LLM説明文とQ-Formerの活用による CLIP Few-shot Adaptationの精度向上

川越 壮[1,a)]　平野 甫[1]　岩村 雅一[1,b)]

概要：CLIP は，画像とテキストの特徴量を統合的に学習した Visio Language Model であり，Zero-shot 分類において高い汎用性を示す．一方で，詳細分類レベルの認識や専門ドメインでは，事前学習データの偏りにより性能が低下しやすい．本研究では，こうした少数データ環境下における CLIP の Few-shot Adaptation を目的とし，Large Language Model (LLM) によって生成された視覚的説明文を活用する手法を提案する．具体的には，各クラスに対して LLM から説明文を生成し，BLIP-2 で導入された Q-Former を用いてテキスト特徴量を画像特徴量空間にマッピングすることで，CLIP が持つモダリティギャップを解消する．さらに，変換後の特徴量を用いて Adapter を学習することで，視覚と言語の統合的な Few-sho Adaptation を実現する．CUB-200-2011 データセットを用いた実験では，既存手法である Linear probe と比較して，特に 1-shot および 2-shot 設定において顕著な性能向上を確認した．

1. はじめに

Vision Language Model（VLM）は，画像処理と自然言語処理を融合させたモデル群であり，近年急速に発展してきた．特に，Contrastive Language-Image Pre-training (CLIP) [1] や ALIGN [2]，BLIP [3] などは，視覚と言語の両モダリティの特徴量を統合し，多様なタスクで高い性能を発揮する．CLIP は，大規模な画像とテキストのペアデータを用いて対照学習を行うことで，学習時に存在しない新規クラスに対しても Zero-shot 推論が可能となり，ラベル付けの手間を省きつつ高い汎用性を実現する．

しかしながら，CLIP の性能は事前学習データの分布に強く依存することが知られており，専門領域や本研究で焦点を当てる詳細分類レベルの認識においては，学習データの不足により性能が低下することが報告されている [4]．例えば，航空機の機体識別を対象とした FGVC Aircraft [5] や，道路環境を扱う KITTI [6] などでは，CLIP の Zero-shot 分類精度は 40%を下回る．この原因は，事前学習データセットにおいて，こうした特定分野のデータが十分に網羅されていないためである．

このような場面では，CLIP を下流タスクへ適応させる追加学習が必要となる．一般に，このような適応には Fine-tuning と Few-shot Adaptation の 2 つのアプローチがある．Fine-tuning では，全てのモデルパラメータを更新して新たなデータセットに適応させるが，大量のラベル付きデータと計算コストが必要になるため，リソース制約のある環境では適用が難しい．一方で，Few-shot Adaptation は，少量のデータで効果的にモデルを適応させる方法として注目される．本研究では特に，後者のアプローチに焦点を当てる．既存の Few-shot Adaptation 手法としては，CLIP の Image-Encoder の出力に線形分類器を付加する Linear probe [1] や，特徴量レベルで調整を行う CLIP-Adapter [7]，プロンプト最適化によりテキスト側を調整する CoOp [8] などが提案されている．これらは計算資源効率の面で優れるが，主に画像特徴量やプロンプトに焦点を当てており，CLIP のもう一方のモダリティであるテキスト特徴量の積極的活用は限定的である．

CLIP のテキスト特徴量を利用してモデルを学習する手法として CLOSE [9] がある．CLOSE は，CLIP の Image-Encoder と Text-Encoder を固定し，出力されたテキスト特徴量にガウシアンノイズを加えることで画像特徴量に近づけ，画像の代わりにテキストのみでモデルを学習できるクロスモーダル転送を可能とした．CLOSE は，テキストのみで学習したモデルが画像とテキストの両方で学習させたモデルと遜色ない性能を，キャプション生成，Visual Question Answering (VQA) [10]，Visual Entailment [11]，Visual News Captioning [12] といった複数の Vision & Language (V&L) タスクで達成できたことを報告した．しか

[1] 大阪公立大学 大学院情報学研究科
Graduate School of Informatics, Osaka Metropolitan University, 1-1, Gakuencho, Naka, Sakai, Osaka 599-8531, Japan
[a)] sp25249p@st.omu.ac.jp
[b)] masa.i@omu.ac.jp

し，テキスト特徴量にノイズを加えるため，テキスト特徴量の持つ繊細な情報を破壊してしまい，詳細分類レベルの認識では実用性がない．

本研究は，これらの先行研究の限界を踏まえ，特に CLIP の苦手とする詳細分類レベルの認識に対し，説明文から得られるテキスト特徴量を積極的に活用した Few-shot Adaptation に焦点を当てる．具体的には，LLM で各クラスの視覚的説明文を生成し，BLIP-2 [13] において導入された Q-Former を用いて説明文由来のテキスト特徴量から画像特徴量への変換を図る．Q-Former は CLOSE と異なり，視覚言語表現を学習し，より効果的なクロスモーダル転送が可能である．これにより詳細な情報を保持したまま，テキストによる学習が可能となり，少数データ環境下において，詳細画像分類タスクで画像のみを学習に用いる既存手法より高い性能を実現した．

