FedCMR: Federated Cross-Modal Retrieval

Linlin Zong*
llzong@dlut.edu.cn
School of Software, Dalian University of Technology
Dalian, Liaoning, China

Qiujie Xie*
qiujie_xie@126.com
School of Software, Dalian University of Technology
Dalian, Liaoning, China

Jiahui Zhou
zjhjixiang@mail.dlut.edu.cn
School of Software, Dalian University of Technology
Dalian, Liaoning, China

Peiran Wu
1294980047@mail.dlut.edu.cn.com
School of Software, Dalian University of Technology
Dalian, Liaoning, China

Xianchao Zhang†
xczhang@dlut.edu.cn
School of Software, Dalian University of Technology
Dalian, Liaoning, China

Bo Xu†
xubo@dlut.edu.cn
School of Computer Science and Technology, Dalian University of Technology
Dalian, Liaoning, China

ABSTRACT

Deep cross-modal retrieval methods have shown their competitiveness among different cross-modal retrieval algorithms. Generally, these methods require a large amount of training data. However, aggregating large amounts of data will incur huge privacy risks and high maintenance costs. Inspired by the recent success of federated learning, we propose the federated cross-modal retrieval (FedCMR), which learns the model with decentralized multi-modal data. Specifically, we first train the cross-modal retrieval model and learn the common space across multiple modalities in each client using its local data. Then, we jointly learn the common subspace of multiple clients on the trusted central server. Finally, each client updates the common subspace of the local model based on the aggregated common subspace on the server, so that all clients participated in the training can benefit from federated learning. Experiment results on four benchmark datasets demonstrate the effectiveness proposed method.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**.

KEYWORDS

Cross-modal retrieval, multi-modal learning, federated learning

ACM Reference Format:
Linlin Zong, Qiujie Xie, Jiahui Zhou, Peiran Wu, Xianchao Zhang, and Bo Xu. 2021. FedCMR: Federated Cross-Modal Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3462989

*Indicates equal contribution
†Corresponding Author

SIGIR '21, July 11–15, 2021, Virtual Event, Canada.

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462989

1 INTRODUCTION

Cross-modal retrieval [22], which takes one type of data as the query to retrieve relevant data of another type, has been a hot topic since the past decade. The existing works can be roughly divided into two categories: the traditional learning methods [5, 10] and the deep learning methods [1, 16, 23, 25]. On account of the amazing performance for feature extraction tasks, deep learning based cross-modal retrieval methods receives much attention in recent years, which usually learn modality-specific transformations to project the data samples from different modalities into a common subspace [1, 16, 23–25]. However, the deep learning methods rely on a large amount of high-quality multi-modal data.

In reality, the multi-modal data are generally scattered in various institutions, and a client can only grasp a small amount of the data. Due to the limitation of network privacy protection and data security management [17], it is tough to aggregate multi-modal data of multiple clients, and the lack of training data will significantly reduce the efficiency of the deep cross-modal retrieval model. To efficiently utilize multi-modal data distributed across multiple clients, we study Federated Cross-Modal Retrieval(FedCMR).

Federated learning, firstly proposed by Google in 2016[11], is a machine learning setting where many clients cooperate in training a model under the coordination of a central server while maintaining the dispersion of data [9]. The existing federated learning models mostly are work on single modality data [3, 13, 15]. Compared with the single modality model, the multi-modal model is more complex in terms of the model functionality and model size. It is necessary to explore how to use the complex local multi-modal data reasonably and aggregate the local multi-modal models efficiently.

In this paper, we propose a cross-modal retrieval framework for distributed data. Taking three clients for example, the process is shown in Figure 1. The proposed FedCMR consists of three steps. (1) Local training: Each client trains the cross-modal retrieval model using the local data. (2) Aggregation: The server aggregates the common space of the clients. (3) Local update: Each client updates the common subspace of the local model based on the aggregated model computed by the last step.

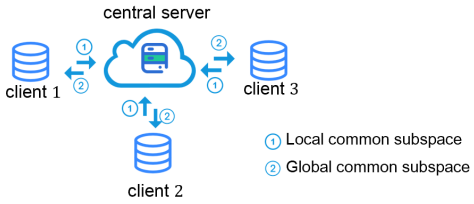

Figure 1: The framework of FedCMR

To the best of our knowledge, our framework is the first attempt to combine federated learning with cross-modal retrieval, trying to provide a solution for cross-modal retrieval in the distributed data storage scenario. We validate our approach on four benchmark datasets. The main contributions of this work can be summarized as follows: (1) By aggregating the updates of all common subspaces instead of aggregating the client model, we reduce the training communication overhead. (2) We present a smooth transition from global common subspace to local common subspace to reduce the impact on the conversion effect of data from feature space to common subspace.

