計算機はオーディオとビジュアルの相関を学習できるか？

Can we learn the correlation between audio and visual?

深視聴覚クロスモーダル検索のための共同埋め込みの学習

Learning Joint Embedding for audio-visual Cross-Modal Retrieval

Donghuo Zeng (SOKENDAI), Yi Yu, Keizo Oyama

背景 Background

深層学習は、異なるデータモダリティ間の統合表現を学習するには、優れた性能を示してきました。しかしながら、オーディオやビジュアルなどのデータモダリティに重要である時間的構造は、クロスモーダル相関学習に関するほとんどの研究には考慮されていなかった。

Deep learning has successfully showed excellent performances in learning joint representations between different data modalities. Unfortunately, little research focuses on cross-modal correlation learning where temporal structures of different data modalities such as audio and visual should be taken into account.

目標 Target

ビジュアルのクロスモーダル検索を目指し、我々は、オーディオ信号とビジュアル信号の時間的構造の特性を利用して、その深層シーケンス相関モデルを学習します。

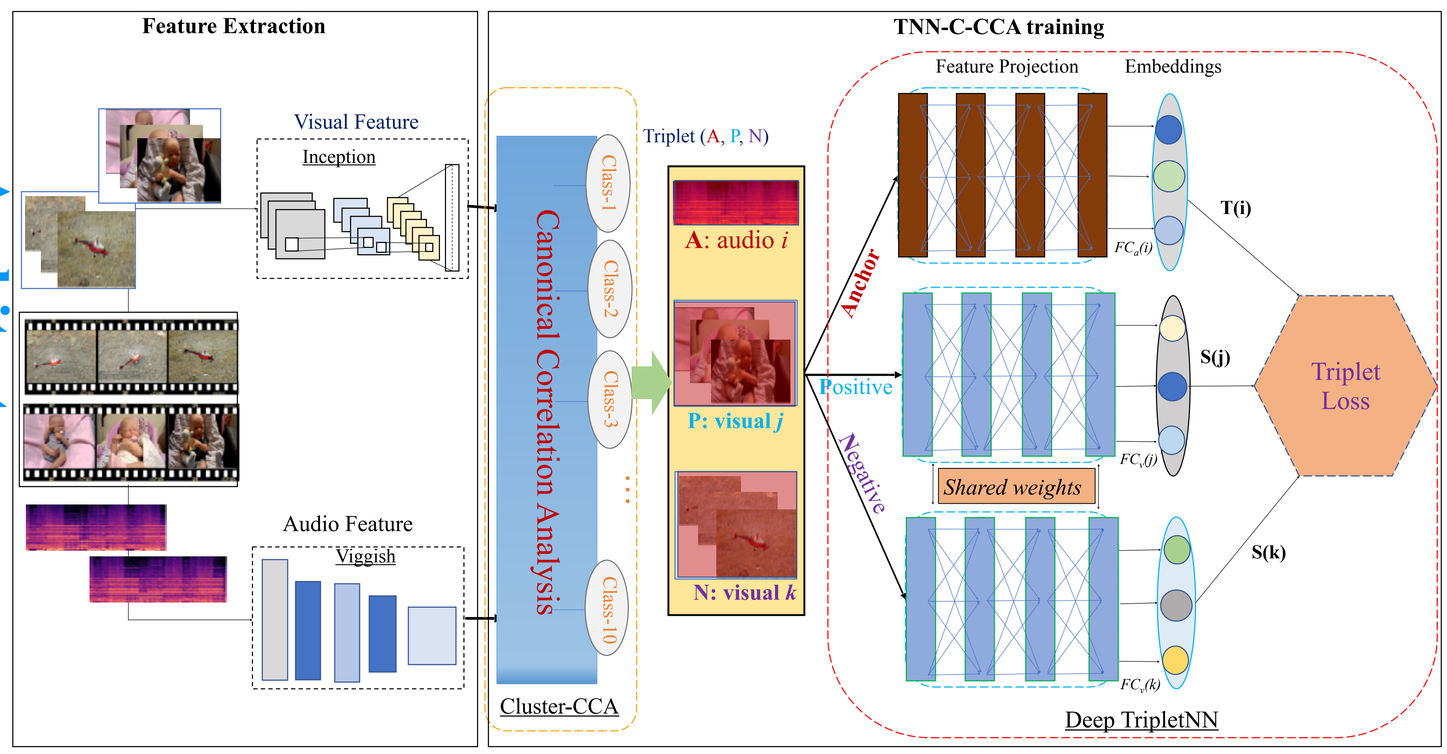
Aiming at the cross-modal retrieval between audio and visual, we try to exploit the temporal structure of audio and visual signal, and learn a deep sequential correlation model between them.

