ADVERSARIAL CROSS-MODAL RETRIEVAL VIA LEARNING AND TRANSFERRING SINGLE-MODAL SIMILARITIES

Anonymous ICME submission

ABSTRACT

Cross-modal retrieval aims to retrieve relevant data across different modalities (e.g., texts vs. images). The common strategy is to apply element-wise constraints between manually labeled pair-wise items to guide the generators to learn the semantic relationships between the modalities, so that the similar items can be projected close to each other in the common representation subspace. However, such constraints often fail to preserve the semantic structure between unpaired but semantically similar items (e.g. the unpaired items with the same class label are more similar than items with different labels). To address the above problem, we propose a novel cross-modal similarity transferring (CMST) method to learn and preserve the semantic relationships between unpaired items in an unsupervised way. The key idea is to learn the quantitative similarities in single-modal representation subspace, and then transfer them to the common representation subspace to establish the semantic relationships between unpaired items across modalities. Experiments show that our method outperforms the state-of-the-art approaches both in the class-based and pair-based retrieval tasks.

Index Terms— Cross-modal, retrieval

1. INTRODUCTION

Cross-modal retrieval aims to retrieve relevant data across different modalities, which enables flexible search across multiple modalities. The common strategy is to apply element-wise constraints on manually labeled cross-modal pairs to bridge the semantic gaps between different modalities. However, such manually labeled pairs can only reflect a small part of the semantic structure in the common representation subspace, while the abundant semantic relationships between unpaired items are often failed to be preserved. As demonstrated in Figure 1, only the similarity between the red car and its corresponding description are manually labeled (green solid line), as well as the similarity between the blue car and its description. However, the similarity between the red car and the blue car's description are missing (orange dot line), although they are semantically similar to some extent because of the same label (the label of car) they share.

To solve the above problem, methods like DeViSe [1], DCML [2] and ACMR [3] try to utilize the intra-class and inter-class labels of the cross-modal items [4, 3] by directly assigning the highest similarity (e.g., the value of 1) to the items with the same class label and the lowest similarity (e.g., the value of 0) to the items with different class labels. The problem is, these labels cannot quantitatively reflect the semantic relationships between the intra-class items, which is especially important for the retrieval tasks, because the rank list should indicate the discriminative order of all the retrieved items, especially for the items in the same class. The samples that matches the query better should be ranked higher compared to other samples even in the same class. On the other hand, using the intra-class and inter-class labels as the supervision information also makes the cross-modal retrieval methods sensitive to dataset noises such as mislabeled samples.

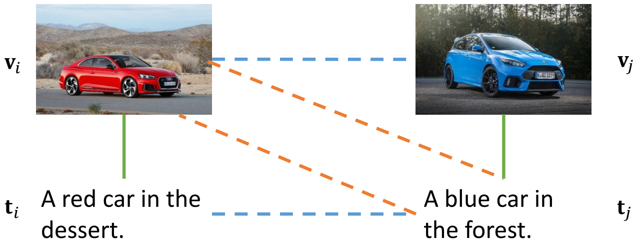

Fig. 1: Demonstration of the missing semantic relationships. The green solid line is the manual label of paired items cross modalities, and the dot line is the missing label for inter-modal items (orange) and inner-modal items (blue). The proposed CMST considers learning the missing inner-modal similarity (blue dot line) and transferring it to the inter-modal similarity (orange dot line) based on paired items (green solid line).

In this paper, to address the above-mentioned issues, a novel cross-modal similarity transferring (CMST) method is proposed for cross-modal retrieval. The main idea is to employ unsupervised strategy to learn the endogenous semantic relationships between unpaired items in each single-modal representation subspace, and then, transfer the learned relationships to the common representation subspace to establish the semantic structure between unpaired cross-modal items. In detail, the CMST first employs a similarity learning network to establish the finer similarity metric to capture the semantic structure of training items in each single-modal space. Then, three similarity transferring approaches are proposed to transfer the learnt single-modal similarities to the common representation subspace. CMST works in an adversarial framework to utilize the ability of distribution generation from generative adversarial networks. Our main contributions can be summarized as follows.