2. 関連研究

2.1 Vision Language Model

VLM は，画像処理と自然言語処理を融合させたモデルであり，特に CLIP [1] や ALIGN [2] は大規模な画像とテキストの組で事前学習することで，一般的な視覚表現を学習し，テキストを介して様々な下流タスクでの Zero-shot 分類を可能とした．

CLIP のアーキテクチャを図 1 に示す．CLIP は物体認識で通常用いられるような画像とラベルのペアからなるデータセットではなく，画像とその説明文からなるデータセットを用いて学習する．これにより，学習時に見たことがないクラスの画像であっても，正しく分類できる Zero-shot 推論を可能としたモデルである．

CLIP では，バッチサイズ N の画像 $\{\boldsymbol{x}_i\}_0^N$ とテキスト $\{\boldsymbol{t}_i\}_0^N$ のペアについて，画像は Image-Encoder $\mathcal{F}_{\text{image}}$ で，テキストは Text-Encoder $\mathcal{F}_{\text{text}}$ でそれぞれを同一の次元を持つ特徴量に変換する．すなわち，

$$\boldsymbol{I}_i = \mathcal{F}_{\text{image}}(\boldsymbol{x}_i) \tag{1}$$

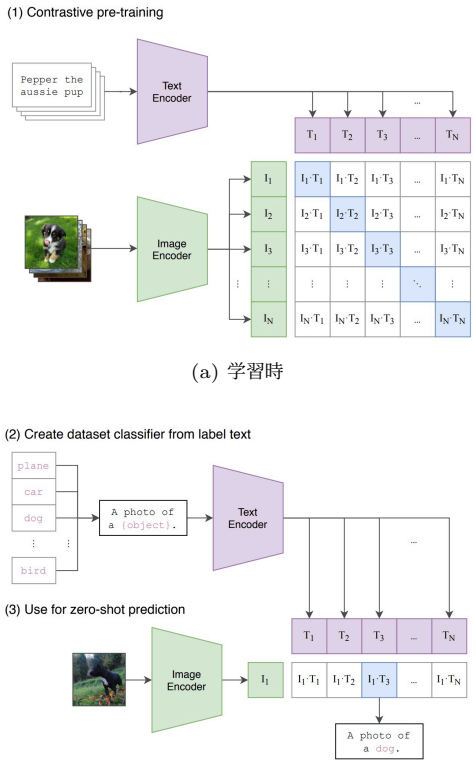
$$\boldsymbol{T}_i = \mathcal{F}_{\text{text}}(\boldsymbol{t}_i) \tag{2}$$

である．その後，$\{\boldsymbol{I}_i\}_0^N$，$\{\boldsymbol{T}_i\}_0^N$ を正規化後，図 1(a) に示すように，全ての組み合わせでコサイン類似度を計算する．すなわち，正規化した $\{\boldsymbol{I}_i\}_0^N$，$\{\boldsymbol{T}_i\}_0^N$ をそれぞれ $\{\boldsymbol{I}'_i\}_0^N$，$\{\boldsymbol{T}'_i\}_0^N$ とすると，

$$\boldsymbol{I}'_i = \frac{I_i}{|I_i|} \tag{3}$$

$$\boldsymbol{T}'_i = \frac{T_i}{|T_i|} \tag{4}$$

となる．なお，$\{\boldsymbol{I}'_i\}_0^N$，$\{\boldsymbol{T}'_i\}_0^N$ は正規化されているため，$\boldsymbol{I}'$，$\boldsymbol{T}'^T$ の行列積を計算すれば次式のようにコサイン類似度 S を計算することができる．



(a) 学習時

(b) 推論時

図 1: CLIP の学習と推論 [1]

$$\boldsymbol{S} = \boldsymbol{I}'\boldsymbol{T}'^T \tag{5}$$

得られた $N \times N$ 行列 $\boldsymbol{S}$ に対して，行方向，列方向両方の交差エントロピー誤差を計算する．なお，交差エントロピー誤差を計算する際のラベルは，図 1(a) の水色部分のインデックスが用いられる．すなわち，$\boldsymbol{S}$ の対角成分が大きくなるように学習する．これにより，画像と対応するテキストからそれぞれの特徴を抽出すると，これらの特徴は特徴空間上で相対的に近くに位置するようになる．推論時には，図 1(b) に示すように，入力画像と “a photo of a <class label>“の類似度を計算し，最も類似度の高いクラスを推論結果とする．なお <class label> にはカテゴリ名が入る．しかし，CLIP の性能は事前学習データセットの分布に依存し，特定の専門領域やニッチなクラスにおいては十分な性能が得られないことが知られている．このような場合，CLIP を下流タスクに適応させる手法が必要となる．

