MSSPQ: Multiple Semantic Structure-Preserving Quantization for Cross-Modal Retrieval

Lei Zhu, Liewu Cai, Jiayu Song
Xinghui Zhu, Chengyuan Zhang*, Shichao Zhang*

College of Information and Intelligence, Hunan Agricultural University
Changsha, Hunan, P.R. China
School of Computer Science and Engineering, Central South University
Changsha, Hunan, P.R. China

Content

- Motivation
- Related Work
- Contribution
- Problem Definition
- The Method
- Experiment
- Conclusion

Motivation

we investigate how to capture multiple semantic correlation to boost cross-modal hashing learning.

Related Work

According to the representation type of multimedia instances, cross-modal retrieval can be divided into two groups: **real-valued representation based retrieval** and **binary representation (hash code) based retrieval**.

- **real-valued representation based retrieval**
  - **CCA, LDA**
- **binary representation based retrieval**
  - **MDCH**

Contributions

I. A very efficient end-to-end cross-modal hashing framework, named **MSSPQ**

II. A novel multiple deep semantic correlation learning method, which contains inter-modal pairwise correlation learning, intra-modal pairwise correlation learning, Cosine correlation learning and hashing learning.

III. We conduct extensive experiments on three commonly used multimedia datasets to comprehensively evaluate the performance of our method.

Problem Definition

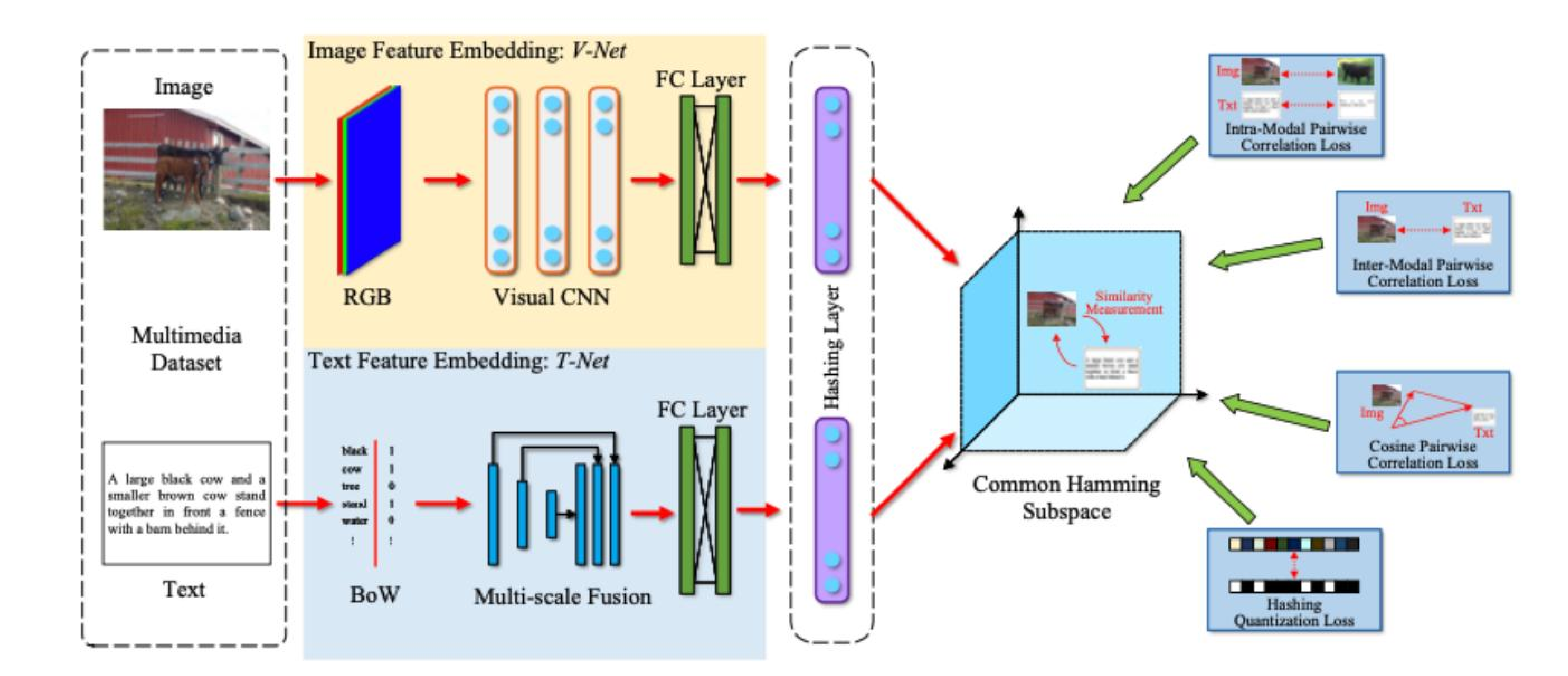
- Let $O = \{< I_i, T_i, L_i >\}_{i=1}^{n}$ be a multimedia dataset containing *n* pairs of an image and a text with a category label, where $< I_i, T_i, L_i >$ denotes a triplet of i-th pair of image I_i and text T_i with their corresponding category label L_i . For $Txt2Img$ retrieval:

