Deep Similarity Preserving and Attention-based Hashing for Cross-Modal Retrieval

Shubai Chen[1], Song Wu[1,*], Yu Chen[2]

[1] College of Computer and Information Science, Southwest University, Chongqing 400715, China

[2] College of Engineering and Technology, Southwest University, Chongqing 400715, China

E-mail: chansuba,cy1034429543@email.swu.edu.cn, songwuswu@swu.edu.cn

Abstract

With the fast progress of deep neural networks and the quick search efficiency of hashing, deep cross-modal hashing (CMH) methods have obtained more and more attention. Generally speaking, the existing CMH methods simultaneously learn hash functions and hash codes in an end-to-end architecture. However, they mostly focus on the hash codes generation stage neglected the losing of rich semantic information in the hash representations learning stage. Besides, the multi-label constraint is ignored, and the single-label criterion is leveraged. Thus, we propose a novel Deep Semantic Preserving Attention-based Hashing (DSPAH) for cross-modal retrieval. In the DSPAH, we first use a cross-level attention block to emphasize meaningful parts of hash representations and oversee unnecessary ones. Moreover, a Fine-Grained Similarity Criterion (FGSC) is proposed to explore the multiple semantic of image or text instances, helping to learn robust and optimal hash codes. Extensive experiment results on two large-scale public datasets have shown the competition of our proposed DSPAH.

I. Introduction

Due to the rapid development of search engines and social networks, exponential growth can be seen in multimedia data such as images, text, audio, and video. Thus how to efficiently and effectively retrieve information across these modalities has become a hot spot called multi-modal retrieval. To be specific, one may want to obtain all semantically related instances from the datasets given in a paragraph. However, due to the discrepancies in distribution and inconsistent representations among different modalities, this has raised a significant challenge to unify the gap effectively and efficiently.

Especially, cross-modal retrieval is the most pervasive method of multi-modal retrieval, which aims to map original data (images or text) into similarity, preserving embedding in a common latent space[1]. In this way, instances that share similar semantic information may have shorter distances, dissimilar otherwise. The cross-modal retrieval methods can be grossly split into two classes. Traditionally, real-value latent representations is adopted such as [2], [3], [4], [5]. However, the real value methods may cause high computational costs and heavy storage burdens. Thus, another popular method called cross-modal hashing (CMH) is proposed to save storage and accelerate the retrieval speed, which leverages Manifold Learning to generate compact hash codes from original high-dimension data.

As the Superior performance of deep learning, Deep Neural Networks (DNN) has shown robust capability in various applications such as [6], [7], [8]. Thus, the recent cross nodal hashing methods are all based on DNN and achieve appealing results. To take advantages of DNN, many cross-modal hashing methods are proposed including deep cross-modal hashing (DCMH) [9], self-supervised adversarial hashing (SSAH) [10], self-constraint and attention-based hashing network (SCAHN) [11], triplet-based deep hashing (TDH) [12] and multi-label semantics preserving hashing (MLSPH) [13]. However, there are still some issues that need to be solved in the deep CMH community. Firstly, the existing deep CMH methods use a 'hard' metric policy to measure the similarity between instances, judged by if two instances share at least one label. However, the simple approximation has neglected the fact that most instances in large-scale cross-modal datasets have multiple labels. Secondly, the hash representations generation and hash codes projection is the equally important part of cross-modal hashing methods. Furthermore, traditionally, deep CMH methods concentrate more on the hash codes generation stage. However, hash representation with less semantic information and spatial relevance may fail to generate optimal hash codes.

A superior Deep Similarity Preserving and Attention-based Hashing (DSPAH) is proposed to solve these problems mentioned above. The framework of DSPAH is illustrated in which corporately learns hash representations and binary codes in an end-to-end architecture. The DSPAH consists of two main components in the hash representations generation stage. CNN model is leveraged to learn rich semantic information from image-modality and text-modality. Moreover, the CNN model is followed by a cross-level attention level where multi-level hash representations are concatenated together as the input. Thus the context relationship and informative information can be obtained by the final hash representations. Moreover, to take advantage of multi-label information. A novel dice formula is proposed to build similarity matrixs, which can better explore the fine-grained relationship among multiple labels.

