living in an information era where the amount of image and video data increases exponentially. It is important to develop appropriate information systems to store, manage, and distribute such large data collections. Among them, intelligent image retrieval is one of the most indispensable techniques to be considered. It satisfies our needs for searching information of interest. To enable intelligent image retrieval (including high accuracy and high efficiency retrieval), feature representations are at the core of most retrieval algorithms.

For humans, it is easy to find similar images from an image gallery according to a given query image. However, it is difficult for a computer to search as accurately as humans due to the existing semantic gap between the high-level concepts used by humans and the typically low-level features derived from images (*i.e.* pixels or symbols). In addition, it will be more difficult for the computer to search accurately if the query contains multiple modalities (*e.g.* text, audio *etc*). This is caused by the second challenge: the heterogeneity gap. Deep learning, especially for convolutional neural networks has made progress in addressing these challenges and significantly facilitated the process of intelligent image retrieval.

The first theme in this thesis is to explore cross-modal retrieval by considering visual and textual modalities. This theme is hard to realize because it involves both the above mentioned semantic gap and heterogeneity gap. We design an information entropy loss function based on Shannon information theory to regularize the learning of a shared latent space for paired image and text inputs. The common practice of cross-modal retrieval is to construct a shared space where image features and text features are highly intermixed, thereby the similarity between image and text can be further associated. This property of the shared space is consistent with Shannon information theory by measuring the information entropy. This idea is demonstrated for cross-modal hashing retrieval where real-valued features and binary hash codes are constrained by the information entropy loss.

Next, we explore the integration of Shannon information theory and adversarial learning for cross-modal retrieval. This adversarial mechanism achieves a better feature distribution agreement for the two modalities thereby bridging the heterogeneity gap and enabling a more accurate retrieval. To reduce the semantic gap, Kullback-Leibler (KL) divergence and bi-directional triplet loss are used to associate

the intra- and inter-modality similarity between features in the shared space. Also, we design a regularization term based on KL-divergence with temperature scaling to calibrate the bias of the label classifier that is caused by the data imbalance issue.

The second theme of this thesis is to explore the continuous retrieval capacity of deep neural networks where three important sub-questions are studied: incremental learning for retrieval on the same fine-grained dataset, feature estimation for sequences of deep models in incremental learning, and lifelong learning for image retrieval on different datasets, respectively. Unlike the learning process of humans, training previously trained deep networks on new data leads the networks to forget what was learned before. For the first sub-question, we employ incremental learning for the find-grained image retrieval task. This is achieved through regularizing the retrieval representations and classification probabilities by using a maximum mean discrepancy loss function and knowledge distillation loss function. To evaluate the proposed method, we split a dataset into two parts, one is used as the old data (or old tasks) and the other is used as the new data for incremental training (or new tasks).

For the second sub-question, we focus on the sequence of deep neural networks which have been trained when new tasks are added sequentially. This multi-task scenario will suffer from more severe catastrophic forgetting. Saving the sequence of models for transferring previously learned knowledge is memory-consuming. Instead, we propose a simple but effective feature estimation method to alleviate this limitation.

For the third sub-question, we consider a more practical lifelong image retrieval scenario where the deep model is trained successively on different datasets. The semantic drifts between different datasets make minimizing the forgetting ratio more difficult. We address this limitation by using a dual knowledge distillation framework that includes two professional teachers and a self-motivated student. One teacher model has its parameters fixed and is used for transferring previously learned knowledge on the proceeding tasks while another on-the-fly teacher is trained jointly with the student and is responsible for transferring knowledge learned on the newly added tasks. Furthermore, we also use the statistics on the BatchNorm layers of the frozen teacher model to generate some representative images for the successive training tasks.

We conduct thorough experiments to verify the efficacy of the proposed methods for the two themes. The results demonstrate significant improvements over various baselines and state-of-the-art methods. Therefore, this thesis provides novel contributions, insights, and findings for the research community and future applications in the field of intelligent image retrieval.