2 THE PROPOSED METHOD

Given N clients which denoted as $C = \{C^1, C^2, \cdots, C^N\}$, with a set of data $D = \{D^1, D^2, \cdots, D^N\}$, where $D^k = \{x_i^k\}_{i=1}^{n^k}$ is the set of instances in k-th client and n^k is the number of instances in the k-th client. Each instance contains m modalities, i.e., $x_i^k = \{x_{1,i}^k, x_{2,i}^k, \cdots, x_{m,i}^k\}$. We introduce the three key steps of FedCMR in the following.

2.1 Local Cross-modal retrieval Network

We employ DSCMR [26] as the local model to handle the multi-modal data in each client effectively. Taking image modality and text modality for example, the net structure is shown in Figure 2. It is composed of two sub-networks to generate feature vector, one linear layer to ensure the two sub-networks learn a shared common subspace for image and text modalities, and one linear classifier to learn discriminative features by exploiting the label information.

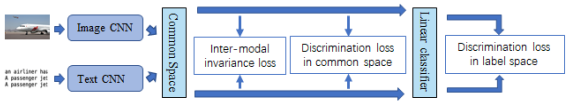

Figure 2: The cross-modal network structure

In the k-th client, define the representation matrix as $U_i^k = \left[u_{i,1}^k, u_{i,2}^k, \cdots, u_{i,n^k}^k\right]$, $i \in \{1, 2\}$ and $u_{i,j}^k$ is the learned representation for the j-th instance of i-th modality in the k-th client. Denote the label matrix $Y^k = \left[y_1^k, y_2^k, \cdots, y_{n^k}^k\right]$ and $y_i^k = [y_i^k(1), y_i^k(2), \cdots, y_i^k(c^k)]^T$ where $y_i^k(j) = 1$ if the i-th instance belongs to the j-th category, $y_i^k(j) = 0$ otherwise. c^k is the number of categories in the k-th client. The objective function of DSCMR is as follows:

$$L = l_1 + \lambda l_2 + \eta l_3 \tag{1}$$

where $l_1 = \frac{1}{n^k} \sum_{i=1}^{2} \|P^T U_i^k - Y^k\|_F$ is the discrimination loss in the label space, P is the projection matrix of the linear classifier. $l_2 = \frac{1}{(n^k)^2} \sum_{\alpha,\beta=1}^{2} \sum_{i,j=1}^{n^k} (log(1 + e^{\Gamma_{ij}^{\alpha\beta}}) - S_{ij}^{\alpha\beta} \Gamma_{ij}^{\alpha\beta})$ is the loss in the common representation space, $\Gamma_{ij}^{\alpha\beta}$ is the similarity between two instances in the α-th modality and the β-th modality, $S_{ij}^{\alpha\beta}$ is an indicator function whose value is 1 if the two elements are the representations of intra-class samples, otherwise 0. Notice that $\alpha = \beta$ shows the intra-modality relationship, and $\alpha \neq \beta$ shows the inter-modality relationship. $l_3 = \frac{1}{n^k} \|U_1^k - U_2^k\|_F$ is the modality invariance loss. The hyper-parameters λ and η control the contributions of the last two components.

2.2 Model Aggregation

2.2.1 Aggregation Framework. In the traditional federated learning algorithm, multiple clients train the model collaboratively under the coordination of a trusted central server. During each round of communication, each selected client computes an update to the model using the entire local dataset and uploads the training result model to the server for security aggregation. However, when the local net architecture is complex and the number of network parameters is large, such an approach usually results in a large communication overhead during the training process.

Considering that the cross-modal network uses the common subspace to connecting multiple modalities, we share the knowledge among clients by using the common space. Specifically, the server collects the linear layer parameters which learn the common space and combines the updates of all common subspaces as eq.(2) to find a globally consistent potential common subspace.

$$W_t = \sum_{k=1}^{K} q_t^k W_t^k \tag{2}$$

where K is the number of selected clients during the t-th communication, W_t^k is the parameters of the penultimate linear layer from the local cross-modal retrieval network which can be seen as the updated local common subspace, W_t is the global common space and q_t^k is the weight of the k-th client.

