

# Deep Triplet Neural Networks with Cluster-CCA for Audio-Visual Cross-Modal Retrieval

DONGHUO ZENG, National Institute of Informatics, SOKENDAI, Japan

YI YU, National Institute of Informatics, SOKENDAI, Japan

KEIZO OYAMA, National Institute of Informatics, SOKENDAI, Japan

Cross-modal retrieval aims to retrieve data in one modality by a query in another modality, which has been a very interesting research issue in the field of multimedia, information retrieval, and computer vision, and database. Most existing works focus on cross-modal retrieval between text-image, text-video, and lyrics-audio. Little research addresses cross-modal retrieval between audio and video due to limited audio-video paired datasets and semantic information. The main challenge of audio-visual cross-modal retrieval task focuses on learning joint embeddings from a shared subspace for computing the similarity across different modalities, where generating new representations is to maximize the correlation between audio and visual modalities space. In this work, we propose a novel deep triplet neural network with cluster canonical correlation analysis (TNN-C-CCA), which is an end-to-end supervised learning architecture with audio branch and video branch. We not only consider the matching pairs in the common space but also compute the mismatching pairs when maximizing the correlation. In particular, two significant contributions are made: i) a better representation by constructing deep triplet neural network with triplet loss for optimal projections can be generated to maximize correlation in the shared subspace. ii) positive examples and negative examples are used in the learning stage to improve the capability of embedding learning between audio and video. Our experiment is run over 5-fold cross-validation, where average performance is applied to demonstrate the performance of audio-video cross-modal retrieval. The experimental results achieved on two different audio-visual datasets show the proposed learning architecture with two branches outperforms existing six CCA-based methods and four state-of-the-art based cross-modal retrieval methods.

CCS Concepts: • **Information systems** → **Music retrieval**; Information extraction.

Additional Key Words and Phrases: Deep triplet neural networks, cluster-cca, cross-modal retrieval, triplet loss

## ACM Reference Format:

Donghuo Zeng, Yi Yu, and Keizo Oyama. 2020. Deep Triplet Neural Networks with Cluster-CCA for Audio-Visual Cross-Modal Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1 (March 2020), 23 pages. <https://doi.org/10.1145/3387164>

## 1 INTRODUCTION

With the high-speed development of innovative technology and user interaction on the Internet, various multimedia data and information have been aggregated. It results in a heterogeneous gap between different modality data, which brings a big challenge for efficiently and effectively cross-modal retrieval between data from different modalities. In the past, researches have focused on building the similarity link between every two data points from different modalities for cross-modal

---

Authors' addresses: Donghuo Zeng, National Institute of Informatics, SOKENDAI, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 100-0003, Japan; Yi Yu, National Institute of Informatics, SOKENDAI, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 100-0003, Japan; Keizo Oyama, National Institute of Informatics, SOKENDAI, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 100-0003, Japan.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1551-6857/2020/3-ART \$15.00

<https://doi.org/10.1145/3387164>

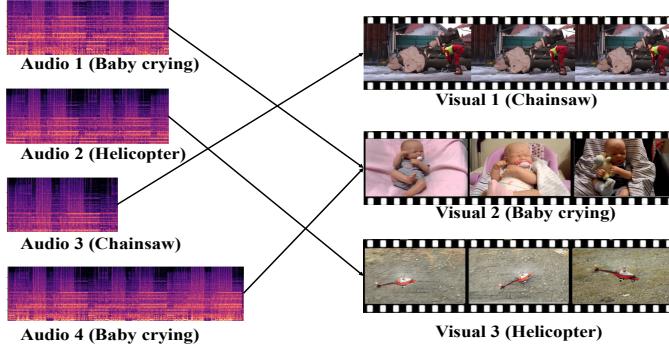


Fig. 1. The overview of audio-to-visual cross-modal retrieval. Given an audio to find similar visual contents, the matched visual contents which have the same semantic label as an audio query. The audio query can be any kind of arbitrary length.

retrieval tasks, which has made big successes in cross-modal retrieval, such as image-text [42, 44], audio-text [56], and video-text [35]. In particular, the visual and auditory senses of human being are the most important ways to understand the living environment and understand the world, seen in Fig. 1. For instance, when hearing a helicopter sound, a helicopter can be imagined in your mind. When you see lightning, subconsciously the thunder is coming soon. Unfortunately, due to the limited audio-video paired dataset and semantic category information, little research works on audio-visual cross-modal retrieval [29]. This motivates us to mimic the mutual-aid based learning process and extract cognitive patterns from human being.

Cross-modal retrieval between every two data points from different modalities has a challenge of the heterogeneous gap of data structure among the modalities, which requires us to formulate a joint representation space, where the similarity of different data modalities reflects the semantic closeness between their corresponding original inputs by correlation learning.

Recently, most methods for correlation learning are to bridge the gap of different modalities by learning joint embedding or representation learning, which has achieved great success in cross-modal retrieval tasks [8, 21, 24, 43]. The purpose of representation learning is to find projections of data examples from a different data set into a shared subspace, where the similarity relationship between them can be measured.

The typical representation learning method CCA [8] is to find linear transformations of two-view of data as inputs via maximizing the pairwise correlation. However, if there is a nonlinear relation between two instances, CCA has no capability to always extract useful features. Kernel-CCA [17] uses the kernel method to CCA, which enables the nonlinear transformation for two-view of data. With the rapid growth of deep neural network (DNN) techniques, the DNN model has been progressively applied in cross-modal retrieval tasks [2, 6, 20, 51]. For example, Deep Canonical Correlation Analysis (DCCA) [2], which is used for learning complex nonlinear transformations of the different datasets. DCCA can learn nonlinear transformations without the inner product computation of Kernel-CCA. Also, DCCA has no hyper-parameters limited in the representation learning unlike kernel-CCA limited in the fixed kernel. The current cross-modal retrieval model also tries to keep the pairwise correlation with the joint predefined semantic categories, where each category contains many pairs of cross-modal data. CCA, Kernel-CCA and DCCA cross-modal retrieval methods focus on the pairwise correlation only. However, the different samples with the same category convey the same semantic information which might be neglected. In theory,

to solve this issue, it requires a model that can preserve all the semantic information during the representation learning, where the heterogeneous gap in the pairwise samples is minimized while non-pairwise samples with the same semantic categories are maximized.

Cluster-CCA [25] can preserve all the semantic information by applying a one-to-one correspondence between all pairs from the cross-modal dataset and use standard CCA to learn the projections. Cluster-CCA can learn joint representations that maximize the correlations between the two different modalities and segregating the different categories in the shared subspace. Cluster-CCA tries to enhance the similarity inside the category between data from different modalities. Inspired by Cluster-CCA and DCCA, an improved C-DCCA[55] is proposed to learn the nonlinear correlation between data from different modalities and simultaneously consider the similarity within the category across modality data. However, the above methods cannot guarantee all the similarity distance of two instances from different modalities of the same category is similar than that of two instances from different modalities of the different categories.

To settle this problem, it needs to completely consider all the positions of data points in the common space. The previous joint embedding learning methods, after the two branch networks are optimized, the CCA projections are calculated only one time. It is impossible to completely focus on the distribution of all the data points in the shared subspace.

To figure out this issue, our first contribution is that deep TNN is proposed to maximize the correlation between every two instances from different modalities with the same category while minimizing the correlation between data from different modalities from different categories during training. In other words, each data point from one modality is more close to samples with the same semantic category from the other modality (namely positive samples). Simultaneously, the data point is farther from instances with different categories. (namely negative samples). The deep TNN used here is to apply deep neural networks with backpropagating errors and use triplet loss to update the weights of the neural network during the training. The second contribution is that all the data points within a batch size is considered to meet storage limitation instead of using all the position of data points space. Finally, our architecture is evaluated on two video datasets. MV-10K dataset is selected from the YouTube-8M video dataset by us, which is utilized in our previous work [57]. To evaluate the extendability of our algorithms, VEGAS dataset [60] is used in the experiments. The experimental results demonstrate that the proposed embedding learning architecture significantly surpasses the existing six CCA-based methods and four state-of-the-art methods in cross-modal retrieval.

The rest of this paper is organized as follows. In Section 2, we show some works related to our proposed approach. Section 3 presents our proposed TNN-C-CCA model. Section 4 introduces the experimental results of our approach achieved on two audio-visual cross-modal datasets with the analyses. Finally, Section 5 makes a conclusion of this paper.

## 2 RELATED WORK

In this section, we briefly present two main research lines related to our method, including cross-modal retrieval and triplet neural networks.

### 2.1 Cross-modal Retrieval

Different from retrieval in the same modality, such as image retrieval [45], cross-modal retrieval is used for implementing a retrieval task across different modalities. such as image-text[15, 42, 44, 51], video-text[35], and audio-text[56] cross-modal retrieval. The main challenge of cross-modal retrieval is the modality gap and the key solution of cross-modal retrieval is learning joint embedding for different modalities. Learning joint embedding is not only a solution of cross-modal retrieval and also applied for other multimedia tasks, such as image classification [52], video question and

answering [61]. As for our task, cross-modal retrieval aim at generating new representations from different modalities in the shared subspace, such that new generated features can be applied in the computation of distance metrics, such as Cosine distance and Euclidean distance. Generally, the output space of the architecture after training can be real-valued common space and binary representation learning space.

### 2.1.1 Real-valued common space.

In this paper, we focus on common space learned by the real-valued representation learning for cross-modal retrieval. This kind of representation learning can be classified into three categories as follows.

