氏　　名	Donghuo Zeng
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 2196 号
学位授与の日付	2020 年 9 月 28日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Deep Alignment Representation Learning for Multimodal Information Retrieval
論 文 審 査 委 員	主　査　教授　大山　敬三
	教授　相澤　彰子
	教授　山岸　順一
	助教　ユ　イ
	教授　佐藤　真一 国立情報学研究所　コンテンツ科学研究系

Summary of Doctoral Thesis

Name in full: Donghuo Zeng

Title: Deep Alignment Representation Learning for Multimodal Information Retrieval

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. 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 use a query in one modality to obtain relevant data 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 have inconsistent distributions and representations, which leads to the features is unable to be directly compared with each other to accomplish the retrieval achievements. The objective of this dissertation is to develop new architecture to project the low-level features of different modalities to high-level semantic representation in a common space to bridge the gap. 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 high-speed development of innovative technology and user interaction on the Internet, various multimedia data and information have been 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 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 should have similar representations for MIR.

Among the various ways of learning aligned representation, Canonical Correlation Analysis (CCA) is a classical linear method to learn the correlation of two sets of variables (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 item pairs are optimized, supposed the linear transformsW_1 and W_2 are the matrices and $\sum_{v1v1}$ and $\sum_{v2v2}$ are the covariance matrices of V_1 and V_2 and $\sum_{v1v2}$ is the cross-covariance matrix, which uses to maximize the correlation in the latent subspace as follows.

$$(W_1, W_2) = \arg\max_{(W_1,W_2)} \frac{W_1^T \Sigma_{v_1 v_2} W_2}{\sqrt{W_1^T \Sigma_{v_1 v_1} W_1 \cdot W_2^T \Sigma_{v_2 v_2} W_2}} \tag{1.1}$$

Based on the CCA, some extension methods are proposed. Before projecting the features into a common space, Kernel CCA first map the features into a higher dimensional feature space. In order to be beneficial from the deep learning method, Deep CCA learns complex nonlinear transformations for two different sets of variants. Unlike the standard pair in the CCA training, in Cluster-CCA 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, we propose S-DCCA model to learn aligned representations in a shared latent space by finding nonlinear transforms for audio and visual to optimize the correlation between them. 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.

In the second work, we present TNN-C-CCA method that can be viewed as an improvement of S-DCCA with a deep triplet network to optimize the correlation between audio and visual by establishing triplet as training based on the similar or dissimilar semantic paired. The implication of this work is to learn a better aligned representation for audio-visual cross-modal retrieval. Compared with other state-of-the-art methods, the proposed method can achieve better performances.

In the third work, we introduce 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 experiments indicate the results can beneficial from the representations.

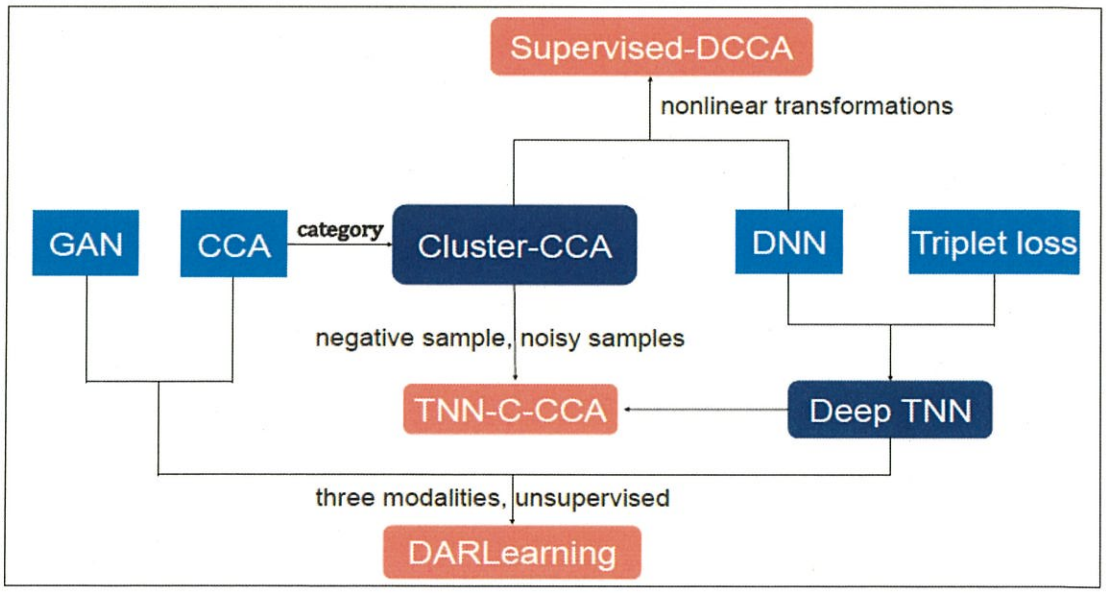

Figure 1.1: Overview of techniques used in this dissertation.

