Regularized Residual Quantization: a multi-layer sparse dictionary learning approach

Sohrab Ferdowsi, Slava Voloshynovskiy, Dimche Kostadinov

Department of Computer Science, University of Geneva, Switzerland

{Sohrab.Ferdowsi, svolos, Dimche.Kostadinov}@unige.ch

I. INTRODUCTION

Quantizing the residual errors from a previous level of quantization has been considered in signal processing for different applications, e.g., image coding. This problem was extensively studied in the 80's and 90's (e.g., see [1] and [4]). However, due to strong over-fitting, its efficiency was limited for more modern applications with larger scales. In practice, it was not feasible to train more than a couple of layers. Particularly at high dimensions, the codebooks learned on a training set were not able to quantize a statistically similar test set.

Inspired by an insight from rate-distortion theory, we introduce an effective regularization for the framework of Residual Quantization (RQ), making it capable to learn multiple layers of codebooks with many stages. Moreover, the introduced framework effectively deals with high dimensions making it feasible to go beyond patch level processing and deals with entire images. The proposed regularization makes use of the problem of optimal rate allocation for asymptotic case of Gaussian independent sources, which is reviewed next.

II. BACKGROUND: QUANTIZATION OF INDEPENDENT GAUSSIAN SOURCES

Given n independent Gaussian sources X_j's each with variance σ_j^2 distributed as $X_j \sim \mathcal{N}(0, \sigma_j^2)$, the optimal rate allocation from the rate-distortion theory is derived for this setup as (Ch. 10 of [2]):

$$D_j = \begin{cases} \gamma, & \text{if } \sigma_j^2 \geqslant \gamma, \\ \sigma_j^2, & \text{if } \sigma_j^2 < \gamma, \end{cases} \quad (1)$$

where γ should be chosen to guarantee that $\sum_{j=1}^n D_j = D$. Hence, the optimal codeword variance $\sigma_{C_j}^2$ is soft-thresholding of σ_j^2:

$$\sigma_{C_j}^2 = \left(\sigma_j^2 - \gamma\right)^+ = \begin{cases} \sigma_j^2 - \gamma, & \text{if } \sigma_j^2 \geqslant \gamma, \\ 0, & \text{if } \sigma_j^2 < \gamma. \end{cases} \quad (2)$$

This means that sources with variances less than γ should not be assigned any rate at all. We next incorporate this phenomenon for codebook learning and enforce it as a regularization for the codebook variances. This, in fact, will be an effective way to reduce the gap between the train and test distortion errors. Moreover, the inactivity of the dimensions with variances less than γ will also lead to a natural sparsity of codewords, which lowers the computational complexity.

III. THE PROPOSED APPROACH: RRQ

Instead of the standard K-means used in RQ, we first propose its regularized version and then use it as the building-block for RRQ.

A. VR-Kmeans algorithm

After de-correlating the data points, e.g., using the pre-processing proposed in Fig. 2, and gathering them in in columns of X with σ_j^2 at each dimension, define $S \triangleq \text{diag}([\sigma_{C_1}^2, \cdots, \sigma_{C_n}^2])$ from Eq. 2. For codebook C, to regularize only the diagonal elements of CC^T, define

P_j with all elements as zeros except at $P_{(j,j)} = 1$. We formulate the variance-regularized K-means algorithm with parameter λ as:

$$\underset{C, A}{\text{minimize}} \quad \frac{1}{2}||X - CA||_F^2 + \frac{1}{2}\lambda||\sum_{j=1}^n P_j CC^T P_j - S||_F^2, \quad (3)$$

$$\text{s.t.} \quad ||\boldsymbol{\alpha}_i||_0 = ||\boldsymbol{\alpha}_i||_1 = 1.$$

Like the standard K-means algorithm, we iterate between fixing C and updating A, and then fixing A and updating C.

Fix C, **update** A: Exactly like the standard K-means.

