# Product Quantization to Reduce Entropy of Labels for Fast and Accurate Image Retrieval

Fuga Nakamura, Ryosuke Harakawa and Masahiro Iwahashi
Department of Electrical, Electronics and Information Engineering, Nagaoka University of Technology, Niigata, Japan
E-mail: s171062@stn.nagaokaut.ac.jp, {harakawa, iwahashi}@vos.nagaokaut.ac.jp

***Abstract*—Product quantization (PQ) is a popular technique for fast image retrieval from a large-scale database. PQ methods quantize image features into short codes and realize fast retrieval using lookup tables based on the codes. Although the entropy of labels (i.e., ground truths for retrieval) is crucial for the retrieval performance, existing PQ methods focus only on the quantization errors. This paper proposes a novel PQ method that reduces the entropy of labels to improve the retrieval performance. We assume that correct labels for each training sample are known; then, we train the codes so that we can minimize the label errors as well as the quantization errors to reduce the entropy of labels. This enables fast and accurate retrieval when queries (i.e., images whose labels are unknown) are given.**

## I. Introduction

Content-based image retrieval (CBIR) is a fundamental technique in the multimedia and computer vision fields [1], [2]. CBIR uses an image rather than a text as a query, and retrieves visually similar images from the database. The simple method is to calculate distances between a feature vector of the query and those of all images in the database and present the closest one. Time complexity of this method is $O(ND)$, where $N$ is the number of feature vectors in the database, and $D$ is their dimensions. Therefore, the large-scale database with large $N$ and $D$ requires a huge amount of calculation time.

Product quantization (PQ) [3] is a popular technique for fast image retrieval from a large-scale database. PQ decomposes each feature vector in the database ($\in \mathbb{R}^D$) into $M$ subvectors ($\in \mathbb{R}^{D/M}$) and applies vector quantization (VQ) [4] to them. Thus, each feature vector is represented as short codes. By using a lookup table, the distance between an uncompressed vector (query) and many compressed vectors in the database can be computed quickly. As a result, PQ can retrieve approximate neighbors quickly. However, if subspaces are not mutually independent, the retrieval performance is degraded. To overcome this limitation, optimized product quantization (OPQ) [5], additive quantization (AQ) [6], and composite quantization (CQ) [7] have been proposed. Specifically, OPQ applies optimal rotation to feature vectors in advance, and AQ and CQ represent feature vectors as sums of several codewords of dimension $D$. While PQ, OPQ, AQ, and CQ are unsupervised methods, supervised and semi-supervised methods [8]–[13] have been proposed. These methods [8]–[13] can achieve high retrieval performance when labels of all training data or its part are known. However, there is still room for improvement in these conventional methods [8]–[13]. Specifically, although the entropy of labels is crucial for the retrieval performance (whose details are explained in Section I), these methods focus only on the quantization errors.

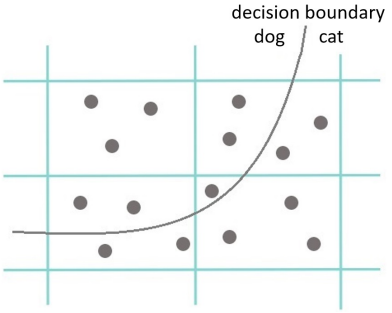



**Fig. 1:** Example of clustering by PQ. Black dots denote feature vectors in a database. The black line is the decision boundary that separates the dog domain from the cat domain.

Motivated by this, this paper proposes a novel PQ method to reduce the entropy of labels. Specifically, when retrieving a query, data are presented randomly from the cluster (Voronoi region of a codeword) to which the query belongs. For example, Fig. 1 shows an example of clustering the training data by PQ. If the query belongs to the upper-left clusters (entropy is 0), a vector of the correct label is presented. If it belongs to the lower-left or upper-right cluster (entropy is larger than 0), a vector of the wrong label may be presented. Therefore, the retrieval performance would be improved by reducing the entropy of labels in each cluster. To reduce the entropy, we assume that correct labels for each training sample are known, and learn the codebook by using these labels. Specifically, instead of clustering the training data with a *k*-means algorithm [14] as in the PQ, our method clusters training data to minimize errors of the labels (degrees to which each cluster includes different labels) as well as the quantization errors. To the best of our knowledge, this is the first work to reduce the entropy of labels in PQ. In the experiment, we confirmed that the proposed method achieves higher performance than conventional methods.

