Multi-head Hashing with Orthogonal Decomposition for Cross-modal Retrieval

Wei Liu[1], Jun Li[1(✉)], Zhijian Wu[2], Jianhua Xu[1], and Bo Yang[3]

[1] School of Computer and Electronic Information, Nanjing Normal University, Nanjing 210023, China
lijuncst@njnu.edu.cn

[2] School of Data Science and Engineering, East China Normal University, Shanghai 200062, China

[3] School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China

Abstract. Recently, cross-modal hashing has become a promising line of research in cross-modal retrieval. It not only takes advantage of complementary multiple heterogeneous data modalities for improved retrieval accuracy, but also enjoys reduced memory footprint and fast query speed due to efficient binary feature embedding. With the boom of deep learning, convolutional neural network (CNN) has become the de facto method for advanced cross-model hashing algorithm. Recent research demonstrates that dominant role of CNN is challenged by increasingly effective Transformer architectures due to their advantages of long-range modeling by relaxing local inductive bias. However, the absence of inductive bias shatters the inherent geometric structure, which inevitably leads to compromised neighborhood correlation. To alleviate this problem, in this paper, we propose a novel cross-modal hashing method termed Multi-head Hashing with Orthogonal Decomposition (MHOD) for cross-modal retrieval. More specifically, with the multi-modal Transformers used as the backbones, MHOD leverages orthogonal decomposition for decoupling local cues and global features, and further captures their intrinsic correlations through our designed multi-head hash layer. In this way, the global and local representations are simultaneously embedded into the resulting binary code, leading to a comprehensive and robust representation. Extensive experiments on popular cross-modal retrieval benchmarking datasets demonstrate the proposed MHOD method achieves advantageous performance against the other state-of-the-art cross-modal hashing approaches.

Keywords: Cross-modal Retrieval · Transformer · Orthogonal Decomposition · Multi-head Hashing · Aggregated Hash Codes

Supported by the National Natural Science Foundation of China under Grant 62173186, 62076134, 62303230 and Jiangsu provincial colleges of Natural Science General Program under Grant 22KJB510004.

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14555, pp. 170–183, 2024.
https://doi.org/10.1007/978-3-031-53308-2_13

1 Introduction

With the rapid growth of multi-modal data including images, texts, videos and audios, cross-modal retrieval [6,9,15,30] aims to perform fast and accurate retrieval among different modalities, e.g., text-to-image or image-to-text retrieval. By fully exploiting the complementarity among multiple data modalities, inter-modality correlation can be uncovered in depth for significantly improved retrieval accuracy. With the dramatic expansion of multi-modal data volume, achieving efficient cross-modal retrieval is becoming increasingly urgent. Emerging as a popular line of research in cross-modal retrieval, cross-modal hashing [2,10,24] aims to project data of different modalities onto a Hamming space, yielding compact hash codes of maintained similarity for binary feature embedding.

As deep learning has prospered in recent years, the most representative convolutional neural network (CNN) has considerably advanced the cross-modal hashing for unprecedented performance improvements [1,2,10]. In particular, with the rise of Transformer architecture and large-scale pre-trained models [14,19,22], there is a major shift from the CNN-based to the Transformer-based methods [7,8,24], since the latter demonstrates superior performance in cross-modal hashing. In particular, the Transformer-based vision-language models including BERT [4,14] and CLIP [19] can well interpret and encode semantic representations of both images and text for accurate cross-modal retrieval. Aiming to capture global dependencies in sequential data, the Transformer model [25] learns global contextual information by employing a self-attention mechanism to assess the relative importance of each position with respect to other positions, making the individual local contents downplayed in the feature embedding. Although massive efforts are devoted to combining global and local information for generating more discriminative hashing codes, most studies [16,18,26] perform feature fusion prior to hashing process without exploring intrinsic correlation between global and local cues in the process of binary embedding. In this sense, the feature fusion is relatively independent of the binary embedding, leading to the hashing codes which lack global-local perception capability.

To address the above-mentioned drawback, in this study, we propose a novel cross-modal hashing method which is multi-head hashing with orthogonal decomposition (MHOD) for cross-modal retrieval. In MHOD, an orthogonal decomposition module is imposed on local tokens and global features derived from multi-modal Transformer backbones, resulting in a set of tokens serving as the local features. Next, both the local and global features are delivered to our designed multi-head hashing layer to generate separate hashing codes. The resulting binary codes are aggregated using a pooling-like operation, enabling the combination of local and global information in a unified binary representation. To summarize, the contributions of our work are threefold as follows:

– We leverage orthogonal decomposition for decoupling the local cues and the global features, such that both local and global information can be fully encoded in our cross-modal hashing framework.

