Probabilistic Semi-Supervised Multi-Modal Hashing

Behnam Gholami
bb510@cs.rutgers.edu
Computer Science Department
Rutgers university

Abolfazl Hajisami
hajisamik@cac.rutgers.edu
Department of Electrical and Computer Engineering
Rutgers university

In this paper, we propose a non-parametric Bayesian framework for multi-modal hash learning that takes into account the distance supervision (similarity/dissimilarity constraints). Our model embeds data of arbitrary modalities into a single latent binary feature with the ability to learn the dimensionality of the binary feature using the data itself. Given supervisory information (labeled similar and dissimilar pairs), we propose a novel discriminative term and develop a new Variational Bayes (VB) algorithm which incorporates that term into the proposed Bayesian framework.

Let $\boldsymbol{T} = [\boldsymbol{X}, \boldsymbol{Y}]$ be the observed bi-modal data matrix where $\boldsymbol{X} = [x_1, x_2, ..., x_d]_{M \times d}$ and $\boldsymbol{Y} = [y_1, y_2, ..., y_d]_{N \times d}$ denote the first modal and the second modal data matrix respectively, and $\boldsymbol{Z} = [z_1, z_2, ..., z_d]_{K \times d}$ denotes the latent binary code matrix.

In our VB framework, we truncate the length of the binary codes (K) and we set it to a finite but large number. If K is large enough, the analyzed multi-modal data using this number of bits, will reveal less than K bits. In order to incorporate the information of the similarity/dissimilarity constraints into the VB algorithm, we first define a regularizer for the binary code z_i as

$$\alpha(z_i) = \frac{1}{|\mathcal{D}_i|} \sum_{j:(i,j)\in\mathcal{D}} KL(q_{z_i}(z_i)||q_{z_j}(z_j)) - \frac{1}{|\mathcal{S}_i|} \sum_{j:(i,j)\in\mathcal{S}} KL(q_{z_i}(z_i)||q_{z_j}(z_j)) \quad (1)$$

where $KL(p||q)$ denotes the KL divergence between two distributions p and q, and $\mathcal{S}(\mathcal{D})$ denotes the set of similar (dissimilar) pairwise constraints. Intuitively, for each binary code z, $\alpha(z)$ should be large such that it best agrees with those constraints.

By defining the regularizer $\Omega(\boldsymbol{Z}) = \sum_{i=1}^{d} \alpha(z_i)$ for the binary code matrix $\boldsymbol{Z}$ using the set of similar/dissimilar pairs, we add this regularizer to the objective function of VB and solve the new optimization problem using the Coordinate Descent method.

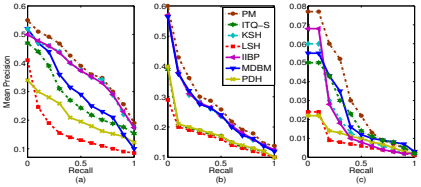

Figure 1: The result of category retrieval for image-to-image queries. (a) PASCAL-Sentence Dataset; (b) SUN Dataset (Euclidean ground truth computed from visual data); (c) SUN Dataset (Class label ground truth)

We evaluate the proposed method on two benchmark bi-modal datasets: (1) The PASCAL-Sentence 2008 dataset [1] consists of 1000 images categorized into 20 classes. (2) The SUN-Attribute dataset [2] contains 102 attribute labels for each of the 14340 images from 717 categories. We compare the performance of the proposed method against five state-of-the-art hashing methods (Fig. 1) using precision-recall curve as an accuracy measure. As can be seen, the proposed method outperforms the other state of the art (multi-modal) hashing methods.

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010.

[2] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758. IEEE, 2012.