Deep Normalized Cross-Modal Hashing with Bi-Direction Relation Reasoning

Changchang Sun[1], Hugo Latapie[2], Gaowen Liu[2], Yan Yan[1]
[1]Illinois Institute of Technology, [2]Cisco Research

ILLINOIS TECH

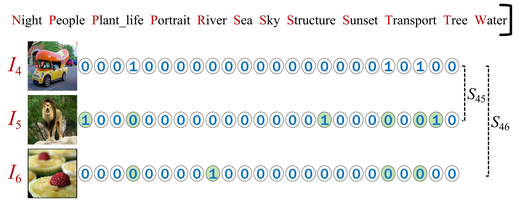
CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Introduction:

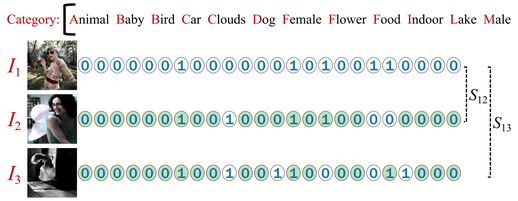
- Most of supervised cross-modal hashing methods [1] simply treat two instances similar as long as they share a common relevant category, and dissimilar otherwise.
- Although some pioneering approaches [2] have considered the multiple categories in defining the similarity between instances, they merely focus on how similar two instances are, while thoroughly overlooking the dissimilarity between them. The dissimilar information in fact delivers pivotal cues regarding the complex relation between instances.
- Existing deep hashing methods directly adopt the inner product between them as the similarity between two instances. Nevertheless, due to the distribution difference between heterogeneous modalities, this kind of measurement may hurt the model performance.

.

- **Consistent direction.**



- **Inconsistent direction.**



Bi-direction Relation Reasoning

- To better depict the complex multi-level relations between two similar instances that share at least one category, we adopt the consistent direction reasoning, i.e., the overlapped information is utilized for similarity estimation. If two instances do not share any category, we select another direction reasoning to infer their scores.
- **Consistent Direction**: Let S_{ij} denote the similarity score between instances e_i and e_j , whose label vectors are y_i and y_j . We resort to the XOR operation to obtain the overlapped label information of two instances. Formally, we have,

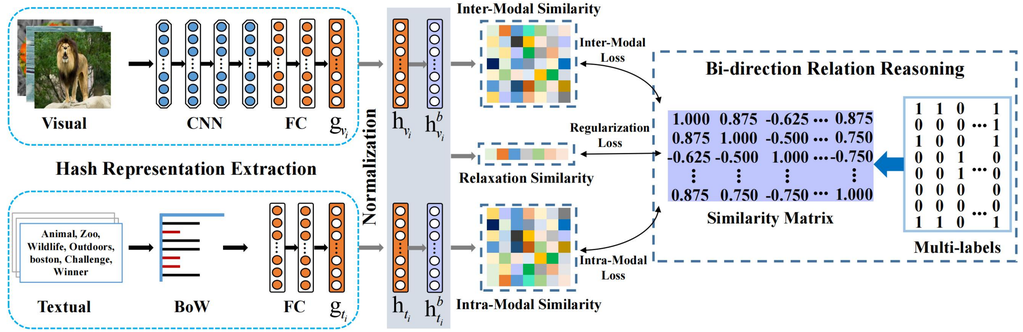
$$S_{ij} = \frac{K - (y_i \oplus y_j)}{K}$$

where K is the number of categories.

- **Inconsistent Direction**: As the irrelevant information plays a critical role in estimating the dissimilarity score, we assume that the more non-overlapped label information two instances have, the more dissimilar they are. We define the score as follows,

$$S_{pq} = -\frac{y_p \oplus y_q}{K}$$

where S_{pq} ranges from −1 to 0.



Hash Representation Normalization:

- If magnitudes of these multi-modal hash representations vary greatly, the similarity will be determined by the one with the larger magnitude, therefore causing poor retrieval performance.
- To mitigate this issue, we execute normalization on the learnt hash representations to compress them ranging from −1 to 1, while keeping the modules of them with 1.

$$\begin{cases} h_{v_i} = \dfrac{g_{v_i}}{\|g_{v_i}\|_2} \\ h_{t_j} = \dfrac{g_{t_j}}{\|g_{t_j}\|_2} \end{cases}$$

Table 1. The MAP performance comparison between our proposed model and the state-of-the-art baselines on two datasets. The CNN-F features are utilized for shallow learning models, and the best results are highlighted in bold.

Method	MIRFLICKR-25K								NUS-WIDE							
	Image→Text				Text→Image				Image→Text				Text→Image			
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
CCA	0.553	0.545	0.548	0.547	0.554	0.583	0.549	0.548	0.306	0.299	0.294	0.290	0.301	0.295	0.290	0.287
SCM-Or	0.594	0.580	0.572	0.560	0.605	0.590	0.567	0.555	0.330	0.311	0.300	0.289	0.313	0.298	0.286	0.281
SCM-Se	0.686	0.691	0.691	0.694	0.698	0.727	0.713	0.716	0.428	0.434	0.442	0.449	0.362	0.364	0.362	0.363
DCH	0.638	0.642	0.662	0.669	0.636	0.643	0.659	0.638	0.331	0.330	0.339	0.347	0.397	0.399	0.419	0.424
DCMH	0.730	0.741	0.748	0.726	0.759	0.767	0.775	0.749	0.426	0.413	0.440	0.446	0.477	0.491	0.498	0.524
SSAH	0.767	0.775	0.782	0.772	**0.767**	0.774	0.753	0.739	0.486	0.501	0.512	0.529	0.506	0.520	0.525	0.531
Bi_NCMH	**0.770**	**0.781**	**0.796**	**0.780**	0.760	**0.776**	**0.780**	**0.781**	**0.511**	**0.528**	**0.540**	**0.557**	**0.526**	**0.542**	**0.545**	**0.546**

Conclusion

- We design a novel bi-direction relation reasoning scheme to capture the complex multi-level semantic similarity relying on multi-labels. Moreover, under this supervision, hash codes could preferably maintain original similarity relations.
- We execute the feature normalization on the hash representation. It can effectively reduce the modality distribution gap and binarization penalization.
- Extensive experiments on two multimodal benchmark datasets demonstrate the superiority of our model over several state-of-the-art methods.

References

[1] Deep cross-modal hashing. Qing-Yuan Jiang and Wu-Jun Li. CVPR 2017.

[2]Self-supervised adversarial hashing networks for cross-modal retrieval. Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. CVPR 2018.