$$R = \{I | I \in O, I^{'} \in O \setminus R, D_H(\Phi_T(T_q; \theta_T), \Phi_I(I; \theta_I)) > D_H(\Phi_T(T_q; \theta_T), \Phi_I(I^{'}; \theta_I))\}$$

where R is the result set, θ_T and θ_I are the parameter vectors. $D_H(\cdot\,,\cdot)$ is the Hamming distance function which is to measure the semantic similarity between two different modal objects in Hamming subspace. To realize the projections $Txt2Img$ and $Img2Txt$, we propose a very efficient end-to-end cross-modal hashing framework, named MSSPQ, which aims to generate high-quality cross-modal hash codes by enhancing semantic similarity preserving.

The Method

- **The Framework**

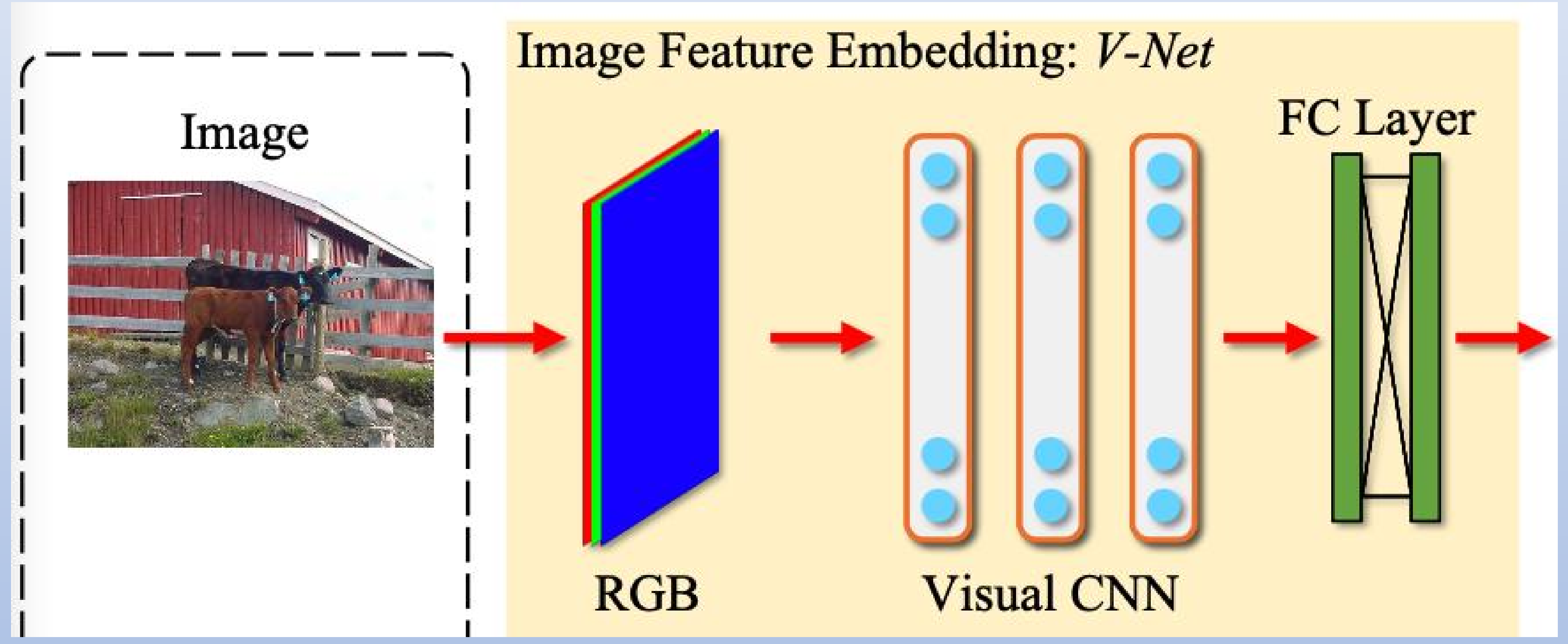


This model consists of three components: (1) cross-modal embedding; (2) hashing learning module; (3) multiple correlation learning module.

The Method

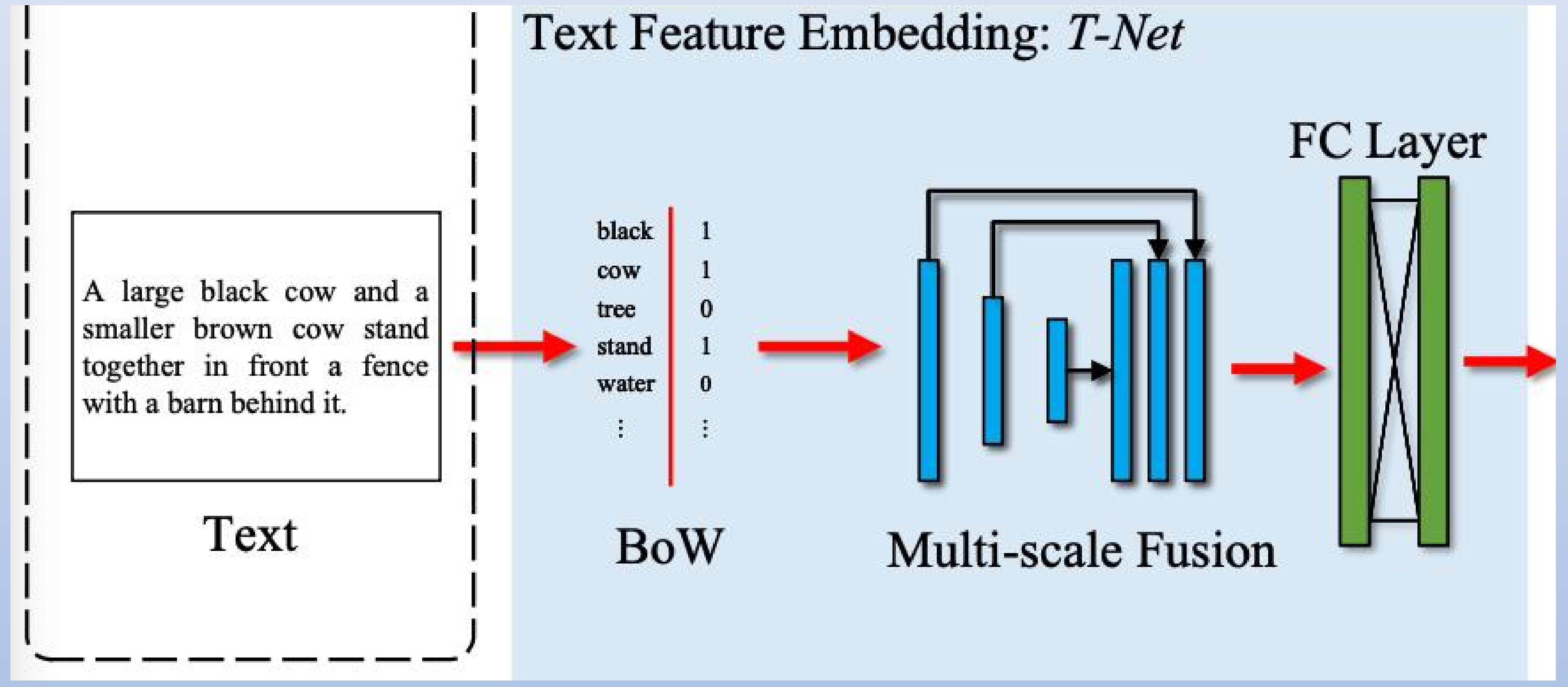
- **Cross-Modal Embedding**

Image Feature Embedding



For image modality, ResNet34 is used to extract visual features.

