マルチモーダル自己教師あり学習におけるモーダルの組み合わせによる学習への影響の分析

ER20038 小林　亮太　　指導教授：藤吉 弘亘

1. はじめに

マルチモーダル自己教師あり学習では，ビデオ，オーディオ，テキストなどの様々なモーダルを用いて学習をする．その中で，3 つのモーダルはそれぞれ異なる内容のデータを持っている．ビデオは動作の主役となる物体と背景，オーディオではナレーションの音声と動作に伴う音やその他の雑音，テキストはナレーションの内容というような形をしている．以上のことからオーディオやビデオと比較してテキストは抽象的な特徴を持つことが分かり，モーダル間の性質の違いが確認できる．このことから，モーダルの組み合わせ方によって背景や雑音を維持・軽減した学習が可能であると考えられる．

本研究では，モーダルの性質に応じた対照学習の設計による事前学習の性能改善を目的とし，その事前調査としてマルチモーダル自己教師あり学習において，近づけるモーダルの組み合わせによる学習効果への影響について調査する．

2. Multimodal Clustering Network (MCN)

マルチモーダル自己教師あり学習の手法のひとつとして Multimodal Clustering Network (MCN) [1] がある．ここで MCN のアーキテクチャを図 2 に示す．

MCN では，図 1 に示すように学習で使用する 3 つのモーダルの各入力はそれぞれ特徴抽出器を用いて抽出した特徴を使用する．入力された各モーダルの特徴は線形射影により埋め込み表現へ変換される．

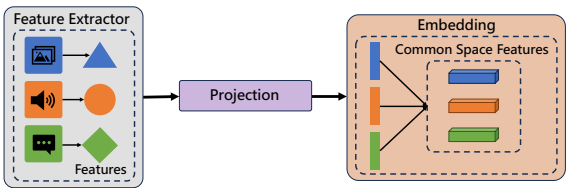

図 1：MCN のネットワーク構造

MCN では，式 1 に示す様に 3 つ損失を用いて学習する．ここで，$L_{Combine}$ は 3 つの損失の和であり，それぞれ以下のような損失である．

$$L_{Combine} = L_{MMS} + L_{Clustering} + L_{Reconstruction} \quad (1)$$

L_{MMS} は単一の共通の埋め込み空間において 3 モーダルにおける 3 つのペア（テキスト，オーディオ），（オーディオ，ビデオ），（ビデオ，テキスト）の時間的な距離を近づける損失を表す．

$L_{Clustering}$ は K-means 法でクラスタリングをして各クラスタの重心を算出した上での各モーダルと重心を意味的に近づける損失を表す．

$L_{Reconstruction}$ はオートエンコーダによる再構築の前後のデータを近づける損失を表す．これらの損失を同時に使用することによって，異なる時系列における行動や動作の類似性を確保することができるようになっている．この手法では，全てのモーダルの共通空間を利用していることにより検索タスクなどをモーダル間の隔たりなく実行できるという利点が存在している．しかし，逆に単一の共通空間を利用していることにより，オーディオやビデオに比べテキストが抽象的であることからオーディオやビデオに含まれているきめ細かい表現や情報が失われる可能性があるという欠点がある．

3. 評価実験

本実験では，モーダル毎の性質の違いによるマルチモーダル自己教師あり学習への影響を調べることを目的として，モーダルの組み合わせによる学習効果を調査する．

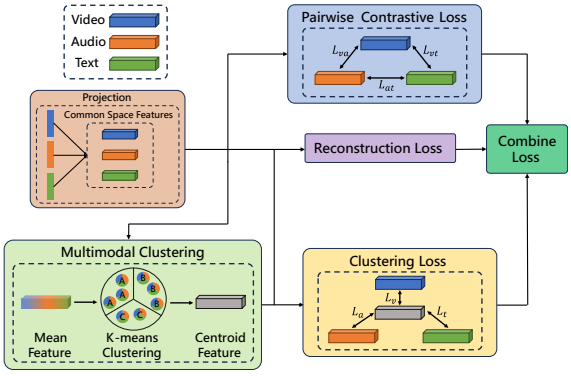

図 2：MCN のアーキテクチャ

3.1 実験概要

本実験では，マルチモーダル自己教師あり学習として MCN を用いてテキストからビデオの検索および時系列行動検出による評価を行う．ここで時系列行動検出とは特定の行動のビデオ内のおける開始時刻と終了時刻の範囲を検出するものである．また，学習用データセットには HowTo100M[2] を，評価用データセットには YouCook2[3] および msrvtt[4] を用いる．HowTo100M は，Youtube 上のビデオを利用しているナレーション付きビデオの大規模なデータセットである．HowTo100M は論文中では約 123 万本のビデオからなるとされているが，Youtube 上での削除や非公開化によって利用可能なデータの総数が減少傾向にある．今回の実験では，現時点で残存する約 90 万本のビデオを用いる．

3.2 実験条件

データセットに含まれるビデオは全て 454 × 256 の解像度と 30FPS のフレームレートで統一した．特徴量抽出には，それぞれのモーダルに対して学習済みのモデルを用いる．ビデオには ResNet152，オーディオには DaveNet[5]，テキストには Word2vec を使用する．学習条件は，学習率 0.0001，バッチサイズ 128，エポック数 30，最適化手法は Adam とした．

4. おわりに

本研究では，マルチモーダル自己教師あり学習におけるモーダルの性質に応じた対照学習の設計による事前学習の性能改善を目的として，モーダル間の影響についての調査を行った．MCN は単一の空間を利用した手法だったが，複数の埋め込み空間に分けることでモーダル特有の情報を活用が期待できると考えている．今後も引き続きモーダルの組み合わせによる学習効果への影響について他の手法に調査も加えつつ行う．

参考文献

[1] B. Chen, *et al.*, “Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos”, ICCV, 2021.

[2] A. Miech, *et al.*, “HowTo100M:Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips”, ICCV, 2019.

[3] L. Zhou, *et al.*, “Towards automatic learning of procedures from web instructional videos”, AAAI, 2018.

[4] J. Xu, *et al.*, “MSR-VTT: A large video description dataset for bridging video and language”, CVPR, 2016.

[5] D. Harwath, *et al.*, “Jointly discovering visual objects and spoken words from raw sensory input”, ECCV, 2018.