2.2.2 Weighting Scheme. Since the model will be more generalized with more samples and label categories, the weight q_t^k of each client is proportional to the number of samples and the number of label categories owned by the client, i.e., $q_t^k \propto S^k$,

$$S^k = \frac{n^k}{\sum_{k=1}^{K} n^k} \cdot \frac{c^k}{\sum_{k=1}^{K} c^k}. \tag{3}$$

Moreover, given that in each round of communication, under the same loss calculation mechanism, the retrieval ability of the local cross-modal model is stronger when the loss value of the model is lower. Thus, we believe that the variable weight should be inversely proportional to the loss value. Inspired by Gompertz function [12], we calculate V_t^k of each client as eq.(4).

$$V_t^k = e^{-e^{f_t^k / (\frac{\sum_{k=1}^{K} f_t^k}{K})}}, \tag{4}$$

where f_t^k represents the loss value of cross-modal retrieval model trained by local data in the t-th round of communication. We could conclude that $q_t^k \propto V_t^k$.

Overall, the weight q_t^k of the client is as follows. The hyper-parameters α controls the weight of the loss of the model.

$$q_t^k = \frac{e^{S^k + \alpha V_t^k}}{\sum_{k=1}^{K} e^{S^k + \alpha V_t^k}} \tag{5}$$

2.3 Local update

In the traditional federated learning algorithm, after the central server obtaining the global model W_t, the central server then sends W_t back to all clients, and each client replaces the local model W_t^k with W_t. However, in a multi-modal setting, it is very stiff to directly replace the W_t^k with W_t. To mitigate this issue, we present a smooth transition for W_{t+1}^k, which integrates the local training update and improves the stability of the overall framework.

Firstly, at the beginning of each communication round, the clients record the local common subspace as W_0^k.

Then, each client trains the model with local data to get the updated local common subspace W_t^k and uploads it to the central server. The server returns the aggregated common subspace W_t.

Finally, after receiving W_t from the server, as shown in eq.(6), each client adds the modifications $W_t^k - W_0^k$ to W_t to get W_{t+1}^k. The client uses W_{t+1}^k as the common subspace of the local retrieval model to achieve a smooth transition from the global common subspace to the local common subspace.

$$W_{t+1}^k = W_t + \gamma(W_t^k - W_0^k) \tag{6}$$

2.4 The Algorithm

Algorithm 1 presents the implementation process of FedCMR. For the federated optimization process, to make each client obtains a high-quality model suitable for the local objective function, we divide the training process into two stages: 1) joint training and 2) independent enhancement training.

First, in the joint training stage, each client will randomly select 80% of the local dataset to complete the model aggregation process.

Second, during the independent enhancement training phase, each client complete the local update process and continue to iterate over the local model with the remaining 20% of the data, so as to shrink W_{t+1}^k and make it more suitable for the local model to measure the similarity between samples from different modalities.

Algorithm 1 The proposed FedCMR

procedure Federated Optimization
  Input: The set of clients, $C = \{C^1, C^2, \cdots, C^N\}$; The set of dataset, $D = \{D^1, D^2, \cdots, D^N\}$; The number of communication round, T; The number of selected clients, K;
  for each round t = 1, ..., T - 1 **do**
    $C_t, D_t \leftarrow$ (random set of K clients);
    $W_t \leftarrow$ Model Aggregation(C_t,D_t);
    for each client $c \in C_t$ **in parallel do**
      LocalUpdate(W_t);
    end for
  end for
end procedure
procedure Model Aggregation
  Input: C_t, D_t
  for each client $c \in C_t$ **in parallel do**
    Performs local training using 80% data;
    Obtains the updated local common subspace W_t^k;
    Calculates client weight q_t^k using eq.(5);
  end for
  Gets global common subspace W_t using eq.(2);
return W_t;
end procedure
procedure LocalUpdate
  Input: W_t
  Calculates common subspace W_{t+1}^k using eq.(6);
  Iterate over the local model with the remaining 20% data;
end procedure

3 EXPERIMENT

3.1 Experimental Setup

3.1.1 Datasets. We conduct experiments on four benchmark datasets: the MS-COCO dataset [14], the MIR-Flickr25K dataset [7], the Wikipedia dataset [19] and the Pascal Sentence dataset [18]. The statistics of the datasets are summarized in Table 1. The last column stands for the number of training/test subsets separately.