*Unsupervised methods.* Canonical correlation analysis (CCA) is one of the most popular cross-modal embedding models, which aims at finding a pair of linear transformations to maximize the correlation between two different modalities. The work [26] uses CCA to calculate the cross-modal correlations between image and text. A novel method for a cross-modal association is Cross-modal Factor Analysis (CFA) [18], which is used for audio-image cross-modal retrieval task. Another way to reduce the dimension is to find unified feature subspace by a principle of collective component analysis (CoCA) [34], where two different modality data points should be correspondence among the projections in the shared subspace and the similarity between the paired data points should be maximized. The paper [42] proposed a cross-modal projection matching (CMPM) loss and projection classification loss for learning the discernible different embeddings space. The CMPM loss minimizes the KL divergence between each pair and groups the new representations into different clusters.

*Semi-supervised methods.* Different from the unsupervised methods computing the correlation of projections without any category information during the training, semi-supervised methods solve the problem of limited labels dataset. The approach MVML-GL[46] with semi-supervised learning is to reveal the latent feature space by keeping global consistency structure and local geometric architecture. The paper[36] proposed a semi-supervised model which is suitable for few labeled images and large unaligned textual documents to locate image regions to texts. GSS-SL[31] is a semi-unsupervised method, which is via predicting more related labels for unlabeled data with label graph constraint and the labels directly are regarded as the semantic information of multimedia data. The paper [53] proposed LRGA ranking method and long-term RF algorithm to learn multimedia data representation by exploiting the history of RF information and multimedia data distribution by users.

*Supervised methods.* During the correlation learning of cross-modal retrieval, the semantic class information can be applied for the similarity learning. GAM [30] is an extension of the canonical correlation analysis method, which focuses on keeping the function of popular supervised and unsupervised feature extraction approaches by developing a quadratic program to get a single nonlinear subspace over different feature spaces. This paper [15] proposed a model focusing on a finer level and fragment of both images and sentences in the shared subspace. In this paper [40], they utilized deep CNN features as the visual inputs and topic features as textual semantic inputs. They proposed a regularized deep neural network for nonlinear semantic correlation across modalities. They put forward the intra-modal regularization to learn joint embeddings with Intra- and Inter-modal relation. Topic modeling on Latent Dirichlet Allocation (LDA) is always applied in multi-modal data, especially for the Document Neural Autoregressive Distribution Estimator (DocNADE) which is the best for document modeling. In this paper [59], they put forward a method called SupDocNADE which is a supervised improvement method of DocNADE. This method enhances the discriminative power of topic features. CM-GANs [22] learns discriminative joint embedding

to bridge the heterogeneity gap. Zero-shot can be regarded as an extension of supervised learning, the TANSS [49] model applied zero-shot to obtain a self-supervised semantic sub-network that can enhance sub-network generalization for unseen labels.

### 2.1.2 Binary representation learning space.

As the huge multimodal data increase on the Internet, it is becoming difficult to satisfy the requirement of storage and retrieval capabilities over a big cross-modal multimedia dataset. A solution to speeding up the search is cross-modal hashing [13, 27], which is widely employed to learn binary representations in the common space. Cross-modal hashing methods try to learn correlations between two modalities of data, which can improve the accuracy of cross-modal search by projecting the data into a common Hamming space. However, hashing code generated by traditional methods [27] is less discriminative to different categories. The DCH [50] model directly learns discriminative hashing code by discriminative classification. The MDBE [41] model learns discriminative hash code by preserving discriminability with classification method and keeping similarity with shared structure among the data from different modalities. The CRE [11] model proposes different modality-specific models to bridge heterogeneity gap, meanwhile, projecting different modalities of data into Hamming space to reconstruct embedding, which reduces the optimization complexity and preserves the inter-category information. The paper [48] aims at overcoming the loss information of quantization in hashing code learning, by rotating the joint space to produce better unified hashing codes.

Using common space learning methods usually to preserve the characteristics of original feature explicitly is a challenging issue. Attention mechanism in DNN [23] is used to construct modality-specific semantic space, which is directly generated without common space learning. Adversarial learning based model DAML [47] not only learns latent feature subspace through minimizing the intra-category correlations and maximizing the inter-category correlations, but also proposed a modality classifier to guarantee the output embedding is also statistically indistinguishable.

## 2.2 Triplet Neural Network

Triplet neural network model is the extension of the Siamese Neural Network (SNN) [16], SNN has typically succeeded in object tracking [3], image recognition [16], person re-identification [10, 33], and face recognition [14, 28] tasks. SNN consists of twin networks, and the network is symmetric, which tries to employ a unique network to rank similarity between two different modality inputs. This system can generate discriminative features and enhance the generalization power of the network. Triplet neural network consists of three networks, where the weights of them are updated by the triplet loss. During the training, the triplet loss requires that these anchor samples are more similar to positive samples than to negative samples by a hyper-parameter margin.

Some ranking based approaches, it is effective for the representation learning in the cross-modal retrieval tasks. ACMR [39] model applies triplet loss method to keep the intra-modal distinguishable and inter-modal unchangeable in the feature projector, which minimizes the gap between all samples of different modalities with the same semantic labels, in the meanwhile, maximizing the distances of samples belonging to different labels. Methods with triplet loss in cross-modal retrieval can better preserve the representation structure of images and texts when projected into the shared subspace by means of the triplet constraint in the new representation generation by triplet loss. Deep image-text embedding [44] is trained by applying a large margin objective function with cross-modal ranking constraints. Instead of computing on all triplets, they sampled triplets in each mini-batch by selecting top K most hard matches according to the similarity distance and exploit Stochastic Gradient Descent (SGD) to minimize the loss function. The average of the loss of all batches is used. Learning ranking functions during the training of cross-modal retrieval gets popular in recent

researches, which has been regarded as a fundamental problem. In this paper [54], they proposed a ranking canonical correlation analysis (RCCA) method to solve the limitation of two paradigms. The image vector model highly requires the quality of textual description and the manual labeling is hard to obtain. Through establishing triplet from click-through data to optimize the objective function, the triplet consists of a query, a higher clicked image, and a lower clicked image. In order to preserve the relations in the triplets, they build a triplet loss to optimize the weights. To better understand and model the networks for connecting the information of cross-views in the shared spaces, the proposed structure-preserving metric model (SPML) [32] learns a Mahalanobis distance metric so that the structure of the network is preserved by a hinge-loss over triplets, which consists of a node, its non-neighbor node, and its neighbor node. They apply stochastic subgradient descent to optimize the triplet loss function and sample a batch of triplets to train.

Nevertheless, these models for learning embeddings to preserve the joint relation of examples are not designed for audio-visual cross-modal retrieval. In this work, we propose a deep triplet neural network to optimize the embedding generated from linear cluster-CCA and simultaneously solve the problem that the distance of similar samples is farther than the distance of dissimilar samples during the training.

Table 1. Configuration of TNN-C-CCA

log mel-spectrogram audio inputs	96x64
Output of visual branch	$L^1 \times 1024$
Output of audio branch	$L^2 \times 128$
Output of Cluster-CCA	10
Fully connected layers for audio	[100, 100, 100, 10]
Fully connected layers for visual	[200, 200, 200, 10]
Output of TNN-C-CCA	10

### 3 ARCHITECTURE

Our deep architecture generally can be divided into two different parts: feature extraction and TNN-C-CCA training, as shown in Fig. 2. The configuration of TNN-C-CCA used in this work is shown in Table 1. Outputs of visual branch and audio branch respectively are 1024-dimensional and 128-dimensional, which are mapped to 10-dimensional by cluster-CCA. Deep triplet neural network consists of 4 fully connected layers respectively for audio embedding and visual embedding and outputs a feature vector with a size of 10. The motivation of our architecture is to take advantage of the two models. Cluster-CCA is to establish one-to-one correspondences between all possible pairs by given categories information across the two modalities to maximize the correlation between the latent representation of two different modalities via CCA. Deep TNN aims to enforce the relation of similar samples and simultaneously weaken the relation of dissimilar samples. Particularly, using more negative samples and positive samples during the training of Deep TNN improves the discriminative capability of the embedding space.

#### 3.1 Distance Metric

In previous face identity tasks [10, 28], they used Euclidean distance  $D(T(i) - T(j)) = \|T(i) - T(j)\|_2^2$  to calculate the distance between the image anchor and the positive image or the negative image, where  $i$  and  $j$  are from the same modality of image. However, in our experiment, we apply a cosine

<sup>1</sup>L is the number of frames in a video, by decoding each video at one frame per second.