2.2 Few-shot Adaptation

少量のデータを用いて CLIP を下流タスクに適合させる手法はいくつか存在する．本節では今回の提案手法の元となった Linear probe に加え，テキストプロンプトの最適化を行う Context Optimization (CoOp) [8] と特徴量レベルでの適合を行う CLIP-Adapter [7] について述べる．

CLIP の Linear probe [1] は，事前学習済みの画像エンコーダの特徴量を固定し，その上に線形分類器を付加して下流タスクを解くアプローチである．本手法は，Image-

Encoder の出力特徴量を固定し，タスクごとの分類器のみを学習することで計算効率を向上させることを目的とする．この手法はシンプルかつ計算効率が高い一方，画像特徴量のみに依存しているため，CLIP のテキスト特徴量の活用は行われない．

CLIP が推論する際にはテキストプロンプトとして固定の形式（例: A photo of a {class}）を用いる．それに対し CoOp [8] では，このプロンプトに学習可能なコンテキストトークンを挿入する．コンテキストトークンは、少量のデータを使って最適化され，下流タスク固有のテキストプロンプトが構築される．CLIP の Image-Encoder と Text-Encoder のパラメータは固定され，コンテキストトークンのみを学習することで，少量のデータで下流タスクに適合させることが可能である．

CLIP-Adapter [7] は，CLIP の Image-Encode と Text-Encoder それぞれの出力に対して，追加の特徴 Adapter 層を挿入する形式を取る．この Adapter は，ボトルネック構造を持ち，通常 2 層の Multi-Layer Perceptron (MLP) で構成され，下流タスク固有の新しい特徴を学習する．また，Adapter の出力は元の CLIP の特徴量に残差接続を通じて統合される．この残差の特徴統合により，事前学習済み知識の破壊を防ぎつつ，タスク適合が可能となる．尚，ここでも CLIP の Image-Encoder と Text-Encoder のパラメータは固定される．

これらの手法は，画像特徴量の利用やプロンプト最適化に焦点を当て，リソース効率の観点で優れている一方で，CLIP のもう一方のモダリティであるテキスト特徴量を積極的に活用した手法は限定的である．

2.3 モダリティギャップの解消

CLIP のような VLM は画像とテキストを対照学習することで同一の特徴空間にエンコードすることができるが，実際には両者のベクトルは大きく離れている可能性がある．これはモダリティギャップ [14] と呼ばれる現象である．例えば，COCO Captions [15] では，画像とテキストのペアの特徴量の平均コサイン類似度はわずか 0.26 だが，無関係な 2 つの特徴量の平均コサイン類似度は 0.35 であることが報告されている [14]．この原因として，対照モデルでは画像とテキストの正例に対しては近く，負例に対しては遠くなるように相対的に学習されるため，正例が絶対的に近くなるというわけではないことが挙げられる．このギャップを解消し，テキスト特徴量を有効活用するためには，画像特徴量とテキスト特徴量の絶対的な距離を縮める必要がある．本節ではテキストから画像への特徴量変換を行う CLOSE [9] と画像からテキストへの特徴量変換を行う BLIP-2 [13] について述べる．

Cross ModaL transfer On Semantic Embeddings (CLOSE) [9] は，図 2 に示すように，テキスト特徴量にガウシアンノイズを加えることで，モダリティギャップを解消する．ノイズを加えたテキスト特徴量を下流の LLM に流すことで，画像を用いずに VQA や Image Captioning，Visual Entailment などの，画像または画像とテキストを入力として受け取り，答えを言語で出力する Vision & Language (V&L) タスクを学習する．図 3 は画像とテキストを同一の特徴空間に写像したときの分布を示した図である．テキスト特徴量にガウシアンノイズを加えることで，加えなかった場合に比べて画像特徴量とテキスト特徴量の分布が近くなり，ギャップが解消されていることがわかる．しかし，テキスト特徴量にガウシアンノイズを加えるため，テキスト特徴量の持つ繊細な情報を破壊してしまうという問題がある．そのため，本研究で扱う CLIP の苦手な詳細画像分類タスクに CLOSE は適切ではない．

