Large-Scale Approximate Nearest Neighbor Search

Jingdong Wang
Senior Researcher
Microsoft Research, Beijing, China
Outline

• Overview
• Neighborhood graph Search
• Quantization
  • Composite quantization
  • Supervised quantization
  • Multi-modality quantization
• Application
Outline

• Overview
• Neighborhood graph Search
• Quantization
  • Composite quantization
  • Supervised quantization
  • Multi-modality quantization
• Application
Similar Image Search
Particular Object Retrieval
Similarity Search

Problem definition

$$\text{NN}(q) = \arg \min_{x \in \mathcal{X}} \text{dist}(q, x)$$

$$\{x_1, x_2, \ldots, x_n\}$$

Euclidean distance

Time complexity: $O(nd)$
Principles

Reduce #(distance computations)
- Time complexity: $O(n'd)$, $n' \ll n$
- Tree, neighborhood graph, inverted index
  
1. High efficiency 😊
2. Large memory cost 😞

Reduce the cost of each distance computation
- Time complexity: $O(nd')$, $d' \ll d$
- Compact codes (hashing, quantization)
  
1. Small memory cost 😊
2. Low efficiency 😞

Three key factors: Time, accuracy, memory
Our work

Index Structure

- Trinary-projection tree (CVPR10, TPAMI14)

  - KD tree: Single coordinate axis
  - TP tree: Trinary combination of coordinate axes
  - PCA tree: arbitrary axes

- Neighborhood graph construction and search (ACMMM12, CVPR12, ICCV13)

Compact Coding

- Complementary hashing (ICCV11)
- Order preserving hashing (ACMMM13)
- Optimized distance for binary code ranking (ACMMM14)
  - Composite quantization (ICML14, TPAMI)
  - Sparse composite quantization (CVPR15)
  - Collaborative Quantization for Cross-Modal Similarity Search (CVPR16)
  - Supervised Quantization for Similarity Search (CVPR16)
  - A survey on learning to hash (2015, TPAMI)

| x1   | 010101010110111 |
| x2   | 101010101010101 |
| x3   | 110110101010010 |
| x4   | 001101011010110 |
| x5   | 101011010101010 |
| x6   | 101010101010101 |
| x7   | 110101011010110 |
Outline

• Overview
• Neighborhood graph Search
• Quantization
  • Composite quantization
  • Supervised quantization
  • Multi-modality quantization
• Application
Neighborhood graph search

Point b is a nearest neighbor of point a
Point a is near to query q
⇒ b is most likely to be near to q

\[ \|q - b\|_2 \leq \|q - a\|_2 + \|a - b\|_2 \]

Neighborhood graph as index structure:
Each database point is connected with its k-nearest points
Greedy search

1. Visit the neighborhood points of point $v$
2. Select the point as $v$ that is the nearest to the query from the neighborhood points
3. Iterate step 1 and step 2

Too greedy to make good use of the other points in the neighborhood except the point nearest to the query
Search

**Greedy search**

1. Visit the neighborhood points of point \( v \)
2. Select the point as \( v \) that is the nearest to the query from the neighborhood points
3. Iterate step 1 and step 2

Too greedy to make good use of the other points in the neighborhood except the point nearest to the query
Best-First Search

Best-first search

1. Push the neighborhood points of $v$ into a priority queue
2. Pop out the best point $v$ from the queue
3. Iterate step 1 and step 2

A priority queue storing the points in the neighborhood
Easy Implementation

Modify few codes of depth-first search

1. procedure DFS(G, v):
   2.   let S be a Stack
   3.   S.push(v)
   4.   while S is not empty
   5.       v = S.pop()
   6.       label v as discovered
   7.       for all edges from v to w in G.adjacentEdges(v) do
   8.           if w is not labeled as discovered:
   9.               S.push(w)

1. procedure BFS(G, s, q, R):
   2.   let S be a priority queue
   3.   S.key = dist(s, q)
   4.   R.push(s)
   5.   S.push(s)
   6.   while S is not empty
   7.       v = S.pop()
   8.       label v as discovered
   9.       for all edges from v to w in G.adjacentEdges(v) do
   10.          if w is not labeled as discovered:
   11.             w.key = dist(w, q)
   12.             R.push(w)
   13.             S.push(w)
Nonlocal Search

• **Best-first search is local**
  - Stuck at a locally optimal point
  - Exhaustive neighborhood expansions

• **Iterative local search**
  - Iterative query-driven new starting point generation when the local search cannot find better points
  - Generate new starting points using kd-trees (ACMNNM12), product quantization (ICCV13)

SIFT 1M

Accuracy vs. average query time for different methods:

- ICCV13
- ACM12
- CVPR10

Legend:
- Our Approach
- Iterative Graph
- Arya
- Inverted Multi-Index
- TP Tree
- FLANN
- Split Tree
- VP Tree
GIST 1M

![Graph showing accuracy over average query time for different approaches including ICCV13, ACMMM12, CVPR10, and others. The X-axis represents average query time ranging from 0 to 5, and the Y-axis represents accuracy ranging from 0.1 to 1.0. The graph compares the performance of our approach with Iterative Graph, Arya, Inverted Multi-Index, TP Tree, FLANN, Split Tree, and VP Tree.]
Evaluation on 20M 100D CNN Features

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Query time (ms)</th>
<th>Precision@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamming distance + ITQ</td>
<td>83</td>
<td>0.93</td>
</tr>
<tr>
<td>FLANN (OpenCV)</td>
<td>46</td>
<td>0.93</td>
</tr>
<tr>
<td>FLANN (ours)</td>
<td>29</td>
<td>0.93</td>
</tr>
<tr>
<td>Composite NN Search</td>
<td>13.5</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Neighborhood Graph Search</strong></td>
<td><strong>7.4</strong></td>
<td><strong>0.95</strong></td>
</tr>
</tbody>
</table>