Fix A, **update** C: Eq. 3 can be re-written as:

$$\underset{C}{\text{minimize}} \ \text{Tr}\Big[-XA^T C^T + \frac{1}{2}CAA^T C^T + \frac{1}{2}\lambda(\sum_{j=1}^n P_j CC^T P_j)(CC^T - 2S)\Big]. \quad (4)$$

$\sum_{j=1}^n P_j CC^T P_j$, and due to its structure $AA^T \triangleq \text{diag}([a_1, \cdots, a_k])$ are diagonal. Therefore Eq. 4 will reduce to minimizing independent sub-problems at each (active) dimension:

$$\underset{\mathbf{c}(j)}{\text{minimize}}\Big[-\mathbf{z}(j)^T \mathbf{c}(j) + \frac{1}{2}\big(a_1 c_1(j)^2 + \cdots + a_k c_k(j)^2\big) + \frac{1}{2}\lambda||\mathbf{c}(j)||^2(||\mathbf{c}(j)||^2 - 2\sigma_{C_j}^2)\Big], \quad (5)$$

where $Z \triangleq XA^T = [\mathbf{z}(1), \cdots, \mathbf{z}(n)]^T$. These independent problems can be solved easily using the Newton's algorithm, for which the derivation of the required gradient and Hessian is straightforward.

B. Regularized Residual Quantization (RRQ) algorithm

For a fixed number of centroids $K^{(l)}$ at layer l and $D_j^{(l-1)}$ the distortion of the previous stage of quantization for each dimension, the RRQ first specifies $\gamma^* = \underset{\gamma}{\text{argmin}}\Big(|\log_2 K^{(l)} - \sum_{j \in \mathcal{A}_\gamma} \frac{1}{2}\log_2^+ \frac{D_j^{(l-1)}}{\gamma}|\Big)$ followed by calculation of an active set of dimensions $\mathcal{A}_{\gamma^*}^{(l)} = \{j : 1 \leqslant j \leqslant n | \sigma_j^2 \geqslant \gamma^*\}$. The algorithm then continues with quantizing the residual of stage $l - 1$ with the *VR-Kmeans* algorithm described above, until a desired stage L which can be chosen based on distortion constraints or an overall rate budget allowed.

IV. EXPERIMENTS

Fig. 1 and Table I compare the performance of *VR-Kmeans* with K-means in quantization of high-dimensional variance-decaying independent data. In fact, in many practical cases, the correlated data behaves similarly after an energy-compacting and de-correlating transform. As is seen in this figure, the *VR-Kmeans* regularizes the variance resulting in a reduced train-test distortion gap.

Fig. 3 demonstrates the performance of the RRQ in super-resolution of similar images. It is clear from this figure that the high-frequency content lost in down-sampling can be reconstructed from a multi-layer codebook learned from face images with full resolution.
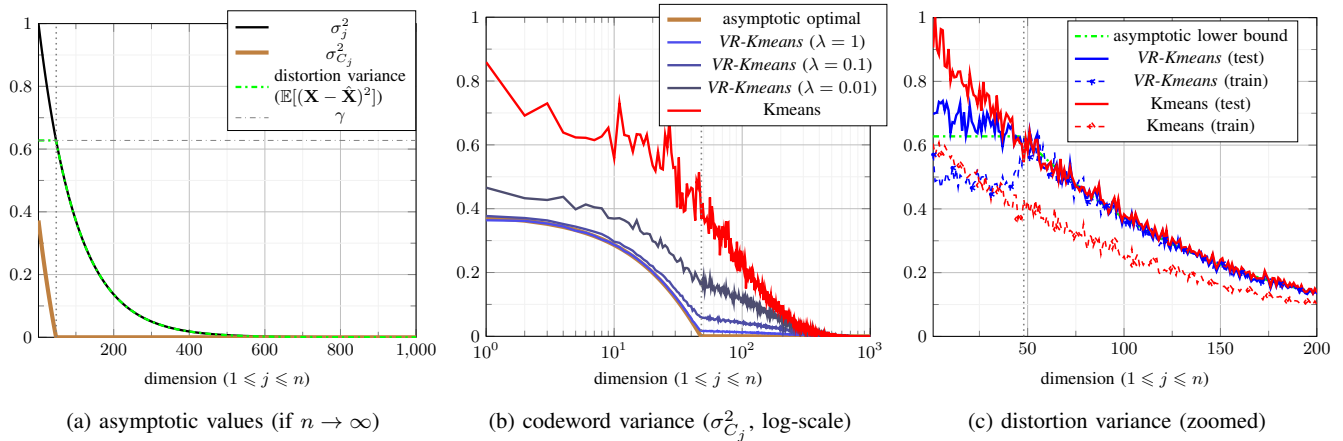
(a) asymptotic values (if $n \to \infty$)   (b) codeword variance ($\sigma^2_{C_j}$, log-scale)   (c) distortion variance (zoomed)

