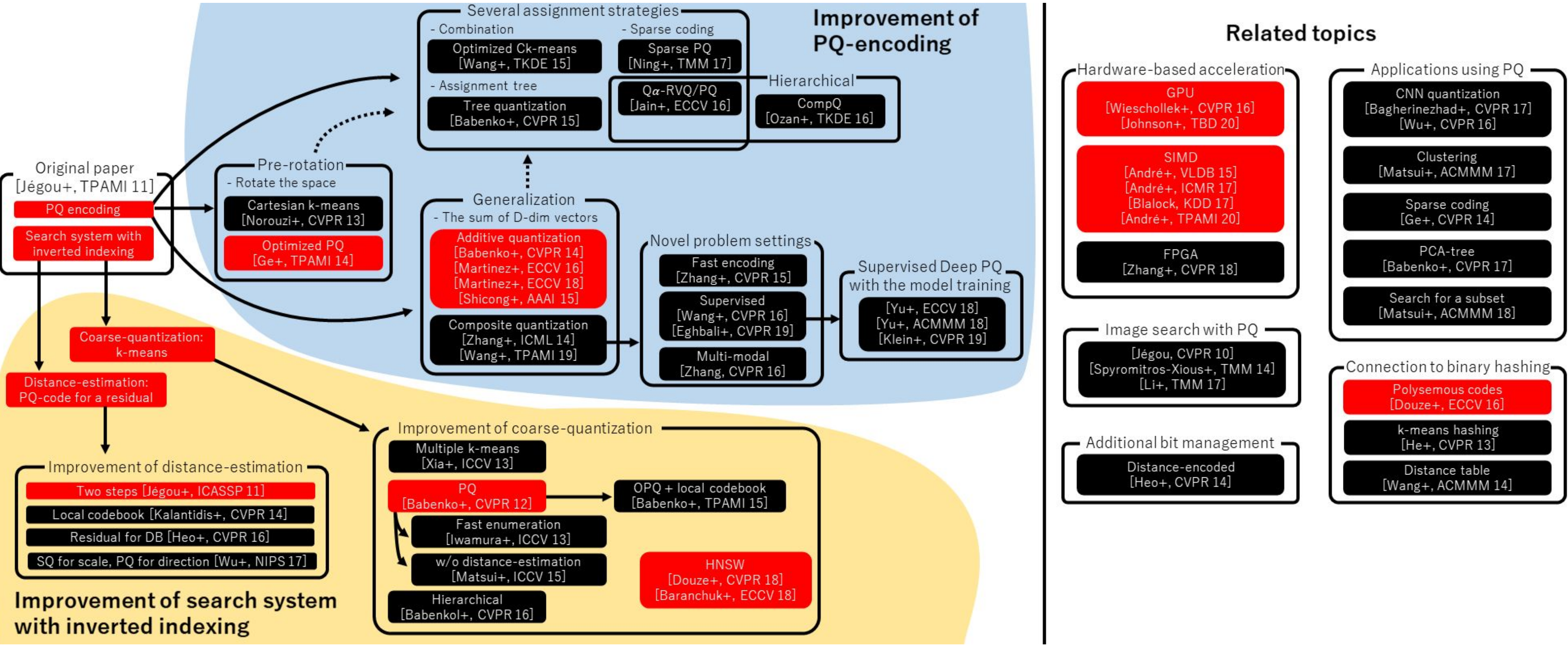
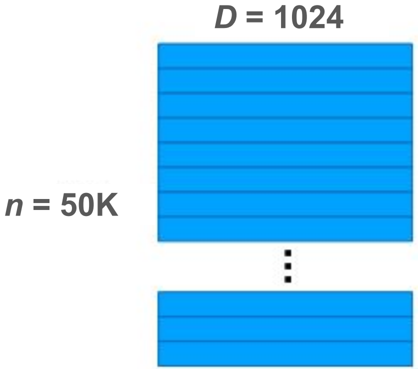


credit: Yusuke Matsui and FAISS

Optimized Product Quantization for Approximate Nearest Neighbor Search

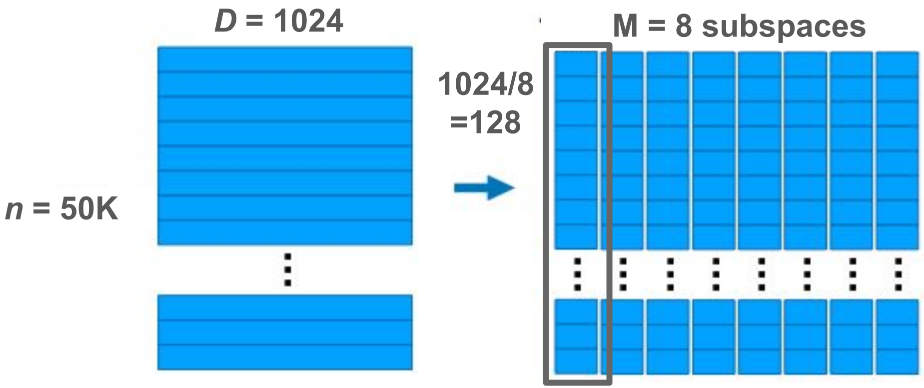
Tiezheng Ge[1]* Kaiming He[2] Qifa Ke[3] Jian Sun[2]
[1]University of Science and Technology of China [2]Microsoft Research Asia [3]Microsoft Research Silicon Valley

Recap: Product Quantization



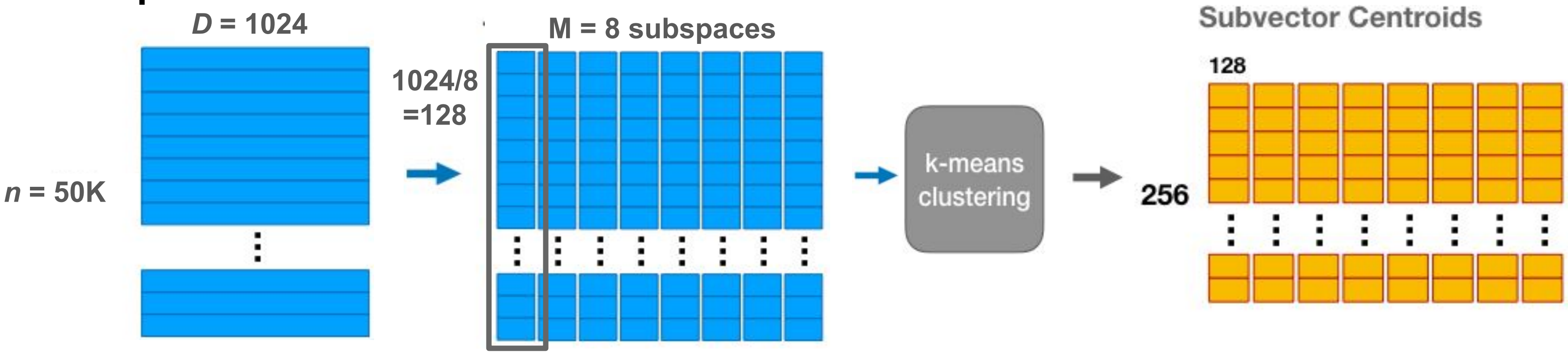
Source: https://mccormickml.com/2017/10/13/product-quantizer-tutorial-part-1/