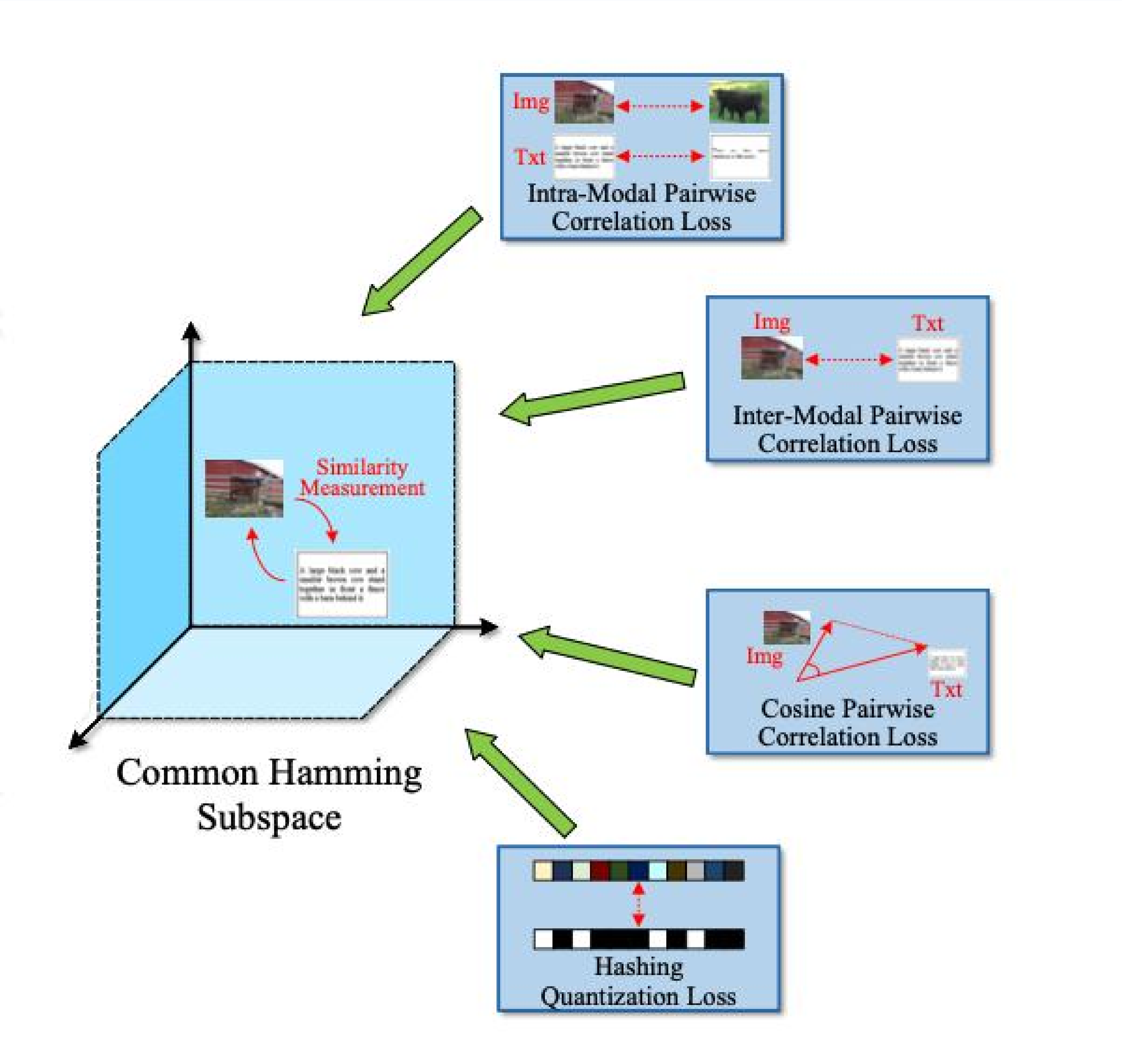
Text Feature Embedding



For text modality, each sample is firstly encoded by BoW model, and then fed into a multi-scale fusion model.

