第2回ミーティング
論文調査

ER20038 小林 亮太

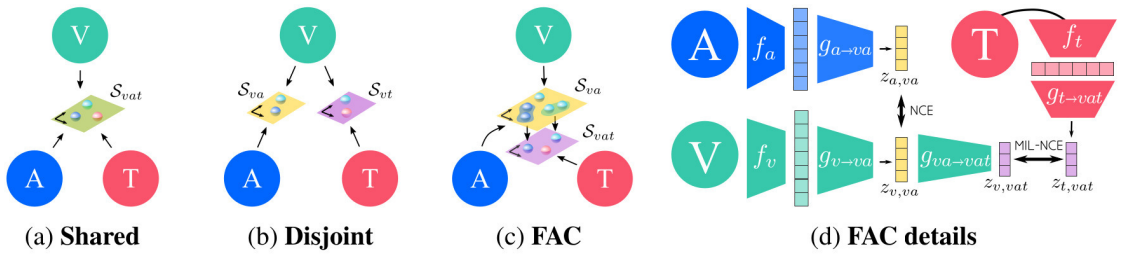
2023年4月8日

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，Multimodal Versatile Networks (MMV)[1] についての論文調査を行った．

2　アプローチ

MMV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．また，今回はテキストのモダリティはオーディオに自動音声認識 (ASR) を用いて生成を行っている．
次に学習を行う上で4つの特性を備えるように設計を行うことが目的である．(i)3つのモダリティのいずれかを入力とすることができること，(ii) モダリティの特異性を重視すること，(iii) 学習中に一度も見たことがなくても，異なるモダリティを容易に比較できること，最後に (iv) 動画または静的画像として得られる視覚データを効率的に適用できること，の4点である．そして，論文内ではこの4つの特性を備えたネットワークの設計方法が3つ検討されている．次からそれぞれの方法について説明していく．



(a) **Shared**　(b) **Disjoint**　(c) **FAC**　(d) **FAC details**

図 1: (a)-(c)Modality Embedding Graph，(d)Projection head and losses for FAC

2.1 Shared spaces

図 1(a) に示すように，Shared spaces では全てのモダリティが単一の共有ベクトル空間 S_{vat} に組み込まれる．これによって全てのモダリティを直接比較することが可能となり，先述した特性 (iii) を満たすことができる．しかしながら，デメリットとして全てのモダリティが均一な粒度をもつことを暗黙の裡に仮定している点が挙げられる．粒度とは，入力データの分割の程度の事であり，例えばテキストであれば単語ごとに分割を行うなど，各モダリティによって異なる粒度をもつ．その粒度が均一であることによって，モダリティの特異性を無視することになり特性 (ii) が欠如することになる．

2.2 Disjoints spaces

図 1(b) に示すように，Disjoints spaces ではビジュアルとオーディオ，ビジュアルとテキストのペアをそれぞれ比較するために 2 つの別々の空間 S_{va}，S_{vt} を使用する．このような不連続な空間でのアプローチは，異なるモダリティのペアの特異性をある程度重視する．これによって Shared spaces の欠点は補っていると言える．しかしながら，2 つの空間同士の比較をすることができないため，テキストからオーディオへの検索などは不可能となっている．これは特性 (iii) の欠如となってしまっている．

2.3 Fine and Coarse spaces (FAC)

図 1(c) に示すように，Fine and Coarse spaces ではオーディオとビジュアルの 2 つとテキストではモダリティの粒度が大きく異なるという点に着目して，オーディオとビジュアルを比較する空間 S_{va} と空間 S_{va} から投影してテキストと比較を行う空間 S_{vat} の 2 つを使用する．オーディオとテキストやビジュアルとテキストを比較する際は，微細な空間 S_{va} を介して，粗大な空間 S_{vat} に投影する．この手法では，設計する上で備えるべき特性をすべてカバーしている．そのためこの手法が用いられる．

3 Multimodal Constrastive Loss

ここからは FAC を用いて，学習を行う際に導入する損失関数について説明していく．まず，損失関数は以下 1 のように与えられている．

$$L(x) = \lambda_{va} NCE(x_{[}v], x_a) + \lambda_{vt} MIL - NCE(x_v, x_t) \quad \cdots\cdots\cdots\cdots\cdots\cdots ①$$

ここで λ_{va}，λ_{vt} はそれぞれのペアの重みを表している．次に NCE，MIL-NCE について説明していく．

3.1 Noise Constrastive Estimation (NCE)

NCE では，与えられたサンプルが真のペアであるかを推定するために偽のペアと比かっくして確率を出力する．NCE は以下 2 のように与えられている．

$$NCE(x_v, x_a) = -log(\frac{exp(\frac{z_{v,va} \cdot z_{a,va}}{\tau})}{exp(\frac{z_{v,va} \cdot z_{a,va}}{\tau}) + \sum_{z' \sim N(x)} exp(\frac{z'_{v,va} \cdot z'_{a,va}}{\tau}))} \quad \cdots\cdots\cdots\cdots\cdots\cdots ②$$

ここで，τ は温度パラメータである．P(x) と N(x) はそれぞれ，真のペアの集合と偽のペアの集合を表している．$z_{k,mn}$ はモダリティペア mn の空間におけるモダリティk を表している．これは，図 1(d) に示すように 1 つ目のビジュアルとオーディオを比較する空間における式であることから添え字が va となっている．加えて，z 同士のドット積は真のペアを表しており，$z^{'}$ 同士のドット積は偽のペアを表している．

3.2 Multiple Instance Learning (MIL)-NCE

MIL-NCE は MIL と NCE を組み合わせた手法である．MIL は，複数のインスタンスが存在するようなデータを扱う際に使用される手法であり，NCE は MIL と NCE を組み合わせた手法である組み合わせてモダリティのデータを一度に複数用いて計算を行う．MIL-NCE は以下 3 のように与えられる．

$$MIL-NCE(x_v,x_t) = -log(\frac{\sum_{z\in P(x)} exp(\frac{z_{v,vat}\cdot z_{t,vat}}{\tau})}{\sum_{z\in P(x)} exp(\frac{z_{v,vat}\cdot z_{t,vat}}{\tau}) + \sum_{z'\sim N(x)} exp(\frac{z^{'}_{v,va}\cdot z^{'}_{a,va}}{\tau}))} \cdots\cdots ③$$

ここでは 2 の式と同様に z の添え字は 1 つのモダリティどの空間であるかを示しており，図 1 に示すように空間 S_{va} からの投影とテキストを比較する空間 S_{vat} に関する式である．

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について論文調査を行った．今後，前回と続けて調査した 2 つの論文の再現実験を行い，より理解を深める必要がある．

参考文献

[1] Jean-Baptiste Alayrac, et al.,“Self-Supervised MultiModal Versatile Networks”, NeurIPS2020.