Composite Correlation Quantization for Efficient Multimodal Retrieval

Mingsheng Long\textsuperscript{1}, Yue Cao\textsuperscript{1}, Jianmin Wang\textsuperscript{1}, and Philip S. Yu\textsuperscript{1,2}

\textsuperscript{1}School of Software
Tsinghua University

\textsuperscript{2}Department of Computer Science
University of Illinois, Chicago

ACM Conference on Research and Development in Information Retrieval, SIGIR 2016
Outline

1. Introduction
   - Problem
   - Effectiveness and Efficiency
   - Previous Work

2. Composite Correlation Quantization
   - Multimodal Correlation
   - Composite Quantization
   - Optimization Framework

3. Evaluation
   - Results
   - Discussion

4. Summary
Multimodal Understanding

- How to utilize multimodal data to understand our real world?
  - Isomorphic space: integration, fusion, correlation, transfer, ...
Multimodal Retrieval

- Nearest Neighbor (NN) similarity retrieval across modalities
  - Database: $\mathcal{X}^{\text{img}} = \{x_1^{\text{img}}, \ldots, x_N^{\text{img}}\}$ and Query: $q^{\text{txt}}$
  - Cross-modal NN: $\text{NN} \left( q^{\text{txt}} \right) = \min_{x^{\text{img}} \in \mathcal{X}^{\text{img}}} d \left( x^{\text{img}}, q^{\text{txt}} \right)$

**Figure:** Cross-modal retrieval: similarity retrieval across media modalities.

(a) $I \rightarrow T$ (Image Query on Text DB)  
(b) $T \rightarrow I$ (Text Query on Image DB)

Precision: 0.625
Multimodal Embedding

Multimodal embedding reduces cross-modal heterogeneity gap

- **Coupling**: $\min \sum_{i=1}^{N} d(z_{i}^{\text{img}}, z_{i}^{\text{txt}}) \rightarrow$ more flexible
- **Fusion**: $z_{i} = f(z_{i}^{\text{img}}, z_{i}^{\text{txt}}) \rightarrow$ tighter relationship

“A Tabby cat is leaning on a wooden table, with one paw on a laser mouse and the other on a black laptop”
Indexing and Hashing

- Approximate Nearest Neighbor (ANN) Search
  - Exact Nearest Neighbor Search: linear scan $O(NP)$
  - Efficient, acceptable accuracy, practical solutions

- Reduce the number of distance computations: $O(N'P)$, $N' \ll N$
  - Indexing: tree, neighborhood graph, inverted index, ...

- Reduce the cost of each distance computation: $O(NP')$, $P' \ll P$
  - Hashing: Locality-Sensitive Hashing, Spectral Hashing, ...
    - Produce a few distinct distances (curse of dimensionality)
    - Limited ability and flexibility of distance approximation

- Quantization: Vector Quantization (VQ), Iterative Quantization (ITQ), Product Quantization (PQ), Composite Quantization (CQ)
  - K-means: Impossible for medium and long codes (large $K$)
Multimodal Hashing

Previous work: separate pipeline for Multimodal Embedding and Binary Encoding → large information loss, unbalanced encoding
Problem Definition

Definition (Composite Correlation Quantization, CCQ)

Given an image set \( \{x_n^1\}_{n=1}^{N_1} \in \mathbb{R}^{P_1} \) and a text set \( \{x_n^2\}_{n=1}^{N_2} \in \mathbb{R}^{P_2} \), learn two correlation mappings \( f^1 : \mathbb{R}^{P_1} \mapsto \mathbb{R}^D \) and \( f^2 : \mathbb{R}^{P_2} \mapsto \mathbb{R}^D \) that transform images and texts into a \( D \)-dimensional isomorphic latent space, and jointly learn two composite quantizers \( q^1 : \mathbb{R}^D \mapsto \{0, 1\}^H \) and \( q^2 : \mathbb{R}^D \mapsto \{0, 1\}^H \) that quantize latent embeddings into compact \( H \)-bits binary codes.
A Latent Semantic Analysis (LSA) optimization framework

\[ x_n^v \approx R^v C^v b_n^v, \]

where \( R^v \) is correlation-maximal mapping, \( C^v \) is similarity-preserving codebook, \( b_n^v \) is compact binary code

- Multimodal Embedding: Correlation Mapping & Code Fusion
- Composite Quantization: Isomorphic Space (shared codebook)