20M 100D float in database, 75K queries

Summary

- Neighborhood graph is better than KD-tree, Hierarchical k-means tree
- Memory cost is large

Construction

- Improvement by relative neighborhood graph
Outline

• Overview
• Neighborhood graph Search

• Quantization
  • Composite quantization
  • Supervised quantization
  • Multi-modality quantization

• Application
Outline

• Overview
• Neighborhood graph Search
• Quantization
  • Composite quantization
  • Supervised quantization
  • Multi-modality quantization
• Application

Composite Quantization for Approximate Nearest Neighbor Search. Ting Zhang, Chao Du, Jingdong Wang. ICML 2014, TPAMI 2018
Product quantization

- Approximate \( \mathbf{x} \) by the **concatenation** of \( M \) subvectors
- Code presentation: \((i_1, i_2, \ldots, i_M)\)
- Distance computation:
  - \( d(q, \mathbf{x})^2 = d(q_1, p_{1i_1})^2 + d(q_2, p_{2i_2})^2 + \cdots + d(q_M, p_{Mi_M})^2 \)
  - \( M \) additions using a pre-computed distance table

\[
\mathbf{x} = \begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_M
\end{bmatrix}
\]

\[
\begin{align*}
  p_{1i_1} & \quad \overleftarrow{d(q, \{p_{11}, p_{12}, \ldots, p_{1K}\}, q_1) = \{d(q_1, p_{11}), d(q_1, p_{12}), \ldots, d(q_1, p_{1K})\}} \quad \overleftarrow{d(q, \{p_{11}, p_{12}, \ldots, p_{1K}\}, q_1)}
  \\
  p_{2i_2} & \quad \overleftarrow{d(q, \{p_{21}, p_{22}, \ldots, p_{2K}\}, q_2)}
  \\
  \vdots & \quad \overleftarrow{d(q, \{p_{Mi_M}, \ldots, p_{Mi_M}\}, q_M)}
  \\
  p_{Mi_M} & \quad \overleftarrow{d(q, \{p_{M1}, p_{M2}, \ldots, p_{MK}\}, q_M)}
\end{align*}
\]
Product quantization

- Approximate $x$ by the concatenation of $M$ subvectors
- Codebook generation
  - Do k-means for each subspace
Product quantization

- Approximate $\mathbf{x}$ by the concatenation of $M$ subvectors
- Codebook generation
  - Do k-means for each subspace
Product quantization

- Approximate $\mathbf{x}$ by the concatenation of $M$ subvectors
- Codebook generation
  - Do k-means for each subspace
- Result in $K^M$ groups
  - The center of each is the concatenation of $M$ subvectors

$$p_{23} = p_{11}$$

$$M = 2, K = 3$$
Product quantization

• Approximate \( \mathbf{x} \) by the concatenation of \( M \) subvectors

• Codebook generation
  • Do k-means for each subspace

• Result in \( K^M \) groups
  • The center of each is the concatenation of \( M \) subvectors

\[ M = 2, K = 3 \]
Product quantization

- Approximate $\mathbf{x}$ by the concatenation of $M$ subvectors
- Codebook generation
  - Do k-means for each subspace
- Result in $K^M$ groups
  - The center of each is the concatenation of $M$ subvectors

$M = 2, K = 3$
Cartesian K-means

- Extended product quantization
  - Optimal space rotation
  - Perform PQ over the rotated space

\[
\hat{x} \approx \bar{x} = R \begin{bmatrix} p_{1i_1} \\ p_{2i_2} \\ \vdots \\ p_{Mi_M} \end{bmatrix}
\]
Cartesian K-means

- Extended product quantization
  - Optimal space rotation
  - Perform PQ over the rotated space

\[
x \approx \bar{x} = \begin{bmatrix} p_{1i_1} \\ p_{2i_2} \\ \vdots \\ p_{Mi_M} \end{bmatrix}
\]
Composite quantization (ICML 2014)

• Approximate $\mathbf{x}$ by the addition of $M$ vectors

• Code representation: $i_1 i_2 \cdots i_M$
  • length: $M \log K$

\[ x \approx \bar{x} = c_1 i_1 + c_2 i_2 + \cdots + c_M i_M \]

\[
\{c_{11}, c_{12}, \ldots, c_{1K}\} \quad \{c_{21}, c_{22}, \ldots, c_{2K}\} \quad \ldots \quad \{c_{M1}, c_{M2}, \ldots, c_{MK}\}
\]

Source codebook 1    Source codebook 2    Source codebook $M$
Composite quantization

2 source codebooks:
\{c_{11}, c_{12}, c_{13}\}
\{c_{21}, c_{22}, c_{23}\}
Composite quantization

2 source codebooks:
\{c_{11}, c_{12}, c_{13}\}
\{c_{21}, c_{22}, c_{23}\}

Composite center:
\[\star = c_{11} + c_{21}\]
Composite quantization

2 source codebooks:
\{c_{11}, c_{12}, c_{13}\}
\{c_{21}, c_{22}, c_{23}\}

Composite center:
\* = c_{11} + c_{22}
Composite quantization

2 source codebooks:
\[
\{c_{11}, c_{12}, c_{13}\} \\
\{c_{21}, c_{22}, c_{23}\}
\]

Composite center:
\[
\times = c_{11} + c_{23}
\]
Composite quantization

\[ \{c_{11}, c_{12}, c_{13}\} \]
\[ \{c_{21}, c_{22}, c_{23}\} \]

More composite centers
Composite quantization

Composite codebook: 9 composite centers
Composite quantization

Source codebook:
\{c_{11}, c_{12}, c_{13}\}
\{c_{21}, c_{22}, c_{23}\}

Space partition:
9 groups
Approximate Distance Computation

Approximate distance:

\[ \| q - x \|_2^2 \approx \| q - \sum_{m=1}^{M} c_{m}m(x) \|_2^2 \]