- A novel CMST method is proposed to learn and preserve the semantic structure between unpaired items across different modalities in the common representation subspace in an unsupervised way.
- Three similarity transferring approaches are proposed by the observation on how cross-modal relationship is built from the similarities in single modalities, which are proven effective in the experiments.

2. RELATED WORK

Cross-modal retrieval methods can be roughly divided into *joint representation learning methods* [3] and *cross-modal hashing methods* [5, 6]. The proposed CMST model falls into the category of joint representation based methods. It aims to learn a real common representation subspace of multimodal data, where cross-modal data can be directly compared to each other through predefined similarity measurement. Cross-modal retrieval methods like CCA-based methods [7, 8], LDA-based methods [9] and neural network based methods [1, 2, 3] also fall into this category. On the other hand, the cross-modal hashing methods mainly focus on the retrieval efficiency by mapping the items of different modalities into a common binary Hamming space.

Benefited from the strong ability of distribution modeling and discriminative representation learning, some recent cross-modal retrieval methods have collaborated with GAN models [10, 11, 3]. In this work, our method also follows the similar adversarial learning framework that uses the single-modal similarities to guide the cross-modal representation learning.

More recently, methods to explore the semantic relationships between unpaired items have been proposed. The ACMR [3] method proposes the triplet loss and the modality classifier for preserving the modality level semantic structures. The MHTN [11] is proposed to minimize the maximum mean discrepancy between modalities, which preserves more flexibility for the generator to project vectors into a new space. The difference between CMST and the previous work is that CMST can learn the item-level semantic relationships between unpaired items in an unsupervised way.

3. CMST-BASED CROSS-MODAL RETRIEVAL

3.1. Problem Formulation

Without losing generality, we consider the images and texts pairs in this paper. Let $V = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n] \in \mathbb{R}^{d_v \times n}$ be a collection of image features, and $T = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n] \in \mathbb{R}^{d_t \times n}$ be the corresponding collection of text features, in which $\mathbf{v}_i$ and $\mathbf{t}_i$ form a pair, where d_v and d_t denote the dimension of the image features and the text features, respectively. Each sample pair is assigned a semantic label vector denoted as $\mathbf{y_i} = [y_{i1}, y_{i2}, \dots, y_{ic}] \in \mathbb{R}^c$, where c indicates the semantic classes. If the i-th sample pair in V and T belongs to the semantic class j, $y_{ij} = 1$; otherwise, $y_{ij} = 0$. We denote the collection of semantic label vectors as $Y = [\mathbf{y_1}, \mathbf{y_2}, \dots, \mathbf{y_n}] \in \mathbb{R}^{d_c \times n}$. The goal of our proposed CMST method is to learn a common semantic space $S = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n] \in \mathbb{R}^{d_s \times n}$, where the features from different modalities can be directly compared in terms of semantic similarity, and d_s denotes the dimension of the common semantic space.

3.2. Learning the Single-modal Similarity

The Siamese networks [12] are adopted as similarity learning networks to learn the semantic relationships in single-modal representation subspace. Siamese networks can learn similarity metrics discriminatively and effectively. In addition, Siamese networks are also robust to data noises because they consider not only the label relationship but also the distance between the features of each sample pair. For simplicity, we only detail the network in image modality. Given two image features and their labels $(\mathbf{v}_i, \mathbf{y}_i)$ and $(\mathbf{v}_j, \mathbf{y}_j)$, let $u_{ij} = 1$ if $\mathbf{y}_i = \mathbf{y}_j$, otherwise $u_{ij} = 0$. The loss function of our similarity learning network can be formulated as:

$$\mathcal{L}_{sia} = \sum_{i,j} u_{ij} \cdot s(\mathbf{v}_i, \mathbf{v}_j) + (1 - u_{ij}) \cdot \max(C - s(\mathbf{v}_i, \mathbf{v}_j), 0). \quad (1)$$

The similarity measurement in single-modal representation subspace is given as

$$s(\mathbf{v}_i, \mathbf{v}_j) = ||H_V(\mathbf{v}_i) - H_V(\mathbf{v}_j)||_2^2, \quad (2)$$

where the H_V denotes the image similarity learning network consists of feed-forward layers with ReLU activations. The text similarity learning network has the same structure as the image similarity learning network. The Siamese network is illustrated in the middle bottom part of Figure 1.