The core contributions of DSPAH are listed as follows:

- Firstly, an cross-level attention block is proposed to explore intensive semantic information. In this module, hash representations generated from multi-level are concatenated and further integrated by the bi-attention module, which explores the context correlation and global dependence from both channel and spatial view.
- Secondly, a dice formula is proposed to effectively obtain the multi-label information constraint, further generating robust hash codes.
- Finally, the DSPAH is applied on two large-scale cross-modal datasets, and the experimental results illustrate the superiority of our proposed DSPAH compared with other state-of-the-art methods.

II. Proposed Method

A.Problem Defination

We use G^T denotes the transpose of G and $\|\cdot\|_F$ denotes the Frobenius norm. The $\operatorname{sign}(\cdot)$ is an element-wise sign function defined as follows:

$$\operatorname{sign}(x)=\begin{cases}1 & x\geq 0\\ -1 & x<0\end{cases} \tag{1}$$

We use $O=\{o_i\}_{i=1}^N$ to denote the training-set with N instances which are image-text data labeled at least one tag. The proposed DSPAH can be expanded to all kinds of modality (e.g. image, text, audio and video) and we mainly concentrate on image-modality and text-modality in this paper. Thus we use $o_i=(v_i,t_i,l_i)$ to denote the ith training instance, $v_i\in R^{d_v}$, $t_i\in R^{d_t}$ and $l_i\in R^{d_l}$ are image, text and label feature vector with dimension d_v, d_t and d_l. Moreover, the fine-grained similarity matrix is defined as $S=\{S^{vt},S^{vv},S^{tt}\}$, where $S^{vv}=\left\{S_{ij}^{vv}\mid i,j=1,2,\ldots,N\right\}\in R^{N\times N}$ and $S^{tt}=\left\{S_{ij}^{tt}\mid i,j=1,2,\ldots,N\right\}\in R^{N\times N}$ denotes the intra-modality similarity matrix of image and text, $S^{vt}=\left\{S_{ij}^{vt}\mid i,j=1,2,\ldots,N\right\}\in R^{N\times N}$ denotes the inter-modality similarity matrix between image and text.

The most important task of our proposed DSPAH is learning two discriminative hash functions $h^{(v)}(\mathbf{v})$ and $h^{(t)}(\mathbf{t})$ for image-modality and text-modality using the training-set O and similarity matrix S. In the hash representations generation stage, hash representations learned from image-modality and text-modality are represented by $F=\{f_{v_i}\mid i=1,2,\cdots,N\}\in R^{N\times c}$ and $G=\{g_{t_i}\mid i=1,2,\cdots,N\}\in R^{N\times c}$. In hash codes projection stage, $B=\{B_i\mid i=1,2,\cdots,N\}\in R^{N\times c}$ denotes the final hash codes from F and G by simply using a sign function $B=\operatorname{sign}(F+G)$.

B.Network Architecture of DSPAH

The overview architecture of DSPAH is illustrated in Fig. 1, which consists of the multi-level hash representations generation and attention-based interaction module.

Speaking of multiple-level hash representations generation, both the image-network and text-network use Resnet as the bone-network because of its remarkable performance on computer vision applications. Especially, the original text data is represented as Bag-of-Words (BoW) vectors and fused into multi-scale BoW representations. To be specific, a multi-scale pooling policy is conducted on the BoW vectors to explore global features, and these vectors are resized into the same length. Furthermore, to facilitate the Resnet[14], these vectors are stacked together to make up a matrix. Therefore, the rich semantics context in text-modality is further explored. For both image-modality and text-modality, we propose cross-level attention to capture the context relationship and global dependency. To be specific, the hash representations from intermediate layers are generated by global average pooling (GAP) and convolution layer with a kernel size of 1×1. The novel CBAM [15] is leveraged to capture the context relationship and global dependency in intermediate layers. Finally, all of these hash representations are weighted together as the final hash representations by multiplying the adaptive attention matrix. Therefore, the final hash representations can fully obtain the semantic information.

