Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Problems in Large-Scale Image Classification

Yuchen Guo
School of Software, Tsinghua University
Beijing 100084, China
yuchen.w.guo@gmail.com

Introduction

The number of images is growing rapidly in recent years because of development of Internet, especially the social networks like Facebook, and the popularization of portable image capture devices like smart phone. Annotating them with semantically meaningful words to describe them, i.e., classification, is a useful way to manage these images. However, the huge number of images and classes brings several challenges to classification, of which two are 1) how to measure the similarity efficiently between large-scale images, for example, measuring similarity between samples is the building block for SVM and kNN classifiers, and 2) how to train supervised classification models for newly emerging classes with only a few or even no labeled samples because new concepts appear every day in the Web, like Tesla's Model S.

The research of my Ph. D. thesis focuses on the two problems in large-scale image classification mentioned above. Formally, these two problems are termed as *large-scale similarity search* (Gong et al. 2013) which focuses on the large scale of samples/images and *zero-shot/few-shots learning* (Lampert, Nickisch, and Harmeling 2014) which focuses on the large scale of classes. Specifically, my research considers the following three aspects: 1) hashing based large-scale similarity search which adopts hashing to improve the efficiency; 2) cross-class transfer active learning which simultaneously transfers knowledge from the abundant labeled samples in the Web and selects the most informative samples for expert labeling such that we can construct effective classifiers for novel classes with only a few labeled samples; and 3) zero-shot learning which utilizes no labeled samples for novel classes at all to build supervised classifiers for them by transferring knowledge from the related classes.

Hashing Based Large-scale Similarity Search

Given a large-scale database, storing the feature vectors and performing similarity search in it are both quite expensive. For example, suppose we have 10 million 128-dimensional SIFT points with float representation. We need about 5GB memory as storage and searching kNN in the database requires at least several minutes. However, if we can transform the float representation into binary codes while preserving the similarity structure, i.e., hashing, the efficiency can be significantly improved based on the extremely fast bit operation in modern computer. For example, in the same database, if we transform the each vector into 128-bit binary codes, we only need about 160MB memory and searching kNN using the bit operations in the database takes only tens of milliseconds. Because of its efficiency in space and speed, hashing has received considerable attention in recent years.

Iterative Quantization (ITQ) (Gong et al. 2013) is one of the state-of-the-art hashing approaches. However, it suffers from two problems. First, its $\ell_{2,2}$-norm loss is very sensitive to noise in data. Second, it only focuses on preserving ℓ_2-norm (Euclidean) distance while ignoring other distance measures. To address these issues, we extended ITQ into a more robust and general formulation using a $\ell_{p,q}$-norm loss ($0 < q \leq p \leq 2$). It is robust to noise by using $q \leq 1$ and supports general ℓ_p-norm similarity search. Minimizing the $\ell_{p,q}$-norm loss with orthogonality constraint is a challenging non-convex and non-smooth problem. As an important theoretical contribution, we proposed a novel iterative optimization algorithm which could efficiently solve this problem, and we rigorously proved the correctness of the algorithm. Extensive experiments on benchmarks showed the proposed approach achieved state-of-the-art performance and was superior to the original ITQ, which validated its effectiveness. This work was published in IJCAI 2016 (Guo et al. 2016b).

In real-world applications, the images can be represented as multiple modalities, like SIFT, GIST, CNN features. And because of the difference of capturing devices, images can be represented by different features. How to combine multi-modality features to improve the retrieval performance, and how to connect different modalities such that cross-modality search is feasible, i.e., using a SIFT represented query image to retrieve GIST represented image database, are two important problems in cross/multi-modality hashing community (Song et al. 2013). To address these problems, we proposed a novel Collective Matrix Factorization Hashing framework, which could effectively build connections across modalities and combine information from multiple modalities. We conducted comprehensive experiments and the results demonstrated the superiority of the proposed framework to the state-of-the-arts. This work was published in CVPR 2014 (Ding, Guo, and Zhou 2014) and the extended version has been accepted by IEEE TIP (Ding et al. 2016).

Cross-class Transfer Active Learning

Given the large scale of the classes in the Web, it is difficult to collect sufficient labeled samples for the uncommon class. In addition, new concepts emerge every day. Therefore, how to construct classifiers for these classes using as few labeled samples as possible is a critical real-world problem. Active Learning (Settles 2009) is an effective way. In addition, we can also notice that there are abundant labeled samples from different but related classes. Therefore, if we can transfer knowledge from the related classes to the target classes, the labeling efforts for the target classes can be further saved. This strategy, termed as cross-class transfer active learning, is an interesting solution to the large-scale classes training.