## II. Proposed method

The proposed method represents feature vectors as short codes of length $M$ like the original PQ. In PQ, when a query is given, images in the cluster to which the query belongs are presented. Therefore, we consider that the smaller the entropy

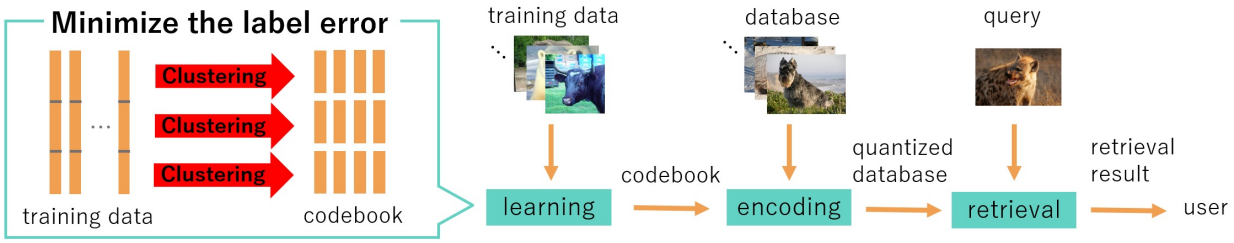



**Fig. 2:** Overview of our method.

of labels in each cluster, the higher the retrieval accuracy is. Motivated by this consideration, our method learns the codebook so that the entropy of the Voronoi regions of each codeword is smaller than that of PQ.

Figure 2 shows an overview of our method. The difference between the conventional method and the proposed method lies in the clustering method used in the learning process. In the conventional method, codewords are obtained by applying $k$-means clustering [14] to the training data to minimize the quantization error. In our method, codewords are obtained by using a new clustering method (explained in Section II-C) that minimizes the error of labels, to reduce the entropy of labels. Because the retrieval process is the same as in the conventional method, the retrieval speed does not change.

### A. Learning Codebook

This section describes a method for learning codebooks. First, let $\{\boldsymbol{x}_n | 1 \leq n \leq N\}$ be the training data that consist of $N$ training samples. We assume that correct labels for each training sample $\boldsymbol{x}_n$ are known, and $\{\boldsymbol{p}_n | 1 \leq n \leq N\}$ denotes one-hot vectors that represent them. More specifically, $\boldsymbol{p}_n$ is a distribution that represents the label of $\boldsymbol{x}_n$. As shown in Eq. (1), by dividing each training sample into $M$ pieces, the first subvector $\boldsymbol{x}_n^1$, the second subvector $\boldsymbol{x}_n^2$, ... , the M-th subvector $\boldsymbol{x}_n^M$ are obtained.

$$\boldsymbol{x}_n = (\underbrace{x_1, ..., x_{D/M}}_{\boldsymbol{x}_n^1}, ..., \underbrace{x_{D-D/M+1}, ..., x_D}_{\boldsymbol{x}_n^M})^T. \quad (1)$$

We apply the new clustering method (explained in Section II-C) to $\{\boldsymbol{x}_n^1 \in \mathbb{R}^{D/M} | 1 \leq n \leq N\}$. This clustering method uses $\boldsymbol{p}_n$ to reduce the entropy of labels. We obtain $K$ clusters by this clustering, and use centroids of clusters $\boldsymbol{\mu}_k$ as the codewords. The same procedure is applied to $\boldsymbol{x}_n^2$, $\boldsymbol{x}_n^3$, ..., $\boldsymbol{x}_n^M$, to obtain the codewords in each subspace.

