Composite Correlation Quantization for Efficient Multimodal Retrieval

Mingsheng Long[1], Yue Cao[1], Jianmin Wang[1], and Philip S. Yu[12]

[1]School of Software
Tsinghua University

[2]Department of Computer Science
University of Illinois, Chicago

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

Outline

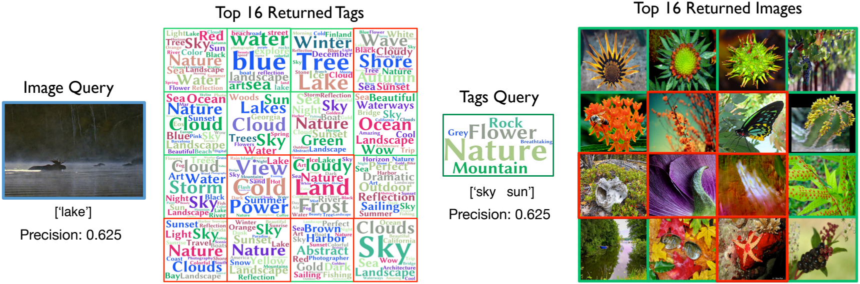
1. Introduction
   - Problem
   - Effectiveness and Efficiency
   - Previous Work
2. Composite Correlation Quantization
   - Multimodal Correlation
   - Composite Quantization
   - Optimization Framework
3. Evaluation
   - Results
   - Discussion
4. Summary

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: $\mathcal{X}^{img} = \{\mathbf{x}_1^{img}, \ldots, \mathbf{x}_N^{img}\}$ and Query: $\mathbf{q}^{txt}$
  - Cross-modal NN: $\mathrm{NN}\left(\mathbf{q}^{txt}\right) = \min_{\mathbf{x}^{img} \in \mathcal{X}^{img}} d\left(\mathbf{x}^{img}, \mathbf{q}^{txt}\right)$

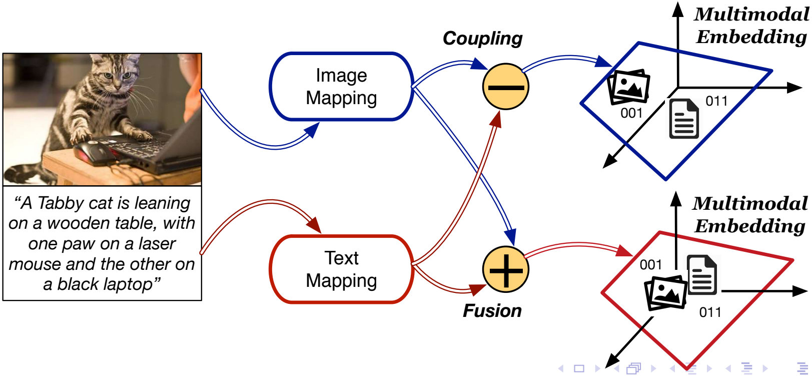


(a) $I \rightarrow T$ (Image Query on Text DB) (b) $T \rightarrow I$ (Text Query on Image DB)

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

Multimodal Embedding

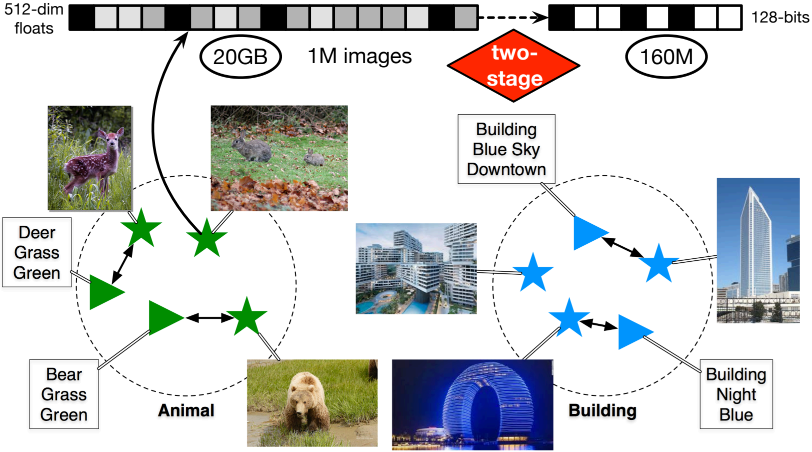
- Multimodal embedding reduces cross-modal heterogeneity gap
  - **Coupling:** $\min \sum_{i=1}^{N} d(\mathbf{z}_i^{img}, \mathbf{z}_i^{txt}) \rightarrow$ more flexible
  - **Fusion:** $\mathbf{z}_i = f(\mathbf{z}_i^{img}, \mathbf{z}_i^{txt}) \rightarrow$ tighter relationship



