大規模言語モデルを用いた Vison-Language Model による詳細画像分類

1　はじめに

Vison-Language Model(VLM) とは，画像処理と自然言語処理を融合させたモデルであり，近年では CLIP [1] や ALIGN，BLIP などの台頭によって様々な研究がされている．特に CLIP は，画像分類のタスクにおいて，インターネット上の大量の画像とテキストを対照学習することで，新しいデータセットに対して追加学習不要なゼロショット推論が可能となっている．

しかし，このような事前学習された大規模なモデルは，詳細な画像分類タスクにおいては性能が良くないことがわかっている．例えば，CUB [6] データセットの鳥類の種名に視覚属性の説明を付与すると，ゼロショット分類精度が 50.5% から 50.7% に僅かに向上するが，Cars [3] データセットでは僅かに低下することがわかっている．

そこで，アダプタチューニングというものがある．アダプタとは，基盤モデルのような大きな事前学習済みモデルに対し、小さなネットワークを差し込み、その部分のみ訓練することで任意の訓練データに対して適用させるものである．同じように事前学習済みモデルを再学習する手法としてファインチューニングがあるが，ファインチューニングは元のモデルの一部も学習し直すため，特に事前学習済みモデルが大きい場合に効率が良くない．さらに元のモデルが変わってしまうため破滅的忘却が起こる．

また，事前学習済みモデルを再学習する際に問題となるのはデータセットの収集である．近年ではモデルサイズの肥大化が顕著であり，それに応じて学習に必要な画像数も増加傾向にあるため，十分な量の画像を収集することは大きなボトルネックになる．特に詳細画像分のように分類する対象が特殊な場合，十分な量の画像の収集はより困難となる．

そこで，テキストを用いてアダプタを学習する手法に平野 [7] の手法がある．この手法では，固定の学習済み画像エンコーダと大規模言語モデルの間を軽量な変換器で繋ぐことで Vision-Language タスクを解く BLIP-2 [4] で用いられている Q-Former を用いてテキスト特徴量から画像特徴量への変換を学習させる．これにより，テキスト特徴量から画像特徴量への変換を可能にし，Q-Former によって変換されたテキスト特徴量を用いてモデルをファインチューニングすることで，テキストから詳細画像分類モデルを学習させている．

しかし，平野の実験ではアダプタを学習する際に，Reed ら [5] の人の手によって作成されたテキストを用いていた．1 枚の画像に対して 10 個のテキストを作成しているので，この手法では膨大な量のテキストを人の手で作成する必要があり，大きなデータセットを用いる場合に現実的ではない．そこで，今回の実験の目的は，大規模言語モデル (LLM) を用いて画像に対するテキストを生成し，そのテキストを用いて平野の手法でアダプタを学習させることである．

2　既存研究

本節では CLIP [1] に加えて，本実験で用いた平野の手法について述べる．

2.1　CLIP

CLIP(Contrastive Language-Image Pre-training) [1] とはインターネット上のテキストと画像の組のデータをもとに，テキストとそれに対応する画像を用いて対比学習を行う手法である．バッチサイズ N の画像とテキストのペアから，画像を Image Encoder で，テキストを Image Encoder でそれぞれ特徴量 $I_{\mathrm{i}}, T_{\mathrm{i}}$ に変換し，コサイン類似度を計算する．推論時には複数の候補クラスラベルをプロンプトテンプレートを用いて文に変換し，Text Encoder を通して特徴量 T_{i} に変換する．その後，画像の特徴量 I_{i} と T_{i} とのコサイン類似度を計算し，最大となったクラスを推論結果とする．

2.2　平野の手法

全体の概要は図 2, 図 3 の通りである．この手法では，図 2 に示すように，画像エンコーダとテキストエンコーダに事前学習済みの CLIP を用い，テキスト特徴量から画像特徴量への変換に BLIP-2 の Q-Former を用いる．まず，Q-Former の学習について説明する．学習済みの画像エンコーダを f_{image}，テキストエンコーダを f_{text} として，バッチサイズ N の画像 x_{i} とテキスト t_{i} のペアを各エンコーダに入力して，画像特徴量 I_{i} とテキスト特徴量 T_{i} を得る．すなわち，