### B. Label Error

Here, we explain the label error that the clustering method used in our PQ method minimizes. The aim of our clustering is to reduce the entropy of labels in each cluster. Eq. (2) shows the average value of the entropy, where $V_k$ is the Voronoi region of the codeword $\boldsymbol{\mu}_k$.

$$\frac{1}{K}\sum_{n=1}^{N} H(V_k) = \\ -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\left[\frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{p}_n}{\sum_{n=1}^{N} r_{nk}}\right]_c \log_2\left[\frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{p}_n}{\sum_{n=1}^{N} r_{nk}}\right]_c. \quad (2)$$

Here, $r_{nk} = 1$ for $k$ where $\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2$ is minimized, and $r_{nk} = 0$ for the other $k$. This expression cannot be differentiated by $\boldsymbol{\mu}_k$. Therefore, it is difficult to find a solution of $\boldsymbol{\mu}_k$ that minimize this function.

Therefore, instead of the entropy, the proposed method minimizes the label error in Eq. (3).

$$Labelerror(V_k) = \sum_{n=1}^{N} \|\boldsymbol{p}_n - \boldsymbol{n}_k\|^2, \quad (3)$$

$$\boldsymbol{n}_k = \frac{\sum_{\boldsymbol{x}_n \in V_k} \boldsymbol{p}_n}{|V_k|}, \quad (4)$$

where $\boldsymbol{n}_k$ is an intermediate value representing the label of the cluster $V_k$. Here we use the value in Eq. (4) for simplicity. The label error in Eq. (3) represents the degree to which the cluster $V_k$ includes different labels. This is because the larger the mixture of labels in the cluster, the more ambiguous the average label is. For example, consider the cluster $V_1$, which contains four images of dogs, and the cluster $V_2$, which contains two images of dogs and two images of cats. Let us assume that the label of dogs is $(1,0)^{\mathrm{T}}$ and that for cats is $(0,1)^{\mathrm{T}}$. In this case, the average label of $V_1$ is $(1,0)^{\mathrm{T}}$ and matches each label, so the error is zero. The average label of $V_2$ is $(0.5,0.5)^{\mathrm{T}}$, and the mixture of labels leads to an ambiguous label that is far from each label, and the error is larger than zero. Because the label error indicates the degree to which each cluster includes different labels, we can reduce the entropy by minimizing it.

### C. Clustering

In order to reduce the entropy of labels in each cluster, our method performs clustering in such a way that the label error is minimized. Specifically, we solve the problem in Eq. (5).

$$\begin{aligned}\arg\min_{\rho_{nk},\boldsymbol{\mu}_k,\boldsymbol{\nu}_k} &\frac{t}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\rho_{nk}^{m}\|\boldsymbol{x}_n-\boldsymbol{\mu}_k\|^2 \\ &+\frac{1-t}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\rho_{nk}^{m}\|\boldsymbol{p}_n-\boldsymbol{\nu}_k\|^2,\end{aligned} \tag{5}$$

where $\rho_{nk}$ is the probability that the point $\boldsymbol{x}_n$ belongs to the $k$-th cluster, and $\sum_{k=1}^{K}\rho_{nk}=1$ must be satisfied. $\boldsymbol{\mu}_k$ is the mean of the $k$-th cluster and, $\boldsymbol{\nu}_k$ is the mean label of the $k$-th cluster. $t$ $(0<t<1)$ and $m$ $(\geq 1)$ are hyperparameters. The first half of Eq. (5) represents the quantization error, the second half represents the label error, and $t$ is the weight for the quantization error. Also, $m$ is the degree of ambiguity of the affiliation probability.

