様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ５ 月 ２０ 日現在

機関番号：６２６１５
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１１９８７

研究課題名（和文）Zero-shot Cross-modal Embedding Learning

研究課題名（英文）Zero-shot Cross-modal Embedding Learning

研究代表者
ュ　イ（Yu, Yi）

国立情報学研究所・コンテンツ科学研究系・特任助教

研究者番号：００７５４６８１

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：視・聴覚クロスモーダル学習のために、i）オーディオデータとビジュアルデータを2つの異なるスペースに別々にエンコードし、さらに正準相関分析によって特徴量を共通のサブスペースにマッピングする、ii）確率的モデリング手法を利用して、データにおけるノイズや欠落を処理する、新しい変分オートエンコーダー（VAE）アーキテクチャを提案・評価した。

研究成果の学術的意義や社会的意義
The distribution of data in different modalities are inconsistent, which makes it difficult to directly measure the similarity across different modalities. The proposed technique of cross-modal embedding learning can help improve the performance of cross-modal retrieval, recognition, and generation.

研究成果の概要（英文）：This project focused on cross-modal embedding learning for cross-modal retrieval. The main challenge is how to learn joint embeddings in a shared subspace for computing the similarity across different modalities. 1) We proposed a novel deep triplet neural network with cluster canonical correlation analysis (TNN-C-CCA), which is an end-to-end supervised learning architecture with audio branch and video branch. 2) We proposed a novel variational autoencoder (VAE) architecture for audio-visual cross-modal retrieval, by learning paired audio-visual correlation embedding and category correlation embedding as constraints to reinforce the mutuality of audio-visual information. 3) We proposed an unsupervised generative adversarial alignment representation (UGAAR) model to learn deep discriminative representations shared across three major musical modalities: sheet music, lyrics, and audio, where a deep neural network based architecture on three branches is jointly trained.

研究分野：データベース関連

キーワード：Cross-Modal Correlation　Cross-Modal Embedding

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

Brain science tells us that human sense and cognize the world through the fusion of multiple sensory organs. Accordingly, a same object is represented in different modalities (vision, sound, and language, etc.), and the complementary information helps users better understand the real world. Evolution of mobile technologies has enabled people to share their multimedia data (e.g., image, video, audio, blog) in different modalities, to show what they are interested in. Consequently, these multimodal data and information are accumulated on the Internet at an unprecedented scale, which motivates us to use artificial intelligence to model the cognitive process of human from these big data.

Cross-modal retrieval is to retrieve data in one modality by a query in another modality, which has been a popular topic in information retrieval, machine learning, and database. It has two major issues: i) The distribution and representations of data in different modalities are inconsistent, which makes it difficult to directly measure the similarity among data in different modalities. ii) The training data only cover all known categories, which makes it difficult to handle emerging data with new categories. To solve the issue i), existing cross-modal retrieval methods mainly focus on multimodal correlation learning where data in different modalities are projected in a common subspace to calculate the similarity. Unfortunately, the issue ii) is not addressed before.

２．研究の目的

This project studies cross-modal embedding learning which can be applied for cross-modal retrieval over large-scale dataset. Our goal is to enhance the scalability of cross-modal retrieval in intelligent or advanced reasoning systems where new data with unknown categories are continuously emerging. To this end, various deep architectures are developed to learn and predict categories in the model training stage. Moreover, adversarial contrastive learning and attention mechanism techniques are used to enhance modality invariance in the embedding and better correlate data in different modalities with inconsistent semantics and heterogeneous distributions. In addition, all these are designed in a unified framework.

３．研究の方法

In this project, we mainly studied embedding learning between different modality data. This kind of embedding learning can be established as a supervised, semi-supervised or unsupervised task depending on how relations between modalities are used. We focus on supervised learning and unsupervised learning between audio-video and melody-lyrics.

In this research project, various deep embedding models are investigated to capture and synchronize semantic information between different modality data, which facilitates to build the cross-modal embedding space. Several techniques are investigated. Canonical correlation analysis (CCA) is a very popular method for linearly embedding multimodal data in a shared common space, and Deep CCA (DCCA) is an extension of CCA which tries to capture the non-linear correlation. These methods aim to maximize the correlation between different modality data.

Attention: In neural networks, the effect of attention aims to enhance some parts of the input data while enervating other parts. It can be utilized for enhancing semantic cross-modal information to align audio and visual sequences through a dynamic pooling process that allows to learn relations between sub-elements.

Transformer: It is an encoder-decoder model with self-attention mechanism as well as positional encoding, which can be exploited to capture sequential cross-modal semantic information between audio and visual cross-modal data.

Generative Adversarial Networks: The generative model is learned for subspace embedding by capturing real data distribution, and the discriminative model is used to discriminate between real data and generated fake data via a minimax game. When the discriminative model is used for modality recognition on the embedding of real data and generated fake data, adversarial learning helps to generate modality-invariant embedding in the target subspace.

Auto-Encoders: AEs are unsupervised encoder-decoder models that create latent representations through reconstruction. The latent representation reflects the structural distribution of the original data (similar to summarizing/reducing dimensionality).

４．研究成果

In the past three years, this project focused on cross-modal embedding learning applied to audio-visual cross-modal retrieval and cross-modal music retrieval. The main challenge of audio-visual cross-modal retrieval task is how to learn joint embeddings in a shared subspace for computing the similarity across different

modalities, where generating new representations is to maximize the correlation between audio and visual modalities space.