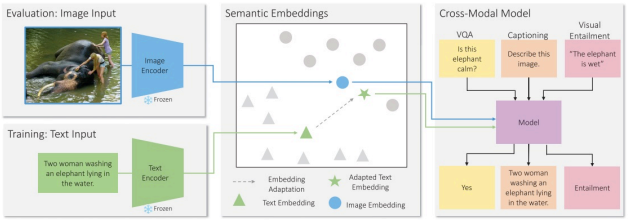
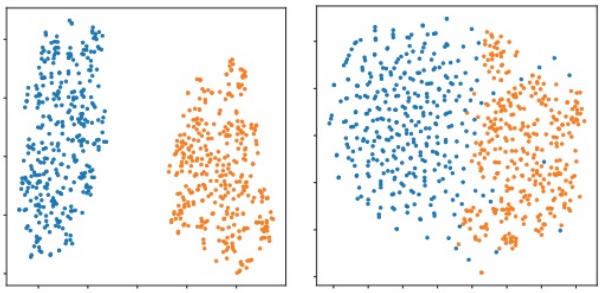


図 2: CLOSE の構造 [9]



(a) ノイズを加えない場合　　(b) ノイズを加えた場合

図 3: ノイズ付加によるモダリティギャップの解消 [9]

BLIP-2 は図 4 に示すように，事前学習済みの Image-Encoder と LLM を固定し，その間を画像特徴量からテキスト特徴量へと変換する Querying Transformer (Q-Former) を用いて繋ぐことで，画像処理と自然言語処理を跨ぐ V&L タスクを学習するモデルである．Q-Former は Transformer ベースのアーキテクチャであり視覚から言語への橋渡しとして用いる．また，Image-Encoder と LLM を事前学習された状態で固定して学習することで，BLIP や Flamingo [16] などの既存モデルと比較して学習パラメータを大幅に削減できる．図 5 に示すように，Q-Former は画像，画像に対応するテキスト，学習可能なクエリの 3 つを入力とし，画像特徴量をクロスアテンションを通じて処理し，テキスト特徴量に変換する役割を担う．この仕組みにより，事前学習済みの Image-Encoder と LLM を変更せずに連携させることが可能となる．特に，Q-Former のクエリは自己アテ

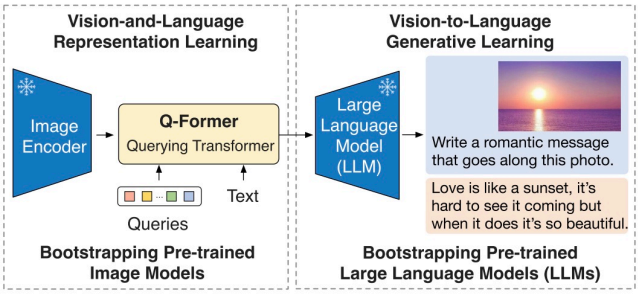


図 4: BLIP-2 の構造 [13]

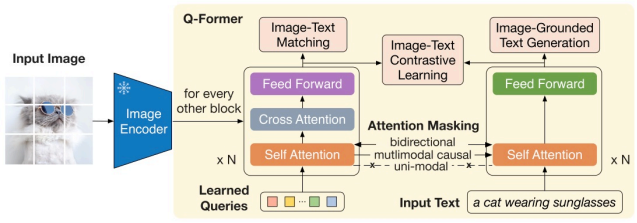


図 5: Q-Former の構造 [13]

ンション機構を持ち，画像特徴量と相互作用しながら最適な表現を学習する．これにより，視覚情報が効果的に言語空間へとマッピングされる．本研究では，この Q-Former を適用し，説明文のテキスト特徴量を画像特徴量に変換するアプローチを採用する．

2.4 説明文を用いた Adaptation

近年，CLIP の Zero-shot 性能改善のために，LLM によって生成された説明文を活用する研究が進められている．Mayug ら [17] は，GPT-4 を用いてデータセットごとに詳細な視覚的説明文を生成し，それを CLIP の Text-Encoder に入力することで，Zero-shot および Few-shot 分類性能の向上を達成した．特に Few-shot 設定では，複数の説明文から分類に効果的なテキスト特徴量を選択するためのアダプターを導入し，既存手法（CoOp）を上回る精度を報告している．ただし，彼らの手法は Base-to-New 設定に基づき，事前に定義された Base クラスの十分なデータを用いた後，New クラスに対する適応性能を評価する形式である．

Bansal ら [18] は，LLM で生成した説明文と画像の組み合わせにおいて，各ペアをあえて対応づけせずにロバストな学習を行う手法を提案した．これにより，iNaturalist [19] 等のデータセットで Zero-shot 性能の向上を実現したが，その目的は主に Zero-shot 性能の向上であり，Few-shot 環境下で説明文を用いることで CLIP の適応性能を向上させるという本研究の目的とは方向性が異なる．

