6V-04

Vision & Language Model におけるユニモーダルの特徴間のアラインメントによる VQA の学習速度の改善

高平 凛† 楊 陽‡ 小松 瑞果‡ 大川 剛直‡

神戸大学† 工学部情報知能工学科 神戸大学 システム情報学研究科‡

1. はじめに

近年、画像とテキストなど、複数の情報を活用するマルチモーダル深層学習が注目されている。例えば、Image Captioning や Visual Question Answering(以下、VQA)のタスクに対しては、画像と関連テキストのペアからなるデータを扱う Vision & Language モデル(以下、V&L モデル)などが用いられている。このモデルは、Attention と呼ばれる機構で対応する画像とテキストを融合させることで、言語信号に基づき画像から対応する情報を抽出できることが期待されている。しかし、個々のデータはそれぞれ異なる統計的特性をもつので、同じ意味をもつ言語と画像の特徴の関連度が低く、そのため特徴量を融合させるための学習に時間がかかることが多い。そこで本研究では、融合する前に対応づけられた画像特徴と言語情報の関連度を高める、つまりアライメントを行うことで学習の収束を早めることを試みる。

2. 関連研究

2.1 Vision & Language モデル

V&L モデルとは画像とテキストを入力とし、異なるモダリティのデータから共通表現を学習するモデルを表す。学習は事前学習とファインチューンの二段階に分かれている。事前学習では、大規模な言語と画像のデータペアを使用し、Image Text Matching などの事前学習タスクで画像とテキストの一般的な表現を学習する。ファインチューンでは、VQA などの下流タスクのデータと教師情報で学習し、モデルの下流タスクへの性能を高める。V&L モデルの構造は Embedding と Modality Interaction に分かれている。図 1 に V&L モデルの概要と、その一例である Vilt[1]の構造を示す。Embedding ではそれぞれの入力を埋め込みベクトルに変換する。Visual Embedding の方法は主に 3 種類ある。V&L における各手法の課題として、画像とラベルから成るデータセットで学習を行う region-based は、学習されていない物体は検出できないこと、grid-based はテキストに関連しない不要な特徴も抽出してしまうこと、patch projection は推論が早い一方で同様に不要な情報が抽出されることが挙げられる。Modality Interaction では、抽出した各特徴を融合する。基本的に Transformer が使用されており、主にその内部の Attention 機構によって融合が行われる。

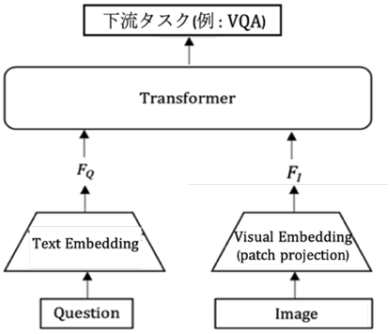

図 1. V&L モデルの基本的な構造(例: Vilt)

2.2 CLIP

CLIP[2]は V&L の事前学習モデルの一つである。インターネットから収集された大量の画像とテキストのデータペアに対し内積を計算し、対応するペアの積を最大化するように事前学習を行なっている。そのためテキストに基づき画像から特徴を抽出できる。CLIP の学習は各特徴の内積とそれぞれの画像特徴と言語特徴の cross-entropy 損失関数を使用している。そのため言語特徴と画像特徴の関連度が考慮されていないという側面をもつ。

2.3 infoNce 損失関数

言語特徴と画像特徴の関連度の向上を目的とし、本研究では infoNce[3]という損失関数に着目する。infoNce は、ペアになる画像X_iと言語X_jの特徴I_i, T_j間の類似性を最大にし、異なるサンプルの特徴I_i, T_k間の類似性を最小にする。言語特徴と画像特徴の関連度を制約条件として学習できる。

$$\tilde{\mathcal{L}}(X_i, X_j) = -log\frac{\exp(sim(I_i T_j)/\mathcal{T})}{\sum_{k=1}^{2M} exp(sim(I_i T_k))} \quad (k \neq i) \quad (1)$$

ここで$\mathcal{T}$は温度パラメータ、$sim()$はcos類似度を表す。

3. 提案手法

本研究では事前学習済みモデルの下流タスクでのファインチューンの学習時間を減らすことを目的とする。時間がかかる原因として、画像と言語それぞれの特徴が異なる統計的特徴を持つ点が挙げられ

Reducing training time on VQA by alignment in Vision & Language Model

† Rin Takahira, Facility of Engineering, Kobe University

‡ Yang Yang, Mizuka Komatsu, Takenao Ohkawa, Graduate School of System Informatics, Kobe University

る。そこで融合前に各モダリティの関連度を向上させるという着想のもとベースラインモデルである V&L モデルへの CLIP と infoNce の導入、VQA 適用のための修正版 infoNce 損失関数の導入を提案する。提案手法の概要を図 2 に示す。