Recap: Product Quantization



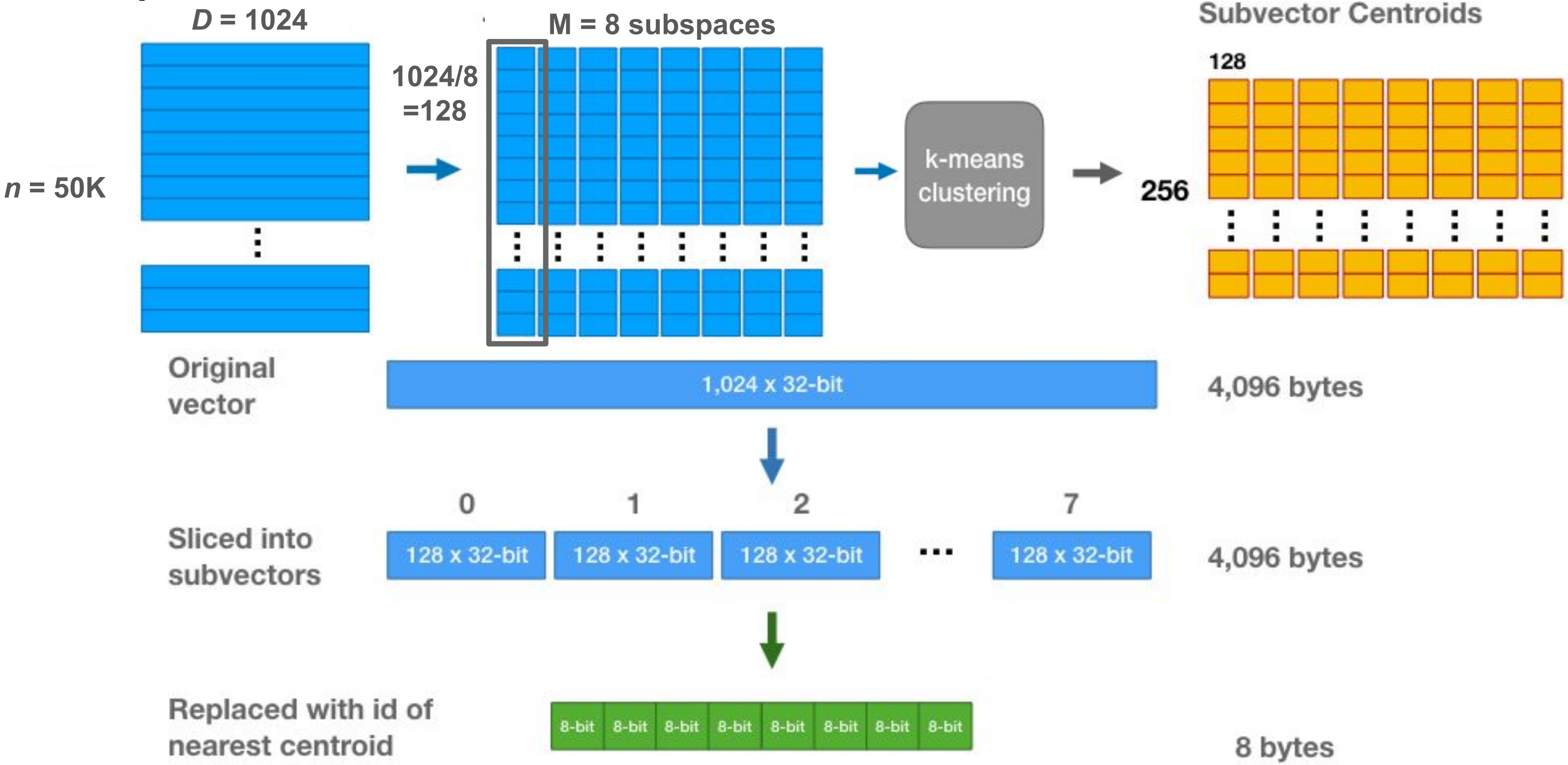
Source: https://mccormickml.com/2017/10/13/product-quantizer-tutorial-part-1/

Recap: Product Quantization



Source: https://mccormickml.com/2017/10/13/product-quantizer-tutorial-part-1/

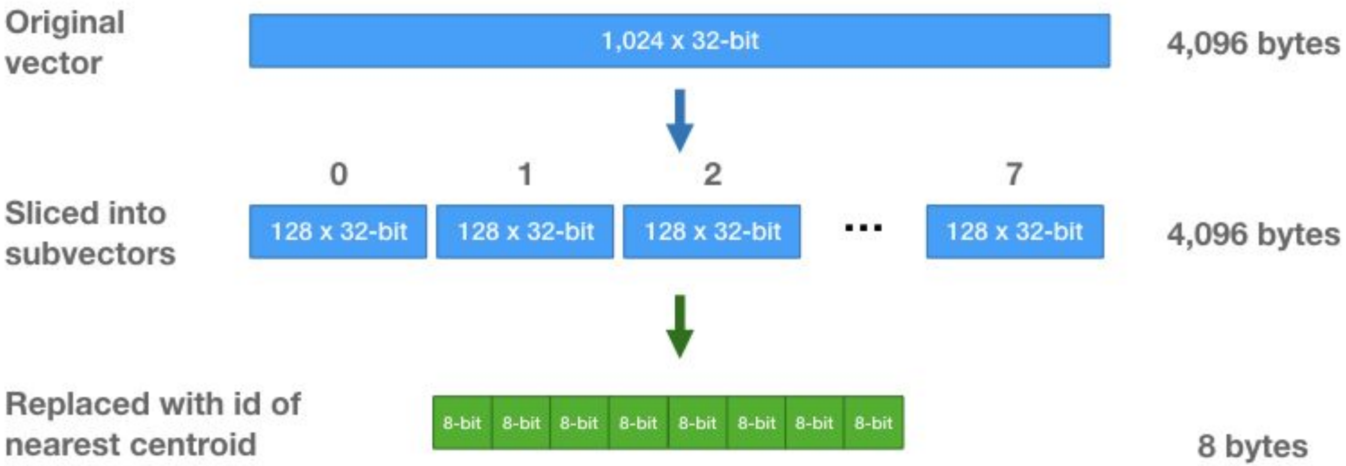
Recap: Product Quantization



Source: *https://mccormickml.com/2017/10/13/product-quantizer-tutorial-part-1/*

Recap: Product Quantization

- Formally:
  - Quantizer: $\mathbf{x} \rightarrow \mathbf{c}(i(\mathbf{x}))$
    M subvectors: $\mathbf{x} = [\mathbf{x}^1, ...\mathbf{x}^m, ...\mathbf{x}^M]$
    M sub-codewords: $\mathbf{c} = [\mathbf{c}^1, ...\mathbf{c}^m, ...\mathbf{c}^M]$