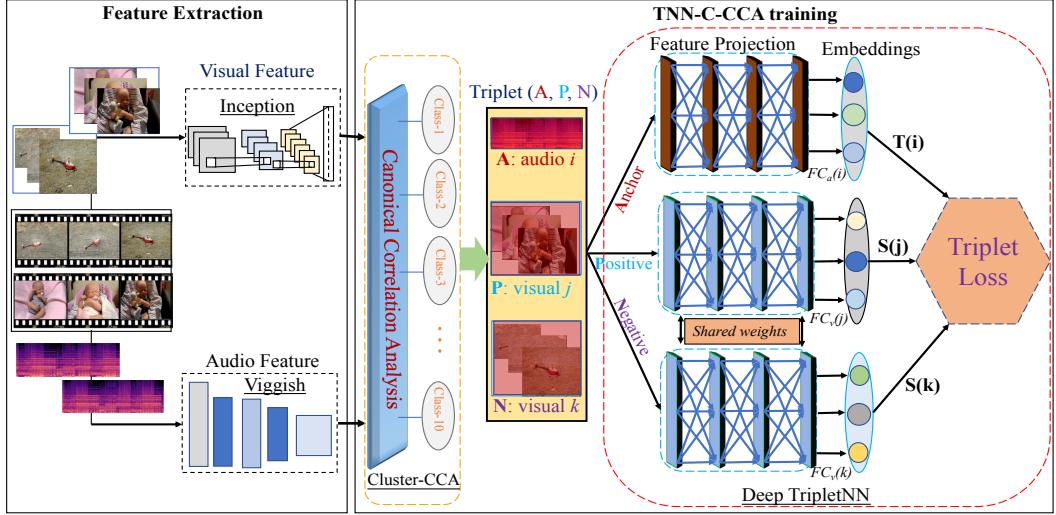


Fig. 2. The overall framework of our TNN-C-CCA model. It consists of two parts: feature extraction and TNN-C-CCA training. We apply Inception V3 and Vggish model to extract feature, then explore cluster-CCA to learn the correlation with cluster segregating and select triplets as input for deep TNN training. In the deep TNN, there are three branches: anchor, positive, and negative. Positive and negative branches shared the same weights. Anchor branch is trained by audio data, positive and negative branches are trained by visual data. The detailed description is shown in section 3.3.

similarity for the final representation comparison at the end of the whole architecture. Our distance metric can be defined as follows:

$$\|x, y\|_{cosine-distance} = 1 - \frac{\sum_{k=1}^n x_k y_k}{\sqrt{\sum_{k=1}^n x_k} \sqrt{\sum_{k=1}^n y_k}}, \quad (1)$$

where  $n$  is the dimension of vector  $x$  and  $y$ , its iteration  $k$  ranges from 1 to  $n$ . The scale of the Cosine distance ranges from 0 to 2 and the effective margin shares the same scale, normally it is set to 0.5.

### 3.2 Input Feature Representation

The audio feature is extracted by a pre-trained VGGish model with the Mel spectrogram feature as input. We apply the librosa<sup>2</sup> library to achieve Mel spectrogram feature extraction with default parameters: hop size=512, nfft=2,048. We use mel-frequency cepstrum (MFC) to do sound processing by applying linear cosine transform of a log power spectrum to represent the short term power spectrum of audio.

Deep learning pre-trained audio feature extractors can be divided into two categories: training with audio label and training without audio label.

We choose the VGGish model training with audio label which can capture the label information in the final feature obtained and is suitable for our dataset with labels. VGGish model is a VGG-like model, which is trained on a large-scale dataset named AudioSet for the audio classification task. Compared with the VGG model, the VGGish model changes the input size and cuts the last group

<sup>2</sup><https://librosa.github.io/librosa/>

of convolutional layers and max pool layers, and uses a 128-wide fully connection layer at the end of the neural network. The inputs of audio features are re-sampled to 16kHZ. The window size of FFT is 25ms, with a window hop of 10ms.

VGGish model converts the audio feature inputs into 128-D semantic high-level feature, which is used for audio-visual cross-modal retrieval.

Visual feature is extracted by the state-of-the-art deep CNN model named Inception V3 [37, 38]. The Inception V3 model <sup>3</sup> is popularly-used in the image recognition task which can reach high accuracy on the ImageNet dataset [5, 12]. Recently, the pre-trained Inception V3 model is exploited as a video feature extractor [1]. The output of the Inception V3 model is frame-level features along with the input of pre-processing videos. By decoding each video at one frame per second, these decoded videos are fed to the Inception V3 network and adopt the ReLU activation in the last hidden layer before the prediction layer. The feature representation is 2,048 dimensional per frame of videos and keeps the front 360 frames. After that, PCA approach is applied to reduce the dimensions per frame to 1,024 by using the quantization method.

Finally, we apply the general semantic audio and visual features as the input of our TNN-C-CCA architecture, the input vector is the global average of frame-level feature which is computed by the number of frames, so that the input audio feature is 128 dimensions and the visual feature is 1024 dimensions.

### 3.3 Cluster-CCA

CCA is used for exploring the relationship between two multivariate sets of vectors, such as  $x \in R^A$  and  $y \in R^B$  with zero-mean, and the pair format is like  $(x_i, y_i)$ . The goal of CCA is to find a new coordinate for  $x$  and  $y$  by direction  $w \in R^A$  and  $u \in R^B$  respectively, such that the correlation between these two sets is maximized. The correlation can be defined as follows:

$$\text{corr} = \frac{w' C_{xy} u}{\sqrt{w' C_{xx} w} \sqrt{u' C_{yy} u}}, \quad (2)$$

$$C_{xx} = E[xx^T] = \frac{1}{n} \sum_{i=1}^n x_i x_i^T, \quad C_{yy} = E[yy^T] = \frac{1}{n} \sum_{i=1}^n y_i y_i^T, \quad C_{xy} = E[xy^T] = \frac{1}{n} \sum_{i=1}^n x_i y_i^T, \quad (3)$$

Where  $\text{corr}$  is the correlation,  $C_{xx}$ , and  $C_{yy}$  are the co-variance metrics,  $C_{xy}$  is the cross-variance metrics. Here  $E(*)$  is the expectation function. Normally, the problem is regarded as an eigenvalue problem, suppose  $w$  is the top eigenvector, the problem can be represented as follows:

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} w = \lambda^2 w, \quad (4)$$

CCA has been successfully applied to several multimedia problems, such as cross-modal retrieval. However, CCA is suitable for calculating pairwise correlation similarity from different modalities and not available for calculating correlation similarity within a cluster. CCA will be ineffective for learning representation with a cluster in this case. Cluster-CCA is a variant of CCA [8] with consideration of the cluster segregating by establishing one-to-one correspondences from all pairs of data points in a given cluster across the two different modalities, then apply CCA to learn the projections.

$$\text{corr} = \frac{w' C'_{xy} u}{\sqrt{w' C'_{xx} w} \sqrt{u' C'_{yy} u}}, \quad (5)$$

---

<sup>3</sup>[https://github.com/google/youtube-8m/tree/maste /feature\\_extractor](https://github.com/google/youtube-8m/tree/maste /feature_extractor)

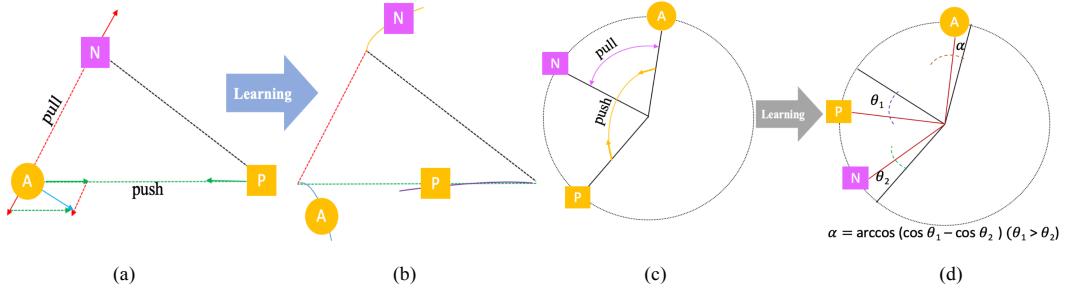


Fig. 3. (a) and (b) show the traditional triplet loss minimizes the Euclidean distance between (anchor, positive) and (anchor, negative) with a fixed margin and optimal gradient back-propagation direction; (c) and (d) present our triplet loss through minimizing the Cosine distance between (anchor, positive) and (anchor, negative) with fixed margin and optimal gradient back-propagation direction.

The three types of variances can be formulated as follows:

$$C'_{xx} = \frac{1}{L} \sum_{c=1}^C \sum_{i=1}^{|X_c|} |Y_c| x_i^c x_i^{cT}, \quad C'_{yy} = \frac{1}{L} \sum_{c=1}^C \sum_{j=1}^{|Y_c|} |X_c| y_j^c y_j^{cT}, \quad C'_{xy} = \frac{1}{L} \sum_{c=1}^C \sum_{i=1}^{|X_c|} \sum_{j=1}^{|Y_c|} x_i^c y_j^{cT}, \quad (6)$$

Where  $L = \sum_{c=1}^C |X_c||Y_c|$  is the sum number of all pairs. Similar to CCA, the optimization problem can be regarded as an eigenvalue problem like formulation (4). Here we assume that the covariance is calculated for the zero-mean random variables.

### 3.4 Deep Triplet Neural Network

The Deep Triplet Neural Network is an end-to-end training, as shown in Fig. 2, which is optimized by triplet loss [28] at the end of cross-modal retrieval architecture. For example, in audio-to-visual retrieval process, we try to obtain an audio  $i$  represented by  $T(i)$  and a visual  $j$  represented by  $S(j)$ , a visual  $k$  ( $k \neq i$ ) represented by  $S(k)$ , where  $T(\cdot)$  and  $S(\cdot)$  are the output of Cluster-CCA model,  $i$  and  $j$  from the same category,  $i$  and  $k$  from different categories. Here we want to guarantee audio sample  $i$  (Anchor) of one specific category is closer to visual sample  $j$  (Positive) of the same category than any visual sample  $k$  (Negative) of any other category. As shown in Fig. 3. Triplet loss will pull Anchor and Positive samples, simultaneously push Anchor and Negative samples. The condition is represented as follows.

$$\begin{aligned} & ||T(i) - S(j)||_{\text{cosine-distance}} + \alpha < ||T(i) - S(k)||_{\text{cosine-distance}}, \\ & Lab^i = Lab^j, \quad Lab^i \neq Lab^k (i \neq k), \quad \forall (i, j, k) \in \Lambda, \end{aligned} \quad (7)$$

Where  $\alpha$  is a margin that is used for reinforcing the Cosine distance among anchor, positive and negative.  $\Lambda$  is the collection of all possible triplets in the training dataset. The triplet loss can be defined as follows:

$$\text{Loss} = \text{Max}\left\{\sum_i^N [|T(i) - S(j)|_{\text{cosine-distance}} - ||T(i) - S(k)||_{\text{cosine-distance}} + \alpha], 0\right\}, \quad (8)$$

Where  $N$  is the sum of all possible triplets. The collection of all the possible triplets is generated by the output of Cluster-CCA model, it is easy to fulfill the condition defined in Eq.(7), because the new audio/visual representations have already learned pairwise-based correlation and cluster-based correlation which results in almost pairwise examples of the same class group more closer

than the pairwise example from different classes. The triplet loss values of most triplets are zero and these triplets have no contribution to the sum of triplet loss, which lead to the final average of loss values close to zero. In particular, when a loss has  $\|T(i) - S(j)\|_{cosine-distance} + \alpha < \|T(i) - S(k)\|_{cosine-distance}$ , it is equal to zero, the loss has no contribution to optimizing the final loss. Our experiment follows [10], a better triplet loss optimization is to ignore all the triplet when its loss is zero, so that the triplet loss can be fast converged and the optimization will be more effective [28].