The Method

- **Multiple Correlation Learning**



The Method

To capture comprehensive semantic correlation, multiple correlation loss, including inter-modal pairwise correlation loss, intra-modal pairwise correlation loss, as well as Cosine correlation loss are involved.

Inter-Modal Pairwise Correlation Loss: the pairwise correlation is implemented by negative log likelihood function of similarity probability, which is formulated as:

$$L_1 = -\sum_{i,j=1}^{n} \left(S_{ij}^{IT} \cdot \phi_{ij}^{IT} - \log\left(1 + e^{\phi_{ij}^{IT}}\right) \right)$$

where $S_{ij}^{IT} \in \{0,1\}$ denotes the similarity between image I_i and text T_j . $\phi_{ij}^{IT} = \frac{1}{2} F_{i*}^{T} G_{j*}$ is the inner product of representations generated by V-Net and T-Net.

The Method

To capture comprehensive semantic correlation, multiple correlation loss, including inter-modal pairwise correlation loss, intra-modal pairwise correlation loss, as well as Cosine correlation loss are involved.

Intra-Modal Pairwise Correlation Loss: we utilize intra-modal pairwise correlation learning to training the cross-modal embedding model:

$$L_2^I = -\sum_{i,j=1}^{n} (S_{ij}^{II} \cdot \phi_{ij}^{II} - \log(1 + e^{\phi_{ij}^{II}}))$$

$$L_2^T = -\sum_{i,j=1}^{n} (S_{ij}^{TT} \cdot \phi_{ij}^{TT} - \log(1 + e^{\phi_{ij}^{TT}}))$$

where L^I denotes the intra-modal pairwise correlation loss for image modality, L^T denotes the intra-modal pairwise correlation loss for text modality.

The Method

To capture comprehensive semantic correlation, multiple correlation loss, including inter-modal pairwise correlation loss, intra-modal pairwise correlation loss, as well as Cosine correlation loss are involved.

Cosine Correlation Loss: the cosine similarity between sample labels and the cosine similarity between sample features are defined as:

$$S_{ij}^{c} = \frac{l_{i*}}{||l_{i*}||_F^2} \cdot \frac{l_{j*}^T}{||l_{j*}||_F^2} \qquad Cos(F_{i*}, G_{j*}) = \frac{F_{i*}}{||F_{i*}||_F^2} \cdot \frac{G_{j*}^T}{||G_{j*}||_F^2}$$

Thus, the inter-modal label pairwise Cosine similarity loss can be defined as:

$$L_3 = \sum_{i,j=1}^{n} (S_{ij}^{c} - Cos(F_{i*}, G_{j*}))^2$$

The Method

To capture comprehensive semantic correlation, multiple correlation loss, including inter-modal pairwise correlation loss, intra-modal pairwise correlation loss, as well as Cosine correlation loss are involved.

Cosine Correlation Loss: To preserve more semantic Cosine similarity between samples within modality, the intra-modal label pairwise Cosine similarity loss can be defined as $L_4 = L_4^I + L_4^T$:

$$L_4^I = \sum_{i,j=1}^{n} (S_{ij}^c - Cos(F_{i*}, F_{j*}))^2$$

$$L_4^T = \sum_{i,j=1}^{n} (S_{ij}^c - Cos(G_{i*}, F_{j*}))^2$$

Minimizing L_3 and L_4 can learn more multiple label Cosine similarity so as to preserve much more semantic structure information.

