Deep Alignment Representation Learning for Multimodal Information Retrieval

by

Donghuo Zeng

Dissertation

submitted to the Department of Informatics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

The Graduate University for Advanced Studies, SOKENDAI

September 2020

Acknowledgments

I would appreciate extending my genuine thanks to my two supervisors, Yi Yu and Keizo Oyama, for their effective guidance through each stage of my Ph.D program. With their instruction in our regular group meeting for each two weeks in the almost past three years, I have made a great progress in my research and gained a deeper understanding of the significance of scientific research. Moreover, they convincingly conveyed a spirit of adventure in regard to research. Without their persistent instruct this dissertation may impossible to be done.

In particular, professor Yu inspires my interests in the development of innovative technologies and helps me improve the basic skill of research, such as paper writing, presentation, research proposal, architecture designing and even some trips of research life. It's absolutely inaccessible to finish the dissertation without professor Yu's guidance.

In addition, professor Oyama guild me to generally and deeply think about my research, he always put forward some brilliant questions that really broad my mind.

Financial support was provided by JSPS KAKENHI Grant Number 16K16058.

Abstract

Multimedia data has a complex structure including various modalities, such as audio, image, text, and video. In order to satisfy the requirement of users to obtain useful information from the increasingly complicated multimedia collections, which often rely on multimodal information retrieval (MIR). MIR takes one modality of data as the query to retrieve relevant data of another modalities. Learning robust representation is essential for MIR, which requires the MIR system to capture the alignment that contains similar semantic concepts across modalities. Such alignment representation learning allows us to learn a shared latent space, where the data samples of different modalities have similar representations and can be directly compared with each other.

The main challenge of MIR remains in diminishing the differences between two data points of different modalities or bridging the heterogeneous gap, which has been widely studied in video-text, image-text, and audio-text. Unfortunately, because of the lack of available temporal structures of multimodal dataset, learning aligned representation of the temporal structure of different data modalities is impossible. Moreover, learning aligned representations for multimodal information retrieval on three data modalities are rarely reported, such as alignment representation learning in sheet music, lyrics, and audio.

The target of this dissertation is to learn the correlation between the data set of different modalities for MIR. We introduce three architectures: Supervised-Deep Canonical Correlation Analysis (S-DCCA) and Triplet Neural Networks with Cluster Canonical Correlation Analysis (TNN-C-CCA) for audio-visual cross-modal retrieval, Deep alignment representation learning methods (DARLearning) for sheet music, audio, and lyrics.

S-DCCA model learns aligned representations in a shared latent space by finding

nonlinear transforms for audio and visual to optimize the correlation between them. In particular, the model exploits the temporal structure of data to achieve the music video retrieval with audio chunks as query to retrieve the full-length visual. The contribution of this work is that we utilize the temporal structure of our collected MV-10K dataset to retrieve full-length visual with audio chunks as query by using attention mechanism to capture local properties of audio. The experiment results show that the aligned representations for audio and visual of our proposed architecture is useful for music video retrieval.

TNN-C-CCA method can be viewed as an improvement of S-DCCA with audio-visual special loss function to promote CCA-variant methods by establishing triplets as training based on the similar or dissimilar semantic pairs on the Cluster-CCA embeddings. The implication of this work is to learn a better aligned representation for audio-visual cross-modal retrieval by applying audio-visual special loss function to improve Cluster-CCA method. Compared with other state-of-the-art methods, the proposed method can achieve better performance.

DARLearning approach transfers strong semantic relevant pairs from two different modalities to the weak relevant data of another modality by adversarial learning. The contribution of this work is that our approach can learn useful representations of three different modalities for MIR. The learned discriminative aligned representations of this approach in the experiment indicates the results can beneficial from the representations.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

Deep alignment representation learning is to map low-level features of different modalities into a semantic shared space based on latent concept alignment by the deep learning method, seen in the Fig. 1.1. This dissertation aims at learning deep alignment representations for multimodal information retrieval (MIR), including audio-visual cross-modal retrieval and cross-modal retrieval between every two modalities from sheet music, audio, and lyrics, which is to retrieve the relevant data in one modality with a query in another modality. The challenge of MIR mainly discussed here is the semantic gap or the heterogeneous gap. Especially, the widely used low-level features of different modalities possess inconsistent distributions and representations, which causes the features are unable to be directly compared with each other to accomplish the retrieval achievement. The objective of this dissertation is to develop new architecture to project the low-level data representations of different modalities to high-level semantic representation in a common space to bridge the modality gap.

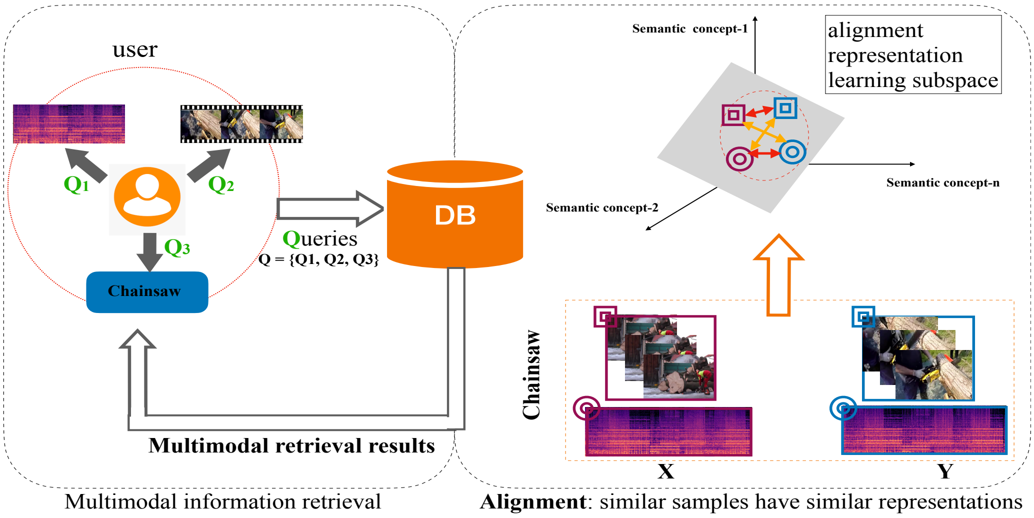

user
Q1
Q2
Q3
Chainsaw
Queries
Q = {Q1, Q2, Q3}
DB
Multimodal retrieval results
Multimodal information retrieval
Semantic concept-1
alignment representation learning subspace
Semantic concept-2
Semantic concept-n
Chainsaw
X
Y
Alignment: similar samples have similar representations

Figure 1.1: The general framework of alignment representation learning for multimodal information retrieval.

The contribution of this dissertation is that we proposed three different advanced approaches for deep alignment representation learning in MIR areas. Our experiments suggest that these kinds of representations are useful for MIR.

With the rapid growth in web technologies and user applications on the Internet, web has increasingly become the platform of various multimedia data aggregated. In order to enable MIR system to perceive and understand the unstructured multimedia data and conduct indistinguishable multimodal information interaction from a large amount of data, it requires the multimodal models can abstract the data and build similarity link from one modality item to anther modalities of the items there are semantically related by representation learning.

Different from learning representations for single modality, this dissertation learns representations across modalities. Imagine a scene: when there is lightning in the air visually, the same concept also appears aurally, such as a thunder sound, and the concept also can be written in a sentence "Lightning flashed around and thunder rumbled". In the case of representation learning, a robust representation of modality is often the one that captures the alignment in representations across modalities for the observed inputs. In this way, the lightning video or image, the thunder sound, and the sentence description are expected to produce similar representations for MIR. In the Fig. 1.1, the dataset for training contains videos and their labels (i.e., chainsaw and etc.), a video contains two tracks: audio track and visual track. The alignment representation learning method projects the low-level features of different modalities into a alignment representation learning subspace, where supports the data points of different modalities can be compared with each other, then ranking the items of retrieved database. When the user inputs a data point from one modality, the MIR system will rank the items of the database in the corresponding semantic space and return the multimodal retrieval results.

Among the various ways of learning aligned representation, CCA [1] is a classical linear method to learn the correlation of two variable sets (V_1, V_2) by utilizing two views of the same semantic object to learn the aligned representation of the semantics. In order to find linear transforms to map V_1 and V_2 into a common space, where the correlation of similar pairs are optimized, supposed the linear transforms W_1 and W_2 are the matrices and $\Sigma_{v_1v_1}$ and $\Sigma_{v_2v_2}$ are the covariance matrices of V_1 and V_2 and $\Sigma_{v_1v_2}$ is the cross-covariance matrix, which uses to maximize the correlation in the latent subspace as follows.

$$(W_1, W_2) = \arg \max_{(W_1,W_2)} \frac{W_1^T \Sigma_{v_1v_2} W_2}{\sqrt{W_1^T \Sigma_{v_1v_1} W_1 \cdot W_2^T \Sigma_{v_2v_2} W_2}} \tag{1.1}$$

Based on the CCA, some extension methods are proposed. Before projecting the features into a common space, KCCA [2] first map the features into a higher dimensional feature space. To be beneficial from deep learning method, DCCA [3] learns complex nonlinear transformations for two different sets of variants. Unlike the standard pair in the CCA training, in Cluster-CCA [4] divided each set into several clusters, the new pairs between two sets defined by the label then applied CCA to optimize the correlation between new pairs.

This dissertation focused on learning aligned representation based on deep learning by the composition of multiple non-linear transformations. The first two works can be viewed as two different nonlinear extension ways of CCA. The third work use CCA embedding to transfer one close relationship to the other two relationships.

In the first work, since previous researches are required the query and the retrieved content shared the same length of time, we achieve cross-modal music video retrieval concerning emotion similarity, which is to obtain full-length music silence video using an audio snippet. Therefore, we introduce a novel audio-visual aligned representation learning approach by the Supervised Deep Canonical Correlation Analysis (S-DCCA) that maps audio and visual into a latent shared subspace to bridge the heterogeneous gap between audio and visual data. The method is not only learning the aligned representation across the modalities but also preserves the similarity between data points with the same label in each modality and temporal structure information. Due to little off-the-shelf music video dataset is available, we collect a 10,000 music video from the YouTube-8M dataset to evaluate our proposed architecture. The performance of our experiment including the MAP and PRC suggest that our novel model can be implemented to music video retrieval.

Two main contributions were made in this work: i) We apply the emotion feature extraction model to select top k audio chunks to summary the audio content with local properties. ii) We establish an end-to-end supervised learning model for audio-visual cross-modal embedding where the model can acquire the semantic correlation between audio and visual content.

In the second work, on account of establishing alignment representation across modalities in previous works is trained on matched data pairs, which overlooks the unpaired data will weaken the alignment. We present a novel deep triplet neural network with cluster canonical correlation analysis (TNN-C-CCA) that is an end-to-end deep model with audio and visual branches. We utilize the correlation optimization during learning a latent shared subspace. The experimental results implemented on two audio-visual datasets demonstrate the presented model with two branches exceeds other state-of-the-art cross-modal retrieval methods.

In particular, two significant contributions include: i) A novel alignment representation learning method is benefit from a deep triplet neural network and cluster-CCA method. ii) We take the positive and negative examples into account to enhance the alignment learning between audio and visual during the training process. Our experiment uses 5-fold cross-validation to evaluate the learned predictive model.

In the third work, alignment learning in two different modalities limited in special modalities, in reality, alignment may appear in each modality of multimedia. In this case, we try to learn deep discriminative representations across three major musical modalities: sheet music, lyrics, and audio, where a deep learning network based on

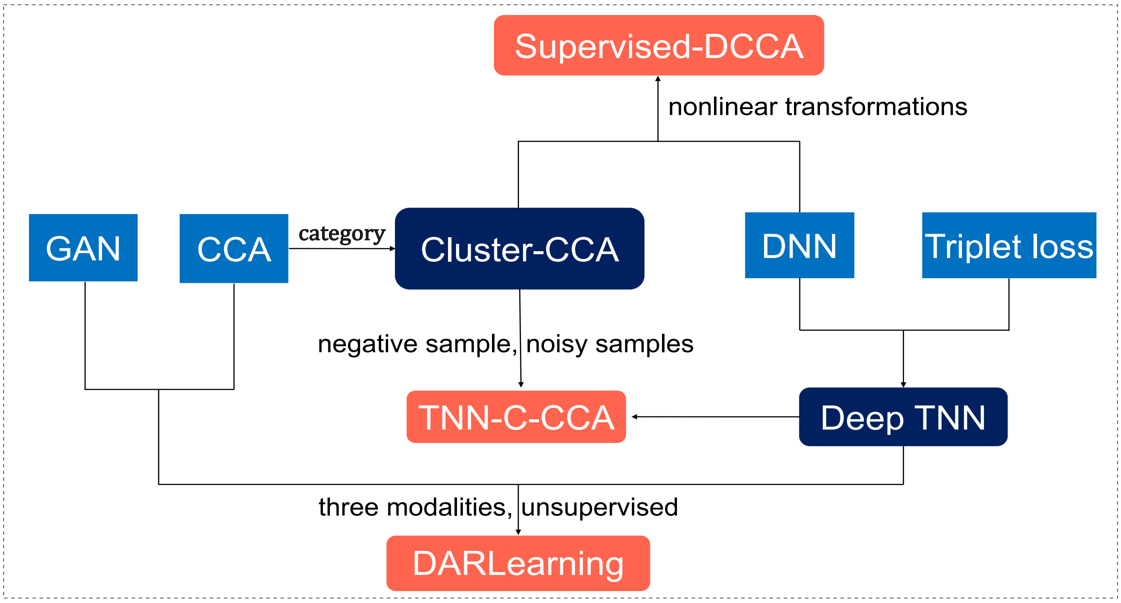

Supervised-DCCA
nonlinear transformations
GAN
CCA
category
Cluster-CCA
DNN
Triplet loss
negative sample, noisy samples
TNN-C-CCA
Deep TNN
three modalities, unsupervised
DARLearning

Figure 1.2: Overview of techniques used in this dissertation.

three branches is jointly trained. Our experiment result suggests that our model is useful for cross-modal retrieval tasks.

Two main contributions are achieved in the third work: i) our model has the capability of transferring a known pair to the other two unknown pairs by adversarial learning, and one of the unknown pairs is from the known pair. ii) we explore the manifold structure of data points on CCA embedding, which enhances the generator model to generate discriminative representations.

Fig. 1.2 presents the relationship between the proposed methods and their related existing methods. The first model supervised-DCCA extends cluster-CCA by finding nonlinear transformations instead of a linear projection model to optimize the correlation between audio and visual modalities, which can be used for retrieving full-length visual with audio chunk. Compared with supervised-DCCA, our second model TNN-C-CCA uses negative samples to reduce the noisy samples that the samples in the supervised-DCCA shared subspace are grouped into the wrong cluster. The former two architectures constrain on two cross-modal data and it highly relies on the user's annotation, our third approach can learn alignment representation for three different modalities by unsupervised learning.

1.2 Organization

The remainder of this dissertation is organized as follows:

Chapter 2 summarises the related works of our previous three works, we respectively display related works of cross-modal retrieval and multimodal information retrieval; alignment representation learning in special domain. Chapter 3 introduces the detailed process of dataset collection, the methods of feature extraction and the applied evaluation metrics. Chapter 4 derives from one of our papers [5] and presented a new audio-visual cross-modal retrieval system, the system is based on the supervised deep canonical analysis. Chapter 5 is based on previous work [6], proposed an audio-visual cross-modal embedding learning system which consists of cluster canonical analysis algorithm and triplet neural networks. Chapter 6 is based on our work [7], which shows how we develop a system for unsupervised generative adversarial multimodal alignment learning for sheet music audio and lyrics. Chapter 7 summarizes the current works on multimodal joint embedding learning and described the feasibility of our vision on future research.

2 Related Work

In this section, we discuss some close related works, which promotes our motivation of the architecture developed and are useful for the explanation of our research background. We organize this section with related works by two subsections. Section 2.1 introduces the cross-modal retrieval and multimodal information retrieval to explain their difference and relevance. Section 2.2 shows some typical related cross-modal retrieval tasks in special domains.

2.1 Cross-modal Retrieval and Multimodal Information Retrieval

2.1.1 Cross-modal Retrieval Techniques

Different from retrieval in the same modality, such as image retrieval [8], cross-modal retrieval is used for implementing a retrieval task across different modalities. such as image-text[9, 10, 11, 12], video-text[13], and audio-text[14] cross-modal retrieval. The main challenge of cross-modal retrieval is the modality gap and the key solution of cross-modal retrieval is to learn aligned embedding for different modalities. Learning aligned representation is not only a solution of cross-modal retrieval and also applied for other multimedia tasks, such as image classification [15], video question and answering [16]. As for our task, cross-modal retrieval aim at generating new representations from different modalities in the shared subspace, such that newly generated features can be applied in the computation of distance metrics, such as Cosine distance and Euclidean distance.

Canonical Correlation Analysis Variant Methods

Some methods such as CFA [17], CoCA [18], CMPM [10], MVML-GL[19], GSS-SL[20] and LRGA [21] are to learn the cross-modal association to reduce the dimension. Canonical correlation analysis (CCA) is one of the most prevailing cross-modal embedding models, which aims at finding a pair of linear transformations to maximize the correlation between two different modalities. CCA [22] can be used to calculate the cross-modal correlations between image and text. Kernel canonical correlation analysis (KCCA) [2, 23] is to extend CCA by finding nonlinear transforms for the data to a feature common space and then applying linear-CCA. KCCA is aspired by support Vector regression [24] to perform a nonlinear mapping of the data set into a high-dimensional feature space. Compared with CCA to learn the pairwise correspondence correlation between the data points from two modalities, Cluster-CCA (C-CCA) [4, 25, 26] partitions all the data points into multiple classes or clusters, where the data sets shares the correspondences. By the same way of extending CCA to KCCA, extending C-CCA to C-KCCA [4, 2] to account for non-linear correlations. Instead of exploring linear method CCA to learn the correlation, Deep CCA [3, 27] is an alternative to the non-parametric method KCCA, which is to learn complex nonlinear transformations for the data set, such that the newly generated representations are highly linearly correlated. Similar to the extension of CCA to DCCA, extending C-CCA to category-based deep canonical correlation analysis (C-DCCA) [28] by mapping venue image and text into the same semantic space, which can strengthen their pairwise correlation. Besides, some CCA variant methods [29, 30, 31, 32] are applied in cross-modal retrieval tasks.