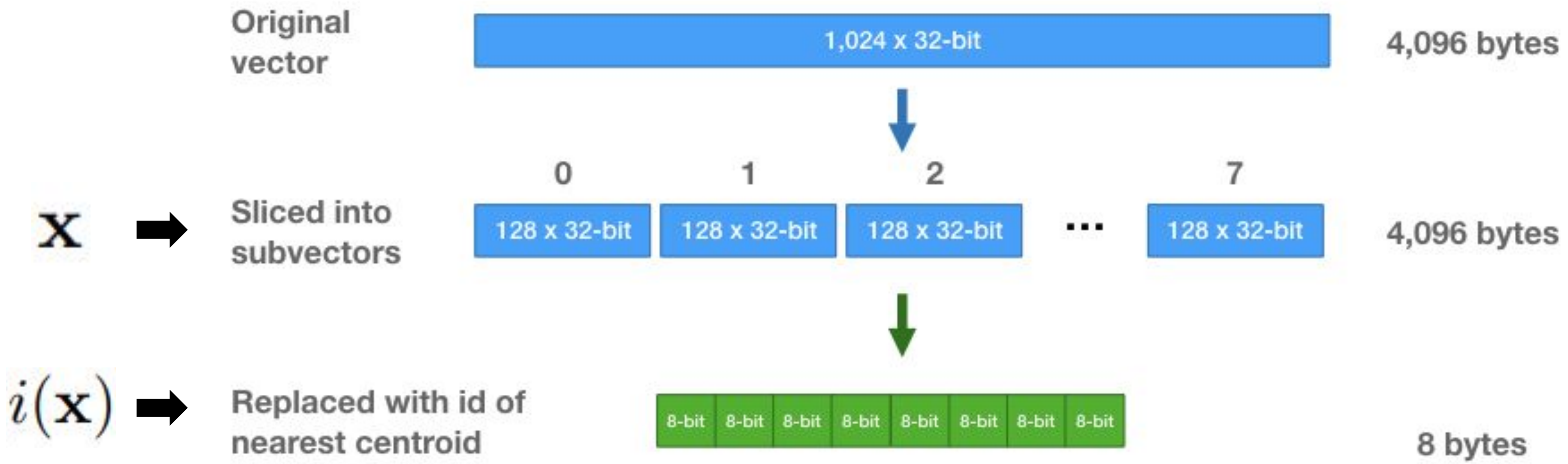


Source: https://mccormickml.com/2017/10/13/product-quantizer-tutorial-part-1/

Recap: Product Quantization

- Formally:
  - Quantizer: $\mathbf{x} \rightarrow \mathbf{c}(i(\mathbf{x}))$

$$M \text{ subvectors: } \mathbf{x} = [\mathbf{x}^1, \ldots \mathbf{x}^m, \ldots \mathbf{x}^M]$$
$$M \text{ sub-codewords: } \mathbf{c} = [\mathbf{c}^1, \ldots \mathbf{c}^m, \ldots \mathbf{c}^M]$$



Source: https://mccormickml.com/2017/10/13/product-quantizer-tutorial-part-1/

Quantization Distortion

- Formally:
  - Quantizer: $\mathbf{x} \to \mathbf{c}(i(\mathbf{x}))$
    M subvectors: $\mathbf{x} = [\mathbf{x}^1, ...\mathbf{x}^m, ...\mathbf{x}^M]$
    M sub-codewords: $\mathbf{c} = [\mathbf{c}^1, ...\mathbf{c}^m, ...\mathbf{c}^M]$
- PQ's quantization distortion (i.e. loss function)

$$\min_{\mathcal{C}^1,\ldots,\mathcal{C}^M} \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{c}(i(\mathbf{x}))\|^2,$$

$$s.t. \quad \mathbf{c} \in \mathcal{C} = \mathcal{C}^1 \times ... \times \mathcal{C}^M.$$

Quantization Distortion

- Formally:
  - Quantiz
- PQ's quantiz

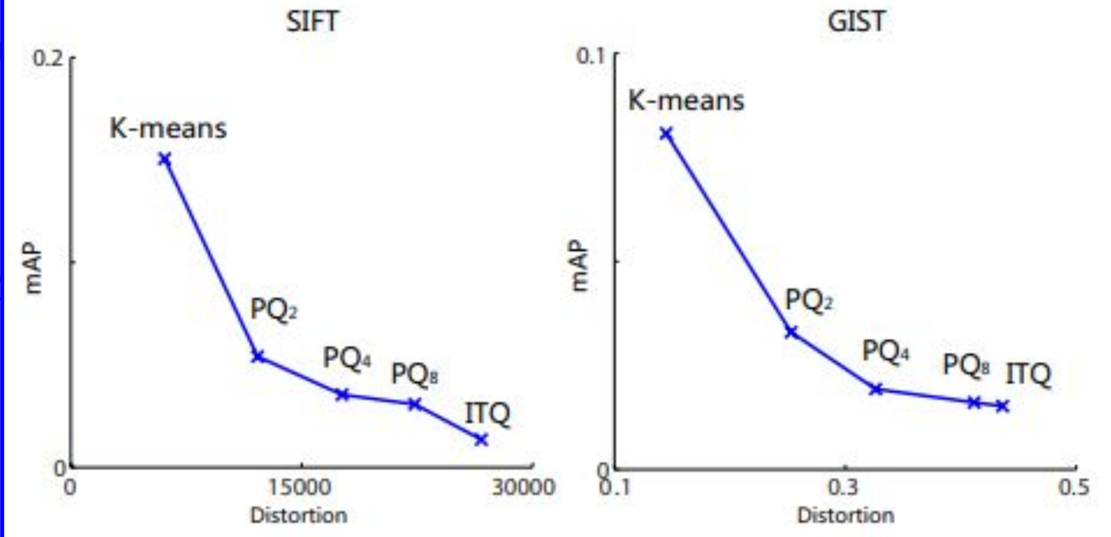
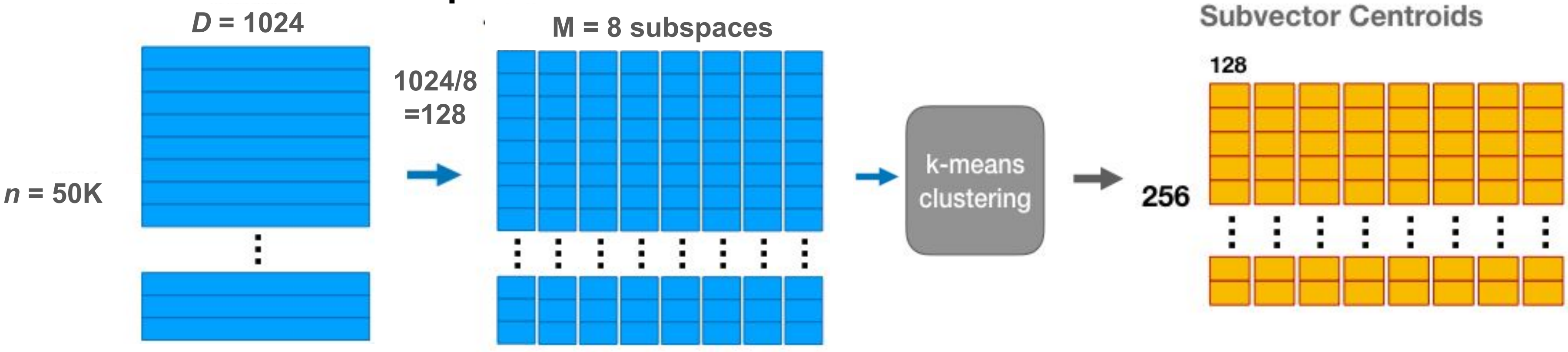
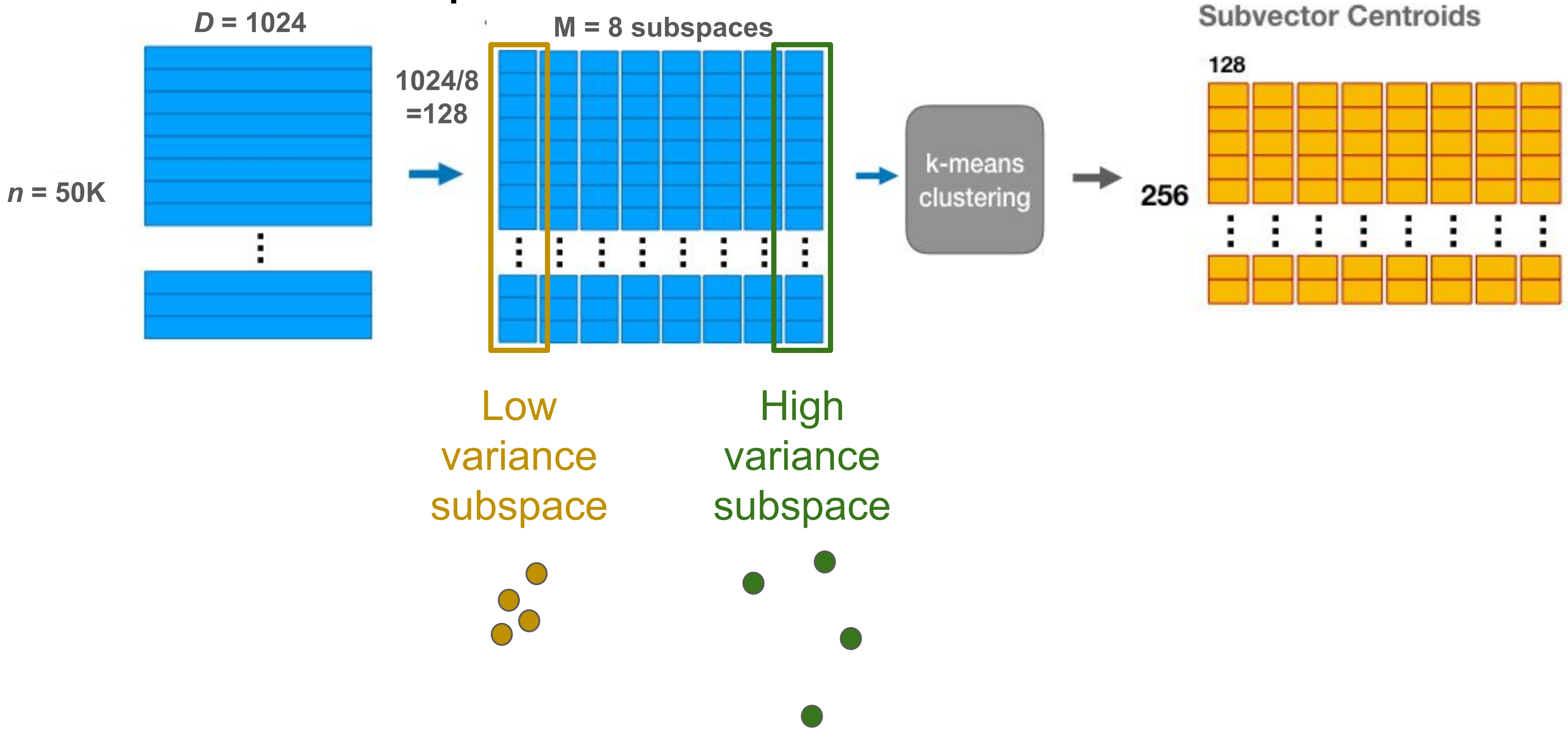