Because the class name is described by words, we can utilize word2vec tool to generate a vector representation for each class as its feature. In this way, different classes are embedded into the same feature space and we can adopt this intermediate space to transfer knowledge across classes (Socher et al. 2013). Based on the cross-class knowledge transfer, we can significantly reduce the number of labeled samples for target classes to achieve satisfactory classification accuracy, which was demonstrated by the experiments compared to the state-of-the-art conventional active learning approaches. This work was published in AAAI 2016 (Guo et al. 2016d). In addition, based on the space, we can train a projection function that projects images into this space using labeled samples. In this way, the similarity between each sample and each target class can be directly measured and then we can select from related classes the most similar samples to the target classes and assign pseudo labels to them. Based on the pseudo labels, we extend the labeled set of target classes such that we can expect better performance. In fact, it is believed that such a sample-transfer scheme is very logical because human being often uses one category to deduce another one as long as they share the same characteristics. This cross-class sample transfer based active learning approach was published in IJCAI 2016 (Guo et al. 2016a).

The above approaches consider the relationship between samples and classes, but the relationship between samples is ignored. In fact, because there are a few labeled samples for target classes, we can utilize them to help selects the samples from related classes that can well capture the characteristics of target classes. The above approaches focus on the class aspect while this approach focuses on the sample aspect and they are complementary to each other. We can expect better results by considering this extra information. This work is currently in progress and anticipated to finish in Feb. 2017.

Zero-shot Learning

In the above section we consider the situation where target classes has a few labeled samples, i.e., few-shots learning. In zero-shot learning (Lampert, Nickisch, and Harmeling 2014), we consider that there is no labeled samples at all. To transfer knowledge, we need the attributes the word vector of classes for knowledge transfer. The work on attribute learning was published in AAAI 2015 (Guo et al. 2015). To construct classifiers for target classes, we assume that the classifiers' parameters for target classes and related classes share a same model space and the attribute of a class is the seed parameters to produce the classifier's parameters by a function. Based on this assumption, we can learn the function using the labeled samples in the related classes and the corresponding attributes. Then, we can directly produce the classifiers for target classes by their attributes and the learned function. Our experiments on benchmarks revealed that this approach performed better than the state-of-the-arts. This work was published in AAAI 2016 (Guo et al. 2016c).

Currently, I am working on applying the sample transfer method into zero-shot learning. Because there is no labeled samples for target classes, the transfer method requires more elaborate designs and this work is submitted to AAAI 2017.

The recent years has witnessed the success of deep learning, especially deep convolutional neural network, on image classification. However, training CNN model needs a large number of labeled samples. Even fine-tuning it requires sufficient label information. In the future, I plan to investigate how to train/fine-tune CNN model in the zero-shot manner to save training cost, which is meaningful for large-scale image classification in the Web if I could complete this work.

References

Ding, G.; Guo, Y.; Zhou, J.; and Gao, Y. 2016. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Trans. Image Processing*.

Ding, G.; Guo, Y.; and Zhou, J. 2014. Collective matrix factorization hashing for multimodal data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2083–2090.

Gong, Y.; Lazebnik, S.; Gordo, A.; and Perronnin, F. 2013. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE TPAMI* 35(12):2916–2929.

Guo, Y.; Ding, G.; Jin, X.; and Wang, J. 2015. Learning predictable and discriminative attributes for visual recognition. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 3783–3789.

Guo, Y.; Ding, G.; Gao, Y.; and Wang, J. 2016a. Semi-supervised active learning with cross-class sample transfer. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 1526–1532.

Guo, Y.; Ding, G.; Han, J.; and Jin, X. 2016b. Robust iterative quantization for efficient ℓ_p-norm similarity search. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3382–3388.

Guo, Y.; Ding, G.; Jin, X.; and Wang, J. 2016c. Transductive zero-shot recognition via shared model space learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*.

Guo, Y.; Ding, G.; Wang, Y.; and Jin, X. 2016d. Active learning with cross-class knowledge transfer. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 1624–1630.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(3):453–465.

Settles, B. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. Y. 2013. Zero-shot learning through cross-modal transfer. In *NIPS*.

Song, J.; Yang, Y.; Yang, Y.; Huang, Z.; and Shen, H. T. 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *ACM SIGMOD*, 785–796.