Ze-bin WU, Jun-qing YU, 2019. Vector quantization: a review. *Frontiers of Information Technology & Electronic Engineering*, 20(4):507-524. https://doi.org/10.1631/FITEE.1700833

Vector quantization: a review

Key words: Approximate nearest neighbor search; Image coding; Vector quantization

Corresponding author: Jun-qing YU
E-mail: yjqing@hust.edu.cn
ORCID: http://orcid.org/0000-0001-7057-0402

Motivation

- Vector quantization (VQ) is a very effective way to save bandwidth and storage for speech coding and image coding.
- Over the past decade, quantization-based approximate nearest neighbor (ANN) search has been developing very fast, and many methods have emerged for searching images with binary codes in the memory for large-scale datasets.

Main idea

1. Traditional vector quantization methods:

(1) Optimal vector quantization;

(2) Tree-structured VQ, direct sum VQ, Cartesian product VQ, lattice VQ, classified VQ, feedback VQ, and fuzzy VQ.

2. New codebook structure:

(1) Linear combination codebook;

(2) Joint codebook.

Classification of vector quantization

Table 1 Classification of vector quantization (VQ) methods

Codebook structure	VQ method
Tree	TSVQ (Buzo et al., 1980) HKM (Nister and Stewenius, 2006) RPT (Dasgupta and Freund, 2009) TQ (Babenko and Lempitsky, 2015)
Lattice	LVQ (Gersho, 1979) PVQ (Fischer, 1986)
Classified	CVQ (Ramamurthi and Gersho, 1986) QCVQ (Chen et al., 2014)
Feedback	Feedback VQ (Kieffer, 1982) FSVQ (Foster et al., 1985)
Direct sum	MSVQ/RVQ (Juang and Gray, 1982) ERVQ (Ai et al., 2014) PRVQ (Wei et al., 2014) RVQ-NP (Guo et al., 2016) GRVQ (Liu et al., 2017)
Cartesian product	PQ (Jégou et al., 2010) TC (Brandt, 2010) OPQ (Ge et al., 2013) CKM (Norouzi and Fleet, 2013) DPQ (Heo et al., 2014) LOPQ (Kalantidi and Avrithis, 2014) OCKM (Wang et al., 2014) PTQ (Yuan and Liu, 2015a) KSQ (Ozan et al., 2016a)
Joint	JII (Xia et al., 2013)
Linear combination	AQ (Babenko and Lempitsky, 2014) CQ (Zhang T et al., 2014) SCQ (Zhang et al., 2015) TQ (Babenko and Lempitsky, 2015) LSQ (Martinez et al., 2016) CompQ (Ozan et al., 2016b)

Tree-structured vector quantization

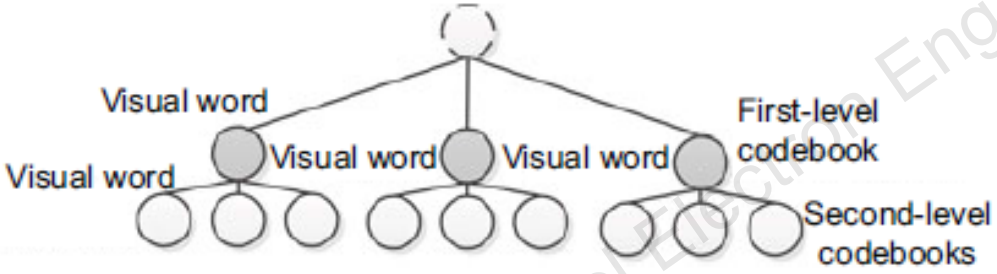


Fig. 1 Vocabulary tree of hierarchical k-means

Merits:

(1) The code length is decided by the branch factor and the tree height;

(2) The search time is also decided by the tree height.

Drawbacks:

(1) Search strategy: locally optimal algorithm for finding the nearest neighbor at each level;

(2) Codebook space: the codebook at each needs to be stored.