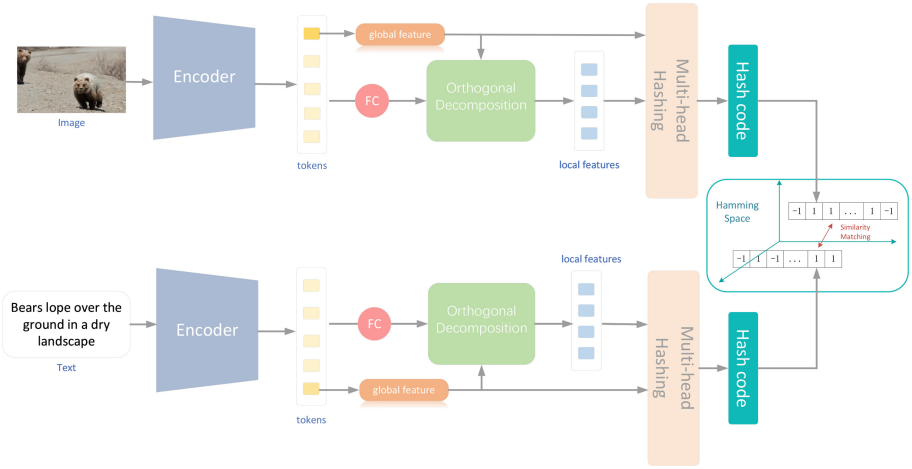

Fig. 1. The network architecture of the proposed MHOD. It consists of three primary blocks including feature encoder used as Transformer backbone, orthogonal decomposition module used for decoupling local cues and global features along with multi-head hashing layer for generating aggregated hash codes. The resulting hash codes can be exploited for accurate and fast cross-modal retrieval in the Hamming space. Different from the existing methods, our method is capable of simultaneously integrating local and global features into binary hashing within Transformer-based cross-modal hashing framework, and considerably benefits mining the intrinsic correlation among different modalities for accurate retrieval.

- To further explore the intrinsic correlation between the local and global features, we simultaneously integrate them into our designed multi-head hashing layer to generate aggregated hash codes with preferable global-local perception capability. This is in contrast to the previous methods in which feature fusion and binary hashing are separately handled.
- Extensive experiments on two public benchmarking cross-modal retrieval datasets demonstrate the superiority of our proposed MHOD against the other state-of-the-art cross-modal hashing models.

The remainder of this paper is organized as follows. We elaborate on our proposed MHOD model in Sect. 2 and carry out extensive experimental evaluations in Sect. 3. The paper is finally concluded in Sect. 4.

2 The Proposed Method

2.1 The Model Framework

While Transformer-based cross-modal hashing methods have achieved considerable success, they either overlook the local clues during the hashing process or perform local-global coupling independent of binary embedding, leading to

hash codes with degraded local-global perception. To address the drawback, we propose a cross-modal hashing method termed MHOD for cross-modal retrieval in this study. As shown in Fig. 1, our MHOD first leverages feature encoder modules for generating global features from class tokens across data modalities. Subsequently, local clues are decoupled from the local tokens via orthogonal decomposition and combined with global features in the multi-head hashing layer, producing aggregated hash codes for cross-modal retrieval. Next, We will discuss these key modules and the training loss functions in details.

2.2 Feature Encoder

For notation, $D = \{X_i, Y_i\}_{i=1}^{N}$ denotes a batch of pairwise data modalities, while N is the number of instances. Since we mainly focus on two different modalities of image and text in cross-modal hashing, X_i and Y_i indicate the original i^{th} image and text instance respectively. In each training batch, $F \in R^{N \times d}$ denotes the feature embedding extracted from the feature encoder where d is the dimension of feature embedding. In addition, F^I and F^T respectively represent the feature of image and text modality.

With the help of the successful pre-trained large models, we adopt the pre-trained CLIP model as our feature encoder for both image and text input. More specifically, image encoder is used as the vision Transformer structure ViT [5], while the text encoder is GPT2 [20] which is a modified architecture developed from Transformer. For brevity, the CLIP encoders are denoted as $CLIP$ which takes the cross-modal image and text as input. Mathematically, the feature encoding process of raw data can be formulated as:

$$F_g, F_t = CLIP(D) \tag{1}$$

where F_g is essentially the object-specific class token obtained by feature encoding from a global perspective, while F_t denotes the remaining local-aware tokens.

