大規模言語モデルを用いた Vison-Language Model による Few-shot 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 Contrastive Language-Image Pre-training(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．

このような基盤モデルは Few-shot 分類や教師あり学習の文脈から活用したいという需要がある．Few-shot 分類とは，少数のラベル付きデータのみを用いて分類を行うタスクである．一般に，別のデータセットで学習済みのモデルのパラメータを初期値として，ファインチューニングをする手法が取られる．しかし，よく使われるものでも 600MB，大きなもので 10GB と膨大なサイズであるため，元のモデルの一部を再学習するファインチューニングは効率が良くない．また，膨大な事前訓練から細かなコンテクストを学習しているため、ファインチューニングをしてしまうと、破滅的忘却により学習した細かなコンテクストを忘れてしまうという問題がある．

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．CLIP のアダプタを差し込んだモデルとして Linear-Probe [1] という手法が CLIP と同じ論文で提案されている．Linear-Probe では CLIP の Image-Encoder の末端にロジスティック回帰を用いて学習している．しかし，ここでは Image-Encoder とともに対照学習された Text-Encoder は用いられていない．

そこで，テキスト特徴量も用いてアダプタを学習する手法に平野 [6] 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 [2] 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 Few-shot 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 [4] 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] と Linear-Probe [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 [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 Linear-Probe

Linear-Probe [1] では CLIP の事前学習済みエンコーダを固定し，エンコーダが出力するベクトルに対して，単純な線形分類器を訓練する．損失関数に交差エントロピーが使われる．

2.3 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

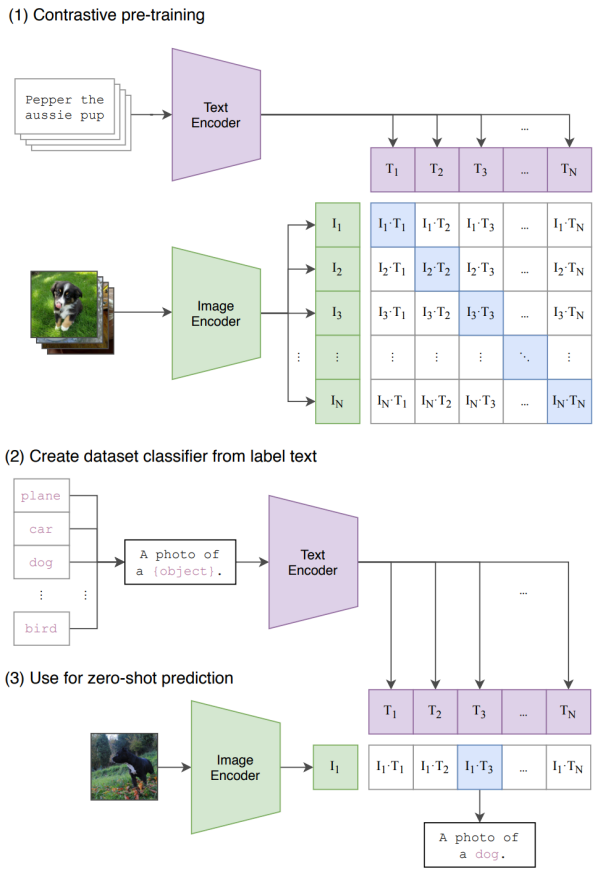

図 1: CLIP [1]

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，

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

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

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．

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

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

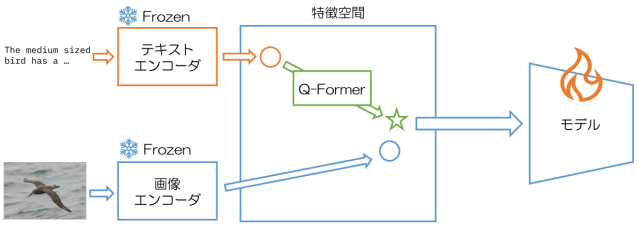

図 2: Q-Former の学習

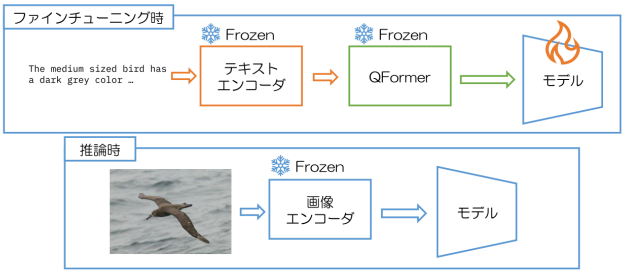

図 3: アダプタの学習

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，

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

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

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

である．得られた結果 $output_{\text{i}}$ 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，画像データセットと合わせた画像テキストデータセットを構築する．そして学習された Q-Former を用いることで，LLM で生成させたテキストを画像と同様に学習することが可能となり，アダプタに対し Few-shot 学