Artificial neural networks methods

With the development of artificial neural networks related methods [33], such as deep neural networks (DNN) [34], recurrent neural network (RNN) [35], Long short-term memory (LSTM) [36], convolutional neural network (CNN) [37], Attention mechanisms [38] and generative adversarial network (GAN) [39], which can learn nonlinear transforms for data representation has success in single-modality tasks, such as pattern classification and recognition [40, 41, 42], person re-identification [43, 44, 45]. **DNN** method recently has increasingly explored in cross-modal correlation learning by finding nonlinear transformations of data points to optimize the correlation in a shared subspace. DCCA computes feature representations of the two data modalities via feeding them into two layers of nonlinear transformation. Corr-AE [46] architecture learns the correlation of hidden representations for two modality autoencoders, which minimizes linear transforms of representation and correlation error in hidden representations of modalities. [47] develops a novel application of deep neural networks to learn feature representation across modalities. [48] proposes deep neural networks for cross-modal retrieval, which is divided into two steps: by considering intra-media information and inter-media correlation to generate a new representation for each modality. Then, learning the cross-modal correlation in shared feature representation space through a complex cross-modal multiple deep networks. [48] is for matching images and captions, which learns a joint latent space with DCCA to get the high dimensionality of the feature representations.

LSTM method applied in [49, 50, 51] for learning language representation and AlexNet [52], VGGNet [53] or ResNet [54] for learning image representations, by learning a shared latent space to generate compact binary codes for image and sentence. HM-LSTM [55] explores the hierarchical relations between sentences and phrases by learning joint embedding of sentences, phrases, images, and image regions.

GAN approach has increasingly exploited in the cross-modal retrieval task. [56] presents an architecture, which includes the interplay between feature projector part and a modality classifier part for adversarial learning, in order to learn aligned representation for image and text modalities. SCH-GAN [57] model is for cross-modal hashing, where the generative model can select margin samples from unlabeled data in one modality by a query from other modality. Such that the model can solve the problem that cross-modal architecture highly relies on the user's annotation.

2.1.2 Multimodal Information Retrieval

Only one modality retrieval such as image retrieval [58, 59], textual retrieval [60, 61] and audio retrieval [62, 63], has successfully shown excellent performance, where the query and the content have the same format and can be matched almost directly. The cross-modal representation learning or retrieval also has revolutionized, which through learning a common latent space to make it possible to compare the query and the content. However, many applications in the artificial intelligence field involve multiple modalities even more than two modalities [64]. In reality, flexible and applicable retrieval is more like the retrieval system can retrieve the content from any modalities by a query from one modality. Similar to learning aligned representations for cross-modal retrieval, representation learning for MIR into a shared space is a more challenging task, where query processing in MIR must fill a tremendous gap. Bridging the semantic gap between query and content for the features, which includes the low-level features such as color, shape, object, action, and high-level features like user's annotation, content-based semantic features, requires the system to learn a common latent space.

In the previous research reports, cross-modal aligned representation has made a breakthrough improvements successfully, while little research reported in three modalities. In the paper [65], they developed a deep convolutional neural networks model for audio, image, and text aligned representation learning. This model was trained by audio-image pairs and image-text pairs and is can be transfer between audio-text pairs by two alignment steps: 1) The alignment was a unsupervised teacher-student model transfer by optimizing the gaps with KL-divergence loss. 2) They applied the transfer discriminative visual model to transfer into the other two models by ranking loss. The learned representation showed the hidden units can automatically detect the concepts among each modality. In this paper [66], due to general methods learn shared subspace require that the data point from different modalities should share the same labels. However, it's not suit for a zero-shot learning based cross-modal retrieval, when the samples of target include unseen classes during the training. The TANSS model addresses the difficulty in the dataset for cross-modal zero-shot learning.

Firstly, semantic features learning to preserve the data structure of different modalities and simultaneously keep the relationship of different modalities. Then, the proposed self-supervised semantic model can leverage the relation of seen and unseen classes. The above correlation learning in common space was optimized by adversarial learning.

2.2 Alignment Representation Learning in Special Domain

2.2.1 Audio and Visual

Understanding the relationship between audio and visual is crucial for multimodal intelligence, which allows the system to learn the association between their contribution of high-level semantics. In the paper [67], they applied the linear correlation model such as CCA and CFA, to learn the relationship across modalities in a synchronized audiovisual signal. [68] learns correlations between audio and visual datasets and the correlations can be adopted for the clustering on the datasets. [69] uses KCCA and MV-HCRF to learn the relationship between audio and visual by using a multi-chain structured latent variable discriminative model. [70] presents a novel deep networks to learn the relationship between audio and visual. [71] overcomes the shortage of the Maximum covariance analysis method, which requires perfectly paired data as input. The proposed architecture can accept weakly paired data on large datasets and learn efficient representation for materials and its sound. In these works [72, 73, 74], they use the property of materials to learn audio representations from the visual features.

2.2.2 Audio and Lyrics

Recently, the audio-lyrics alignment techniques are getting trendy. The target of the techniques is to leverage the relationship between audio and lyrics, such as temporal relation [75], deep sequential correlation [76]. [77] presents two different novel models to learn the aligned representations for audio and lyrics. In the first model, using chord change in the Markov chain and a audio feature to extend the HMM architecture. In the second model, applying the repetition in the audio to stand for the lost chord information. [78] proposes a model for learning aligned representation for audio-lyrics by learning the correlation between the synthesized voice and the vocal track extracted from the song resource.

2.2.3 Sheet music and Audio

The popular problem is to automatically generate musically relevant linking structures between sheet music and audio. In [79], where aims to establish the linking of sheet music snippet to the corresponding clip in an audio recording of the same piece. [80] puts forward an end-to-end convolutional neural network for multi-modal learning, which takes short music audio snippets as input to find the relevant pixel location in the image of sheet music. However, the global and local tempo deviations in music recordings will effect the accuracy of the retrieval system in the temporal context. To address that, [81] introduces an additional soft-attention mechanism on audio modality. Instead of correlation learning with high-level representations, [82] matches musical audio to sheet music directly, the method learns latent shared subspace for short excerpts of the audio and corresponding section in sheet music.

2.2.4 Lyrics and Sheet music

Learning the correlation between lyrics and sheet music is a challenging research issue, which requires learning latent relationships with high-level representations. The automatic composition techniques are considerable for upgrading the musical applications. [83] proposes a novel deep generative model LSTM-GAN to learn the correlation in lyrics and melody for generation task. Similarly, [84] presents an approach that can automatically generate songs from Japanese lyrics. [85] presents a novel data-driven language model that can generate entire lyrics for a given input sheet music. [86] proposed an improved query by QBSH system with melody and lyrics information, which take advantage of extra lyrics information by combining the scores from pitch-only melody recognition and lyrics recognition. Accept that, “singing voice synthesis,” which is for generating singing voice has been drawing attention in the last years. [87] explores a novel architecture that the musical audio generation with no consideration of pre-assigned melody and lyrics.

3

Data Collection, Feature Extraction and Evaluation Metrics

In this section, we explain the motivation and contribution of our data collection, then present the process of dataset collection applied in our experiments and the data feature extraction is discussed following. In the end, we explain all the evaluation metrics that apply to leverage our models.

3.1 Dataset Collection

In our dissertation, we introduce three different models for different multimodal learning tasks. To evaluate our these new architectures of learning joint embedding, we collected three different datasets: music video 10K (MV-10K) dataset, VEGAS dataset and Musical Ternary Modalities (MTM) dataset. Fig 3.1 shows a few examples of each dataset we applied. The detailed description is as follows.

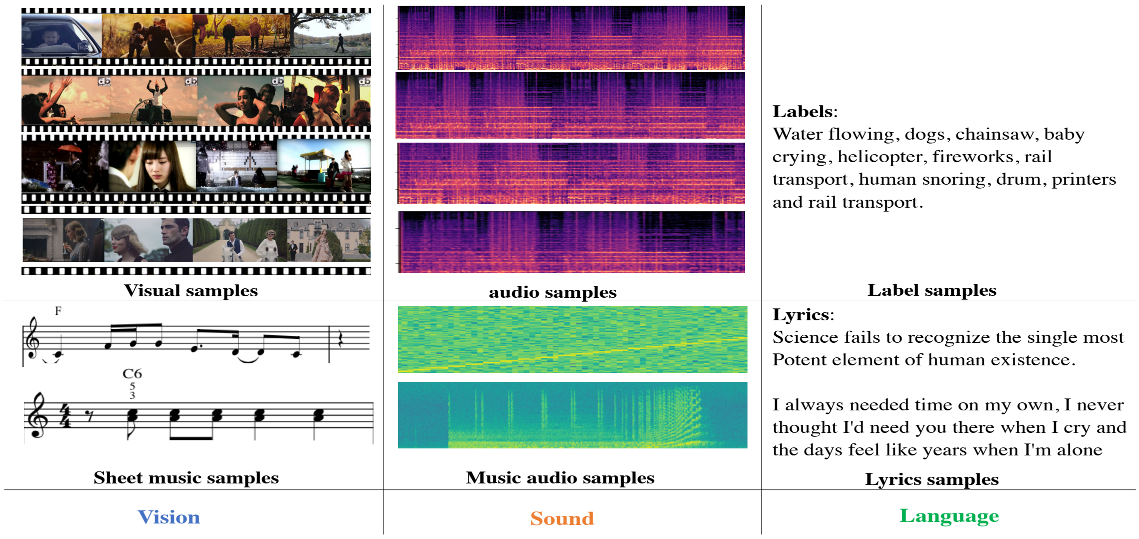

Visual samples
audio samples
Labels:
Water flowing, dogs, chainsaw, baby crying, helicopter, fireworks, rail transport, human snoring, drum, printers and rail transport.
Label samples
F
C6
Sheet music samples
Music audio samples
Lyrics:
Science fails to recognize the single most Potent element of human existence.
I always needed time on my own, I never thought I'd need you there when I cry and the days feel like years when I'm alone
Lyrics samples
Vision
Sound
Language

Figure 3.1: A few examples of our dataset applied in our experiments.

3.1.1 Dataset 1: Music Video 10K Dataset

We chose to learn aligned representation for musical audio and musical visual, which is two tracks extracted from a music video. Because the growing availability of music video allows us to build a certain scale audio-visual dataset to achieve our goal.

YouTube-8M is the largest video understanding multi-label dataset, mainly used for video classification, where has released the audio-visual time-localized frame-level and it's globe mean video-level features that is extracted from the Inception V3 model. We are interested in the music video, so we download the audio-visual feature pairs where contains the "music video" label. We collect 10K audio-visual feature pairs and neglect other annotations by two rules as follows:

1) In order to reduce the noisy in the video and focus on the music video, the selected video is required to contain "music video" entity only and not including other top entity in the hierarchical tree.

2) The time length of the selected video should span from 213 to 219 seconds, which allows us to cut the audio evenly and try to keep original time-located feature in each audio chunk without removing excess frames or padding extra frames.

Our goal is to retrieve the full length musical visual samples with the musical audio chunk as query, the initial pre-processing for our dataset is to slide audio into three chunks on average. In order to preserve sufficient information in each chunk, we set the number of audio chunks as 3, 6, 9. The time length of selected video in our MV-10K dataset is around 216 seconds, because 216 is the common multiple of 3, 6, 9. Frame-level visual feature is extracted by Inception version 3 model trained on the large amount of ImageNet dataset. The dimension of the frame-level visual feature is $L1024$, where L is the music video lengths in seconds. The frame-level audio feature is extracted by a Vggish model [88], the globe average of frame-level feature is as the video-level audio feature.

3.1.2 Dataset 2: VEGAS Dataset

In the MV-10K dataset, the label of the dataset is generated by music video content with K-mean methods, which annotates video with mood categories, the detailed description is in section 4.2.3. The precision achieved on the MV-10K dataset is not enough to leverage our second architecture TNN-C-CCA in section 5. Because the mood categories labeled by the content highly rely on the feature extraction model. However, the pre-trained model is not trained on the musical knowledge dataset. Except for investigating the musical video, the relationship between musical audio and visual based on the high-level semantic features, we are interested in the relationship between audio and visual with human/environment sound labels by learning object alignment from low-level features behind the data. We download the video from the Visually Engaged and Grounded AudioSet (VEGAS) [89] with manual annotated 10 labels (water flowing, dog, chainsaw, and etc.). A video in our dataset should satisfy two conditions:

1) To keep enough information for each, the length of each video should be longer than 1 second.

2) The selected video under 1) should be available to extract audio features from the audio track with the mel-frequency cepstrum (MFC) method.

Finally, the length of a video in the VEGAS dataset spans from 2 seconds to 10 seconds and the average of all videos is around 7 seconds. Finally, we obtain 28,103 videos to evaluate our proposed model in the second work.

3.1.3 Dataset 3: Musical Ternary Modalities Dataset

The available musical dataset with three modalities, which can be applied in multimodal information retrieval based on the high-level semantic features is rarely reported. We try to learn aligned representation for Sheet music image, musical audio and lyrics because they frequently appear in the music data collection. We follow the work [90] to collect our musical dataset by extending two modalities (lyrics and music notes) to three modalities: sheet music, audio, and lyrics.

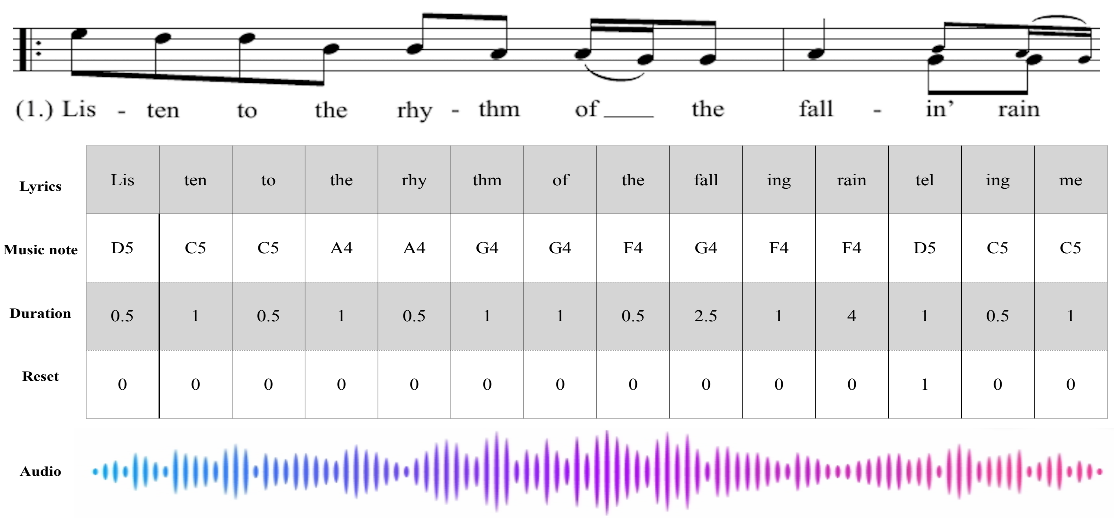
In [90], the music dataset consists of lyrics and music notes. the lyrics is parsed as syllable level collection, such as the lyrics: 'Listen to the rhythm of fall ...' will parse as 'Lis ten to the rhy thm of fall'. A music note is a ternary structure that includes three attributions: pitch, duration, and rest. The pitch is a frequency-related scale of sounds, for example, piano keys MIDI number ranges from 21 to 108, each MIDI number correspond to a pitch number, such as MIDI number '76' represents pitch number 'E5'. Duration in music notes denotes the time span of the pitch, for example, a pitch number 'E5' with its duration 1.0, means this music note will last 0.5 seconds in the playing. The rest of the pitch is the intervals of silence between two adjacent music notes and share the same unit with duration. The dataset used for the melody generation from lyrics, to consider the time-sequence information in the pairs, the syllable-level lyrics and music notes are aligned by pairing a syllable and a note.

The initial pre-processing for our dataset is to get the beginning of music notes and corresponding syllables. In our Musical Ternary Modalities (MTM) dataset collection, two rules are as follows.

1) We ensure that each syllable-note paired sample contains 20 notes, it keeps the former first 20 notes as a sample or first 40 notes as two samples.

2) we removed the samples if existing the rest attribute of one note are longer than 8 (around 4 seconds).

Music audio and sheet music can be generated from music notes that satisfies our objective of music three modalities data establishment. Once we get the music note and the syllable-level lyrics, we can extend them to generate the pairs of music audio and sheet music by some high-quality present technologies. All the music data modalities contain temporal structure information, which motivates us to establish fine-grained alignment across different modalities, as seen in Fig. 3.2. In detail, the syllable of lyrics, the audio snippet, and sheet music fragment generated from music notes are aligned.



Lyrics	Lis	ten	to	the	rhy	thm	of	the	fall	ing	rain	tel	ing	me
Music note	D5	C5	C5	A4	A4	G4	G4	F4	G4	F4	F4	D5	C5	C5
Duration	0.5	1	0.5	1	0.5	1	1	0.5	2.5	1	4	1	0.5	1
Reset	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Figure 3.2: An example of fine-grained alignment across three modalities: sheet music, lyrics and music audio.

Music audio is also music sound transmitted in signal form. We add piano instrument in the music channel to create new midi files, and synthesize audios with TiMidity++ tool [1]

Sheet music is created by music note with Lilypond tools [2]. Lilypond is a compiled system that runs on a text file describing the music. The text file may contain music notes and lyrics. The output of Lilypond is sheet music can be viewed as an image. Lilypond is like programming language system, Music notes are encoded with letters and numbers and commands are entered with backslashes. It can combine melody with lyrics by adding the "\addlyrics" command. In our MTM Dataset, sheet music (visual format) for one note and entire sheet music (visual format) for 20 notes are created respectively. Accordingly, each song has single note-level and sequential note-level (sheet fragment) visual formats.