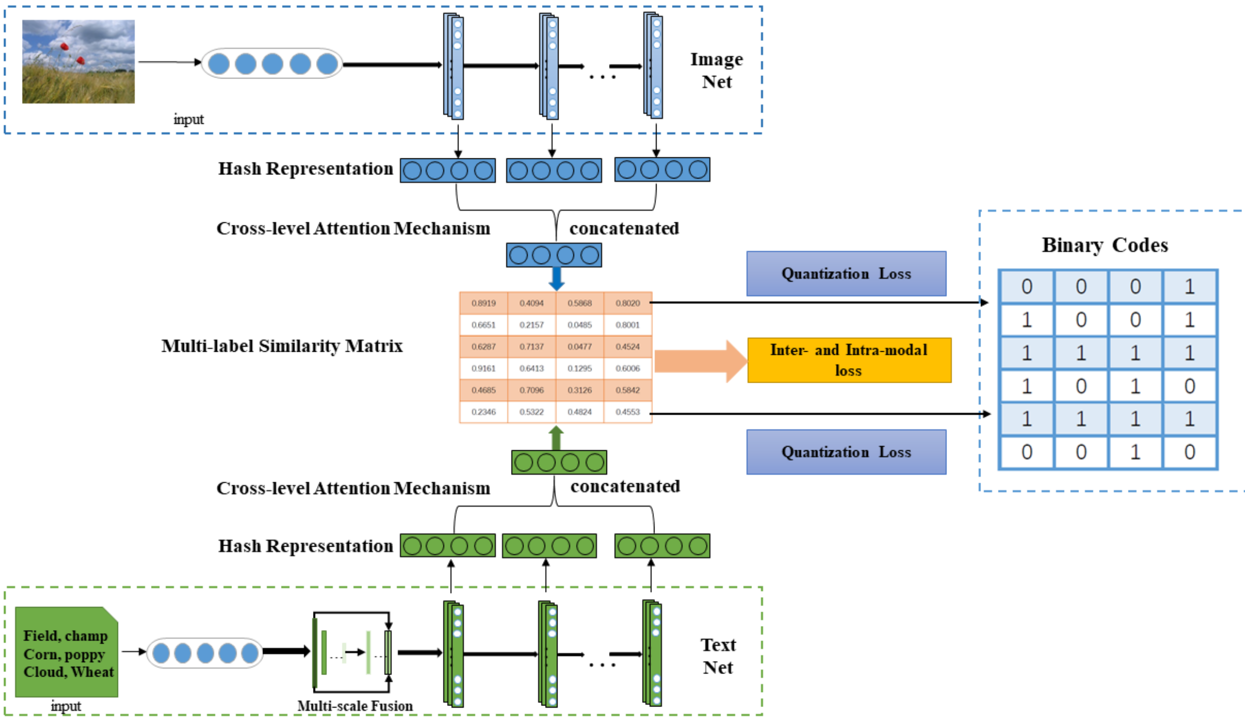

Fig. 1. The overview architecture of our proposed DSPAH consists of two parts: (1) multi-level hash representations generation: the networks are divided into several blocks which are weighted by CBAM attention, and then the multi-level hash representations are multiplied by an adaptive attention matrix. Finally, these multiple layers are concatenated together as the final hash representations. (2) multi-label similarity preserving loss.

C.Hash Function Learning

In large-scale cross-modal datasets, multi-labels for a single instance(e.g., image and text) are quite common. However, most previous cross-modal retrieval methods measure the similarity by only one shared label, neglecting the fine-grained similarity among instances. Thus, we propose a new similarity measurement policy called Fine-Grained Similarity Criterion (FGSC) to explore the semantic relationship among instances better. The FGSC of inter-modality can be defined as follows:

$$S_{ij}^{vt} = \frac{l_i^v \cap l_j^t}{\sqrt{l_i^v \times l_j^t}} \tag{2}$$

where l_i^v denotes the label vector of ith image instance and l_j^t denotes the label vector of jth text instance. $l_i^v \cap l_j^t$ denotes the number of shared labels of vectors ith and text. $\sqrt{l_i^v \times l_j^t}$ is the geometric mean of these two label vectors. Similarly, the FGSCs of intra-modality instances are defined as follows:

$$S_{ij}^{vv} = \frac{l_i^v \cap l_j^v}{\sqrt{l_i^v \times l_j^v}} \tag{3}$$

$$S_{ij}^{tt} = \frac{l_i^t \cap l_j^t}{\sqrt{l_i^t \times l_j^t}} \tag{4}$$

where S_{ij}^{vv} denotes the similarity across image-modality and S_{ij}^{tt} denotes the similarity across text-modality. Besides, $S = \{S^{vt}, S^{vv}, S^{tt}\} \in (0, 1)$. Thus, the hamming-based loss function is no longer suitable for the continuous similarity value. In this paper, the Mean Square Error (MSE) based loss function is adopted to fit the FGSC. Following the common protocol proposed in DCMH, the inner product $< *, * >, * \in (f, g)$ are leveraged to measure the semantic similarity of hash representations. Therefore, the MSE loss can be defined as follows:

$$\begin{aligned}
\mathcal{L}_{\text{inter}} &= \sum_{i=1,j=1}^{n} \left\| \frac{\langle f_i, g_j \rangle + c}{2} - s_{ij}^{vt} \cdot c \right\|^2 \\
\mathcal{L}_{\text{intra-image}} &= \sum_{i=1,j=1}^{n} \left\| \frac{\langle f_i, f_j \rangle + c}{2} - s_{ij}^{vv} \cdot c \right\|^2 \\
\mathcal{L}_{\text{intra-text}} &= \sum_{i=1,j=1}^{n} \left\| \frac{\langle g_i, g_j \rangle + c}{2} - s_{ij}^{tt} \cdot c \right\|^2
\end{aligned} \tag{5}$$

where f_i and g_j are used to denote the hash representations of the ith image instance and jth text instance. c is the length of hash codes. Since the inner product $\langle *, * \rangle \in [-c, c]$, the value range of $\frac{\langle *,* \rangle + c}{2}$ will be the same as $s_{ijr}^{**} \cdot c$.

The purpose of FGSC-based MSE loss is to generate modal-specific and discriminative hash representations G and F. However, there is a gap between the hash codes and hash representations. Moreover, during the learning procedure of FGSCC-based MSE loss, the similarity between $B^{(v)} = sign(F)$ and $B^{(t)} = sign(g)$ has been ignored. Since the aim of CMH methods is to learn high-quality hash functions and hash codes, we also need to keep the semantic similarity of $B_{(v)}$ and $B_{(t)}$. Another constraint $B^{(v)} = B^{(t)} = B$ is added to keep the modal invariance. Accordingly, the quantization loss is defined as follows:

$$\mathcal{L}_q = \frac{1}{c} \left(\|B - F\|_F^2 + \|B - G\|_F^2 \right) \tag{6}$$

III. Optimization

By assembling the above loss functions, the final overall loss function is given as follows:

$$\min_{B,\theta_x,\theta_y} \mathcal{L} = \mathcal{L}_{inter} + \mathcal{L}_{intra-image} + \mathcal{L}_{intra-text} + \mathcal{L}_q \tag{7}$$

where θ_x, θ_y denote the network parameters of the image-modality and text-modality. An alternating optimization strategy is employed to optimize equation 7. Some parameters will be optimized while others are fixed. The whole optimization algorithm for DSPAH is outlined in Algorithm 1.

IV. Experiment and Discussion

This section evaluates the proposed DSPAH on two large-scale public datasets, MIRFlickr-25K [16], and NUS-WIDE [17] compared with other state-of-the-art methods.

A. Datasets

MIRFLICKR-25K [16] is a standard benchmark which contains 25,000 image-text pairs collected from Flickr website of different group. Each image is related to several textual descriptions. 20,015 instances of image-text pair with at least one of twenty-four labels are selected, which is similar to DCMH[9]. The text-modality instances are transferred into 1,386-dimensional BoW vectors.