3.3. Transferring Learned Similarities to Common Representation Subspace

3.3.1. Value transferring

A direct way of using the intra-modal similarity as a reference for cross-modal similarity learning is to use the value as the learning goal. In this case, we use $\mathbf{t}_j$ as the anchor sample, only the intra-modal similarity that is different from it (image-modality similarity) is used for clear explanation, but the training process contains both directions. Also, the range

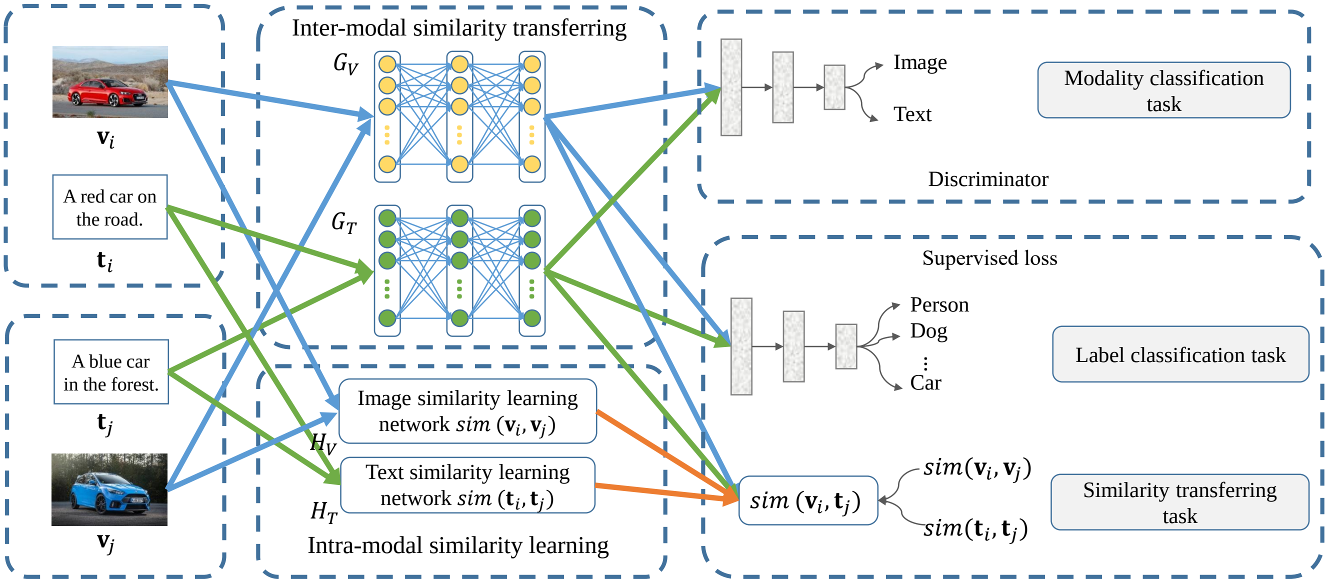
Inter-modal similarity transferring
G_V
G_T
$\mathbf{v}_i$
A red car on the road.
$\mathbf{t}_i$
A blue car in the forest.
$\mathbf{t}_j$
$\mathbf{v}_j$
Image
Text
Modality classification task
Discriminator
Supervised loss
Person
Dog
Car
Label classification task
Image similarity learning network $sim\ (\mathbf{v}_i, \mathbf{v}_j)$
H_V
Text similarity learning network $sim\ (\mathbf{t}_i, \mathbf{t}_j)$
H_T
Intra-modal similarity learning
$sim\ (\mathbf{v}_i, \mathbf{t}_j)$
$sim(\mathbf{v}_i, \mathbf{v}_j)$
$sim(\mathbf{t}_i, \mathbf{t}_j)$
Similarity transferring task