2.3 Orthogonal Decomposition

Since local tokens F_t in Eq. (1) also characterize certain global information via self-attention mechanism for exploring local interaction, it is necessary to decouple the local clues and the global information for deriving local features from F_t. Following [27], consequently, we leverage the orthogonal decomposition module for decoupling local cues and the global features. Specifically, F_g and F_t are treated as the input, while F_t^k denotes the k^{th} token in F_t. At first, the tokens pass through consecutive Fully-Connected layers (FCs) such that they have the same dimension as the global features. Afterwards, each token F_t^k is projected onto the global feature F_g, which can be mathematically expressed as follows:

$$F_{t,proj}^k = \frac{F_t^k \cdot F_g}{\|F_g\|_2^2} F_g \tag{2}$$

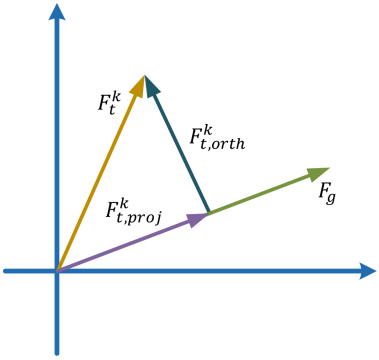

Fig. 2. Illustration of orthogonal decomposition for deriving the local features from local tokens. For notation, F_g represents global feature and F_t^k indicates a local token. In order to decouple F_t^k and F_g, we first project F_t^k onto F_g to obtain $F_{t,proj}^k$ which encodes the global component related to F_g. Thus, the orthogonal component $F_{t,orth}^k$ obtained by subtracting $F_{t,proj}^k$ from F_t^k can be treated as the local feature that is independent of F_g.

where $F_t^k \cdot F_g$ indicates dot product operation and $\|\cdot\|_2$ is ℓ_2 norm. As demonstrated in Fig. 2, the orthogonal component can be calculated as the difference between F_t^k and its projection onto F_g, which is formulated as:

$$F_{t,orth}^k = F_t^k - F_{t,proj}^k \tag{3}$$

In this way, $F_{t,orth}^k$ is independent of the global feature F_g, and can be treated as the local feature F_l that is separated from the global clues.

2.4 Multi-head Hashing

To leverage both the global and local information effectively, the decoupled global and local features are forwarded to hashing layer for generating and aggregating efficient binary hashing codes. In MHOD, the hashing layer includes multiple heads as shown in Fig. 3. Each head comprises a MLP, a *tanh* activation function, and a *sign* function. The MLP maps high-dimensional features to low-dimensional ones. Using the *tanh* function, the low-dimensional features are rescaled from -1 to 1. Finally, the *sign* function converts these features into discrete binary embeddings for generating the hash code b:

$$b = sign(tanh(MLP(feat))) \tag{4}$$

Each hashing head within our multiple heads receives different input features $feat$, with the first head taking global features. Each of the remaining local feature groups is averaged, producing averaged local features delivered to individual hashing head. Thus a hashing matrix $H \in R^{L \times M}$ can be derived, where M is the number of hashing heads and L denotes the length of hash code. We generate multiple hash codes from global and local vectors, but only one hash code