[1] http://timidity.sourceforge.net/.
[2] http://lilypond.org/

3.2 Feature Extraction

In this section, we will explain the motivation and the method of feature extraction for different modality, including audio, visual, text and sheet music.

3.2.1 Audio Feature Extraction

Generally, audio feature extraction is to extract feature from audio signal, which plays main role in speech processing [91, 92], music genre classification [93], and so on. Here, we present a typical model for audio feature extraction, the supervised trained model Vggish.

Vggish model [3] is released by Google Audioset trained on a large YouTube dataset. Firstly, we exploit the librosa2 library to extract the Mel spectrogram feature with some parameters setting like hop size=512, nftt=2,048. Secondly, input the extract Mel-spectrogram feature into the Vggish model, seen in Fig. [**?**].

We choose the supervised learning Viggish model trained with users' annotations to catch the predefined audio label information for audio representation from the output of the model. Finally, the Vggish model project the Mel-spectrogram feature into 128-D audio representation for the input of alignment representation learning of audio-visual cross-modal embedding.

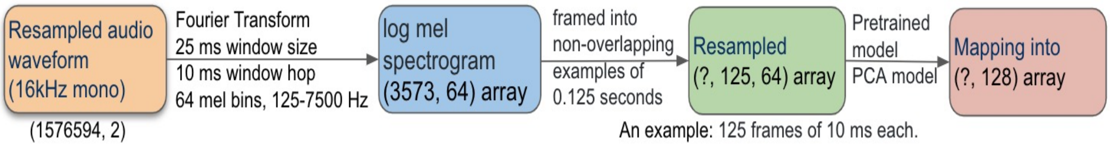

Resampled audio waveform (16kHz mono)
(1576594, 2)
Fourier Transform
25 ms window size
10 ms window hop
64 mel bins, 125-7500 Hz
log mel spectrogram (3573, 64) array
framed into non-overlapping examples of 0.125 seconds
Resampled (?, 125, 64) array
Pretrained model
PCA model
Mapping into (?, 128) array
An example: 125 frames of 10 ms each.

Figure 3.3: The audio feature extracted process with vggish model.

3.2.2 Visual Feature Extraction

In our dataset, hand-crafted features are hard and time-consuming to obtain. With the deep learning model success in the visual feature extraction, we expect to chose Inception V3 pre-trained model to extract useful visual feature.

[3] https://github.com/tensorflow/models/tree/master/research/audioset/vggish

Inception V3 is widely applied in visual feature extraction from image [94, 95], for object recognition and can get good performance on the ImageNet dataset [96, 97]. Visual track in the video can be viewed as an image sequence, the visual features are also can be extracted by the pre-trained Inception V3 model [96]. The input of the Inception V3 model is the pre-processing video that each video is decoded with one frame one second. After feed the decoded videos into the deep CNN architecture and use the ReLU activation and PCA technique in the last layer, the output of Inception V3 model is the frame-level semantic features, the output dimension of visual feature is 1,024.

3.2.3 Sheet music Feature Extraction

Different from other image feature extraction, our feature extraction of sheet music image tries to catch pitches and the segments. In this dissertation, our information extraction of sheet music has two levels, pitch detection, and semantic segments. We apply ASMCMR [98] model trained in audio-sheet retrieval tasks, which learns the correlation between audio clips and corresponding sheet snippet. In our work, the shape of extracted note-level feature and sheet snippet-level feature are (100, 32) and (32,) respectively

3.2.4 Lyrics Feature Extraction

We follow [90] to keep the alignment between syllable and note by representing lyrics in the form of syllable and word level. The syllable-level feature extracted with the syllable skip-gram model, the word-level feature extracted with word skip-gram model used in [90]. These two pre-trained skip-gram models is trained on all the lyrics data, as a logistic regression task with stochastic gradient descent (SGD) optimization. The input of the syllable-level skip-gram model is a sequence of syllable in a sentence, while the input of the word-level model is word sequence in a sentence. The output of syllable-level and word-level skip-gram model is 20 dimensional embedding for each syllable and word, respectively.

The overall statistics of our musical data are shown in Table 3.2. We divided the dataset into 3 parts as training, validation, and testing set by 70%, 15% and 15%. The number of training, validation and testing set are 13,535, 2800 and 2800 respectively.

Table 3.1: General statistics of two modalities used different pre-trained models in our experiments

Modalities	Feature Extractor	Dimension
audio	Vggish	(10, 128)
	Soundnet	(10, 400)
visual	Inception V3	(10, 1024)
	I3D	(10, 400)

Table 3.2: General statistics of three data modalities used in our experiments

Modality	Feature Extractor	Dimension	Number
Audio	Vggish	(20, 128)	14,454
Lyrics	Skip-gram	(20, 20)	14,454
Sheet music	Lilypond&ASMCMR	(20, 100, 32)	14,454

3.3 Evaluation Metrics

3.3.1 The Difference of Distance and Similarity

Normally, the higher similarity the shorter distance, here we try to discuss Euclidean distance and cosine similarity.

Euclidean distance used in previous face identity tasks [99, 100], the distance like $D(T(i) - T(j)) = ||T(i) - T(j)||_2^2$ 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. In our experiment, we apply a cosine similarity for the final representation comparison at the end of the whole architecture. Our distance metric is defined by following equation.

$$||x, y||_{cosine-distance} = 1 - \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k}\sqrt{\sum_{k=1}^{n} y_k}}, \tag{3.1}$$

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.

Cosine similarity and cosine distance have 1.0 difference. Which all very popular in cross-modal embedding metrics. It is defined as follows.

$$Similarity = cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}} \tag{3.2}$$

where x_i and y_i are the components of vector A and B respectively.

Seen in Fig.3.4. the overview of Euclidean distance and cosine similarity. While cosine (θ) focuses on the angle between vectors, euclidean distance (d) like applying a ruler to measure the distance.

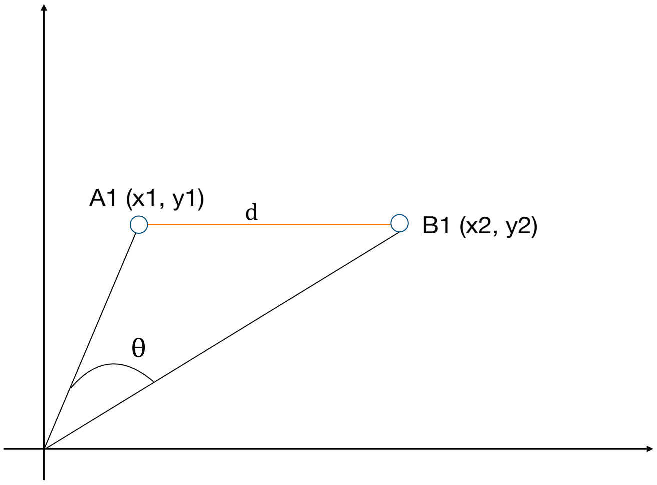


Figure 3.4: The overview of Euclidean distance (d) and cosine similarity (cos θ)

3.3.2 The Evaluation Metrics in our Experiments

This section is to introduce some metrics to evaluate the performance of our approaches in our experiments. They can be summarized as two groups. One group is for supervised cross-modal retrieval, such as mean average precision (MAP), Precision and recall curve (PRC), other group is for unsupervised cross-modal retrieval: Recall@K, MedR and MeanR. We exploit the common evaluation metrics in most previous work [101] on unsupervised cross-modal retrieval.

Mean Average Precision

is the mean of average precision (AP) for all the queries. To calculate the MAP value, it firstly required to compute the AP for each query in the rank list. The AP is calculated by

$$AP = \frac{1}{Rel} \sum_{k=1}^{N} p(k) \cdot rel(k) \tag{3.3}$$

where Rel is the number of relevant documents that shares the same label with the query. $p(k)$ is the precision in the top k of rank list, $rel(k)$ is a indicator function. When the value is 1 if the k^{th} candidate in the rank list has the same label as query. If the value is 0, both the k^{th} candidate and query have the different labels.

Precision-Recall Curve

Precision-Recall Curve (PRC) a graph with precision on y axis and recall on x axis. Precision is about the percentage of the numbers of relevant items in the top k of rank list, while recall relate to the percentage of the numbers of retrieved items in all relevant items in the database.

Recall@K (K=1, 5, 10)

R@K (Recall at K) is to compute the percentage of relevant items appear in the top-k of rank list for the query. We calculate the performance of our unsupervised architecture by the average of R@1, R@5 and R@10 for all queries respectively.

Median Rank and Median Rank

Median Rank compute the median of the relevant items in the rank list. As a popular metric to evaluate the performance of unsupervised task, the lower value it obtain the better performance it achieve. Similarly, Mean Rank measure the mean rank of all relevant items and the higher value means the better performance of proposed approach.

4

Audio-Visual Aligned Representation Learning for Music Video Retrieval with Supervised Deep CCA

4.1 Background and Motivation

Deep alignment learning is a very important research topic in the area of multimedia and computer vision, which aims at learning aligned representations between different modalities to bridge the modality gap. It has widely discussed in same special domain, such as image-text [12, 28]. The cross-modal retrieval in the music area, applying music audio clip to obtain corresponding visual content is a imaginative application to raise the experience of users. Image when you go cross a mall, a fantastic song you heard and you want to record the song clips to find the music video, as shown in Fig. 4.1. Learning aligned representation for audio-visual is non-trivial. Unfortunately, few works has reported that the aligned learning within temporal structure of two

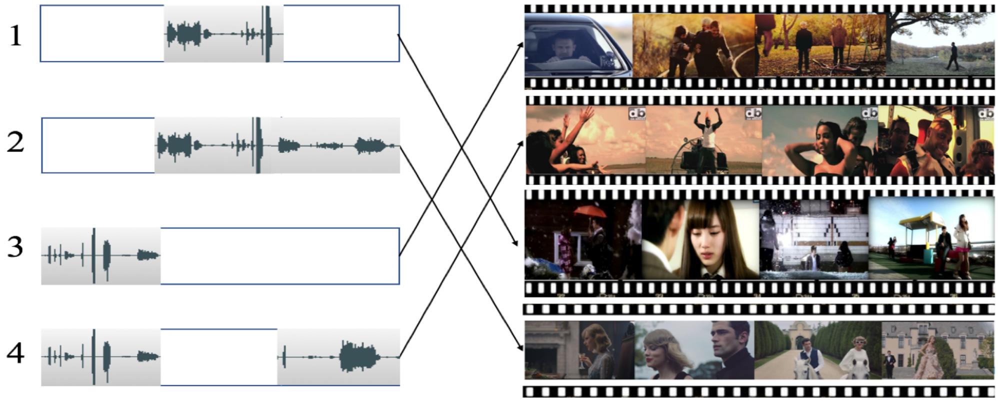


Figure 4.1: The framework of music video retrieval: Applying a short audio clip to find related music video by computing the similarity between the audio clip and the visual items in the video collection.

modalities, which will be next generation for cross-modal representation learning.

A music video contains visual and audio modalities, which are embedded in musical temporal sequences to express music theme and story. Moreover, as a special form of expression, a music video also conveys strong feelings and emotions, which are semantically contained in audio and visual modalities. That is to say, music emotion is delivered by both audio and visual modalities in music video. This motivates us to learn a aligned representation subspace where music audio and visual contents are assumed with same semantically meaning.

The rapid growth of music videos on the Internet allow us to learn the align representation between audio and visual sequences. Audio and visual are two tracks of music video to together convey the music mood and feelings. Moreover, as a special form of expression, a music video also conveys strong feelings and emotions, which are semantically contained in audio and visual modalities. That is to say, music emotion is delivered by both audio and visual modalities in music video. This motivates us to learn a aligned representation subspace where music audio and visual contents are assumed with same semantically meaning.

In this work, we study how to use audio to retrieve music video under a realistic situation: with a segment of music audio that has a variable length as a query, the system automatically finds the music video that is similar to this audio with regard to emotions. In other words, an audio with an arbitrary length can retrieve a longer or full-length music video. It is natural for users to search music video in this way. However, this is a challenging research issue because audio and video are different modalities that have different low-level features with different properties of temporal structures. To this end, we propose a novel audio-visual embedding algorithm by Supervised Deep Canonical Correlation Analysis (S-DCCA) that projects audio and video into a joint feature space to bridge the gap across different modalities. This also preserves the similarity among audio and visual contents from different videos with the same class label and the temporal structure. In addition to selecting 10K music video data from the YouTube-8M dataset, most importantly, several contributions are made in this paper as follows:

i) We are the first try to study how to retrieve a full-length silence music video by an variable length of audio clips as query.

ii) We propose to select k representative audio chunks based on emotion features extracted by a Long Short-Term Memory (LSTM)-based attention model, which serve as audio summary meanwhile conserving the temporal structure.

iii) We develop an end-to-end deep learning architecture for audio-visual aligned representation learning by learning the semantic correlation.

iv) The experimental results suggest that our algorithm has competitive performance compared with the state-of-the-art methods.

4.2 Architecture

The section is to explain the detailed architecture of our musical audio-visual aligned representation learning method.

4.2.1 Neural Attention Modeling

Aspired by the [36], they applied a bi-directional LSTM to achieve the selection of short audio clips candidates for audio summarization. We exploit the same pre-trained model to select the top k chunks based on the contribution of the audio clips for emotion similarity.

LSTM understand the present audio frame assisting from previous audio frames, and can preserve the "long-term dependencies" information because it consist of self-loops units. The weights of self-loops are updated by four components at the same time.

1) **Input gate** determine what kind of values should be updated. It relies on the present input x_i and the previous hidden state h_{t-1} as follows:

$$s_t = \sigma(b_i + W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1}). \tag{4.1}$$

2) **Forget gate** judge which information will be forgotten from the present cell, seen as follows:

$$f_t = \sigma(b_f + W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1}). \tag{4.2}$$

3) **Cell state** c_t renovate the old state c_{t-1}, shown as follows:

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c). \tag{4.3}$$

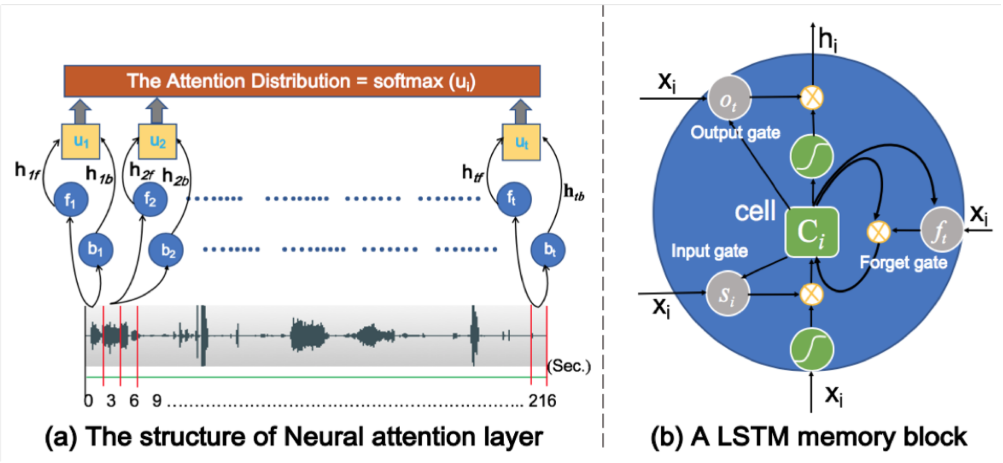
4) **Output gate** determines what the next hidden state will be, seen as follows:

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o). \tag{4.4}$$

$$h_t = o_t tanh(c_t) \tag{4.5}$$

where x_t denotes the present input variable, h_{t-1} represents the last hidden state, W and b are the weight and bias of the updated function.

LSTM is a one way computation method. In order to consider both past and future information, the extension of LSTM networks adds one more layer with the opposite temporal sequence and is named bi-directional LSTM, as shown in Fig. 4.2. In our works, each audio is divided into 72 chunks, each with 3 seconds. Then, the bi-directional LSTM model is applied on each chunk. In the attention model, the input of bi-directional LSTMs is the output of global max-pooling layer, which is the first attention layer to compute the contribution scores of different audio chunks. The attention score u_t of the t-th chunk can be computed as follows.



(a) The structure of Neural attention layer

(b) A LSTM memory block

Figure 4.2: (a) The neural attention model can select the top k audio chunks, which have the top k contribution for the emotions. (b) A LSTM memory block consists of four components, the cell state and three gates.

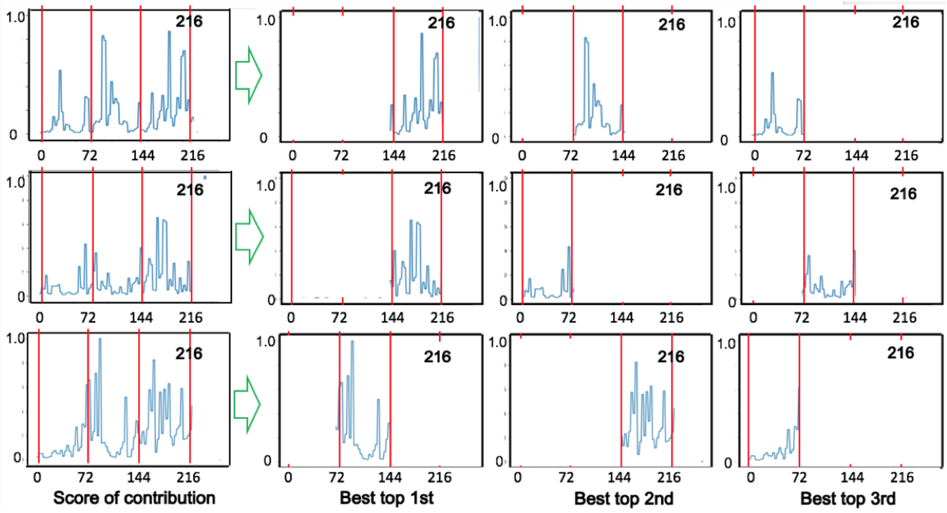


Figure 4.3: Emotion learning model for evaluating the contribution of each chunk to emotions. When an original 216 seconds audio is divided into 3 chunks, the model calculates the contribution score of each chunk, which helps to obtain the top $k - th$ chunk.

$$u_t = W^T tanh(W_f h_{tf} + W_b h_{tb} + \beta), \tag{4.6}$$

where h_{tf}, h_{tb} are the outputs of forward and backward LSTM for the t-th chunk, the W^T, W_f, W_b and β are the weight parameters of attention score function. When the attention score is obtained, the attention distribution θ is calculated by a softmax function:

$$\theta = softmax(u_t). \tag{4.7}$$

We regard this architecture as an emotion learning model [102], which is trained over the MER31K dataset, using emotion tags from AllMusic[1]. The detail of selecting audio segments achieved by emotion learning model is shown in Fig.4.3. Firstly, the emotion learning model is used to evaluate the contributions of each chunk to emotions. The contribution score allows us to rank the chunks. Secondly, in the ranked chunks, the best top k are selected. For instance, the first audio in the Fig. 4.3 is divided into 3 chunks, and depending on the contribution scores, the third chunk is selected as the best one, because it has the highest score within the audio.

