Cross-media Hash Retrieval using Multi-head Attention Network

Paper ID:510

Zhixin Li[1],*, Feng Ling[1], Canlong Zhang[1], and Huifang Ma[2]

[1]Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China

[2] College of Computer Science and Engineering,Northwest Normal University,Lanzhou 730070, China

***Corresponding Author. E-mail: lizx@gxnu.edu.cn**

INTRODUCTION

Cross media retrieval is an important part in research of cross media intelligent computing. It aims to search semantic related samples from different media instances.

However, most unsupervised methods learn hash function by keeping the correlation between modals or within modals, but they have not achieved good experimental results. In addition, the existing work mainly focuses on pairwise relation modeling.

In this paper, the cooperative learning of auxiliary matrix and hash matrix is used to model the whole relationship, so as to reduce the distance within the same modals and expand the distance between the different modals, and finally enhance the crossmodal hash learning.

First, our method needs to generate the auxiliary similarity matrix by using the original image and text features, and then use it to guide the generation hash coding matrix of the whole training sample, and conduct collaborative learning with the auxiliary similarity matrix to ensure the accuracy.

The method in this paper is called “Cross-media Hash Retrieval using Multi-head Attention Network” (UMHA), and it has the following main contributions:
Although the current supervised cross-media hashing method has high performance, its practicability is not strong. Because there is no semantic label for massive cross-media data in reality, manual labeling is time-consuming and laborious, so it needs unsupervised methods to learn from it. UMHA proposes using the auxiliary similarity matrix to combine the information of different modals. The matrix carefully integrates the original neighborhood relations from different modals, so it can capture the potential semantic similarity between instances. Because UMHA is unsupervised, it can be applied to large-scale retrieval and has useful practical significance.

METHODS

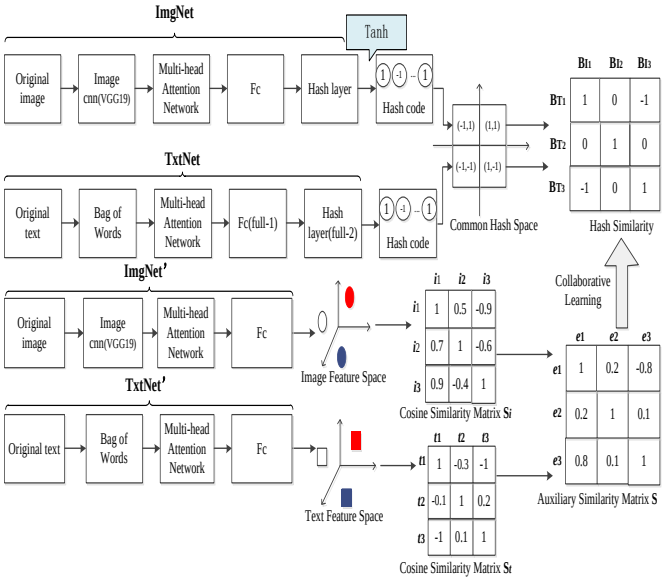

Fig. 1. Schematic diagram of the model.

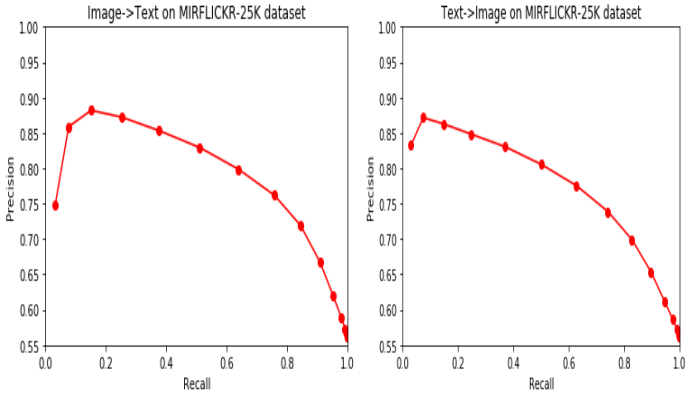

Fig. 2. Precision-recall curve on MIRFlickr-25K data set

RESULTS

TABLE II
EXPERIMENTAL RESULTS ON WIKIPEDIA DATASETS.

