ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447517

ADAPTIVE CONFIDENCE MULTI-VIEW HASHING FOR MULTIMEDIA RETRIEVAL

Jian Zhu[1,2], *Yu Cui*[2], *Zhangmin Huang*[2], *Xingyu Li*[3], *Lei Liu*[1,*], *Lingfang Zeng*[2,*], *Li-Rong Dai*[1]

[1] University of Science and Technology of China, Hefei, China {liulei13, lrdai}@ustc.edu.cn
[2] Zhejiang Lab, Hangzhou, China {qijian.zhu, cui.yu, zmhuang, zenglf}@zhejianglab.com
[3] Lin Gang Laboratory, Shanghai, China {lixingyu0404}@lglab.ac.cn

ABSTRACT

The multi-view hash method converts heterogeneous data from multiple views into binary hash codes, which is one of the critical technologies in multimedia retrieval. However, the current methods mainly explore the complementarity among multiple views while lacking confidence in learning and fusion. Moreover, in practical application scenarios, the single-view data contains redundant noise. To conduct confidence learning and eliminate unnecessary noise, we propose a novel Adaptive Confidence Multi-View Hashing (ACMVH) method. First, a confidence network is developed to extract useful information from various single-view features and remove noise information. Furthermore, an adaptive confidence multi-view network is employed to measure the confidence of each view and then fuse multi-view features through a weighted summation. Lastly, a dilation network is designed to further enhance the feature representation of the fused features. To the best of our knowledge, we pioneer the application of confidence learning into the field of multimedia retrieval. Extensive experiments on two public datasets show that the proposed ACMVH performs better than state-of-the-art methods (maximum increase of 3.24%). The source code is available at https://github.com/HackerHyper/ACMVH.

***Index Terms*—** Multi-view Hash, Adaptive Confidence Multi-view Learning, Multi-modal Hash, Multi-view Fusion

1. INTRODUCTION

Due to the advantages of fast retrieval speed and low storage resources, hash representation learning [1–8] is widely used in the field of multimedia retrieval. Multi-view hashing utilizes the fusion of data from various views to generate a binary hash code with stronger semantic expression capabilities. How to effectively integrate multi-view data is an important research direction.

Current multi-view hashing methods suffer from the issue of untrustworthy fusion. The main reasons are detailed as follows. First, the single-view data generally contain some redundant noise features. For instance, Flexible Graph Convolutional Multi-modal Hashing (FGCMH) [7] is a GCN-based [5] multi-view hashing method. It first constructs the edges of the graph based on similarity and then the GCN aggregates features of adjacent nodes. Unfortunately, the noisy features of neighboring nodes are also introduced and aggregated to generate new features of the nodes during this procedure. Therefore, it becomes necessary to remove noise and help the multi-view hashing method achieve better performance. Second, multi-view feature fusion lacks a measure of the confidence of single-view features and the importance of measuring the confidence of each single view is underrated. To get a global representation, typical multi-view hashing methods such as Bit-aware Semantic Transformer Hashing (BSTH) [9] use a simple sum operation to fuse the multi-view features. However, the view-level confidence is ignored during the fusing process, which incurs a weak expressiveness of fused features. The facts above result in the problem of untrustworthy fusion.

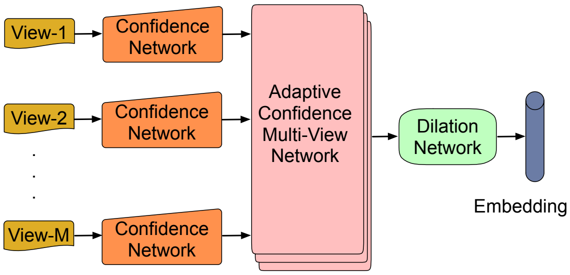

Fig. 1: Adpative Confidence Multi-View Learning. Firstly, perform confidence network on individual view features to extract useful features and suppress redundant features. Secondly, automatically learn the confidence values of each single view feature and then fuse these features by a weighted summation. Finally, a dilation network is implemented on the fused feature to generate global representation.