Fig. 1: mAP *vs.* quantization distortion.

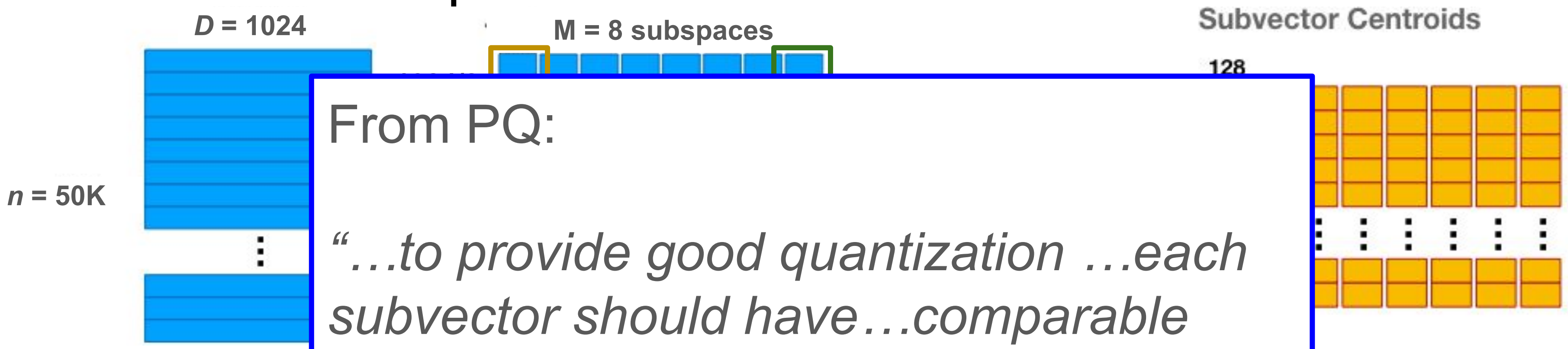
How can we improve PQ?



How can we improve PQ?



How can we improve PQ?

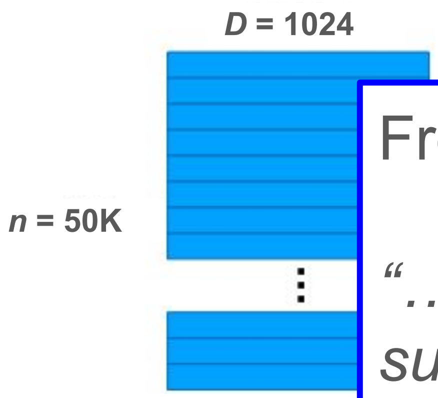


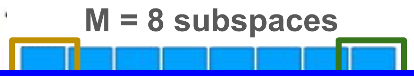
From PQ:

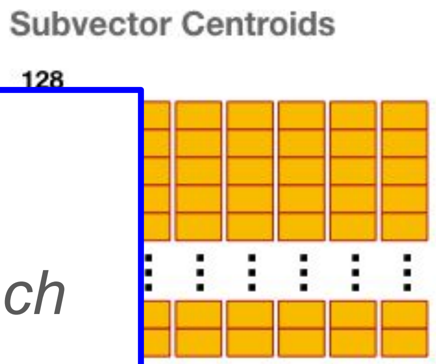
"…to provide good quantization …each subvector should have…comparable energy"

→ multiply by a random rotation matrix"

How can we improve PQ?







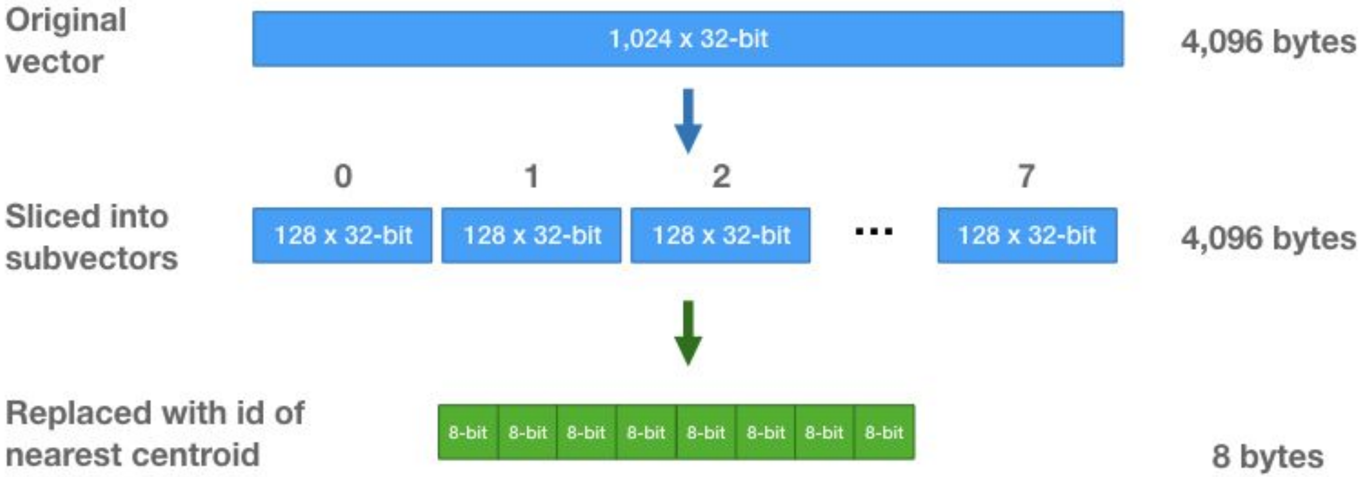
From PQ:

"…to provide good quantization …each subvector should have…comparable energy"

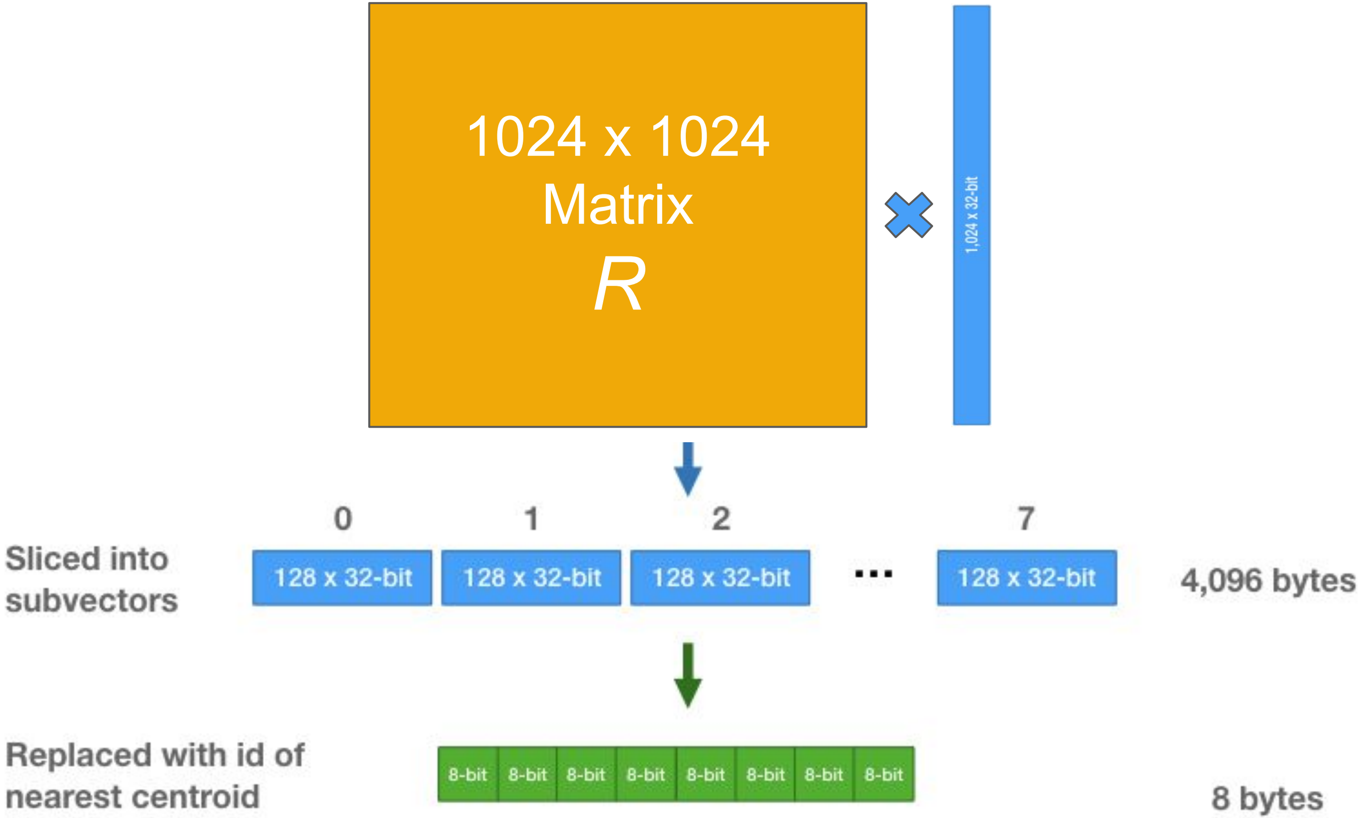
⟶ multiply by a random rotation matrix"

OPQ: Find optimal rotation matrix

Optimized Product Quantization



Optimized Product Quantization



Optimized Quantization Distortion

- PQ's quantization distortion (i.e. loss function)

$$\min_{\mathcal{C}^1,\ldots,\mathcal{C}^M} \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{c}(i(\mathbf{x}))\|^2,$$

$$s.t. \quad \mathbf{c} \in \mathcal{C} = \mathcal{C}^1 \times \ldots \times \mathcal{C}^M.$$

- Optimized PQ proposes to minimize:

$$\min_{R,\mathcal{C}^1,\ldots,\mathcal{C}^M} \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{c}(i(\mathbf{x}))\|^2,$$

$$s.t. \quad \mathbf{c} \in \mathcal{C} = \{\mathbf{c} \mid R\mathbf{c} \in \mathcal{C}^1 \times \ldots \times \mathcal{C}^M,\ R^{\mathrm{T}}R = I\}$$

How do we find matrix *R*?

1. Non-parametric
   - Optimize two easier subproblems
2. Parametric
   - Gaussian assumption
   - Still works on non-Gaussian data

Non-parametric solution to minimize distortion

$$\min_{R,\mathcal{C}^1,\ldots,\mathcal{C}^M} \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{c}(i(\mathbf{x}))\|^2,$$

$$s.t. \quad \mathbf{c} \in \mathcal{C} = \{\mathbf{c} \mid R\mathbf{c} \in \mathcal{C}^1 \times \ldots \times \mathcal{C}^M,\ \ R^{\mathrm{T}}R = I\}$$

Non-parametric solution to minimize distortion

$$\min_{R,\mathcal{C}^1,\ldots,\mathcal{C}^M} \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{c}(i(\mathbf{x}))\|^2,$$

$$s.t. \quad \mathbf{c} \in \mathcal{C} = \{\mathbf{c} \mid R\mathbf{c} \in \mathcal{C}^1 \times \ldots \times \mathcal{C}^M, \;\; R^{\mathrm{T}}R = I\}$$

1. Fix *R,* optimize codebooks

$$\min_{\mathcal{C}^1,\ldots,\mathcal{C}^m} \sum_{\hat{\mathbf{x}}} \|\hat{\mathbf{x}} - \hat{\mathbf{c}}(i(\hat{\mathbf{x}}))\|^2,$$

$$s.t. \quad \hat{\mathbf{c}} \in \mathcal{C}^1 \times \ldots \times \mathcal{C}^M.$$

Non-parametric solution to minimize distortion

$$\min_{R,\mathcal{C}^1,\ldots,\mathcal{C}^M} \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{c}(i(\mathbf{x}))\|^2,$$

$$s.t. \quad \mathbf{c} \in \mathcal{C} = \{\mathbf{c} \mid R\mathbf{c} \in \mathcal{C}^1 \times \ldots \times \mathcal{C}^M, \;\; R^{\mathrm{T}}R = I\}$$