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

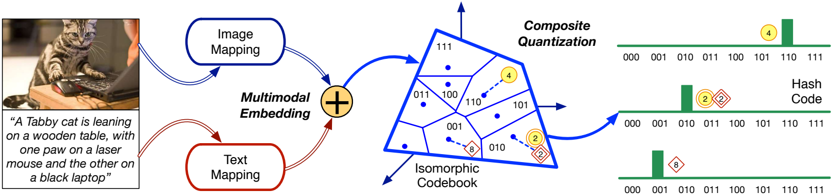
Problem Definition

Definition (Composite Correlation Quantization, CCQ)

Given an image set $\{\mathbf{x}_n^1\}_{n=1}^{N_1} \in \mathbb{R}^{P_1}$ and a text set $\{\mathbf{x}_n^2\}_{n=1}^{N_2} \in \mathbb{R}^{P_2}$, 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.

Overview

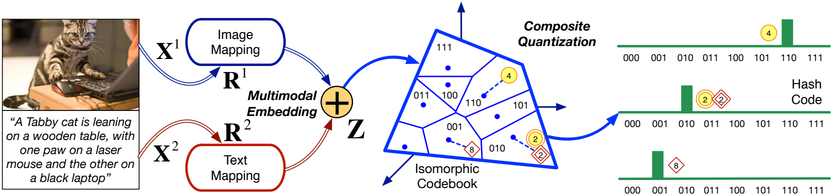
- A Latent Semantic Analysis (LSA) optimization framework
  - $\mathbf{x}_n^v \approx \mathbf{R}^v \mathbf{C}^v \mathbf{b}_n^v$, where $\mathbf{R}^v$ is correlation-maximal mapping, $\mathbf{C}^v$ is similarity-preserving codebook, $\mathbf{b}_n^v$ is compact binary code
  - Multimodal Embedding: Correlation Mapping & Code Fusion
  - Composite Quantization: Isomorphic Space (shared codebook)
- *A "simple and reliable" approach to efficient multimodal retrieval*



Multimodal Correlation

- Paired data matrices: $\mathbf{X}^1 = [\mathbf{x}_1^1, \ldots, \mathbf{x}_N^1], \mathbf{X}^2 = [\mathbf{x}_1^2, \ldots, \mathbf{x}_N^2]$
- Fusion representation matrix: $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]$
- Transformation matrices: $\mathbf{R}^1$, $\mathbf{R}^2$, which transform $\mathbf{X}$ into $\mathbf{Z}$

$$\min_{\mathbf{R}^1, \mathbf{R}^2, \mathbf{Z}} \lambda_1 \left\| {\mathbf{R}^1}^{\mathrm{T}} \mathbf{X}^1 - \mathbf{Z} \right\|_F^2 + \lambda_2 \left\| {\mathbf{R}^2}^{\mathrm{T}} \mathbf{X}^2 - \mathbf{Z} \right\|_F^2 \tag{1}$$



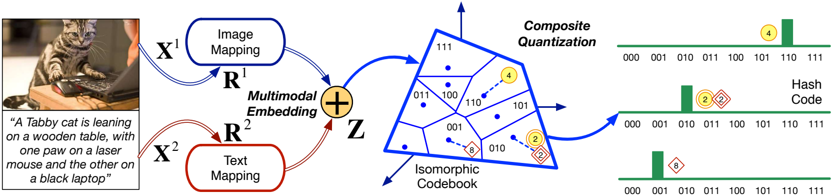
Multimodal Correlation

- This problem is ill-posed, which cannot be solved successfully

$$\min_{\mathbf{R}^1,\mathbf{R}^2,\mathbf{Z}} \lambda_1 \left\| {\mathbf{R}^1}^{\mathsf{T}} \mathbf{X}^1 - \mathbf{Z} \right\|_F^2 + \lambda_2 \left\| {\mathbf{R}^2}^{\mathsf{T}} \mathbf{X}^2 - \mathbf{Z} \right\|_F^2 \tag{2}$$