To eliminate the redundant noise information and realize the confidence multi-view learning [1–3], we propose a novel multi-view hashing method termed *Adaptive Confidence Multi-View Hashing* (ACMVH). As shown in Fig. 1, adaptive confidence multi-view learning aims to learn an effective representation for the multi-view hashing task. Firstly, we utilize a confidence network for the single-view feature extraction, which can sift through useful features and suppress noise features in each single view. Then, an adaptive confidence multi-view network is used to implement confidence fusion, which can automatically learn the confidence of each view. Furthermore, based on the learned confidence of each view, we can obtain the trustworthy fusion through a weighted summation. Finally, we develop a dilation network to perform on the fused representation and enhance semantic representation further.

We evaluate the proposed ACMVH method on MIR-Flickr25K and NUS-WIDE datasets in multi-view hash representation learning benchmarks. Our ACMVH yields an improvement of up to 3.24% in mean average precision (mAP), according to benchmark results. Our main contributions are summarized as follows:

- To the best of our knowledge, this paper is the first to apply the confidence learning to the multi-view retrieval tasks.
- We conduct experiments to validate the efficiency of our method and achieve SOTA results in multimedia retrieval.

[*] Corresponding author. Lei Liu (liulei13@ustc.edu.cn) and Lingfang Zeng (zenglf@zhejianglab.com).

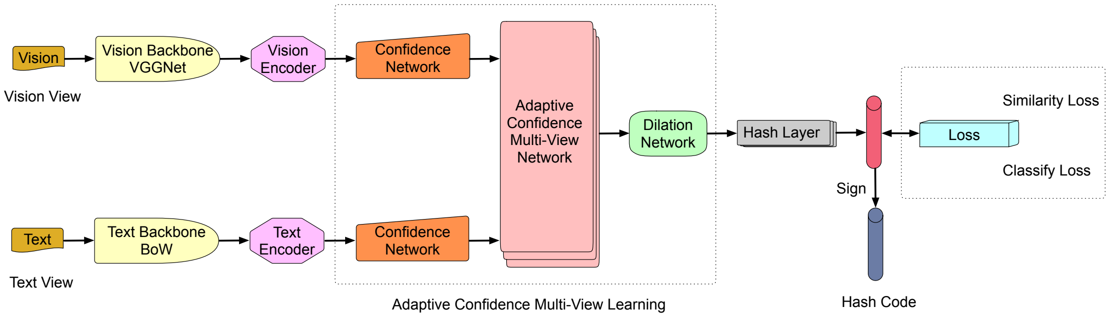

Fig. 2: The flow chart of ACMVH method. The vision and text features are extracted by backbones respectively. Each single view feature needs to be mined for useful information through the confidence network. Then, view-level adaptive confidence learning is performed, and multiple view features are adaptively fused. Subsequently, the dilation network is performed on fused features to enhance the semantic representation. Finally, the hash layer outputs the binary hash codes based on the enhanced semantic representation.

2. THE PROPOSED METHODOLOGY

We propose adaptive confidence multi-view learning (ACMVL) to credibly fuse multi-view features and apply ACMVL to the field of multimedia retrieval. In this section, we detail the neural structure of our method and its objective.

2.1. Deep Multi-view Hashing Network

The deep multi-view hashing network transforms multi-view data into binary hash code. As shown in Fig. 2, ACMVH consists of (1) backbones, (2) confidence networks, (3) an adaptive confidence multi-view network, (4) a dilation network, and (5) a hash layer.

2.1.1. Backbones

Let the training dataset be $\mathcal{X} = \left\{\{x_i\}_{i=1}^{N}, Y\right\}$, where $x_i \in \mathbb{R}^D$ is a multi-view instance, N represents the number of samples. $Y = \{y_1, y_2, \ldots, y_N\}$ is a sequence set, where y_i denotes the category information of x_i. We set $x_i = \{x_i^1, x_i^2, \ldots, x_i^M\}$ and M is the number of views. Assume that

$$Z_i^m = Backbone^m(x_i^m), \tag{1}$$

x_i^m represents the original data of m-th view. The m-th view data has a backbone network responsible for its respective feature Z_i^m. In our experiments, we utilize VGGNet [10] for vision feature extraction and Bag-of-Words model [11] for text feature extraction.

Then, we utilize a two-layer fully connected network as an encoder module. First, it can represent each view feature at a high level. Next, each view feature is normalized to the same dimension and threshold. Let

$$E_i^m = Encoder^m(Z_i^m), \tag{2}$$

where $E_i^m \in \mathbb{R}^d$ represents the extracted features of the sample through the encoder module in the m-th view and d denotes the embedding dimension of the m-th view.