The Method

- **Hashing Learning**

For cross-modal representations F and G, we utilize a sign function $sign(\cdot)$ to generate binary hash codes, namely, $H^I = sign(F)$ and $H^T = sign(G)$. Using the same hash code for different modalities can improve the training effect, in this article we set $H^I = H^T = H$. To make the approximate hash code output by the V-Net and T-Net similar to the binary representation, we define quantization loss L_5 as follows:

$$L_5 = ||H - F||_F^2 + ||H - G||_F^2$$

Overall, the total objective function is:

$$\arg\min_{H,\theta_I,\theta_T} L_{total} = L_1 + L_2 + \lambda(L_3 + L_4) + L_5$$

$$s.t. H \in \{-1, +1\}^{n \times k}$$

Experiments

- **Datasets**
  I. NUS-WIDE
  II. MS COCO
  III. MIRFLICKR-25K

- **Baselines**
  SCM, SEPH, PRDH, CMHH, CHN, DCMH, MLCAH

Experiments

- The results (mAP) on NUS-WIDE

Methods	NUS-WIDE					
	Img2Txt			Txt2Img		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
SCM [43]	0.4626	0.4792	0.4886	0.4261	0.4372	0.4478
SEPH [17]	0.4796	0.4858	0.4906	0.6078	0.6022	0.6288
PRDH [37]	0.5918	0.6058	0.6116	0.6155	0.6284	0.6342
CMHH [2]	0.5531	0.5698	0.5920	0.5738	0.5782	0.5882
CHN [3]	0.5754	0.5966	0.6015	0.5816	0.5967	0.5992
DCMH [14]	0.5445	0.5597	0.5802	0.5793	0.5922	0.6014
MLCAH [18]	**0.6440**	0.6410	0.6430	**0.6620**	**0.6730**	0.6870
MSSPQ	0.6346	**0.6478**	**0.6615**	0.6312	0.6631	**0.6882**

Experiments

- The results (mAP) on MS COCO

Methods	MS COCO					
	Img2Txt			Txt2Img		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
SCM [43]	0.3601	0.3574	0.3562	0.4118	0.4183	0.4345
SEPH [17]	0.4295	0.4353	0.4726	0.4348	0.4606	0.5195
PRDH [37]	0.5538	0.5672	0.5572	0.5122	0.5190	0.5404
CMHH [2]	0.5463	0.5675	0.5674	0.4884	0.4554	0.4846
CHN [3]	**0.5763**	0.5822	0.5805	0.5198	0.5320	0.5409
DCMH [14]	0.5229	0.5438	0.5419	0.4883	0.4942	0.5145
MLCAH [18]	0.5700	0.5620	0.5620	0.5440	0.5470	0.5940
MSSPQ	0.5710	**0.5881**	**0.5862**	**0.5472**	**0.5630**	**0.5985**