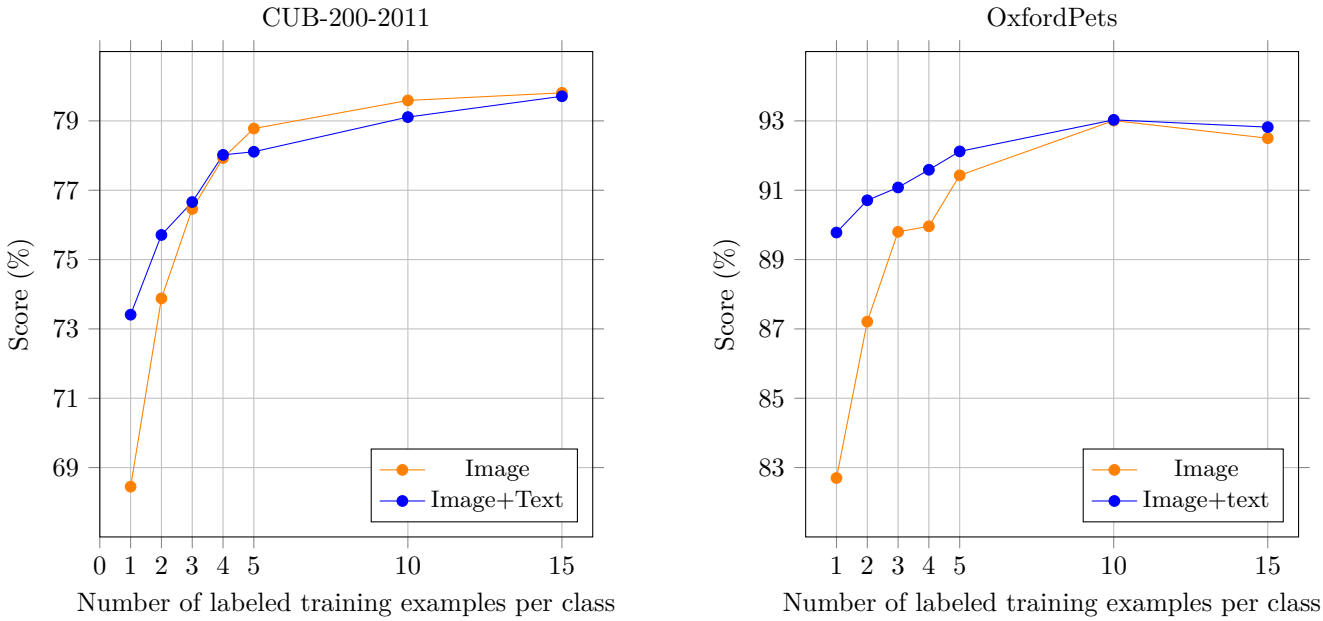

図 4: CUB と OxfordPets における画像のみ学習した場合とテキストを加えて学習した場合の比較

表 1: CUB と OxfordPets における Few-shot 分類の精度

学習設定		1-shot	2-shot	3-shot	4-shot	5-shot	10-shot	15-shot
CUB	画像	68.45	73.88	76.46	77.93	**78.78**	**79.59**	**79.81**
	画像 + テキスト (数)	**73.41**(5)	**75.71**(2)	**76.66**(4)	**78.02**(3)	78.11(1)	79.11(1)	79.71(1)
OxfordPets	画像	82.7	87.21	89.8	89.96	91.43	93.01	92.5
	画像 + テキスト (数)	**89.78**(20)	**90.71**(3)	**91.08**(30)	**91.59**(5)	**92.12**(10)	**93.03**(1)	**92.82**(1)

習をすることを提案する．

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．本実験では，データセットとして 200 種類の鳥の画像からなる CUB-200-2011 [5] と 100 種類の動物の画像からなる OxfordPets [3] を用いた．また，アダプタの学習については 1 クラス当たりに学習する画像のサンプル数を 1,2,3,4,5,10,15 として Few-shot 学習を行い，推論時には学習されたモデルで全てのテスト用画像を用いて評価した．また，アダプタとして 1 層の全結合層を採用した．アダプタとハイパーパラメータについては平野の実験と同じである．

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"""
Describe the visual characteristics of a Black Footed Albatross in 10 lines.
"""
```

4.3 実験結果

本実験で得られた結果を図 4 に示す．CUB においては，1 クラス当たりに学習する画像のサンプル数が 1 枚から 4 枚の場合に，画像に加えてテキストを学習に用いた精度が画像のみを学習に用いた精度を上回った．OxfordPets においては，全ての Few-shot 設定でテキストを追加で学習した場合の精度が画像のみを学習した精度を上回った．また，表 1 では，画像に加えテキストを学習した場合の精度に，括弧書きで 1 クラス当たりに学習したテキストのサンプル数を示した．shot 数によってテキストのサンプル数が異なっているのは，同じ shot 数の条件下でテキストのサンプル数を変えて実験したものの中で最も精度の良いサンプル数における精度を示しているためである．ここで，関連研究にもあ

る CLIP との比較をする．CLIP の Zero ー shot 分類の精度は CUB で 52.9%，OxfordPets で 97.0% であった．提案手法と比べると，OxfordPets では全ての設定で Zero-shot を下回る制度となっているが，CUB では全ての設定で上回っている．OxfordPets が比較的容易な分類データセットである一方，CUB は詳細画像分類タスクで用いられるデータセットとして知られる．そのようなデータセットにおいては Zero-shot では精度が出ないため，今回のような追加学習する手法が有効であり，且つテキストを追加で学習することが有効であることがわかった．

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．Few-shot 分類では，CUB と OxfordPets の両方のデータセットで画像に加えテキストをアダプタの学習に用いることが有効であ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] 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.

[3] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

[4] 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.

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

[6] 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.