2.1.2. Confidence Networks

To reduce the influence of noise features, we propose a confidence network for each view to extract useful features and eliminate noise features, which improves feature confidence in each view. Let

$$w_i^m = \sigma(w_c E_i^m + b_c), \tag{3}$$

where σ refers to the sigmoid activation function, $w_c \in \mathbb{R}^{d\times d}$ and $b_c \in \mathbb{R}^d$ are trainable parameters. The vector of weights $w_i^m \in [0,1]^d$ represents a set of learned gates applied to the individual dimensions of the encoded feature E_i^m. With the learned weight vector w_i^m, the filtered features are obtained by the element-wise production between the encoded feature E_i^m and the weight vector w_i^m for each sample in each view as:

$$C_i^m = w_i^m \odot E_i^m. \tag{4}$$

To recap, the confidence network transforms the backbone feature Z_i^m into a new representation C_i^m.

2.1.3. Adaptive Confidence Multi-View Network

The importance of individual views varies in multimedia retrieval tasks. We learn the confidence value of each view as:

$$A_i = \sum_{m=1}^{M} (p^m * C_i^m), \tag{5}$$

where p^m represents the confidence weight of m-th view. In fact, p^m is also a part of the neural network parameters, therefore, the optimal p^m can be obtained through training. Further, A_i is the result of the confidence fusion of multiple view features.

2.1.4. Dilation Network

Lastly, a dilation network structure is developed for the semantic enhancement of fused features. This module first increases the dimension of the fused features and then reduces them to the original dimension. More precisely, the specific structure consists of two layers: U_i and G_i. First, U_i is defined as follows:

$$U_i = ReLU(w_{u1} A_i + b_{u1}), \tag{6}$$

where $ReLU$ refers to the ReLU activation function, $w_{u1} \in \mathbb{R}^{d\times 4d}$ and $b_{u1} \in \mathbb{R}^{4d}$ are deep network parameters. Then, we define G_i by

$$G_i = w_{u2} U_i + b_{u2} + A_i \tag{7}$$

$w_{u2} \in \mathbb{R}^{4d\times d}$ and $b_{u2} \in \mathbb{R}^d$ are trainable parameters. Notice that, G_i is the final global representation of multi-view features.

Table 1: General statistics of two 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	4096-D	1386-D
NUS-WIDE	21000	193749	2085	21	4096-D	1000-D

2.1.5. *Hash Layer*

A linear layer with a tanh activation is hired as the hash layer, which can be represented as:

$$h_i = \tanh(w_h G_i + b_h), h_i \in \mathbb{R}^{1\times k}, \tag{8}$$

$$b_i = \operatorname{sign}(h_i), b_i \in \{-1, 1\}^{1\times k}, \tag{9}$$

where $sign$ is the signum function, $w_h \in \mathbb{R}^{d\times k}$ and $b_h \in \mathbb{R}^k$ are network parameters. k indicates that the hash layer generates k-bit hash code.

2.2. Loss Functions

The loss function shown below is used to learn the hash codes while taking the similarity metric between samples into account:

$$L_{sim} = \|\cos(h_i, h_j) - \phi_{ij}\|_2^2, \tag{10}$$

where ϕ is the affinity matrix, which can model the relation between relevant samples. ϕ_{ij} is calculated as follows:

$$\phi_{ij} = \frac{2}{1 + e^{-y_i y_j^T}} - 1. \tag{11}$$

Notice that, the category information is not completely utilized even if pairwise category information is used to train the hash function in Eq.(10). We believe that the learned binary codes should be suitable for categorization. To describe the connection between the learned binary codes and the category information, we utilize a simple linear classifier. The classifier loss function can be formulated as:

$$L_{clf} = \left\|y'_i - y_i\right\|_2^2, \tag{12}$$

where

$$y'_i = Linear(h_i), \tag{13}$$

is the predicted value of the linear classifier and the squared L2 norm is used as the loss for classification.

We can derive the total loss function as

$$L_{total} = L_{sim} + \mu L_{clf}, \tag{14}$$

where μ is the hyper-parameter obtained through grid search in our work.