$$\mathbf{Z} = \frac{\lambda_1 {\mathbf{R}^1}^{\mathsf{T}} \mathbf{X}^1 + \lambda_2 {\mathbf{R}^2}^{\mathsf{T}} \mathbf{X}^2}{\lambda_1 + \lambda_2} \tag{3}$$

$$\mathbf{R}^1 = \left(\mathbf{X}^1 {\mathbf{X}^1}^{\mathsf{T}} \right)^{-1} \mathbf{X}^1 \mathbf{Z}^{\mathsf{T}} \qquad \mathbf{R}^2 = \left(\mathbf{X}^2 {\mathbf{X}^2}^{\mathsf{T}} \right)^{-1} \mathbf{X}^2 \mathbf{Z}^{\mathsf{T}} \tag{4}$$

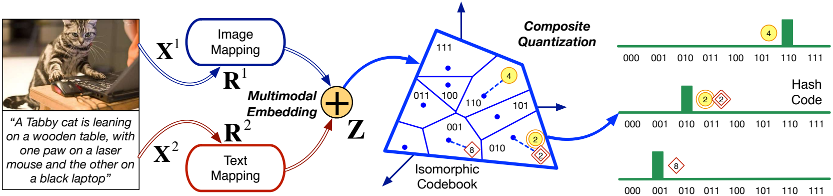


Multimodal Correlation

- Add the covariance maximization with orthogonal constraints

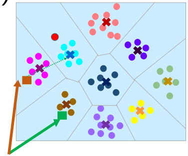
$$\begin{aligned}\min_{\mathbf{R}^1,\mathbf{R}^2,\mathbf{Z}} \lambda_1 & \left(\left\| \mathbf{R}^{1\mathsf{T}} \mathbf{X}^1 - \mathbf{Z} \right\|_F^2 + \left\| \mathbf{R}_\perp^{1\ \mathsf{T}} \mathbf{X}^1 \right\|_F^2 \right) \\ + \lambda_2 & \left(\left\| \mathbf{R}^{2\mathsf{T}} \mathbf{X}^2 - \mathbf{Z} \right\|_F^2 + \left\| \mathbf{R}_\perp^{2\ \mathsf{T}} \mathbf{X}^2 \right\|_F^2 \right)\end{aligned} \tag{5}$$

$$\min_{\mathbf{R}^1,\mathbf{R}^2,\mathbf{Z}} \lambda_1 \left\| \mathbf{X}^1 - \mathbf{R}^1 \mathbf{Z} \right\|_F^2 + \lambda_2 \left\| \mathbf{X}^2 - \mathbf{R}^2 \mathbf{Z} \right\|_F^2 \tag{6}$$



Composite Quantization

- Learn M codebooks: $\mathbf{C} = [\mathbf{C}_1, \ldots, \mathbf{C}_M]$, each codebook has K codewords $\mathbf{C}_m = [\mathbf{c}_{m1}, \ldots, \mathbf{c}_{mK}]$ (cluster centroids of K-means)
- Each $\mathbf{z}_i$ is approximated by the addtion of M codewords
- One per codebook, each selected by the binary assignment $\mathbf{b}_{mi}$
- Code representation: $i_1 i_2 \ldots i_M$, where $i_m = \mathrm{nz}(\mathbf{b}_{mi})$
- Code length: $M \log_2 K$ (1-of-K encoding)



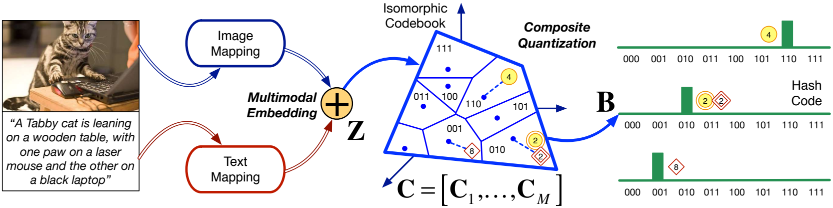
$$\begin{aligned} \mathbf{z} \approx \hat{\mathbf{z}} &= \mathbf{C}_1\mathbf{b}_1 + \mathbf{C}_2\mathbf{b}_2 + \ldots + \mathbf{C}_M\mathbf{b}_M \\ &= \mathbf{c}_{1i_1} + \mathbf{c}_{2i_2} + \ldots + \mathbf{c}_{Mi_M} \end{aligned} \tag{7}$$