Fig. 2: The overall structure of the CMST method, including two intra-modal similarity learning networks and an inter-modal similarity transferring network. This figure shows the procedure to learn the cross-modal similarity between $\mathbf{v}_i$ and $\mathbf{t}_j$. Different from the existing methods which simply assign the highest similarity to intra-class cross-modal samples, the CMST method uses the similarity value between $\mathbf{v}_i$ and $\mathbf{v}_j$ to guide cross-modal similarity learning, which is learnt by the intra-modal similarity learning network in the image modality.

of the original features and the projected features are not limited, as the fully connected networks have the ability of scaling between the input and the expected output. The typical form of similarity transferring loss based on value transferring is defined as

$$\mathcal{L}_{val} = |s(\mathbf{v}_j, \mathbf{v}_j) - s(\mathbf{v}_j, \mathbf{t}_j)| + |s(\mathbf{v}_i, \mathbf{v}_j) - s(\mathbf{v}_i, \mathbf{t}_j)|, \quad (3)$$

where the similarity between $\mathbf{v}_j$ and itself is set to value 1.

3.3.2. *Difference transferring*

Compared to the absolute value of intra-modal similarity itself, it is the relationship between cross-modal samples that we are really interested in. In other words, the relation of $(s(\mathbf{v}_i, \mathbf{t}_j) - s(\mathbf{v}_j, \mathbf{t}_j))$ can measure the difference of $\mathbf{v}_i$ to $\mathbf{t}_j$ and $\mathbf{v}_j$ to $\mathbf{t}_j$ in the common semantic space, where the difference is expected to be related to the difference of $\mathbf{v}_i$ and $\mathbf{v}_j$ in their original image space. By difference transferring, the range of the common semantic space and the original modality space can be decoupled. The learnt joint subspace can be more flexible to modality divergence. The typical form of similarity transferring loss based on difference transferring is defined as

$$\mathcal{L}_{diff} = |(s(\mathbf{v}_j, \mathbf{t}_j) - s(\mathbf{v}_i, \mathbf{t}_j)) - (s(\mathbf{v}_j, \mathbf{v}_j) - s(\mathbf{v}_i, \mathbf{v}_j))|, \quad (4)$$

where we assume $s(\mathbf{v}_j, \mathbf{v}_j) = 1$.

3.3.3. *Product transferring*

As transitivity exists in the nature of similarity measurement, similarity transitivity across modalities can be expected. Following this motivation, product transferring can be done by the multiplication on the similarity chain. We can say that the similarity of $\mathbf{v}_i$ and $\mathbf{t}_j$ is generated by the chain of $\mathbf{v}_i$ to $\mathbf{v}_j$ and then $\mathbf{v}_j$ to $\mathbf{t}_j$. In this case, the typical form of similarity transferring loss based on product transferring is defined as

$$\mathcal{L}_{prod} = |s(\mathbf{v}_i, \mathbf{t}_j) - s(\mathbf{v}_i, \mathbf{v}_j)s(\mathbf{v}_j, \mathbf{t}_j)|, \quad (5)$$

where some implicit linear function is assumed to enable the direct arithmetical operation between inter-modal and intra-modal similarities. The linear function is absorbed into the function approximation ability of neural networks.

The proposed similarity transferring approaches provide a reference value or reference relationship for the unpaired items of two different modalities, so the loss function can be defined as the difference between the similarity calculated from learnt cross-modal features and the reference values. The similarity transferring task is applied together with other widely used tasks within a generative adversarial framework.

3.4. Total Loss for Training

Follow the common practice of cross-modal retrieval, the proposed CMST introduces the adversarial learning and classification task for learning a better semantic structure in common representation subspace.

