Computationally Bounded Retrieval

Mohammad Rastegari[1], Cem Keskin[2], Pushmeet Kohli[2], Shahram Izadi[2]
[1]University of Maryland. [2]Microsoft Research.

In this paper, we focus primarily on image retrieval using binary codes [3]. In this approach, each image is encoded into a binary code such that the nearest neighbors in the corresponding Hamming space remain the same as the actual nearest neighbors in the original feature space. Calculating the binary code, *i.e.* encoding, is in this case evaluating a mapping function f from some d-dimensional inputs to k-dimensional outputs. However if d and k are large, computational efficiency of f becomes a bottleneck at test time.

Typically, a single bit in a binary code is obtained by a dot product between the normal vector of a hyperplane and the feature vector, which implies $\mathcal{O}(dk)$ time complexity. In [1] a bilinear projection is employed to reduce the time complexity to $\mathcal{O}(d\sqrt{k})$. Recently, [2] proposed to use the columns of a circulant matrix as linear projections. This enables faster projection to generate k-bits ($k > 1$) by Fast Fourier Transform (FFT) which is $\mathcal{O}(d\log(d))$. In this paper, we propose an optimization that learns a sparse projection to binary codes with a constraint on the computational budget.

We propose to find a mapping f that quantizes the data into binary values while: 1-*minimizing the quantization error*, and 2-*minimizing the computational cost of* f.

We formulated the problem as learning of a prediction function $f_w : \mathcal{X} \rightarrow \mathcal{Y}$, which is a many-one mapping between some input space $\mathcal{X}$ and an output space $\mathcal{Y}$ that is parameterized by a parameter vector $\mathbf{w}$. Given a set of n training input-output pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, the conventional empirical risk minimization approach for learning the optimal parameter vector $\mathbf{w}^*$ involves solving the following optimization problem:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(\mathbf{y}_i, f_w(\mathbf{x}_i)). \tag{1}$$

where $\ell(\mathbf{y}, \hat{\mathbf{y}})$ is a loss function that measures the discrepancy between the prediction $\hat{\mathbf{y}}$ and the ground truth output $\mathbf{y}$.

A simple but popular representation for the prediction function for binary classification problems is $f_w(\mathbf{x}) = \operatorname{sign}(\mathbf{w}^\mathsf{T}\mathbf{x})$, where sign is element wise sign function $\{\operatorname{sign} : \mathbb{R} \mapsto \{+1, -1\}\}$. In the case of binary codes learning the parameter vector $\mathbf{w}$ is replaced by a parameter matrix $\mathbf{W}$. In this paper, we focus on binary code predictors that tie themselves to a fixed computational budget. More formally, we want to solve the computation-bounded risk minimization problem that is defined as:

$$\underset{\mathbf{W},\mathbf{b}}{\operatorname{argmin}} \quad \sum_{i=1}^{n} \ell(\mathbf{b}_i, \operatorname{sign}(\mathbf{W}^\mathsf{T}\mathbf{x}_i)) \tag{2}$$

$$st. \qquad \|\mathbf{w}\|_0 \leq l \tag{3}$$

where $\mathbf{b}_i \in \{-1, 1\}^k$ is the desired binary code for $\mathbf{x}_i$ and $\|\mathbf{w}\|_0$ denotes the ℓ_0 norm that counts the number of non-zero components of $\mathbf{W}$ since only these many multiplications are needed to evaluate the function.

Sparse Projection When Binary Codes Are Given

To improve the computation cost of f, a relatively straightforward idea is to make the matrix $\mathbf{W}$ sparse. When $\|\mathbf{W}\|_0 \leq l$, *i.e.* when number of non-zero entries of $\mathbf{W}$ is at most l, then clearly f can be computed in $\mathcal{O}(l)$. However, directly solving for ℓ_0-norm is intractable. Therefore, sparsity is often incorporated by introducing an ℓ_1 penalty on the parameter matrix $\mathbf{W}$ followed by thresholding.