4.2.2 Supervised Deep Canonical Correlation Analysis and Distance Similarity

CCA[103] is a classical approach for correlation analysis among two or more modalities. Its core idea is to learn projection matrices that map features of different modalities into the same space, where the correlation between similar items of different modalities are maximized.

Denote $X \in R^k$ as an audio feature, $Y \in R^l$ as a visual feature, and denote W_x, and W_y as matrices that linearly map X and Y to the same space, then W_x and W_y are found by maximizing the correlation between $W_x^T X$ and $W_y^T Y$, as follows:

[1] http://www.allmusic.com/moods

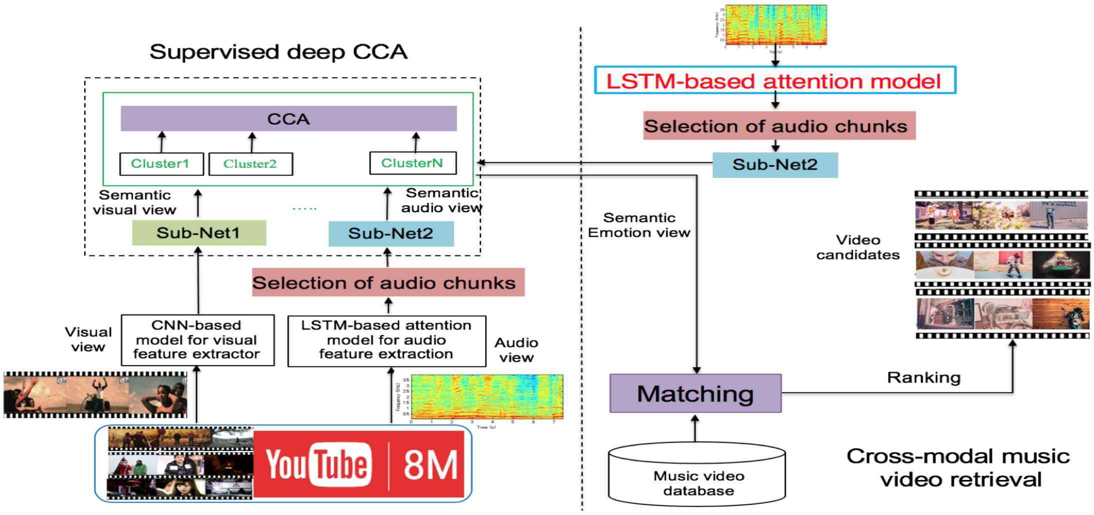


Figure 4.4: Audio-visual embedding architecture through S-DCCA. (left) During the training process, the model learns the correlation between audio and visual content. (right) Using audio chunks as input to retrieve music videos.

$$(W_x, W_y) = \arg\max_{(W_x, W_y)} \frac{W_x^T \Sigma_{xy} W_y}{\sqrt{W_x^T \Sigma_{xx} W_x \cdot W_y^T \Sigma_{yy} W_y}} \tag{4.8}$$

where Σ_{xx} and Σ_{yy} represent the covariance matrices of X and Y, respectively and Σ_{xy} is their cross covariance matrix.

DCCA extends CCA, realizing non-linear projections by deep neural networks (DNN). Assume the output of $(i-1)^{th}$ layer is X_{i-1} and Y_{i-1} ($X_0 = X$ and $Y_0 = Y$), and $W_{xi}, W_{yi}, b_{xi}, b_{yi}$ are the weights and biases of the i^{th} layers. Then, the i^{th} layer outputs $X_i = s(W_{xi}^T X_{i-1} + b_{xi})$, $Y_i = s(W_{yi}^T Y_{i-1} + b_{yi})$ at two branches, where s: $R \longrightarrow R$ is a nonlinear function. The output of the final (d^{th}) layer are $f_x = s(W_{xd} X_{d-1} + b_{xd})$, $f_y = s(W_{yd} Y_{d-1} + b_{yd})$. Let θ_x represent the parameters W_{xi}, b_{xi}, i = 1, ..., d, and θ_y represent the parameters W_{yi}, b_{yi}, i = 1, ..., d. They are optimized by

$$(\theta_x^*, \theta_y^*) = \arg\max_{(\theta_x, \theta_y)} corr(f_x(X, \theta_x), f_y(Y, \theta_y)). \tag{4.9}$$

Supervised deep CCA does not merely consider one-to-one match between all pairs of audio-visual data and apply deep CCA to learn the correlation. In order to preserve the similarity among items with the same class label, audio and visual contents from different videos with the same class label are formed as new relevant pairs to increase the number of training samples.

In the training process, maximizing the CCA objective function $G(W_x^T \Sigma_{xy} W_y)$ to obtained the linear projections weight W_x, W_y and non-linear function f_x, f_y as follow.

$$\begin{aligned} (W_x, W_y, f_x, f_y) = \underset{(W_x, W_y, f_x, f_y)}{\arg\max} \; G(W_x^T \Sigma_{xy} W_y), \\ s.t. W_x^T \Sigma_{xx} W_x = I, W_y^T \Sigma_{yy} W_y = I. \end{aligned} \tag{4.10}$$

where the covariance matrices Σ_{xx}, Σ_{xy} and Σ_{yy} are computed as.

$$\Sigma_{xx} = E_i(f_X^{(i)} f_X^{(i)T}) + r\mathbf{I}, \tag{4.11}$$

$$\Sigma_{yy} = E_i(f_Y^{(i)} f_Y^{(i)T}) + r\mathbf{I}, \tag{4.12}$$

$$\Sigma_{xy}^{(1)}(d) = E_{i \in d}(f_X^{(i)} f_Y^{(i)T}), \tag{4.13}$$

$$\Sigma_{xy}^{(2)}(d) = E_{i,j \in d, i \neq j}(f_X^{(i)} f_Y^{(j)T}), \tag{4.14}$$

$$\Sigma_{xy} = \sigma \cdot E_{d \subset D}(\Sigma_{xy}^{(1)}(d)) + (1 - \sigma) \cdot E_{d \subset D}(\Sigma_{xy}^{(2)}(d)). \tag{4.15}$$

where N is the number of all pairs. The σ value decide two factor of the number of training dataset, different from DCCA, S-DCCA considers pairs between audio and visual contents from videos with the same class label, including those pairs formed from different videos, as shown in (4.14). similar to DCCA, all parameters are optimized by formulation (4.13). The left side of Fig. 5.1 shows the whole process.

4.2.3 K-means Clustering

k-means clustering is a very popular unsupervised learning method for cluster analysis in data mining. k-means clustering enables n variables to be separated into k clusters based on the nearest mean, where k is usually pre-defined by users.

Given a set of variables X=(x_1, x_2, $\cdots$, x_n), where each variable $x_i \in X$ is a d-dimensional vector. In order to cluster them into k groups $G = g_1, g_2, ..., g_k$ $(k < n)$, firstly, a common method is to randomly choose k values from X as initial cluster centers, then iteratively update the cluster center after assigning each variable x_i to its closest cluster till the cluster center never changes. The objective function is defined as follows:

$$\arg\max_{G} \sum_{i=1}^{k} \sum_{x \in g_i} ||x - u_i||^2 \tag{4.16}$$

where u_i is the mean of points or cluster center of G_i. In our experiments, we allocate 3 annotated audios for each 10 predefined categories (angry, tender, bitter, cheerful, fun, bright, happy, anxious, calm and warm) to compute the initiated mean u_0. We use the k-means method to cluster all audios into 10 semantic classes based on the emotion features.

4.2.4 Matching and Ranking

It is not easy to recognize emotion inside the visual modality, because the visual feature of the dataset is high-level semantic features without clear emotion expression like facial expression changes or body movement. However, the high-level semantic information extracted or trained from complicated deep network is able to represent emotion attributes contained in music. Based on this background, we design a S-DCCA model to learn the correlation between audio and video, which enables us to use audio to retrieve video clip.

The audio-visual embedding is to map audio chunks and visual features to a common space. This space links audio chunks and visual feature in terms of emotion, and enables us to implement cross-modal music video retrieval based on emotion similarity. In the cross-modal retrieval, given an audio chunk or multiple chunks as query, we calculate the similarity between the query audio chunks and each of the visual features from the database in the emotion-based embedding space. We use the cosine similarity between $f_x(X, \theta_x)$ and $f_y(Y, \theta_y)$ as the similarity metric, which is defined as follows.

$$Cos(f_x, f_y) = \frac{f_x f_y}{||f_x||.||f_y||} \tag{4.17}$$

The detail of our architecture is shown in Fig. 5.1. which consists of 2 branches: audio branch and visual branch. Firstly, the pre-trained VGG16 model is used to extract frame-level audio feature and the pre-trained Inception model is used to extract frame-level visual feature, for all data in the dataset. Secondly, the frame-level visual feature is represented as video-level feature by the max pooling method. As for audio branch, we load frame-level audio feature into the pre-trained emotion learning model [102] to extract emotion features , based on which the best top k chunks are selected to do music video retrieval, then feed them into Sub-Net1 and Sub-Net2 respectively. Thirdly, based on the extracted emotion features, we apply k-means to cluster the audio into 10 groups. Fourthly, the visual video-level feature and emotion of top k audio chunks are fed into 4 fully connected layers, which generates compact features. Finally, CCA components of these compact features are used to compute the similarity between video and audio chunks.

4.3 Experiments

The performance of the proposed S-DCCA for cross-modal music video retrieval are evaluated in this section, with the studies on the influence of the number of chunks and cross-modal music video retrieval by audio.

4.3.1 Experiment Setting

The frame-level video feature in YouTube-8M is computed one frame per second, according to the pre-trained emotion learning model. We divide the 216 second frame-level audio feature into 72 chunks.The attention model is applied to each chunk to calculate the contribution score of emotion, and each 3 second share the same score. Finally, the result of max pooling is regarded as the score of emotion for each chunk.

The following parameters are used in our experiments:

- Network parameter. Both the audio and the branch have 4 hidden layers. The number of units per layer is 512, 512, 256, 256 in the visual branch, and 128, 128, 64, 64 in the audio branch. The number of CCA component is 30. We set the probability of dropout to 0.2 and apply *tanh* as the activation function in each hidden layer and use *sigmoid* function in the final layer.

- Experiment parameter. Train batch size is 512 and test batch size is 64. The number of training epochs is 50.

- We run the experiments with 5 fold cross-validation and get the average performance.

- The *RMSProp* optimizer is used and the learning rate is set to 0.001.

4.3.2 Baseline

Multi-view [104] learning is a technology in machine learning that learn one function per view to model multiple views and optimizes all functions to remove the cross-view gap.

CCA [105] algorithm is to find the correlations between two multivariate sets of vectors by linear projections, which depends on singular value decomposition.

KCCA [106] is also a method to extract common features from two data sets Instead of the linear correlation KCCA tries to obtain non-linear correlation through the kernel method, which uses Gaussian kernel and set parameter β=0.4.

DCCA [103] is to learn the nonlinear transformations of two data sets such that outputs are highly correlated.

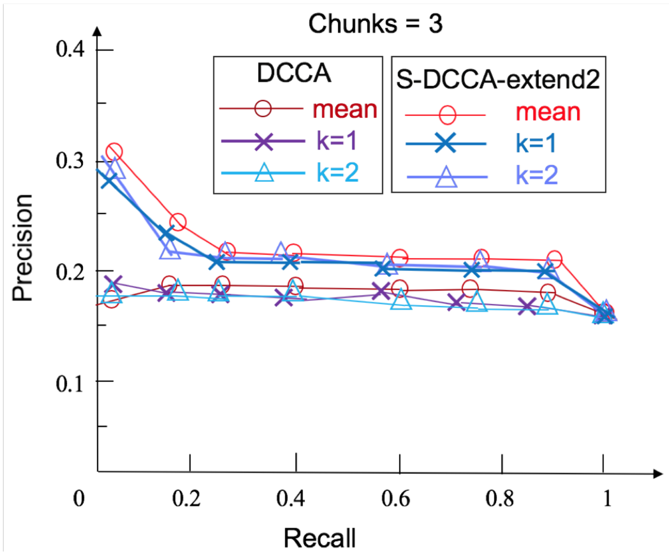


Figure 4.5: Precision-recall curve with the number of chunks set to 3, where “mean" denotes using the average of frame level audio feature as query, k (=1, 2) is the number of audio chunks selected as query.

C-CCA [4] (Cluster-CCA) is a CCA variant. Different from standard CCA. C-CCA algorithm clusters each data set into several groups or classes and tries to enhance the intra-cluster correlation.

4.3.3 Evaluation and Analysis

Our experiments of S-DCCA use three different training data sets to obtain three different models. The basic C-CCA and S-DCCA model are trained by the 8000 one-to-one pairs. To enhance to intra-cluster correlation, we further consider the correlation between audios and visual contents from different videos of the same cluster, to learn the relationship between the two modalities. We also try to construct more audio-visual pairs during the training. The C-CCA-extend1 and S-DCCA-extend1 are trained by around 0.8 million pairs, C-CCA-extend2 and S-DCCA-extend2 models by around 1.5 million pairs. where the former -extend1 model uses 50% of all music videos of a cluster to form training pairs with each audio in the cluster, and the latter -extend2 model applies 100% of all music videos in the same cluster to form training pairs.

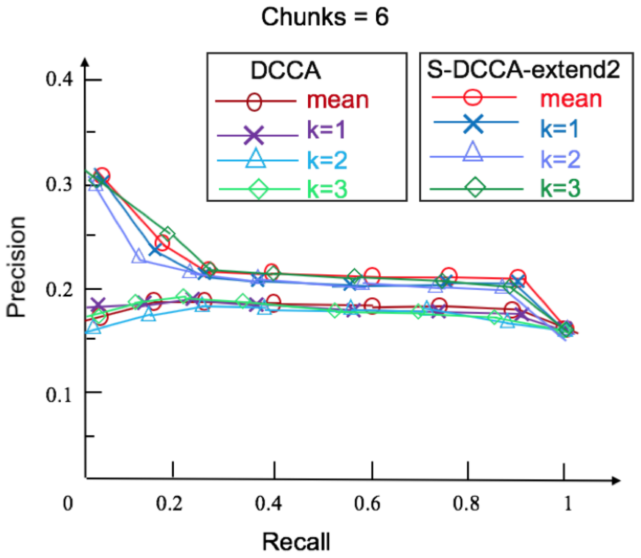


Figure 4.6: Precision-recall curve with the chunks=6, where “mean" denotes using the average of frame level audio feature, k(=1, 2, 3) is the number of audio chunks selected as query.

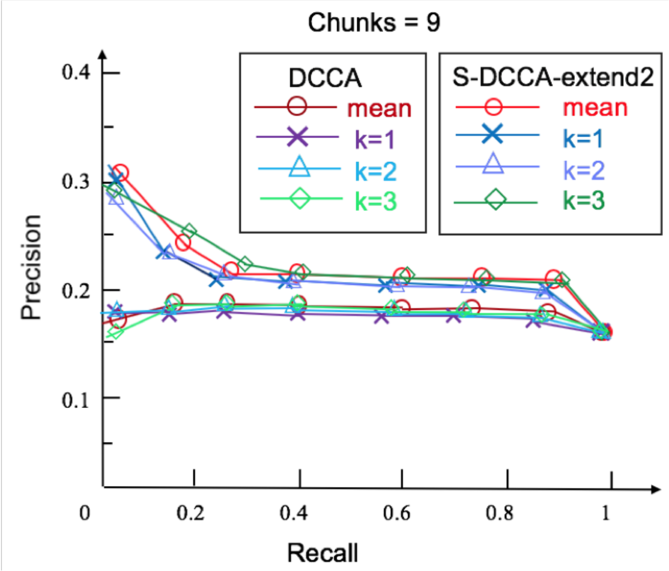


Figure 4.7: Precision-recall curve with the chunks=9, where “mean" denotes using the average of frame level audio feature, k (=1, 2, 3) is the number of audio chunks selected as query.

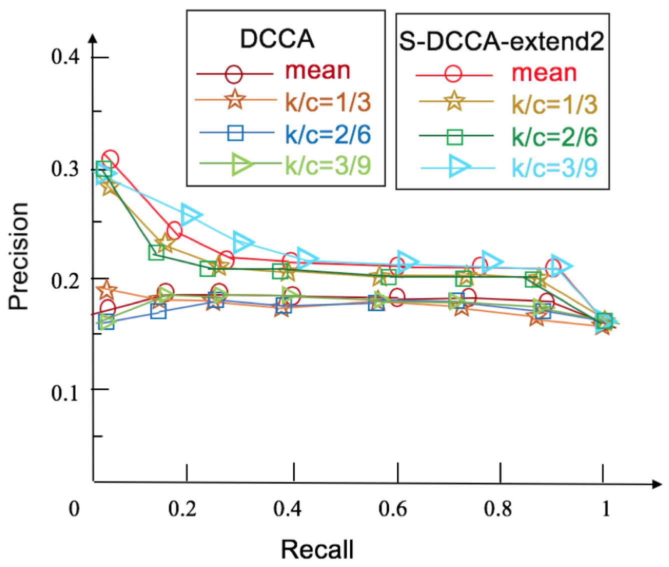


Figure 4.8: Precision-recall curve, achieved by changing the number of output, where k (=1, 2, 3) is the number of chunks selected from all chunks (c) of an audio as query; for example, k/c=1/3 denotes selecting 1 chunk from an audio that is divided into 3 chunks. "mean" denotes using the average of the whole audio as query.

