An Adversarial Learning and Canonical Correlation Analysis based Cross-Modal Retrieval Model

Thi-Hong Vuong[1], Thanh-Huyen Pham[1,2], Tri-Thanh Nguyen[1], and Quang-Thuy Ha[1]

[1] Vietnam National University, Hanoi (VNU), VNU-University of Engineering and Technology (UET), No. 144, Xuan Thuy, Cau Giay, Hanoi, Vietnam
{hongvt57, ntthanh, thuyhq}@vnu.edu.vn
[2] Ha Long University, Quang Ninh, Vietnam
{phamthanhhuyen}@daihochalong.edu.vn.

Abstract. The important of cross-modal retrieval approaches is to find a maximally correlated subspace between multimodal data. This paper introduces a novel Adversarial Learning and Canonical Correlation Analysis based Cross-Modal Retrieval (ALCCA-CMR) model, which seeks an effective learning representation. We train two-branch for each multimodal data to seek an effective common subspace by the adversarial learning. Cross-modal correlation learning identifies a relationship between different modalities in sets of variables on an effective common subspace by canonical correlation analysis. We demonstrate an application of ALCCA-CMR model implemented for bi-modal data. Experimental results on real music data show the efficacy of the proposed method in comparison with other existing ones.

Keywords: Cross-modal retrieval · adversarial learning · canonical correlation analysis.

1 Introduction

Cross-modal retrieval has drawn much attention due to the explosion multimodal data. The different types of media data such as text, image, and video are used for describing the same events or topics. In order to optimally benefit from the source of multimodal data and make maximal use of the developing multimedia technology, automated mechanisms are to set up a similarity link from one multimedia item to another if cross datasets semantically correlated. Constructing a joint representation invariant across different modalities is of significant importance in many multimedia applications. Previous studies have focused mainly on single modality scenarios [2, 7, 11]. However, these techniques mainly use metadata such as keywords, tags or associated descriptions to calculate similarity than content-based information. In this study, we use content-based multimodal data for cross-modal retrieval as [5, 13, 14, 18]. There are various approaches have

been proposed to deal with this problem, which can be roughly divided into two categories as [16]: real-value representation learning [13, 14, 18] and binary representation learning [5, 17, 22]. The approach in this paper focuses on in the category of real-value representation.

Features of multi-modal data have inconsistent distribution and representation, therefore a modality gap needs to be bridged which ways need to be found to access the semantic similarity of items across modolities. A common approach to bridge the modality gap is representation learning. The goal is to find projections of data items from different modalities into common feature representation subspace in which the similarity between them can be assessed directly. Recently, the study have focused on maximize the cross-modal pairwise item correlation or item classification accuracy like canonical correlation analysis [10, 19, 20]. However, the existing approaches fail to explicitly address the statistical aspect of the transformed features of multi-modal data, the similarity between their distributions must be measured in a certain way. The practical challenge is the difficulty of obtaining well-matched cross datasets that are essential for data-driven learning as deep learning [12, 15, 18].

We focus on real-value approach for the supervised representation learning by the adversarial learning and CCA for cross-modal retrieval (ALCCA-CMR). The adversarial learning was inspired by the effectiveness of for image applications [6, 21, 14]. On the one hand, CCA and DNN combined together to deep representations in computer vision, like DCCA method [1]. Therefore, we use a deep learning with the adversarial learning and CCA to find a common subspace effectively. We evaluate the proposed approach on music dataset and show that it significantly outperforms the state-of-the-art in cross-modal retrieval. Section 2 shows the detail of ALCCA-CMR method and evaluate it in Section 3. Section 4 describes the related existing work. Section 5 concludes the paper.

2 ALCCA-CMR Model

2.1 Problem Formulation

The ALCCA-CMR contains two sub-problems: ALCCA and CMR. The ALCCA build CCA to seek an common subspace effectively by adversarial learning and CCA. Then, CMR retrieve cross-modal base on the common subspace.