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}} \{\|\mathbf{W}^\mathsf{T}\mathbf{X} - \mathbf{B}\|_F + \lambda|\mathbf{W}|_{\ell_1}\} \tag{4}$$

It can be shown that the optimal solution for $\mathbf{W}$ can be computed independently for each column of $\mathbf{W}$.

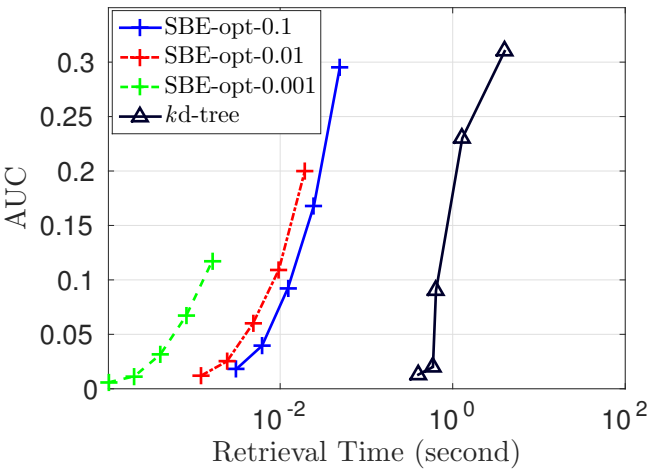

Figure 1: Comparison of the accuracy (area-under-the-curve) and computational cost of the proposed algorithm (SBE) with the *kd*-tree method for different sparsity values. The curves correspond to increasing code lengths for SBE and decreasing bucket size for *kd*-tree. Our method achieves the same accuracy as a *kd*-tree while being 100 times faster.

Joint Optimization

There is no guarantee that the exact codes can be reconstructed by sparse mapping, we need to incorporate the search for binary codes into the main objective. Similar to the Equation 4, we aim to minimize the quantization error while maintaining the low ℓ_1-norm. In contrast to Equation 4, $\mathbf{B}$ is an unknown variable.

$$(\mathbf{W}^*, \mathbf{B}^*) = \underset{\mathbf{W},\mathbf{B}}{\operatorname{argmin}} \{\|\mathbf{W}^\mathsf{T}\mathbf{X} - \mathbf{B}\|_F + \lambda|\mathbf{W}|_{\ell_1}\} \tag{5}$$

To solve the optimization of Equation 5, we replace the matrix $\mathbf{B}$ with an explicit sign function of an orthogonal projection of the data as follows:

$$\begin{aligned}(\mathbf{W}^*, \mathbf{P}^*) = &\underset{\mathbf{W},\mathbf{P}}{\operatorname{argmin}} \{\|\mathbf{W}^\mathsf{T}\mathbf{X} - \operatorname{sign}(\mathbf{P}^\mathsf{T}\mathbf{X})\|_F + \lambda|\mathbf{W}|_{\ell_1}\} \\ &s.t. \quad \mathbf{P}^\mathsf{T}\mathbf{P} = \mathbf{I}\end{aligned} \tag{6}$$

where $\mathbf{P} \in \mathbb{R}^{d \times k}$ is an orthogonal matrix. This orthogonal projection ensures the low correlation between the bits, and also in contrast to Equation 5, it provides an update for binary codes which is not directly dependent on $\mathbf{W}$.

[1] Yunchao Gong, Sanjiv Kumar, Henry A Rowley, and Svetlana Lazebnik. Learning binary codes for high-dimensional data using bilinear projections. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 484–491. IEEE, 2013.

[2] Felix X Yu, Sanjiv Kumar, Yunchao Gong, and Shih-Fu Chang. Circulant binary embedding. *arXiv preprint arXiv:1405.3162*, 2014.

[3] Liang Zheng, Shengjin Wang, and Qi Tian. Coupled binary embedding for large-scale image retrieval. *Image Processing, IEEE Transactions on*, 2014.

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