The above datasets have been collected together. Due to the limited amount of Pascal Sentence and Wikipedia dataset, we distribute the data equally among three clients to simulate the federated cross-modal retrieval process in the experiments.

Table 1: Statistics of the datasets.

Dataset	# of Categories	# of Instance
MS-COCO	80	82081/30137
MIRFlickr-25K	24	16012/2002
Wikipedia	10	2173/462
Pascal Sentence	20	800/100

3.1.2 Baselines. To show the effectiveness of the proposed federated learning method, we compare FedCMR with the following methods: (1)DSCMR [26], which conducts DSCMR on each client, but does not aggregate the clients. (2)FedAvg [15], which conducts DSCMR on each client, and then aggregates the clients using FedAvg. (3)FedProx [13], which conducts DSCMR on each client, and then aggregates the clients using FedProx.

3.1.3 Parameter setting. In our experiments, we adopt a 19-layer VGGNet [21] to learn the representations of the image samples and obtain a 4,096-dimensional representation vector outputted by the fc7 layer of the VGGNet for each image. For representing text samples, we use the sentence BERT [2] to learn a 1024-dimensional representation vector for each text. For the data partition of these datasets, we follow the data partition strategy of [4, 6, 8, 26].

All the parameter settings of the baselines are based on the original papers. For the proposed FedCMR, γ is set by the grid $[-2, -1, -0.5, 0.5, 1, 2]$ and α is set by the grid $[1, 5, 10, 15, 20, 30, 50, 100]$. We utilize the widely-used federated learning framework PySyft[20] to simulate the federated cross-modal retrieval process.

Table 2: Results on the Pascal Sentence dataset.

	Client	DSCMR	FedAvg	FedProx	FedCMR
Image→Text	A	0.695	0.683	0.657	**0.715**
	B	0.711	0.681	0.648	**0.726**
	C	0.689	0.671	0.626	**0.701**
Text→Image	A	0.706	0.665	0.587	**0.725**
	B	0.702	0.664	0.615	**0.732**
	C	**0.699**	0.639	0.575	0.683
Average	A	0.700	0.674	0.622	**0.720**
	B	0.706	0.673	0.632	**0.729**
	C	**0.695**	0.655	0.600	0.692

Table 3: Results on the Wikipedia dataset.

	Client	DSCMR	FedAvg	FedProx	FedCMR
Image→Text	A	0.469	0.456	0.400	**0.480**
	B	0.479	0.454	0.399	**0.487**
	C	0.450	0.445	0.393	**0.463**
Text→Image	A	**0.434**	0.391	0.355	0.429
	B	0.437	0.404	0.355	**0.454**
	C	0.414	0.375	0.334	**0.422**
Average	A	0.452	0.424	0.377	**0.455**
	B	0.458	0.430	0.377	**0.471**
	C	0.432	0.410	0.364	**0.443**

Table 4: Results on the MIRFlickr-25K dataset.

	Client	DSCMR	FedAvg	FedProx	FedCMR
Image→Text	A	**0.745**	0.740	0.739	0.742
	B	0.737	0.737	0.736	**0.742**
	C	0.734	0.736	0.734	**0.738**
Text→Image	A	0.758	0.761	0.752	**0.776**
	B	0.753	0.762	0.753	**0.771**
	C	0.752	0.759	0.746	**0.775**
Average	A	0.752	0.751	0.745	**0.759**
	B	0.745	0.749	0.745	**0.756**
	C	0.743	0.747	0.740	**0.757**

3.2 Experimental Result

We evaluate the baselines using the widely-used mean Average Precision (mAP) score for all four datasets. Table 2~Table 5 present the mAP scores of FedCMR and the compared methods. In each column of the tables, the best result is highlighted in boldface. We have the following observations from the tables: (1) FedCMR significantly outperforms the two benchmark federated learning methods on all of the four datasets. (2) FedCMR outperforms DSCMR generally. (3) By comparing with the mAP result on the MIRFlickr-25K and MS-COCO datasets, the advantage of FedCMR is more evident on the Pascal Sentence dataset and the Wikipedia dataset. The results indicate that it is reasonable to study multi-modal federated learning, especially in a scenario with a small number of data, and the multi-modal federated learning is more effective for aggregating cross-modal retrieval models.