In ALCCA, input is feature matrices of two modalities as $\mathbf{A} = \{a_1, ..., a_n\}$ and $\mathbf{T} = \{t_1, ..., t_n\}$ with label matrix $\mathbf{Y} = \{y_1, ..., y_n\}$, where n is the number of samples. Output is ALCCA model which find an common subspace $\boldsymbol{S}$ for mapping cross-modal. In $\boldsymbol{S}$, the similarity of different points reflects the semantic closeness between their corresponding original inputs. We assume that f_A and f_T can take $\mathbf{A}$ and $\mathbf{T}$ in $\boldsymbol{S} = \{\boldsymbol{S}_A, \boldsymbol{S}_T\}$ such as $\boldsymbol{S}_A = f_A(\mathbf{A}; \theta_A)$ and $\boldsymbol{S}_T = f_T(\mathbf{T}; \theta_T)$. We have two mappings $f_A(\mathbf{a}; \theta_A)$ and $f_T(\mathbf{t}; \theta_T)$ that transform audio and lyrics text features into d dimensional vector $\boldsymbol{s}_A$ and $\boldsymbol{s}_T$ with $\boldsymbol{s}_A^i = f_A(\mathbf{a}_i; \theta_A)$ and $\boldsymbol{s}_T^i = f_T(\mathbf{t}_i; \theta_T)$. In the subspace, we use CCA with the number of components from 10 to 100.

In CMR, input gives a audio/lyrics as query. Output takes a lyrics/audio list which relevant with the audio/lyrics query.

2.2 Proposed Framework

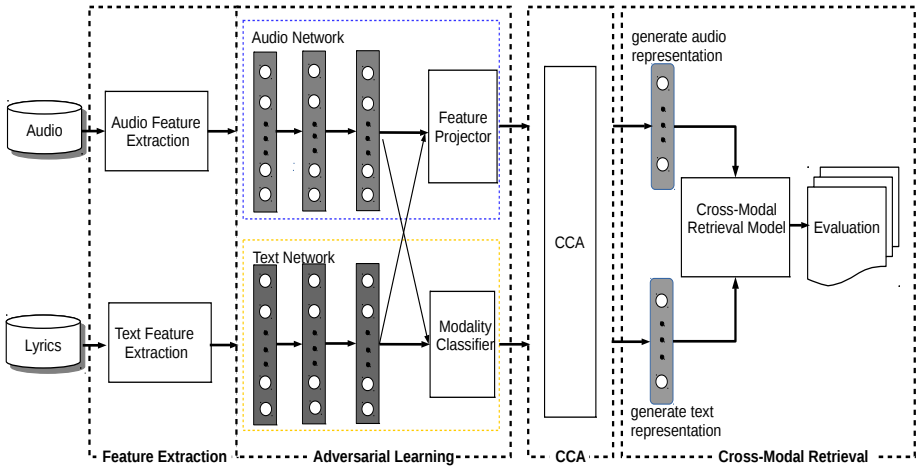

Fig. 1. The general flowchart of the proposed method. Given audio and lyrics, the feature extraction phase extracts audio features and lyrics text features. For each modality, ALCCA seek an effective common subspace in the adversarial learning phase and calculate their similarity by CCA embedding for CMR.

The process of cross-modal retrieval is showed in Figure 1. The feature extraction phase extracts audio feature and lyrics text feature. The ALCCA phase tries to generate a common subspace for supervised multi-modal data. Adversarial learning is the interplay between feature projector and modality classifier D with parameter θ_D, conducted as a minimax game. The feature projector and classifier trained under the adversarial leaning. Audio and lyrics features first pass through respective transformation f_A and f_T. The goal modality classifier is to maximize its prediction precision given a transformed feature vector. Whereas, the feature projector are trained to generate modality invariant features minimizing the classifier's prediction precision. Then, transformed features are calculated their similarity by CCA function. The CMR implement cross-modal retrieval and evaluate performance of CMR.

2.3 Adversarial Learning and CCA

Adversarial Learing. We based on the adversarial learning as [14] to design for audio and lyrics text. In the adversarial learning, *feature projector* are trained

to generate modality invariant features to maximize the modality classifier error while *modality classifier* is trained to minimize its error.