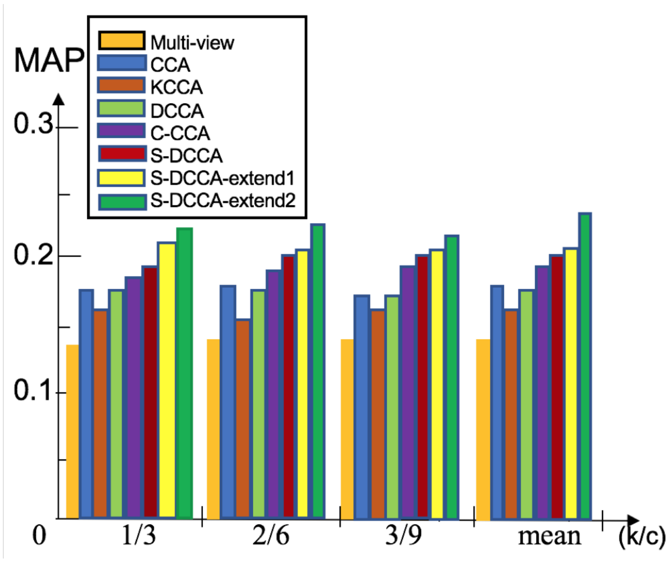


Figure 4.9: Mean average precision when using different numbers of audio chunks selected as query for video retrieval, k denotes the number of chunks selected as query, c denotes the number of overall chunks that the audio is divided into.

Table 4.1: The MAP results of different methods under different configurations.

k/chunks	**1/3**	**2/6**	**3/9**	**mean**
Multi-views	14.02	14.36	14.25	14.58
CCA	18.34	18.39	18.32	18.35
KCCA	17.54	17.04	17.49	17.80
DCCA	18.35	18.39	18.22	18.40
C-CCA	18.51	19.60	19.73	19.72
S-DCCA-extend1	20.19	20.04	20.00	20.14
S-DCCA	**21.38**	**21.43**	**21.24**	**21.76**

We use the precision-recall curve to draw the tendency of results as the number of outputs increases so as to compare our S-DCCA model with DCCA model and S-DCCA-extend2 model. Our model tries to leverage the temporal structure inside the query audio, and each query audio is divided into 3, 6, or 9 chunks, from which k chunks are selected as the actual query. In order to investigate the overall performance of our S-DCCA, we use MAP as the metric and compare S-DCCA with others CCA variants (DCCA, C-CCA, KCCA), we set the same dimension of embedding for all methods, and set the same hidden layers structure for DCCA, S-DCCA, S-DCCA-extend1, and S-DCCA-extend2. The correct retrieved video in the rank list which has the same category as query, otherwise it is incorrect video.

Figs. 4.5, 4.6, 4.7 demonstrates the precision-recall curve, comparing DCCA and S-DCCA-extend2 model. The pair of precision and recall value is achieved by changing the number of music videos output. Generally, with the increase of the number of music videos output, the recall increases and the precision decreases. In the S-DCCA-extend2 model, these three figures show that precision starts with the highest value and then sharply decreases before recall arrives at 0.2, then precision almost remains stable as recall increases to 1.0. As is known, the query and the model as two main factors control the curve trend. As for the query factor, when each audio is divided into 3 or 6 chunks, the precision and recall curves of the selected chunks and full-length audio are very close. But when each audio is divided into 9 chunks, and 3 chunks are selected as query, the performance is better than other configurations when the number of output is small. This infers that the 3 chunks have most contribution of emotion and this kind of information is helpful for cross-modal retrieval. As for the model factor, S-DCCA-extend2 is better than DCCA, which indicates that more videos in the output belong to the same cluster as the query in S-DCCA-extend2, than in DCCA.

We also investigate the influence of the number of overall chunks and the number of chunks selected. Fig. 4.8, shows that with the same volume of audio information as query, when the audio is divided into 9 chunks and 3 chunks are selected as the query the S-DCCA-extend2 model achieves the best performance (precision ranges from 26.6% to 23.8%; recall ranges from 0.20 to 0.41).

In order to further study the influence of the number of overall chunks and the number of chunks selected as query , the MAP results of different models are compared in Table 4.1 and Fig. 4.9. As for the number of chunks selected, generally there is no big difference in MAP when the same model is used. When the same audio information is used as query, comparing the MAP results among different models, it shows that the training process explicitly exploiting the cluster information generally outperforms the one without cluster information.

As a result, S-DCCA (and S-DCCA-extend1, S-DCCA-extend2) and C-CCA (and C-CCA-extend1, C-CCA-extend2) can get higher MAP than Multi-views, CCA, KCCA, and DCCA. It indicates that the correlation learning based on both cluster information and instance features is better than those using instance features only. With the increases in the volume of the training data, from two groups, group 1: C-CCA, C-CCA-extend1, C-CCA-extend2, and group 2: S-DCCA, S-DCCA-extend1, S-DCCA-extend2, the MAP gets higher and higher. It proves that considering all possible pairs within two data sets for each label cluster can get better performance than one-to-one pairs, and it also illustrates the limited training data cannot well learn the correlation between audio and visual feature in this case. Generally, using parts of audio as queries to do retrieval can get close performance as in this case where full-length audio is used as queries.

4.3.4 Summary

We proposed a supervised deep CCA model to learn a semantic space where audio and visual data from music video, which are in different modalities, are linked to learn the cross-modal correlation. Besides the pairwise similarity, the semantic similarity or alignment between audio and visual contents from different videos in the same cluster is also explicitly considered. An end-to-end deep architecture that represents an audio sequence as representative chunks is studied. The experimental evaluation run over MV-10K data selected from You Tube-8M proves the effectiveness of the proposed deep audio-visual aligned representation learning algorithm in cross-modal music video retrieval. We proposed a more advanced architecture for the audio-visual aligned representation learning in the next section, instead of training with audio-visual pairs, we established triplet with audio-visual modalities based on the component of triplet belong to the same category or not, and our model are trained by the built triplet.

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

5.1 Background and Motivation

The web has progressively become the multimedia content platform, to be beneficial from the relationship of multimedia content will 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 learning aligned representation between every two modalities of data for cross-modal retrieval tasks, which has made big successes in multimodal information retrieval, such as image-text [10, 12], audio-text [14], and video-text [13]. In particular, the visual and auditory senses of human being are the most important ways to understand the living environment and understand the world. 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 [107]. This motivates us to mimic the mutual-aid based learning process and extract cognitive patterns from human being.

Cross-modal retrieval between data from different modalities has a challenge of the heterogeneous gap of data structure among the modalities, which requires us to formulate a aligned representation space, where the similarity of different data modalities suggests the semantic matched pair between their former inputs by correlation learning. Recently, most methods for correlation learning are to bridge the gap of different modalities by learning aligned representation, which has achieved great success in cross-modal retrieval tasks [108, 109, 48, 110].

The typical representation learning method CCA [109] 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 [23] 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 [103, 46, 47, 11]. For example, Deep Canonical Correlation Analysis (DCCA) [103], 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 aligned 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 [4] 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 aligned 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[28] 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 alignment representation 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 every two instances with 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. To evaluate the extendability of our algorithms, VEGAS dataset [89] 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.

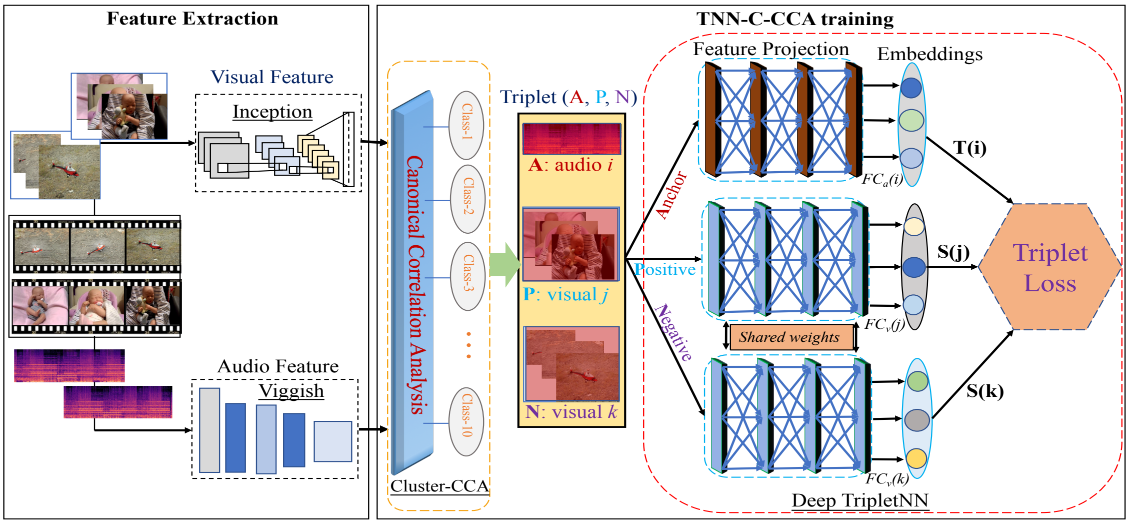


Figure 5.1: The overall framework of our TNN-C-CCA model. It consists of two parts: feature extraction and TNN-C-CCA training. We apply Inception V3 and Vggish model to extract feature, then explore cluster-CCA to learn the correlation with cluster segregating and select triplets as input for deep TNN 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.

5.2 Architecture

Table 5.1: Configuration of TNN-C-CCA

log mel-spectrogram audio inputs	96x64
Output of visual branch	L[1]*1024
Output of audio branch	L*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

Our deep architecture generally can be divided into two different parts: feature extraction and TNN-C-CCA training, as shown in Fig. 5.1. The configuration of

[1] L is the number of frames in a video, by decoding each video at one frame per second.

TNN-C-CCA used in this work is shown in Table 5.1. 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.

5.2.1 The 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 new axis for x and y by the weight $w \in R^A$ and $u \in R^B$ respectively, such that the correlation between these two sets is maximized. The correlation can be defined as follows:

$$corr = \frac{w' C_{xy} u}{\sqrt{w' C_{xx} w} \sqrt{u' C_{yy} u}}, \tag{5.1}$$

$$C_{xx} = E[xx^T] = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T, \quad C_{yy} = E[yy^T] = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T, \quad C_{xy} = E[xy^T] = \frac{1}{n}\sum_{i=1}^{n} x_i y_i^T, \tag{5.2}$$

Where *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:

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} w = \lambda^2 w, \tag{5.3}$$

CCA has been successfully utilized to solve the 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 [109] with consideration of the cluster segregating by training on all possible pairs in the cluster across modalities, then apply CCA to learn the projections.

$$corr = \frac{w^{'}C^{'}_{xy}u}{\sqrt{w^{'}C^{'}_{xx}w}\sqrt{u^{'}C^{'}_{yy}u}}, \tag{5.4}$$

The three types of variances can be formulated as follows:

$$C^{'}_{xx} = \frac{1}{L}\sum_{c=1}^{C}\sum_{i=1}^{|X_c|}|Y_c|x_i^c x_i^{cT}, \quad C^{'}_{yy} = \frac{1}{L}\sum_{c=1}^{C}\sum_{j=1}^{|Y_c|}|X_c|y_j^c y_j^{cT}, \quad C^{'}_{xy} = \frac{1}{L}\sum_{c=1}^{C}\sum_{i=1}^{|X_c|}\sum_{j=1}^{|Y_c|}x_i y_j^{cT}, \tag{5.5}$$

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 (5.9). Here we assume that the covariance is calculated for the zero-mean random variables.

5.2.2 Deep Triplet Neural Network

The Deep Triplet Neural Network is an end-to-end training, as shown in Fig. 5.1, which is optimized by triplet loss [99] 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(.) and S(.) 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. 5.2. Triplet loss will pull Anchor and Positive samples, simultaneously push Anchor and Negative samples. The condition is represented as follows.

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

$$Loss = Max\{\sum_{i}^{N}[||T(i) - S(j)||_{cosine-distance} - ||T(i) - S(k)||_{cosine-distance} + \alpha], 0\}, \quad (5.6)$$

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.(5.12), 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 [100], 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 [99].

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 [99] and select triplets to remove all negative/positive samples in a batch when its triplet loss is zero.

5.2.3 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:

$$corr = \frac{w^{'} C_{xy} u}{\sqrt{w^{'} C_{xx} w} \sqrt{u^{'} C_{yy} u}}, \tag{5.7}$$

$$C_{xx} = E[xx^T] = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T, \quad C_{yy} = E[yy^T] = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T, \quad C_{xy} = E[xy^T] = \frac{1}{n}\sum_{i=1}^{n} x_i y_i^T, \tag{5.8}$$

Where $corr$ 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:

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} w = \lambda^2 w, \tag{5.9}$$

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 [109] 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.

$$corr = \frac{w^{'} C_{xy}^{'} u}{\sqrt{w^{'} C_{xx}^{'} w} \sqrt{u^{'} C_{yy}^{'} u}}, \tag{5.10}$$

The three types of variances can be formulated as follows:

$$C_{xx}^{'} = \frac{1}{L}\sum_{c=1}^{C}\sum_{i=1}^{|X_c|} |Y_c| x_i^c x_i^{cT}, \quad C_{yy}^{'} = \frac{1}{L}\sum_{c=1}^{C}\sum_{j=1}^{|Y_c|} |X_c| y_j^c y_j^{cT}, \quad C_{xy}^{'} = \frac{1}{L}\sum_{c=1}^{C}\sum_{i=1}^{|X_c|}\sum_{j=1}^{|Y_c|} x_i y_j^{cT}, \tag{5.11}$$

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 (5.9). Here we assume that the covariance is calculated for the zero-mean random variables.

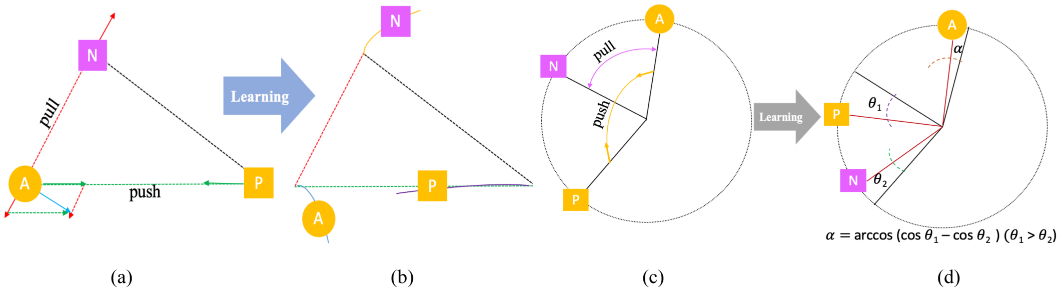


Figure 5.2: (a) and (b) show the traditional triplet loss minimizes the Euclidean distance between (anchor, positive) and (anchor, negative) with a fixed margin and optimal gradient back-propagation direction; (c) and (d) present our triplet loss through minimizing the Cosine distance between (anchor, positive) and (anchor, negative) with fixed margin and optimal gradient back-propagation direction.

5.2.4 Deep Triplet Neural Network

The Deep Triplet Neural Network is an end-to-end training, as shown in Fig. 5.1, which is optimized by triplet loss [99] 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(.) and S(.) 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. 5.2. Triplet loss will pull Anchor and Positive samples, simultaneously push Anchor and Negative samples. The condition is represented as follows.

$$\begin{gathered} ||T(i) - S(j)||_{cosine-distance} + \alpha < ||T(i) - S(k)||_{cosine-distance}, \\ Lab^i = Lab^j, \quad Lab^i \neq Lab^k(i \neq k), \quad \forall(i,j,k) \in \Lambda, \end{gathered} \tag{5.12}$$

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

$$Loss = Max\{\sum_{i}^{N}[||T(i) - S(j)||_{cosine-distance} - ||T(i) - S(k)||_{cosine-distance} + \alpha], 0\}, \tag{5.13}$$

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.(5.12), 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 [100], 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 [99].

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 [99] and select triplets to remove all negative/positive samples in a batch when its triplet loss is zero.

5.3 Experiments

5.3.1 Training Setting

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

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.

Table 5.2: 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 (%)**	
	audio→visual	visual→audio	audio→visual	visual→audio
CCA [109]	32.43	32.11	18.38	18.17
KCCA [23]	28.65	27.24	17.81	17.03
DCCA [103]	41.43	42.15	18.43	18.21
C-CCA [4]	65.16	64.35	19.71	19.62
C-KCCA [4]	32.41	32.74	18.38	18.11
C-DCCA [5]	70.34	69.27	21.79	20.08
UGACH [111]	17.18	17.07	11.11	11.40
AGAH [112]	57.82	56.16	20.74	20.19
UCAL [113]	42.68	41.53	18.82	18.47
ACMR [56]	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

5.3.2 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 5.2 with MAP metric and Fig. 5.3 with PRC. We implement our architecture compared with some existing CCA-variant approaches and non-CCA methods: CCA [109], DCCA [103], KCCA [23], C-CCA [4], C-KCCA [4] C-DCCA [28], AGAH [112] 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.