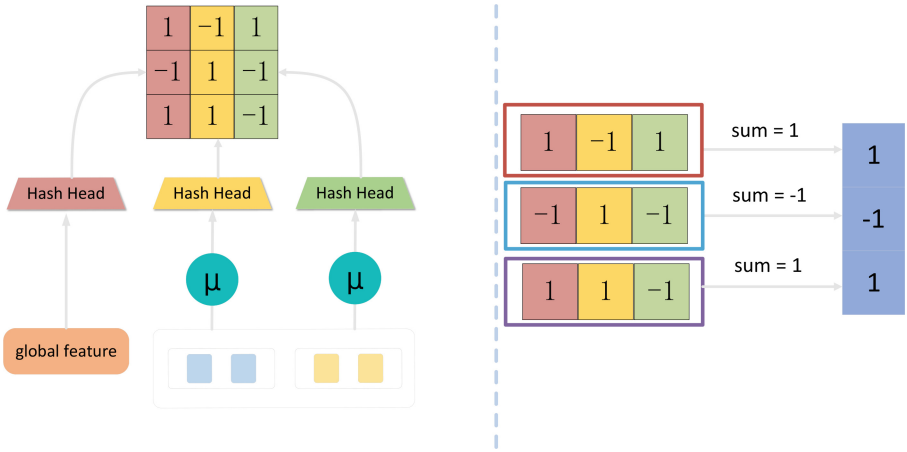

Fig. 3. The structure of our designed multi-head hashing module (left) and accumulation mechanism for aggregating multi-head output into the final hash code (right). Each hashing head contains a MLP, a *tanh* activation function, and a *sign* function. μ represents the mean value operation which is used to average the local features. With the hash codes derived from multiple hashing heads, a simple accumulation strategy is adopted to generate the aggregated hash code.

contributes to the final feature matching. The matching is refined at each bit, handling each bit of the final hash code via a voting mechanism. This ensures that each bit of the hash code is optimized. As illustrated in Fig. 3, we adopt a simple fusion mechanism to aggregate the hashing codes resulting from different heads. Mathematically, it can be formulated as:

$$B_i = \begin{cases} 1 & \sum_{m=1}^{M} O(i,m) > 0 \\ -1 & otherwise \end{cases} \tag{5}$$

Specifically, i^{th} element of the hash code B_i is set as 1 when the sum of i^{th} row in H is greater than 0, and -1 conversely.

2.5 Loss Function

The loss function of our MHOD network for model training includes two critical components, namely similarity loss and hashing loss. Let f_I^i denote the feature for image i and f_T^j for the corresponding text j. With label a_i for each sample, the semantic relation of two different samples can be defined as:

$$A_{ij} = \begin{cases} 1 & a_i \cdot a_j > 0 \\ 0 & a_i \cdot a_j = 0 \end{cases} \tag{6}$$

Similarity Loss. For similarity loss, we evaluate both the inter-modality and intra-modality feature similarity. The similarity S based on Euclidean distance is defined as:

$$S_{ij}^{IT} = \left\| f_I^i - f_T^j \right\|_2 \tag{7}$$

For each sample within the same batch, there exist positive samples with the same label and negative samples with different labels. Consequently, the positive sample similarity P^{IT} and negative sample similarity N^{IT} can be respectively computed as:

$$P^{IT} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(S_{ij}^{IT} \cdot A_{ij}^{IT} \right)^2 \tag{8}$$

$$N^{IT} = \frac{1}{N^2} \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\left(\sqrt{L} - S_{ij}^{TT} \right) \cdot \left(1 - A_{ij}^{IT} \right) \right)^2 \tag{9}$$

Therefore, cross-modal similarity can be formulated as:

$$L_{sim}^{IT} = P^{IT} + N^{IT} \tag{10}$$

Similarly, the image-related and text-specific intra-modality similarity L_{sim}^{II} and L_{sim}^{TT} can be obtained respectively, and the complete similarity loss L_{sim} is:

$$L_{sim} = L_{sim}^{IT} + L_{sim}^{II} + L_{sim}^{TT} \tag{11}$$

Hashing Loss. In addition to similarity loss, hashing loss L_{hash} aims to calculate the information loss of binary embedding:

$$L_{hash} = \frac{1}{M} \left(\sum_{m=1}^{M} H_I^m + \sum_{m=1}^{M} H_T^m \right) \tag{12}$$

where H represents modality-specific hashing loss. More specifically, image-related hashing loss H_I can be formulated as:

$$H_I = \frac{1}{N} \sum_{n=1}^{N} \sqrt{\sum_{l=1}^{L} \left(f_I^{(n,l)} - h_I^{(n,l)} \right)^2} \tag{13}$$

where h is computed as $sign(f)$. Besides, H_T can be calculated analogously. Notably, the modality-specific hashing loss is computed independently for each hashing head.

Overall Loss. The overall loss function is the weighted sum of the above-mentioned similarity loss and hashing loss:

$$L = L_{sim} + \lambda L_{hash} \tag{14}$$

where λ is a balancing hyper-parameter to compromise between the two terms. It will be discussed in the parameter analysis in the following section of experiments.

Table 1. Comparison of different cross-modal hashing methods in the two public datasets (mAP@all). The best results are highlighted **in bold** and the second best are underlined. The results demonstrate that our method is superior to the other state-of-the-art models in both MS-COCO and NUS-WIDE.