$$\mathbf{C}_1 = [\mathbf{c}_{11}, \ldots, \mathbf{c}_{1K}] \quad \mathbf{C}_2 = [\mathbf{c}_{21}, \ldots, \mathbf{c}_{2K}] \quad \ldots \quad \mathbf{C}_M = [\mathbf{c}_{M1}, \ldots, \mathbf{c}_{MK}]$$

Composite Quantization

- Learn M codebooks: $\mathbf{C} = [\mathbf{C}_1, \ldots, \mathbf{C}_M]$, each codebook has K codewords $\mathbf{C}_m = [\mathbf{c}_{m1}, \ldots, \mathbf{c}_{mK}]$ (cluster centroids of K-means)
- Binary code matrices: $\mathbf{B} = [\mathbf{B}_1; \ldots; \mathbf{B}_M]$, $\mathbf{B}_m = [\mathbf{b}_{m1}; \ldots; \mathbf{b}_{mN}]$
- Control binary codes quality by quantization error minimization

$$\min_{\mathbf{Z},\mathbf{C},\mathbf{B}} \left\| \mathbf{Z} - \sum\nolimits_{m=1}^{M} \mathbf{C}_m \mathbf{B}_m \right\|_F^2 = \sum_{i=1}^{N} \left\| \mathbf{z}_i - \sum\nolimits_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mi} \right\|_2^2 \tag{8}$$



Composite Correlation Quantization

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

$$\begin{aligned}
\min_{\mathbf{R}^v,\mathbf{C},\mathbf{B}^v} & \sum_{v=1}^{V}\sum_{n=1}^{N_v}\lambda_v\left\|\mathbf{x}_n^v - \mathbf{R}^v\sum_{m=1}^{M}\mathbf{C}_m\delta\left(\mathbf{b}_{mn}^v\right)\right\|_2^2 \\
\text{s.t.} \quad & \mathbf{R}^{v\mathsf{T}}\mathbf{R}^v = \mathbf{I}_{D\times D}, \mathbf{R}^v \in \mathbb{R}^{P_v\times D} \\
& \left\|\delta\left(\mathbf{b}_{mn}^v\right)\right\|_0 = 1, \delta\left(\mathbf{b}_{mn}^v\right) \in \{0,1\}^K \\
& \delta\left(\mathbf{b}_{mn}^v\right) = \begin{cases}\mathbf{b}_{mn}, & n = 1\ldots N_0 \\ \mathbf{b}_{mn}^v, & \text{otherwise}\end{cases} \\
& v = 1\ldots V, m = 1\ldots M, n = 1\ldots N_v
\end{aligned} \tag{9}$$

Approximate Distance Computation

- Asymmetric Quantizer Distance: $\left\|\mathbf{q}^{\bar{v}} - \mathbf{x}_n^v\right\|_2^2 \approx \mathrm{AQD}\left(\mathbf{q}^{\bar{v}}, \mathbf{x}_n^v\right)$

$$\begin{aligned}\mathrm{AQD}\left(\mathbf{q}^{\bar{v}}, \mathbf{x}_n^v\right) &= \left\|\mathbf{q}^{\bar{v}} - \mathbf{R}^{\bar{v}} \sum_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mn}^v\right\|_2^2 \\ &= -2 \sum_{m=1}^{M} \left\langle \tilde{\mathbf{q}}^{\bar{v}}, \mathbf{C}_m \mathbf{b}_{mn}^v \right\rangle + \left\|\sum_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mn}^v\right\|_2^2 \\ &+ \left\|\tilde{\mathbf{q}}^{\bar{v}}\right\|_2^2 + \left\|\mathbf{R}_{\perp}^{\bar{v}\mathsf{T}} \mathbf{q}^{\bar{v}}\right\|_2^2\end{aligned} \tag{10}$$