Time-consuming
Fast Approximate Distance Computation

\[ \| q - \sum_{m=1}^{M} c_{m \cdot m(x)} \|^2 \\
= \sum_{m=1}^{M} \| q - c_{m \cdot m(x)} \|^2 - (M - 1) \| q \|^2 + \sum_{m \neq l} c_{m \cdot m(x)}^T c_{l \cdot l(x)} \]
Fast Approximate Distance Computation

\[ \| q - \sum_{m=1}^{M} c_{m_i m(x)} \|^2_2 = \sum_{m=1}^{M} \| q - c_{m_i m(x)} \|^2_2 - (M - 1) \| q \|^2_2 + \sum_{m \neq l} c_{m_i m(x)}^T c_{l_i l(x)} \]

Constant

If constant

Computing this is enough for search
Fast Approximate Distance Computation

\[ \|q - \sum_{m=1}^{M} c_{m} x_{m}(x)\|^2_2 \]

\[ = \sum_{m=1}^{M} \|q - c_{m} x_{m}(x)\|^2_2 - (M-1)\|q\|^2_2 + \sum_{m \neq l} c_{m}^{T} c_{l} x_{m}(x) \]

Minimize quantization error:
\[ \|x - \sum_{m=1}^{M} c_{m} x_{m}(x)\|^2_2 \]

Subject to
the third term is a constant

Near-orthogonal composite quantization (NOCQ)
Formulation

- Constrained formulation

\[
\begin{align*}
\min_{\{c_m\}, \{i_m(x)\}, \epsilon} & \quad \sum_x \| x - \sum_{m=1}^M c_{m(i_m(x))} \|^2_2 \\
\text{s. t.} & \quad \sum_{m \neq l} c_{m(i_m(x))}^T c_{l(i_l(x))} = \epsilon
\end{align*}
\]

Minimize quantization error for search accuracy

Constant constraint for search efficiency
Connection

- Constrained formulation

\[
\min_{\{c_m\}, \{i_m(x)\}, \epsilon} \quad \sum_x \| x - \sum_{m=1}^{M} c_{mi_m(x)} \|^2_2 \\
\text{subject to} \quad \sum_{m \neq l} c_{mi_m(x)}^T c_{li_l(x)} = \epsilon
\]

Minimize quantization error for search accuracy

Constant constraint for search efficiency

- Product quantization and Cartesian k-means: suboptimal solutions of our approach (NOCQ)

Non-overlapped space partitioning

Codebooks are mutually orthogonal

Product quantization and Cartesian k-means

\[
\sum_{m \neq l} c_{mi_m(x)}^T c_{li_l(x)} = \epsilon
\]
Connection

Near-orthogonal composite quantization generalizes product quantization and Cartesian k-means.
A Distance Preserving View of Quantization

• Quantization
  • Data approximation: $\bar{x} \approx x$

• Better search
  • If better distance preserving: $\|q - \bar{x}\|_2 \approx \|q - x\|_2$

• Distance preserving view
  • Triangle inequality: $|\|q - x\|_2 - \|q - \bar{x}\|_2| \leq \|x - \bar{x}\|_2$
  • Minimize the upper bound: $\sum_x \|x - \bar{x}\|_2^2$
A Joint Minimization View

Generalized triangle inequality

\[ |\hat{d}(q, \bar{x}) - \hat{d}(q, x)| \leq \|x - \bar{x}\|_2 + |\delta|^{1/2} \]

Triangle inequality

\[ \|q - x\|_2 - \|q - \bar{x}\|_2 \leq \|x - \bar{x}\|_2 \]

Distortion Efficiency

\[
\begin{align*}
\hat{d}(q, \bar{x}) &= (\sum_{m=1}^{M} \|q - c_{m \cdot}(x)\|_2^2)^{1/2} \\
\hat{d}(q, x) &= (\|q - x\|_2^2 + (M - 1)\|q\|_2^2)^{1/2} \\
\delta &= \sum_{m \neq l} c_{m \cdot}(x)^{T} c_{l \cdot}(x) \\
\bar{x} &= \sum_{m=1}^{M} c_{m \cdot}(x) \\
x &= \sum_{m=1}^{M} c_{m \cdot}(x)
\end{align*}
\]
A Joint Minimization View

Generalized triangle inequality

\[ |\tilde{d}(q, \bar{x}) - \hat{d}(q, x)| \leq \|x - \bar{x}\|_2 + |\delta|^{1/2} \]

Distortion  Efficiency

Our formulation

\[
\min_{\{c_m\}, \{i_m(x)\}, \epsilon} \sum_x \|x - \bar{x}\|_2^2 \\
\text{s.t.} \quad \delta = \epsilon
\]

Minimize distortion for search accuracy

Constant constraint for search efficiency
Experiments

• **Datasets**
  • 1 million of 128D SIFT vectors, 10000 queries
  • 1 million of 960D GIST vectors, 1000 queries
  • 1 billion of 128D SIFT vectors, 1000 queries

• **Evaluation**
  • Recall@R
    • the fraction of queries for which the ground-truth Euclidean nearest neighbor is in the R retrieved items
Comparison on 1M SIFT and 1M GIST
Comparison on 1M SIFT and 1M GIST

Recall@10:
- 64 btis: 71.59%
- Our: 71.59%
- CKM: 63.83%
Comparison on 1M SIFT and 1M GIST

Our: 71.59% 64 bits
ITQ: 53.95% 128 bits

ITQ without asymmetric distance underperformed ITQ with asymmetric distance
Our approach with 64 bits outperforms (A) ITQ with 128 bits, with slightly smaller search cost
Relatively small improvement on 1M GIST might be that CKM has already achieved large improvement.
Comparison on 1B SIFT