A “simple and reliable” approach to efficient multimodal retrieval
Multimodal Correlation

- Paired data matrices: \( X^1 = [x^1_1, \ldots, x^1_N], X^2 = [x^2_1, \ldots, x^2_N] \)
- Fusion representation matrix: \( Z = [z_1, \ldots, z_N] \)
- Transformation matrices: \( R^1, R^2 \), which transform \( X \) into \( Z \)

\[
\min_{R^1, R^2, Z} \lambda_1 \left\| R^1^T X^1 - Z \right\|_F^2 + \lambda_2 \left\| R^2^T X^2 - Z \right\|_F^2
\]

(1)
This problem is **ill-posed**, which cannot be solved successfully.

\[
\min_{R^1, R^2, Z} \lambda_1 \left\| R^1 X^1 - Z \right\|_F^2 + \lambda_2 \left\| R^2 X^2 - Z \right\|_F^2
\]  

\[
Z = \frac{\lambda_1 R^1 X^1 + \lambda_2 R^2 X^2}{\lambda_1 + \lambda_2}
\]

\[
R^1 = \left( X^1 X^{1T} \right)^{-1} X^1 Z^T
\]

\[
R^2 = \left( X^2 X^{2T} \right)^{-1} X^2 Z^T
\]
Multimodal Correlation

Add the covariance maximization with orthogonal constraints

\[
\begin{align*}
\min_{R^1, R^2, Z} & \quad \lambda_1 \left( \left\| R^1^T X^1 - Z \right\|_F^2 + \left\| R^1_\perp^T X^1 \right\|_F^2 \right) \\
& \quad + \lambda_2 \left( \left\| R^2^T X^2 - Z \right\|_F^2 + \left\| R^2_\perp^T X^2 \right\|_F^2 \right) \\
\min_{R^1, R^2, Z} & \quad \lambda_1 \left\| X^1 - R^1 Z \right\|_F^2 + \lambda_2 \left\| X^2 - R^2 Z \right\|_F^2
\end{align*}
\]
Composite Quantization

- Learn $M$ codebooks: $C = [C_1, \ldots, C_M]$, each codebook has $K$ codewords $C_m = [c_{m1}, \ldots, c_{mk}]$ (cluster centroids of K-means)
- Each $z_i$ is approximated by the addition of $M$ codewords
- One per codebook, each selected by the binary assignment $b_{mi}$
- Code representation: $i_1i_2\ldots i_M$, where $i_m = \text{nz}(b_{mi})$
- Code length: $M \log_2 K$ (1-of-$K$ encoding)

$$z \approx \hat{z} = C_1b_1 + C_2b_2 + \ldots + C_Mb_M$$
$$= c_{1i_1} + c_{2i_2} + \ldots + c_{Mi_M}$$

$C_1 = [c_{11}, \ldots, c_{1K}]$  $C_2 = [c_{21}, \ldots, c_{2K}]$  \ldots  $C_M = [c_{M1}, \ldots, c_{MK}]$
Composite Quantization

- Learn $M$ codebooks: $C = [C_1, \ldots, C_M]$, each codebook has $K$ codewords $C_m = [c_{m1}, \ldots, c_{mK}]$ (cluster centroids of K-means)
- Binary code matrices: $B = [B_1; \ldots; B_M], B_m = [b_{m1}; \ldots; b_{mN}]$
- Control binary codes quality by quantization error minimization

$$
\min_{Z, C, B} \left\| Z - \sum_{m=1}^{M} C_m B_m \right\|_F^2 = \sum_{i=1}^{N} \left\| z_i - \sum_{m=1}^{M} C_m b_{mi} \right\|_2^2 \quad (8)
$$

"A Tabby cat is leaning on a wooden table, with one paw on a laser mouse and the other on a black laptop"
Composite Correlation Quantization

- Pro 1: Joint optimization: correlation, covariance & quantization
- Pro 2: Semi-Paired Data Quantization through the $\delta$ function
- Pro 3: Shared codebook & coding enables multimodal retrieval
- Pro 4: Easy configurations $H = M \log_2 K$, $D = \min(V, H)$

$$\min_{R^v, C, B^v} \sum_{v=1}^{V} \sum_{n=1}^{N_v} \lambda_v \left\| x_n^v - R^v \sum_{m=1}^{M} C_m \delta (b_{mn}^v) \right\|_2^2$$

s.t. $R^v^T R^v = I_{D \times D}$, $R^v \in \mathbb{R}^{P_v \times D}$

$$\| \delta (b_{mn}^v) \|_0 = 1, \delta (b_{mn}^v) \in \{0, 1\}^K$$

$$\delta (b_{mn}^v) = \begin{cases} b_{mn}, & n = 1 \ldots N_0 \\ b_{mn}^v, & \text{otherwise} \end{cases}$$