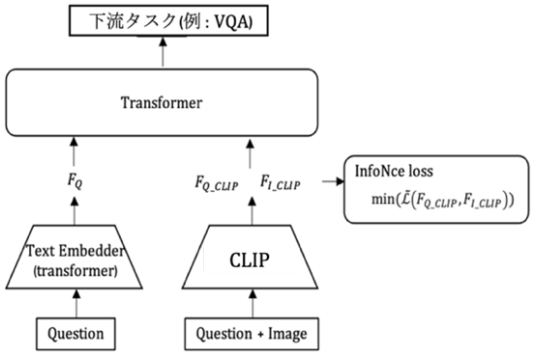

図 2. 提案手法の構造

CLIP の事前学習時の特性から、各モダリティに対する出力特徴の関連度は間接的に高くなる。しかし CLIP において導入されている損失関数を用いた場合、各特徴の直接的な関連度は考慮されない。そこで出力特徴に対して infoNce を導入することで言語特徴と画像特徴の直接的な関連度も高めることが狙いである。具体的には、V&L 事前学習済みモデルの一つである Vilt をベースラインモデルとし、Visual Embedding を patch projection から CLIP に変更する。CLIP に画像とテキストを入力し、Vilt の Text Embedder にテキストを入力、そしてそれぞれの出力を融合する。また CLIP の出力の画像特徴と言語特徴I_i, T_iの関連度の算出に infoNce を導入し、それぞれのモダリティの関連度向上を図る。

VQA データセットに適用する際の特有の問題点として、画像と質問の内容が一致していないペアにも同じ手法を適用した時、類似度最大化に悪影響を及ぼしてしまうという点がある。この問題を解消すべく infoNce における、cos 類似度を特徴ベクトルの内積に変更する。下式において$s(z_i, z_j)$は内積を表す。

$$\tilde{\mathcal{L}}(X_i, X_j) = -log \frac{\exp(s(z_i, z_j)/\mathcal{T})}{\sum_{k=1}^{2M} exp(s(z_i, z_k))} \quad (k \neq i) \quad (2)$$

以上により特徴融合にかかる時間を抑え、収束を早められることが期待される。

4. 実験

4.1 データセット

実験には VQA2.0 データセットを使用する。画像と質問文、解答がセットになったアノテーションデータセットを利用する。画像は MS COCO のものであり、一つの画像に対し複数の質問がセットになっている。質問のタイプは“what”で始まるものや、“yes/no”で答えるものなどの種類がある。

4.2 実験方法

Vilt の事前学習済みモデルを用い、VQA データセットで同じハイパーパラメータ設定でファインチューンを実施する。ベースラインモデルである Vilt と、infoNce の式を変更せずそのまま使用した提案モデル、提案モデル+infoNce の式を修正、の三つの実験を行い、学習ステップ数に対する損失の値から学習速度を比較する。

4.3 実験結果・考察

実験結果を図 3 に示す。結果より、初めは Vilt モデルの学習速度は早いが途中で提案手法のモデル二つの方が速度が向上していることがわかる。これは VQA の答えを出力するための分類層から、融合部分のパラメータ調整に移行した後に学習速度が速くなっていることを示しており、CLIP と infoNce による各モダリティの関連度向上の効果と考えられる。また提案モデルでは安定後にも損失の振れが生じていることがわかる。一方で修正版 infoNce のモデルでは比較的安定しているように見られる。ここから内積に変更することでデータセットの画像と質問の内容が一致していないペアによる影響を抑えることができたといえる。

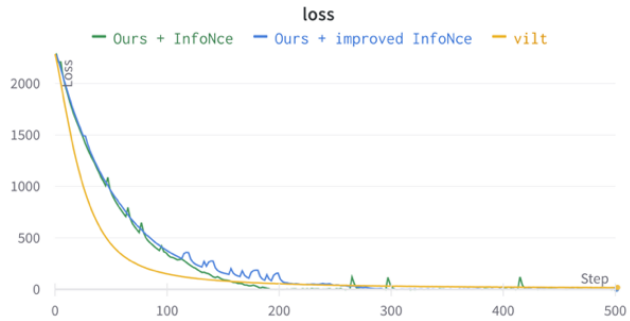

図 3. 実験結果

5. おわりに

CLIP と infoNce により各モダリティの特徴を融合する前に類似度を高めておくことで学習速度を向上させることができた。

謝辞
本研究の一部は JSPS 科研費 21H04914 の助成による。

参考文献

[1] Kim, W., Son, B. and Kim, I. "Vilt: Vision-and-language transformer without convolution or region supervision." International Conference on Machine Learning. PMLR, 2021.
[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning. PMLR, 2021.
[3] Oord, A., Li, Y. and Vinyals, O. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748.