Direct-sum vector quantization

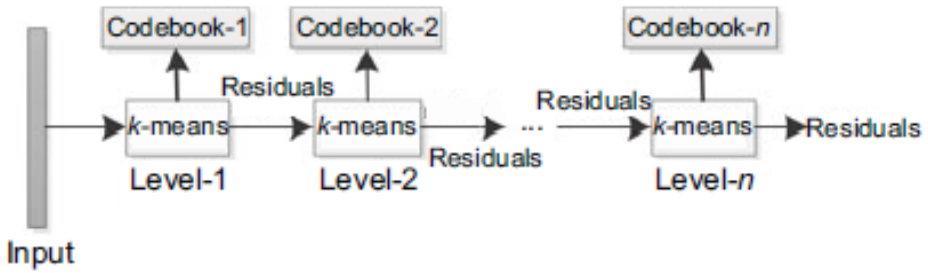


Fig. 2 Training process of multiple stage vector quantization

$$C = C_1 + C_2 + \cdots + C_L,$$

$$\hat{q} = \sum_{j=1}^{L} q_j,$$

$$I = I_1 \times I_2 \times \cdots \times I_L,$$

Properties:

(1) The overall codebook is the direct-sum of the sub-codebook of each stage.

(2) The overall index is the direct product of the index obtained at each stage.

(3) Codebook space is the product of all the stage-codebook space.

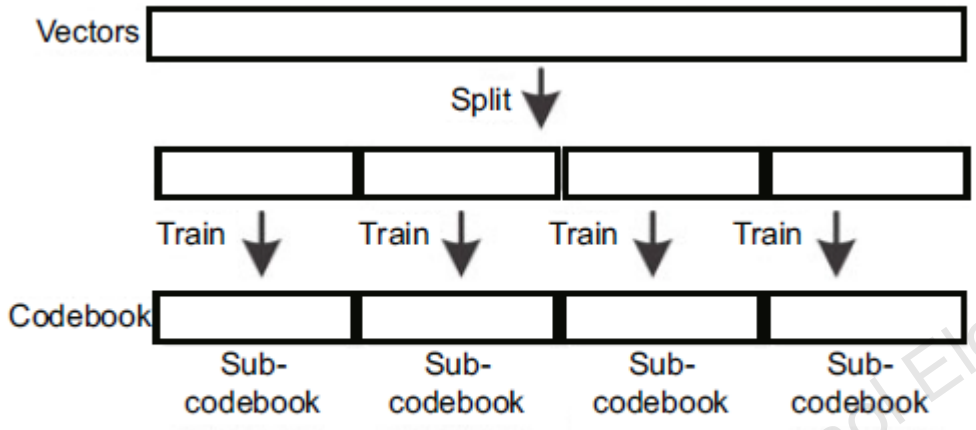
(4) Quantization loss is small.

Drawbacks:

(1) Search strategy: sequential search, locally optimal.

(2) Training overhead: the number of stages cannot be too large.

Cartesian product VQ



$$C = C_1 \times C_2 \times \cdots \times C_m.$$

Fig. 3 Training process of product vector quantization (PQ)

Merits:

(1) Training time: the computational complexity is cut down by splitting the vectors.

(2) Long codes are possible.

Drawbacks:

Space splitting: the splitting of the space may not be optimal.

Feedback VQ

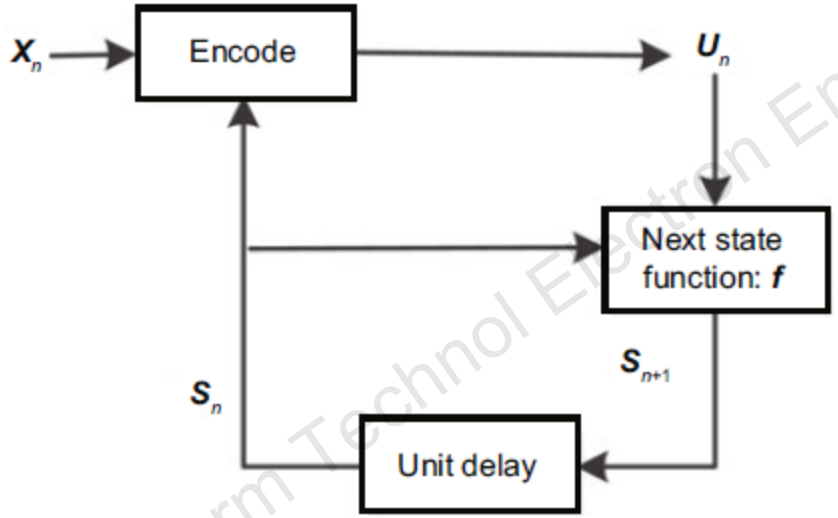


Fig. 5 Encoding process of the feedback vector quantization method

Properties:

(1) A kind of quantization method with memory;

(2) Codebook: time-varying.