$v = 1 \ldots V$, $m = 1 \ldots M$, $n = 1 \ldots N_v$
Approximate Distance Computation

- Asymmetric Quantizer Distance: \( \| q^\tilde{v} - x_n^v \|_2^2 \approx \text{AQD} \left( q^\tilde{v}, x_n^v \right) \)

\[
\text{AQD} \left( q^\tilde{v}, x_n^v \right) = \left\| q^\tilde{v} - R^\tilde{v} \sum_{m=1}^{M} C_mb_{mn}^v \right\|_2^2 \\
= -2 \sum_{m=1}^{M} \langle \tilde{q}^\tilde{v}, C_mb_{mn}^v \rangle + \left\| \sum_{m=1}^{M} C_mb_{mn}^v \right\|_2^2 \\
+ \left\| \tilde{q}^\tilde{v} \right\|_2^2 + \left\| R_{\perp}^\tilde{v} q^\tilde{v} \right\|_2^2
\] (10)

- Query-specific distance lookup table: Store the distances from all \( M \times K \) codebook elements in \( C = [C_1, \ldots, C_M] \) to query \( q^\tilde{v} \)

- \( O(M) \) additions for term 1, \( O(M^2) \) or \( O(1) \) additions for term 2

- Alternative: Cosine Distance \( \cos \left( q^\tilde{v}, x_n^v \right) = \sum_{m=1}^{M} \langle \tilde{q}^\tilde{v}, C_mb_{mn}^v \rangle \)
Approximation Error Analysis

**Theorem (Approximation Error Bound)**

The error of approximating Euclidean distance with AQD is bounded by

\[ |d(\tilde{q}^v, \tilde{x}_n^v) - d(\tilde{q}^v, \hat{x}_n^v)| \leq \left\| x_n^v - R^v \sum_{m=1}^{M} C_m b_{mn}^v \right\|_2. \]  

(11)

From triangle inequality, \( |d(\tilde{q}^v, \tilde{x}_n^v) - d(\tilde{q}^v, \hat{x}_n^v)| \leq d(\tilde{x}_n^v, \hat{x}_n^v) \). Then

\[
d^2(\tilde{x}_n^v, \hat{x}_n^v) = \left\| R^{vT} x_n^v - \sum_{m=1}^{M} C_m b_{mn}^v \right\|_2^2 \\
\leq \left\| R^{vT} x_n^v - \sum_{m=1}^{M} C_m b_{mn}^v \right\|_2^2 + \left\| R_{\perp}^{vT} x_n^v \right\|_2^2 \]  

(12)

\[
= \left\| x_n^v - R^v \sum_{m=1}^{M} C_m b_{mn}^v \right\|_2^2,
\]

Quantize by max cross-modal correlation & within-modal covariance.
# Experiment Setup

- **Datasets:** NUS-WIDE, Wiki, and Flickr1M
- **Tasks:** $I \rightarrow I$, $T \rightarrow T$, $I \rightarrow T$, $T \rightarrow I$, $I \rightarrow IT$, and $T \rightarrow IT$
- **Methods:**
  - **Unsupervised hashing:** CVH, IMH
  - **Deep hashing:** CorrAE + Sign
  - **Supervised hashing:** CMSSH, SCM, QCH
- **Metrics:** MAP@$R$, Precision-Recall, Precision@$R$, Efficiency

## Table: The Statistics of Three Multimodal Benchmark Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>NUS-WIDE</th>
<th>Wiki</th>
<th>Flickr1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Set</td>
<td>195,834</td>
<td>2,866</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Query Set</td>
<td>2,000</td>
<td>693</td>
<td>1,000</td>
</tr>
<tr>
<td>Database</td>
<td>193,834</td>
<td>2,173</td>
<td>24,000</td>
</tr>
<tr>
<td>Training Set</td>
<td>10,000</td>
<td>2,173</td>
<td>975,000</td>
</tr>
</tbody>
</table>
Evaluation

