第4回ミーティング
MCNの再現実験と論文調査

ER20038 小林 亮太

2023年5月20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．それに引き続き，調査した論文内で紹介されていた手法で論文と同じ精度が達成できるかを確認するために再現実験を行う．また，MCN とは別の手法について論文の調査を行った．

2　実験準備

再現実験を行うにあたり，その準備を可能な限り行った．以下に完了したものとそうでないものを示す．

2.1　完了した準備

2.1.1　プログラムのダウンロード

論文で紹介されている手法のソースコードが github に存在しているため，clone コマンドでダウンロードした．また，ダウンロードしたファイルを解凍した．

2.1.2　Dockerfile の作成

実行予定のプログラムから必要なモジュールなどの情報を読み取り，適切な dockerfile を作成を行う．しかしながら，一部のモジュールはビルド段階ではエラーとなり実装できなかったため，ビルド後にコンテナ内で手動にてインストールをした．

2.2 未完了の準備

2.2.1 データセットのダウンロード

今回の手法で利用する HowTo100M データセットのダウンロードが必要である．そのため，公式サイトから All-in-One とされている zip ファイルをダウンロードした．しかしながら，ダウンロードしたファイルはデータセットそのものではなく，別途フォームから申請を行い，ユーザー ID とパスワードを受け取る必要がある．以下がデータセットのダウンロードに関する現在までの進捗である．

- 4/13：フォームを送信
- 4/27：著者にメールを送信
- 現在は返答待ち

また，返答があまりにも来ないこと場合，youtube-dl を利用することも検討している．
データセットについての情報を以下にまとめる．

- ナレーション付きビデオの大規模データセット
- 120 万本の Youtube ビデオのビデオにキャプションを付けた 1 億 3600 万本のビデオクリップ
- 23000 のカテゴリが存在
- 全てをダウンロードする場合，12TB ほどの容量が必要

3 論文調査

今回は，再現実験のためのデータセットの準備に加えて，Everything at Once - Multimodal Fusion Transformer for Video Retrieval[2] について論文調査を行った．

3.1 アプローチ

この手法では，多様なモダリティ間の学習をする Multimodal Fusion Transformer (MFT) を提案している．また，Transformer を用いていることから，self-attention の各要素が他のすべての要素と比較されるという処理によって異なる長さの入力を処理することが可能となっている．加えて，Combinatorial Loss という入力モダリティのすべての組み合わせに対して対照損失を計算するような損失関数を用いている．

3.1.1 Multimodal Fusion Transformer (MFT)

MFT では，各モダリティから得られた特徴量を，それぞれキー，クエリ，バリューに変換し，キーとクエリを用いて各モダリティから得られた特徴量を比較し，各トークンのセルフアテンションを計算す

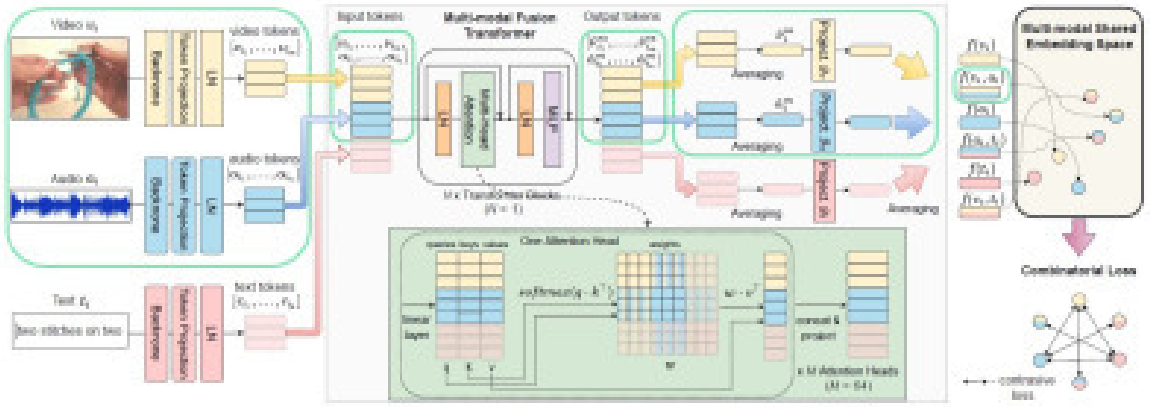

図 1: Schematic visualization the proposed method

る．その後，得られた重みとバリューを結合して得た特徴量に対して MLP を適応する．これらの処理を複数回繰り返すような処理となっている．
複数のモダリティから得られた特徴量を収集，結合することによってより豊かな表現空間を作成することが可能となる．

3.1.2 Combinatorial Loss

Combinatorial Loss は，すべての入力の組み合わせにおける対照損失を計算するものであり，その組み合わせは単一のモダリティ，モダリティのペアを含めたすべての組み合わせとなっている．また，この損失関数では比較対象同士が同じラベルを持つかどうかを比較するものである．
処理の流れとしては MFT によって出力トークンが生成されたのち，その出力を各モダリティに分割してそれぞれにおいて平均化をする．次に平均化された特徴量を用いて損失を計算する．そして，これらを最小化するように学習を行う．

3.2 学習時間の見込み

この手法の再現実験を行う場合に予想される所要時間を以下に示す．

- HowTo100M を用いたモデルのトレーニング
  論文著者の GPU 環境と所要時間は 4 つの Nvidia v100 32GB で 2 日となっているので，これと自分の環境 3 つの Nvidia A100 40GB を参考にすると 1 日半から 2 日かかる予想となる．
- ファインチューニング
  こちらは論文著者の GPU 環境は同じで 30 分となっているので，こちらも同じ GPU 環境で考えると十数分程度の予想となる．

データ容量の大きい HowTo100M を用いているのにも関わらず，比較的短時間で学習が完了する．これは HowTo100M のデータ容量が大きいことが含まれるデータが動画であるという点が起因しており，画像を用いた学習と比較する際単純なデータセットの容量で比較することができないと考えられる．そのため，学習時間がデータ容量に対して比較的短いことがある．

4 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の再現実験の準備を行った．進捗としては，プログラム実行時にプログレスバーが表示されるところまでは進んでいるので，データセットを入手すればすぐにでも動かすことができると考えられる．今後としては，データセットの入手を待ちつつ，再現実験の対象を今回調査した論文の手法のファインチューニングに切り替えていくことも考えられる．そのため，ファインチューニングのためのデータセットの準備を優先して進める必要がある．

参考文献

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

[2] Nina Shvetsova, et al.,“Everything at Once - Multimodal Fusion Transformer for Video Retrieval”, CVPR2022.