Drawbacks:

(1) Structure is very complex;

(2) Less studied .

Linear combination VQ

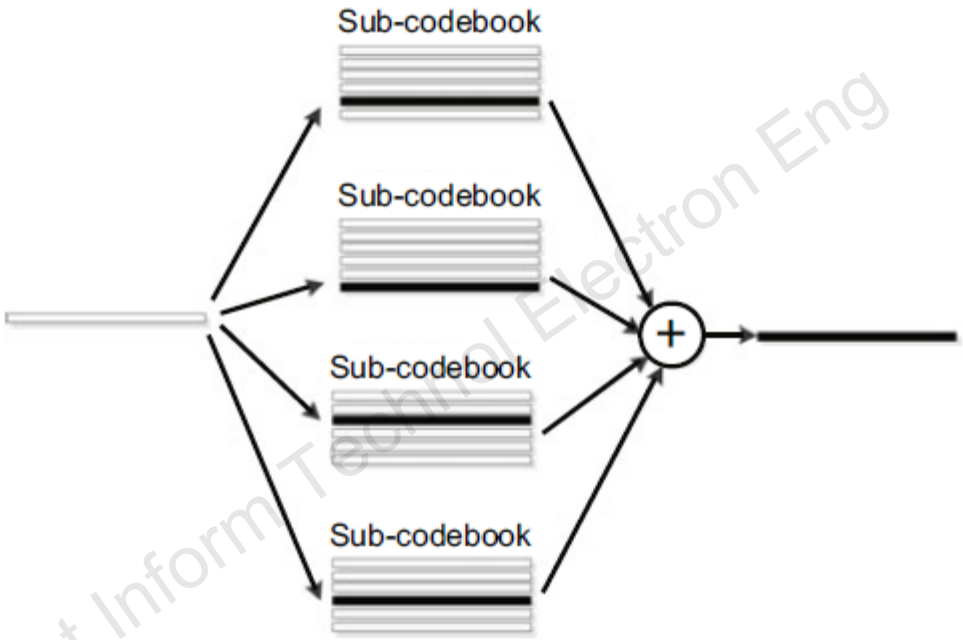


Fig. 6 Encoding process of additive quantization

Merits:

(1) Accuracy: quantization loss is small;

(2) Codebook space can be large.

Drawbacks:

Training: multiple codebooks are needed.

Joint codebook

Properties:

(1) Quantizer: multiple quantizers are used.

(2) Training: multiple quantizers are trained together to make the quantizers joint optimal.

Example:

JII: joint inverted index

Drawback:

Space consumption: larger index space is needed than in other methods.

Conclusions

1. The key difference of VQ methods lies in the codebook structure.

2. There are mainly six codebook structures, namely, tree (TSVQ), direct sum (RVQ), Cartesian product(PQ), lattice (LVQ), feedback (feedback VQ), and classied (CVQ).

3. There are two new codebook structures: linear combination and joint.

4. For ANN search application, the most extensively studied four VQ methods are tree-structured VQ, Cartesian product VQ, direct sum VQ, and linear combination VQ.

Challenges

1. Tree codebooks can be very large to increase the accuracy without loss of speed.

2. The direct sum VQ and linear combination VQ are not practical for high-dimensional large-scale datasets performed on the raw vectors.

3. A method with a lattice codebook is very fast for the regular structure of the codebook, but it is not easy to generalize for the uniform distribution assumption.

4. The feedback codebook is not very common because of its complex structure and time-varying property.

5. The multi-quantizer is not sufficiently studied because of its relatively large storage cost.