It is impossible for us to calculate all the argmin and argmax among all the training dataset. Because in our experiment dataset, we have around 1K examples in MV-10K dataset and more than 2K examples in the VEGAS dataset for each class, which result in a large number of possible triplets. And computation in this way may bring bad generation and over-fitting. In this paper, we follow the FaceNet method [28] and select triplets to remove all negative/positive samples in a batch when its triplet loss is zero.

## 4 EXPERIMENTS

### 4.1 Dataset and Evaluation Metric

We evaluate our model on two different video datasets: VEGAS [60] and MV-10K [57]. Each video from these datasets contains audio track and visual track, and both data can be represented with high-level features. Our goal is to learn the correlation between this two-view of high-level features. We adopted Mean Average Precision (MAP) and Precision-Recall Curve (PRC) for quantitative performance evaluation.

MV-10K Dataset refers to our previous work [57]. This dataset is a small subset of large-scale video dataset YouTube-8M<sup>4</sup> which contains 10,000 (10k) videos with a "music video" label, and each video ranging from 213 to 219 seconds. Audio and visual track features are extracted in the MV-10K dataset. The YouTube-8M dataset has already released the audio and visual features, respectively extracted by VGGish model trained on Audioset<sup>5</sup>, and pre-trained Inception V3 model trained on ImageNet dataset<sup>6</sup>. Based on the frame-level audio features, we applied 10 pre-defined music categories to annotate all videos, we assume these videos have certain knowledge with the music categories and audio-visual pairs shared a single music category.

The VEGAS dataset[60] selected videos from Google Audioset by 10 categories and applied Amazon Mechanical Turk to do data cleaning, the 10 categories are human/animal sounds (chainsaw, helicopter, drum, printers, fireworks, dog, rail transport, baby crying, human snoring, water flowing and rail transport). The length of a video ranges from 2 to 10 seconds and the average is 7 seconds, the percentage of all video which ranges from 8 to 10 seconds is above 55%. In our experiments, we use 28,103 videos to evaluate our architecture.

**Evaluation Metric.** In our work, we use MAP and PRC as metrics to leverage our architecture. We focus on category-based cross-modal retrieval, where the system generates a ranked list of documents in one modality by a query in another modality. The samples with the same category as that of the documents are regarded as relevant. Moreover, it takes the location of retrieved document in the rank list into account. The more related documents appear in the top rank list, the higher MAP value it has.

### 4.2 Training Setting

In our experiments, we set parameters for our deep TNN-C-CCA model as follows.

---

<sup>4</sup><https://research.google.com/youtube8m/>

<sup>5</sup><https://research.google.com/audioset/>

<sup>6</sup><http://www.image-net.org/>

- (1) For deep TNN, there are three branches: anchor branch, positive branch, and negative branch. For each branch, they will go through a full connection. Anchor branch has its own parameters, positive and negative branches share the same parameters. When taking audio sample as an anchor, the positive and negative are visual samples. We set four hidden layers for each full connection. The number of units per layer is respectively set to 100, 100, 100, 10 for audio branch and 200, 200, 200, 10 for visual branch. If taking visual as the anchor, the positive and negative samples are from audio samples. We set the number of units per layer for visual branch to 200, 200, 200, 10, and 100, 100, 100, 10 for audio branch.
- (2) We set the correlation component for all the following experiments as 10. We set the probability of dropout as 0.2 and use *tanh* as activation function for each hidden layer and use *sigmoid* as the activation function in the last layer.
- (3) We separately divided the training set ranges from 300 to 1,000, and select the best one. The number of training epochs is 20.
- (4) Our result is the average performance via 5-fold cross-validation. We consider the category balance when we evenly group all the dataset into 5 folds.
- (5) The Adam optimizer is used for our experiment. The learning rate is set as 0.001.

### 4.3 Results on the VEGAS Dataset

We report the result of audio-visual cross-modal retrieval task on the VEGAS dataset in the left part of Table 2 with MAP metric and Fig. 4 with PRC. We implement our architecture compared with some existing CCA-variant approaches and non-CCA methods: CCA [8], DCCA [2], KCCA [17], C-CCA [25], C-KCCA [25] C-DCCA [55], AGAH [7] and etc. as baselines, to show the improvement of our model. For these baselines, we separately implement all of them with the same dimension of outputs and the same parameters.

Table 2. The MAP scores of cross-modal retrieval between audio and visual contents for our TNN-C-CCA method and some existing state-of-the-art methods on VEGAS dataset and MV-10K dataset.

Models	VEGAS Dataset (%)		MV-10K Dataset (%)	
	audio2visual	visual2audio	audio2visual	visual2audio
CCA [8]	32.43	32.11	18.38	18.17
KCCA [17]	28.65	27.24	17.81	17.03
DCCA [2]	41.43	42.15	18.43	18.21
C-CCA [25]	65.16	64.35	19.71	19.62
C-KCCA [25]	32.41	32.74	18.38	18.11
C-DCCA [55]	70.34	69.27	21.79	20.08
UGACH [58]	17.18	17.07	11.11	11.40
AGAH [7]	57.82	56.16	20.74	20.19
UCAL [9]	42.68	41.53	18.82	18.47
ACMR [39]	45.46	43.12	19.02	18.63
LSTM_C_CCA	66.62	71.34	19.11	18.89
TNN-C-CCA	74.66	73.77	23.34	21.32

According to the experience of our experiments, when the correlation component is set to 10, the CCA-variant approaches can get the best performance[55, 56]. Here we use the MAP value as our main performance metric, the MAP of 10 correlation components is much better than the other number of ten multiples correlation components. We set the dimension of outputs of all baselines as 10. The dimensions of the audio feature as inputs are  $L * 128 (L \in [2, 10])$ , the dimensions of visual

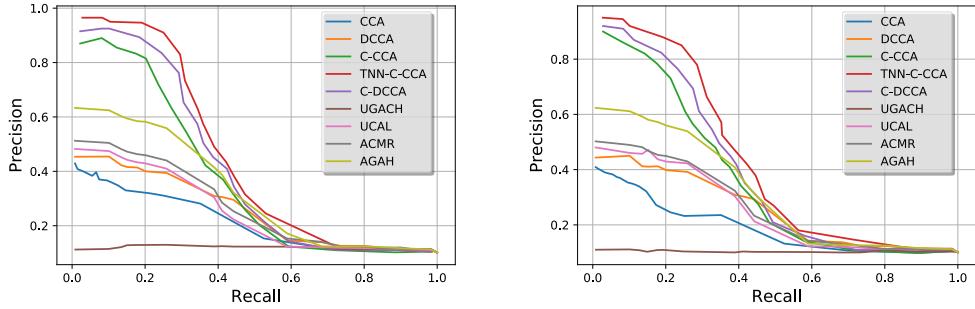


Fig. 4. The PRC achieved on the VEGAS dataset with our TNN-C-CCA model and other eight different models. The left figure is for audio-to-visual retrieval, the right figure is for visual-to-audio retrieval.

feature as inputs are  $L * 1024 (L \in [2, 10])$ . For each audio-visual pairwise,  $L$  for the audio and the visual are the same. Then feed them into a mean layer to make all the audios and all the visual samples respectively have the same dimensions, to make it possible to calculate the correlation in the shared space with CCA-variant approaches. Especially, the DCCA and the C-DCCA have the same structures of hidden layers. We did all the experiments for each model with 5-fold cross-validation. All models were done by the same structure of folds and the structure established considers balance factor. Each fold contains the same number of samples in each category and 10 categories are kept simultaneously in each fold. Table 2 shows that all CCA variants with category information as training such as C-CCA, C-KCCA, LSTM-C-CCA, and C-DCCA are much better than training without any class as inputs such as CCA, DCCA, and KCCA. The best performance without category information training is DCCA. The MAP of audio-to-visual retrieval is 41.43% and the MAP of visual-to-audio is 42.15% over VEGAS dataset, which outperforms the CCA method: the MAP of audio-to-visual retrieval is 32.43% and the MAP of visual-to-audio retrieval is 32.11%, and are much better than the KCCA method: the MAP of audio-to-visual retrieval is 28.65% and the MAP of visual-to-audio is 27.24%. Compared with the above unsupervised CCA-variant method, the supervised CCA variants can get higher MAP performance. Taking C-CCA as an example, the MAP of audio-to-visual retrieval is 65.16% which has 23.63% improvement and the MAP of visual-to-audio retrieval is 64.35% which has 22.20% improvement. C-DCCA not only considers the pairwise correlation but also learns the category-based similarity correlation by enlarging the number of pairs if the paired data points have the same category information. In our experiment with this dataset, we establish new possible pairs within the same category for each sample in the train set, then select 50% pairs for each sample to enlarge the train set. There are three main shortages of C-DCCA method: 1) because it deeply relies on the balance of pairwise correlation and category-based correlation which is adjusted by a hyper-parameter  $\beta$ , it is very hard to set the best  $\beta$  during the training. 2) when we do model generation for new dataset input, the method can not reduce the noisy pairs which belong to the paired data from other categories closer than the paired data from its category. 3) it is really time-consuming and space-consuming during the training.