1. Fix *R*, optimize codebooks

$$\min_{\mathcal{C}^1,\ldots,\mathcal{C}^m} \sum_{\hat{\mathbf{x}}} \|\hat{\mathbf{x}} - \hat{\mathbf{c}}(i(\hat{\mathbf{x}}))\|^2,$$

$$s.t. \quad \hat{\mathbf{c}} \in \mathcal{C}^1 \times \ldots \times \mathcal{C}^M.$$

2. Fix codebooks, optimize *R*

$$\min_{R} \sum_{\mathbf{x}} \|R\mathbf{x} - \hat{\mathbf{c}}(i(\hat{\mathbf{x}}))\|^2,$$

$$s.t. \quad R^{\mathrm{T}}R = I.$$

Non-parametric solution to minimize distortion

$$\min_{R,\mathcal{C}^1,\ldots,\mathcal{C}^M} \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{c}(i(\mathbf{x}))\|^2,$$

$$s.t. \quad \mathbf{c} \in \mathcal{C} = \{\mathbf{c} \mid R\mathbf{c} \in \mathcal{C}^1 \times \ldots \times \mathcal{C}^M, \;\; R^{\mathrm{T}} R = I\}$$

1. Fix *R*, optimize codebooks

Same as PQ

⟶Solve with K-means

2. Fix codebooks, optimize *R*

$$\min_{R} \sum_{\mathbf{x}} \|R\mathbf{x} - \hat{\mathbf{c}}(i(\hat{\mathbf{x}}))\|^2,$$

$$s.t. \quad R^{\mathrm{T}} R = I.$$

Non-parametric solution to minimize distortion

$$\min_{R, \mathcal{C}^1, \ldots, \mathcal{C}^M} \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{c}(i(\mathbf{x}))\|^2,$$

$$s.t. \quad \mathbf{c} \in \mathcal{C} = \{\mathbf{c} \mid R\mathbf{c} \in \mathcal{C}^1 \times \ldots \times \mathcal{C}^M, \; R^{\mathrm{T}} R = I\}$$

1. Fix *R*, optimize codebooks

$\min_{\mathcal{C}^1,\ldots,}$ $))\|^2,$

Same as PQ

⟶Solve with K-means

$s.t.$ $^M.$

2. Fix codebooks, optimize *R*

$\min_R$ $))\|^2,$

$$\min_R \|RX - Y\|_{\mathrm{F}}^2$$

$$XY^{\mathrm{T}} = USV^{\mathrm{T}}$$

$$R = VU^{\mathrm{T}}$$

Non-parametric solution to minimize distortion

$$\min_{R, \mathcal{C}^1, \dots, \mathcal{C}^M} \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{c}(i(\mathbf{x}))\|^2,$$

$$s.t. \quad \mathbf{c} \in \mathcal{C} = \{\mathbf{c} \mid R\mathbf{c} \in \mathcal{C}^1 \times \dots \times \mathcal{C}^M, \; R^{\mathrm{T}}R = I\}$$

1. Fix *R*, optimize codebooks

Same as PQ

$\longrightarrow$Solve with K-means

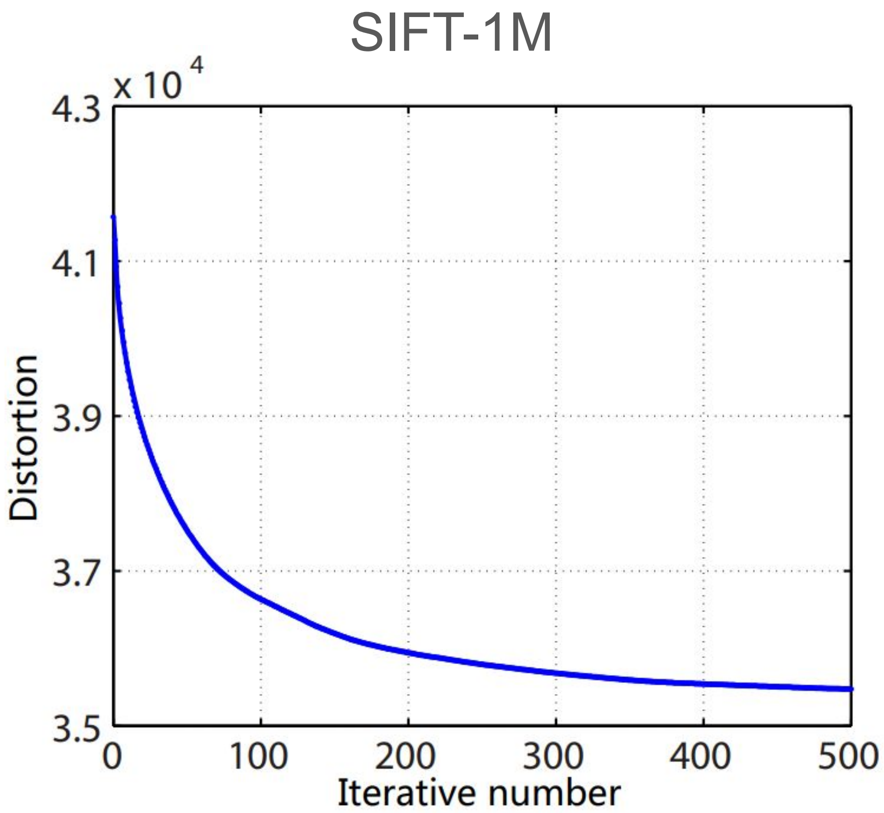
2. Fix codebooks, optimize *R*

$$\min_R \|RX - Y\|_{\mathrm{F}}^2$$

$$XY^{\mathrm{T}} = USV^{\mathrm{T}}$$

$$R = VU^{\mathrm{T}}$$

Non-parametric solution to minimize distortion



Parametric Solution

- If data is Gaussian, distortion *E* of PQ is:

$$E_{\mathrm{PQ}} = k^{-\frac{2M}{D}} \frac{D}{M} \sum_{m=1}^{M} |\Sigma_{mm}|^{\frac{M}{D}}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1M} \\ \vdots & \ddots & \vdots \\ \Sigma_{M1} & \cdots & \Sigma_{MM} \end{pmatrix}.$$

- Lower bound of distortion:

$$\sum_{m=1}^{M} |\hat{\Sigma}_{mm}|^{\frac{M}{D}} \geq M|\Sigma|^{\frac{1}{D}}.$$

Parametric Solution

- If data is Gaussian, distortion *E* of PQ is:

$$E_{\mathrm{PQ}} = k^{-\frac{2M}{D}} \cdots \begin{matrix} \cdots & \Sigma_{1M} \\ \ddots & \vdots \\ \cdots & \Sigma_{MM} \end{matrix} \Bigg) .$$