Feature projector. The goal of feature projector implements the process of modality-invariant embedding of audio and lyrics into a common subspace. In the feature projector, we use embedding loss L_{emb} that it is formulated as the combination of the intra-modal discrimination loss L_{imd} and the inter-modal invariance loss L_{imi} with regularization L_{reg}.

$$L_{imd}(\theta_{imd}) = -\frac{1}{n}\sum_{i=1}^{n}(m_i.(log\hat{p}_i(a_i) + log(1-\hat{p}_i(t_i)). \tag{1}$$

where m_i is the ground-truth modality label of each instance, expressed as one-hot vector, $\hat{p}$ is probability distribution of semantic categories per item.

$$L_{emd}(\theta_A, \theta_T, \theta_{imd}) = \alpha.L_{imi} + \beta.L_{imd} + L_{reg}. \tag{2}$$

$$L_{imi}(\theta_A, \theta_T) = L_{imi}(\theta_A) + L_{imi}(\theta_T). \tag{3}$$

$$= \sum_{i,j,k} l2(a_i, t_j) + \sum_{i,j,k} l2(t_i, a_j)) \tag{4}$$

where the hyper-parameters α and β control the contributions of the two terms. All distance between the feature mapping $f_A(A;\theta_A)$ and $f_T(T;\theta_T)$ per couple item pair were used l2 norm.

$$L_{reg} = \sum_{l=1}^{L}(||W_a^l||F + ||W_t^l||F) \tag{5}$$

where F denotes the Frobenius norm and W_a, W_t represent the layer-wise prameters of DNNs.

Modality Classifier. A modality classifer D with paramter θ_D which actives as discriminator The adversarial loss L_{adv} is cross-entropy loss of modality classification.

$$L_{adv}(\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}(m_i.(logD(a_i;\theta_D) + log(1-D(t_i;\theta_D))). \tag{6}$$

Optimization. The optimization goals of the two objective functions are opposite, the process runs as minimax game [6] as follow:

$$\hat{\theta_A}, \hat{\theta_T}, \hat{\theta_{imd}} = \underset{(\theta_A,\theta_T,\theta_{imd})}{argmin}\ (L_{emd}(\theta_A, \theta_T, \theta_{imd}) - L_{adv}(\hat{\theta_D})). \tag{7}$$

$$\hat{\theta_D} = \underset{(\theta_D)}{argmax}(L_{emd}(\hat{\theta_A}, \hat{\theta_T}, \hat{\theta_{imd}}) - L_{adv}(\theta_D)). \tag{8}$$

As in [14], minimax optimization was performed efficiently by incorporating Gradient Reversal Layer (GRL). If GRL is added before the first layer of the modality classifier, we update the model parameters using following rules

$$\theta_A \leftarrow \theta_V - \mu.\nabla_{\theta_A}(L_{emb} - L_{adv}), \tag{9}$$

$$\theta_T \leftarrow \theta_T - \mu.\nabla_{\theta_T}(L_{emb} - L_{adv}), \tag{10}$$

$$\theta_{imd} \leftarrow \theta_{imd} - \mu.\nabla_{\theta_{imd}}(L_{emb} - L_{adv}), \tag{11}$$

$$\theta_D \leftarrow \theta_D + \mu.\nabla_{\theta_{imd}}(L_{emb} - L_{adv}). \tag{12}$$

where μ is learning rate. The results of the adversarial learning learn representation in common subspace: $f_A(A)$ and $f_T(T)$.
The procedure is shown in Algorithm 1: pseudocode of the proposed method use ALCCA for cross-modal retrieval.

