Deep Semantic Ranking Based Hashing for Multi-Label Image Retrieval

Fang Zhao, Yongzhen Huang, Liang Wang, Tieniu Tan
Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences

With the rapid growth of web images, hashing has received increasing interests in large scale image retrieval. Research efforts have been devoted to learning compact binary codes that preserve semantic similarity based on labels. However, most of these hashing methods are designed to handle simple binary similarity. The complex multilevel semantic structure of images associated with multiple labels have not yet been well explored. Besides, the learning capability of the standard pipeline followed by most hashing method, i.e., firstly extracting features like GIST and SIFT as image representations, and then learning mappings from these representations to binary codes, is inadequate for dealing with relatively complex semantic structure due to the semantic information loss in the hand-crafted features. Thus more effective semantic feature representation is also desirable.

In this paper, we introduce a novel framework based on semantic ranking and deep learning model for learning hash functions that preserve multilevel similarity between multi-label images in the semantic space. An overall view of the proposed framework termed deep semantic ranking based hashing (DSRH) is illustrated in Fig. 1.

We use deep convolutional neural network (CNN) [2] to construct hash functions $\mathbf{h}(\mathbf{x};\mathbf{W})$ to learn directly from images, which provides much richer sematic information than hand-crafted features. As shown in Fig. 2, we add a bypassing connection between the first fully connected layer (FCa) and the hash layer to reduce the possible information loss. We argue that the features from the second fully connected layer (FCb) of CNN are dependent on classes too much and have strong invariance, which is unfavorable for capturing subtle sematic distinction. Thus we connect the hash layer to both the two fully-connected layers to enable it encoding more diverse information biased toward visual appearance.

Meanwhile, to optimize ranking measures such as NDCG [1], we learn such deep hash functions with semantic ranking supervision which is the order of a ranking list derived from shared class labels between query and database images. The learning is a joint optimization of feature representation and mappings from them to hash codes, and it is more effective than the conventional two-stage pipeline.

A ranking loss defined on a set of triplets is used as surrogate loss to solve the optimization problem resulting from nonsmooth and multivariate ranking measures:

$$L_\omega(\mathbf{h}(\mathbf{q}),\{\mathbf{h}(\mathbf{x}_i)\}_{i=1}^M) = \sum_{i=1}^{M}\sum_{j:r_j<r_i}\omega(r_i,r_j)\left[\delta d_H(\mathbf{h}(\mathbf{q}),\mathbf{h}(\mathbf{x}_i),\mathbf{h}(\mathbf{x}_j))+\rho\right]_+. \quad (1)$$

where M is the length of the ranking list, r_i is the similarity level of the i-th database point in the ranking list, $[.]_+ = \max(0,.)$, $\delta d_H(\mathbf{h},\mathbf{h}_1,\mathbf{h}_2) = d_H(\mathbf{h},\mathbf{h}_1) - d_H(\mathbf{h},\mathbf{h}_2)$, $d_H(.,.)$ is the Hamming distance and ρ is a margin parameter which controls the minimum margin between the distances of the two pairs. The weight ω can be given by:

$$\omega(r_i,r_j) = \frac{2^{r_i}-2^{r_j}}{Z} \quad (2)$$

where Z is the normalization constant in NDCG. The higher the relevance of $\mathbf{x}_i$ and $\mathbf{q}$ is than that of $\mathbf{x}_j$ and $\mathbf{q}$, the larger decline the NDCG score would suffer if $\mathbf{x}_i$ is ranked behind $\mathbf{x}_j$. And thus the larger weight should be assigned to this triplet.

Given the dataset $\mathcal{D}$ as a training set, we wish to learn hash functions that optimize the rankings for all query points $\mathbf{q}$ from $\mathcal{D}$. Based on the surrogate loss (1) and the hash function in Fig. 2, the objective function can be given by the empirical loss subject to some regularization:

$$\mathcal{F}(\mathbf{W}) = \sum_{\mathbf{q}\in\mathcal{D},\{\mathbf{x}_i\}_{i=1}^M\subset\mathcal{D}} L_\omega(\mathbf{h}(\mathbf{q};\mathbf{W}),\{\mathbf{h}(\mathbf{x}_i;\mathbf{W})\}_{i=1}^M) + \frac{\alpha}{2}\left\|\operatorname{mean}_{\mathbf{q}}(\mathbf{h}(\mathbf{q};\mathbf{W}))\right\|_2^2 + \frac{\beta}{2}\|\mathbf{W}\|_2^2. \quad (3)$$

The second term is the balance penalty, and the third term is the L_2 weight decay. Stochastic gradient descent is used to minimize the objective function (3), and the derivatives of (3) with respect to hash code vectors can be fed into the underlying CNN via the back-propagation algorithm to update the parameters of each layer.

Experimental results demonstrate that our method is able to capture complex multilevel semantic structure and significantly outperforms other hashing methods based on hand-crafted features, pre-trained and fine-tuned CNN in ranking quality.

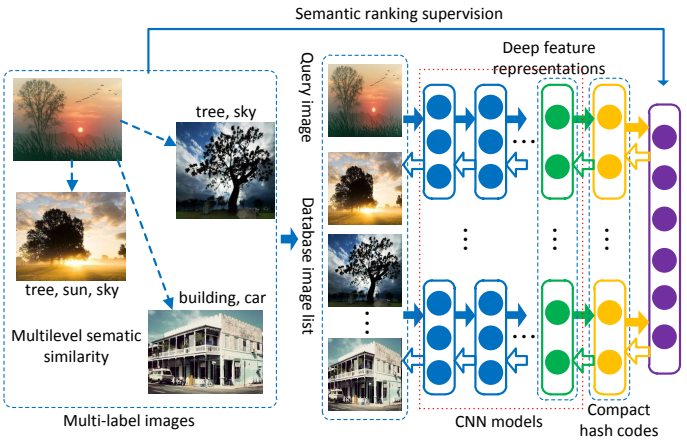

Figure 1: The proposed deep semantic ranking based hashing. Solid and hollow arrows indicate forward and backward propagation directions of features and gradients respectively. Hash functions consist of deep convolutional neural network (CNN) and binary mappings of the feature representation from the top hidden layers of CNN. Multilevel semantic ranking information is used to learn such deep hash functions to preserve the semantic structure of multi-label images.

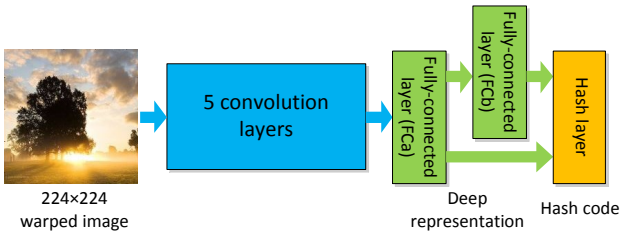

Figure 2: The structure of deep hash functions. CNN provides a deep feature representation, and the hash layer generates a compact binary code from the deep representation.

[1] K. Jarvelin and J. Kekalainen. Ir evaluation methods for retrieving highly relevant documents. In *Proc. ACM SIGIR*, 2000.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.