Dataset	Method		I to T			T to I	
		16bit	32bit	64bit	16bit	32bit	64bit
MS-COCO	DCMH [10]	0.5533	0.5540	0.5667	0.5272	0.5467	0.5521
	SCAHN [12]	0.6095	0.6502	0.6435	0.6035	0.6403	0.6435
	MSSPQ [31]	0.5710	0.5862	0.5881	0.5472	0.5630	0.5985
	DADH [1]	0.6388	0.6668	0.6812	0.6027	0.6334	0.6528
	DCHMT [24]	0.6447	0.6757	0.6915	0.6531	0.6832	0.7025
	MHOD (Ours)	**0.6595**	**0.6870**	**0.7056**	**0.6713**	**0.6940**	**0.7130**
NUS-WIDE	DADH [1]	0.6492	0.6662	0.6664	0.6501	0.6679	0.6808
	DMFH [18]	0.6065	0.6212	0.6396	0.6307	0.6468	0.6798
	TEACH [28]	0.6512	0.6643	0.6704	0.6732	0.6871	0.6893
	DCHMT [24]	0.6799	0.6992	0.7038	0.6876	0.7104	0.7253
	SCAHN [12]	0.6155	0.6403	0.6662	0.6446	0.6702	0.6980
	MHOD (Ours)	**0.6992**	**0.7083**	**0.7168**	**0.7041**	**0.7135**	**0.7299**

3 Experiments

3.1 Datasets and Experimental Settings

We have evaluated our approach in two popular benchmarking datasets for cross-modal retrieval, i.e., MS-COCO [13] and NUS-WIDE [3]. On both datasets, we randomly select 10,000 samples for training data, 5,000 samples as queries and the rest as retrieval database. We initialize both the image and text encoders with a pre-trained CLIP(ViT-B/32) model. In our proposed MHOD, Adam optimizer [11] is used for model training. The initial learning rate is set as 0.001 in MS-COCO, 0.0001 in NUS-WIDE, and 1e-7 for the Transformer encoders. The batch size is set to 64 and the number of hashing heads is 3. In terms of evaluation metric, mAP@all and mAP@50 are used for performance measures. All the experiments are conducted on a server with Intel i9-10900K CPU and one NVIDIA RTX3090 GPU using PyTorch framework.

3.2 Results

Comparative Studies. As demonstrated in Table 1, we have compared our MHOD model with recent eleven state-of-the-art cross-modal hashing methods including DADH [1], DMFH [18], TEACH [28], DCHMT [24], SCAHN [12],

Table 2. Comparison of different cross-modal hashing methods in NUS-WIDE dataset (mAP@50). The best results are highlighted **in bold**.

Method	I to T			T to I		
	16bit	32bit	64bit	16bit	32bit	64bit
DJSRH [23]	0.724	0.773	0.798	0.712	0.744	0.771
HNH [29]	0.582	0.789	0.800	0.423	0.747	0.781
DUCH [17]	0.753	0.775	0.814	0.726	0.758	0.781
DAEH [21]	0.766	0.789	0.809	0.718	0.751	0.766
MHOD (Ours)	**0.806**	**0.817**	**0.836**	**0.766**	**0.773**	**0.783**

Table 3. Ablation studies in MS-COCO using 64bit hashing code (mAP@all). MH denotes our designed multi-head hashing component for generating aggregated hashing codes. It should be noted that the first two methods corresponding to the first two rows directly employ a straightforward *sign* function mapping instead of MH for binary embedding.

Global	Local	MH	I to T	T to I
✓			0.6998	0.7010
	✓		0.6912	0.6970
✓		✓	0.7013	0.7076
	✓	✓	0.6934	0.7012
✓	✓	✓	0.7056	0.7130

DCMH [10], MSSPQ [31], DUCH [17], DAEH [21], DJSRH [23] and HNH [29] in the two benchmarking datasets for different cross-modal retrieval tasks. More specifically, for image-to-text retrieval task, our MHOD reports the highest mAP@all scores at 65.95%, 68.70% and 70.56% in MS-COCO, surpassing DCHMT model by 1.5%, 1.1% and 1.4% when the length of the hashing code is 16, 32 and 64, respectively. In NUS-WIDE, analogous advantages against DCHMT can also be observed with respective performance gains of 1.9%, 0.9% and 1.3% for various hashing codes of different lengths. Similar results are also shown for text-to-image retrieval task, suggesting that our method beats the other competing models in both MS-COCO and NUS-WIDE. In terms of the mAP@50 metric, our MHOD also reports the highest accuracies of 80.6%, 81.7% and 83.6% in NUS-WIDE when the length of hash code is 16, 32 and 64 respectively, revealing consistent advantages against the state-of-the-arts as shown in Table 2.