Table 5: Results on the MS-COCO dataset.

	Client	DSCMR	FedAvg	FedProx	FedCMR
Image→Text	A	**0.779**	0.728	0.618	0.777
	B	**0.775**	0.696	0.623	0.774
	C	0.775	0.705	0.624	**0.781**
Text→Image	A	0.751	0.703	0.628	**0.755**
	B	0.752	0.698	0.629	**0.759**
	C	0.753	0.708	0.628	**0.755**
Average	A	0.765	0.715	0.623	**0.766**
	B	0.764	0.697	0.626	**0.766**
	C	0.764	0.706	0.626	**0.768**

3.3 Parameter Analysis

In this subsection, we analyze the effect of the hyper-parameters on the performance of FedCMR. Due to space limitation, we only present the influence of α on the Pascal Sentence dataset. We set α vary in the range $[1, 5, 10, 15, 20, 30, 50, 100]$. With $\gamma = 1$, Figure 3 shows the average mAP scores of FedCMR on the three clients versus different values of α. We can see that all three clients perform stable relatively when α vary in $[10, 15, 20, 30]$. Thus α is set as 20 for all experiments.

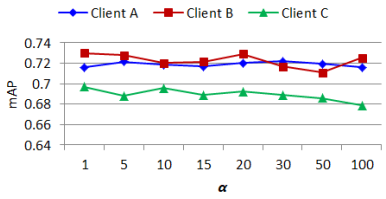

Figure 3: Analysis of parameter α

4 CONCLUSION

This paper proposes a novel framework to provide a solution for cross-modal retrieval in the distributed data storage scenario. Based on the traditional federated learning algorithm, we have innovatively implemented the following steps to form FedCMR: 1) Aggregation of common subspace; 2) Client weight calculation method designed for cross-modal retrieval; 3) Smooth transition from global common subspace to local common subspace; 4) Two-stage enhanced training. Comprehensive experimental results on four widely-used multi-modal datasets have demonstrated the effectiveness of our proposed framework. As for future work, we want to further explore the federated cross-modal retrieval under harsh conditions, such as small data distribution and incomplete data.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No.61806034; No.61876028; No.61632019; No.61972065; No.62006034) and the Provincial College Student Innovation and Entrepreneurship Training Program.

REFERENCES

[1] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, and Jungong Han. 2020. IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. (2020).

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[3] Moming Duan, Duo Liu, Xianzhang Chen, Yujuan Tan, Jinting Ren, Lei Qiao, and Liang Liang. 2019. Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications. In *2019 IEEE 37th International Conference on Computer Design (ICCD)*. IEEE, 246–254.

[4] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*. 7–16.

[5] Harold Hotelling. 1935. Relations Between Two Sets of Variates. *Biometrika* 28 (1935), 321–377.

[6] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 635–644.

[7] Mark J Huiskes and Michael S Lew. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. 39–43.

[8] Qing-Yuan Jiang and Wu-Jun Li. 2017. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3232–3240.

[9] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2019. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977* (2019).

[10] Jon R Kettenring. 1971. Canonical analysis of several sets of variables. *Biometrika* 58, 3 (1971), 433–451.

[11] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

[12] A K Laird, S A Tyler, and A D Barton. 1965. Dynamics of normal growth. *Growth* 29, 3 (1965), 233–248.

[13] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2018. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127* (2018).

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[15] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*. PMLR, 1273–1282.

[16] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2020. Fine-grained Visual Textual Alignment for Cross-Modal Retrieval using Transformer Encoders. (2020).

[17] Marta Otto. 2018. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation–GDPR). In *International and European Labour Law*. Nomos Verlagsgesellschaft mbH & Co. KG, 958–981.

[18] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 139–147.

[19] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. 251–260.

[20] Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. 2018. A generic framework for privacy preserving deep learning. *arXiv preprint arXiv:1811.04017* (2018).

[21] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[22] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*. 154–162.

[23] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2016. On Deep Multi-View Representation Learning: Objectives and Optimization. *arXiv e-prints* (2016), arXiv–1602.

[24] Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. 2016. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454* (2016).

[25] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2013. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 6 (2013), 965–978.

[26] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10394–10403.