CLIP MULTI-MODAL HASHING: A NEW BASELINE

Jian Zhu[1], *Mingkai Sheng*[2], *Mingda Ke*[2], *Zhangmin Huang*[1], *Jingfei Chang*[1]

[1] Zhejiang Lab, China {qijian.zhu, zmhuang, cjf_chang}@zhejianglab.edu.cn
[2] University of Chinese Academy of Sciences, China {shengmingkai22, kemingda21}@mails.ucas.ac.cn

ABSTRACT

The multi-modal hashing method is widely used in multimedia retrieval. It can fuse multi-source data to generate binary hash code. However, the current multi-modal methods have the problem of low retrieval accuracy. The reason is that the individual backbone networks have limited feature expression capabilities and are not jointly pre-trained on large-scale unsupervised multi-modal data. To solve this problem, we propose a new baseline CLIP Multi-modal Hashing (CLIPMH) method. It uses CLIP model to extract text and image features, and then fuse to generate hash code. CLIP improves the expressiveness of each modal feature. In this way, it can greatly improve the retrieval performance of multi-modal hashing methods. In comparison to state-of-the-art unsupervised and supervised multi-modal hashing methods, experiments reveal that the proposed CLIPMH can significantly enhance performance (Maximum increase of 8.38%). CLIP also has great advantages over the text and visual backbone networks commonly used before. The source codes of our CLIPMH is publicly available at: https://github.com/xxx.

Index Terms— Multi-view Hash, CLIP, Multi-modal Hash, Multi-view Fusion

1. INTRODUCTION

Multi-modal hashing is one of the important technologies in the field of multimedia retrieval. It is the fusion of multi-modal heterogeneous data to generate hash codes.

The current multi-modal hashing methods have the problem of low retrieval accuracy. The reason is that the backbone networks lack good feature expression capability. For instance, Flexible Multi-modal Hashing (FDH) [1] and Bit-aware Semantic Transformer Hashing (BSTH) [2] hire a VGG net [3] for the image modal. And these methods use a Bag-of-Words model [4] for the text modal. They are outdated for feature extraction, thus, an update in feature extraction methods is necessary. The fact above results in a degradation of the overall retrieval accuracy for the current multi-modal hashing method.

In recent years, multi-modal large-scale models have achieved great success. Because these models are trained on large-scale data, they have stronger semantic expression ability. Contrastive Language-Image Pre-training(CLIP) [5] is one of the most representative multi-modal models. However, The application of a multi-modal large model in multi-modal retrieval has not been studied. For the first time, we investigate how CLIP affects the retrieval efficiency of multi-view hashing. As shown in Fig. 1, it is pre-trained by contrastive learning on large-scale image text data pairs. It has shown exceptional zero-shot or few-shot learning abilities as well as excellent semantic understanding capabilities. The multi-modal field has been greatly changed by CLIP, and more people are beginning to acknowledge that it is superior at multi-modal tasks. Although CLIP has undergone multiple successful trials, a thorough analysis of its effects and performance on multi-modal hashing retrieval has not yet been conducted.

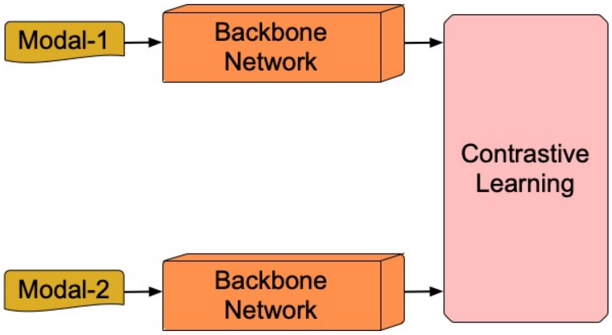

Fig. 1

We perform an in-depth study in this work to examine the potential of CLIP on retrieval from multi-modal hashing. We use the CLIP model to extract text and image features. The extracted modal feature data through the CLIP model performs better. It can significantly improve the retrieval performance of multi-modal hashing methods. Compared with the latest state-of-the-art method, the CLIPMH proposed by us has a maximum improvement of 8.38%.