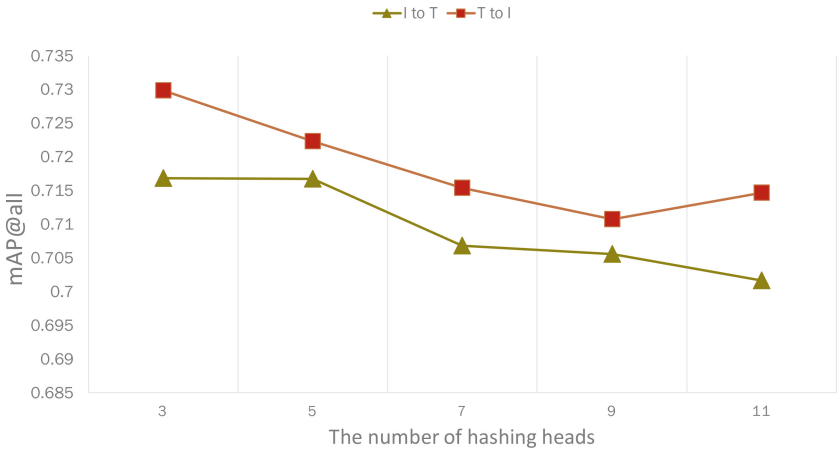

Fig. 4. Performance of the proposed MHOD with different numbers of hashing head in NUS-WIDE using 64bit hashing code.

3.3 Ablation Study

To gain an insight into the effectiveness of each module within our MHOD network, we have carried out comprehensive ablation studies for exploring the effect of individual module on our model. Firstly, we investigate the feature fusion module and compare different strategies of feature embedding. As illustrated in Table 3, combining both local and global features contributes to further performance boost. For text-to-image retrieval task, To be specific, our complete model with local-global fusion provides respective performance gains of 0.5% and 1.2% over the approach with single global feature embedding and single local feature embedding. While the global features plays a dominant role in feature embedding, local contents are supplementary to global feature and conducive to boosting model performance. On the other hand, with our multi-head hashing layer (which is MH for short in Table 3), the retrieval accuracy increases from 70.10% to 70.76% when only global feature embedding is performed. When only considering local cues, similar improvement can also be observed, which substantially suggests the beneficial role of our MH module.

In our MHOD model, the multi-head hashing module consists of multiple hashing heads. In addition to the above ablation studies, we discuss the effect of different head numbers on the model performance. As shown in Fig. 4, the overall declined performance is observed with increasing head number. This can be explained by less tokens assigned to each head when the hashing head increases. Consequently, the amount of information available for each individual hashing head decreases. This reduction in token allocation can adversely affect the retrieval accuracy of the hashing code generated from each head. As a result, the aggregated hashing code combining the outputs of all the heads may still suffer from degraded retrieval accuracy.

