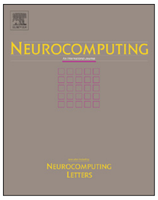
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Dynamic programming based optimized product quantization for approximate nearest neighbor search

Yuanzheng Cai [a,b], Rongrong Ji [a,b,*], Shaozi Li [a,b]

[a] Cognitive Science Department, Xiamen University, China
[b] Fujian Key Lab of the Brain-like Intelligence System, Xiamen University, China

ARTICLE INFO

Article history:
Received 13 August 2015
Received in revised form
1 December 2015
Accepted 28 January 2016

Keywords:
Approximate nearest neighbor search
Optimized product quantization
Dynamic programming
Optimal quantizer
Retrieval
Global optimal solution
Overhead analysis

ABSTRACT

Product quantization and its variances have emerged in approximate nearest neighbor search, with a wide range of applications. However, the optimized division of product subspaces retains as an open problem that largely degenerates the retrieval accuracy. In the paper, an extremely optimized product quantization scheme is introduced, which ensures, both theoretically and experimentally, a much better subspace partition comparing to the existing state-of-the-arts PQ and OPQ. The key innovation is to formulate subspace partition as a graph-based optimization problem, by which dynamic programming is leveraged to pursuit optimal quantizer learning. Another advantage is that the proposed scheme is very easily integrated with the cutting-edge multi-indexing structure, with a nearly eligible overhead in addition. We have conducted a serial of large-scale quantitative evaluations, with comparisons to a group of recent works including PQ, OPQ, and multi-Index. We have shown superior performance gain in the widely used SIFT1B benchmark, which validates the merits of the proposed algorithm.

1. Introduction

Approximate nearest neighbor (ANN) search has become a promising research direction, with emerging applications in image retrieval, video search, location-based service [33–35], and content analysis. Several existing research directions for ANN are widely studied in the literature, for instance, bag-of-words representation with inverted indexing [5,6,19,20], feature hashing [7,8], as well as binary code learning [9–11]. In brief, most existing schemes either adopt compact binary representation or inverted indexing technique to tackle the challenge of massive-scale retrieval. However, advantages in retrieval efficiency typically result in the loss in retrieval accuracy [12,13]. Subsequently, its compensation becomes a key focus in the existing works [14,2,15,16].

Among them, product quantization (PQ) [2] and its variances [1,17,18] are widely regarded as one of the most promising directions in ANN. Basically speaking, PQ represents a codebook as $c = \{\bar{c} \in c^1 \times \ldots \times c^m\}$, which is composed by the product of several sub-codebook in the corresponding dimensions. Since it can generate an exponential product space, PQ can be deployed with large-scale inverted indexing file systems [2]. Subsequently, exhaustive search can be directly conducted after PQ based representation [3,2]. In addition, it can be further integrated with multi-indexing structure [4,17] to reduce the runtime memory requirement.

Recently, PQ has been integrated with hashing in [1,17,18] to fit to the underlying data distribution. For instance, the optimized product quantization proposed in [1,17] introduces an optimized rotation to fit the data distribution. Another work refers to the so-called Cartesian *k*-means proposed by Norouzi and Fleet in [18], which comes from the same idea as [1].

However, it is also widely admitted in [1,3,17] that the subspace division in PQ serves as the key bottleneck and largely degenerates the retrieval accuracy. In the literature, several existing schemes are proposed to study this problem. For instance, OPQ [1] adopts an optimized rotation to the raw data. And for another instance, Ref. [17] adopts a group of optimized rotation to different invented-index. While progressive improvement are reported in [17], the truly "optimal" product quantization retains as an open problem, mainly due to the difficulty in a theoretical proven about the balanced eigenvalue allocation.

In this paper, we propose a novel formulation that conquers this very challenging optimal PQ problem. Our key innovation is a graph-based formulation to explain the eigenvalue allocation and dimension projection. More specially, we adopt dynamic programming to solve the problem of eigenvalue balanced partition. We term the proposed scheme Dynamic Programming based Optimized Product Quantization (DP-OPQ). To further illustrate the proposed DP-OPQ, Fig. 1 shows a graphical explanation depicting

[*] Corresponding author at: Fujian Key Lab of the Brain-like Intelligence System, Xiamen University, China.
E-mail address: jirongrong@gmail.com (R. Ji).

http://dx.doi.org/10.1016/j.neucom.2016.01.112
0925-2312/© 2016 Published by Elsevier B.V.

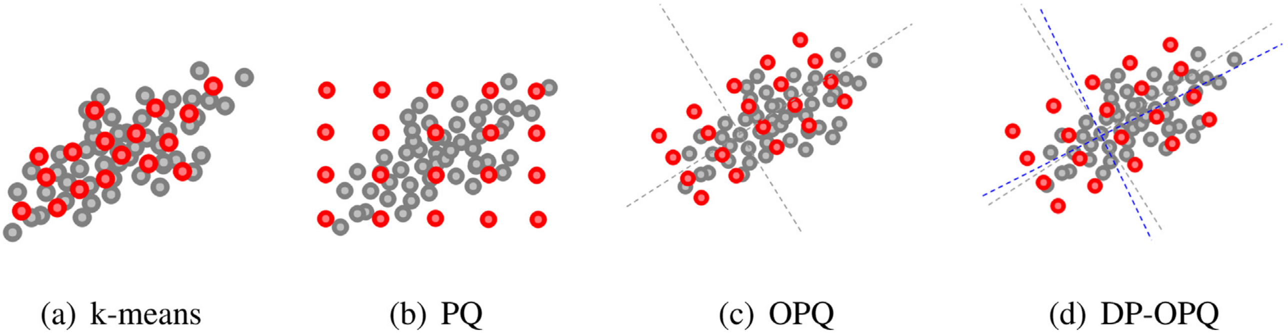


Fig. 1. The quantizers of different quantization methods for 2D data. The gray points present a set of random data, and the red points represent the centroids trained on the random data. The dash line of different colors represent the different rotations. (a) shows centroids of the k-means. (b) shows centroids of PQ. (c) shows the OPQ which optimizes PQ by a rotation. (d) shows DP-OPQ which applies a rotation to fit data better, noticed in this 2D case it is actually not different from OPQ, but we still use the dash line to represent the rotation because they will be different in high dimension space. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