$$I_i = f_{image}(x_i) \tag{1}$$

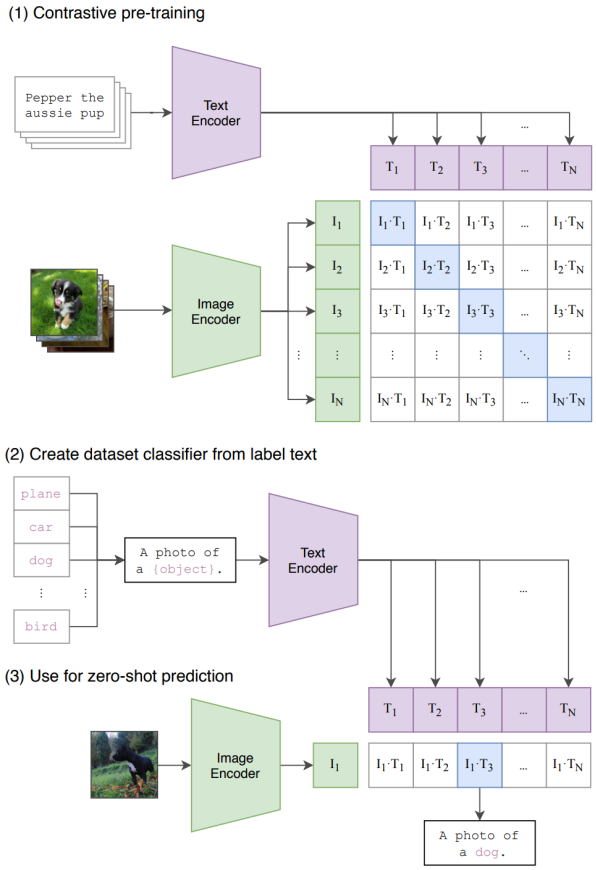

図 1: CLIP

$$T_i = f_{text}(t_i) \tag{2}$$

である．そして得られたテキスト特徴量 T_{i} を Q-Former f_{Q} へ入力して，出力 I'_{i} を次の式で得る．

$$I'_i = f_Q(T_i) \tag{3}$$

Q-Former からの出力 I'_{i} と，画像エンコーダから得られた画像特徴量 I_{i} を用いてコサイン類似度誤差 L_{CS} を計算し，以下のように損失を計算する．

$$L_{CS}(I_i, I'_j) = \begin{cases} 1 - \cos(I_i, I'_j) & (i \neq j) \\ \max(0, \cos(I_i, I'_j)) & (i = j) \end{cases} \tag{4}$$

$$L(I_i, I'_j) = \frac{1}{N^2} \sum_{i,j} L_{CS}(I_i, I'_j) \tag{5}$$

なお，学習の過程では画像エンコーダとテキストエンコーダのネットワーク重みは固定する．Q-Former の学習後，ファインチューニング用データセットを用いてモデルのヘッドをファインチューニングする．以下にファインチューニングの流れを説明する．バッチサイズ N のテキスト t_{i} とラベル y_{i} に対して，t_{i} をテキストエンコーダ f_{text} と Q-Former f_{Q} を通して，画像特徴量 I'_{i} へと変換後，モデルヘッド f_{head} へと入力することでモデルの推論結果を得る．すなわち，

$$T_i = f_{text}(t_i) \tag{6}$$

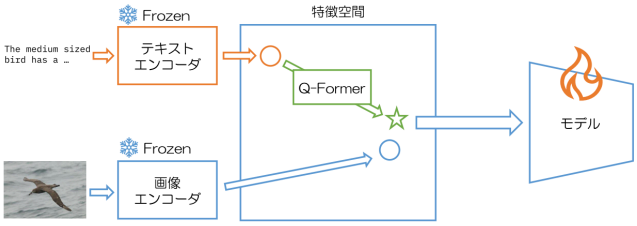

図 2: Q-Former の学習

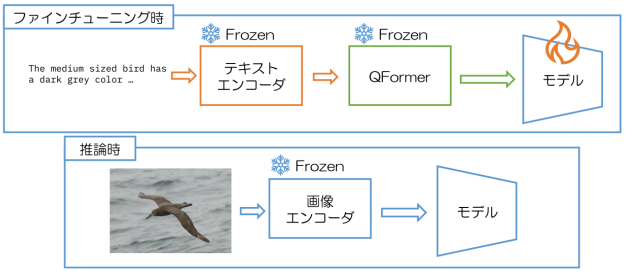

図 3: アダプタの学習

$$I'_i = f_Q(T_i) \tag{7}$$

$$output_i = f_{head}(I'_i) \tag{8}$$

である．得られた結果 $output_{\mathrm{i}}$ とラベル y_{i} を用いて，交差エントロピー誤差で損失を計算する．したがって交差エントロピー誤差を L_{c} とすると，

$$L = \frac{1}{N} \sum_i L_c(output_i, y_j) \tag{9}$$

となる．

3 提案手法

本研究では，LLM を用いて分類対象となるクラスに関するテキストを生成し，そのテキストを平野の手法でのアダプタの学習に用いること提案する．

4 実験

4.1 実験設定

本節では，Q-Former の学習およびアダプタの学習設定について述べる．Q-Former の学習には大規模なキャプション付きデータセットである COCO Captions [2] データセットを用いた．このデータセットでは，約 33 万枚の画像が提