Fig. 1.1 presents the relationship between our three 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 can use negative samples to reduce the noisy samples that the samples in the supervised-DCCA shared subspace are grouped to 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 DARLearning model can learn alignment representation for three different modalities by unsupervised learning.

Form8・Separate Sheet
(様式８・別紙１)

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　Donghuo Zeng

Title
論文題目　　Deep Alignment Representation Learning for Multimodal Information Retrieval

本学位論文は、マルチモーダル情報検索のための深層アライメント表現の学習に関するものであり、全７章から構成され、英語で記述されている。

第１章では序論として、本研究の背景、位置づけや意義を述べている。マルチモーダル情報検索においては、モダリティ間にあるセマンティックギャップを解消するため、モダリティごとの低位のデータ表現から共通空間である高位の意味表現への写像が必要となるが、適合性の高い検索結果を得るためには、その写像において、各モダリティのアイテムの概念的特徴を保存しつつ、モダリティをまたぐアイテム間の概念的類似性を反映した共通空間上の配置にマッピングするための deep alignment representation learning が求められるとし、従来手法と本研究の提案手法について概略を述べるとともに相互関係を示している。

第２章では、クロスモーダル及びマルチモーダル情報検索に関する関連研究を紹介している（本論文ではモダリティが２の場合をクロスモーダル、３以上の場合をマルチモーダルと呼んでいる)。まず、上述の写像について、CCA などの基本的な線形写像から LSTM や GAN などの深層学習による写像までサーベイを行い、次に具体的なモダリティとして映像・音響間及び音響・歌詞・楽譜間のマルチモーダル情報検索手法について研究事例を紹介し、アライメント表現の学習について論じている。

第３章では本研究のために収集・整備したアライメント情報を有するデータセット、低位のデータ表現を取得するための特徴抽出手法、および検索性能の評価指標について述べている。データセットのうち２つは音響と映像のクロスモーダル、１つは音楽コンテンツの楽譜画像、音響、歌詞のマルチモーダルであり、これらの特性が示されている。特徴抽出手法については、音響、映像、楽譜（画像)、歌詞（テキスト）のそれぞれのモダリティについて、既存手法の中から本研究で使用した手法を紹介している。また、情報検索結果の評価指標については、本研究で使用した情報検索の代表的な評価指標について説明している。

第４章から第６章はそれぞれ、出願者が提案した手法と実験結果に関して論じている。

第４章は音楽コンテンツの検索のための、音響と映像のクロスモーダルアライメント表現の学習手法として提案した Supervised Deep Canonical Correlation Analysis (S-DCCA)にして論じている。本手法は時間的構造への対応と感情の類似性に基づくアライメントを特徴としている。本手法を用いて、音響または映像の一方をクエリとし、感情が類似するもう一方のモダリティを出力とする検索システムを実装し、評価実験を行い、検索性能が既存の手法と比較して概ね上回っていることが確認されている。

第５章では音響と映像のクロスモーダルアライメント表現の学習をより高度化するために提案した Deep Triplet Neural Network with Cluster Canonical Correlation Analysis (TNN-C-CCA)に関して論じている。本手法は音響と映像の対応するペアとともに異なるカテゴリの対応しないアイテムを加え、トリプレットとして同時に学習に利用することによって、共通空間での相関を最大化することを特徴としている。本手法を用いた検索実験を行い、検索性能が既存の手法と比較して一貫して上回ることが示されている。

第６章では楽譜・音声・歌詞の３モダリティ間のマルチモーダルアライメント表現の学習手法として提案した Deep Alignment Representation Learning Method (DARLearning)に関して論じている。これまで３モダリティを扱うためには２モダリティごとの学習を組み合わせる手法がとられていたが、本手法では３モダリティを同時に学習し同一の共通空間にマッピングすることを可能としている。また、本手法では教師・生徒モデルを利用した敵対的学習によって、相関の強い音響・楽譜ペア間の関連性を、相関の弱い音響・歌詞ペア間及び楽譜・歌詞ペア間の関連性に転移できることを特徴としている。本手法を用いた検索実験を行い、検索性能が従来の基本的な手法を上回り、一部の評価指標では GAN を用いた最新の手法をも上回ることが示されている。３モダリティを同時に学習する手法として有望な手法であるとしている。

第７章では本研究の主要部分である第４章から第６章で提案した３つの手法について総括するとともに、それぞれの得失や将来の課題について論じている。

以上を要するに本学位論文は、複数のモダリティにまたがるマルチモーダル情報検索に必要とされる、共通空間上でのアライメント表現の効果的学習方法を示したものである。また、本学位論文の成果は、学術雑誌論文１件、フルペーパー査読付き国際会議論文１件として発表され、学術的な貢献も認められる。

公開発表会では博士論文の章立てに従って発表が行われ、その後に行われた論文審査会では、審査員からの質疑に対して適切に回答がなされた。出願者退出後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、出願者の博士研究は独創性が高く検索性能の向上にも有効であることが評価された。

以上の理由により、審査委員会は、本学位論文が学位の授与に値すると判断した。