第1回ミーティング
論文調査

ER20038 小林 亮太

2023年3月25日

1　はじめに

卒業研究を始めるにあたり，マルチモーダルデータを用いた自己教師あり学習のなかでも3モーダルの場合の知識を得るために今回は，ラベルなしビデオデータからの自己教師あり学習のための Multimodal Clustering Networks (MCN)[1] について論文調査を行った．

2　アプローチ

MCN では，ラベル付けされていないナレーション付きビデオからビデオ，テキスト，オーディオの3つのモダリティを使用して自己教師あり学習を行う．モダリティとは情報を伝達する手段や媒体のことである．3つのモダリティの特徴量を低次元の共通の特徴空間に写像することで複数のモダリティを統合的に扱うことができるようになる．ここで効率的に特徴空間を構築するために，ビデオ，オーディオ，テキストから埋め込み表現を導出する．次にデータを空間内に適切に配置するために，Contrastive Loss の L_{MMS}，Clustering Loss の $L_{Cluster}$，Reconstruction Loss の $L_{Reconstruct}$ の3つの損失関数を導入する．最終的には，以下の3つの損失の合計を最小化するように学習する．次にそれぞれの損失関数について説明していく．

$$L = L_{MMS} + L_{Cluster} + L_{Reconstruct} \quad \cdots\cdots\cdots\cdots\cdots\cdots ①$$

2.1　Contrastive Loss L_{MMS}

図 1(b) に示すように，全てのモダリティのペア, (t，a) , (v，a) , (v，t) に対して損失を算出する．ここで，t はテキスト，a はオーディオ，v はビデオのことを指す．この損失は2つのモダリティの類似しているデータポイント間の距離を最小化し、異なるデータポイント間の距離を最大化するように学習する。L_{MMS} は，3つのモダリティそれぞれを用いた損失の合計である．

$$L_{MMS} = L_{ta} + L_{va} + L_{vt} \quad \cdots\cdots\cdots\cdots\cdots\cdots ②$$

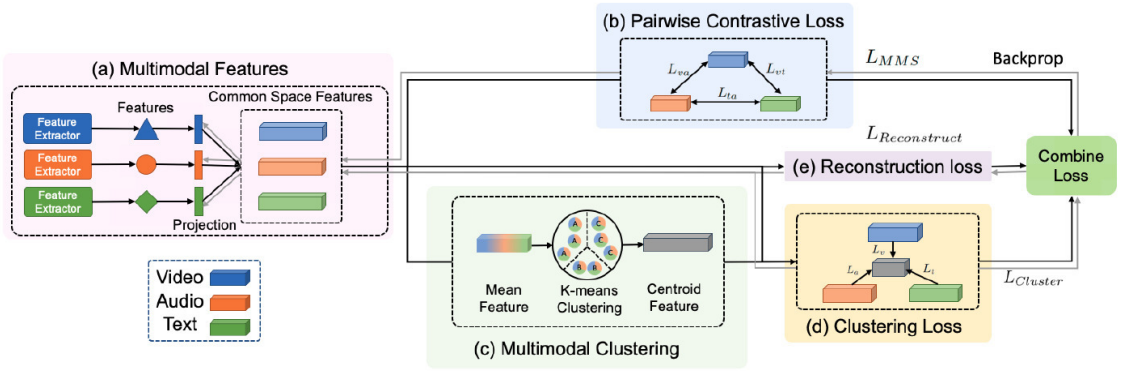

図 1: Illustration of our proposed framework

ここで，L_{ta}，L_{va}，L_{vt} はそれぞれのペア（t，a）,（v，a）,（v，t）の損失を表す．例として，テキストとオーディオの損失 L_{ta} は，以下のように与えられる．

$$L_{ta} = -\frac{1}{B}\sum_{i=1}^{B}[(log\frac{e^{h(t_i)\cdot g(a_i)-\delta}}{h(t_i)\cdot g(a_i)-\delta+\sum_{\substack{k=1\\k\neq i}}^{B} e^{h(t_k^{imp})\cdot g(a_i)}}) \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{③}$$
$$+ (log\frac{h(t_i)\cdot g(a_i)-\delta}{h(t_i)\cdot g(a_i)-\delta+\sum_{\substack{j=1\\j\neq i}}^{B} e^{h(t_i)\cdot g(a_j^{imp})}})]$$

ここで，g(a)，h(t) はそれぞれオーディオとテキストを入力として埋め込み表現を出力とする関数である．また，a^{imp}，t^{imp} は t_i，a_i に対する負のペアとなっており，これらのドット積が小さい，つまり不正解に対する確率が小さいほど損失は少なくなる．加えて，δ はマージンのハイパーパラメータである．

2.2 Clustering Loss $L_{Cluster}$

同一のクラスタに属するものの特徴が似通るようにすることを目的としている．各フレームとクラスタの重心の距離を最小化してより良いクラスタリングを実現する．最初に，k-means 法によってクラスタリングを行い，各クラスタの重心を算出する．ここで k 個のクラスタの重心を行列 C={μ1,...,μk} として出力する．次に 3 つのモダリティからの特徴を，その特徴に近い重心に引き寄せる．ここで，例としてテキストの損失 L_t は次のように与えられる．

$$L_t = -\frac{1}{B} log\frac{e^{h(t_i)\cdot\mu^{'}-\delta}}{\sum_{k=1}^{K} e^{h(t_i)\cdot\mu_k}} \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{④}$$

ここで，$\mu^{'}$ はその時の t_i に最も近い重心であり，μ_k は k 番目のクラスタの重心であることから属するクラスタの重心になるべく近くかつ他のクラスタの重心からはなるべく離れているときほど損失は少なくなることが分かる．

$L_{Cluster}$ は，3 つのモダリティの損失，L_t，L_a，L_v の和である．

$$L_{Cluster} = L_v + L_a + L_t \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{⑤}$$

2.3 Reconstruction Loss $L_{Reconstruct}$

Reconstruct（再構成）は，オートエンコーダで行い，入力の元データと出力の再構成後データを比較して，その差を小さくする．これによって，Contrastive Learning やクラスタリングによって抑制された特徴を捕らえることができる．また，再構成は正則化項を加えることによって汎化性能を改善する補助的なタスクでもある．例として，視覚モダリティについて，損失 $L_{v'}$ は次のように与えられる．

$$L_{v'} = -\frac{1}{B}\sum_{i=1}^{B} \parallel f^{'}(v) - f(v) \parallel_2 \quad \cdots\cdots\cdots\cdots\cdots\cdots ⑥$$

ここで，$f^{'}$ は再構成後で f は再構成前である．
そして，各モダリティの損失を合計したものが $L_{Reconstruct}$ となる．

$$L_{Reconstruct} = L_{v'} + L_{a'} + L_{t^t} \quad \cdots\cdots\cdots\cdots\cdots\cdots ⑦$$

3 おわりに

今回はマルチモーダルデータの自己教師あり学習について論文調査を行った．今後としては，より理解を深めるために関連の論文などの資料の調査を行う必要がある．

参考文献

[1] Brian Chen, et al.,“Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos”, ICCV2021.