- Query-specific distance lookup table: Store the distances from all $M \times K$ codebook elements in $\mathbf{C} = [\mathbf{C}_1, \ldots, \mathbf{C}_M]$ to query $\mathbf{q}^{\bar{v}}$
- $O(M)$ additions for term 1, $O(M^2)$ or $O(1)$ additions for term 2
- Alternative: Cosine Distance $\cos\left(\mathbf{q}^{\bar{v}}, \mathbf{x}_n^v\right) = \sum_{m=1}^{M} \left\langle \tilde{\mathbf{q}}^{\bar{v}}, \mathbf{C}_m \mathbf{b}_{mn}^v \right\rangle$

Approximation Error Analysis

Theorem (Approximation Error Bound)

The error of approximating Euclidean distance with AQD is bounded by

$$\left| d\left(\tilde{\mathbf{q}}^{\bar{v}}, \tilde{\mathbf{x}}_n^v\right) - d\left(\tilde{\mathbf{q}}^{\bar{v}}, \hat{\mathbf{x}}_n^v\right) \right| \leqslant \left\| \mathbf{x}_n^v - \mathbf{R}^v \sum\nolimits_{m=1}^M \mathbf{C}_m \mathbf{b}_{mn}^v \right\|_2. \tag{11}$$

From triangle inequality, $\left| d\left(\tilde{\mathbf{q}}^{\bar{v}}, \tilde{\mathbf{x}}_n^v\right) - d\left(\tilde{\mathbf{q}}^{\bar{v}}, \hat{\mathbf{x}}_n^v\right) \right| \leqslant d\left(\tilde{\mathbf{x}}_n^v, \hat{\mathbf{x}}_n^v\right)$. Then

$$\begin{aligned} d^2\left(\tilde{\mathbf{x}}_n^v, \hat{\mathbf{x}}_n^v\right) &= \left\| \mathbf{R}^{v\mathsf{T}} \mathbf{x}_n^v - \sum\nolimits_{m=1}^M \mathbf{C}_m \mathbf{b}_{mn}^v \right\|_2^2 \\ &\leqslant \left\| \mathbf{R}^{v\mathsf{T}} \mathbf{x}_n^v - \sum\nolimits_{m=1}^M \mathbf{C}_m \mathbf{b}_{mn}^v \right\|_2^2 + \left\| \mathbf{R}_\perp^{v\mathsf{T}} \mathbf{x}_n^v \right\|_2^2 \\ &= \left\| \mathbf{x}_n^v - \mathbf{R}^v \sum\nolimits_{m=1}^M \mathbf{C}_m \mathbf{b}_{mn}^v \right\|_2^2, \end{aligned} \tag{12}$$

Quantize by max cross-modal correlation & within-modal covariance.

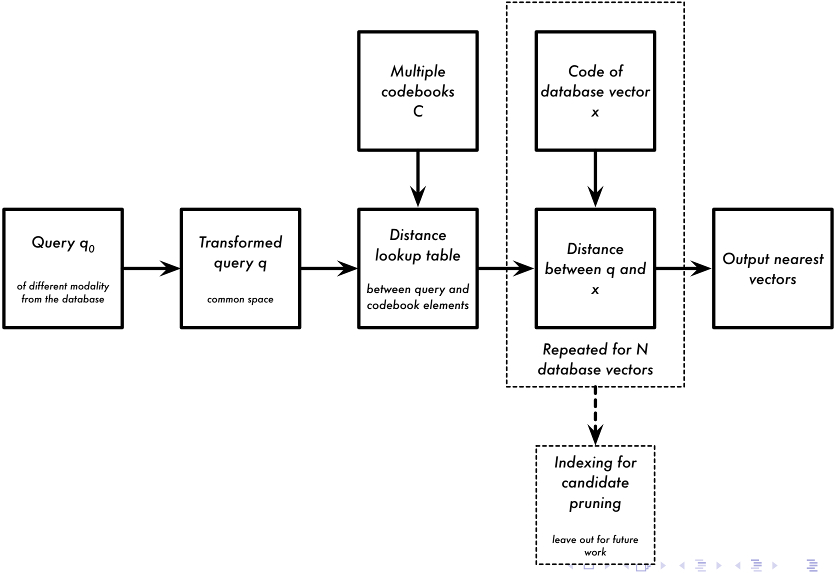
Experiment Setup

- **Datasets:** NUS-WIDE, Wiki, and Flickr1M
- **Tasks:** $I \to I$, $T \to T$, $I \to T$, $T \to I$, $I \to IT$, and $T \to IT$
- **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