To overcome three shortages, we put forward TNN-C-CCA model with the aim of learning a more reliable correlation in the common space and learning better new joint embeddings for each modality to compute the similarity. Table 3 shows that our TNN-C-CCA model can get a MAP of 65.62% for audio-to-visual retrieval and the MAP of 63.30% for visual-to-audio retrieval by randomly selecting the 150 negative samples for each anchor in the training set. Compared

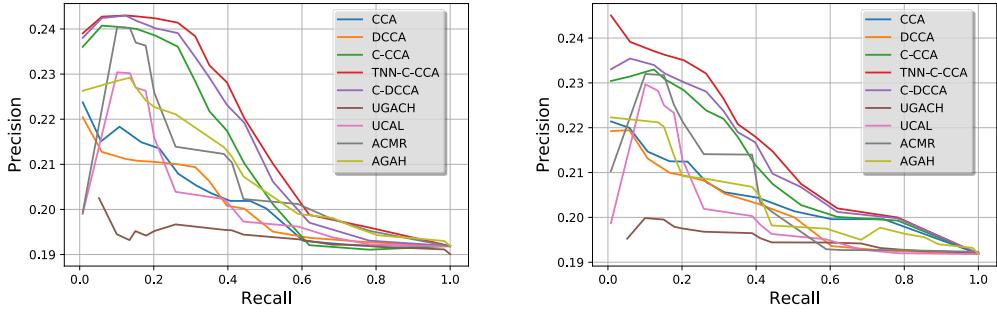


Fig. 5. The PRC achieved on the MV-10K dataset with nine different models. The left figure is for audio-to-visual retrieval, the right figure is for visual-to-audio retrieval.

with Cluster-CCA without considering negative information, the Map of visual-to-audio retrieval is improved by around 7%. However, randomly selecting the negative samples are not statistical reliability, which brings trouble for re-implementing the experiments to get the same result. In theory, we hope to consider all the negative samples, but in fact, for each sample, there almost have 16,800 negative samples and exist  $N^2$  ( $N$  is the size of the training set.) training samples, it is the time- and space- consuming in the case of TNN-C-CCA. In order to balance the time- and space-consuming, and consider the negative samples, according to these works [10, 19, 28], we build triplets (anchor, positive and negative) inside a batch for training. If the size of the training set is  $N$  and the number of the batch is  $B$ , the batch size is the floor of  $N/B$ . The samples of all categories balance in each batch. In each batch, there are  $\sum_{i=1}^{10} \frac{N_i^2(N-N_i)}{B^3}$  triplets, and the training set size is  $\sum_{i=1}^{10} \frac{N_i^2(N-N_i)}{B^2}$ . Established triplets in a batch, it can save  $\sum_{i=1}^{10} \frac{N^2}{B^2}(N-N_i)(B^2-1)$  training set compared with building all triplets one time, where the  $N_i$  is the number of pairs with class  $i$  in train set. And the performance is much better than that of the C-DCCA and other baselines, the MAP of audio-to-visual retrieval is 74.66% which has 4.28% improvement compared to C-DCCA model. The MAP of visual-to-audio retrieval is 73.77% which has 4.5% improvement compared to C-DCCA model. In addition, we compare TNN-C-CCA model with four state-of-the-art cross-modal retrieval methods. As shown in Table 2, the performance of our TNN-C-CCA model is much better than that of novel adversarial learning methods.

Table 3. The MAP scores of audio-visual cross-modal retrieval for our TNN variant methods

Models	VEGAS Dataset		MV-10K Dataset	
	Audio2Visual	Visual2Audio	Audio2Visual	Visual2Audio
C-CCA [25]	65.16	64.35	19.71	19.62
TNN (batch all)	14.18	13.44	13.25	14.02
TNN (batch semi-hard)	15.18	14.22	14.20	14.17
TNN (batch hard)	11.18	12.20	12.06	11.59
TNN-C-CCA (rand)	65.62	63.30	19.23	18.74
TNN-C-CCA (batch semi-hard)	71.35	70.23	20.37	19.97
TNN-C-CCA (batch hard)	60.71	58.39	19.16	18.85
TNN-C-CCA (batch all)	74.66	73.77	23.34	21.32

#### 4.4 Results on the MV-10K Dataset

We report the result of audio-visual cross-modal retrieval on the MV-10K dataset in Table 2 with MAP metric and Fig. 5 with the PRC. We compare our model with some previous models published in [57]. For those models, where the results of audio-visual retrieval are calculated. Based on the previous works, we use the same input features that are used in all models. In Table 3, the TNN-C-CCA (rand) model is achieved by selecting the negative and positive in the training set by random to build the triplet as inputs after obtaining the embedding in the common space with Cluster-CCA method. In the experiment, we randomly select 150 triplets for each sample during the training, as shown in Table 3. Because it is very hard to select the triplet for each sample. Since it is time-consuming to use all the possible triplets, we select all the triplets within a batch. For audio-to-visual retrieval as shown in Table 2, our model gets the improvement of 1.55% for MAP and 1.24% improved for visual-to-audio retrieval task compared with the state-of-the-art model C-DCCA, and the performance of proposed method is much higher than the state-of-the-art non-CCA models: UGACH, AGAH, UCAL and ACMR model.

In Table 2, Fig. 4 and Fig. 5, it is easy to notice that the MAP of VEGAS Dataset is much better than that of MV-10K Dataset. Two main reasons are explained as follows.

- (1) The supervised cross-modal retrieval deeply depends on the accuracy of the label for the samples. In the MV-10K Dataset, the labels are allocated by the feature similarity. It is hard to guarantee the allocated labels are always correct. There exist many noisy labels in this dataset. However, the VEGAS Dataset is annotated by volunteers and the labels are double-checked. The label can accurately reflect the semantic information in both audio and visual modalities.
- (2) Moreover, video in the MV-10K Dataset is about 216 seconds while the VEGAS dataset is 10 seconds or less. The input of our model is high-level features, this kind of feature is more effective for the short length of the video in this case. Because high-level semantic features will filter that unimportant information. We use the same dimension to represent those two datasets, in general, which leads to long videos losing more information than short videos.

#### 4.5 Ablation Study of TNN-C-CCA

To have a good ablation study, we investigate triplet selection for the inputs of TNN model to see how it influences the performance of TNN-C-CCA architecture. We also study the impact of distance using in triplet loss of TNN-C-CCA. Then, we show the visualization of the learned semantic space and display the visualization of retrieval results according to the given audio query. In addition, we discuss the effect of model parameters.

##### 4.5.1 Triplet selection strategies.

According to the relationship between anchor-positive distance and anchor-negative distance, triplets can be divided into three categories. In other words, under the fixed anchor-positive distance, negative samples can be categorized into three classes: easy negative, hard negative and semi negative, as shown in Fig. 6. During the training, the triplet selection for training TNN-C-CCA model is a very important part. We introduce three triplet selection strategies: batch all when selecting all triplets as training, batch hard when selecting one hard negative-based triplet as training, batch semi-hard when selecting all semi-hard as training.

Table 3 shows the MAP scores of audio-visual cross-modal retrieval with three triplet selections strategies that are used for training for TNN and TNN-C-CCA. In TNN model which uses original audio-visual features as input, batch semi-hard as training can achieve the best performance for audio-visual retrieval. However, in TNN-C-CCA model, batch all can obtain the best performance.

On the other hand, it is obviously that C-CCA with TNN embedding is much better than C-CCA embedding and TNN embedding respectively, the best TNN model (batch semi-hard), which can

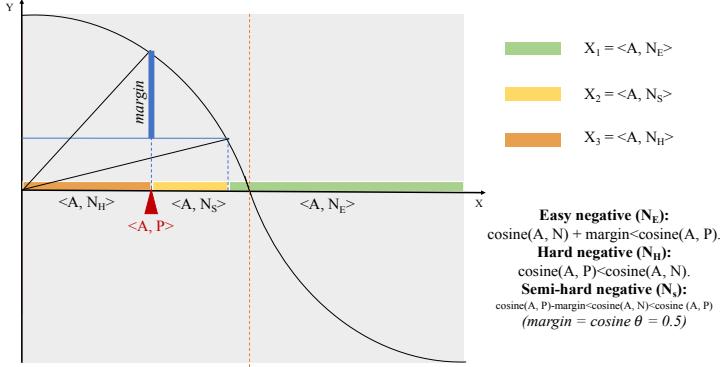


Fig. 6. Given an Anchor-Positive pair with its angle  $\angle A, P$ , those negative samples having the same modality with Anchor as Positive and having different label as Positive, based on the relationship between  $\cosine(A, P)$  and  $\cosine(A, N)$ , can be classified into three categories: 1) Easy negative, 2) Hard negative and 3) Semi-hard negative.

achieve MAP of 15.18% for audio-to-visual retrieval and MAP of 14.22% for visual-to-audio on VEGAS dataset, MAP of 14.02% for audio-to-visual retrieval and MAP of 14.17% for visual-to-audio on MV-10K dataset. The TNN-C-CCA (batch hard) can obtain MAP of 60.71% for audio-to-visual retrieval and MAP of 58.39% for visual-to-audio on VEGAS dataset, MAP of 19.16% for audio-to-visual retrieval and MAP of 18.85% for visual-to-audio on MV-10K dataset. From these results, we can observe that the proposed TNN-C-CCA model gets a significant improvement comparing with C-CCA embedding.