- Lower bound of d

$m=1$

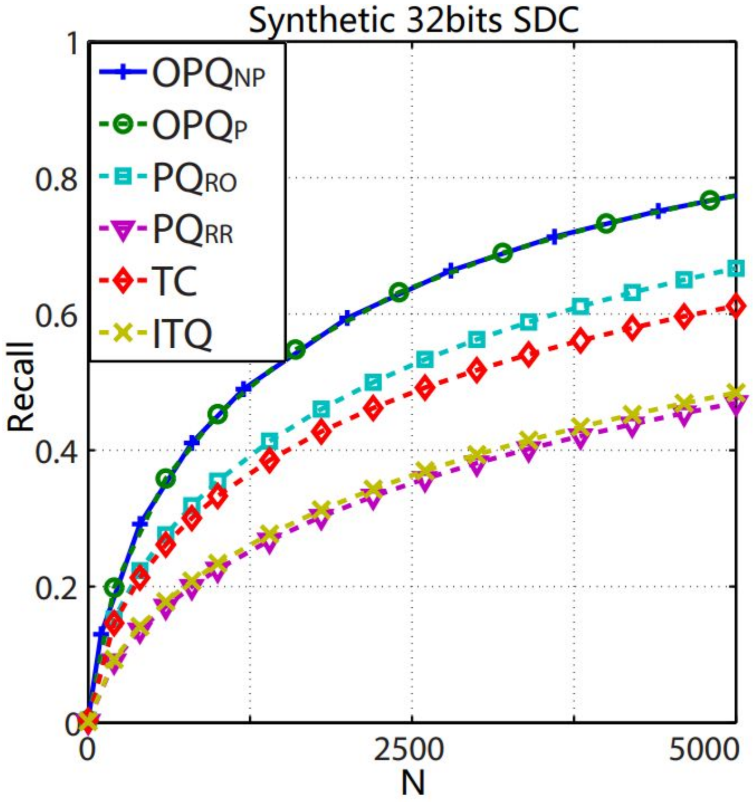
Minimal distortion with:

1. Vector dimension independence
2. Balanced subspaces variance

Evaluation

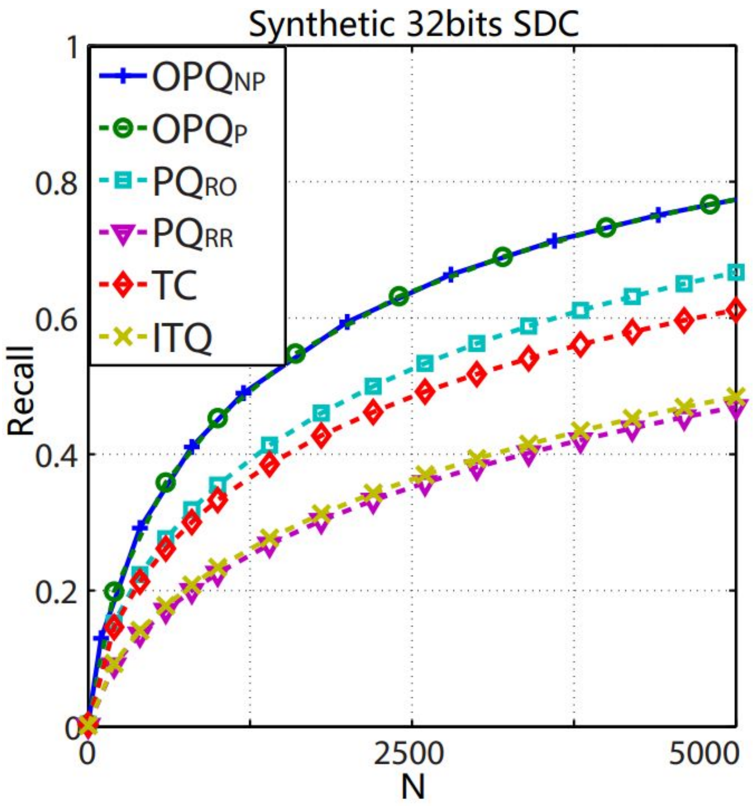
- Datasets: Synthetic Gaussian dataset, GIST1M, SIFT1M, MNIST,
- Compare OPQ_P and OPQ_NP with:
  - PQ_RO: randomly ordered dimensions
  - PQ_RR: PCA alignment then random rotation
  - TC: scalar quantizer
  - ITQ: vector quantizer

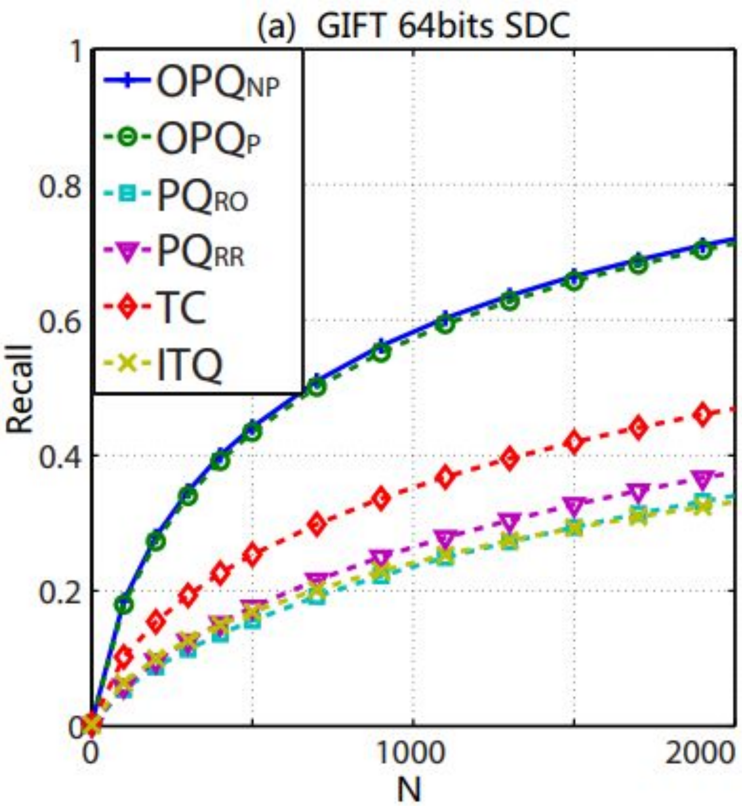
Evaluation



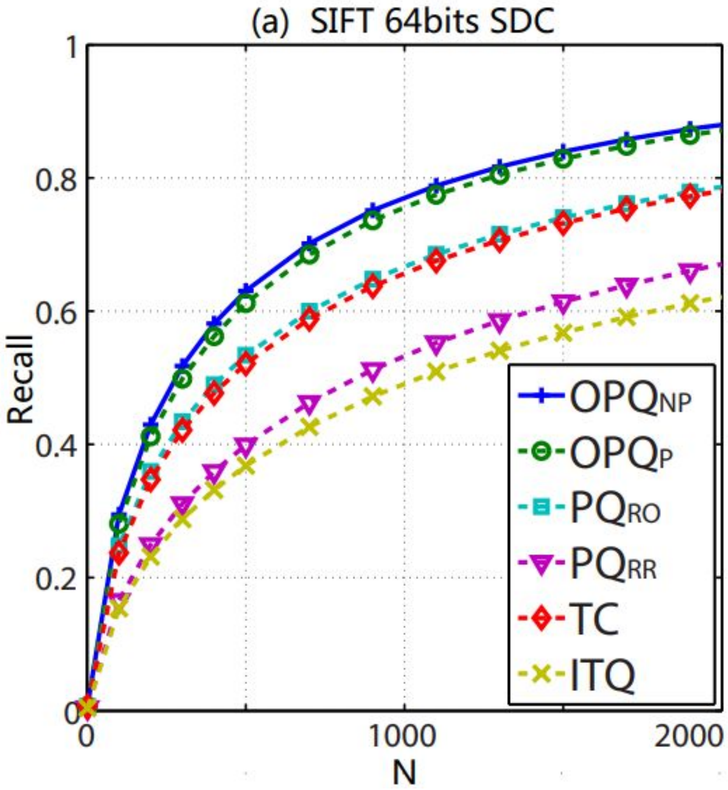
- Synthetic Gaussian data
  - 128D
  - 1M data points
- PQ_RO better than PQ_RR
  - Vector dimension independence