our main idea. In Fig. 1(a), k-means utilizes $\log_2 k$ bits to represent arbitrary d-dimensional vector by k centroids. Despite its simplicity and efficiency, i.e. $O(dk)$, it requires a very large k to ensure high accuracy. In Fig. 1(b), PQ [2] decreases the number of centroids via partition axis into m-parts, and independently generates centroids by k-means in each group. This operation makes the centroids decrease to k^m and retain an $O(dk)$ search.

In Fig. 1(c), OPQ [1] rotates to align or adjust the dimensions of data, which balances their variance in corresponding subspaces to minimize the quantization distortion. However, such a variance balancing is *line-by-line*, a.k.a., a dimension-wised optimization scheme, which is because it uses a greedy eigenvalue allocation to balance the variance of the corresponding covariance matrix.

As shown in Fig. 1(d), our solution takes advantages of both OPQ [1] and multi-index [4] by designing a dynamic programing based scheme for eigenvalue allocation. In such a way, the subsequent PQ can minimize the distortions by a sequence of optimal rotation operations. More specially, we design two DP-OPQ schemes, i.e. a non-parametrized OPQ termed OPQ_p for data-independent quantization, as well as a parameterized OPQ termed for data-driven quantization. Moreover, we further show that this scheme can be seamlessly integrated with the cutting-edge multi-index structure [4]. Correspondingly, this optimized solution trends to make all variances to some groups and keep the sum of each variances group equals. The proposed DP-OPQ scheme merits in three-fold: first, the process of computing does not increase any additional overhead for index structure. Second, for OPQ_p, we prove that the embedding have a theoretical minimal distortion. Third, the group structure of variances would be more continuous, which is beneficial since it keeps the dimension natural order. We have carried out extensive experiments on the SIFT1B benchmark, which is a widely used benchmark for evaluating both OPQ and PQ. For the task of non-exhaustive ANN, we have reported obvious performance over the state-of-the-art OPQ scheme [1].

The rest of this paper is organized as follows: We briefly revisit related work in Section 2. Then, we introduce and analysis PQ and OPQ in Section 3 as the preliminary. In Section 4, we introduce our dynamic programming strategy to optimize the OPQ_p, and combine it with the state-of-the-art multi-index structure, quantitative experiments and comparisons to the state-of-the-arts are given in Section 5. And finally this paper concludes in Section 6.

2. Related work

With the popularity of feature designing method for large-scale image retrieval and mobile search, most state-of-the-art, VLAD [21], Fisher vector [22,23] and their derivative approaches have achieved well performance. Generally, most of these approaches rely on quantizing local feature descriptors into visual vocabulary. Approximate nearest neighbor search can apply to decrease the loss of quantization of local feature, e.g., SIFT [24]. In addition, it also can apply to global feature as a generalized search, such as many compact features [25–27] for mobile search. Approximate nearest neighbor search algorithms can be coarsely subdivided as the hashing or binary embedding based scheme [13,7,28], and the dictionary learning based [4,2,29] schemes with inverted indexing. The former aims to convert vectors into binary representation and search using Hamming distance. The later aims to convert vectors into inverted indexing by dictionary based quantization. Very recently, the works in [30,12,16] also propose to integrate both schemes. Below we briefly review related works in the recent literature and discuss our differences.

The binary encoding is widely adopted in large-scale visual search, due to its efficiencies in both storage and computing. Recent works include, but not limited to, Spectral Hashing [11], ITQ [9] or K-means Hashing [8]. Note that different from Local Sensitive Hashing [28], all above schemes are data dependent, i.e. they investigate the data distribution to learn the optimized codes. Datar et al. [7] proposed to use multiple k-means quantizers and random reinitialization to improve the performance. Norouzi et al. [18] proposed to train the Cartesian k-means for quantization. Norouzi et al. [16] further improves the online search speed by a set of hash tables which focus on binary code substrings.

As one of the most popular quantization based schemes, Product Quantization (PQ) [2] partition vector's dimensions into subdimensions in which quantization is done independently. As its improved version, OPQ is widely regarded as the state-of-the-art solution, which achieves the balance of compactness and speed, and is able to handle billion-scale dataset. Non-parametric OPQ [1] utilizes a rotation to optimize subspace decomposition and centroids of each subspace jointly. In contrast, parametric OPQ gives both derivation and proof, in a sense of local optimal, to balance the assignment of data variances to subspaces, which leads to less distortion. Similarly to OPQ, Cartesian k-means [18] also jointly optimizes the subspace decomposition and centroids. The works in [30] trains codebook and quantizes vector with recursively space decomposition. To further improve the index structure of PQ based methods, IVFADC (inverted file with asymmetric distance computation) [2] is proposed and operated by two steps, i.e. vectors are first quantized by coarse quantizer, then their residuals are quantized by PQ-encoded. Very recently, the inverted multi-index scheme proposed in [4] partitions the original feature space into two subspaces and assigns quantizers to each subspaces respectively. The multi-index also derives from the PQ framework, which can make use of the pre-trained quantizer of each subspace to