Table 1: Comparison with existing cross-modal retrieval methods in aspect of mAP

Category	Methods	Wikipedia			Pascal Sentences			NUS-WIDE-10k		
		Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.
Traditional methods	CCA-3V [8]	0.437	0.383	0.410	0.316	0.270	0.293	-	-	-
	LCFS [13]	0.455	0.398	0.427	0.442	0.357	0.400	0.383	0.346	0.365
	JRL [14]	0.453	0.400	0.426	0.504	0.489	0.496	0.426	0.376	0.401
	JFSSL [15]	0.428	0.396	0.412	-	-	-	-	-	-
DNN-based methods	Corr-AE [16]	0.402	0.395	0.398	0.489	0.444	0.467	0.366	0.417	0.392
	DCML [2]	0.554	0.538	0.546	-	-	-	0.385	0.405	0.395
	CMDN [17]	0.488	0.427	0.458	0.534	0.534	0.534	0.492	0.515	0.504
GAN-based methods	CM-GAN [10]	0.521	0.466	0.494	0.603	0.604	0.604	-	-	-
	MHTN [11]	0.514	0.444	0.479	0.496	0.500	0.498	0.520	0.534	0.527
	ACMR [3]	0.619	0.489	0.554	0.535	0.543	0.539	0.544	0.538	0.541
	CMST (Ours)	**0.632**	**0.505**	**0.569**	**0.621**	**0.586**	**0.604**	**0.628**	**0.562**	**0.595**

Let $\mathcal{L}_{sim}$ denote the similarity transferring task, which is chosen as the one from $\mathcal{L}_{val}$, $\mathcal{L}_{diff}$ and $\mathcal{L}_{prod}$. For classification task, the cross entropy loss is used in CMST and denoted as $\mathcal{L}_{lab}$. For adversarial learning, the discriminator composed of 3 feed-forward layers takes the generated representation as input. The output is the prediction of which modality the input comes from, using the sigmoid activation. Let $\mathcal{L}_V$ denotes the cross entropy loss of predicting the image input, and $\mathcal{L}_T$ denotes the cross entropy loss of predicting the text input. The total loss for generator and the discriminator is given as

$$\mathcal{L}_G = \mathcal{L}_{lab} + \mathcal{L}_{sim} + \mathcal{L}_V - \mathcal{L}_T, \tag{6}$$

$$\mathcal{L}_D = -\mathcal{L}_V + \mathcal{L}_T \tag{7}$$

4. EXPERIMENTS

4.1. Experimental Setup

4.1.1. Datasets and pre-processing

Three widely used datasets for cross-modal retrieval are used in the experiments, including Wikipedia dataset [18], NUS-WIDE-10k dataset [19] and Pascal Sentence dataset [20]. We follow the dataset partition method as [21, 3] for fair comparison. Image features are taken from the fc7 layer of a pre-trained VGGNet-19 model while text features are computed by the classic Bag-of-Words features with tf-idf weighting. The image features for all the datasets are of 4096 dimensions, while the text BoW feature is 5000 dimensions for the Wikipedia dataset, 1000 dimensions for the NUS-WIDE-10k and Pascal datasets.

4.1.2. Evaluation metrics

For the class-based retrieval task, performance is measured in terms of the mean average precision (mAP), the measurement is applied on both directions, i.e. Img2txt and Txt2img. The larger the mAP value is, the better the performance becomes.

For the pair-based retrieval task, we accept only the ground-truth paired sample of the input query as correct retrieval result. The performance is evaluated by top-k accuracy, indicating the times that the correct retrieval result appears within the top-k retrieved results over the test set.

4.2. Results and Analysis

The CMST is compared with three classes of cross-modal retrieval methods, namely traditional methods, DNN-based methods and GAN-based methods, as shown in Table 1. The results of CMST is based on difference transferring.

