Dynamic Masking and Auxiliary Hash Learning for Enhanced Cross-Modal Retrieval

Shuang Zhang, Yue Wu, Lei Shi, Yingxue Zhang, Feifei Kou,Huilong Jin, Pengfei Zhang, Meiyu Liang, Mingying Xu

Challenage

- The significant semantic gap between different modalities often leads to inconsistent cross-modal representations, and much non-critical information or noise can affect matching accuracy, resulting in similar images and texts being mismatched.
- When processing features, traditional hash layers often ignore the importance differences between different channels, which can lead to insufficient capture of key information and difficulty in effectively suppressing noise and redundant information.
- When hash codes are generated, they often rely on a single optimization goal, which can lead to insufficient performance in cross-modal matching.

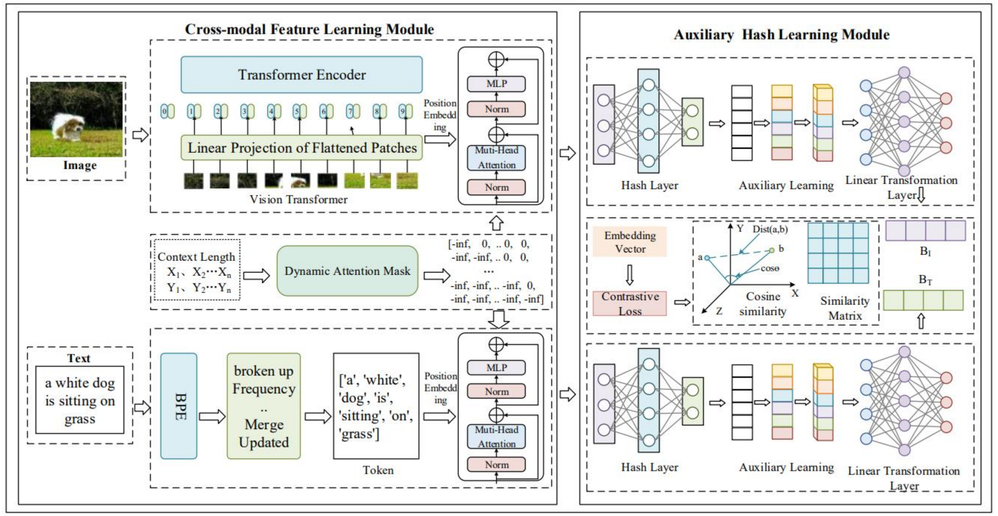
motivation

To deal with this challenge, we proposed a method called auxiliary hashing learning (AHLR) for cross modal retrieval. It signiffcantly improves feature extraction and alignment capability by introducing a dynamic mask mechanism. By constructing an auxiliary hash layer to adaptively weight the features of each channel, the problem of channel information imbalance is solved, while the ability to capture key information is enhanced and noise interference is effectively suppressed.

contribution

- We propose a dynamic masking and auxiliary hash learning (AHLR) method for cross-modal retrieval, which can effectively enhance feature extraction and alignment capabilities, generate more detailed hash codes, and improve the accuracy of cross-modal hashing retrieval.
- We introduce a dynamic masking mechanism to automatically select key information from images and text and weight it, thereby improving the accuracy of feature alignment and matching.
- We design an adaptive auxiliary hash learning cross-modal module that can adaptively weight the features of each channel, enhancing the retention of key information. Moreover, we introduce the contrast loss function to distinguish the similarity and heterogeneity of the samples and improve cross-modal semantic consistency.
- Extensive experiments on three benchmark datasets show that our AHLR outperforms state-of-the-art baselines, demonstrating clear performance advantages.

Network architecture



Model Composition

Cross-modal Feature Learning Module

Image Feature: Use the Vision Transformer to extract deep features from the input image. By introducing a dynamic attention mask M, the key information of the image is automatically selected and weighted according to the length of the input sequence, and the image feature representation $F_I = MIP(X_i)$ is obtained through the Transformer encoder.

Text Feature: Use BPE for subword level encoding, also introduce dynamic attention mask M,and finally obtain text feature representation $F_T = MIP(X_t)$ through text Transformer encoder.

Auxiliary Hash Learning Module

High-level features are mapped to low-dimensional space: $f_i = \mathbf{F}_I\mathbf{W}_I + \mathbf{b}_I, f_t = \mathbf{F}_T\mathbf{W}_T + \mathbf{b}_T$

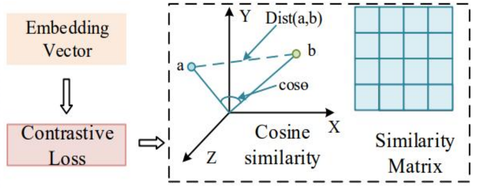
Obtain embedded features after dimensionality reduction and weighting:

$$f_I^* = \mathbf{W}_i \odot f_I', f_T^* = \mathbf{W}_i \odot f_T'$$

The final binary hash code is determined by selecting the maximum probability and obtaining the vector representation of the final hash code at once: $\mathbf{H} = (\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_K) \in \{0,1\}^{\mathbf{K}}$

Loss Function

Contrastive loss is used to calculate the similarity between samples.