3. EXPERIMENTS

3.1. Evaluation Datasets and Metrics

We evaluate the proposed ACMVH method on multimedia retrieval tasks in experiments. Two public datasets are selected: MIR-Flickr25K [12] and NUS-WIDE [13]. These datasets are widely used for evaluating multimedia retrieval performance. We use the mean Average Precision (mAP) as the evaluation metric. The details of two datasets used in experiments are summarized in Table 1.

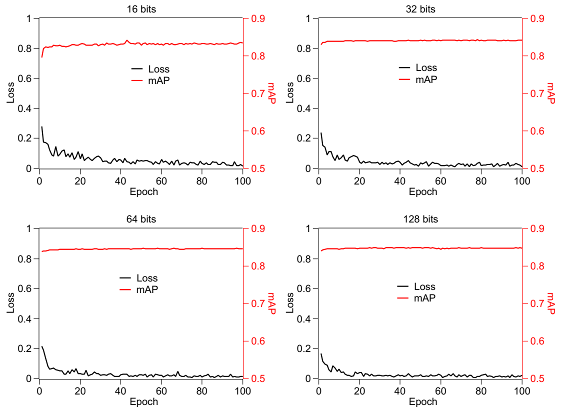

Fig. 3: The training loss and test mAP curves on MIR-Flickr25K dataset.

3.2. Baseline

To evaluate the retrieval metric, we compare the proposed ACMVH method with six multi-view hashing methods (e.g., Flexible Discrete Multi-view Hashing (FDMH) [14], Flexible Online Multi-modal Hashing (FOMH) [15], Deep Collaborative Multi-View Hashing (DCMVH) [16], Supervised Adaptive Partial Multi-view Hashing (SAPMH) [17], Flexible Graph Convolutional Multi-modal Hashing (FGCMH) [7], and Bit-aware Semantic Transformer Hashing (BSTH) [9]).

3.3. Analysis of Experimental Results

The experimental comparisons of all methods are conducted according to the unified conditions of the train set, the retrieval set, and the query set in Table 1. All multi-view hashing methods use the same backbone networks to extract visual and textual features.

The mAP result is shown in Table 2. The results show that the proposed ACMVH is overall better than all the compared multi-view hashing methods by a large margin. For example, compared with the current state-of-the-art multi-view hashing method Bit-aware Semantic Transformer Hashing (BSTH) [9], the average mAP score of our method has increased by 2.12%, and 2.05% on MIR-Flickr25K and NUS-WIDE, respectively. The main reasons for these superior results come from three aspects:

- The confidence network can extract useful features of a single view effectively and suppress noise features.
- Adaptive confidence multi-view network could credibly fuse the multi-view features into a global representation.
- Dilation network enhances the semantic representation of fused multiple view embedding.

Adaptive confidence multi-view learning promotes the discriminative capability of hash codes.

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

Method	Ref.	MIR-Flickr25K*				NUS-WIDE*			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
FOMH	MM19	0.7557	0.7632	0.7564	0.7705	0.6329	0.6456	0.6678	0.6791
FDMH	NPL20	0.7802	0.7963	0.8094	0.8181	0.6575	0.6665	0.6712	0.6823
DCMVH	TIP20	0.8097	0.8279	0.8354	0.8467	0.6509	0.6625	0.6905	0.7023
SAPMH	TMM21	0.7657	0.8098	0.8188	0.8191	0.6503	0.6703	0.6898	0.6901
FGCMH	MM21	0.8173	0.8358	0.8377	0.8606	0.6677	0.6874	0.6936	0.7011
BSTH	SIGIR22	0.8145	0.8340	0.8482	0.8571	0.6990	0.7340	0.7505	0.7704
ACMVH	Proposed	**0.8424**	**0.8573**	**0.8692**	**0.8740**	**0.7314**	**0.7562**	**0.7719**	**0.7762**

Table 3: Ablation Experiments On Two Datasets. Effects of Adaptive Confidence Multi-View Hash Architecture.