Dataset	NUS-WIDE	Wiki	Flickr1M
Complete Set	195,834	2,866	1,000,000
Query Set	2,000	693	1,000
Database	193,834	2,173	24,000
Training Set	10,000	2,173	975,000

Search Pipeline



MAP Results

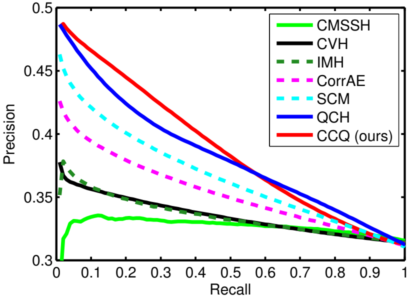
- CCQ significantly outperforms unsupervised hashing methods (CVH, IMH) and deep hashing methods (CorrAE), and generally outperforms supervised hashing methods (CMSSH, SCM, QCH).

Table: MAP Comparison of Multimodal Retrieval on Standard Datasets

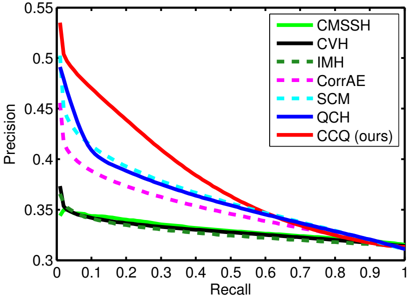
Task	Method	NUS-WIDE	Wiki	Flickr1M
	CorrAE (deep)	0.4699	0.2033	0.6357
$I \to T$	QCH (supervised)	0.5050	0.2368	0.6685
	CCQ (ours)	**0.5165**	**0.2371**	**0.7183**
$I \to IT$	CCQ (ours)	0.5414	0.2529	0.6989
	CorrAE (deep)	0.4634	0.3478	0.6247
$T \to I$	QCH (supervised)	0.5389	**0.4411**	0.6485
	CCQ (ours)	**0.5413**	0.4222	**0.7165**
$I \to IT$	CCQ (ours)	0.7131	0.6394	0.7190

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.



(a) $I \rightarrow T$ @ 32 bits



(b) $T \rightarrow I$ @ 32 bits

Figure: Precision-recall curves on NUS-WIDE cross-modal tasks @ 32 bits.

Wiki

- The low quality of the image modality leads to low cross-modal retrieval performance, which fits supervised hashing methods.

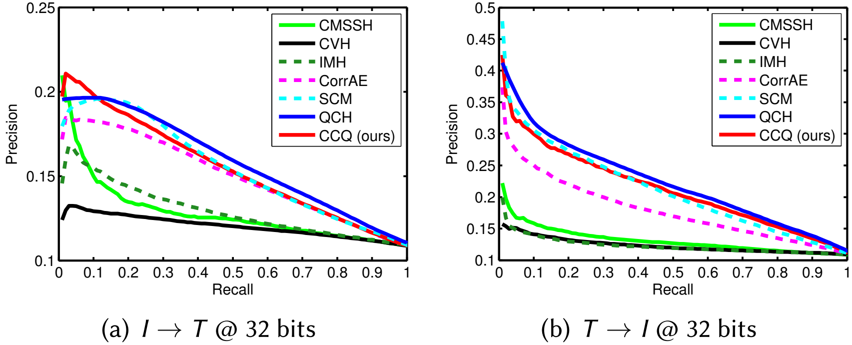
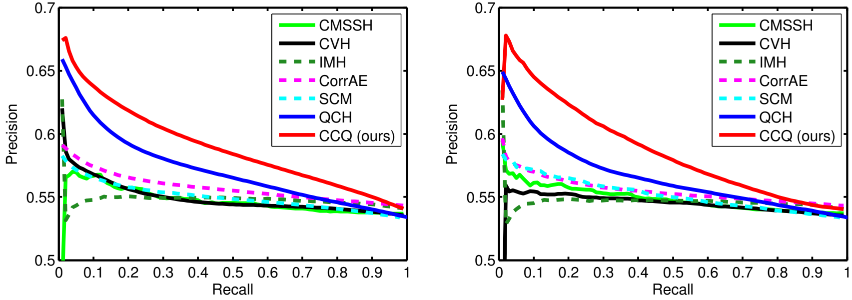


Figure: Precision-recall curves on Wiki cross-modal tasks @ 32 bits.