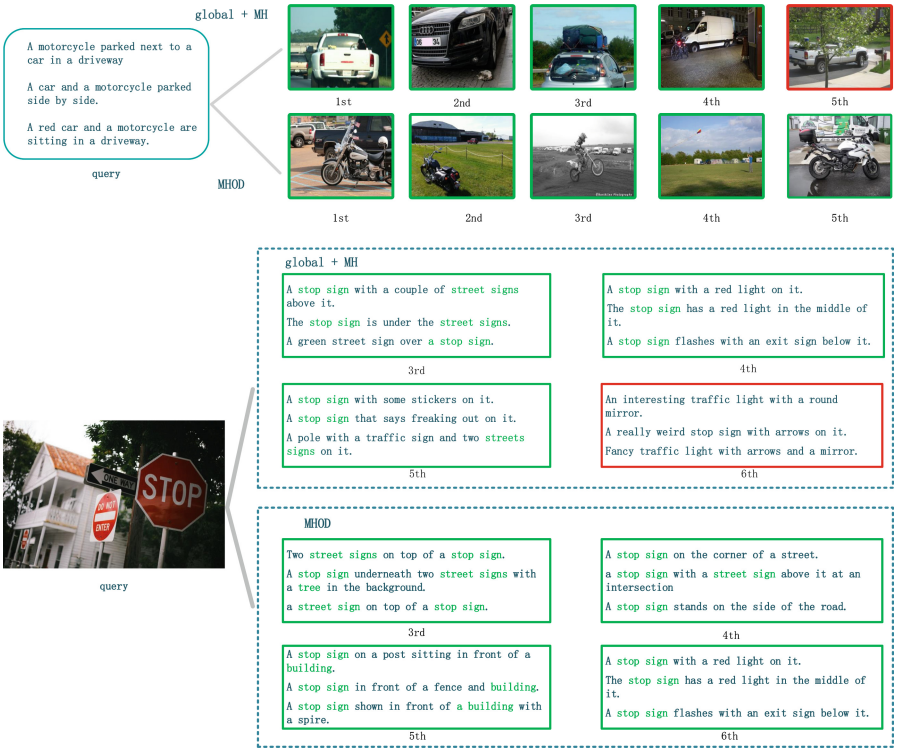

Fig. 5. Demonstration of top returned results achieved by two different methods for different cross-modal retrieval tasks. The query-related ground-truths are highlighted in a green box, whereas the mismatched ones are annotated in a red box. Compared to the method only considering global information, our MHOD model can retrieve more positive images or texts that better match the query by leveraging both local and global contents. (Color figure online)

In addition to the above quantitative results of ablation studies, we also present some qualitative results as shown in Fig. 5. It is shown that compared to only using global feature, our proposed MHOD using both global and local features not only brings a boost in retrieval accuracy, but also helps us find more query-related targets. For instance, for text-to-image retrieval task, the query texts are closely related to two objects, namely car and motorcycle. The model which only focuses on global contents mainly captures the car object while overlooking the other one. In contrast, by taking advantage of both global and local clues, our model can find the images including both car and motorcycle, which implies the beneficial and supplementary role of local information in capturing multiple objects in cross-model retrieval.

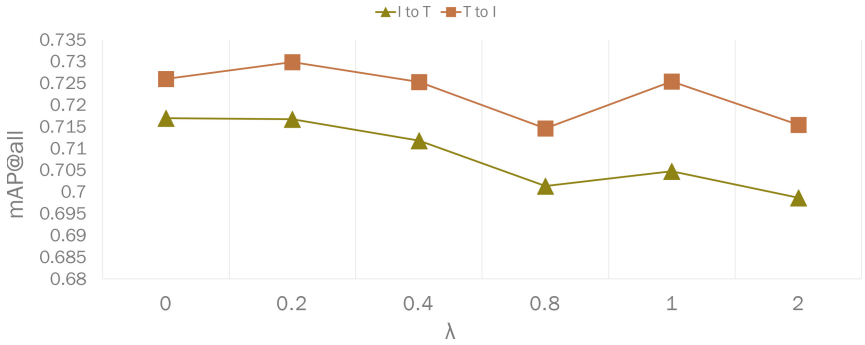

Fig. 6. λ Sensitivity Analysis in NUS-WIDE.

3.4 Parameter Sensitivity Analysis

In the loss function of our MHOD model as formulated in Eq. (14), the hyper-parameter λ needs to be tuned for balancing the similarity loss and hashing loss. Figure 6 demonstrates the performance of our model with varying λ values in NUS-WIDE for different cross-modal retrieval tasks. It can be observed that the best results of 72.99% and 71.68% are reported when λ equals 0.2 for image-to-text retrieval and text-to-image retrieval task. Interestingly, the results even exceed the model performance when λ equals 0 which implies hashing loss is not involved in our model. Profiting from the multi-head hashing module, our model does not severely suffer the information loss resulting from binary embedding of the hashing loss.

4 Conclusions

In this paper, we present a novel cross-modal hashing method termed MHOD for cross-modal retrieval task. More specifically, it leverages the multi-modal Transformer encoders for generating global features and local tokens, which is followed by the orthogonal decomposition module for decoupling the local cues and the global features. Then, the feature fusion is achieved by passing both global and local features through multi-head hashing layer for generating aggregated hash codes. Different from the previous methods in which either local contents are downplayed or the local-global fusion is independent of binary hashing, our method can integrate local-global feature fusion into the hashing process for improved local-global perception capability. Extensive experiments in two public benchmarking datasets show that the proposed MHOD achieves the state-of-the-art performance.