Evaluation


Synthetic 32bits SDC
Recall
N
OPQNP
OPQP
PQRO
PQRR
TC
ITQ
0
2500
5000
0
0.2
0.4
0.6
0.8
1

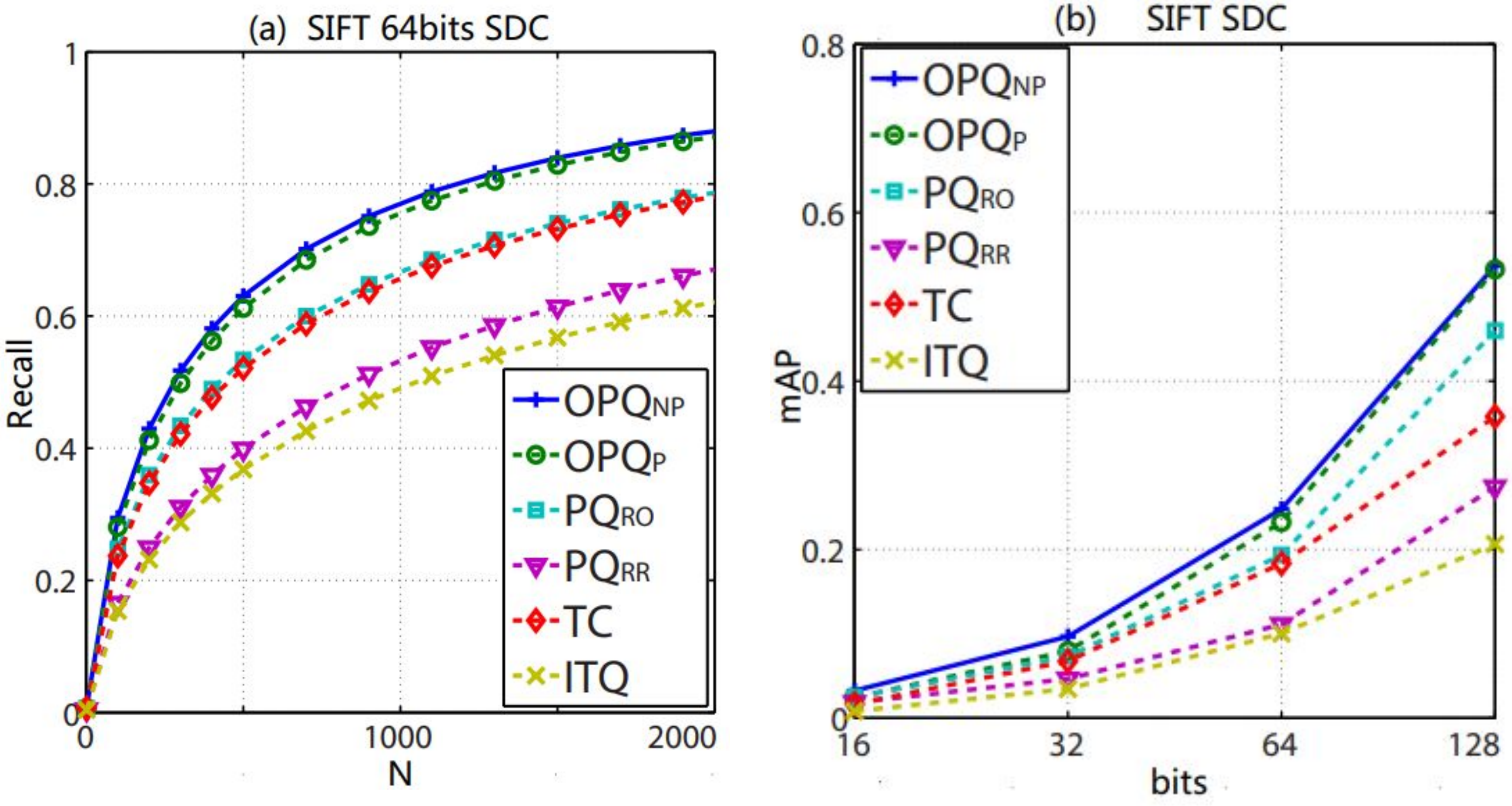

(a) GIFT 64bits SDC
Recall
N
OPQNP
OPQP
PQRO
PQRR
TC
ITQ
0
1000
2000
0
0.2
0.4
0.6
0.8
1

Evaluation

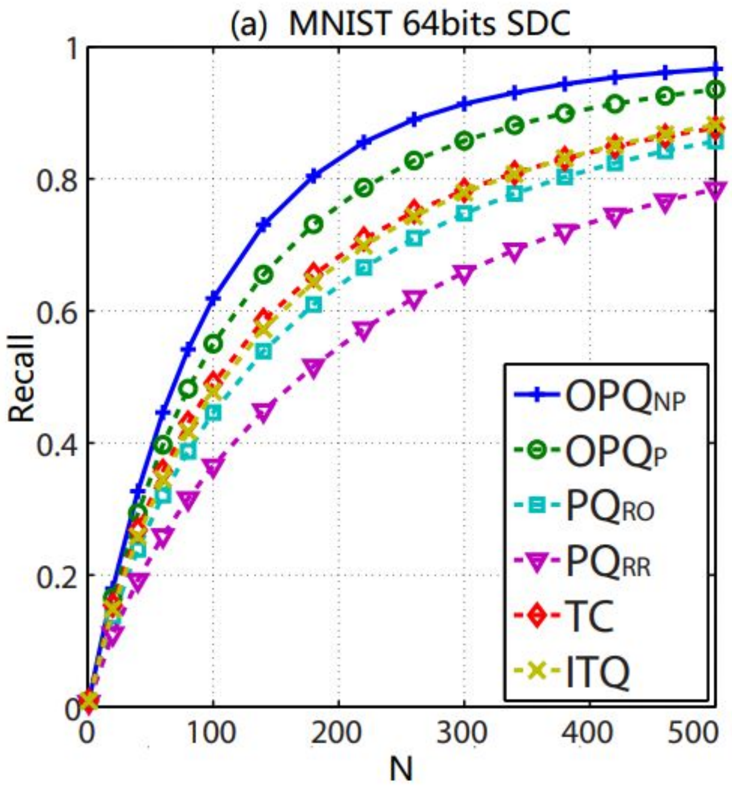


- SIFT1M: two distinct clusters
- OPQ_NP begins to outperform OPQ_P

Evaluation


(a) SIFT 64bits SDC
Recall
N
0
1000
2000
0
0.2
0.4
0.6
0.8
1
OPQNP
OPQP
PQRO
PQRR
TC
ITQ
(b) SIFT SDC
mAP
bits
16
32
64
128
0
0.2
0.4
0.6
0.8
OPQNP
OPQP
PQRO
PQRR
TC
ITQ

Evaluation



- MNIST: ten distinct clusters
- Greater difference between OPQ_NP and OPQ_P
- More complicated data?

Takeaways

- PQ is sensitive to data distribution!
- Optimizing transformation matrix can improve PQ accuracy
- Gaussian solution:
  - Independent vectors dimensions
  - Balanced subspace variance
- Limitations:
  - No non-parametric convergence guarantee
  - No evaluation of overhead
  - Gaussian assumption

Extras: Overhead

Table 4. The Indexing Time for the GIST Dataset

	RaBitQ	PQ	OPQ	LSQ
Time	117s	105s	291s	time-out (>24 hours)

RaBitQ: Quantizing High-Dimensional Vectors with a Theoretical Error Bound for Approximate Nearest Neighbor Search, JIANYANG GAO, CHENG LONG