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効果への影響の分析

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つ．ビデオは動作する物体と背景，オーディオでは動作についてのナレーションの音声や動作音と雑音，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最終目的とし，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，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を目的とする．

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 1 に示す．

MCN では，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特徴空間へ埋め込む．

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 つの損失の和である．

$$L_{Combine} = L_{Contrast} + L_{Clustering} + L_{Recon} \quad (1)$$

$L_{Contrast}$ 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）のビデオ内における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_{Recon} は各モーダルにおいて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．

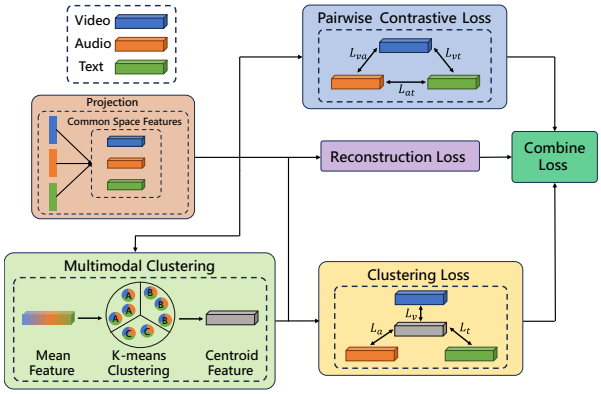

図 1：MCN のアーキテクチャ

3. 調査実験

本実験では，モーダルの組み合わせ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，影響調査のための比較の対象として 3 モーダルを一度に学習した場合の自分の環境での精度を知るために先んじて MCN の再現実験を行う．

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による評価を行う．

3.2 実験条件

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] を用いる．

HowTo100M は，Youtube 上のビデオを利用している大規模なデータセットである．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．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する．

評価には，テキストからのビデオの検索タスクを YouCook2 データセットを用いてゼロショット評価を行う．

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[?], 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，n_pair=32，最適化手法は Adam とした．ここで n_pair はビデオ 1 本あたりの分割数を示しており，バッチサイズがビデオの本数を表す．

4. 実験結果

再現実験の学習結果を表 1 に示す．ここで R は再現率を表しており，@に付随する値は R@k とした場合各クエリの上位 k 個の予測のうちの正解数の総正解数に対する割合を表している．以下の結果より概ね再現できていると考える．

また，この結果をモーダルの組み合わせを変えた実験における比較対象とする．

表 1：精度比較

	R@1	R@5	R@10
論文	18.1	35.5	45.2
再現実験	14.7	34.1	44.7

5. 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最終目的として，その事前調査として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実験を通して調査を行う．

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.