供されており，また，人手によって作成されたキャプションが各画像に対して 5 つずつ提供されている．本研究では，2014 年度のサブセットを用いた．このサブセットには，約 11 万 3 千枚の画像と画像毎に 5 つのキャプションが含まれている．すなわち，約 56 万個の画像とキャプションのペアを用いて学習を行った．アダプタの学習には詳細画像分類データセットとして有名な，200 種類の鳥の画像からなる CUB データセットを用いた．このデータセットには，学習用画像として 5994 枚，テスト用画像として 5794 枚が提供されている．また，アダプタとして 1 層の全結合層を採用した．以上の設定，パラメータについては平野の実験と同じである．また，平野の実験では学習サンプルの選択において，全体から割合を採用していた．しかし，これでは学習サンプルにクラスの偏りが生じる可能性が存在する．よって本実験では，学習サンプルの選択において，クラス単位での割合に修正して実験を行った．後に述べるが，この修正によって平野の結果と本実験の結果は大きく異なった．

4.2 LLM によるテキスト生成

LLM として "gpt-3.5-turbo" API を用いてテキストを生成した．例えば，クラス名が "Black Footed Albatross" の場合，以下のプロンプトを LLM に渡した．

f"""
You are very knowledgeable about bird species, and when given a certain breed, you want to output sentences describing the only bird's visual characteristics such as size, shape, coloration and bill shape and size.
"""

4.3 実験結果

本実験で得られた結果を表 1 に示す．画像のみの列については，学習データセットのうち学習サンプルとして選択された画像のみをアダプタの学習に用いた場合の Top-1 Accuracy．画像＋テキストの Reed らの列については，平野の実験で用いられていた人の手によって作成されたテキストを学習に用いたもの．GPT の列については gpt-3.5-turbo で生成させたテキストを学習に用いたものである．また，比較対象として括弧内に平野の実験結果も示す．まず，テキストのみをアダプタの学習に用いた結果を見ると，人の手で作成されたテキストを用いた場合と，LLM で生成させたテキストを用いた場合とで精度はあまり変わらず，LLM で生成させたテキストが，人工的に作成した事実的に正しいテキストと同じような働きをすることが確認された．また，画像のみをアダプタの学習に用いた結果を見ると，平野の結果より大幅に精度が向上していることがわかる．実験設定としてはサンプルの選択方法のみが異なり，画像の割合が低くなるにつれて精度の差が大きくなることから，平野の実験ではアダプタの学習に用いた学習サンプルにクラスの偏りがあったため精度が低くなっていたと考えられる．また，それによって画像＋テキストの精度が全ての割合で画像のみの精度を下回ることになリ，アダプタの学習にテキストを用いると精度が下がるという結論になった．

表 1: 画像とテキストの割合を変えた実験結果

割合 (%)		画像＋テキスト	
(画像, テキスト)	画像のみ (平野)	Reed ら (平野)	LLM
(100, 0)	**85.07**(84.80)	-	-
(90, 10)	**85.26**(84.67)	83.87(85.05)	84.05
(80, 20)	**84.95**(83.55)	83.13(84.55)	82.90
(70, 30)	**84.71**(82.55)	81.57(84.16)	81.62
(60, 40)	**84.48**(80.48)	80.85(83.33)	80.01
(50, 50)	**84.23**(79.81)	77.64(82.52)	78.68
(40, 60)	**83.66**(77.82)	76.10(80.84)	76.56
(30, 70)	**83.38**(73.33)	72.05(67.50)	72.47
(20, 80)	**82.83**(67.24)	68.08(61.25)	67.24
(10, 90)	**81.88**(53.92)	57.47(49.12)	57.96
(0, 100)	-	**24.12**(27.67)	**23.78**

5 まとめ

本実験では，平野の手法において，アダプタの学習に人の手で作成されたテキストの代わりに LLM で生成させたテキストを用いた実験を行った．また，学習サンプルの選択を修正しクラスの偏りをなくすことで，少量の画像でアダプタの学習を行った場合の精度が向上し，学習にテキストを用いると精度が下がることが確認された．

参考文献

[1] Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Alec Radford, Jong Wook Kim and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions:

Data collection and evaluation server. *ArXiv*, Vol. abs/1504.00325, , 2015.

[3] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

[5] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 49–58, 2016.

[6] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[7] 平野甫. テキスト特徴量から画像特徴量への変換による詳細画像分類. 修士論文, 大阪公立大学大学院工学研究科電気・情報系専攻知能情報工学分野, 2024.