3. 提案手法

本研究では，CLIP の Few-shot Adaptation において，画像データが限られた環境でも効果的な適応を実現するため，LLM によって生成した各クラスの視覚的説明文を活用する．図 6 に示すように，事前学習済みの CLIP を用いて画像とテキストの特徴量を取得し，テキスト特徴量を画像特徴量に近づけるために Q-Former を利用する．その上で，少数の画像および説明文の双方を用いて Adapter を学習する．ここで，BLIP-2 では画像特徴量をテキスト特徴量に変換する際に用いられていた Q-Former を，本研究では逆の変換で用いた．これは，Q-Former が Transformer ベースのアーキテクチャであり，今回学習に用いたテキストは LLM の Temprature を 1.0 に設定して生成したものであるため，より多様なテキスト特徴量から画像特徴量にマッチするような特徴量を学習することが期待できると考えたためである．

3.1 LLM による説明文生成

テキスト特徴量が画像特徴量に近づくためには，説明文が画像の視覚的特徴を的確に表現している必要がある．そこで，本研究では，各クラスの視覚的特徴を詳細に記述する説明文を LLM により生成する．

具体的には，以下のプロンプトを用いて説明文を生成した．

> You are a helpful assistant. What characteristics can be used to differentiate a category from other birds based on just a photo? Produce an exhaustive list of attributes and information that can be used to identify the bird uniquely. Texts should be of the form “[`class`] with <characteristic feature>”. Describe in 10 lines and, one sentence per line.

[class] の部分は下流データセットに含まれるクラス名を入れる．このプロンプトは，各クラス（鳥の種）ごとに，視覚的識別に有効な特徴を明確に抽出することを意図して設計されている．各クラスにつき画像 1 枚あたり 10 個の説明文を生成し，全体としてクラスごとに十分なテキストデータを確保した．

3.2 Q-Former による特徴量変換

説明文のテキスト特徴量を画像特徴量に変換するため，BLIP-2 で用いられている Q-Former を導入する．Q-Former は Transformer ベースのアーキテクチャであり，学習可能なクエリを通じて視覚特徴と相互作用しながら，適切な特徴量変換を実現する．

本研究では，事前学習済みの CLIP を用いて画像とテキストの特徴量を取得し，Q-Former によりテキスト特徴量を画像特徴量にマッピングする．具体的には，以下の手順で特徴量を得る．学習済みの Image-Encoder を $\mathcal{F}_{\text{image}}$，Text-Encoder を $\mathcal{F}_{\text{text}}$ として，バッチサイズ N の画像 $\{\boldsymbol{x}_i\}_0^N$ とテキスト $\{\boldsymbol{t}_i\}_0^N$ のペアを各エンコーダに入力して，式 (1), (2) から計算される画像特徴量 $\{\boldsymbol{I}_i\}_0^N$ とテキス

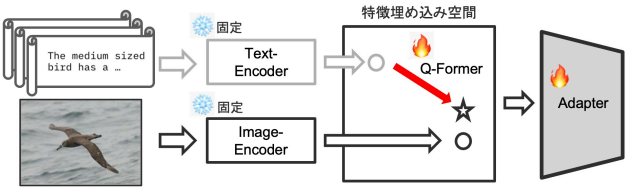


図 6: 提案手法の概要．カテゴリの視覚的説明文をテキストとして用いる．

ト特徴量 $\{\boldsymbol{T}_i\}_0^N$ を得る．そして得られたテキスト特徴量 T_i を Q-Former $\mathcal{F}_{\mathrm{Q}}$ へ入力して，出力 $\{\boldsymbol{I}'_i\}_0^N$ を次式で得る．

$$\boldsymbol{I}'_i = \mathcal{F}_{\mathrm{Q}}(\boldsymbol{T}_i) \tag{6}$$

Q-Former からの出力 I'_i と，Image-Encoder 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．

$$\mathcal{L}_{CS}(\boldsymbol{I}_i, \boldsymbol{I}'_j) = \begin{cases} 1 - \cos(\boldsymbol{I}_i, \boldsymbol{I}'_j) & (i \neq j) \\ \max(0, \cos(\boldsymbol{I}_i, \boldsymbol{I}'_j)) & (i = j) \end{cases} \tag{7}$$

$$\mathcal{L}(\boldsymbol{I}_i, \boldsymbol{I}'_j) = \frac{1}{N^2} \sum_{i,j} L_{CS}(\boldsymbol{I}_i, \boldsymbol{I}'_j) \tag{8}$$