Method	Text to Image				Image to Text			
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
CVH [6]	25.2	23.5	17.1	15.4	17.9	16.2	15.3	14.9
IMH [13]	46.7	47.8	45.3	45.6	20.1	20.3	20.4	19.5
CMFH [2]	59.5	60.1	61.6	62.2	25.2	25.3	25.9	26.3
LSSH [23]	56.9	59.3	59.3	59.5	19.7	20.8	19.9	19.5
DBRC [3]	57.4	58.8	59.8	59.9	25.3	26.5	26.9	28.8
UDCMH [19]	**62.2**	63.3	64.5	**65.8**	30.9	31.8	32.9	34.6
UMHA(Ours)	61.4	**63.8**	**64.7**	65.2	**43.2**	**45.2**	**45.4**	**45.7**

TABLE III
RESULTS ON THE NUS-WIDE DATASET.

Method	Text to Image				Image to Text			
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
CVH [6]	47.4	44.5	41.9	39.8	45.8	43.2	41.0	39.2
PDH [12]	48.9	51.2	50.7	51.7	47.5	48.4	48.0	49.0
CMFH [2]	43.9	41.6	37.7	34.9	51.7	55.0	54.7	52.0
LSSH [23]	51.7	61.7	64.2	66.3	48.1	48.9	50.7	50.7
CCQ [11]	49.9	49.6	49.2	48.8	50.4	50.5	50.6	50.5
CMSSH [1]	51.9	49.8	45.6	48.8	51.2	47.0	47.9	46.6
MGAH [20]	60.3	61.4	64.0	64.1	61.3	62.3	62.8	63.1
UDCMH [19]	63.7	65.3	69.5	71.6	51.1	51.9	52.4	55.8
UMHA(Ours)	**76.4**	**81.4**	**81.2**	**82.8**	**77.9**	**81.6**	**83.3**	**84.1**

TABLE IV
RESULTS ON THE MIRFLICKR-25K DATASET.

Method	Text to Image				Image to Text			
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
CVH [6]	59.1	58.3	57.6	57.6	60.6	59.9	59.6	59.8
PDH [12]	62.7	62.8	62.8	62.9	62.3	62.4	62.1	62.6
IMH [13]	60.3	59.5	58.9	58.0	61.2	60.1	59.2	57.9
CMFH [2]	64.2	66.2	67.6	68.5	62.1	62.4	62.4	62.7
CCQ [11]	62.8	62.8	62.2	61.8	63.7	63.9	63.9	63.8
DBRC [3]	61.8	62.6	62.6	62.8	61.7	61.9	62.0	62.1
MGAH [20]	67.3	67.6	68.6	69.0	68.5	69.3	70.4	70.2
UDCMH [19]	69.2	70.4	71.8	73.3	68.9	69.8	71.4	71.7
UMHA(Ours)	**86.4**	**86.4**	**88.2**	**89.9**	**87.7**	**90.6**	**91.3**	**94.3**

TABLE V
THE MAP@50 RESULTS ON THE MIRFLICKR-25K DATASET TO EVALUATE THE EFFECTIVENESS OF EACH COMPONENT IN THE UMHA.

Method	Text to Image	Image to Text
	128bits	128bits
UMHA-1	83.1	86.4
UMHA-2	86.0	88.6
UMHA-3	86.2	91.3
UMHA-4	87.6	91.5
UMHA-5	87.9	92.0
UMHA	**89.9**	**94.3**

CONCLUSION

In this paper, unsupervised cross media hash retrieval based on multi-head attention network is proposed, which can be used in large-scale cross media retrieval. In general, UMHA uses the idea of deep and shallow parallel to learn the deeper hash matrix encoded by hash network and the auxiliary similarity matrix constructed by the relatively shallow image and text feature network, and learns a lot of semantic information of image and text, so that UMHA can capture the potential connection between different modals and within the same modal, making up for it differences between different modals and within the same modal. A large number of experiments have proved the superiority of the method proposed in this paper, and the effectiveness of each component has been studied carefully by ablation experiments.