Algorithm 1 Pseudocode of the proposed method

procedure ProposedMethod($\boldsymbol{A}, \boldsymbol{T}$)
  Compute spectrogram from audio $\boldsymbol{A}, \rightarrow \boldsymbol{F}_A$
  Compute textual feature from lyrics $\boldsymbol{T}, \rightarrow \boldsymbol{F}_T$
  for each epoch **do**
    Randomly divide $\boldsymbol{F}_A, \boldsymbol{F}_T$ to batches
    for each batch $(\boldsymbol{\omega}_A, \boldsymbol{\omega}_T)$ of audio and lyrics **do**
      for each pair $(\boldsymbol{a}, \boldsymbol{t}) \in (\boldsymbol{\omega}_A, \boldsymbol{\omega}_T)$ **do**
        Compute representations f_A and f_T
        for k steps **do**
          Update parameters θ_A as Eq. 9
          Update parameters θ_T as Eq. 10
          Update parameters θ_{imd} as Eq. 11
        Update parameters θ_D as Eq. 12
        learned representation in $\mathbf{S}$=(f_A, f_T)
        $\boldsymbol{a} \rightarrow \boldsymbol{x}$ byf_A
        $\boldsymbol{t} \rightarrow \boldsymbol{y}$ by f_T
      Get converted batch $(\boldsymbol{X}, \boldsymbol{Y})$
      Apply CCA on $(\boldsymbol{X}, \boldsymbol{Y})$ to compute $\boldsymbol{W}_X, \boldsymbol{W}_Y$ as Eq. 13
      Compute number of canonical components

CCA. CCA is used to maximally correlated between two multi-dimension variables $\boldsymbol{X} \in R^{p \times n}$ and $\boldsymbol{Y} \in R^{q \times n}$. Here n is the number of samples, p and q are the number of features of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. When a linear projection is performed, CCA tries to find two canonical weights $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$, so that the

correlation between the linear projections $\boldsymbol{w}_x\boldsymbol{X}^T$ and $\boldsymbol{w}_y\boldsymbol{Y}^T$ is maximized.The correlation coefficient ρ is given as

$$\begin{aligned}\rho &= \underset{(\boldsymbol{w}_x,\boldsymbol{w}_y)}{argmax}\, corr(\boldsymbol{w}_x^T\boldsymbol{x}, \boldsymbol{w}_y^T\boldsymbol{y}) \\ &= \underset{(\boldsymbol{w}_x,\boldsymbol{w}_y)}{argmax} \frac{\boldsymbol{w}_x^T\boldsymbol{C}_{xy}\boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^T\boldsymbol{C}_{xx}\boldsymbol{w}_x \cdot \boldsymbol{w}_y^T\boldsymbol{C}_{yy}\boldsymbol{w}_y}}.\end{aligned} \tag{13}$$

where C_{xy} is the cross-covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$, while C_{xx} and C_{yy} are covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. CCA obtains two directional basis vectors $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ such that the correltaion between $\boldsymbol{X}^T$ w_x and $\boldsymbol{Y}^T$ w_y is maximum. Regularied CCA (RCCA) [4] is an improved version of CCA which used a ridge regression optimization scheme to prevent over-fitting of insufficient training data. However, RCCA is computationally very expensive because of this regularization process. We use CCA and CCA variants to calculate the similarity between audios and lyrics in the common subspace with number of canonical components for cross-modal retrieval.

2.4 Cross-Modal Retrieval

In the CMR phase, we use 20% data to evaluate the peformance of the ALCCA when using audio or lyrics as query. We evaluate 5 cross-validation on multi-modal data.

Evaluation metric. In the retrieval evaluation, we use the standard evaluation criteria used in most prior work on cross-modal retrieval [20]. We use mean reciprocal rank 1 (MRR1) and recall@N as the metrics. Because there is only one relevant audio or lyrics, MRR1 is able to show the rank of the result. MRR1 is defined by Eq. 14

$$MRR1 = \frac{1}{N_q}\sum_{i=1}^{N_q}\frac{1}{rank_i(1)}, \tag{14}$$

where N_q is the number of the queries and $rank_i(1)$ corresponds to the rank of the relevant item in the i-th query. We also evaluate recall@N to see how often the relevant item is included in the top of the ranked list. Assume S_q is the set of its relevant items ($|S_q| = 1$) in the database for a given query and the system outputs a ranked list K_q ($|K_q| = N$). Then, recall@N is computed by Eq. 15 and is averaged over all queries.

$$recall@N = \frac{|S_q \bigcap K_q|}{|S_q|} \tag{15}$$

3 Experiments

3.1 Experimental Setup

We implement the proposed method on a music dataset and compare with the methods as the same in [20]. First, the music datset have 10,000 pairs of audio and lyrics with 20 most frequent mood categories *(aggressive, angry, bittersweet, calm, depressing, dreamy, fun, gay, happy, heavy, intense, melancholy, playful, quiet, quirky, sad, sentimental, sleepy, soothing, sweet)* .