Flickr1M

- In the presence of big data, there is strong motivation to learn accurate models from large-scale dataset (big model capacity).



(a) $I \to T$ @ 32 bits (b) $T \to I$ @ 32 bits

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.

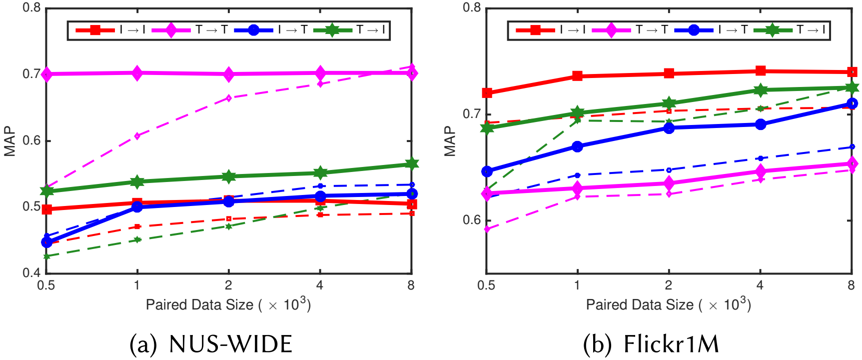
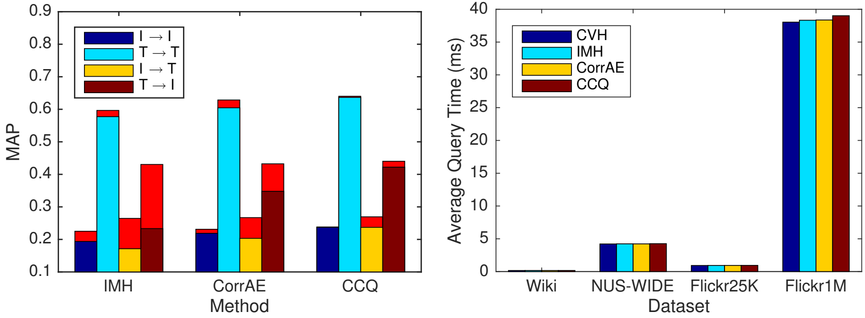


Figure: MAP of CCQ by varying the numbers of paired points for training.

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.

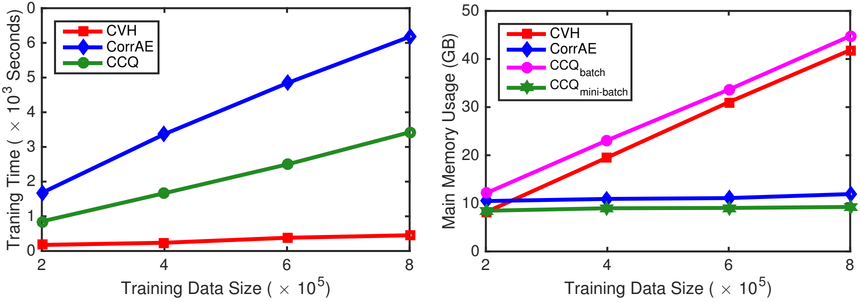


(a) MAP Loss (b) Search Efficiency

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.

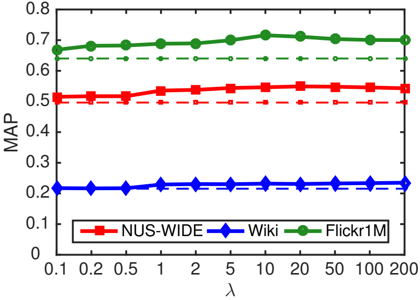


(a) Time (b) Memory

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.



(a) $I \rightarrow T$ @ 32 bits (b) $T \rightarrow I$ @ 32 bits

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 indexding CCQ codes
  - Deep neural networks for multimodal correlation embedding
- `http://ise.thss.tsinghua.edu.cn/~mlong`