Search Pipeline

- Query $q_0$
of different modality from the database
- Transformed query $q$
  common space
- Multiple codebooks $C$
- Distance lookup table
  between query and codebook elements
- Code of database vector $x$
- Distance between $q$ and $x$
- Output nearest vectors

Repeated for $N$
database vectors

Indexing for candidate pruning

leave out for future work

M. Long et al. (Tsinghua University) Composite Correlation Quantization ACM SIGIR 2016 19 / 28
MAP Results

- CCQ significantly outperforms unsupervised hashing methods (CVH, IMH) and deep hashing methods (CorrAE), and generally outperforms supervised hashing methods (CMSSH, SCM, QCH).

<table>
<thead>
<tr>
<th>Task</th>
<th>Method</th>
<th>NUS-WIDE</th>
<th>Wiki</th>
<th>Flickr1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>I → T</td>
<td>CorrAE (deep)</td>
<td>0.4699</td>
<td>0.2033</td>
<td>0.6357</td>
</tr>
<tr>
<td>I → T</td>
<td>QCH (supervised)</td>
<td>0.5050</td>
<td>0.2368</td>
<td>0.6685</td>
</tr>
<tr>
<td></td>
<td>CCQ (ours)</td>
<td><strong>0.5165</strong></td>
<td><strong>0.2371</strong></td>
<td><strong>0.7183</strong></td>
</tr>
<tr>
<td>I → IT</td>
<td>CCQ (ours)</td>
<td>0.5414</td>
<td>0.2529</td>
<td>0.6989</td>
</tr>
<tr>
<td>T → I</td>
<td>CorrAE (deep)</td>
<td>0.4634</td>
<td>0.3478</td>
<td>0.6247</td>
</tr>
<tr>
<td>T → I</td>
<td>QCH (supervised)</td>
<td>0.5389</td>
<td><strong>0.4411</strong></td>
<td>0.6485</td>
</tr>
<tr>
<td></td>
<td>CCQ (ours)</td>
<td><strong>0.5413</strong></td>
<td>0.4222</td>
<td><strong>0.7165</strong></td>
</tr>
<tr>
<td>I → IT</td>
<td>CCQ (ours)</td>
<td>0.7131</td>
<td>0.6394</td>
<td>0.7190</td>
</tr>
</tbody>
</table>
NUS-WIDE

- **Asymmetric difficulty:** $T \rightarrow T \leq T \rightarrow I \leq I \rightarrow T \leq I \rightarrow I$; If the image modality is high quality $\rightarrow$ unsupervised hashing is good.

**Figure:** Precision-recall curves on NUS-WIDE cross-modal tasks @ 32 bits.
The low quality of the image modality leads to low cross-modal retrieval performance, which fits supervised hashing methods.

Figure: Precision-recall curves on Wiki cross-modal tasks @ 32 bits.
In the presence of big data, there is strong motivation to learn accurate models from large-scale dataset (big model capacity).

**Figure:** Precision-recall curves on Flickr1M cross-modal tasks @ 32 bits.
Semi-Paired Data Quantization

- Training with semi-paired data helps as paired data is limited; semi-supervised learning is helpful for partial-modal big data.

Figure: MAP of CCQ by varying the numbers of paired points for training.
Quantization Loss and Query Efficiency

- MAP loss due to binarization/quantization is controlled by CCQ;
  Query processing efficiency is compared to the state of the arts.

(a) MAP Loss

(b) Search Efficiency

Figure: MAP loss by quantization and average search time for each query.
Scalable Training Complexity

- Scales linearly to large samples; large-scale implementation via mini-batch paradigm (load fraction of data each time) is trivial.

**Figure:** Training time and memory costs on complete Flickr1M dataset.
Cross-Modal Tradeoff Sensitivity

- Stable sensitivity is important for unsupervised cross-modal retrieval, as model selection via cross-validation is impossible.

Figure: Stable parameter sensitivity for unsupervised cross-modal retrieval.
Summary

- A composite correlation quantization for multimodal retrieval
- A seamless optimization framework of
  - Multimodal Correlation
  - Composite Quantization
- Learning bound analysis for approximate similarity retrieval

Future Work

- Multimodal Inverted Multi-Index for indexing CCQ codes
- Deep neural networks for multimodal correlation embedding

http://ise.thss.tsinghua.edu.cn/~mlong