Image Classification & Retrieval are ONE (Online NN Estimation)

Lingxi Xie[1], Richang Hong[2], Bo Zhang[1] and Qi Tian[3]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]Department of Computer Science, University of Texas at San Antonio, Texas, USA

ABSTRACT

Both image classification and retrieval receive a query image at a time. Classification tasks aim at determining the class or category of the query, for which a number of training samples are provided and an extra training process is often required. For retrieval, the goal is to rank a large number of candidates according to their relevance to the query, and candidates are considered as independent units, i.e., without explicit relationship between them. Both image classification and retrieval tasks could be tackled by the Bag-of-Visual-Words (BoVW) model. However, the ways of performing classification [10][26] and retrieval [46][38] are, most often, very different. Although all the above algorithms start from extracting patch or regional descriptors, the subsequent modules, including feature encoding, indexing/training and online querying, are almost distinct.

In this paper, we suggest using only ONE (Online Nearest-neighbor Estimation) algorithm for both image classification and retrieval. This is achieved by computing similarity between the query and each category or image candidate. Inspired by [4], we detect multiple object proposals on the query and each indexed image, and extract high-quality features on each object to provide better image description. On the online querying stage, the query's relevance to a category or candidate image is estimated by the averaged nearest distance from querying objects to the objects in that category or candidate image. As shown in experiments, extracting more objects helps to find various visual clues and obtain better results. To improve efficiency, we leverage the idea of approximate nearest-neighbor search, and take advantage of GPU parallelization for fast computation. Experiments on a wide range of image classification/retrieval datasets reveal the state-of-the-art performance of our method.

NOVELTY

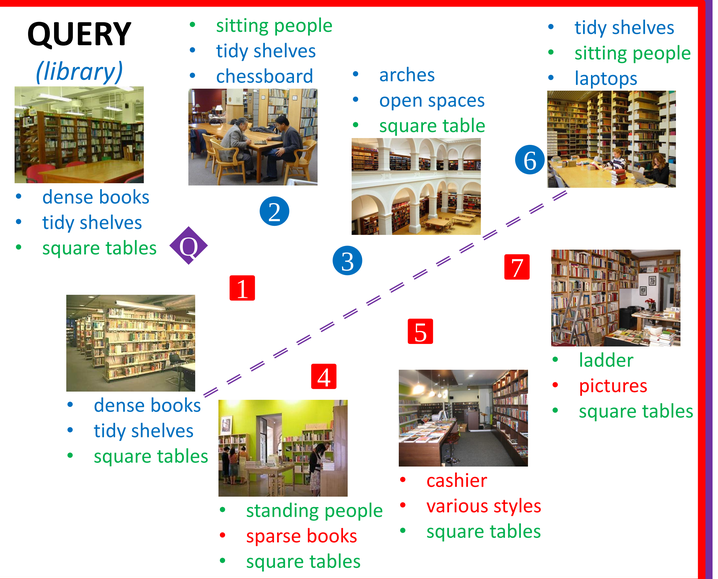
The major contribution of this paper is summarized in the following aspects.

1. We reveal that the possibility of unifying image classification and retrieval systems into ONE (Online NN Estimation).
2. ONE achieves the state-of-the-art accuracy on a wide range of image classification and retrieval tasks, defending both training-free models for image recognition and regional features for near-duplicate object retrieval.
3. We make full use of GPU parallelization to alleviate heavy online computational overheads, which might inspire various multimedia applications and research efforts in the future.

THE PROPOSED FRAMEWORK

Motivation: bridging the difference

Q: Why are we using different algorithms for classification and retrieval?
A: In classification, each training image is annotated with a label. However, retrieval tasks often do not provide such additional information.
Q: How can classification benefit from label information?
A: With labels, images are latently partitioned into concept groups, and classification algorithms can measure image-to-class distance rather than image-to-image distance, which is verified much more stable [4]. Training-based algorithms such as SVM actually uses a different approach to compute the image-to-class distance.
Q: What does the right-handed example imply?
A: This is atypical classification vs. retrieval comparison. Since the query "Q" is most similar to "1", in image retrieval, we might take "1" as its nearest neighbor. However, the label of "1" (*bookstore*) is not the same as "Q" (*library*). However with SVM, we train an optimal classification boundary which takes all the samples into consideration. The classifier detects that "1" is an outlier and produces the correct prediction for "Q".



