Optimized Transform Coding for Approximate KNN Search

Minwoo Park
mpark@objectvideo.com
Kiran Gunda
kgunda@objectvideo.com
Himaanshu, Gupta
hgupta@objectvideo.com
Khurram, Shafique
kshafique@objectvideo.com

Research and Development Services
ObjectVideo
11600 Sunrise Valley Dr, Ste 210
Reston, USA
http://www.objectvideo.com

Transform coding (TC) is an efficient and effective vector quantization approach where the resulting compact representation can be the basis for a more elaborate hierarchical framework for sub-linear approximate search. However, as compared to the state-of-the-art product quantization methods, there is a significant performance gap in terms of matching accuracy. One of the main shortcomings of TC is that the solution for bit allocation relies on an assumption that probability density of each component of the vector can be made identical after normalization. Motivated by this, we propose an optimized transform coding (OTC) such that bit allocation is optimized directly on the binned kernel estimator of each component of the vector. Experiments on public datasets show that our optimized transform coding approach achieves performance comparable to the state-of-the-art product quantization methods, while maintaining learning speed comparable to TC.

Introduction: In the context of general vector quantization, a quantizer encoder $e(\mathbf{x})$ is a real-valued function $\mathcal{E}:\mathbb{R}^n \to \mathcal{I}$ characterized by the region it induces on the input space, $\mathcal{R}^n_{\mathbf{x}} = \{\mathbf{x} \in R^n(i) : e(\mathbf{x}) = i\}$ and $\cup_{i=1}^{L} R^n(i) = \mathbb{R}^n$ where $\mathcal{I} = \{1, \cdots, L\}$ and $\mathbf{x}$ is an input vector. The decoder $d(i)$ is a real-valued function $\mathcal{D}:\mathcal{I} \to \mathbb{R}^n$ characterized by the codebook $\mathcal{C} = \{i \in \mathcal{I} : d(i) = y_i\} \subset \mathbb{R}^n$. The mean distortion error of the given quantization level L (MDE) of the quantization is given as:

$$MDE(L) = \sum_{i=1}^{L} \int_{R^n(i)} f(\mathbf{x}) Dist(\mathbf{x}, d(e(\mathbf{x}))) d\mathbf{x} \tag{1}$$

where f is an estimated probability density function of multi-dimensional vector $\mathbf{x}$ and $Dist(\mathbf{x},\mathbf{x}')$ is a distortion error between $\mathbf{x}$ and $\mathbf{x}'$.

In general, to find the optimal set of region $\mathcal{R}_{\mathbf{x}}$, the codebook $\mathcal{C}$, and the given quantization level L, minimum-distortion quantizer aims to minimize mean distortion error (MDE) as follows:

$$\left(\mathcal{R}^{opt}_{\mathbf{x}}, \mathcal{C}^{opt}\right) = arg \min_{\mathcal{R}_{\mathbf{x}},\mathcal{C}} MDE(L) \tag{2}$$

Although design of such a scalar quantizer to satisfy the minimum distortion criterion is well understood, vector quantization is still an open problem. For instance, it can be challenging to obtain sufficient sample data to characterize $f(\mathbf{x})$. Moreover, solving Eq. (2) is computationally expensive in high dimensions.

However, if $p(\mathbf{x})$ is independent in its components (dimensions), and the metric is of the form given as:

$$Dist(\mathbf{x},\mathbf{x}') = \sum_{k=1}^{D} dist(x_k, x'_k), \tag{3}$$

where D is a dimension of $\mathbf{x}$, $\mathbf{x}_k$ are the k^{th} component of $\mathbf{x}$, and $dist(x_k, x'_k)$ is a distance metric between $\mathbf{x}_k$ and $\mathbf{x'}_k$, we can obtain a minimum distortion quantizer by forming the Cartesian product of the independently quantized components. That is, the vector quantization encoder can be of a form, $e(\mathbf{x}) = [e_1(x_1), \cdots, e_D(x_D)]^T$. In the original PQ [4, 5], D dimensional space is divided into M sub-spaces (typical M is 8) to form given as:

$$e(\mathbf{x}) = \left[e_{1\sim K}(x_{1\sim K}), \cdots, e_{7K+1\sim 8K}(x_{7K+1\sim 8K})\right]^T \text{ where } K = D/M. \tag{4}$$

However, each component is not independent in practice. Therefore, TC [1] and OPQ [2] aim to minimize inter-component dependencies using the principal component analysis (PCA) and show great success over the original PQ [4, 5]. After minimizing the inter-component statistical dependencies using PCA, the quantizer design problem reduces to a set of M number of independent K dimensional problems. In TC, $K = 1$ and $M = D$. The major difference between OPQ and TC lies in the bit-allocation approach used in each method. The key difference is that OPQ assigns the same number of bits per sub-space, while TC assigns a different number of bits per sub-space. Therefore OPQ finds the best combination of components for each sub-space while maintaining the same number of bits for each sub-space while TC finds the number of bits suitable for each sub-space.

In the context of TC, each quantizing encoder e_k at the k^{th} dimension is designed independently for every $1 \le k \le D$ to minimize the expected distortion given as:

$$MDE_k(L_k) = \sum_{i=1}^{L_k} \int_{R_k(i)} f_k(c_k) dist_k(c_k, d_k(e_k(c_k))) dc_k. \tag{5}$$

where c_k is PCA coefficient after projection of $\mathbf{x}$ to PCA subspace k.

Therefore, a vector quantization using B-bits code is summarized as follows:

$$(\mathcal{L}, \mathcal{R}_{\mathbf{c}}, \mathcal{C})^{opt} = arg \min_{\mathcal{L},\mathcal{R}_{\mathbf{c}},\mathcal{C}} \sum_{k=1}^{D} MDE_k(L_k) \text{ subject to } \sum_{k=1}^{D} log_2(L_k) = B. \tag{6}$$

If the number of distinct quantization levels per k^{th} component L_k is known for a total target bit B, a product quantizer can be obtained by using the minimum distortion criterion. Optimal bit allocation is achieved by minimizing the expected distortion due to quantization. However, solution to this optimization problem for general distributions and distortion functions requires computationally prohibitive numerical search [1].

Instead, Brandt [1] adopted greedy integer-constrained allocation algorithm [3] to assign bits. Number of the quantization level set to be proportional to the variance of the data under the two assumptions that 1) probability density of each component can be made identical after the normalization and 2) per-component distortion functions are identical. However, the first assumption can be easily violated in many cases (e.g., non-Gaussian probability density function). Motivated by this problem, we propose to solve Eq. (6) directly in our proposed optimized transform coding (OTC). Details can be found in the paper.

[1] Jonathan Brandt. Transform coding for fast approximate nearest neighbor search in high dimensions. volume 0, pages 1815–1822, Los Alamitos, CA, USA, 2010. IEEE Computer Society.

[2] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 2946–2953, Washington, DC, USA, 2013. IEEE Computer Society.

[3] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991. ISBN 0-7923-9181-0.

[4] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 304–317, Berlin, Heidelberg, 2008. Springer-Verlag.

[5] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

Acknowledgments: Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.