According to the experience of our experiments, when the correlation component is set to 10, the CCA-variant approaches can get the best performance[28, 14]. 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 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 via 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 5.2 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 discusses the pairwise correlation but also studies the category-based similarity

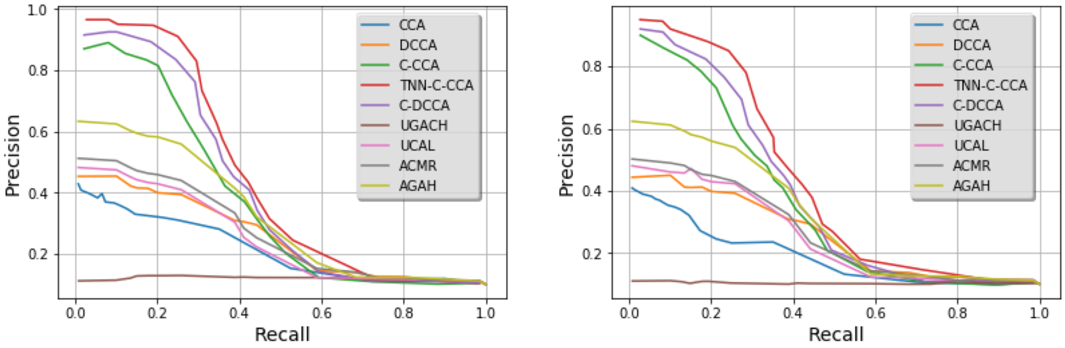


Figure 5.3: The PRC achieved on the VEGAS dataset with our TNN-C-CCA model and other eight different models. The left figure is for audio-to-visual retrieval, the right figure is for visual-to-audio retrieval.

correlation with enlarging the number of pairwise by 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 aligned embeddings for each modality to compute the similarity. Table 5.3 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 with Cluster-CCA without considering negative information, visual-to-audio retrieval can get the MAP of about 7% improved. 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 [99, 100, 114], 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 $\sum_{i=1}^{10} \frac{N_i^2(N-N_i)}{B^2}$. Built triplets based on 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 5.2, the performance of our TNN-C-CCA model is much better than that of novel adversarial learning methods.

Table 5.3: The MAP scores of audio-visual cross-modal retrieval for our TNN variant methods

Models	**VEGAS Dataset**		**MV-10K Dataset**	
	Audio → $Visual$	Visual → $Audio$	Audio → $Visual$	Visual → $Audio$
C-CCA [4]	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

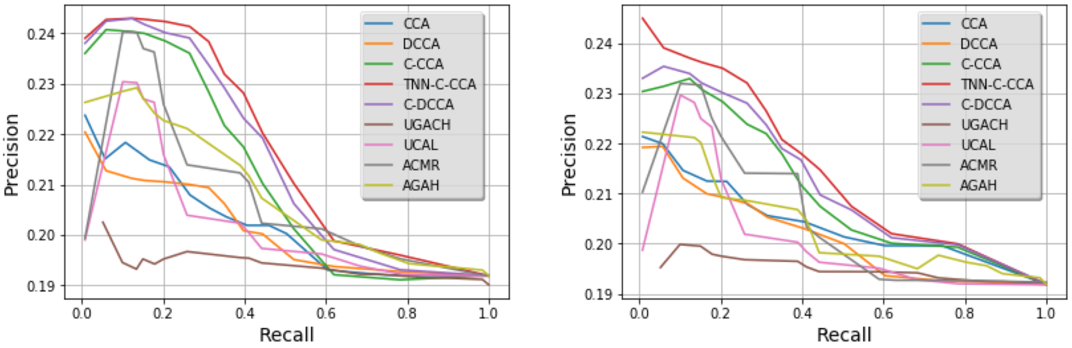


Figure 5.4: The PRC achieved on the MV-10K dataset with nine different models. The left figure is for audio-to-visual retrieval, the right figure is for visual-to-audio retrieval.

5.3.3 Results on the MV-10K Dataset

We report the result of audio-visual cross-modal retrieval on the MV-10K dataset in Table 5.2 with MAP metric and Fig. 5.4 with the PRC. We compare our model with some previous models published in [5]. 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 5.3, 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 5.3. 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 5.2, 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 5.2, Fig. 5.3 and Fig. 5.4, 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.

5.3.4 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.

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. 5.5. During the training, the triplet selection for training TNN-C-CCA model is a very important part. We

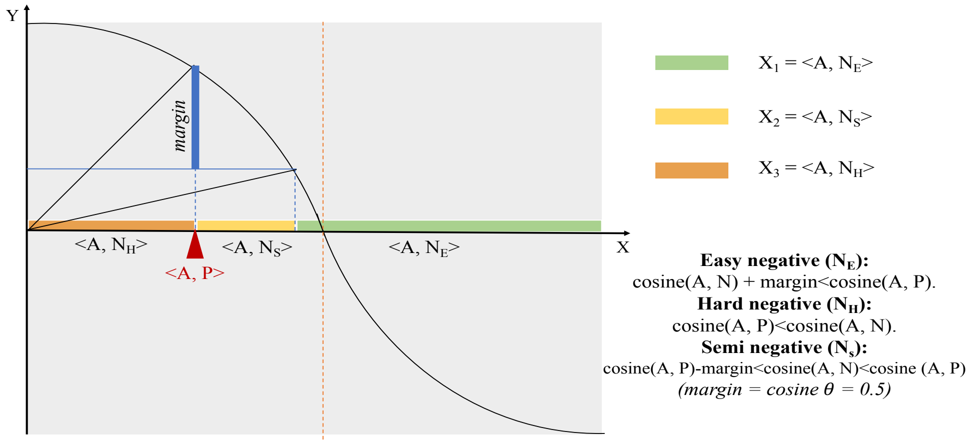

Y
margin
X
<A, N_H>
<A, N_S>
<A, N_E>
<A, P>
🟩 X_1 = <A, N_E>
🟨 X_2 = <A, N_S>
🟧 X_3 = <A, N_H>
Easy negative (N_E):
cosine(A, N) + margin<cosine(A, P).
Hard negative (N_H):
cosine(A, P)<cosine(A, N).
Semi negative (N_s):
cosine(A, P)-margin<cosine(A, N)<cosine (A, P)
(margin = cosine θ = 0.5)

Figure 5.5: Given an Anchor-Positive pair with its 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.

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 5.3 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 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 5.4: MAP with respect to Euclidean distance and Cosine distance in TNN-C-CCA model

Distances	audio-visual	visual-audio
Euclidean distance	0.5300	0.4206
Cosine distance	0.7466	0.7377

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:

$$||X, Y||_{euclidean-distance} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{5.14}$$

where $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$ 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 5.4 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.

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-SNE [115] to implement dimension reduction on the original

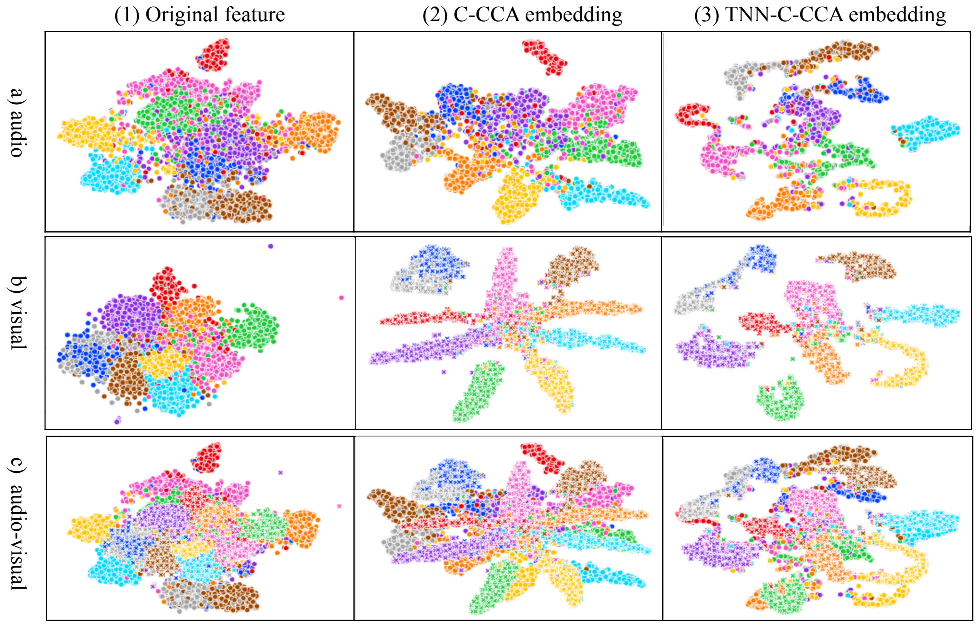


Figure 5.6: The visualization of the two learned subspace with the t-SNE plot, shows audio, visual and audio-visual in the original feature, C-CCA learning subspace, and TNN-C-CCA learning subspace. The circle sign represents audio, the cross sign represents visual.

audio-visual dataset and these features respectively generated from Cluster-CCA and TNN-C-CCA model, where Fig. 5.6 shows audio, visual and audio-visual of their raw features, C-CCA features and TNN-C-CCA features. We can see that in Fig. 5.6, 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, those 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 method with the other three best approaches. Fig. 5.7 provides audio-to-visual retrieval examples

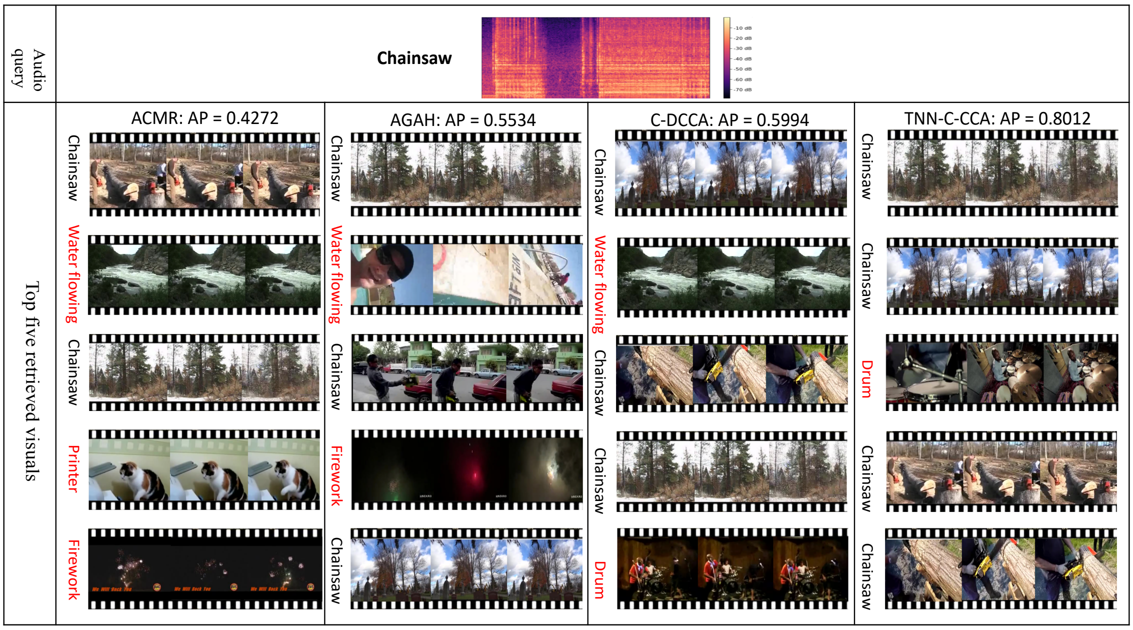


Figure 5.7: The visualized audio-visual retrieval results of our proposed method and other three best approaches, the Cluster-CCA, the AGAH, and the ACMR model. Given an audio as query, the figure shows the top five retrieved visuals.

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.

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 leverage the effect of batch size and margin.

Table 5.5: 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 [116] 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 5.5. 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. 5.8 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 5.6 shows the 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,

Table 5.6: 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

time-consuming take more, 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, 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 lower than that of the C-DCCA model. In the left of Fig. 5.8, 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.

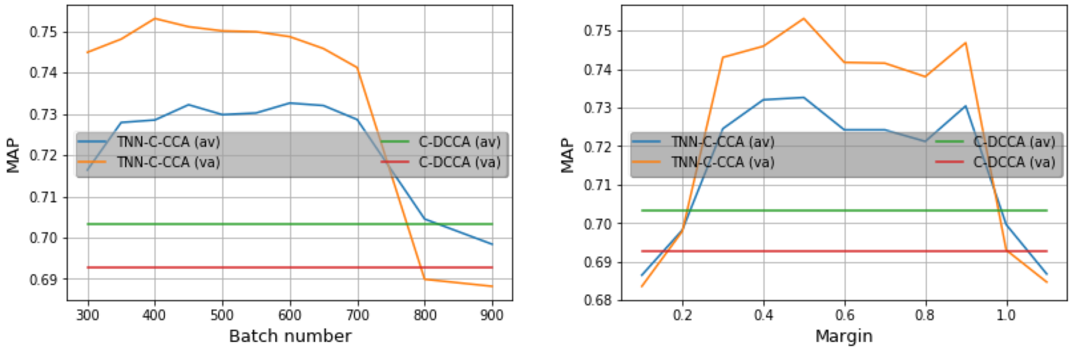


Figure 5.8: The left figure is the MAP curve of TNN-C-CCA and C-DCCA on batch number range from 300 to 700 and the margin are 0.5. The right figure is the MAP curve of TNN-C-CCA and C-DCCA on margin range from 0.3 to 1.0 and the batch number is 500.

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.5.9 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 obvious 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.

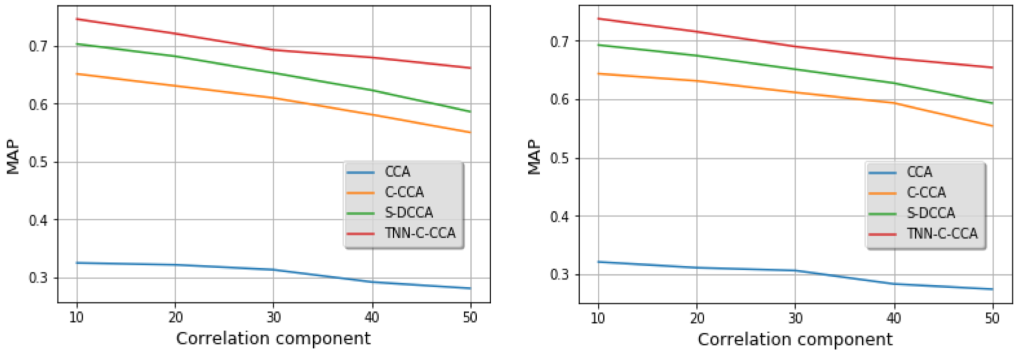


Figure 5.9: The MAP curve of the correlation component changes from 10 to 50, the corner point in the curve represents the correlation component of X-axis and MAP of Y-axis, which use line to connect two adjacent points.The left part is audio-to-visual retrieval and the right part is visual-to-audio retrieval.

5.3.5 Summary

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. We try to apply GAN models to transfer the given cross-modal aligned learning to other one modalities in the next section.

6 Unsupervised Generative Adversarial Multimodal Alignment Learning

6.1 Background and Motivation

With the rapid growth of music contents including users' annotations emerging on the Internet, it is becoming very important to learn common semantics of music alignment representation for facilitating cross-modal music information retrieval. For example, when we input "kids" as query to search a song's audio, video or lyrics, what we expected is an audio that exists kids' voice, video contains kids or lyrics has semantic kids' information. Such semantic concept in audio, video and lyrics are based on explicit concept "kids", which is defined by users. In this paper, sheet music, audio and lyrics are implicitly aligned by high-level semantic concepts, so we develop a content-based representation learning approach by learning alignment across these modalities for retrieval task. The approach ensures the search engine to find the exactly paired music data, without involving the problems of deviation of users' preference.

The main challenge of representation learning across different musical modalities is the heterogeneous gap. In previous works, representation learning for musical cross-modal retrieval focus more on two modalities to bridge the modality gap, such as audio-sheet music [79], which achieved success in musical cross-modal retrieval domain. A classical method series is the CCA-based approaches, which aims at finding transformation to optimize the correlation between the input pairs from two different variable sets. In order to be beneficial from CCA and rank loss for two modalities aligned representation learning, CCA layer [117], combines the existing representations learning like pairwise loss, with the optimal projections of CCA to learn representation between the short snippets of music and the corresponding part of sheet music for the content-based sheet music-audio retrieval scenarios.

With deep learning achieved excellent in modalities aligned representation learning, deep neural networks (DNNs) is introduced in learning aligned representation for cross-modal retrieval task, which provides extensible nonlinear transformations for effective data item representations. Especially, the prevailing architectures combine the DNNs and the CCA that widely apply in the cross-modal aligned representation learning domain, such as deep CCA (DCCA) [118, 103], which demonstrates the possible of learning the aligned representation to retrieval the sheet music image with a short music audio clips as query and vice versa.

Learning aligned representations between two modalities has progressively been arranged in cross-modal retrieval [28, 14, 5], such as learning temporal relation [75] between audio and lyrics for various applications, deep sequential correlation [14] between audio and lyrics for cross-modal retrieval. However, it is hard to satisfy the requirement of real multimodal information retrieval when retrieving one modality by the other two modalities. The goal of this paper is to learn a robust alignment representation for sheet music, audio and lyrics by unsupervised learning, and explore the representations for three groups of cross-modal retrieval tasks to evaluate the performance of our architecture.

Little research has been conducted on the content-based alignment representation learning among musical modalities: audio, sheet music and lyrics, due to the limited available musical dataset. In this paper, we collected a musical dataset with three modalities, including musical audio, sheet music and lyrics. In the dataset, audio and sheet music are paired because they are generated by music notes. Our new architecture for musical alignment representation learning for multimodal information retrieval have achieved two main contributions. Firstly, our architecture can transfer the audio-sheet music pair to audio-lyrics and sheet music-lyrics pair by generative adversarial networks (GANs), and some results achieved by our approach on the MTM dataset prove the feasibility of learning aligned among three modalities by transferring one close relationship to the other two couple of relationship. Secondly, we combine the objective of existing CCA projection with the optimal representations of GANs. In detail, we establish a new ground truth based on the cca embedding and explore generative model (G model) to generate new audio-sheet music pair, the discriminative model of GANs try to distinguish the input is from G model or ground truth, during the adversarial learning, the G model can generated more discriminative and aligned representation for lyrics, audio and sheet music.

6.2 Architecture

6.2.1 Problem Formulation

The goal of our research is to develop a model that has the capability to accept either an audio, a sheet music or a lyric as input to retrieve other two modalities.

Let $U = \{x_1, x_2, ..., x_n\} \in R^n$ be audio set, $V = \{y_1, y_2, ..., y_n\}$ be sheet music set and $W = \{z_1, z_2, ..., z_n\}$ be lyrics set. An example of ternary is $T = \{x_1, y_i, z_i\}$. In particular, the audio features, the sheet music features and lyrics features are represent with different distributions, which cannot be directly compared with each other. Their respectively mappings $f(x_i)$, $g(y_i)$ and $h(z_i)$, that transform audio, sheet music and lyrics features into d-dimensional vectors s_U, s_V and s_W have the same dimension.