To solve Eq. (5), we alternately optimize $\rho_{nk}$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\nu}_k$. When $\boldsymbol{\mu}_k$ and $\boldsymbol{\nu}_k$ are fixed, the minimum solution of $\rho_{nk}$ is shown in Eq. (6) for $m>1$.

$$\begin{aligned}\rho_{nk}&=\frac{A_{nk}}{\sum_{k=1}^{K}A_{nk}},\\ A_{nk}&=\left(t\|\boldsymbol{x}_n-\boldsymbol{\mu}_k\|^2+(1-t)\|\boldsymbol{p}_n-\boldsymbol{\nu}_k\|^2\right)^{\frac{1}{1-m}}.\end{aligned} \tag{6}$$

If $m=1$, $\rho_{nk}=1$ for $k$ where $t\|\boldsymbol{x}_n-\boldsymbol{\mu}_k\|^2+(1-t)\|\boldsymbol{p}_n-\boldsymbol{\nu}_k\|^2$ is minimized, and $\rho_{nk}=0$ for the other $k$. When $\rho_{nk}$ is fixed, the minimum solutions of $\boldsymbol{\mu}_k$ and $\boldsymbol{\nu}_k$ are shown in Eqs. (7) and (8).

$$\boldsymbol{\mu}_k=\frac{\sum_{n=1}^{N}\rho_{nk}^{m}\boldsymbol{x}_n}{\sum_{n=1}^{N}\rho_{nk}^{m}}, \tag{7}$$

$$\boldsymbol{\nu}_k=\frac{\sum_{n=1}^{N}\rho_{nk}^{m}\boldsymbol{p}_n}{\sum_{n=1}^{N}\rho_{nk}^{m}}. \tag{8}$$

### *D. Encoding and Retrieval*

Finally, we describe encoding and retrieval. The method of encoding vectors using the learned codebook is the same as in the original PQ, where vectors are divided into $M$ subvectors and VQ is applied to them. This quantizes the vectors into short codes of length $M$.

Also, the method of retrieval is the same as in PQ. Concretely, the query vector is divided into $M$ subvectors and the distance between each subvector and each codeword is calculated to create a lookup table in advance. Using this table, we calculate the distance between the query and each feature vectors in a database, and select the closest one to the query.

## III. Experimental results

In this section, we demonstrate that the proposed method achieves higher retrieval accuracy than conventional methods. The precision-recall curve was used to evaluate retrieval performance. Precision and recall are defined by Eqs. (9) and (10). Here, $S$ is the number of correct images in the database, and we consider a situation where $R$ correct images are obtained as a result of presenting the top $T$ images as retrieval results.

**TABLE I:** Recall@100 of the proposed method

| $t$ \ $m$ | 1 | **1.1** | 1.2 | 1.3 |
|---|---|---|---|---|
| **0.0001** | N/A | **0.360** | 0.358 | 0.345 |
| 0.001 | 0.356 | 0.359 | 0.357 | 0.345 |
| 0.01 | 0.347 | 0.352 | 0.336 | 0.157 |
| 0.1 | 0.306 | 0.314 | 0.211 | 0.120 |

**TABLE II:** Quantization error and entropy

| | Quantization error | Entropy |
|---|---|---|
| PQ [3] | 1117.46 | 1.08 |
| OPQ [5] | 1531.36 | 1.61 |
| **Ours** | 1147.95 | **0.91** |

$$\text{Precision}=\frac{\text{R}}{\text{T}}, \tag{9}$$

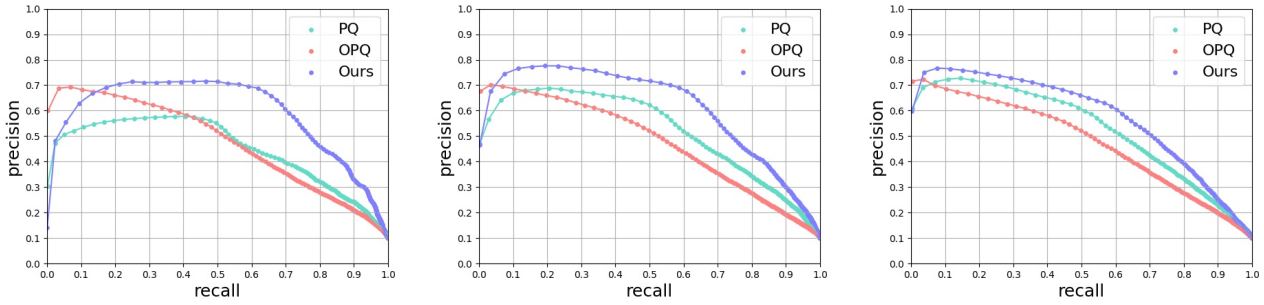
$$\text{Recall}=\frac{\text{R}}{\text{S}}. \tag{10}$$