Audio feature extraction. The audio signal is represented as a spectrogram. We mainly focus on mel-frequency cepstral coefficients (MFCCs). For each audio signal, a slice of 30s is resampled to 22,050Hz with a single channel. Each audio extracted 20 MFCC sequences and 161 frames for each MFCC.

Lyrics text feature extraction. From the sequence of words in the lyrics, textual feature is computed, more specifically, by a pre-trained Doc2vec [8] model, generating a 300-dimensional feature for each song.

Implementation details. We deploy our proposed method as follow: the adversarial learning with three-layer feed-forward neural networks activated by *tanh* function to nonlinearly project the raw audio and lyrics text features into common subspace, i.e., (A $\rightarrow$ 1000 $\rightarrow$ 200 for audio modality and T $\rightarrow$ 200 $\rightarrow$ 200 for lyric text modality). With modality classifier, we stick to the three fully connected layers (f $\rightarrow$ 50 $\rightarrow$ 2). We use the same parameters in [14] with batch size is set to 100 and the training takes 200 epochs for proposed method.

After learned representation in common subspace, we use they calculate their similarity by CCA function for cross-modal retrieval. Here, we evaluate the impact of the number of CCA components, which affects the performance of both the baseline methods and the proposed methods. The number of CCA components is adjusted from 10 to 100.

Comparison with baseline methods. We compare our proposed method against all the methods which used in [20] such as PretrainCNN-CCA, Spotify-DCCA, PretrainCNN-DCCA, JointTrain-DCCA the same dataset. This comparison can be verify the effectiveness of our proposed adversarial and correlation learning for coss-modal retrieval.

3.2 Experimental Results

There are two kinds of MRR1 measures to evaluate the effectiveness as [20]: instance-level MRR1 and category-level MRR1. Instance-level MRR1 is to retrieve items of different datasets without label. Category-level MRR1 is to retrieve multi-modal data within label. I-MRR1-A, C-MRR1-A are instance-level MRR1 and category-level when using audio as query. I-MRR1-L, C-MRR1-L are instance-level MRR1 when using lyrics as query.

Proposed method results. The proposed method results implements five cross-validate on dataset with MRR1, R@1 and R@5 measure when using audio as query or lyrics as query.

Table 1. Performance cross-modal retrieval of the propose method

#CCA	I-MRR1-A	I-MRR1-L	C-MRR1-A	C-MRR1-L	R@1-A	R@1-L	R@5-A	R@5-L
10	0.08	0.081	0.213	0.212	0.045	0.047	0.100	0.099
20	0.200	0.200	0.305	0.305	0.137	0.136	0.251	0.253
30	0.300	0.300	0.387	0.387	0.224	0.224	0.371	0.376
40	0.370	0.366	0.448	0.445	0.288	0.284	0.454	0.447
50	0.415	0.411	0.488	0.484	0.335	0.327	0.498	0.496
60	0.439	0.436	0.506	0.506	0.358	0.354	0.523	0.519
70	0.453	0.449	0.519	0.517	0.371	0.367	0.539	0.535
80	0.456	0.452	0.521	0.519	0.373	0.370	0.540	0.536
90	0.447	0.444	0.515	0.513	0.365	0.362	0.531	0.529
100	0.427	0.425	0.497	0.497	0.349	0.346	0.507	0.505

In Table 1, the performance of the cross-modal retrieval overall measures are approximate between using audio and lyrics as query, which demonstrates that the cross-modal common subspace is useful for both audio and lyrics retrieval. When the number of CCA components increases from 10 to 40, the performance also significantly increases from 10% to 30%. After that, there is a slight increase from 30% to 40% when the number of CCA components gets more 40. The category-level MRR1 and recall@5 are higher and more stable than another measures.