For the results shown in the table, the number of the intra-class samples with the query are counted in top-50 retrieved documents. Results show that our CMST method outperforms the counterparts on all three datasets for cross-modal retrieval tasks. On the Pascal dataset, our method improves the best competitor ACMR by 16.1% and 7.9% in image to text and text to image retrieval tasks, respectively. On NUS-WIDE-10k and Wikipedia dataset, the proposed CMST method achieves a relatively small but solid improvement compared to the state-of-the-art performance. The underlying cause why the method improves a lot on Pascal dataset but not as much on the other two datasets is that the Pascal dataset contains more semantic classes than the other two datasets. On Wikipedia and NUS-WIDE-10k, although the intra-class and inter-class labels are coarse supervision for cross-modal similarity learning, the small number of total classes makes it easy to separate between classes and arrange the joint semantic distribution. For the datasets with more classes, such as Pascal Sentences, the finer similarity structure benefits the generation of common semantic space as we expected.

Table 2 shows the experimental results in the pair-based retrieval task in aspect of top-k accuracy on Pascal dataset, where different k values are tested to examine the method's performance. For smaller values of k (k=1,5,10), our proposed method outperforms the ACMR method by 43.4%, 18.9% and 10.1% for the average measurement on two re-

Table 2: Comparison with ACMR method on Pascal Sentences dataset in aspect of top-k acc

Methods	$k = 1$			$k = 5$			$k = 10$			$k = 50$		
	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.	Img2txt	Txt2Img	Avg.
ACMR	0.140	0.145	0.143	0.445	0.400	0.423	0.675	0.665	0.670	0.910	0.915	0.913
CMST(Ours)	**0.210**	**0.200**	**0.205**	**0.455**	**0.550**	**0.503**	**0.725**	**0.750**	**0.738**	**0.990**	**0.965**	**0.978**

trieval directions. The results indicate that our proposed CMST method works effectively on improving the rank of the most related retrieval results for the given query.

4.3. Further Analysis on CMST

In this subsection, several experiments are conducted to investigate the effectiveness of the important components of CMST method.

Table 3: Similarity transferring with different intra-modal similarity measurements

Methods	mAP		top-1 acc	
	Img2txt	Txt2Img	Img2txt	Txt2Img
Cosine	0.509	0.500	0.150	0.140
Euclidean	0.563	0.554	0.165	0.125
Siamese Network	**0.621**	**0.586**	**0.210**	**0.200**

The intra-modal similarity learnt by the Siamese networks plays an important role in the subsequent similarity transferring procedure. To examine its effectiveness, two traditional similarity measurements are employed as the source of intra-modal similarity for comparison. Table 3 illustrates the performance of traditional similarity metric and the learnt Siamese similarity metric, showing that the Siamese network achieves the best performance among the three similarity measurements.

Table 4: Effects of the similarity transferring approaches

Methods	mAP		top-1 acc	
	Img2txt	Txt2Img	Img2txt	Txt2Img
Value	0.600	0.564	0.145	0.155
Difference	**0.621**	**0.586**	**0.210**	**0.200**
Product	0.510	0.514	0.120	0.140
No transfer	0.485	0.508	0.080	0.105

In order to compare the different methods of similarity transferring, we examine the three approaches on Pascal dataset. The results in the last row in Table 4 with no transferring indicates that the similarity transferring task is not included in the training. Overall, we can see that all the similarity transferring approaches are effective for noticeable performance improvement with similarity transferring training. The best similarity transferring approach contributes an improvement of 21.6% compared to the CMST without similarity transferring. The difference transferring approach performs better than the value transferring method because it keeps the comparative relationship between samples instead of assigning a value to the similarity between samples. The generator is guaranteed with more flexibility for the feature projection by difference transferring method. The product transferring seems to perform badly, as the distance measurement used in the experiment is the Euclidean distance and there is no explicit limitations or normalization methods added to the distance calculation. The product of the Euclidean distances between high dimensional feature vectors varies sharply on minor changes.

Table 5: Effects of different training strategies.