Here is a summary of the key contributions of our method:

- We have studied for the first time the improvement of retrieval performance of multi-modal hashing methods through multi-modal large models.
- We solve the problem of poor semantic representation in the backbone network of multi-modal hashing methods using the CLIP model.
- We propose a new multi-modal hashing method termed CLIPMH, which achieves the state-of-the-art result.

2. THE PROPOSED METHODOLOGY

Deep multi-view hashing network is designed to convert multi-view data into hash code. As shown in Fig. 2, CLIPMH consists of CLIP backbones, a multi-modal fusion module, and a hash layer. These modules are described in detail below.

1. **Vision Backbone:** CLIP [5] is employed to produce visual features.
2. **Text Backbone:** CLIP [5] is utilized to extract text features.
3. **Multi-View Fusion Module:** We employ Context Gating to fuse the concatenated visual and text features. The multi-view

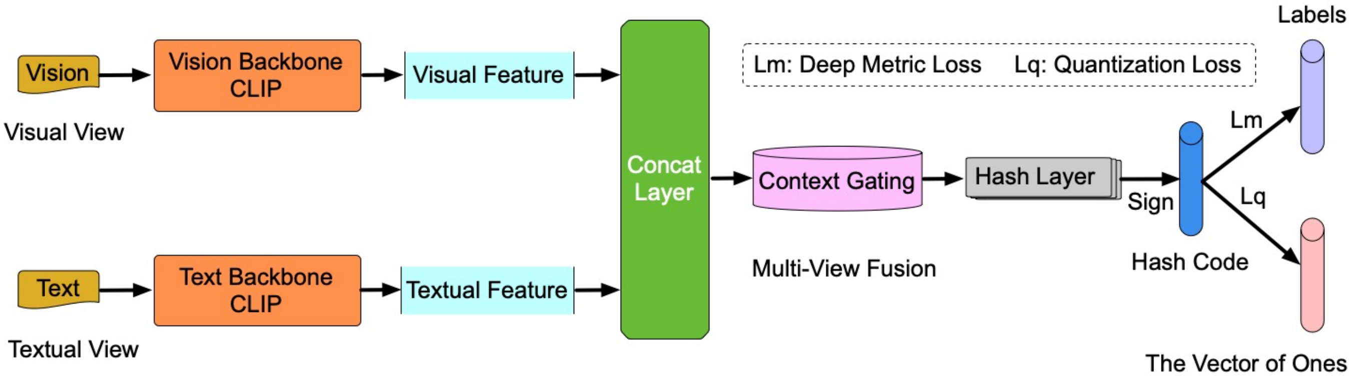

Fig. 2

Table 1: General statistics of three datasets. The dataset size, number of categories, and feature dimensions are included.

Dataset	Training Size	Retrieval Size	Query Size	Categories	Visual Embedding	Textual Embedding
MIR-Flickr25K	5000	17772	2243	24	512-D	512-D
NUS-WIDE	21000	193749	2085	21	512-D	512-D
MS COCO	18000	82783	5981	80	512-D	512-D

fusion module projects the input multi-modal features into a new global representation as:

$$X_{\text{fusion}} = \sigma(w_{\text{fusion}} X_{\text{concat}} + b_{\text{fusion}}) \circ X_{\text{concat}}, \qquad (1)$$

where $X_{\text{concat}} \in \mathbb{R}^n$ is the multi-view feature vector, σ is the element-wise sigmoid activation, and $\circ$ is the element-wise multiplication. $w_{\text{fusion}} \in \mathbb{R}^{n \times n}$ and $b_{\text{fusion}} \in \mathbb{R}^n$ are trainable parameters. The vector of weights $\sigma(w_{\text{fusion}} X_{\text{concat}} + b_{\text{fusion}}) \in [0, 1]$ represents a set of learned gates applied to the individual dimensions of the input feature X_{concat}.

4. **Hash Layer:** A linear layer with a $\tanh$ activation is hired as the hash layer, which can be represented as $h_{\text{k-bit}} = \text{sgn}[\tanh(w_{\text{hash}} X_{\text{fusion}} + b_{\text{hash}})]$, where sgn represents the signum function. $w_{\text{hash}} \in \mathbb{R}^{n \times n}$ and $b_{\text{hash}} \in \mathbb{R}^n$ are trainable parameters. The output has the same number of dimensions as the hash code.