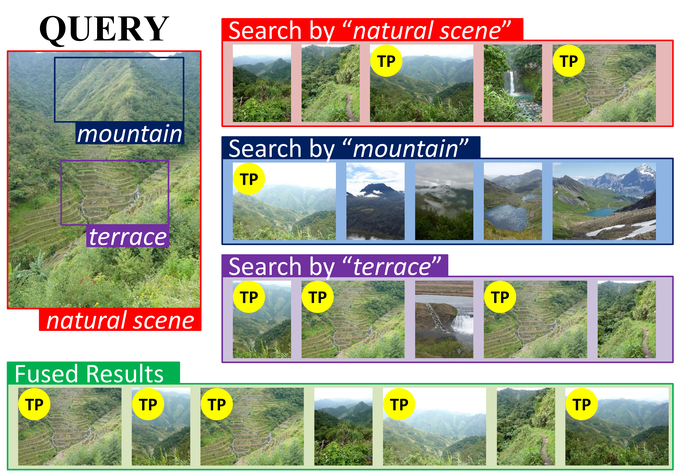
Algorithm: image-to-class distance

Q: How to define class in retrieval tasks?
A: We extract several regions on an image (with either manual definition or automatic detection) to capture more visual information. Each region can be considered as an independent image, and all the regions (images) extracted from the same original image are considered to form a pseudo class.
Q: What is the difference between this formulation and NBNN [4]?
A: NBNN uses local features as basic units, but we use regional features (e.g., deep features) which are verified more effective. We use this model for both classification and retrieval tasks, while NBNN is only tested on classification.
Q: What is the shortcoming of this formulation?
A: Complexity. Computing NBNN needs much more computational resources than conventional methods. We use both PCA and PQ approximation. The highly parallelizable form also allows us to use GPU acceleration.
Q: Does approximation harm the accuracy?
A: We find that approximation does not cause dramatic accuracy drop. With proper parameters, we achieve balance between performance and speed.

$$\underbrace{\mathrm{dist}(\mathbf{I}_0, c)}_{\text{Query-to-Class distance}} \doteq \underbrace{\mathrm{dist}(\mathbf{I}_0, \mathcal{F}_c)}_{\text{Image-to-Class distance}} = \frac{1}{K_0}\sum_{k=1}^{K_0}\underbrace{\mathrm{dist}(\mathbf{f}_{0,k}, \mathcal{F}_c)}_{\text{Feature-to-Class distance}} = \frac{1}{K_0}\sum_{k=1}^{K_0}\min_{\mathbf{f}\in\mathcal{F}_c}\|\mathbf{f}_{0,k}-\mathbf{f}\|_2^2$$

Analysis: fusing visual attributes

Q: How can retrieval tasks benefit from the ONE algorithm?
A: The difference is shown in the right-handed figure. Using a single image (or region) for retrieval often captures one type of visual attributes, with the use of multiple regions, we can extract a lot more information and the retrieval performance is significantly improved.



RESULTS

Scene Recognition

Algorithm	LandUse-21	Indoor-67	SUN-397
Kobayashi [24]	92.8	63.4	46.1
Xie [61]	-	63.48	45.91
Donahue [14]	-	-	40.94
Razavian [43]	-	69.0	-
SVM	**94.52**	68.46	53.00
ONE	93.98	**69.61**	**54.47**
SVM+ONE	**94.71**	**70.13**	**54.87**

Fine-Grained Recognition

Algorithm	Pet-37	Flower-102	Bird-200
Wang [52]	59.29	75.26	-
Murray [32]	56.8	84.6	33.3
Donahue [14]	-	-	58.75
Razavian [43]	-	86.8	61.8
SVM	88.05	68.46	53.00
ONE	**89.50**	**86.24**	**61.54**
SVM+ONE	**90.03**	**86.82**	**62.02**

Near-duplicate Retrieval

Algorithm	Holiday	Holiday+1M	UKBench
Zhang [68]	0.809	0.633	3.60
Zheng [70]	0.881	0.724	3.873
BoVW	0.518	-	3.134
ONE	**0.887**	-	**3.873**
BoVW+ONE	**0.899**	**0.758**	**3.887**

REFERENCES

Key references are numbered as they appear in the paper.
[4] O. Boiman, E. Shechtman, and M. Irani. In Defense of Nearest-Neighbor Based Image Classification. CVPR, 2008.
[25] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS, 2012.
[37] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher Kernel for Large-scale Image Classfication. ECCV, 2010.
[45] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint, arXiv: 1409.1556, 2014.
[72] L. Zheng, S. Wang, and Q. Tian. Lp-Norm IDF for Scalable Image Retrieval. TIP, 2014.

ACKNOWLEDGE.

This work was supported by the 973 Program of China (Grant Nos. 2013CB329403 and 2012CB316301), the NNSF of China (Grant Nos. 61332007, 61273023 and 61429201), and the Tsinghua University Initiative Scientifc Research Program (Grant No. 20121088071). This work was also supported in part to Dr. Hong by State High-Tech Development Plan 2014AA015104, and the Program for New Century Excellent Talents in University under grant NCET-13-0764; and in part to Dr. Tian by ARO grant W911NF-12-1-0057, and Faculty Research Awards by NEC Labs of America.