Our deep alignment representation learning (DARLearning) approach proposed in this section aim at learning effective three couple of representation s_U, s_V and s_W for audio, sheet music and lyrics. The requirement of distribution of s_U, s_V and s_W to be modality-invariant and there is alignment behind the representations. We explain how to achieve the goal of the requirement in the following subsection.

6.2.2 Generative Model

We take advantage of the close relationship of audio and sheet music. The generative model is to generate a new audio and sheet music pair with a fixed lyrics to challenge the discriminative model. The generated new pair combines with the abstract concept of audio and sheet music for lyrics.

Alignment by Model Transfer

Aspired by teacher-student model [119, 120], we assume audio and sheet music are Student models, lyrics the Teacher model. Our model tries to establish a new aligned representation for all of them. Let x_i be a data point from audio set x, the corresponding data point y_i and z_i are from sheet music set y and lyrics set z. The new generated aligned representations $f(x_i)$ and $g(y_i)$ of audio and sheet music from our model are trained with lyrics z_i. Because the three modalities are synchronized, we can learn $h(z_i)$ model for lyrics z_i to predict the feature of $f(x_i)$ and $g(y_i)$. In our work, we separately use the entropy and following KL-divergence[119, 120, 65] as a loss:

$$\begin{aligned}\sum_{i}^{n} D_{KL}(h(z_i)||f(x_i)) &= \sum_{j}^{n} h(z_i)log\frac{h(z_i)}{f(x_i)} \\ \sum_{i}^{n} D_{KL}(h(z_i)||g(y_i)) &= \sum_{i}^{n} h(z_i)log\frac{h(z_i)}{g(y_i)}\end{aligned} \tag{6.1}$$

The model transfer enhances the audio and sheet music to learn the discriminative representation as lyrics. To reinforce three components of a ternary have similar representations, we enable an alignment across different modalities by generative probability.

Alignment by Generative Probability

Given a lyrics, generative model G aims at fitting the contribution over the lyrics-audio and lyrics-sheet music pairs in a shared common space by mapping function $F(x)$, $G(y)$ and $H(z)$ for audio x, sheet music y and lyrics z. Then, the pairs of informative audio and sheet music are selected to test the ability of discriminative model D. The generative probability of G is $p_\theta(x^U, y^V|z)$, which is the foundation of selecting relevant audio-sheet music pair from unpaired data with lyrics. For instance, given a lyrics query z_i, the generative model tries to select relevant audio x_j from X_{db} and sheet music y_k from Y_{db}. The generative probability $p_\theta(J(x^U, y^V)|z)$ is defined as follows.

$$\begin{gathered} p_\theta(J(x^U, y^V)|z) == \frac{exp(-||H(z) - J(x^U, y^V)||^2)}{\sum_J exp(-||H(z) - J(x^U, y^V)||^2)} \\ J(x^U, y^V) = 0.5 * (F(x^U) + G(y^V)) \end{gathered} \tag{6.2}$$

Where the final $p_\theta(J(x^U, y^V)|z)$ decides the possibility of an audio-sheet music to be a relevant sample.

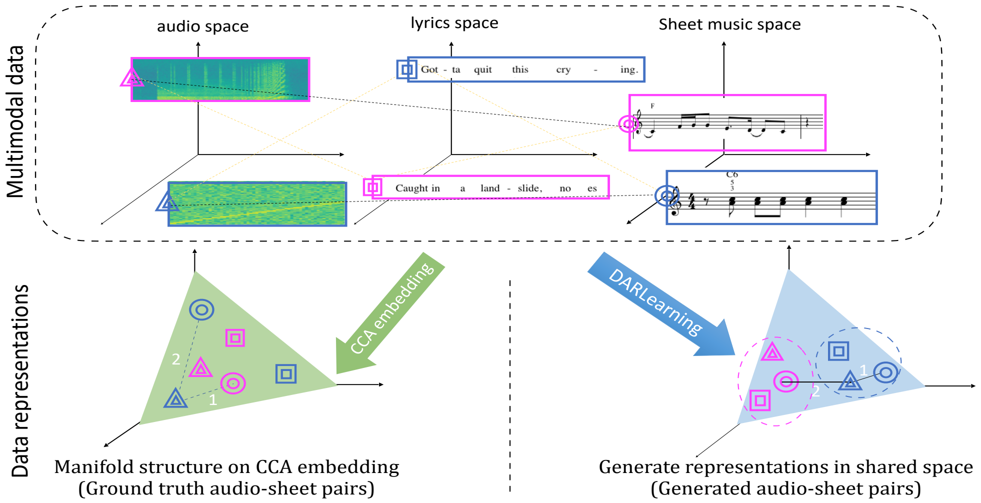

audio space
lyrics space
Sheet music space
Multimodal data
Got - ta quit this cry - ing.
Caught in a land - slide, no es
Data representations
CCA embedding
DARLearning
1
2
1
2
Manifold structure on CCA embedding
(Ground truth audio-sheet pairs)
Generate representations in shared space
(Generated audio-sheet pairs)

Figure 6.1: Schematic of the proposed deep alignment representation learning. The left manifold structure establishment is based on the CCA-based embedding, the right shared space learned by DARLearning model.

6.2.3 Discriminative Model

We apply KNN method to exploit underlying manifold structure for the CCA embeddings of audio-sheet music pairs, we select the top five most close items to establish new pairs as ground truth. The input of the discriminative model is the generated audio and sheet music pair, and the manifold structure based ground truth, seen in Fig. 6.1.

The target of discriminative model D is to discern the input audio-sheet music pair is from ground truth or generated. Once the discriminative model receives the two kinds of input pairs, the model will receive a relevance score for each pair (query and instance i) as the judgment score. The relevance score of $\varphi(p, q)$ is calculated by the following formulation:

$$\begin{aligned} \varphi(p^G, q) = max(0, \alpha + ||\Theta(q) - \Theta(p^T)||^2 - ||\Theta(q) - \Theta(p^G)||^2) \\ \Theta(x) = tanh(W_i x + b_i) \end{aligned} \tag{6.3}$$

where q is the audio query and its generated sheet music instances p^G , p^T is the corresponding ground truth instance. α the margin parameter and it set as 1 in this work. W_i is the weight and b_i is the bias parameter.

The discriminative model use the relevant score to compute the predicted probability of a audio-sheet music pair (x, y) by a sigmoid function.

$$D(p|q) = sigmoid(\varphi(p^G, q)) = \frac{exp(\varphi(p, q)}{1 + exp(\varphi(p, q)} \tag{6.4}$$

6.2.4 Adversarial Learning

Once the concepts of the G model and D models accomplished, they both can be trained by applying a minimax game together. Inspired by the GAN [39], this adversarial process can be defined as follow.

$$\begin{aligned} V(G, D) = min_\theta max_\phi \sum_{j=1}^{n} (E_{x \sim p_{true}(x^T|q_j)}[log(D(x^T)|q_j)] \\ + E_{x \sim p_\theta(x^G|q_j)[log(1-D(x^G)|q_j)])} \end{aligned} \tag{6.5}$$

Fig. 6.2 shows the architecture of our developed algorithm. The architecture has three components: 1) applying advanced pre-trained deep model to extract the features for each modality. 2) generating new data points to fool the D model. 3) distinguishing the input feature belong to generated or ground truth. In detail, we apply different pre-trained model to extract high-level semantic features to bridge the modality gap by alignment representation learning. The features of each modality are 2-Dimensional, which describe more detailed in the section 3. In generative model, the input data is the summary of extracted feature, learning aligned representation is to take lyrics as Teacher model to teach the audio and sheet music model to learn discriminative representation, then transfer the lyrics model into audio and sheet music by GANs. In the discriminative model, the goal is to distinguish the input pair is from generated pair or ground truth pair by computing the relevance score for the judgement result of each generated pair from generative model.

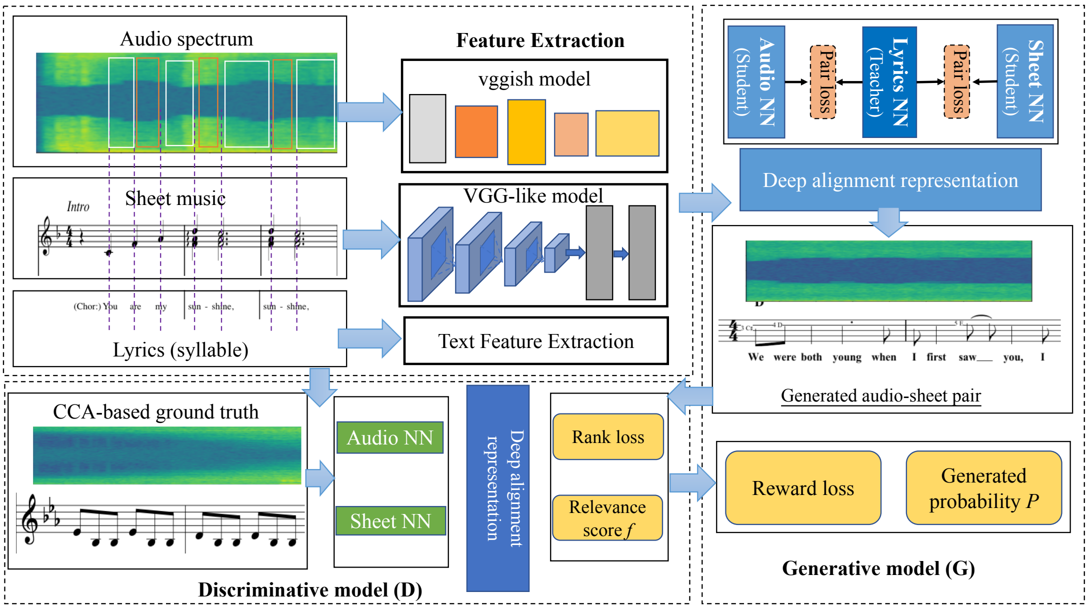


Figure 6.2: The overall framework of our present architecture, which includes three parts: feature extraction, G model and D model.

6.3 Experiments

6.3.1 Implementation Details

This subsection is for the implementation details of our DARLearning model, the model implemented by tensorflow [1]. Audio feature is extracted from the last layer of Vggish model, apply 32 dimensional ASMCMR [?] model extracted feature for sheet music and use the 40 dimensional skip-gram model extracted word-level and syllable-level features for lyrics. The dimensional of feature in the common space is set as 128. Moreover, we train our DARLearning model in a mini-batch with batch size as

[1] https://www.tensorflow.org/

64 for both generative and discriminative model, all the fully connection layers of audio and sheet music in G model and D model share the same structure but learn its own weights and bias. The model is trained iteratively and the G model and D model trained respectively. More detailed setting is shown in the table 6.1.

Table 6.1: The experiment setting of DARLearning model

Music audio feature	128
Sheet music feature	32
Lyrics feature	40
The output dimensional in common space	128
Batch size of generative model	64
Batch size of discriminative model	64
Fully connected layer for audio branch	[1024,1024,128]
Fully connected layers for sheet music branch	[1024,1024,128]
Fully connected layers for lyrics branch	[1024,1024,128]
Global epoch for iterative training	30
Generative epoch	2
Discriminative epoch	2
Initial learning rate	0.001
Decreased factor for each 2 epochs	10

In order to leverage the performance of our proposed model on test sample, we set the same the input and output dimension. In detail, all the above methods apply the same input audio, sheet music and lyrics features and share the same dimension of the output in the common space.

6.3.2 Retrieval Tasks

In our experiments, three couple of cross-modal retrieval are achieved. Specially, retrieving audio by lyrics query ($lyrics \longrightarrow audio$) and retrieving lyrics by audio query ($audio \longrightarrow lyrics$), retrieving audio by sheet music query ($sheetmusic \longrightarrow audio$) and retrieving sheet music by audio query ($audio \longrightarrow sheetmusic$), retrieving lyrics by sheet music query ($sheetmusic \longrightarrow lyrics$) and retrieving sheet music by lyrics query ($lyrics \longrightarrow sheetmusic$). We learn deep alignment representation for audio, lyrics and sheet music in the query and retrieval database for our approach, then take one modality such as using audio as query to compute the cosine similarity with all lyrics in retrieval database. We arrange all lyrics by cosine similarity to obtain rank list and evaluate the rank list by five different standard evaluation criteria, which used in most prior work [101] on unsupervised cross-modal retrieval. We evaluate rank-based performance by *R@K* (K=1,5 and 10), *MedR* and *MeanR*.

6.3.3 Comparison with Existing Methods

We compare our method with two baselines for musical multimodal information retrieval task.

Baseline only discriminative model without generative model and adversarial learning, denoted as *Baseline*. The Baseline model is trained by triplet ranking loss, and the positive item in triplet is only the paired data.

Baseline-GAN is expanded Baseline with adversarial training, denoted as Baseline-GAN. The input of *D* model is the pre-trained model extracted features.

DARLearning model The only difference with Baseline-CCA is the loss of *G* model, we applies the joint generated probability in the *G* model.

The novelty of our architecture is that our proposed DARLearning method projects all the three modalities into a shared subspace to learn aligned representations. We do some initial experiments with the simple structure like applying mean function as joint probability in the *G* model, using the simplest fully connection for each branch.

Compared with traditional cross-modal retrieval methods, our model can use data in one modality to retrieve the samples from another two different modalities in only one learned latent shared subspace. For example, when apply audio as query to retrieve lyrics with traditional cross-modal retrieval methods, we can not use audio to retrieve sheet music because the sheet music is not in this audio-lyrics shared subspace.

Our model projects three modalities into a common space to support the representation can be compared with each other. Some initial results show in the tables. In table I, it verifies the effective of our proposed model for transferring learning, which proves our hypothesis is acceptable. In detail, our model can transfer the relationship of two modalities to another one modality. In table II, the result of three groups of cross-modal retrievals (audio-lyrics, sheet music-lyrics, and audio-sheet music) show the feasibility of further improvement of our proposed model.

In general, our proposed method gets a better performance like R@1, MedR, and

Table 6.2: Cross-modal Retrieval Results on MTM Musical Dataset.

audio-to-lyrics retrieval

Methods	R@1	R@5	R@10	MedR	MeanR
RANDOM [65]	2.77	5.53	7.61	7312.0	7257.19
Our model	5.02	5.51	6.34	715.6	808.02
lyrics-to-audio retrieval					
RANDOM	2.69	5.45	7.59	7316.5	7257.31
Our model	4.14	4.56	5.21	716.0	797.00
sheet music-to-lyrics retrieval					
RANDOM	2.74	5.48	7.53	7311.8	7257.26
Our model	8.36	14.24	16.58	572.5	765.30
lyrics-to-sheet music retrieval					
RANDOM	2.72	5.51	7.66	7313.2	7257.43
Our model	9.95	14.26	17.02	576.0	767.82
audio-to-sheet music retrieval					
RANDOM	2.84	5.57	7.50	7310.0	7257.16
Our model	30.06	33.98	35.02	352.4	600.34
sheet music-to-audio retrieval					
RANDOM	2.63	5.49	7.48	7310.0	7257.37
Our model	33.02	34.12	35.88	330.8	584.42

MeanR than the corresponding performance of GCCA, which suggests the effective of our DARLearning model by applying transfer learning with adversarial learning in *audio* → *lyrics* retrieval, and also for *lyrics* → *audio* retrieval. In other two cross-modal retrievals, when find a shared subspace for all three musical data modalities to obtain aligned representations, the performance of our architectures are higher than Random case and similar to GCCA.

In the table 6.3, it verifies the effective of our proposed generative model for transferring learning, which means our hypothesis is accepted that our DARLearning model can transfer the relationship of one pair to other two relationship within three modalities.

6.3.4 Further Analysis on Our Method

The initial experimental results suggest that our model is viability to learn alignment representations for audio, sheet music and lyrics for cross-modal retrieval task.

Table 6.3: Cross-modal Retrieval Results on MTM Musical Dataset with R@1 and MedR metric.

Method	audio ↓ lyrics		lyrics ↓ audio		audio ↓ sheet music		sheet music ↓ audio		lyrics ↓ sheet music		sheet music ↓ lyrics	
	R@1	MedR	R@1	MedR	R@1	MedR	R@1	MedR	R@1	MedR	R@1	MedR
Baseline	3.69	838.0	2.97	891.0	24.46	562.0	25.13	502.5	3.18	963.0	3.05	982.0
Baseline-GAN	3.92	804.2	3.18	858.0	26.01	529.0	25.74	472.2	3.21	940	3.36	925.8
Our model	5.02	715.6	4.14	716.0	30.06	352.4	33.02	330.8	8.36	572.5	9.95	576.0

Instead of learning representations of two variable sets, our model learns only one shared subspace across three modalities. The learned representations can keep the modality-variant and the paired data should have similar representations.

We expect that our model can surpass CCA model in each couple of cross-modal retrieval in the future. Currently, the shortages of our modal are as follows: i) the loss in G model is not good enough to generate new representation of each modality. Especially, the mean function as the joint probability for the generative probability may weaken the relationship between audio-sheet music by only considering the relevant positions of audio and sheet music. ii) some weights of the fully connection is close to zero. In the future, we would like to use new method to normalize the input features. Overall, it requires us to enhance the relationship between the audio-sheet music during transfer learning with some advanced joint probability, such as considering the local positions of audio and sheet music.

Traditional cross-modal retrieval methods can not learn a deep alignment representation space at once, it constrains on two variable sets. It can not guarantee that our model is lower than this kind of methods, because the learned representation of three modalities with Traditional cross-modal retrieval methods are not in the only one common space. So, what we expected is that our model can surpass traditional cross-modal retrieval methods in each group of cross-modal retrieval task in the future. The aligned representations can keep the modality-variant and the paired data should have similar representations. Currently, the shortages of our modal are as follows: i) the loss in *G* model not good enough to generate new representation of each modality. Especially, the mean function as the joint probability for the generative probability can not ensure that the relationship in audio-sheet music get closer during the learning, which may weaken the relationship between audio-sheet music. ii) same weights of the fully connection is very low, some of them is around 3.1323e-22, this is because the input feature would contains so many value are closed to zero. In the future, I would like to use new method to normalize the input features. Overall, it requires us to enhance the relationship between the audio-sheet music during transfer learning with some advanced joint probability, such as considering the local positions of audio and sheet music.

