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

ER20038 小林　亮太　　指導教授：藤吉 弘亘

1. はじめに

深層学習では，人間の認知プロセスのように複数の感覚 (モーダル) を組み合わせたマルチモーダル学習が注目されている．マルチモーダル学習は，同時刻に収集した異なるモーダルデータを用いて学習する．しかしながら，同時刻における各モーダルデータは完全一致していないことも多い．例えば，テキストにはオーディオやビデオに含まれる雑音のようなノイズとなる情報が存在せず，モーダルごとに性質が異なる．そのため，モーダルの持つ特性を考慮したモデルの設計を行うことで，雑音を軽減した学習が期待できる．また，マルチモーダル学習では使用するデータの多さから，ラベルなしデータのみを用いる自己教師あり学習が用いられることが多い．そこで，本研究では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方による学習への影響調査を行う．

2. Multimodal Clustering Network (MCN)

マルチモーダル自己教師あり学習の代表的な手法として Multimodal Clustering Network (MCN) [1] がある．MCN のアーキテクチャを図 1 に示す．MCN では，各モーダルのデータを特徴抽出器に入力して特徴を抽出し，各モーダルの特徴を線形射影をして特徴空間に埋め込む．

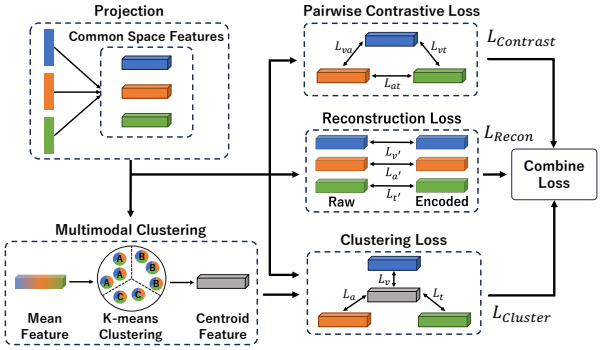

図 1：MCN のアーキテクチャ

損失計算には，式 (1) に示すような $L_{Contrast}$ と $L_{Cluster}$ と L_{Recon} の 3 つの損失関数の総和 $L_{Combine}$ を用いる．

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

$L_{Contrast}$ は，各モーダルのペア（テキスト，オーディオ），（オーディオ，ビデオ），（ビデオ，テキスト）において，動画内で同じ時刻に存在するもの同士を近づけ，そうでないもの同士を遠ざける損失である．$L_{Cluster}$ は K-means 法でクラスタリングを行って各クラスタの重心を算出した上で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．MCN では，全てのモーダルで単一の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する．しかし，共通の空間を利用することは，全てのモーダルが同じ表現力や詳細度もつことを暗黙のうちに仮定しているため，モーダルごとの性質が十分に考慮されていない．

3. 調査実験

ビデオ，オーディオ，テキストの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する．

3.1 実験概要

本実験では，3 つのモーダルを二段階に分けて学習する．一段階目では，2 モーダルのみで学習をし，二段階目では学習済みの 2 つのモーダルを凍結させ，未使用のモーダルを追加して学習する．以上の手順での学習を 3 つのパターンで比較する．

3.2 実験条件

学習用データセットには HowTo100M[2]，評価用データセットには YouCook2[3] と MSR-VTT[4] を用いて，テキストからのビデオの検索タスクでゼロショット評価を行う．HowTo100M は，Youtube 上のビデオを利用する大規模なデータセットである．一部が削除や非公開されているため，利用可能なビデオの数は減少している．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．ビデオは全て 454 × 256 ピクセルの解像度と 30FPS のフレームレートに統一する．ビデオの特徴の抽出には ResNet152，オーディオの特徴の抽出には DaveNet，テキストの特徴の抽出には Word2vec で学習済みのモデルを用いる．学習条件は，学習率が 0.0001，バッチサイズが 128，エポック数が 30，最適化手法は Adam とした．

4. 実験結果

実験の結果を表 1 に示す．ビデオを V，オーディオを A，テキストを T を表す．最初の 2 つのアルファベットは一段階目の学習，最後のアルファベットは二段階目の学習に用いるモーダルを示す．R は再現率であり，@の後の値は，総正解数に対する各クエリの上位 k 個の予測のうちの正解数の割合を表す．表 1 より，YouCook2 では AT_V が最も高い精度である．YouCook2 は料理の調理手順を扱うデータセットであるため，手順を表現する上でテキストが重要な情報を持つ可能性が高く，一段階目でテキストを用いることが効果的だったと考える．また，オーディオも料理手順の順序性を表現できるため、テキストとオーディオの組み合わせが最適だったと考える．一方で，MSR-VTT では AV_T が最も高い精度である．MSR-VTT はビデオとテキストの説明文からなるデータセットであるため，ビデオそのものの内容の理解が重要であると考えられることから，一段階目でビデオを用いることが効果的だったと考える．このことから，タスクによって適切なモーダルの組み合わせは異なると言える．

表 1：精度比較

	YouCook2			MSR-VTT		
	R@1	R@5	R@10	R@1	R@5	R@10
AV_T	1.61	5.97	9.43	**0.18**	**0.92**	**1.63**
AT_V	**5.16**	**11.3**	**15.3**	0.14	0.61	1.35
VT_A	1.10	5.13	7.13	0.14	0.53	1.14

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いて，モーダルの組み合わせ方による学習効果への影響についての調査を行った．データセットにより適切なモーダルの組み合わせが異なることが分かった．今後は，引き続きモーダルの組み合わせ方による学習効果への影響について調査を行う，最終的に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.