NUS-WIDE [17] The NUS-WIDE includes 268,468 image-text pairs which all belong to 81 categories. A 1,000-dimensional BoW vector is generated for each text-modality instance. In this paper, 190,421 image-text pairs with 21 most common labels have remained, and all instances without supervised information are removed.

We use 10,000 and 10,500 image-text pairs in MIRFLICKR-25K and NUS-WIDE for training. Besides, we stochastically choose 2,000 and 2,100 instances for the

---

Algorithm 1: Optimization algorithm of DSPAH.

Input: Training set $\{v_i, t_i, l_i\}_{i=1}^{N}$, intra-modality and inter-modality similarity matrix S^{vv}, S^{tt}, s^{vt};

Output: Optimized parameters θ_x and θ_y of neural networks and binary codes B;

Initialization: Initialize the parameters of neural networks, the batch size is set to $n_x = n_y = 128$, initialize hash representations of each modality: F and G, set iteration number $iter$ and other hyper-parameters.

for *t=1 to iter* **do**

Update the parameter θ_x of image-network by BP algorithm:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial f_{ik}} = &\sum_{j \in N} \left(f_i^T f_j + q - 2 \cdot s_{ij}^{vv} \cdot q \right) \cdot f_{jk} \\
&+ \sum_{j \in N} \left(f_i^T g_j + q - 2 \cdot s_{ij}^{vt} \cdot q \right) \cdot f_{jk} \\
&+ \frac{2}{c}(F - B)
\end{aligned}$$

Update the parameter θ_y of text-network by BP algorithm:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial g_{ik}} = &\sum_{j \in N} \left(g_i^T g_j + q - 2 \cdot s_{ij}^{tt} \cdot q \right) \cdot g_{jk} \\
&+ \sum_{j \in N} \left(f_i^T g_j + q - 2 \cdot s_{ij}^{vt} \cdot q \right) \cdot g_{jk} \\
&+ \frac{2}{c}(G - B)
\end{aligned}$$

end

Update binary codes B

$$B = \text{sign}(\beta(F + G))$$

Until a fixed number of iterations or convergence;

query items, and the remained are treated as the retrieval items.

B.Implementation Details

The DSPAH is conducted on a server with one Nvidia Xp GPU, and the code is written by Pytorch[18] framework. The Resnet-34 with four blocks is utilized to learn rich hash representations. For the image network, the parameters are initialized by the pre-trained model on ImageNet[19]. In terms of the text network, the Normal distribution with $N\left(\mu,\sigma^2\right)$ with $\mu = 0$ and $\sigma = 0.1$ is leveraged to initialize the parameters. Moreover, pooling sizes from 1 to 50 of BoW vectors are implemented to construct the multi-scale text matrix. We use the SGD as the optimization, and the learning rate is set to $10^{-1.5}$ with a mini-batch size of 128.

C.Evaluation and Baselines

To compare the DSPAH with other state-of-the-art methods, we adopt the Mean Average Precision (MAP) and PR Curves to measure the hamming ranking and hash lookup. Several baseline methods are compared with DSPAH including CMSSH [20], SCM [21], GSPH [22], DCMH [9], CMHH [23], PRDH [24], CHN [25], SepH [26] and SSAH [10]. The MAP results is illustrated in Table I and the PR Curves is demonstrated in . From the results, we can get the following observation.

- The DSPAH significantly outperforms other state-of-the-art methods on 16, 32, 64 bits of hash codes in terms of MAP and PR Curves, which clearly shows its superiority. The advance of DSPAH is partly because the cross-level attention dramatically improves the hash representations of interest to concentrate on the vital part and ignore the unconsidered ones.
- The SSAH and DSPAH surpass other deep architecture-based CMH methods and show competitive results, which indicates the importance of preserving multiple semantic labels. The FGSC we proposed in this paper may have the ability to unify the inter-and intra-modality heterogeneity.
- Deep CMH methods such as DCMH, CMHH, SSAH, CHN, and PRDH distinctly attain better performance than other shadow-based CMH methods, including CMSSH, GSPH, SCM, and SePH. This reveals the robust and advanced character of deep neural networks, obtaining richer semantic information than the hand-crafted features. Therefore, better results can be observed.