Table 4. MAP with respect to Euclidean distance and Cosine distance in TNN-C-CCA model

Distances	audio-visual	visual-audio
<b>Euclidean distance</b>	0.5300	0.4206
<b>Cosine distance</b>	0.7466	0.7377

#### 4.5.2 Distance metrics in triplet loss.

To examine the effectiveness of the distances applied in the triplet loss of TNN-C-CCA model, we briefly introduce the Euclidean distance as follows:

$$\|X, Y\|_{\text{euclidean-distance}} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}, \quad (9)$$

where  $X = (x_1, x_2, \dots, x_n)$  and  $Y = (y_1, y_2, \dots, y_n)$  are two points in Euclidean n-space with Cartesian coordinates.

Then, we compared Euclidean distance with Cosine distance in triplet loss of TNN-C-CCA model. Table 4 shows the results on VEGAS dataset, which demonstrates Cosine distance is much better than Euclidean distance. In particular, the MAP score is significantly improved. Euclidean distance value is unlimited which may lead to the triplet loss is too large during the training and it is hard to be converged.

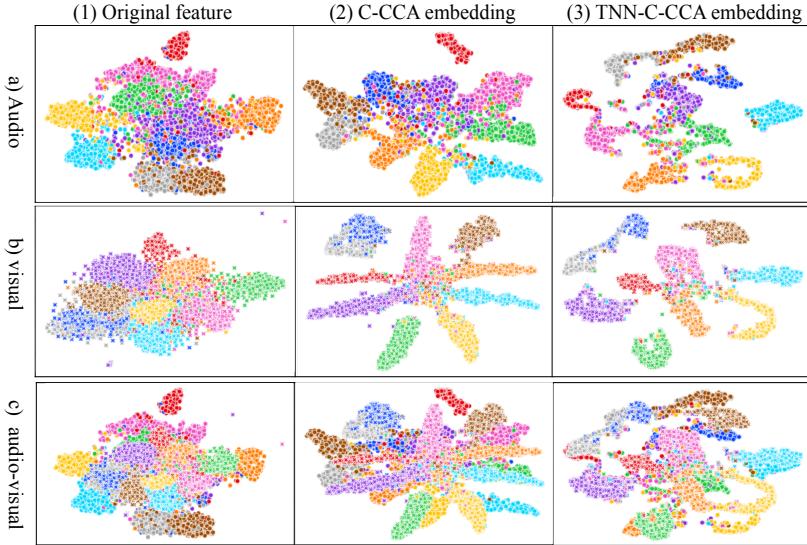


Fig. 7. The visualization of the two learned subspace with the t-SNE plot, shows audio, visual and audio-visual in the original feature, C-CCA learning subspace, and TNN-C-CCA learning subspace. The circle sign represents audio, the cross sign represents visual.

#### 4.5.3 Visualization of the learned semantic space.

The goal is to investigate the effectiveness of TNN-C-CCA model combines C-CCA embedding and TNN model on VEGAS dataset. We select one fold as target set with 5,600 samples. The learned common semantic space from C-CCA to generate the semantic features for all samples and then input them into TNN model is to generate more discriminative semantic features by taking negative samples into the training stage. Then, we use t-distributed Stochastic Neighbor Embedding (t-SNE) to implement dimension reduction on the original audio-visual dataset and these features respectively generated from Cluster-CCA and TNN-C-CCA model, where Fig. 7 shows audio, visual and audio-visual of their raw features, C-CCA features and TNN-C-CCA features. We can see that in Fig. 7, many samples in each category of two modalities scatter and hardly separated, while C-CCA embedding groups into clusters and each cluster represents one category, however, the clusters are not completely discriminative. In the center of space, some samples from different clusters are intersection and hard to be segregated. TNN-C-CCA embedding is much better than C-CCA embedding, those new clusters are more discriminative and samples belonging to the same category are in the same cluster. It indicates that TNN-C-CCA embedding learning effectively improves the performance compared with C-CCA embedding learning.

Furthermore, we investigate the effectiveness of learned semantic space by the audio-visual retrieval task. We try to compare the retrieval results of our model with the other three best models. Fig. 8 provides audio-to-visual retrieval examples generated respectively by ACMR, AGAH, C-DCCA, and our TNN-C-CCA model on VEGAS dataset for given audio with the "Chainsaw" label as the query. we can observe that the matched top 5 visuals by our TNN-C-CCA is 80% related to the label "Chainsaw" and average precision (AP) is 80.12% in all rank lists. For other models, ACMR model is 40% related to the query label and AP is 42.72% in all rank list; AGAH model is 60% related to the query label and AP is 55.34% in all rank list; C-DCCA model is 60% related to the query label and AP is 59.94% in all rank list.

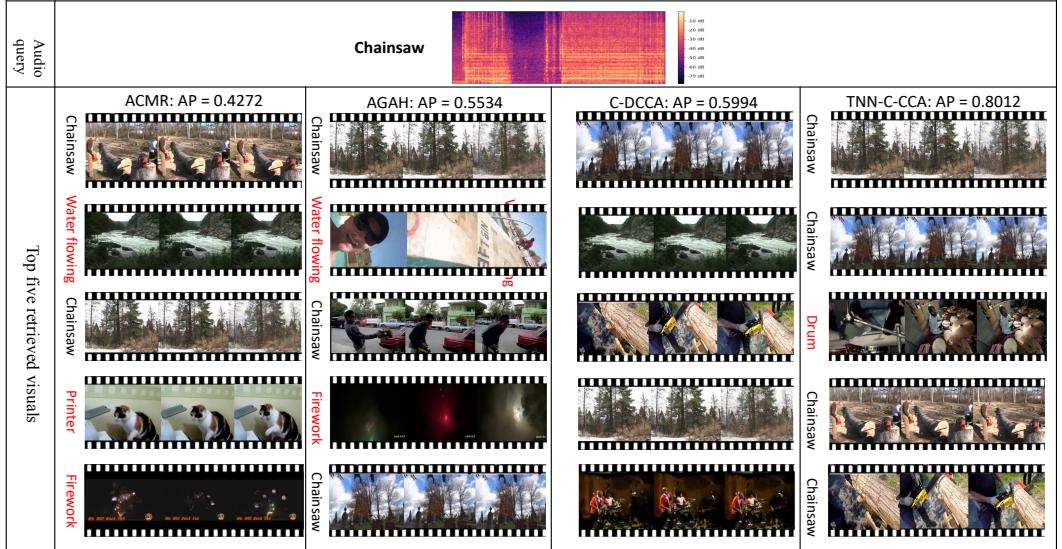


Fig. 8. The visualized audio-visual retrieval results of our proposed method and other three best methods, the Cluster-CCA, the AGAH, and the ACMR model. Given an audio as query, the figure shows the top five retrieved visuals.

#### 4.5.4 Effect of model parameters.

In the deep TNN part, batch size and margin play a leading role in the impact of the performance and time-consuming of the system. In this work, we respectively do some experiments on VEGAS dataset to evaluate the impact of batch size and margin.

Table 5. MAP with respect to different margins with TNN-C-CCA model when batch\_num is 500

Margin	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1
audio-visual	64.73	68.82	74.30	74.59	75.31	74.17	74.15	73.80	74.68	65.30	61.28
visual-audio	64.36	67.29	72.45	73.20	73.26	72.42	72.36	72.12	73.04	62.96	58.47

**Margin.** [4] is a region which is bounded by two hyper-planes in the support-vector machines (SVM), when selecting two hyper-planes to split two categories of data. The goal of SVM optimal is to maximize the margin between the vectors of the two categories. The margin of deep TNN is quite similar to the margin in SVM.

In our work, we use Cosine distance to calculate the difference among anchor, positive and negative samples, according to our loss function of deep TNN, the effective margin ranges from 0.0 to 2.0. In our experiments, we show the MAP of audio-to-visual retrieval and visual-to-audio retrieval based on the margin ranges from 0.1 to 1.1 by a step as 0.1 and set the number of batches to 500. All the results are listed in Table 5. In order to show the change of MAP values more obviously, we draw the MAP curve based on changing the margin. The right of Fig. 9 presents when the margin range from 0.3 to 0.9 by step as 0.1, the MAP value has no big change. When the margin is 0.5 the MAP can get the best performance. As margin increases from 0.1 to 0.5, the MAP increases

from 64.73% to 75.31% for audio-to-visual retrieval and from 64.36% to 73.26% for visual-to-audio retrieval. While the margin ranges from 0.5 to 1.1, the MAP decreases from 75.31% to 61.28% for audio-to-visual retrieval and from 73.26% to 58.47% for visual-to-audio retrieval.