Positive sample pairs:

$$\mathcal{L}_{positive} = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{A}_{ij} \cdot (1 - \mathbf{S}_{ij})$$

Negative sample pairs:

$$\mathcal{L}_{negative} = \sum_{i=1}^{N}\sum_{j=1}^{N} (1 - \mathbf{A}_{ij}) \cdot \mathrm{ReLU}(\mathbf{S}_{ij} - \xi)$$

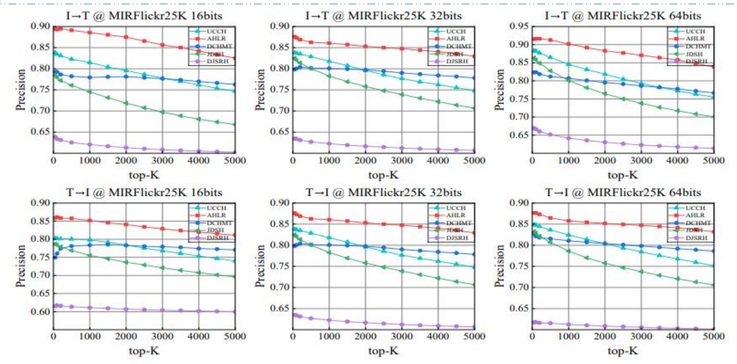
Experiment result

Task	Method	MIRFlickr-25K			NUS-WIDE			MS-COCO		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I→T	DJSRH	0.6652	0.6873	0.6987	0.5271	0.5582	0.6015	0.5257	0.5454	0.5646
	JDSH	0.7276	0.7426	0.7468	0.6536	0.6601	0.6900	0.5928	0.6348	0.6517
	CDTH	0.7317	0.7461	0.7477	0.6596	0.6613	0.6700	0.5853	0.6411	0.6573
	UCCH	0.7606	0.7620	0.7674	0.6718	0.6738	0.6891	0.6039	0.6249	0.6398
	MLCAH	0.7960	0.8080	0.8150	0.6440	0.6410	0.6430	0.5700	0.5620	0.5620
	DCHMT	0.8177	0.8221	0.8261	0.6711	0.6812	0.6932	0.6450	0.6331	0.6647
	AHLR	**0.8203**	**0.8233**	**0.8266**	**0.6777**	**0.6884**	**0.6994**	**0.6454**	**0.6582**	**0.6797**
T→I	DJSRH	0.6710	0.6958	0.7043	0.5575	0.5680	0.5952	0.5590	0.5591	0.5519
	JDSH	0.7304	0.7326	0.7481	0.6439	0.6640	0.6921	0.5888	0.6510	0.6635
	CDTH	0.7315	0.7464	0.7503	0.6788	0.6815	0.6910	0.5846	0.6427	0.6573
	UCCH	0.7343	0.7342	0.7410	0.6740	0.6812	0.6945	0.6023	0.6258	0.6371
	MLCAH	0.7940	0.8050	0.8050	0.6620	0.6730	0.6870	0.5440	0.5470	0.5940
	DCHMT	0.8007	0.8021	0.8065	0.6852	0.6963	0.7009	0.6298	0.6176	0.6616
	AHLR	**0.8046**	**0.8052**	**0.8154**	**0.6952**	**0.7040**	**0.7144**	**0.6451**	**0.6557**	**0.6672**

Ablation Study

Task	Method	MIRFlickr-25K			NUS-WIDE			MS-COCO		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I→T	AHLR-M	0.8086	0.8161	0.8193	0.6753	0.6850	0.6946	0.6396	0.6516	0.6642
	AHLR-A	0.8083	0.8115	0.8179	0.6686	0.6818	0.6894	0.6447	0.6578	0.6688
	AHLR	**0.8203**	**0.8233**	**0.8266**	**0.6777**	**0.6884**	**0.6994**	**0.6454**	**0.6582**	**0.6797**
T→I	AHLR-M	0.7992	0.8005	0.8093	0.6929	0.7011	0.7132	0.6319	0.6525	0.6656
	AHLR-A	0.7988	0.7995	0.8066	0.6865	0.6974	0.7031	0.6422	0.6540	0.6669
	AHLR	**0.8046**	**0.8052**	**0.8154**	**0.6952**	**0.7040**	**0.7144**	**0.6451**	**0.6557**	**0.6672**

top-K curve



Conclusion

We propose an auxiliary hash learning (AHLR) for cross-modal retrieval methods. By introducing a dynamic mask mechanism, the key information between different modalities is automatically selected and weighted to enhance the feature representation and semantic alignment between modalities. In addition, an auxiliary hash layer is constructed to adaptively weight the features of each channel, and combined with the contrast loss function, AHLR can minimize the distance between similar samples and maximize the distance between heterogeneous samples, thereby improving the distinguishing ability of hash codes in cross-modal retrieval tasks and further improving the accuracy of retrieval. Comprehensive experiments have proved the effectiveness of this method.