1B SIFT, 64 bits, T = 1

1B SIFT, 128 bits, T = 1
Comparison on 1B SIFT

Recall@100:
Our: 70.12%
CKM: 64.57%
Comparison on 1M CNN

![Graphs comparing different bit levels of NOCQ, CKM, and PQ for 1MCNN and CNN feature cases with varying thresholds.](image-url)
Inner product search
Sparse Composite Quantization (CVPR 2015)

- Distance computation between a query and the dictionary element

\[ \|q - c\|^2 = \|q\|^2 - 2q^T c + \|c\|^2 \]

- The cost of distance computation: \( O(\|c\|_0) \)
- Our idea: sparsifying the dictionary elements!
CQ + Quantize the third term (TPAMI 2018)

\[
\|q - \sum_{m=1}^{M} c_{m}i_{m}(x)\|_2^2 = \sum_{m=1}^{M} \|q - c_{m}i_{m}(x)\|_2^2 - (M - 1)\|q\|_2^2 + \sum_{m \neq l} c_{m}^{T}c_{m}(x)c_{l}i_{l}(x)
\]

- Perform similarly to NOCQ for long codes
- Worse for short codes
Euclidean distance to inner product

$$\|q - x\|_2^2 = \|q\|_2^2 - 2q^\top x + \|x\|_2^2$$

Worse than NOCQ

$$= \|q\|_2^2 + [q^\top 1] \begin{bmatrix} -2x \\ y\|x\|_2^2 \end{bmatrix}$$

Encode the norm [1]

CQ (TPAMI)

Need to tune $\gamma$ carefully

Outline

• Overview
• Neighborhood graph Search
• Quantization
  • Composite quantization
  • Supervised quantization
  • Multi-modality quantization
• Application

Semantic similarity search

• Visually similar
  • Unsupervised compact coding
  • Euclidean distance

• Semantically and visually similar
  • Supervised compact coding
Background

• Supervised hashing
  • Pairwise similarity preserving
    • Align the similarity over each pair of items with the semantic similarity
  • Multiwise similarity preserving
    • Maximize the agreement of the similarity order over more than two items

• Classification
  • Semantically similar points belong to the same class after encoded

• Supervised Quantization
  • Relatively unexplored
  • First attempt to explore supervised quantization
## Categorization

<table>
<thead>
<tr>
<th>Method</th>
<th>Hash/quantization</th>
<th>Pairwise</th>
<th>Multiwise</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA hashing</td>
<td>Hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimal loss hashing</td>
<td>hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binary reconstructive embedding</td>
<td>Hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-step hashing</td>
<td>Hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FastHash</td>
<td>Hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervised deep hashing</td>
<td>Hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order preserving hashing</td>
<td>Hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triplet loss hashing</td>
<td>Hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise supervision hashing</td>
<td>Hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep semantic ranking based hashing</td>
<td>Hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDH</td>
<td>Hash</td>
<td>▲</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Our approach</strong></td>
<td><strong>Quantization</strong></td>
<td>▲</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supervised quantization

- Perform composite quantization in a discriminative space learned by a transformation matrix $P$
  - $\|P^T x - \bar{x}\|_2^2 = \|P^T x - \sum_{m=1}^{M} c_{mi}m(x)\|_2^2$
Supervised quantization

- Perform composite quantization in a discriminative space learned by a transformation matrix $P$
  
  - $\|P^T x - \bar{x}\|^2_2 = \|P^T x - \Sigma_{m=1}^{M} c_{m} m(x)\|^2_2$

- The encoded points belonging to the same class lie in a cluster
Supervised quantization

• Perform composite quantization in a discriminative space learned by a transformation matrix $P$

$$\|P^T x - \bar{x}\|_2^2 = \|P^T x - \sum_{m=1}^{M} c_{m}m(x)\|_2^2$$

• The encoded points belonging to the same class lie in a cluster

  • The center of cluster is defined by the label vector $y \in \{0,1\}^C$

  $$\|y - W^T \sum_{m=1}^{M} c_{m}m(x)\|_2^2 + \lambda \|W\|_F^2$$
Supervised quantization

- Perform composite quantization in a discriminative space learned by a transformation matrix $P$
  \[ \|P^T x - \bar{x}\|_2^2 = \|P^T x - \sum_{m=1}^{M} c_{mi(x)}\|_2^2 \]

- The encoded points belonging to the same class lie in a cluster
  - The center of cluster is defined by the label vector $y \in \{0,1\}^C$
  \[ \|y - W^T \sum_{m=1}^{M} c_{mi(x)}\|_2^2 + \lambda \|W\|_F^2 \quad \text{e.g., } C=2 \]

\begin{align*}
  y &= [0,1]^T \\
  y &= [1.0]^T
\end{align*}
Supervised quantization

• Perform composite quantization in a discriminative space learned by a transformation matrix P

\[ \| P^T x - \bar{x} \|_2^2 = \| P^T x - \sum_{m=1}^{M} c_{mi_m(x)} \|_2^2 \]

• The encoded points belonging to the same class lie in a cluster

• The center of cluster is defined by the label vector \( y \in \{0,1\}^C \)

\[ \| y - W^T \sum_{m=1}^{M} c_{mi_m(x)} \|_2^2 + \lambda \| W \|_F^2 \]  

\( \text{e.g., } C=2 \)
Supervised quantization

- Formulation

\[
\begin{align*}
\min_{W,P,\{c_m\},\{i_m(x)\},\epsilon} & \quad \sum_x \| y - W^T \sum_{m=1}^M c_{m i_m(x)} \|_2^2 + \lambda \| W \|_F^2 + \gamma \sum_x \| P^T x - \sum_{m=1}^M c_{m i_m(x)} \|_2^2 \\
\text{s.t.} & \quad \sum_{m \neq l} c_{m i_m(x)} c_{l i_l(x)} = \epsilon
\end{align*}
\]