コサイン類似度を損失とすることで，CLIP でモダリティギャップが生じる原因であった「ペアとなる画像とテキストの特徴量がその他のテキストや画像の特徴量よりも相対的に近くなればよい」という学習方法ではなく，画像特徴量とテキスト特徴量の距離を最小化する学習方法となる．なお，学習の過程では Image-Encoder と Text-Encoder のネットワーク重みは固定する．

3.3 Adapter の学習

Q-Former を用いてテキスト特徴量を画像特徴量に変換した後，少数のデータを用いて CLIP の下流タスク適応を行うために，Adapter を学習する．バッチサイズ N のテキスト $\{\boldsymbol{t}_i\}_0^N$ とラベル $\{y_i\}_0^N$ に対して，$\{\boldsymbol{t}_i\}_0^N$ を Text-Encoder $\mathcal{F}_{\mathrm{text}}$ と Q-Former $\mathcal{F}_{\mathrm{Q}}$ を通して，式 (2), (6) から計算される，画像特徴量 $\{\boldsymbol{I}'_i\}_0^N$ へと変換後，Adapter $\mathcal{F}_{\mathrm{adapter}}$ へと入力することでモデルの推論結果を得る．すなわち，

$$\hat{y}_i = \mathcal{F}_{\mathrm{adapter}}(\boldsymbol{I}'_i) \tag{9}$$

である．得られた推論結果 $\{\hat{\boldsymbol{y}}_i\}_0^N$ とラベル $\{y_i\}_0^N$ を用いて，以下のように交差エントロピー誤差 $\mathcal{L}_{\mathrm{adapter}}$ を計算する．

$$\mathcal{L}_{\mathrm{adapter}} = \frac{1}{N} \sum_i y_i \log \hat{y}_i \tag{10}$$

また，画像を用いて Adapter を学習する際は，Image-Encoder $\mathcal{F}_{\mathrm{image}}$ からの出力をそのまま Adapter に入力することで実現する．

4. 実験

本節では，まず実験設定の詳細について述べる．次に Few-shot Adaptation において LLM で生成させたテキストを用いた実験の結果を示す．

4.1 実験設定

データセットには Few-shot Adaptation の性能評価のため，CUB-200-2011 [20] データセットを使用した．CUB-200-2011 は，200 種類の鳥類クラスを含み，各クラスあたり最大 60 枚の画像が含まれている．詳細な画像分類が要求されるデータセットであり，CLIP の Zero-shot 性能が限定的であることが知られる．

本研究では，各クラスごとに k-shot ($k \in 1, 2, 4, 8, 16$) Few-shot 設定 [21] に従い，各クラスからランダムに k 枚の画像を学習データセットから選択し，テストにはテスト用データセットを全て使用した．説明文生成では各クラスに対し，gpt-3.5-turbo [22] および gpt-4o-mini [23] を用いて説明文を生成した．説明文生成の詳細な方法およびプロンプト設計については 3.1 節に記載した．各クラスにつき 10 種類の説明文を生成し，学習で使用した．

モデルは CLIP の ViT-B/16 モデルを基盤とし，Image-Encoder および Text-Encoder のパラメータは全て固定した．モダリティギャップの解消のため，BLIP-2 で導入された Q-Former を利用し，説明文のテキスト特徴量を画像特徴量に変換する．Q-Former の Fine-tuning では，COCO Captions データセット [15] で学習済みモデルのパラメータで初期化し，CUB データセットの少量の画像とテキストの組で Fine-tuning した．AdamW [24] オプティマイザを用い，学習率は 1e-4，バッチサイズは 1024，エポック数は 100，重み減衰は 5e-2 とした．学習率スケジューラには Cosine Annealing を採用し，損失関数としては 2.2 節で定義したコサイン類似度損失を用いた．Adapter の学習は，Q-Former の出力特徴量を入力とし，1 層の全結合層からなるシンプルな構造とした．学習には AdamW オプティマイザを用い，学習率は 5e-3，バッチサイズは 1024，エポック数は 30，重み減衰は 0.01 とした．損失関数にはクロスエントロピー誤差を使用し，画像特徴量および変換後のテキスト特徴量を入力とする．

評価指標は Top-1 Accuracy を用い，Few-shot 設定において画像のサンプリングを 1 回，説明文のサンプリングを 3 回実施し，それぞれの実験結果の平均値を求めた．すべての実験は，NVIDIA GeForce RTX 3090 GPU（24GB メモリ）を 8 枚使用した分散学習環境で実施した．

4.2 ベースラインモデル

比較手法は Linear probe [1] とした．CLIP の Image-Encoder の出力特徴量を固定し，その上に線形分類器を付

表 1: 画像特徴量とテキスト特徴量のユークリッド距離

LLM	1-shot	2-shot	4-shot	8-shot	16-shot
gpt-3.5-turbo	0.4588	0.4434	0.4369	0.4238	0.4160
gpt-4o-mini	0.4548	0.4398	0.4283	0.4152	0.4088