4.1 Main research achievement in audio-visual cross-modal retrieval

In the work published in ACM TOMCCAP journal, we propose a novel deep triplet neural network with cluster canonical correlation analysis (TNN-C-CCA), which is an end-to-end supervised learning architecture with audio branch and video branch (Figure 1). We not only consider the matching pairs in the common space but also compute the mismatching pairs when maximizing the correlation. In particular, two significant contributions are made: i) a better representation by constructing deep triplet neural network with triplet loss for optimal projections can be generated to maximize correlation in the shared subspace. ii) positive examples and negative examples are used in the learning stage to improve the capability of embedding learning between audio and video.

In another work under reviewing, we present a novel variational autoencoder (VAE) architecture for audiovisual cross-modal retrieval, by learning paired audio-visual correlation embedding and category correlation embedding as constraints to reinforce the mutuality of audio-visual information. On the one hand, audio encoder and visual encoder separately encode audio data and visual data into two different latent spaces. Further, two mutual latent spaces are respectively constructed by canonical correlation analysis (CCA). On the other hand, probabilistic modeling methods is used to deal with possible noise and missing information in the data. Additionally, in this way, the cross-modal discrepancy from intra-modal and inter-modal information are simultaneously eliminated in the joint embedding subspace.

We further conclude a survey for recent advances and challenges in deep audio-visual correlation learning, which is available in arXiv.

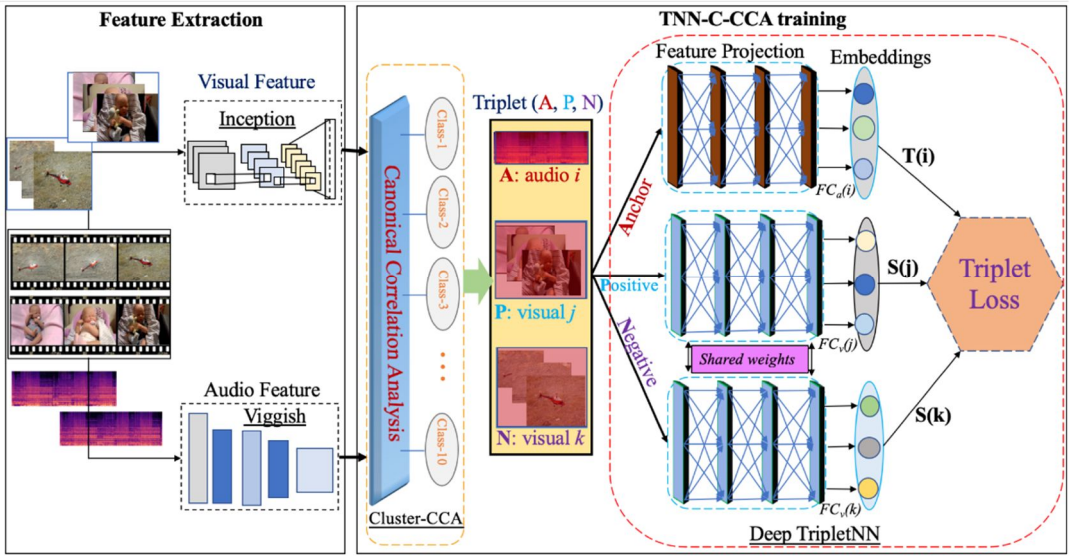

Figure 1 The overall framework of our TNN-C-CCA model. It consists of two parts: feature extraction and TNN-C-CCA training. We apply Inception V3 and Vggish model to extract feature, then explore cluster-CCA to learn the correlation with cluster segregating and select triplets as input for deep TNN training. 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.

4.2 Main research achievement in cross-modal music retrieval

In this work, we propose an unsupervised generative adversarial alignment representation (UGAAR) model to learn deep discriminative representations shared across three major musical modalities: sheet music, lyrics, and audio, where a deep neural network based architecture on three branches is jointly trained. In particular, the proposed model can transfer the strong relationship between audio and sheet music to audio-lyrics and sheet-lyrics pairs by learning the correlation in the latent shared subspace. We apply CCA components of audio and sheet music to establish new ground truth. The generative model learns the correlation of two couples of transferred pairs to generate new audio-sheet pair for a fixed lyrics to challenge the discriminative model. The discriminative model aims at distinguishing the input, to tell whether it is from the generative model or the ground truth. The two models are simultaneously trained in an adversarial way to enhance the ability of deep alignment representation learning.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Donghuo Zeng, Yi Yu, Keizo Oyama	４．巻 16
２．論文標題 Deep Triplet Neural Networks with Cluster-CCA for Audio-Visual Cross-Modal Retrieval	５．発行年 2020年
３．雑誌名 ACM Transaction on Multimedia Computing Communication and Applications	６．最初と最後の頁 1-23
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Donghuo Zeng, Yi Yu, and Keizo Oyama
２．発表標題 Unsupervised generative adversarial alignment representation for sheet music, audio and lyrics
３．学会等名 IEEE International Conference on Multimedia Big Data 2020（国際学会）
４．発表年 2020年

１．発表者名 Donghuo Zeng, Yi Yu, and Keizo Oyama
２．発表標題 MusicTM-Dataset for joint representation learning among sheet music, lyrics, and musical audio
３．学会等名 The 8th Conference on Sound and Music Technology, Lecture Notes in Electrical Engineering
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

https://github.com/yy1lab/Lyrics-Conditioned-Neural-Melody-Generation

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関