Fig. 1: Quantization of variance-decaying independent data ($n = 1000$, $K = 256$): (a) asymptotically optimal values (b) Higher values of γ enforce the asymptotically optimal values of $\sigma^2_{C_j}$ more strongly. (c) The regularization on *VR-Kmeans* results in less extreme distortion minimization on the training set, but a better performance on the test set.
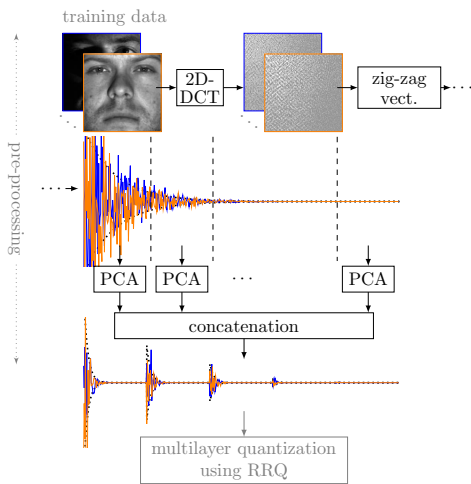

Fig. 2: A general pre-processing for images required before RRQ. Since global PCA for the full-frame images has n^2 parameters, for high-dimensional images it will over-train. Therefore, we propose to use global 2D-DCT which largely de-correlates the images and perform PCA (without dim. reduction) only locally at different sub-bands for further de-correlation. As a result, the data becomes effectively de-correlated with very strong variance-decaying nature, suitable for RRQ.

	Kmeans	random generation	VR-Kmeans ($\lambda = 0.1$)	VR-Kmeans ($\lambda = 10$)	VR-Kmeans ($\lambda = 1000$)
distortion train	0.6727	0.9393	0.8441	0.8520	0.8568
distortion test	1.0054	0.9394	0.9413	**0.9384**	0.9390

TABLE I: Quantization distortion (normalized) on the train and test sets for K-means, random codeword generation (from $\mathcal{N}(\mathbf{0}, S)$) and the *VR-Kmeans* algorithm (average over 5 trails). The theoretically minimum distortion (achieved at $n \to \infty$) is 0.9185. Notice that K-means, while achieves the lowest distortion on the training set, fails to quantize the test set. *VR-Kmeans* with proper γ, on the other hand, performs the best on the test set.

(a) original   (b) bicubic (8 ↑)   (c) RRQ

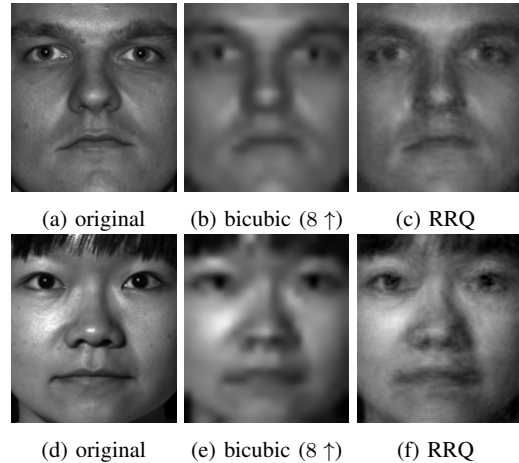(d) original   (e) bicubic (8 ↑)   (f) RRQ

Fig. 3: Super-resolution using RRQ on the *CroppedYale* [3] set with $L = 100$ layers with $K = 256$ centroids each: After pre-processing as proposed in Fig. 2, a randomly chosen train set is quantized with RRQ. Images from a test set are down-sampled and again up-sampled with bi-cubic interpolation and then quantized and reconstructed by the codebooks learned from the training set. Since the data has strong variance-decaying nature at the initial layers, the trained codewords are very sparse. As the number of layers increases, the data becomes more *i.i.d.* and less structured and hence less sparsity in the codewords.

REFERENCES

[1] C. F. Barnes, S. A. Rizvi, and N. M. Nasrabadi. Advances in residual vector quantization: a review. *IEEE Transactions on Image Processing*, 5(2):226–262, Feb 1996.
[2] T. Cover and J. Thomas. *Elements of Information Theory 2nd Edition*. Wiley-Interscience, 2 edition, 7 2006.
[3] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.
[4] N. M. Nasrabadi and R. A. King. Image coding using vector quantization: a review. *IEEE Transactions on Communications*, 36(8):957–971, Aug 1988.