内容 Contents

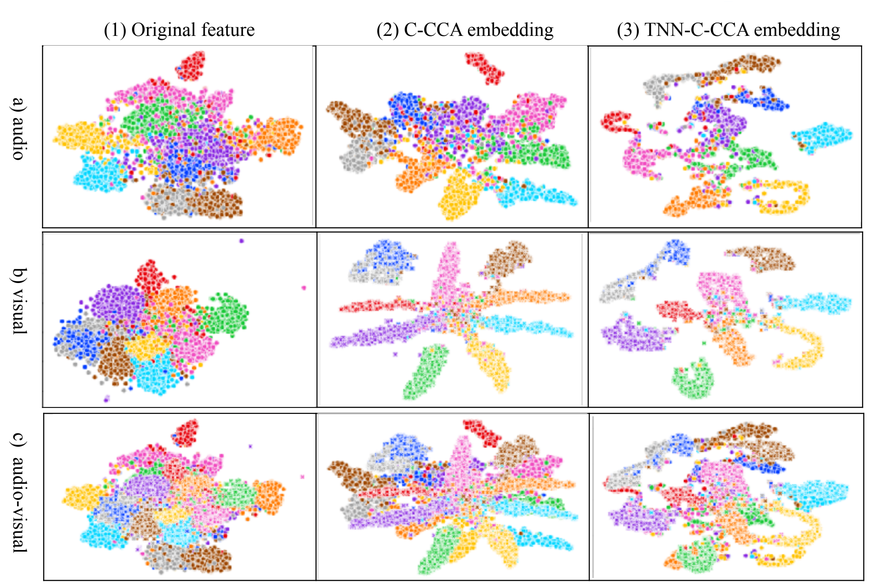
クロスモーダル検索とは、あるモダリティのデータをクエリとして、それに関連する、別のモダリティのデータを取得することです。どのようにモダリティのギャップを跨いで　異質データ間の類似度を算出するか、大きな課題である。イメージとテキスト、オーディオとテキスト、およびビデオとテキストのクロスモーダル検索については近年広く議論されていますが、オーディオとビジュアルのクロスモーダル検索については、時間的クロスモーダル構造の特徴がありますが、アラインメント関係を表すデータがないため、あまり検討されていません。そこで、本研究では、オーディオとビジュアルのクロスモール検索を実現するために、その間の時間的構造を考慮した相関関係を学習することに焦点を当てています。

A cross-modal retrieval process is to use a query in one modality to obtain relevant data in another modality. The challenging issue of cross-modal retrieval lies in bridging the heterogeneous gap for similarity computation, which has been broadly discussed in image-text, audio-text, and video-text cross-modal multimedia data mining and retrieval. However, the gap in temporal structures of different data modalities is not well addressed due to the lack of alignment relationship between temporal cross-modal structures. Our research focuses on learning the correlation between different modalities for the task of cross-modal retrieval. We have proposed an architecture: Supervised-Deep Canonical Correlation Analysis (S-DCCA), for cross-modal retrieval. In this forum paper, we will talk about how to exploit triplet neural networks (TNN) to enhance the correlation learning for cross-modal retrieval. The experimental result shows the proposed TNN-based supervised correlation learning architecture can get the best result when the data representation extracted by supervised learning.

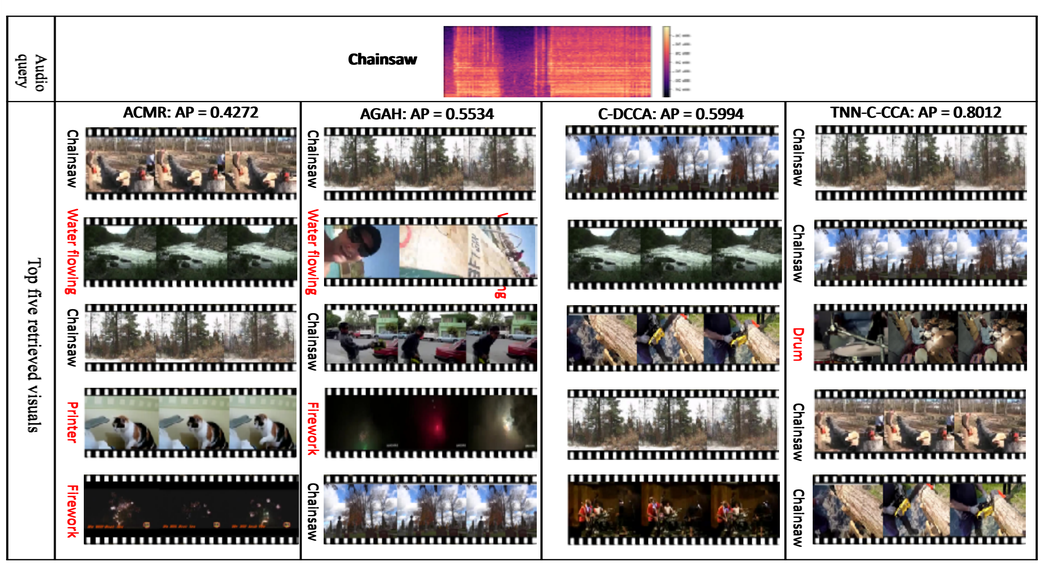
研究アーキテクチャと結果 Research Architecture and results



The visualization of the two learned subspace



The visualized audio-visual retrieval results



The MAP scores of audio-visual cross-modal retrieval

Models	VEGAS Dataset (%)		MV-10K Dataset (%)	
	audio→visual	visual→audio	audio→visual	visual→audio
CCA [8]	32.43	32.11	18.38	18.17
KCCA [17]	28.65	27.24	17.81	17.03
DCCA [2]	41.43	42.15	18.43	18.21
C-CCA [25]	65.16	64.35	19.71	19.62
C-KCCA [25]	32.41	32.74	18.38	18.11
C-DCCA [57]	70.34	69.27	21.79	20.08
UGACH [58]	17.18	17.07	11.11	11.40
AGAH [7]	57.82	56.16	20.74	20.19
UCAL [9]	42.68	41.53	18.82	18.47
ACMR [39]	45.46	43.12	19.02	18.63
LSTM_C_CCA	66.62	71.34	19.11	18,89
TNN-C-CCA	74.66	73.77	23.34	21.32

NII
連絡先：Yu Yi (ユイ）／ 国立情報学研究所　コンテンツ科学研究系
TEL : 03-4212-2574　　FAX : 03-4212-2035　Email : yiyu@nii.ac.jp