**Batch size.** is a hyper-parameter in machine learning, which defines the numbers of samples to update the model weights in one iteration. The number of batches is the number of iterations used in the experiment. Generally, the training dataset can be divided into one or more batches. In our experiments, we defined different batch sizes by changing the number of batches. We divided our training set into different batches ranging from 300 to 900 by a step as 50. Table 6 shows the

Table 6. MAP in respect to different batch sizes with TNN-C-CCA model when margin is 0.5

Batches	300	350	400	450	500	550	600	650	700	800	900
a-v	74.49	73.63	75.31	74.50	74.51	74.99	74.87	74.58	74.12	62.96	61.28
v-a	73.16	71.47	73.26	72.55	72.98	73.22	72.85	72.79	71.64	65.30	58.47
Time(h)	32	27	21	16	12	9	6	4	3	2	2

MAP and time-consuming (hour) of audio-to-visual retrieval and visual-to-audio retrieval. CCA, KCCA, C-CCA, DCCA, and C-DCCA will take about 2, 3, 3, 4 and 7 hours respectively. In general, time-consuming take more time, the performance will be better. When the number of a batch is 400, the batch size is about 55 (batch size=training set/batch number), which can get the best MAP value of 75.31% for audio-to-visual retrieval and 73.26% for visual-to-audio retrieval compared with other number of a batch. Overall, the MAP value has no big difference when the number of batch ranges from 300 to 700. The big difference of running time in audio-visual cross-modal retrieval is when the number of a batch is 300 and the samples in the batch are balanced, it needs almost 32 hours to finish the experiment. There are around 70 samples in the batch, including 63 negative samples and 6 positive samples combination, totally in the batch there are  $6*63*70=264640$  triplets. When the training set is divided into 700 batches, the batch size is about 30. In the same situation,

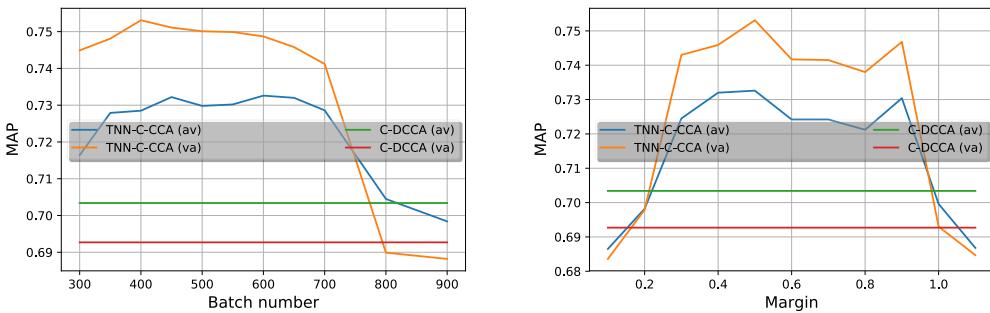


Fig. 9. The left figure is the MAP curve of TNN-C-CCA and C-DCCA on batch number range from 300 to 700 and the margin are 0.5. The right figure is the MAP curve of TNN-C-CCA and C-DCCA on margin range from 0.3 to 1.0 and the batch number is 500.

in the batch, there are  $2*27*30=1620$  triplets, it saves more time compared with 300 batches, only taking 3 hours. When the number of batches is set to 800, the MAP will decrease a lot and the performance is close to that of the C-DCCA model. When the batch number is 900, the MAP will be

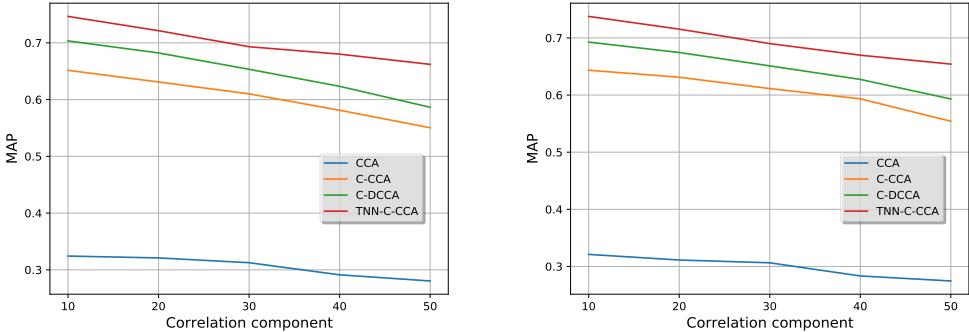


Fig. 10. The MAP curve of the correlation component changes from 10 to 50, the corner point in the curve represents the correlation component of X-axis and MAP of Y-axis, which use line to connect two adjacent points. The left part is audio-to-visual retrieval and the right part is visual-to-audio retrieval.

lower than that of the C-DCCA model. In the left of Fig. 9, the top MAP is 400 batches. In the left part of the curve, as the batches increase from 300 to 400, the MAP will get a bit larger. In the left part of the curve, the number of a batch from 500 to 900, the MAP is degraded. When the number of batches reaches 800, our model gets the same performance as C-DCCA. When the number of a batch is smaller than 800, it will get lower than that of C-DCCA.

The above experiment results show that our model can outperform other methods when we set effective parameters (margin and batch size). We respectively do the experiments based on one of them as the main variable. There are a lot of combinations between batch size and margin. In our experiments, we fixed the margin as 0.5 and make the batch size as a variable. Better batch size is obtained based on better MAP. Secondly, when batch size is fixed and the margin is made as a variable, we can get a better margin.

**Correlation components.** In addition, the number of correlation components in the CCA-variant method are very important, in order to investigate the correlation structure of learned representation among the four approaches. Fig.10 shows the MAP curve based on the change of the number of components for all the four models. In our experiments, as for our architecture TNN-C-CCA, the dimension of Cluster-CCA and the dimension of output in deep TNN are the same. It is very clear that the number of correlation components is set to 10 which can achieve the best MAP 74.66% for audio-to-visual retrieval and 73.77% for visual-to-audio. As the component decreases, the performance will go down. Especially, it is not a big change in the CCA paradigm at 10, 20, 30, but with the decrease at 40 and 50.

## 5 CONCLUSION

In this work, we propose a new deep architecture that consists of Cluster-CCA and deep TNN model. Our architecture can get both benefits of the Cluster-CCA and deep TNN such that completely consider the suitable location of each data point in the shared subspace based on the pairwise correlation and semantic label allocation. The deep TNN model is a supplement of Cluster-CCA model by learning the similarity distance between all pairs within the same class and compares the similarity distance with all possible pairs cross different views. This can help to learn more discriminative embedding space between audio and visual. We applied two different audio-visual datasets to evaluate the performance of our architecture with the PRC and MAP metrics. Audio and

visual features are respectively represented by the advanced pre-trained deep CNN based feature extractors for both datasets. The result of the experiments proved that our model can outperform other state-of-the-art cross-modal retrieval models. In order to further investigate the capability of cross-modal embedding learning, we design more extensive experiments for ablation studies where triplet selection strategies, distance metrics, visualization of learned semantic space, and effect of model parameters are investigated.

In the future, we would like to extend our model to support retrieval across other different multi-modalities, such as image-text, audio-text, and video-text cross-modal retrieval. We would like to explore generative adversarial networks (GAN) methods to improve our architecture, and we attempt to extend our current framework to achieve unsupervised cross-modal retrieval to solve the problem of weakly annotated labels like our MV-10K dataset.

## ACKNOWLEDGEMENTS

This work was supported by JSPS Grant-in-Aid for Scientific Research (C) under Grant No. 19K11987.

## REFERENCES

- [1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. In *arXiv:1609.08675*. <https://arxiv.org/pdf/1609.08675v1.pdf>
- [2] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. 2013. Deep Canonical Correlation Analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. pp.1247–1255. <https://doi.org/10.4324/9781315755649-3>
- [3] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. 2016. Fully-Convolutional Siamese Networks for Object Tracking. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*. pp.850–865. [https://doi.org/10.1007/978-3-319-48881-3\\_56](https://doi.org/10.1007/978-3-319-48881-3_56)
- [4] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning*. Vol.20, no.3 (1995), pp.273–297. <https://doi.org/10.1007/BF00994018>
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. pp.248–255. <https://doi.org/10.1109/CVPRW.2009.5206848>
- [6] Fangxiang Feng, Xiaojie Wang, Ruifan Li, and Ibrar Ahmad. 2015. Correspondence Autoencoders for Cross-Modal Retrieval. *TOMCCAP* Vol.12, no.1s (2015), pp.26:1–26:22. <https://doi.org/10.1145/2808205>
- [7] Wen Gu, Xiaoyan Gu, Jingzi Gu, Bo Li, Zhi Xiong, and Weiping Wang. 2019. Adversary Guided Asymmetric Hashing for Cross-Modal Retrieval. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019*. 159–167. <https://doi.org/10.1145/3323873.3325045>
- [8] David R. Hardoon, Sándor Szemrák, and John Shawe-Taylor. 2004. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*. Vol.16, no.12 (2004), pp.2639–2664. <https://doi.org/10.1162/0899766042321814>
- [9] Li He, Xing Xu, Huimin Lu, Yang Yang, Fumin Shen, and Heng Tao Shen. 2017. Unsupervised cross-modal retrieval through adversarial learning. In *2017 IEEE International Conference on Multimedia and Expo, ICME 2017, Hong Kong, China, July 10-14, 2017*. pp.1153–1158. <https://doi.org/10.1109/ICME.2017.8019549>
- [10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. *CoRR* abs/1703.07737 (2017). arXiv:1703.07737 <http://arxiv.org/abs/1703.07737>
- [11] Mengqiu Hu, Yang Yang, Fumin Shen, Ning Xie, Richang Hong, and Heng Tao Shen. 2019. Collective Reconstructive Embeddings for Cross-Modal Hashing. *IEEE Trans. Image Processing* Vol.28, 6 (2019), pp.2770–2784. <https://doi.org/10.1109/TIP.2018.2890144>
- [12] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. pp.448–456.
- [13] Go Irie, Hiroyuki Arai, and Yukinobu Taniguchi. 2015. Alternating Co-Quantization for Cross-Modal Hashing. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. pp.1886–1894. <https://doi.org/10.1109/ICCV.2015.219>
- [14] Yanli Ji, Yuhan Hu, Yang Yang, Fumin Shen, and Heng Tao Shen. 2019. Cross-domain facial expression recognition via an intra-category common feature and inter-category Distinction feature fusion network. *Neurocomputing* Vol.333