Methods	MIR-Flickr25K				NUS-WIDE			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
ACMVH-text	0.6921	0.7047	0.7201	0.7252	0.5876	0.6069	0.6308	0.6474
ACMVH-vision	0.8076	0.8213	0.8356	0.8464	0.6677	0.7041	0.7252	0.7413
ACMVH-concat	0.8108	0.8235	0.8442	0.8542	0.6924	0.7235	0.7576	0.7621
ACMVH-adaptive	0.8262	0.8390	0.8548	0.8662	0.7150	0.7455	0.7613	0.7721
ACMVH-confidence	0.8109	0.8341	0.8509	0.8587	0.7007	0.7339	0.7599	0.7710
ACMVH-dilation	0.8317	0.8394	0.8583	0.8670	0.7222	0.7512	0.7653	0.7735
ACMVH	**0.8424**	**0.8573**	**0.8692**	**0.8740**	**0.7314**	**0.7562**	**0.7719**	**0.7762**

3.4. Ablation Studies

To evaluate our method component by component, we perform an ablation of the proposed ACMVH with different experiment settings and report the performance. The experiment settings are as follows:

- *ACMVH-text*: Only the text feature is used for retrieval.
- *ACMVH-vision*: Only the vision feature is used for retrieval.
- *ACMVH-concat*: Vison and text features are fused with concatenation without adaptive confidence multi-view learning.
- *ACMVH-adaptive*: The adaptive confidence multi-view network is removed.
- *ACMVH-confidence*: The confidence network is removed.
- *ACMVH-dilation*: The dilation network is removed.
- *ACMVH*: Our full method.

The comparison results are presented in Table 3. ACMVH-vision performs better than ACMVH-text in all tasks by a large margin indicating the vision features contain more useful information than text. By comparing ACMVH-concat with ACMVH-vision, we performed a basic concatenation of visual and text features to achieve a slight performance improvement. ACMVH is the full use of our adaptive confidence multi-view learning, which greatly improves the performance of mAP compared to ACMVH-concat. Based on the performance of ACMVH-adaptive, ACMVH-confidence, and ACMVH-dilation, it is evident that the confidence network holds the highest significance, followed by the adaptive confidence multi-view network, and finally, the dilation network ranks last.

3.5. Convergence Analysis

To verify the generalization performance and convergence of ACMVH, we conduct some experiments. We run hash benchmarks with varying code lengths on the MIR-Flickr25K dataset. Fig. 3 shows training loss and test mAP. As the training goes on, the loss steadily decreases. The loss is steady after 60 epochs, proving that the local minimum reaches. The mAP for the test metric rapidly rises when the experiment begins. After 40 epochs, the test mAP stays stable. Further training does not result in a deterioration of the test MAP, indicating good generalization capability. We observe similar results for different datasets

4. CONCLUSION AND FUTURE WORK

To enhance the feature representation, adaptive confidence multi-view learning (ACMVL) is developed. Under multiple experiment settings, it delivers up to 3.24% performance gain over the current state-of-the-art methods. However, we notice some issues, for instance, the performance gain is not quite significant as the length of the hash code increases. We will work on these issues to further improve the proposed method.

5. ACKNOWLEDGMENT

This work is supported in part by the Zhejiang provincial “Ten Thousand Talents Program” (2021R52007), the National Key R&D Program of China (2022YFB4500405), the Science, Technology Innovation 2030-Major Project (2021ZD0114300), Sponsored by Zhejiang Lab Open Research Project (NO.K2022QA0AB04) , Supported by “the Fundamental Research Funds for the Central Universities” and supported by the National Natural Science Foundation of China (62306291).

6. REFERENCES

[1] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 2, pp. 2551–2566, 2022.

[2] Z. Han, F. Yang, J. Huang, C. Zhang, and J. Yao, "Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 707–20 717.

[3] X. Zheng, C. Tang, Z. Wan, C. Hu, and W. Zhang, "Multi-level confidence learning for trustworthy multimodal classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, 2023, pp. 11 381–11 389.

[4] J. Zhu, P. Hu, B. Li, and Y. Zhou, "Fast metric multi-view hashing for multimedia retrieval," *Information Fusion*, vol. 103, p. 102130, 2024.

[5] M. Welling and T. N. Kipf, "Semi-supervised classification with graph convolutional networks," in *J. International Conference on Learning Representations (ICLR 2017)*, 2016.

[6] J. Zhu, X. Ruan, Y. Cheng, Z. Huang, Y. Cui, and L. Zeng, "Deep metric multi-view hashing for multimedia retrieval," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 1955–1960.

[7] 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.

[8] J. Zhu, W. Cheng, Y. Cui, C. Tang, Y. Dai, Y. Li, and L. Zeng, "Central similarity multi-view hashing for multimedia retrieval," *arXiv preprint arXiv:2308.13774*, 2023.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.