Strategies	mAP		top-1 acc	
	Img2txt	Txt2Img	Img2txt	Txt2Img
Two-stage	**0.621**	**0.586**	**0.210**	**0.200**
Fine-tuning	0.596	0.571	0.180	0.180
End-to-end	0.585	0.561	0.140	0.165

A two-stage training strategy is employed by firstly learning the single-modal similarity metric and then transferring the single-modal similarity into cross-modal semantic space. Note that all previous results are based on the two-stage strategy. On the other hand, fine-tuning and end-to-end training strategies are also available for our proposed CMST method. To examine the influence of different training strategies, we conduct an additional experiment by using different training method and compare the testing results at the same epoch (100). In the two-stage training and fine-tuning training, the Siamese networks are trained for 50 epoches in advance. For two-stage training, the parameters of Siamese networks are fixed after the 50th epoch, while for fine-tuning, the parameters are still learnable with a low learning rate (0.0001). Table 5 shows the effect of different training method, and we can draw from the table that two-stage training yields the best performance. Fine-tuning does not provide additional performance improvement. End-to-end training has negative influence on cross-modal similarity learning because the single-modal similarity learning should not be heavily affected by cross-modal information.

5. CONCLUSIONS

In this paper, a novel cross-modal retrieval method named CMST is proposed. The proposed method efficiently learns similarity metrics in each single modality space by the intra-modal similarity learning networks and then guide the cross-modal similarity learning with the learnt single modal similarity metric. Our proposed similarity transferring approaches successfully transfer finer similarity structure captured in single modal space to cross-modal space. Experiments demonstrate that the CMST method outperforms state-of-the-art cross-modal retrieval methods in both class-based and pair-based retrieval tasks.

6. REFERENCES

[1] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al., "Devise: A deep visual-semantic embedding model," in *NIPS*, 2013, pp. 2121–2129.

[2] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1234–1244, 2017.

[3] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen, "Adversarial cross-modal retrieval," in *ACM MM*. ACM, 2017, pp. 154–162.

[4] Alexis Mignon and Frédéric Jurie, "Cmml: A new metric learning approach for cross modal matching," in *ACCV*, 2012.

[5] Jingkuan Song, Tao He, Lianli Gao, Xing Xu, Alan Hanjalic, and Heng Tao Shen, "Binary adversarial networks for image retrieval," in *AAAI*, 2018.

[6] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *TIP*, vol. 26, no. 5, pp. 2494–2507, 2017.

[7] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. 1247–1255.

[8] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.

[9] Duangmanee Putthividhy, Hagai T Attias, and Srikantan S Nagarajan, "Topic regression multi-modal latent dirichlet allocation for image annotation," in *CCVPR*. IEEE, 2010, pp. 3408–3415.

[10] Yuxin Peng, Jinwei Qi, and Yuxin Yuan, "Cmgans: Cross-modal generative adversarial networks for common representation learning," *arXiv preprint arXiv:1710.05106*, 2017.

[11] Xin Huang, Yuxin Peng, and Mingkuan Yuan, "Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval," *arXiv preprint arXiv:1708.04308*, 2017.

[12] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*. IEEE, 2005, pp. 539–546.

[13] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan, "Learning coupled feature spaces for cross-modal matching," in *ICCV*. IEEE, 2013, pp. 2088–2095.

[14] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 965–978, 2014.

[15] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *TPAMI*, vol. 38, no. 10, pp. 2010–2023, 2016.

[16] Fangxiang Feng, Xiaojie Wang, and Ruifan Li, "Cross-modal retrieval with correspondence autoencoder," in *ACM MM*. ACM, 2014, pp. 7–16.

[17] Yuxin Peng, Xin Huang, and Jinwei Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," in *IJCAI*, 2016, pp. 3846–3853.

[18] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *TPAMI*, vol. 36, no. 6, pp. 521–535, 2014.

[19] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao 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*. ACM, 2009.

[20] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *NAACL HLT*. ACL, 2010, pp. 139–147.

[21] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan, "Ccl: Cross-modal correlation learning with multi-grained fusion by hierarchical network," *IEEE Transactions on Multimedia*, vol. 20, pp. 405–420, Aug. 2017.