V. Conclusion

In this paper, cross-level attention and a Fine-Grained Similarity Criterion (FGSC) are proposed, with the vision of learning context-relevant hash representations and generating optimal hash codes. Besides, the leveraged attention mechanism can better enhance the ability to focus on the image's and text's 'right' area. Evaluations conducted on two datasets demonstrate the significant performance of DSPAH compared with other CMH methods. In the future, we are going to use different metrics to investigate the similarity of embeddings.

Acknowledgment

The authors appreciate helpful comments from the reviewers on improving our work. This work was supported by the NSFC(61806168), Fundamental Research Funds for the Central Universities (SWU117059), and Venture & Innovation Support Program for Chongqing Overseas Returnees (CX2018075).

References

[1] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: A review," *Computer Science Review*, vol. 39, p. 100336, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013720304366

[2] Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang, "Multi-modal mutual topic reinforce modeling for cross-media retrieval," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 307–316.

[3] X. Mao, B. Lin, D. Cai, X. He, and J. Pei, "Parallel field alignment for cross media retrieval," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 897–906.

[4] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in neural information processing systems*, 2014, pp. 1889–1897.

[5] S. Wu, A. Oerlemans, E. M. Bakker, and M. S. Lew, "Deep binary codes for large scale image retrieval," *Neurocomputing*, vol. 257, pp. 5–15, 2017. [Online]. Available: https://doi.org/10.1016/j.neucom.2016.12.070

[6] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016. [Online]. Available: https://doi.org/10.1016/j.neucom.2015.09.116

[7] X.-F. Han, H. Laga, and M. Bennamoun, "Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[9] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3232–3240.

[10] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4242–4251.

Method	MIRFLICKR-25K						NUS-WIDE					
	Image query Text			Text query Image			Image query Text			Text query Image		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
CMSSH[20]	0.5600	0.5709	0.5836	0.5726	0.5776	0.5753	0.3092	0.3099	0.3396	0.3167	0.3171	0.3179
SCM[21]	0.6354	0.5618	0.5634	0.6340	0.6458	0.6541	0.3121	0.3111	0.3121	0.4261	0.4372	0.4478
SePH[26]	0.6740	0.6813	0.6830	0.7139	0.7258	0.7294	0.4797	0.4859	0.4906	0.6072	0.6280	0.6291
DCMH[9]	0.7316	0.7343	0.7446	0.7607	0.7737	0.7805	0.5445	0.5597	0.5803	0.5793	0.5922	0.6014
CHN[25]	0.7504	0.7495	0.7461	0.7776	0.7775	0.7798	0.5754	0.5966	0.6015	0.5816	0.5967	0.5992
PRDH[24]	0.6952	0.7072	0.7108	0.7626	0.7718	0.7755	0.5919	0.6059	0.6116	0.6155	0.6286	0.6349
SSAH[10]	0.7745	0.7882	0.7990	0.7860	0.7974	0.7910	0.6163	0.6278	0.6140	0.6204	0.6251	0.6215
CMHH[23]	0.7334	0.7281	0.7444	0.7320	0.7183	0.7279	0.5530	0.5698	0.5924	0.5739	0.5786	0.5889
DAPSH	**0.7978**	**0.8097**	**0.8179**	**0.7802**	**0.7946**	**0.8115**	**0.6498**	**0.6787**	**0.6834**	**0.6396**	**0.6529**	**0.6792**