References

1. Bai, C., Zeng, C., Ma, Q., Zhang, J., Chen, S.: Deep adversarial discrete hashing for cross-modal retrieval. In: International Conference on Multimedia Retrieval, pp. 525–531 (2020)
2. Cao, Y., Liu, B., Long, M., Wang, J.: Cross-modal hamming hashing. In: European Conference on Computer Vision, pp. 207–223 (2018)
3. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from national university of Singapore. In: ACM International Conference on Image and Video Retrieval, pp. 368–375 (2009)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics, pp. 4171–4186 (2019)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations, pp. 1–22 (2021)
6. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improved visual-semantic embeddings with hard negatives. In: British Machine Vision Conference, pp. 1–14 (2018)
7. Hong, J., Liu, H.: Deep cross-modal hashing retrieval based on semantics preserving and vision transformer. In: International Conference on Electronic Information Technology and Computer Engineering, pp. 52–57 (2022)
8. Huo, Y., et al.: Deep semantic-aware proxy hashing for multi-label cross-modal retrieval. IEEE Trans. Circuits Syst. Video Technol. (2023)
9. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, pp. 4904–4916 (2021)
10. Jiang, Q.Y., Li, W.J.: Deep cross-modal hashing. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3270–3278 (2016)
11. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference for Learning Representations, pp. 1–15 (2015)
12. Liang, M., et al.: Semantic structure enhanced contrastive adversarial hash network for cross-media representation learning. In: ACM International Conference on Multimedia, pp. 277–285 (2022)
13. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision, pp. 740–755 (2014)
14. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: petraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Neural Information Processing Systems, pp. 13–23 (2019)
15. Ma, L., Li, H., Meng, F., Wu, Q., Ngi Ngan, K.: Global and local semantics-preserving based deep hashing for cross-modal retrieval. Neurocomputing **312**, 49–62 (2018)
16. Ma, X., Zhang, T., Xu, C.: Multi-level correlation adversarial hashing for cross-modal retrieval. IEEE Trans. Multimedia **22**, 3101–3114 (2020)
17. Mikriukov, G., Ravanbakhsh, M., Demir, B.: Deep unsupervised contrastive hashing for large-scale cross-modal text-image retrieval in remote sensing. arXiv preprint arXiv:2201.08125 (2022)
18. Nie, X., Wang, B., Li, J., Hao, F., Jian, M., Yin, Y.: Deep multiscale fusion hashing for cross-modal retrieval. IEEE Trans. Circuits Syst. Video Technol. **31**, 401–410 (2021)

19. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021)
20. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
21. Shi, Y., et al.: Deep adaptively-enhanced hashing with discriminative similarity guidance for unsupervised cross-modal retrieval. IEEE Trans. Circuits Syst. Video Technol. **32**, 7255–7268 (2022)
22. Singh, A., et al.: FLAVA: a foundational language and vision alignment model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 15617–15629 (2022)
23. Su, S., Zhong, Z., Zhang, C.: Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: IEEE International Conference on Computer Vision, pp. 3027–3035 (2019)
24. Tu, J., Liu, X., Lin, Z., Hong, R., Wang, M.: Differentiable cross-modal hashing via multimodal transformers. In: ACM International Conference on Multimedia, pp. 453–461 (2022)
25. Vaswani, A., et al.: Attention is all you need. In: Neural Information Processing Systems, pp. 6000–6010 (2017)
26. Wang, H., Zhao, K., Zhao, D.: A triple fusion model for cross-modal deep hashing retrieval. Multimedia Syst. **29**, 347–359 (2022)
27. Yang, M., et al.: DOLG: single-stage image retrieval with deep orthogonal fusion of local and global features. In: IEEE International Conference on Computer Vision, pp. 11752–11761 (2021)
28. Yao, H.L., Zhan, Y.W., Chen, Z.D., Luo, X., Xu, X.S.: TEACH: attention-aware deep cross-modal hashing. In: International Conference on Multimedia Retrieval, pp. 376–384 (2021)
29. Zhang, P., Luo, Y., Huang, Z., Xu, X.S., Song, J.: High-order nonlocal hashing for unsupervised cross-modal retrieval. World Wide Web **24**, 563–583 (2021)
30. Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10386–10395 (2019)
31. Zhu, L., Cai, L., Song, J., Zhu, X., Zhang, C., Zhang, S.: MSSPQ: multiple semantic structure-preserving quantization for cross-modal retrieval. In: International Conference on Multimedia Retrieval, pp. 631–638 (2022)