Semantic similarity preserved by classification

- Unconstrained formulation

\[
\begin{align*}
\min_{W,P,\{c_m\},\{i_m(x)\},\epsilon} & \quad \sum_x \| y - W^T \sum_{m=1}^M c_{m i_m(x)} \|_2^2 + \lambda \| W \|_F^2 + \gamma \sum_x \| P^T x - \sum_{m=1}^M c_{m i_m(x)} \|_2^2 \\
& \quad + \mu \sum_x \left( \sum_{m \neq l} c_{m i_m(x)} c_{l i_l(x)} - \epsilon \right)^2
\end{align*}
\]
Alternative optimization

• Update $\mathbf{W}$
  • Closed form solution, $\mathbf{W} = (\mathbf{X}\mathbf{X}^T + \lambda \mathbf{I}_r)^{-1} \mathbf{X}\mathbf{Y}^T$

• Update $\mathbf{P}$
  • Closed form solution, $\mathbf{P} = (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{X}^T$

• Update $\epsilon$
  • Closed form solution, $\epsilon = \frac{1}{\#\{\mathbf{x}\}} \sum_{\mathbf{x}} \sum_{m \neq l} \mathbf{c}_{m \mathbf{m}}^T(\mathbf{x}) \mathbf{c}_{ll}(\mathbf{x})$

• Update $\{\mathbf{C}_m\}$
  • L-BFGS algorithm

• Update $\{\mathbf{i}_m(\mathbf{x})\}$
  • Iteratively alternative optimization, fixing $\{\mathbf{i}_l(\mathbf{x})\}_{l \neq m}$, update $\mathbf{i}_m(\mathbf{x})$
Experiments

- Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dimension</th>
<th>Feature type</th>
<th>#training samples</th>
<th>#test samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-10</td>
<td>512</td>
<td>GIST feature</td>
<td>59,000</td>
<td>1,000</td>
</tr>
<tr>
<td>MNIST</td>
<td>784</td>
<td>Raw pixel feature</td>
<td>69,000</td>
<td>1,000</td>
</tr>
<tr>
<td>NUS-WIDE</td>
<td>500</td>
<td>Bag-of-words feature</td>
<td>191,652</td>
<td>2,100</td>
</tr>
</tbody>
</table>
Experiments

• Compared methods
  • Supervised hashing
    • Supervised discrete hashing (SDH)
    • FastHash
    • Supervised hashing with kernels (KSH)
    • CCA-ITQ
    • Semi-supervised hashing (SSH)
    • Minimal loss hashing (MLH)
    • Binary reconstructive embedding (BRE)

  • Unsupervised quantization
    • Composite quantization (CQ)

• Evaluation: Mean average precision (MAP)
Comparison with SDH

MAP Improvement on 64 bits: **23.66%**

Relatively small improvement on MNIST because SDH already achieves a high performance

MAP Improvement on 16 bits: **4.65%**
Supervision indeed benefits the search performance

Supervision indeed benefits the search performance
Euclidean results outperforming other hashing suggests that CNN feature has powerful discriminative ability.

Our approach learns better quantizer through the supervision information.
Obtain higher performance for the same query time
Outline

• Overview
• Neighborhood graph Search
• Quantization
  • Composite quantization
  • Supervised quantization
  • Multi-modality quantization
• Application

Collaborative quantization for cross-modal similarity search. Ting Zhang, Jingdong Wang. CVPR 2016
Cross-modal similarity search

Image space $\mathcal{R}^I$

Text space $\mathcal{R}^T$
Cross-modal similarity search

Image space $\mathcal{R}^I$

Text space $\mathcal{R}^T$
Cross-modal similarity search

Image space $\mathcal{R}^I$

Text space $\mathcal{R}^T$
Compact coding approach

Image space $\mathcal{R}^I$  Common space $\mathcal{R}^C$  Text space $\mathcal{R}^T$

Perform hashing or quantization
Cross-modal similarity search

• Compact coding approaches
  • Map data from different modalities into a common space (by exploring the relations between the modalities)
    • Intra-modality relation (image vs. image and text vs. text)

• Obtain codes by performing hashing or quantization
Cross-modal similarity search

Intra-modality relation (image vs. image and text vs. text)
Cross-modal similarity search

• Compact coding approaches
  • Map data from different modalities into a common space (by exploring the relations between the modalities)
    • Intra-modality relation (image vs. image and text vs. text)
    • Inter-modality relation (image vs. text)

• Obtain codes by performing hashing or quantization
Cross-modal similarity search

Image space $\mathcal{R}^I$

Text space $\mathcal{R}^T$

Inter-modality relation (image vs. text)
Cross-modal similarity search

• Compact coding approaches
  • Map data from different modalities into a common space (by exploring the relations between the modalities)
    • Intra-modality relation (image vs. image and text vs. text)
    • Inter-modality relation (image vs. text)
    • Intra-document relation (the correspondence of an image and a text forming a document)

• Obtain codes by performing hashing or quantization
Cross-modal similarity search

Intra-document relation (the correspondence of an image and a text forming a document)
A pair of an image and a text
Cross-modal similarity search

Intra-document relation (the correspondence of an image and a text forming a document)

A pair of an image and a text

A special kind of inter-modality relation
Cross-modal similarity search

• Compact coding approaches
  • Map data from different modalities into a common space (by exploring the relations between the modalities)
    • Intra-modality relation (image vs. image and text vs. text)
    • Inter-modality relation (image vs. text)
    • Intra-document relation (the correspondence of an image and a text forming a document)
      • Inter-document relation (document vs. document)
  