加して Few-shot 分類を行う．本手法とは，画像のみを線形分類器の学習に用いる点で異なる．

4.3 実験結果

本節では，まず LLM 生成テキストを用いた Q-Former の Fine-tuning の結果を示し，次に Adapter の Few-shot Adaptation の結果を示す．

4.3.1 Q-Former の学習結果

Q-Former の学習時にはロス関数としてコサイン類似度を用いたが，テスト時の評価指標としては，両特徴量のベクトルの方向に加え，大きさも比較することのできるユークリッド距離を用いた．表 1 に各 Few-shot 設定において Q-Former を学習させた結果を示す．shot 数が増えるにつれて画像特徴量とテキスト特徴量のユークリッド距離が小さくなっていることが確認できる．また，gpt-4o-mini で生成させたテキストを用いた方が gpt-3.5-turbo よりもユークリッド距離が小さくなっていることがわかる．

4.3.2 Adapter の学習結果

表 2 および表 3 に，各 k-shot ($k \in 1, 2, 4, 8, 16$) 設定における Top-1 Accuracy を示す．学習には k 枚の画像と LLM で生成した複数個のテキストを用いた．表 2 は gpt-3.5-turbo によって生成した説明文を用いた場合，表 3 は gpt-4o-mini を用いた場合の結果である．それぞれ，テキストの利用数を 1 から 10 まで変化させて実験を行った．テキストを使用しない場合（テキスト数 0）は，画像特徴量のみによる学習，すなわち Linear probe である．

1-shot および 2-shot の設定では，テキストを追加することで大幅に Accuracy が向上していることが確認できる．特に gpt-4o-mini による説明文は，1-shot で 49.35%，2-shot で 58.00%と，テキスト無しの 33.29%（1-shot），48.50%（2-shot）と比べて大幅な性能改善を達成している．また，テキスト数の増加に伴い精度が徐々に向上し，一定の数（おおむね 5〜6 文以降）で飽和する傾向が見られる．これは，説明文の追加がクラス識別に有効な特徴を補完し，特に少数データ環境下で顕著に効果を発揮することを示している．

gpt-3.5-turbo と gpt-4o-mini で生成した説明文の比較からは，gpt-4o-mini の方が全体的に高い精度を達成していることがわかる．例えば，4-shot 設定において，gpt-3.5-turbo では 62.91%，gpt-4o-mini では 67.94%と，約 5 ポイントの差が確認できる．これは，より高性能な LLM によって生成された説明文が，画像特徴とより良く整合し，適応性能を高めていることを示唆している．

5. 考察

表 2 および表 3 の結果から，説明文の数を増加させることで分類精度が向上する傾向が確認された．しかし，説明文の数が 5〜6 文を超えたあたりから性能改善が緩やかになり，ほぼ飽和することが見て取れる．これは，Few-shot 環境下において，ある程度の説明文が与えられることで少数の画像特徴だけでは表現しきれないクラス間の違いが十分補完され，それ以上の情報追加が性能に大きく寄与しないことを示唆している．特に 1-shot や 2-shot といった少数データ環境では，説明文の追加による補完効果が顕著であり，画像とテキストの相補的な利用が有効であることが示された．

LLM の違いが分類性能に与える影響については，gpt-4o-mini によって生成された説明文の方が，gpt-3.5-turbo よりも一貫して高い精度を示した．より高性能な LLM は，各クラスの視覚的特徴をより具体的かつ識別的に表現できると考えられ，その結果，少数データ環境においても画像特徴量を補完する効果が高まったと推察される．今後，さらに大規模な LLM や，データセット特化型の LLM を用いることで，説明文の質をさらに高め，分類性能の改善が期待できる．

本研究で実施した Q-Former を利用した明示的な特徴量変換の効果を可視化するため，Q-Former の変換前後におけるテキスト特徴量と画像特徴量の分布を t-SNE [25] により 2 次元に射影し，その関係性を図 7 に示す．ここで用いた Q-Former は 1 クラス 16 枚の画像とテキストの組で Fine-tuning したものである．同じクラスの特徴量は同じ色であり，画像特徴量は “×”，テキスト特徴量は “○” でプロットした．図 7(a) の Q-Former で変換前では，テキスト特徴量と画像特徴量は明確に異なるクラスタを形成し，モダリティギャップが存在している．一方，Q-Former で変換後は両特徴量の分布範囲が一致し，モダリティギャップが解消されていることが確認できる．特に，gpt-4o-mini で生成したテキストを Fine-tuning に用いた Q-Former で変換した図 7(c) では，両特徴量がクラスごとに接近した位置に配置されていることがわかる．このように，Q-Former を用いることで，説明文の情報が損なわれることなく画像特徴量と整合され，詳細な画像分類に効果的に寄与していると考えられる．