Comparison with baseline methods. The ALCCA-CMR model performance is more effective than the baseline methods on the same music dataset overall measures when using audio/lyrics as query.

The Figure 2 demonstrates that the our proposed method significantly outperforms PretrainCNN-CCA, DCCA, PretrainCNN-DCCA and JointTrainDCCA on the instance-level MRR1 measure when the number of components gets than 30. The results of the proposed method are high and stable about 40% while the results are about 25% with JointTrainDCCA, 20% with PretrainCNN-DCCA, about 15% with DCCA and about 10% with PretrainCNN-CCA.

The results in Figure 3 show that the our proposed method is better than PretrainCNN-CCA, DCCA, PretraiinCNN-DCCA and JointTrainDCCA on the category-level MRR1 measure when the number of component gets than 30. The results of the proposed method are high from 40% to 50% while the results are about 35% with JointTrainDCCA, 32% with PretrainCNN-DCCA, about 25% with DCCA and about 20% with PretrainCNN-CCA.

The results Figure 4 show that the our proposed method is more effective than JointTrainDCCA on the recall@1 and recall@5 when the number of component gets than 40. The results of the proposed method are high from 40% to 50% with R@5 and about 35% with R@1. While the results of JointTrainDCCA are stable about 25% both R@1 and R@5.