• Obtain codes by performing hashing or quantization
Cross-modal similarity search

Inter-document relation (document vs. document)
Cross-modal similarity search

• Compact coding approaches
  • Map data from different modalities into a common space (by exploring the relations between the modalities)
    • Intra-modality relation (image vs. image and text vs. text)
    • Inter-modality relation (image vs. text)
    • Intra-document relation (document vs. document)
    • Inter-document relation (the correspondence of an image and a text forming a document)
  
• Obtain codes by performing hashing or quantization
  • One unified code for each document
  • Two separate codes, each corresponding to a modality
## Categorization

<table>
<thead>
<tr>
<th>Method</th>
<th>Multi-modal data relations</th>
<th>Codes</th>
<th>Coding method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intra-modality</td>
<td>Inter-modality</td>
<td>Intra-document</td>
</tr>
<tr>
<td>CMSSH</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>SCM</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>CRH</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>MMNN</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>SM^2H</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>MLBE</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>IMH</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>CVH</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>MVSH</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>SPH</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>LSSH</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>CMFH</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>STMH</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>QCH</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td>CCQ</td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>▲</td>
<td>▲</td>
<td></td>
</tr>
</tbody>
</table>
Formulation

• Common space mapping

• Text data $Y$ to common space $Y'$
  • Matrix factorization $\|Y - UY'\|$
Formulation

- Common space mapping
  - Image data X to common space
    - Sparse coding $\|X - BS\|_F^2 + \rho|S|_{11}$
  - Text data Y to common space $Y'$
    - Matrix factorization $\|Y - UY'\|$
Formulation

- Common space mapping
  - Image data X to common space $X' = RS$
    - Sparse coding $\|X - BS\|_F^2 + \rho |S|_{11}$
  - Text data Y to common space $Y'$
    - Matrix factorization $\|Y - UY'\|$
Formulation

- Common space mapping
  \[ M = \eta + \lambda \|Y' - RS\|_F^2 \]
  - Image data X to common space \( X' = RS \)
  - Sparse coding \( \|X - BS\|_F^2 + \rho |S|_{11} \)
  - Text data Y to common space Y'
    - Matrix factorization \( \|Y - UY'\| \)

Intra-document correlation
Formulation

- **Common space mapping**
  \[ M = \|X - BS\|_F^2 + \rho |S|_{11} + \eta \|Y - UY'\|_F^2 + \lambda \|Y' - RS\|_F^2 \]
  - Image data X to common space \( X' = RS \)
    - Sparse coding \( \|X - BS\|_F^2 + \rho |S|_{11} \)
  - Text data Y to common space \( Y' \)
    - Matrix factorization \( \|Y - UY'\| \)

*Image data X*  
*Text data Y*  
Intra-document correlation
Formulation

• Common space mapping
  \[ M = \|X - BS\|_F^2 + \rho \|S\|_{11} + \eta \|Y - UY'\|_F^2 + \lambda \|Y' - RS\|_F^2 \]
  • Image data X to common space \( X' = RS \)
    • Sparse coding \( \|X - BS\|_F^2 + \rho \|S\|_{11} \)
  • Text data Y to common space \( Y' \)
    • Matrix factorization \( \|Y - UY'\| \)

• Collaborative quantization
  • Image quantization \( \|X' - CP\|_F^2 \)
Formulation

- **Common space mapping**
  \[ M = \|X - BS\|_F^2 + \rho |S|_{11} + \eta \|Y - UY'\|_F^2 + \lambda \|Y' - RS\|_F^2 \]
  - Image data X to common space \( X' = RS \)
    - Sparse coding \( \|X - BS\|_F^2 + \rho |S|_{11} \)
  - Text data Y to common space \( Y' \)
    - Matrix factorization \( \|Y - UY'\| \)

- **Collaborative quantization**
  - Image quantization \( \|X' - CP\|_F^2 \)
  - Text quantization \( \|Y' - DQ\|_F^2 \)
Formulation

- **Common space mapping**
  \[ \mathcal{M} = \|X - BS\|_F^2 + \rho \|S\|_{11} + \eta \|Y - UY'\|_F^2 + \lambda \|Y' - RS\|_F^2 \]
  - Image data X to common space \(X' = RS\)
  - Sparse coding \(\|X - BS\|_F^2 + \rho \|S\|_{11}\)
  - Text data Y to common space \(Y'\)
  - Matrix factorization \(\|Y - UY'\|\)

- **Collaborative quantization**
  \[ Q = \|X' - CP\|_F^2 + \gamma \|CP - DQ\|_F^2 \]
  - Image quantization \(\|X' - CP\|_F^2\)
  - Text quantization \(\|Y' - DQ\|_F^2\)
Formulation

• Common space mapping
  \[ \mathcal{M} = \|X - BS\|_F^2 + \rho |S|_{11} + \eta \|Y - UY'\|_F^2 + \lambda \|Y' - RS\|_F^2 \]
  • Image data X to common space \(X' = RS\)
  • Sparse coding \(\|X - BS\|_F^2 + \rho |S|_{11}\)
  • Text data Y to common space \(Y'\)
    • Matrix factorization \(\|Y - UY'\|\)

• Collaborative quantization
  \[ \mathcal{Q} = \|X' - CP\|_F^2 + \|Y' - DQ\|_F^2 + \gamma \|CP - DQ\|_F^2 \]
  • Image quantization \(\|X' - CP\|_F^2\)
  • Text quantization \(\|Y' - DQ\|_F^2\)

• Overall objective function \(\mathcal{F} = \mathcal{Q} + \mathcal{M}\)
Alternative optimization \( F = Q + M \)

- Fix \( M \), update \( Q \)
  - Update each variable in \( M \) when fixing others

- Fix \( Q \), update \( M \)
  - Update each variable in \( Q \) when fixing others
Experiments

• Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Classes</th>
<th>Image feature type</th>
<th>Text feature type</th>
<th>Training samples</th>
<th>Test samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>10</td>
<td>128D SIFT vectors</td>
<td>10D topic vectors</td>
<td>2173</td>
<td>693</td>
</tr>
<tr>
<td>FLICKR25K</td>
<td>38</td>
<td>3857D vectors</td>
<td>2000D vectors</td>
<td>22500</td>
<td>2500</td>
</tr>
<tr>
<td>NUS-WIDE</td>
<td>10</td>
<td>500D BoW vectors</td>
<td>1000D vectors</td>
<td>182577</td>
<td>4000</td>
</tr>
</tbody>
</table>

• Evaluation
  • Mean average precision (MAP@T)
    • Compute MAP at T retrieved items
  • Precision@T
    • Compute precision at T retrieved items
### Results on Wiki

- **MAP@50 comparison**

<table>
<thead>
<tr>
<th>Task</th>
<th>Method</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16 bits</td>
</tr>
<tr>
<td><strong>Img toTxt</strong></td>
<td>CMSSH [1]</td>
<td>0.2110</td>
</tr>
<tr>
<td></td>
<td>CVH [8]</td>
<td>0.1947</td>
</tr>
<tr>
<td></td>
<td>MLBE [29]</td>
<td><strong>0.3537</strong></td>
</tr>
<tr>
<td></td>
<td>QCH [23]</td>
<td>0.1490</td>
</tr>
<tr>
<td></td>
<td>LSSH [30]</td>
<td>0.2396</td>
</tr>
<tr>
<td></td>
<td>CMFH [4]</td>
<td>0.2548</td>
</tr>
<tr>
<td></td>
<td>(CMFH [4])</td>
<td>(0.2538)</td>
</tr>
<tr>
<td></td>
<td>(CCQ [10])</td>
<td>(0.2513)</td>
</tr>
<tr>
<td></td>
<td>CMCQ</td>
<td>0.2478</td>
</tr>
<tr>
<td><strong>Txt toImg</strong></td>
<td>CMSSH [1]</td>
<td>0.2446</td>
</tr>
<tr>
<td></td>
<td>CVH [8]</td>
<td>0.3186</td>
</tr>
<tr>
<td></td>
<td>MLBE [29]</td>
<td>0.3336</td>
</tr>
<tr>
<td></td>
<td>QCH [23]</td>
<td>0.1924</td>
</tr>
<tr>
<td></td>
<td>LSSH [30]</td>
<td>0.5776</td>
</tr>
<tr>
<td></td>
<td>CMFH [4]</td>
<td>0.6153</td>
</tr>
<tr>
<td></td>
<td>(CMFH [4])</td>
<td>(0.6116)</td>
</tr>
<tr>
<td></td>
<td>(CCQ [10])</td>
<td>(0.6351)</td>
</tr>
<tr>
<td></td>
<td>CMCQ</td>
<td><strong>0.6397</strong></td>
</tr>
</tbody>
</table>
Results on Wiki

- Precision@T comparison on 64 bits
Results on FLICKR25K

- MAP@50 comparison

<table>
<thead>
<tr>
<th>Task</th>
<th>Method</th>
<th>FLICKR25K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16 bits</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Img to Txt</td>
<td>CMSSH [1]</td>
<td>0.6468</td>
</tr>
<tr>
<td></td>
<td>CVH [8]</td>
<td>0.6450</td>
</tr>
<tr>
<td></td>
<td>MLBE [29]</td>
<td>0.6085</td>
</tr>
<tr>
<td></td>
<td>QCH [23]</td>
<td>0.5722</td>
</tr>
<tr>
<td></td>
<td>LSSH [30]</td>
<td>0.6328</td>
</tr>
<tr>
<td></td>
<td>CMFH [4] (CMFH [4])</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(CCQ [10])</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>CMCQ</td>
<td>0.6705</td>
</tr>
</tbody>
</table>

|            |                 |          |         |         |          |
|            |                 |          |         |         |          |
|            | CMSSH [1]       | 0.6123   | 0.6400  | 0.6382  | 0.6242   |
|            | CVH [8]         | 0.6595   | 0.6507  | 0.6463  | 0.6580   |
|            | MLBE [29]       | 0.5937   | 0.6182  | 0.6550  | 0.6392   |
|            | QCH [23]        | 0.5752   | 0.6002  | 0.5757  | 0.5723   |
|            | LSSH [30]       | 0.6504   | 0.6726  | 0.6965  | 0.7010   |
|            | CMFH [4] (CMFH [4]) | —       | —      | —      | —        |
|            | (CCQ [10])      | —        | —      | —      | —        |
|            | CMCQ            | 0.7248   | 0.7335  | 0.7394  | 0.7550   |
Results on FLICKR25K