3. EXPERIMENTS

3.1. Evaluation Datasets and Metrics

In the experiments, we evaluate the performance of the proposed CLIPMH model on large-scale multimedia retrieval tasks. We utilize three well-known datasets: MIR-Flickr25K [6], NUS-WIDE [7], and MS COCO [8]. These datasets have gained widespread usage for evaluating the performance of multimedia retrieval systems. The mean Average Precision (mAP) is employed as the evaluation metric. Table 1 provides a summary of the dataset statistics used in the experiments.

3.2. Baseline

To evaluate the retrieval metric, we compare the proposed CLIPMH method with thirteen multi-view hashing methods, including four unsupervised methods (e.g., Multiple Feature Hashing (MFH) [9], Multi-view Alignment Hashing (MAH) [10], Multi-view Latent Hashing (MVLH) [11], and Multi-view Discrete Hashing (MvDH) [12]) and nine supervised methods (e.g., Multiple Feature Kernel Hashing (MFKH) [13], Discrete Multi-view Hashing (DMVH) [14], Flexible Discrete Multi-view Hashing (FDMH) [15], Flexible Online Multi-modal Hashing (FOMH) [16], Deep Collaborative Multi-View Hashing (DCMVH) [17], Supervised Adaptive Partial Multi-view Hashing (SAPMH) [18], Flexible Graph Convolutional Multi-modal Hashing (FGCMH) [19], Bit-aware Semantic Transformer Hashing (BSTH) [2] and Deep Metric Multi-View Hashing (DMMVH) [20]).

3.3. Analysis of Experimental Results

The mAP results are presented in Table 2. The results demonstrate that the proposed CLIPMH method outperforms all the compared multi-view hashing methods by a significant margin. Specifically, when compared to the current state-of-the-art multi-modal hashing method DMMVH [20], our method achieves an average mAP improvement of 2.00%, 1.43%, and 4.80% on the MIR-Flickr25K, NUS-WIDE, and MS COCO datasets, respectively. These superior results can be attributed to three main factors:

- We use the CLIP model to build a multi-modal hashing method.
- The CLIP model extracts image features and enhances semantic expression.
- The CLIP model extracts text features and enhances semantic expression.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a new multi-modal hashing method termed CLIPMH. Experiments show that our method achieves state-of-the-art results. CLIP model is used to solve the problem that backbone network extraction feature semantic expression is insufficient in the

Table 2: The comparable mAP results on MIR-Flickr25K, NUS-WIDE, and MS COCO. The best results are bolded, and the second-best results are underlined. The * indicates that the results of our method on this dataset are statistical significance.

Method	Ref.	MIR-Flickr25K*				NUS-WIDE*				MS COCO*			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
MFH	TMM13	0.5795	0.5824	0.5831	0.5836	0.3603	0.3611	0.3625	0.3629	0.3948	0.3699	0.3960	0.3980
MAH	TIP15	0.6488	0.6649	0.6990	0.7114	0.4633	0.4945	0.5381	0.5476	0.3967	0.3943	0.3966	0.3988
MVLH	MM15	0.6541	0.6421	0.6044	0.5982	0.4182	0.4092	0.3789	0.3897	0.3993	0.4012	0.4065	0.4099
MvDH	TIST18	0.6828	0.7210	0.7344	0.7527	0.4947	0.5661	0.5789	0.6122	0.3978	0.3966	0.3977	0.3998
MFKH	MM12	0.6369	0.6128	0.5985	0.5807	0.4768	0.4359	0.4342	0.3956	0.4216	0.4211	0.4230	0.4229
DMVH	ICMR17	0.7231	0.7326	0.7495	0.7641	0.5676	0.5883	0.6902	0.6279	0.4123	0.4288	0.4355	0.4563
FOMH	MM19	0.7557	0.7632	0.7564	0.7705	0.6329	0.6456	0.6678	0.6791	0.5008	0.5148	0.5172	0.5294
FDMH	NPL20	0.7802	0.7963	0.8094	0.8181	0.6575	0.6665	0.6712	0.6823	0.5404	0.5485	0.5600	0.5674
DCMVH	TIP20	0.8097	0.8279	0.8354	0.8467	0.6509	0.6625	0.6905	0.7023	0.5387	0.5427	0.5490	0.5576
SAPMH	TMM21	0.7657	0.8098	0.8188	0.8191	0.6503	0.6703	0.6898	0.6901	0.5467	0.5502	0.5563	0.5672
FGCMH	MM21	0.8173	0.8358	0.8377	0.8606	0.6677	0.6874	0.6936	0.7011	0.5641	0.5273	0.5797	0.5862
BSTH	SIGIR22	0.8145	0.8340	0.8482	0.8571	0.6990	0.7340	0.7505	0.7704	0.5831	0.6245	0.6459	0.6654
DMMVH	ICME23	0.8587	0.8707	0.8798	0.8827	0.7714	0.7820	0.7879	0.7916	0.6716	0.7030	0.7122	0.7244
CLIPMH	Proposed	**0.8862**	**0.8921**	**0.8956**	**0.8975**	**0.7802**	**0.7986**	**0.8029**	**0.8085**	**0.6806**	**0.7450**	**0.7693**	**0.8082**