- (2019), pp.231–239. <https://doi.org/10.1016/j.neucom.2018.12.037>
- [15] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. pp.1889–1897.
- [16] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-shot Image Recognition. In *ICML Deep Learning Workshop*, Vol. 7, no.11. pp.956–963. <https://doi.org/10.1017/CBO9781107415324.004>
- [17] Pei Ling Lai and Colin Fyfe. 2000. Kernel and Nonlinear Canonical Correlation Analysis. *Int. J. Neural Syst.* Vol.10, no.5 (2000), pp.365–377. <https://doi.org/10.1142/S012906570000034X>
- [18] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. 2003. Multimedia content processing through cross-modal association. In *Proceedings of the Eleventh ACM International Conference on Multimedia, Berkeley, CA, USA, November 2-8, 2003*. pp.604–611. <https://doi.org/10.1145/957013.957143>
- [19] R. Mammatha, Chao-Yuan Wu, Alexander J. Smola, and Philipp Krähenbühl. 2017. Sampling Matters in Deep Embedding Learning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. pp.2859–2867. <https://doi.org/10.1109/ICCV.2017.309>
- [20] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*. pp.689–696.
- [21] Yuxin Peng, Xin Huang, and Jinwei Qi. 2016. Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. pp.3846–3853.
- [22] Yuxin Peng and Jinwei Qi. 2019. CM-GANs: cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* Vol.15, 1 (2019), pp.22.
- [23] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. 2018. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing* Vol.27, 11 (2018), pp.5585–5599.
- [24] Viresh Ranjan, Nikhil Rasiwasia, and C. V. Jawahar. 2015. Multi-label Cross-Modal Retrieval. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. pp.4094–4102. <https://doi.org/10.1109/ICCV.2015.466>
- [25] Nikhil Rasiwasia, Dhruv Mahajan, Vijay Mahadevan, and Gaurav Aggarwal. 2014. Cluster Canonical Correlation Analysis. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*. pp.823–831. <https://doi.org/10.1201/b18358-8>
- [26] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Covillo, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*. pp.251–260. <https://doi.org/10.1145/1873951.1873987>
- [27] Mohammad Rastegari, Jonghyun Choi, Shobeir Fakhraei, Hal Daumé III, and Larry S. Davis. 2013. Predictable Dual-View Hashing. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. 1328–1336. <http://proceedings.mlr.press/v28/rastegari13.html>
- [28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. pp.815–823. <https://doi.org/10.1109/CVPR.2015.7298682>
- [29] Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann. 2014. ADVISOR: Personalized Video Soundtrack Recommendation by Late Fusion with Heuristic Rankings. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03-07, 2014*. pp.607–616. <https://doi.org/10.1145/2647868.2654919>
- [30] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. 2012. Generalized multiview analysis: A discriminative latent space. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp.2160–2167.
- [31] Abhishek Sharma, Abhishek Kumar, Hal Daumé III, and David W. Jacobs. 2012. Generalized Multiview Analysis: A discriminative latent space. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. pp.2160–2167. <https://doi.org/10.1109/CVPR.2012.6247923>
- [32] Blake Shaw, Bert Huang, and Tony Jebara. 2011. Learning a Distance Metric from a Network. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*. pp.1899–1907.
- [33] Chen Shen, Zhongming Jin, Wenqing Chu, Rongxin Jiang, Yaowu Chen, Guo-Jun Qi, and Xian-Sheng Hua. 2019. Multi-level Similarity Perception Network for Person Re-identification. *TOMM* 15, 2 (2019), 32:1–32:19. <https://doi.org/10.1145/3309881>
- [34] Xiaoxiao Shi and Philip S. Yu. 2012. Dimensionality Reduction on Heterogeneous Feature Space. In *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*. pp.635–644. <https://doi.org/10.1109/ICDM.2012.30>

- [35] Josef Sivic and Andrew Zisserman. 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France*. pp.1470–1477. <https://doi.org/10.1109/ICCV.2003.1238663>
- [36] Richard Socher and Fei-Fei Li. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*. pp.966–973. <https://doi.org/10.1109/CVPR.2010.5540112>
- [37] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. pp.4278–4284. <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806>
- [38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. pp.2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
- [39] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial Cross-Modal Retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. pp.154–162. <https://doi.org/10.1145/3123266.3123326>
- [40] Cheng Wang, Haojin Yang, and Christoph Meinel. 2015. Deep Semantic Mapping for Cross-Modal Retrieval. In *27th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2015, Vietri sul Mare, Italy, November 9-11, 2015*. pp.234–241. <https://doi.org/10.1109/ICTAI.2015.45>
- [41] Di Wang, Xinbo Gao, Xiumei Wang, Lihuo He, and Bo Yuan. 2016. Multimodal discriminative binary embedding for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing* Vol.25, 10 (2016), pp.4540–4554.
- [42] Jian Wang, Yonghao He, Cuicui Kang, Shiming Xiang, and Chunhong Pan. 2015. Image-Text Cross-Modal Retrieval via Modality-Specific Feature Learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015*. pp.347–354. <https://doi.org/10.1145/2671188.2749341>
- [43] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. 2013. Learning Coupled Feature Spaces for Cross-Modal Matching. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. pp.2088–2095. <https://doi.org/10.1109/ICCV.2013.261>
- [44] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. pp.5005–5013. <https://doi.org/10.1109/CVPR.2016.541>
- [45] Zhangcheng Wang, Ya Li, Richang Hong, and Xinmei Tian. 2019. Eigenvector-Based Distance Metric Learning for Image Classification and Retrieval. *TOMM* 15, 3 (2019), 84:1–84:19. <https://doi.org/10.1145/3340262>
- [46] Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *Neural Comput. Appl.* Vol.23 (2013), pp.2031–2038.
- [47] Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. 2019. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web* Vol.22, 2 (2019), pp.657–672.
- [48] Xing Xu, Li He, Atsushi Shimada, Rin-ichiro Taniguchi, and Huimin Lu. 2016. Learning unified binary codes for cross-modal retrieval via latent semantic hashing. *Neurocomputing* Vol.213 (2016), pp.191–203. <https://doi.org/10.1016/j.neucom.2015.11.133>
- [49] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li. 2019. Ternary Adversarial Networks With Self-Supervision for Zero-Shot Cross-Modal Retrieval. *IEEE Transactions on Cybernetics* Vol.49, 7 (2019), pp.1–14. <https://doi.org/10.1109/TCYB.2019.2928180>
- [50] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. 2017. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing* Vol.26, 5 (2017), pp.2494–2507.
- [51] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. pp.3441–3450. <https://doi.org/10.1109/CVPR.2015.7298966>
- [52] Yan Yan, Feiping Nie, Wen Li, Chenqiang Gao, Yi Yang, and Dong Xu. 2016. Image Classification by Cross-Media Active Learning With Privileged Information. *IEEE Trans. Multimedia* Vol.18, 12 (2016), pp.2494–2502. <https://doi.org/10.1109/TMM.2016.2602938>
- [53] Yi Yang, Feiping Nie, Dong Xu, Jiebo Luo, Yueling Zhuang, and Yunhe Pan. 2012. A Multimedia Retrieval Framework Based on Semi-Supervised Ranking and Relevance Feedback. *IEEE Trans. Pattern Anal. Mach. Intell.* Vol.34, 4 (2012), pp.723–742. <https://doi.org/10.1109/10.1109/TPAMI.2011.170>
- [54] Ting Yao, Tao Mei, and Chong-Wah Ngo. 2015. Learning query and image similarities with ranking canonical correlation analysis. In *Proceedings of the IEEE International Conference on Computer Vision*. pp.28–36.
- [55] Yi Yu, Suhua Tang, Kiyoharu Aizawa, and Akiko Aizawa. 2018. Category-based deep CCA for fine-grained venue discovery from multimodal data. *IEEE transactions on neural networks and learning systems*. Vol.30, no.99 (2018), pp.1–9.

- [56] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen. 2019. Deep Cross-Modal Correlation Learning for Audio and Lyrics in Music Retrieval. *TOMCCAP*. Vol.15, no.1 (2019), pp.20:1–20:16. <https://doi.org/10.1145/3281746>
- [57] Donghuo Zeng, Yi Yu, and Keizo Oyama. 2018. Audio-Visual Embedding for Cross-Modal Music Video Retrieval through Supervised Deep CCA. In *2018 IEEE International Symposium on Multimedia, ISM 2018, Taichung, Taiwan, December 10-12, 2018*, pp.143–150. <https://doi.org/10.1109/ISM.2018.00-21>
- [58] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. 2018. Unsupervised Generative Adversarial Cross-Modal Hashing. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp.539–546. <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16746>
- [59] Yin Zheng, Yu-Jin Zhang, and Hugo Larochelle. 2014. Topic Modeling of Multimodal Data: An Autoregressive Approach. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp.1370–1377. <https://doi.org/10.1109/CVPR.2014.178>
- [60] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. 2018. Visual to Sound: Generating Natural Sound for Videos in the Wild. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp.3550–3558. <https://doi.org/10.1109/CVPR.2018.00374>
- [61] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. 2017. Uncovering the Temporal Context for Video Question Answering. *International Journal of Computer Vision* Vol.124, 3 (01 Sep 2017), pp.409–421. <https://doi.org/10.1007/s11263-017-1033-7>

Received Oct 2019; revised Jan 2020; accepted March 2020