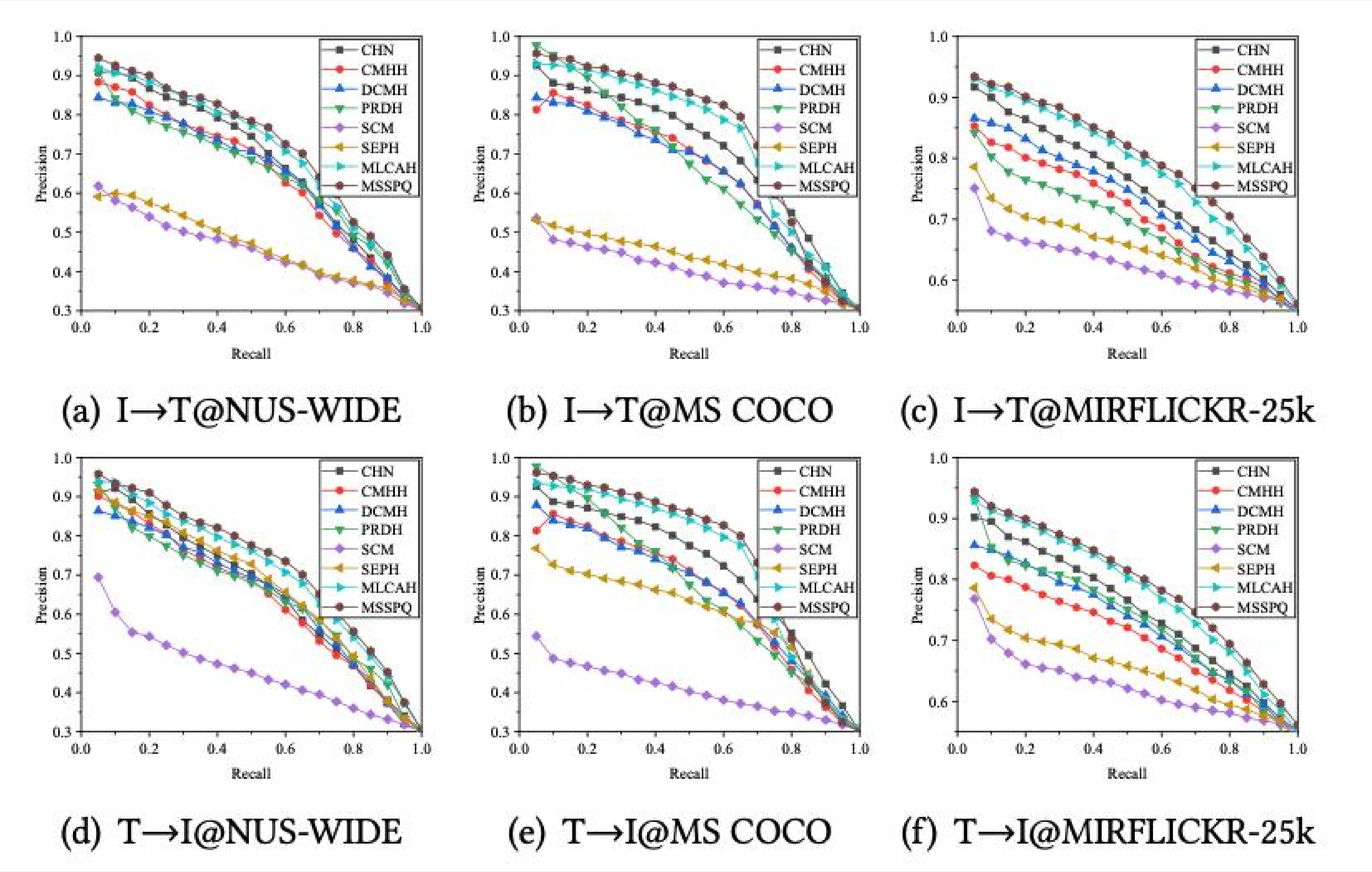
Experiments

- The results (mAP) on MIRFLICKR-25k

Methods	MIRFLICKR-25k					
	Img2Txt			Txt2Img		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
SCM [43]	0.6354	0.6407	0.6556	0.6340	0.6458	0.6541
SEPH [17]	0.6740	0.6813	0.6830	0.7139	0.7252	0.7294
PRDH [37]	0.6952	0.7072	0.7108	0.7626	0.7718	0.7755
CMHH [2]	0.7334	0.7280	0.7441	0.7320	0.7182	0.7276
CHN [3]	0.7504	0.7495	0.7461	0.7776	0.7772	0.7798
DCMH [14]	0.7406	0.7415	0.7434	0.7617	0.7716	0.7748
MLCAH [18]	**0.7960**	**0.8080**	0.8150	0.7940	**0.8050**	0.8010
MSSPQ	0.7868	0.8011	**0.8172**	**0.7946**	0.7885	**0.8022**

Experiments

- The PR-Curves on NUS-WIDE, MS COCO and MIRFLICKR-25k



(a) I→T@NUS-WIDE (b) I→T@MS COCO (c) I→T@MIRFLICKR-25k

(d) T→I@NUS-WIDE (e) T→I@MS COCO (f) T→I@MIRFLICKR-25k

Conclusion

- This article proposes an efficient end-to-end cross-modal hashing learning method, termed as Multiple Semantic Structure-Preserving Quantization (MSSPQ).
- Our method considers multiple semantic correlation learning across different modalities for realizing semantic similarity structure-preserving.
- Extensive experiments are conducted on three commonly used multimedia dataset show that the proposed MSSPQ achieves state-of-the-art performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62072166, 61836016, 61672177).

Thank You !