multi-modal hashing method. We provide a new baseline method for the multimodal hashing domain. In the future, we will study more application issues of multi-modal large models in the field of multimedia retrieval.

5. REFERENCES

[1] L. Zhu, X. Lu, Z. Cheng, J. Li, and H. Zhang, "Flexible multi-modal hashing for scalable multimedia retrieval," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 2, pp. 1–20, 2020.

[2] W. Tan, L. Zhu, W. Guan, J. Li, and Z. Cheng, "Bit-aware semantic transformer hashing for multi-modal retrieval," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 982–991.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International journal of machine learning and cybernetics*, vol. 1, no. 1, pp. 43–52, 2010.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[6] 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.

[7] 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 image and video retrieval*, 2009, pp. 1–9.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[9] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1997–2008, 2013.

[10] L. Liu, M. Yu, and L. Shao, "Multiview alignment hashing for efficient image search," *IEEE Transactions on image processing*, vol. 24, no. 3, pp. 956–966, 2015.

[11] X. Shen, F. Shen, Q.-S. Sun, and Y.-H. Yuan, "Multi-view latent hashing for efficient multimedia search," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 831–834.

[12] X. Shen, F. Shen, L. Liu, Y.-H. Yuan, W. Liu, and Q.-S. Sun, "Multiview discrete hashing for scalable multimedia search," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 5, pp. 1–21, 2018.

[13] X. Liu, J. He, D. Liu, and B. Lang, "Compact kernel hashing with multiple features," in *Proceedings of the 20th ACM international conference on multimedia*, 2012, pp. 881–884.

[14] R. Yang, Y. Shi, and X.-S. Xu, "Discrete multi-view hashing for effective image retrieval," in *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, 2017, pp. 175–183.

[15] L. Liu, Z. Zhang, and Z. Huang, "Flexible discrete multi-view hashing with collective latent feature learning," *Neural Processing Letters*, vol. 52, no. 3, pp. 1765–1791, 2020.

[16] X. Lu, L. Zhu, Z. Cheng, J. Li, X. Nie, and H. Zhang, "Flexible online multi-modal hashing for large-scale multimedia retrieval," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1129–1137.

[17] L. Zhu, X. Lu, Z. Cheng, J. Li, and H. Zhang, "Deep collaborative multi-view hashing for large-scale image search," *IEEE Transactions on Image Processing*, vol. 29, pp. 4643–4655, 2020.

[18] C. Zheng, L. Zhu, Z. Cheng, J. Li, and A.-A. Liu, "Adaptive partial multi-view hashing for efficient social image retrieval," *IEEE Transactions on Multimedia*, vol. 23, pp. 4079–4092, 2020.

[19] X. Lu, L. Zhu, L. Liu, L. Nie, and H. Zhang, "Graph convolutional multi-modal hashing for flexible multimedia retrieval," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1414–1422.

[20] J. Zhu, Z. Huang, X. Ruan, Y. Cui, Y. Cheng, and L. Zeng, "Deep metric multi-view hashing for multimedia retrieval," *arXiv preprint arXiv:2304.06358*, 2023.