TABLE I. Mean Average Percision (MAP) comparison results

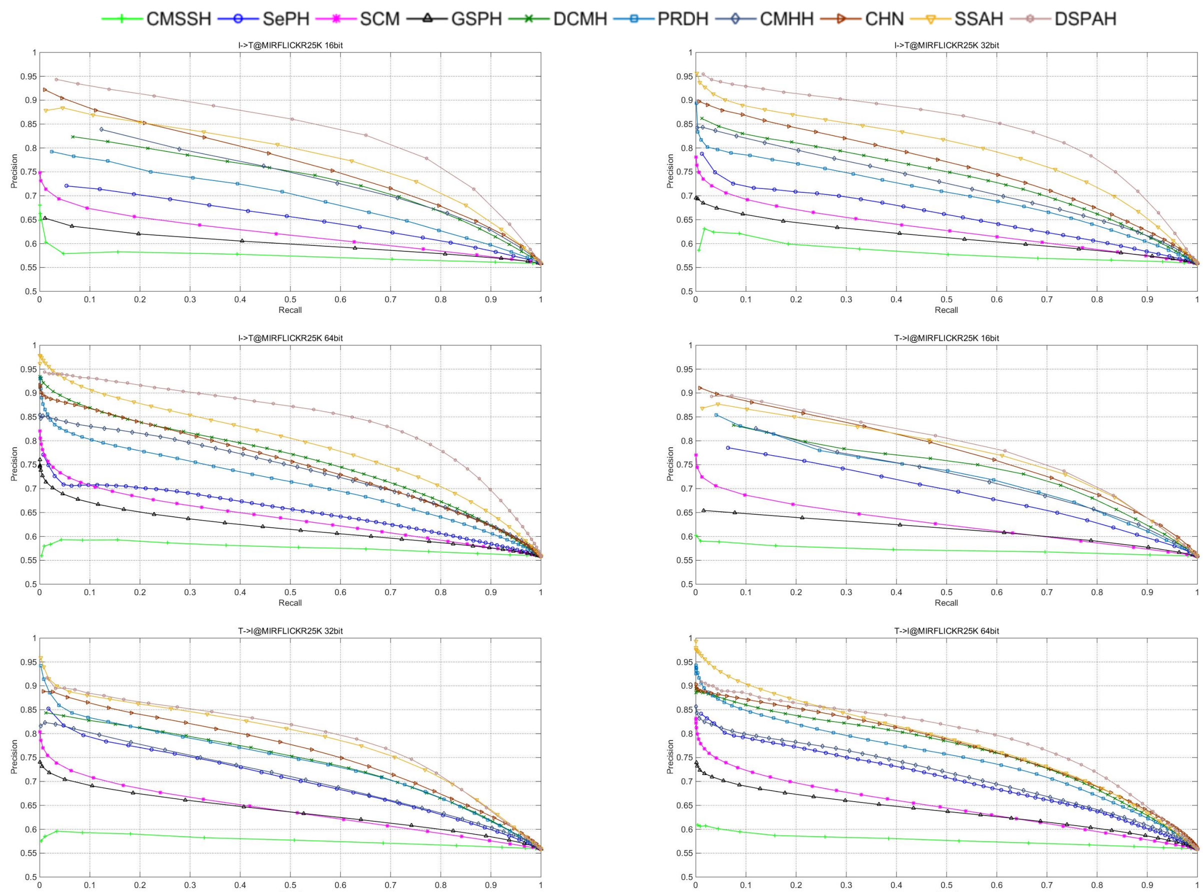

Fig. 2. Performance on MIRFlickr-25K evaluated by PR Curves

[11] X. Wang, X. Zou, E. M. Bakker, and S. Wu, "Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval," *Neurocomputing*, 2020.

[12] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3893–3903, 2018.

[13] X. Zou, X. Wang, E. M. Bakker, and S. Wu, "Multi-label semantics preserving based deep cross-modal hashing," *Signal Processing: Image Communication*, vol. 93, p. 116131, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596520302344

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2016.

[15] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," 2018.

[16] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 39–43.