Because there is a trade-off between precision and recall, the higher the precision-recall curve is in the upper right corner, the more the retrieval performance increases. We used 6500 images of 10 classes extracted from ImageNet [15] (650 images for each class) as a dataset. From this dataset, 400 images of each class were selected as the training data. Also, 200 images of each class were selected as the database. Finally, 50 images of each class were selected as the query images. We used intermediate layer outputs (4096 dimension) of the VGG16 [16] pre-trained on ImageNet as the image features.

Here, we describe parameters of the proposed method. We trained a codebook ($M=4, K=8$) with some values of $t$ and $m$, and measured recall@100 (recall when the top 100 results are presented as retrieval results). Then, $t$ and $m$ that maximize recall@100 were selected as the parameters of the proposed method. Table I shows recall@100. From this table, we use the value of $(t,m)=(0.0001,1.1)$ as the parameters of the proposed method.

Table II shows the values of quantization error and entropy for each quantization method at $M=4, K=8$. This table shows that the proposed method reduces the entropy compared with the conventional method. Fig. 3 shows the precision-recall curve. We can see that the retrieval performance of both methods increases as well as $K$, and the difference of retrieval performance between them decreases. When $K=8$, OPQ is the most accurate if $T\leq 40$, and the proposed method is the most accurate if $T\geq 50$. When $K=16$, the proposed method is the most accurate if $T\geq 20$, and when $K=32$, the proposed method is the most accurate if $T\geq 10$. The retrieval accuracy when $T\leq 100$, is especially important for practical use. Therefore, when $K$ is large, the proposed method is the most accurate in practice use.

Figure 4 shows an example of the retrieval results at $M=4, K=16$. While both PQ and OPQ present images of wrong labels, the proposed method presents three images of the correct labels. This is because our method reduces the entropy of the labels.