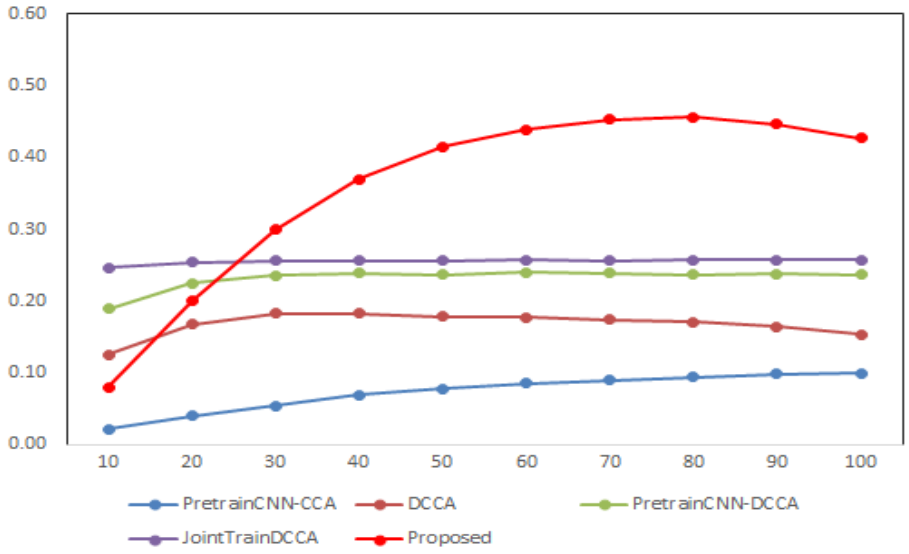

Fig. 2. Comparison with the baseline methods on instance-level MRR1

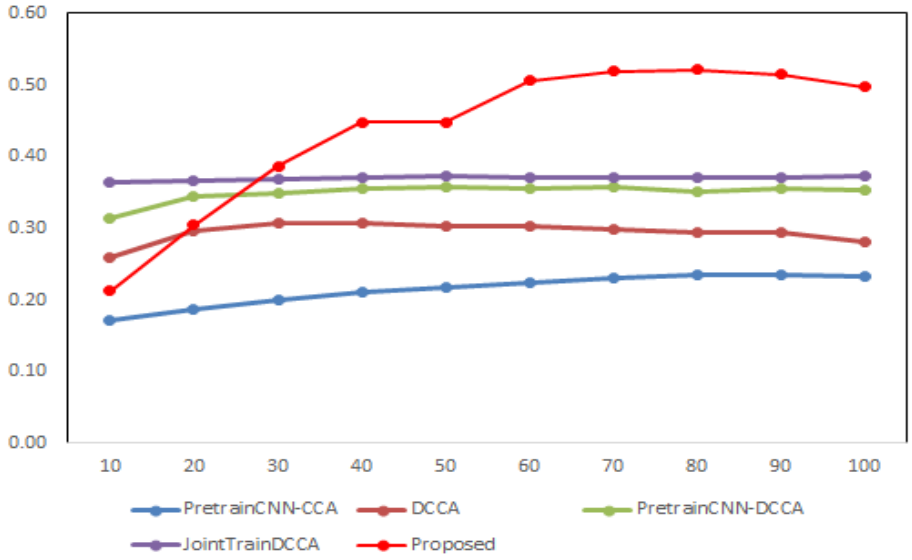

Fig. 3. Comparison with the baseline methods on category-level MRR1

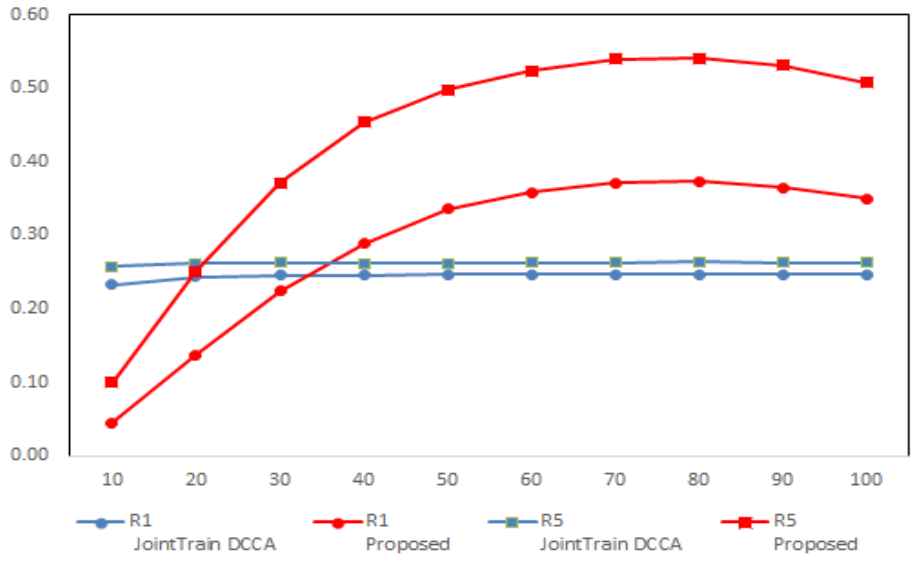

Fig. 4. Comparison with the baseline methods on Recall

4 Related Work

This section on presents the fundamental concepts in the theories of deep learning and CCA. With the rapid development of deep neural network (DNN) models, DNN has increasingly been deployed in the cross-modal retrieval context as well [5, 14, 15, 18]. The existing DNN-based cross multimedia retrieval models mainly focus on ensuring the pairwise similarity of the item pairs in a common subspace which multi-modal data can be compared directly. However, a common representation learned in this way fails to fully preserve the underlying cross-modal semantic structure in data. In [14], a adversarial corss-modal retrieval (ACMR) method used adversarial learning which was proposed by Goodfellow et al.[6] in GAN for image generation, as regularization into cross-modal retrieval for image and text. The adversarial learning used maximize the correlation through features projections and regularize their distributions on modality classifier. Through the joint exploitation of two processes in [14] such as minimax game, the underlying cross-modal semantic structure of bimodal data is better preserved when this data is projected into the common subspace. The adversarial approach learn effective subspace representation for image and text retrieval.