- Precision@T comparison on 64 bits
Results on NUS-WIDE

- MAP@50 comparison

<table>
<thead>
<tr>
<th>Task</th>
<th>Method</th>
<th>NUS-WIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16 bits</td>
</tr>
<tr>
<td>Img to Txt</td>
<td>CMSSH [1]</td>
<td>0.5243</td>
</tr>
<tr>
<td></td>
<td>CVH [8]</td>
<td>0.5352</td>
</tr>
<tr>
<td></td>
<td>MLBE [29]</td>
<td>0.4472</td>
</tr>
<tr>
<td></td>
<td>QCH [23]</td>
<td>0.5090</td>
</tr>
<tr>
<td></td>
<td>LSSH [30]</td>
<td>0.5368</td>
</tr>
<tr>
<td></td>
<td>CMFH [4]</td>
<td>0.4740</td>
</tr>
<tr>
<td></td>
<td>(CMFH [4])</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(CCQ [10])</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>CMCQ</td>
<td>0.5637</td>
</tr>
<tr>
<td>Txt to Img</td>
<td>CMSSH [1]</td>
<td>0.4177</td>
</tr>
<tr>
<td></td>
<td>CVH [8]</td>
<td>0.5601</td>
</tr>
<tr>
<td></td>
<td>MLBE [29]</td>
<td>0.4352</td>
</tr>
<tr>
<td></td>
<td>QCH [23]</td>
<td>0.5099</td>
</tr>
<tr>
<td></td>
<td>LSSH [30]</td>
<td>0.6357</td>
</tr>
<tr>
<td></td>
<td>CMFH [4]</td>
<td>0.5109</td>
</tr>
<tr>
<td></td>
<td>(CMFH [4])</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(CCQ [10])</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>CMCQ</td>
<td>0.6898</td>
</tr>
</tbody>
</table>
Results on NUS-WIDE

- Precision@T comparison on 64 bits
Empirical analysis

\[ \mathcal{F} = \|X' - \text{CP}\|^2_F + \|Y' - \text{DQ}\|^2_F + \gamma \|\text{CP} - \text{DQ}\|^2_F + \|X - \text{BS}\|^2_F + \rho |S|_{11} + \eta \|Y - UY'\|^2_F + \lambda \|Y' - \text{RS}\|^2_F \]

- The effect of intra-document correlation, i.e., $\gamma = 0$ vs. $\gamma \neq 0$ and $\lambda = 0$ vs. $\lambda \neq 0$

Figure 1. Illustrating the effect of the intra-document relation. The MAP is compared among CMCQ, CMCQ ($\gamma = 0$) (without correlation in the quantized space), and CMCQ ($\lambda = 0$) (without correlation in the common space) on the three datasets briefly denoted as W (Wiki), F (FLICKR25K), and N (NUS-WIDE) in the legend.
Empirical analysis

\[ \mathcal{F} = \|X' - CP\|_F^2 + \|Y' - DQ\|_F^2 + \gamma \|CP - DQ\|_F^2 + \|X - BS\|_F^2 + \rho |S|_{11} + \eta \|Y - UY'\|_F^2 + \lambda |Y' - RS\|_F^2 \]

- The effect of intra-document correlation, i.e., \( \gamma = 0 \) vs. \( \gamma \neq 0 \) and \( \lambda = 0 \) vs. \( \lambda \neq 0 \)

Figure 1. Illustrating the effect of the intra-document relation. The MAP is compared among CMCQ, CMCQ (\( \gamma = 0 \)) (without correlation in the quantized space), and CMCQ (\( \lambda = 0 \)) (without correlation in the common space) on the three datasets briefly denoted as W (Wiki), F (FLICKR25K), and N (NUS-WIDE) in the legend.
Empirical analysis

\[ \mathcal{F} = \|X' - CP\|_F^2 + \|Y' - DQ\|_F^2 + \gamma \|CP - DQ\|_F^2 + \|X - BS\|_F^2 + \rho |S|_{11} + \eta \|Y - UY'\|_F^2 + \lambda \|Y' - RS\|_F^2 \]

• The effect of dictionary, i.e., \( C = D \) vs. \( C \neq D \)

Figure 2. Illustrating the effect of the dictionary. The MAP is compared between CMCQ and CMCQ \( (C = D) \) (using one dictionary for both modalities) on the three datasets.
Empirical analysis

\[ \mathcal{F} = \|X' - CP\|_F^2 + \|Y' - DQ\|_F^2 + \gamma \|CP - DQ\|_F^2 + \|X - BS\|_F^2 + \rho |S_{11}| + \eta \|Y - UY'\|_F^2 + \lambda \|Y' - RS\|_F^2 \]

- Parameter sensitive analysis

Figure 4. Parameter sensitive analysis of our algorithm with respect to (a) \(\gamma\), (b) \(\rho\), (c) \(\eta\), and (d) \(\lambda\) over image to text (task1) and text to image (task2) on two datasets: FLICKR25K (F) and NUS-WIDE (N) with 32 bits. The dashdot line shows the best results obtained by other baseline methods and is denoted as B, e.g., B-Task1_F denotes the best baseline results over the image to text task on FLICKR25K.
Take-home message

- A quantization-based compact coding approach for cross-modal similarity search

- Learns the quantizers for both modalities by exploring the intra-document correlation
Outline

• Overview
• Neighborhood graph Search
• Quantization
  • Composite quantization
  • Supervised quantization
  • Multi-modality quantization
• Application
Application

- Bing image search
- Bing web search
- Bing cluster
- Bing Ads
- Xiaolce
- ...

Xiaolce 2.0: Image-Based Chat
[SimilarImage]突破自身的秘诀是“不要命”...0.6254609
瞅你咋地？

你瞅啥？
“这么多 ...... 很撑吧?”

“口水三千丈，肥肉日日长。”
“恭喜啊，啥时候请我们吃喜糖啊，呵呵~~”

“在顶层办公会是一种什么样的体验。”
“不是空气污染就知足吧。。。”

“你！又！粗！去！玩！还！不！带！着！我！”
微软小冰可以评价她看到的任何图像，她具有自己的观察角度，而不仅仅是解释图像内容

Beyond Image Recognition

“瞧这小舌头。。。”
“这朵黄色得都有些透明了，真美”
“大叔真努力！” “不是重度污染就知足吧”
Collaborators

• Jing Wang, You Jia, Naiyan Wang

• Ting Zhang, Xiaojuan Wang

• Guo-jun Qi, Jinhui Tang, Shipeng Li

• ...
Thanks!

Q&A

Neighborhood graph search:
Coming soon

CQ Code:
https://github.com/hellozting/
CompositeQuantization

Homepage:
https://jingdongwang2017.github.io/