**(a)** $M = 4, K = 8$ **(b)** $M = 4, K = 16$ **(c)** $M = 4, K = 32$

**Fig. 3:** Precision-Recall curves. We confirmed that the proposed method has the highest retrieval performance.

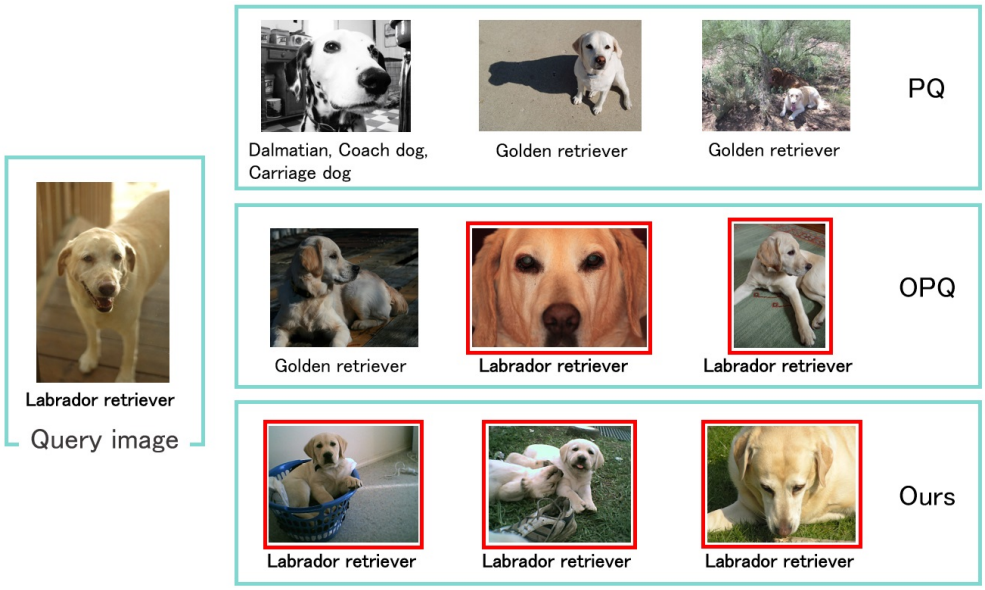



**Fig. 4:** Example of the retrieval results at $M = 4, K = 16$. The query image is shown on the left, and the top three results presented by each quantization method are shown on the right. The label of the query image is "Labrador Retriever".

Finally, we have conducted a minimum required experiments to confirm that the retrieval performance is improved by reducing the entropy. In the future, we will compare our method with other supervised and semi-supervised PQ methods [8]–[13].

## IV. Conclusions

We proposed a PQ method to reduce the entropy of labels. We assume that correct labels for each training sample are known, and learn the codebook that minimizes the label error to reduce the entropy of labels. This improves the discriminative power of images and the retrieval performance. We have confirmed that the retrieval performance is better than conventional methods by the experiment using ImageNet.

Finally, we explain the remaining issues of our method. In our proposed method, clustering is performed in each subspace during the training of the codebook. However, there is no guarantee that the solution of codebook that minimizes the entropy of the clusters in the subspaces coincides with the solution that minimizes the entropy of the clusters in the original space. Therefore, there is still room for improvement in the learning algorithm of the codebook.

## Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number JP21K17861. We wish to thank Tokyo Electric Power Company Holdings, Incorporated for supporting our research.

## References

[1] R. S. Choras, "Image feature extraction techniques and their applications for cbir and biometrics systems," *Int. Journal of Biology and Biomedical Engineering*, vol. 1, no. 1, pp. 6–16, 2007.

[2] S. Deb and Y. Zhang, "An overview of content-based image retrieval techniques," in *Proc. Int. Conf. Advanced Information Networking and Applications*, 2004, vol. 1, pp. 59–64.

[3] H. Jegou et al., "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2010.
[4] R. Gray, "Vector quantization," *IEEE ASSP Magazine*, vol. 1, no. 2, pp. 4–29, 1984.
[5] T. Ge et al., "Optimized product quantization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 744–755, 2013.
[6] A. Babenko and V. Lempitsky, "Additive quantization for extreme vector compression," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 931–938.
[7] T. Zhang, C. Du, and J. Wang, "Composite quantization for approximate nearest neighbor search," in *Proc. Int. Conf. Machine Learning*, 2014, pp. 838–846.
[8] B. Klein and L. Wolf, "End-to-end supervised product quantization for image search and retrieval," in *Proc. the IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 5041–5050.
[9] L. Gao, X. Zhu, J. Song, Z. Zhao, and H. T. Shen, "Beyond product quantization: deep progressive quantization for image retrieval," *arXiv preprint arXiv:1906.06698*, 2019.
[10] M. Liu, Y. Dai, Y. Bai, and L-Y. Duan, "Deep product quantization module for efficient image retrieval," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4382–4386.
[11] Y. K. Jang and N. I. Cho, "Generalized product quantization network for semi-supervised image retrieval," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 3420–3429.
[12] T. Yu, J. Yuan, C. Fang, and H. Jin, "Product quantization network for fast image retrieval," in *Proc. European Conf. Computer Vision (ECCV)*, 2018, pp. 186–201.
[13] X. Wang, W. Zhu, and C. Liu, "Semi-supervised deep quantization for cross-modal search," in *Proc. ACM Int. Conf. on Multimedia*, 2019, pp. 1730–1739.
[14] J. MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Proc. Berkeley Symposium on Mathematical Statistics and Probability*, 1967, vol. 1, pp. 281–297.
[15] J. Deng, W. Dong, R. Socher, L-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.