6.3.5 Summary

Modality-invariant and discriminative representations empower multimodal intelligence to manipulate unrestricted and real world environment. Learning aligned representation is critical for the next generation of multimodal intelligence to learn each cross-modal data on multimodal content. Learning aligned representation between two data modalities has reached outstanding achievement. In this section, we introduce a representation learning model on three data modalities. The experimental results show the feasibility for align representation learning across three different music modalities. Even though there are not audio-lyrics and lyrics-sheet music pairs for training our model, the results demonstrate the alignment can be learned by modalities-level transfer learning.

An open issue for future research is to develop a new generative model which can enhance the relationship of audio-sheet music pairs.

7 Conclusion

In this dissertation, we present three different works to learn aligned representations which are discriminative and modality-invariant for multimodal information retrieval, involving sheet music image, audio, lyrics and visual four modalities. In detail, the proposed S-DCCA architecture for audio-visual cross-modal alignment representation learning capitalizes the temporal structure of the dataset for visual retrieval with audio clips as query, through learning the contributions of the pre-defined semantic labels. Instead of focusing solely on keeping the standard pairwise correspondence between samples, S-DCCA can fully preserve the latent cross-modal semantic structure by considering the gap among the representation of all data points from two modalities with the same class.

In order to further investigating the audio-visual representation learning. Especially, even though S-DCCA considers all the pairs of data points to train within a class, the learned representation of S-DCCA still unavoidably exists the noisy data points that are participated into the wrong class in the learned common space. To address the problem, we propose the TNN-C-CCA model that is an end-to-end supervised learning architecture with audio branch and visual branch. We build triplets based on the pairwise correspondence under partitioning data set into multiple classes as the model input. The experiment results show that TNN-C-CCA model can achieve a better performance than S-DCCA by the pairwise correspondences with the addition of negative samples. Unfortunately, the limitation of GPU memory constrains us to consider all the triplets in the train set. So we established triplets within a mini-batch. With the case of training within a mini-batch, the performance of the model is easily influenced by the size of mini-batch. When the size of mini-batch gets larger, the performance will be better but it will be harder to train and more time-consuming. Moreover, the above two architectures highly rely on the user's annotation and the query and the content of cross-modal retrieval limit in two special modalities.

Our DARLearning model can be viewed as a modality extension of the cross-modal representation learning, which is a new architecture of deep alignment representation for multimodal information retrieval on the musical ternary dataset. To solve the above problem of TNN-C-CCA, the DARLearning model is a no-label involving unsupervised learning which the pairwise correspondence is the standard pairwise and the number of triplets only depends on the number of negative samples. The model can transfer the strong relationship in audio-sheet music pairs to lyrics modality by adversarial learning. However the current model constrains on the transfer learning. In detail, when the model transfers two modalities to another modality, the model only consider the relevant positions of audio and sheet music and ignore the local positions of them. In this case, we need to further investigate our DARLearning model compared with other basic representation learning for multimodal information retrieval in the future, so that when we learn aligned representation among three modalities, we can guarantee each cross-modal representation learning can surpass other state-of-the-art methods.

Invariant representations is essential for multimodal intelligence, which allows the system to work in the real-world environment. Alignment representations for multimodal will be the next generation of MIR when it starts to explore the correlation across modalities. In the future, we look forward to proposing more advanced algorithms for the deep alignment representation for multimodal information retrieval in the audio-visual and musical ternary dataset. For example, we hope to consider more low-level features in our architecture, such as object and action in the visual-audio cross-modal aligned learning. Also, we want to apply some artificial neural networks to exhibit temporal dynamic information in our methods, such as RNN, LSTM, and attention model to capture the sequence information.

Bibliography

[1] David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[2] Pei Ling Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(05):365–377, 2000.

[3] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1247–1255, 2013.

[4] Nikhil Rasiwasia, Dhruv Mahajan, Vijay Mahadevan, and Gaurav Aggarwal. Cluster canonical correlation analysis. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pages pp.823–831, 2014.

[5] Donghuo Zeng, Yi Yu, and Keizo Oyama. 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*, pages pp.143–150, 2018.

[6] Donghuo Zeng, Yi Yu, and Keizo Oyama. Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval. *CoRR*, abs/1908.03737, 2019.

[7] Donghuo Zeng, Yi Yu, and Keizo Oyama. Unsupervised generative adversarial alignment representation for sheet music, audio and lyrics. *arXiv preprint arXiv:2007.14856*, 2020.

[8] Zhangcheng Wang, Ya Li, Richang Hong, and Xinmei Tian. Eigenvector-based distance metric learning for image classification and retrieval. *TOMM*, 15(3):84:1–84:19, 2019.

[9] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. 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*, pages pp.1889–1897, 2014.

[10] Jian Wang, Yonghao He, Cuicui Kang, Shiming Xiang, and Chunhong Pan. 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*, pages pp.347–354, 2015.

[11] Fei Yan and Krystian Mikolajczyk. 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*, pages pp.3441–3450, 2015.

[12] Liwei Wang, Yin Li, and Svetlana Lazebnik. 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*, pages pp.5005–5013, 2016.

[13] Josef Sivic and Andrew Zisserman. 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*, pages pp.1470–1477, 2003.

[14] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen. Deep cross-modal correlation learning for audio and lyrics in music retrieval. *TOMCCAP.*, Vol.15(no.1):pp.20:1–20:16, 2019.

[15] Yan Yan, Feiping Nie, Wen Li, Chenqiang Gao, Yi Yang, and Dong Xu. Image classification by cross-media active learning with privileged information. *IEEE Trans. Multimedia*, Vol.18(12):pp.2494–2502, 2016.

[16] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, Vol.124(3):pp.409–421, Sep 2017.

[17] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. Multimedia content processing through cross-modal association. In *Proceedings of the Eleventh ACM International Conference on Multimedia, Berkeley, CA, USA, November 2-8, 2003*, pages pp.604–611, 2003.

[18] Xiaoxiao Shi and Philip S. Yu. Dimensionality reduction on heterogeneous feature space. In *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*, pages pp.635–644, 2012.

[19] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *Neural Comput. Appl.*, Vol.23:pp.2031–2038, 2013.

[20] Abhishek Sharma, Abhishek Kumar, Hal Daumé III, and David W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages pp.2160–2167, 2012.

[21] Yi Yang, Feiping Nie, Dong Xu, Jiebo Luo, Yueting Zhuang, and Yunhe Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.34(4):pp.723–742, 2012.

[22] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages pp.251–260, 2010.

[23] Pei Ling Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *Int. J. Neural Syst.*, Vol.10(no.5):pp.365–377, 2000.

[24] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

[25] Yoshihiro Yamanishi, J-P Vert, Akihiro Nakaya, and Minoru Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19(suppl_1):i323–i330, 2003.

[26] Matthew B Blaschko, Christoph H Lampert, and Arthur Gretton. Semi-supervised laplacian regularization of kernel canonical correlation analysis. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 133–145. Springer, 2008.

[27] Weiran Wang, Raman Arora, Karen Livescu, and Nathan Srebro. Stochastic optimization for deep cca via nonlinear orthogonal iterations. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 688–695. IEEE, 2015.

[28] Yi Yu, Suhua Tang, Kiyoharu Aizawa, and Akiko Aizawa. Category-based deep cca for fine-grained venue discovery from multimodal data. *IEEE transactions on neural networks and learning systems.*, Vol.30(no.99):pp.1–9, 2018.

[29] Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257, 2011.

[30] David R Hardoon and John Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

[31] Viresh Ranjan, Nikhil Rasiwasia, and CV Jawahar. Multi-label cross-modal retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4094–4102, 2015.

[32] Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora. Deep generalized canonical correlation analysis. *arXiv preprint arXiv:1702.02519*, 2017.

[33] Michael W Roth. Survey of neural network technology for automatic target recognition. *IEEE Transactions on neural networks*, 1(1):28–43, 1990.

[34] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[35] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5528–5531. IEEE, 2011.

[36] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[37] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[39] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[40] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[41] Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007.

[42] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[43] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[44] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[45] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3908–3916, 2015.

[46] Fangxiang Feng, Xiaojie Wang, Ruifan Li, and Ibrar Ahmad. Correspondence autoencoders for cross-modal retrieval. *TOMCCAP*, Vol.12(no.1s):pp.26:1–26:22, 2015.

[47] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages pp.689–696, 2011.

[48] Yuxin Peng, Xin Huang, and Jinwei Qi. 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*, pages pp.3846–3853, 2016.

[49] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1445–1454, 2016.

[50] Zheng Yu and Wenmin Wang. Learning dalts for cross-modal retrieval. *CAAI Transactions on Intelligence Technology*, 4(1):9–16, 2019.

[51] Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, and Qi Tian. Multi-networks joint learning for large-scale cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 907–915, 2017.

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

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

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages pp.770–778, 2016.

[55] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1881–1889, 2017.

[56] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages pp.154–162, 2017.

[57] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Sch-gan: Semi-supervised cross-modal hashing by generative adversarial network. *IEEE transactions on cybernetics*, 50(2):489–502, 2018.

[58] Daniel L Swets and John Juyang Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on pattern analysis and machine intelligence*, 18(8):831–836, 1996.

[59] Yong Rui, Thomas S Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of visual communication and image representation*, 10(1):39–62, 1999.

[60] Mei Kobayashi and Koichi Takeda. Information retrieval on the web. *ACM Computing Surveys (CSUR)*, 32(2):144–173, 2000.

[61] Ellen M Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 171–180, 1993.

[62] Jacobus Cornelis Haartsen, Sten Minör, Bengt Stavenow, William O Camp Jr, Ronald A Louks, and Björn Martin Gunnar Lindquist. Adaptive display for enhancing audio playback, March 18 2014. US Patent 8,676,869.

[63] Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder. Features for content-based audio retrieval. In *Advances in computers*, volume 78, pages 71–150. Elsevier, 2010.

[64] Mohammad Ubaidullah Bokhari and Faraz Hasan. Multimodal information retrieval: Challenges and future trends. *International Journal of Computer Applications*, 74(14), 2013.

[65] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *CoRR*, abs/1706.00932, 2017.

[66] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li. Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. *IEEE Transactions on Cybernetics*, Vol.49(7):pp.1–14, 2019.

[67] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. Multimedia content processing through cross-modal association. In Lawrence A. Rowe, Harrick M. Vin, Thomas Plagemann, Prashant J. Shenoy, and John R. Smith, editors, *Proceedings of the Eleventh ACM International Conference on Multimedia, Berkeley, CA, USA, November 2-8, 2003*, pages 604–611. ACM, 2003.

[68] Hong Zhang, Yueting Zhuang, and Fei Wu. Cross-modal correlation learning for clustering on image-audio dataset. In Rainer Lienhart, Anand R. Prasad, Alan Hanjalic, Sunghyun Choi, Brian P. Bailey, and Nicu Sebe, editors, *Proceedings of the 15th International Conference on Multimedia 2007, Augsburg, Germany, September 24-29, 2007*, pages 273–276. ACM, 2007.

[69] Yale Song, Louis-Philippe Morency, and Randall Davis. Multimodal human behavior analysis: learning correlation and interaction across modalities. In Louis-Philippe Morency, Dan Bohus, Hamid K. Aghajan, Justine Cassell, Anton Nijholt, and Julien Epps, editors, *International Conference on Multimodal Interaction, ICMI '12, Santa Monica, CA, USA, October 22-26, 2012*, pages 27–30. ACM, 2012.

[70] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 689–696. Omnipress, 2011.

[71] Christoph H. Lampert and Oliver Krömer. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II*, volume 6312 of *Lecture Notes in Computer Science*, pages 566–579. Springer, 2010.

[72] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016.

[73] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. *Int. J. Comput. Vis.*, 126(10):1120–1137, 2018.

[74] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Frédo Durand, and William T Freeman. The visual microphone: Passive recovery of sound from video. 2014.

[75] Hiromasa Fujihara and Masataka Goto. Lyrics-to-audio alignment and its application. In Meinard Müller, Masataka Goto, and Markus Schedl, editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 23–36. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, 2012.

[76] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen. Deep cross-modal correlation learning for audio and lyrics in music retrieval. *CoRR*, abs/1711.08976, 2017.

[77] Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. Integrating additional chord information into hmm-based lyrics-to-audio alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):200–210, 2011.

[78] Sang Won Lee and Jeffrey Scott. Word level lyrics-audio synchronization using separated vocals. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 646–650. IEEE, 2017.

[79] Verena Thomas, Christian Fremerey, Meinard Müller, and Michael Clausen. Linking sheet music and audio - challenges and new approaches. In Meinard Müller, Masataka Goto, and Markus Schedl, editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 1–22. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, 2012.

[80] Matthias Dorfer, Andreas Arzt, and Gerhard Widmer. Towards score following in sheet music images. In Michael I. Mandel, Johanna Devaney, Douglas Turnbull, and George Tzanetakis, editors, *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, pages 789–795, 2016.

[81] Stefan Balke, Matthias Dorfer, Luis Carvalho, Andreas Arzt, and Gerhard Widmer. Learning soft-attention models for tempo-invariant audio-sheet music retrieval. In Arthur Flexer, Geoffroy Peeters, Julián Urbano, and Anja Volk, editors, *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, pages 216–222, 2019.

[82] Matthias Dorfer, Jan Hajič Jr, Andreas Arzt, Harald Frostel, and Gerhard Widmer. Learning audio–sheet music correspondences for cross-modal retrieval and piece identification. *Transactions of the International Society for Music Information Retrieval*, 1(1), 2018.

[83] Yi Yu and Simon Canales. Conditional LSTM-GAN for melody generation from lyrics. *CoRR*, abs/1908.05551, 2019.

[84] Satoru Fukayama, Kei Nakatsuma, Shinji Sako, Takuya Nishimoto, and Shigeki Sagayama. Automatic song composition from the lyrics exploiting prosody of the japanese language. In *Proc. 7th Sound and Music Computing Conference (SMC)*, pages 299–302, 2010.

[85] Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. A melody-conditioned lyrics language model. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 163–172. Association for Computational Linguistics, 2018.

[86] Chung-Che Wang, Jyh-Shing Roger Jang, and Wennen Wang. An improved query by singing/humming system using melody and lyrics information. In J. Stephen Downie and Remco C. Veltkamp, editors, *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, pages 45–50. International Society for Music Information Retrieval, 2010.

[87] Jen-Yu Liu, Yu-Hua Chen, Yin-Cheng Yeh, and Yi-Hsuan Yang. Score and lyrics-free singing voice generation. *arXiv preprint arXiv:1912.11747*, 2019.

[88] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[89] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. 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*, pages pp.3550–3558, 2018.

[90] Yi Yu and Simon Canales. Conditional lstm-gan for melody generation from lyrics. *arXiv preprint arXiv:1908.05551*, 2019.

[91] S Watanabe, T Hori, S Karita, T Hayashi, J Nishitoba, Y Unno, NEY Soplin, J Heymann, M Wiesner, N Chen, et al. Espnet: End-to-end speech processing toolkit. arxiv 2018. *arXiv preprint arXiv:1804.00015.*

[92] Craig Lambert, Judit Kormos, and Danny Minn. Task repetition and second language speech processing. *Studies in Second Language Acquisition*, 39(1):167–196, 2017.

[93] Takuya Kobayashi, Akira Kubota, and Yusuke Suzuki. Audio feature extraction based on sub-band signal correlations for music genre classification. In *2018 IEEE International Symposium on Multimedia (ISM)*, pages 180–181. IEEE, 2018.

[94] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 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.*, pages pp.4278–4284, 2017.

[95] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 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*, pages pp.2818–2826, 2016.

[96] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *arXiv:1609.08675*, 2016.

[97] Sergey Ioffe and Christian Szegedy. 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*, pages pp.448–456, 2015.

[98] Matthias Dorfer, Andreas Arzt, and Gerhard Widmer. Learning audio-sheet music correspondences for score identification and offline alignment. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, pages 115–122, 2017.

[99] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 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*, pages pp.815–823, 2015.

[100] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[101] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I*, pages 651–667, 2016.

[102] Yu-Siang Huang, Szu-Yu Chou, and Yi-Hsuan Yang. Music thumbnailing via neural attention modeling of music emotion. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*, pages 347–350. IEEE, 2017.

[103] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages pp.1247–1255, 2013.

[104] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[105] Bruce Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 2005.

[106] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge university press, 2000.

[107] Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann. 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*, pages pp.607–616, 2014.

[108] Viresh Ranjan, Nikhil Rasiwasia, and C. V. Jawahar. Multi-label cross-modal retrieval. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages pp.4094–4102, 2015.

[109] David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation.*, Vol.16(no.12):pp.2639–2664, 2004.

[110] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages pp.2088–2095, 2013.

[111] Jian Zhang, Yuxin Peng, and Mingkuan Yuan. 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*, pages pp.539–546, 2018.

[112] Wen Gu, Xiaoyan Gu, Jingzi Gu, Bo Li, Zhi Xiong, and Weiping Wang. 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.*, pages 159–167, 2019.

[113] Li He, Xing Xu, Huimin Lu, Yang Yang, Fumin Shen, and Heng Tao Shen. 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*, pages pp.1153–1158, 2017.

[114] R. Manmatha, Chao-Yuan Wu, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages pp.2859–2867, 2017.

[115] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[116] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning.*, Vol.20(no.3):pp.273–297, 1995.

[117] Matthias Dorfer, Jan Schlüter, Andreu Vall, Filip Korzeniowski, and Gerhard Widmer. End-to-end cross-modality retrieval with cca projections and pairwise ranking loss. *International Journal of Multimedia Information Retrieval*, 7(2):117–128, 2018.

[118] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pages 1764–1772, 2014.

[119] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 892–900, 2016.

[120] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2827–2836. IEEE Computer Society, 2016.