[17] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on*

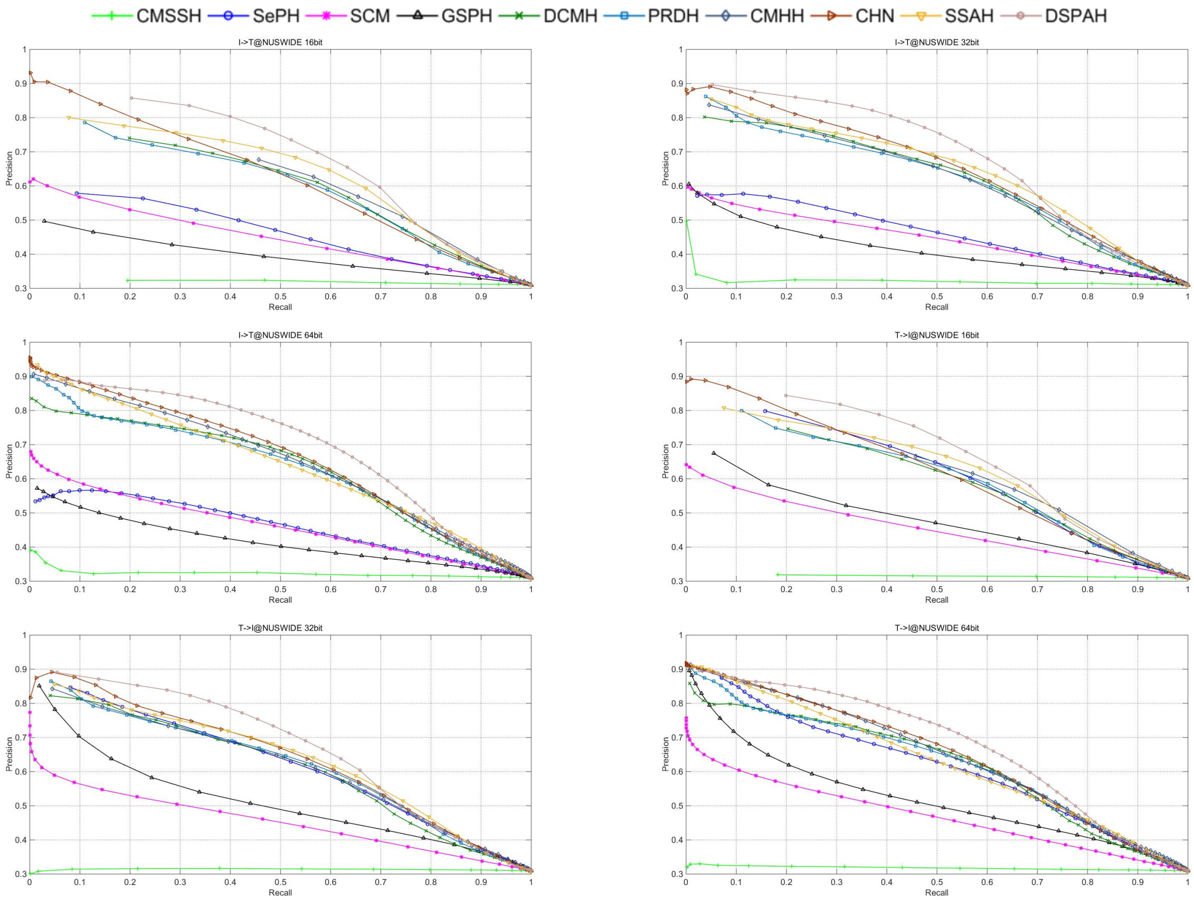

Fig. 3. Performance on NUS-WIDE evaluated by PR Curves

image and video retrieval, 2009, pp. 1–9.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[20] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Computer Vision & Pattern Recognition*, 2010.

[21] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization." in *AAAI*, vol. 1, no. 2. Citeseer, 2014, p. 7.

[22] Devraj, Mandal, Kunal, N, Chaudhury, Soma, and Biswas, "Generalized semantic preserving hashing for cross-modal retrieval." *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, 2018.

[23] Y. Cao, B. Liu, M. Long, and J. Wang, "Cross-modal hamming hashing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 202–218.

[24] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[25] Y. Cao, M. Long, J. Wang, and P. S. Yu, "Correlation hashing network for efficient cross-modal retrieval," *arXiv preprint arXiv:1602.06697*, 2016.

[26] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3864–3872.