CCA is a statistical technique that extracted correlation between two dataset, X and Y, by using cross-covariance matrices [3, 4, 9, 10]. It capitalizes on the knowledge that the different modalities represent different sets of descriptors for characterizing the same object. CCA has many characteristics that make it suitable for analysis of real-world experimental data. First, CCA does not require

that the datasets have the same dimensionality. Second, CCA can be used with more than two datasets simultaneously. Third, CCA does not presuppose the directionality of the relationship between datasets. Fourth, CCA characterizes relationships between datasets in an interpretable way. This is in contrast to correlation methods that merely quantify similarity between datasets. In recent years, deep learning and CCA has used to fuse heterogeneous data such as pixel values of images and text [18], audio and image [3]. Regularized CCA (RCCA) is an advance version of CCA, which used a ridge regression optimiztion scheme [4, 9] in the presence of insufficient training data to prevent overfitting.

The approach proposed in this paper focus on real-value approach for music retrieval. We combine for the supervised representation learning by the adversarial learning and CCA for audio and lyrics retrieval. Our approach was inspired by the effectiveness of the adversarial learning for image applications [6, 21, 14]. On the one hand, CCA and DNN combined together to deep representations in computer vision, like DCCA method [1]. Furthermore, our approach is motivated in music applications instead of focus on image applications.

5 Conclusion

The paper propose the ALCCA-CMR model for cross-modal retrieval. Our approach is inspired by the effectiveness of the adversarial learning and CCA for the supervised multi-modal data. The ALCCA find the common subspace representation which the different data can be compared directly. The results demonstrated that our method is more effective than the baseline methods for both using audio and lyrics as query. In the future, we will advance cross-modal retrieval accuracy by CCA variants and retrieval time.

References

1. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning. pp. 1247–1255 (2013)
2. Boutell, M., Luo, J.: Photo classification by integrating image content and camera metadata. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. vol. 4, pp. 901–904. IEEE (2004)
3. Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th annual international conference on machine learning. pp. 129–136. ACM (2009)
4. De Bie, T., De Moor, B.: On the regularization of canonical correlation analysis. Int. Sympos. ICA and BSS pp. 785–790 (2003)
5. Feng, F., Li, R., Wang, X.: Deep correspondence restricted boltzmann machine for cross-modal retrieval. Neurocomputing **154**, 50–60 (2015)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
7. Hu, X., Downie, J.S., Ehmann, A.F.: Lyric text mining in music mood classification. American music **183**(5,049), 2–209 (2009)

8. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning. pp. 1188–1196 (2014)
9. Mandal, A., Maji, P.: Regularization and shrinkage in rough set based canonical correlation analysis. In: International Joint Conference on Rough Sets. pp. 432–446. Springer (2017)
10. Mandal, A., Maji, P.: Faroc: fast and robust supervised canonical correlation analysis for multimodal omics data. IEEE transactions on cybernetics **48**(4), 1229–1241 (2018)
11. McAuley, J., Leskovec, J.: Image labeling on a network: using social-network metadata for image classification. In: European conference on computer vision. pp. 828–841. Springer (2012)
12. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 689–696 (2011)
13. Peng, Y., Huang, X., Qi, J.: Cross-media shared representation by hierarchical learning with multiple deep networks. In: IJCAI. pp. 3846–3853 (2016)
14. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Proceedings of the 2017 ACM on Multimedia Conference. pp. 154–162. ACM (2017)
15. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2088–2095 (2013)
16. Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215 (2016)
17. Xia, R., Pan, Y., Lai, H., Liu, C., Yan, S.: Supervised hashing for image retrieval via image representation learning. In: AAAI. vol. 1, p. 2 (2014)
18. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3441–3450 (2015)
19. Yao, T., Mei, T., Ngo, C.W.: Learning query and image similarities with ranking canonical correlation analysis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 28–36 (2015)
20. Yu, Y., Tang, S., Raposo, F., Chen, L.: Deep cross-modal correlation learning for audio and lyrics in music retrieval. arXiv preprint arXiv:1711.08976 (2017)
21. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. arXiv preprint (2017)
22. Zhang, J., Peng, Y., Yuan, M.: Unsupervised generative adversarial cross-modal hashing. arXiv preprint arXiv:1712.00358 (2017)