一方で，本手法にはいくつかの限界が存在する．まず，LLM による説明文生成には別途計算リソースとコストが必要であり，大規模データセットや実運用環境での適用時には課題となり得る．また，本研究では CUB-200-2011 を用いた実験を行ったが，他の詳細分類データセットやドメイン特化データセットへの適用性については検証が必要である．

さらに，説明文の品質評価や説明文の自動選別に関する

表 2: gpt-3.5-turbo で生成したテキストを用いた学習結果（Accuracy）
太字は各 Few-shot 設定において最も精度が高いものを示す.

画像数		テキストの数									
shot	0	1	2	3	4	5	6	7	8	9	10
1	33.29	41.29	43.86	46.01	45.70	46.11	46.29	**46.52**	46.41	45.97	46.39
2	48.50	51.01	53.95	54.87	55.46	55.51	55.16	**55.91**	55.74	55.28	55.31
4	60.31	60.97	62.16	62.79	**62.91**	62.76	62.89	62.72	62.49	62.27	61.94
8	69.63	68.56	68.84	68.96	**69.16**	69.11	68.73	68.70	68.58	68.61	68.45
16	74.18	74.49	**74.64**	74.50	74.35	73.89	74.18	74.34	74.21	74.28	74.10

表 3: gpt-4o-mini で生成したテキストを用いた学習結果（Accuracy）
太字は各 Few-shot 設定において最も精度が高いものを示す.

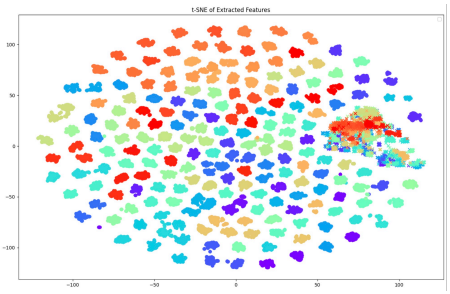
画像数		テキストの数									
shot	0	1	2	3	4	5	6	7	8	9	10
1	33.29	49.35	50.55	51.10	51.10	**51.18**	51.06	50.85	50.77	50.49	50.44
2	48.50	58.00	59.44	59.79	**60.22**	59.90	59.65	59.61	59.69	59.39	59.17
4	60.31	66.64	67.17	67.70	**67.94**	67.87	67.62	67.37	67.05	67.14	66.79
8	69.63	72.15	72.82	73.36	73.61	**73.82**	73.76	73.70	73.49	73.78	73.52
16	74.18	74.37	74.54	74.73	75.06	75.23	75.10	75.17	75.10	75.10	**75.18**

研究も今後の課題である．現在は LLM から生成されたすべての説明文からランダムにサンプリングして使用しているが，モデルの分類性能に特に寄与する説明文を選別する手法や，説明文自体を最適化するアプローチを導入することで，さらなる性能向上と効率化が期待できる．

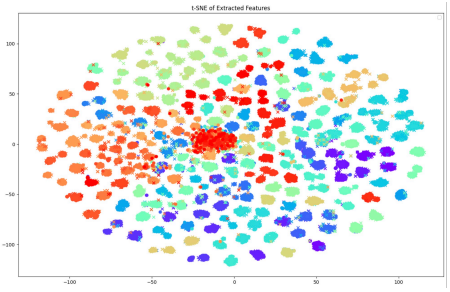
6. まとめ

本研究では，CLIP の Few-shot Adaptation において，LLM で生成した視覚的説明文を活用し，少数データ環境下での分類性能向上を図る手法を提案した．従来の Few-shot Adaptation 手法が画像特徴量やプロンプト最適化に依存していたのに対し，本手法では，LLM 生成の説明文のテキスト特徴量を活用し，さらに Q-Former を用いてモダリティギャップを解消することで，視覚と言語の両モダリティの特性を統合的に利用した．実験では，CUB-200-2011 データセットを用い，各 k-shot 設定において提案手法の有効性を検証した．その結果，従来の Linear probe と比較して，特に 1-shot や 2-shot といった極少数データ環境において顕著な性能向上が確認された．また，LLM の違いによる影響も分析し，高性能な LLM（gpt-4o-mini）によって生成された説明文が，より高い分類精度に寄与することを示した．さらに，説明文特徴量と画像特徴量のモダリティギャップを解消するために導入した Q-Former の有効性についても検証を行い，Fine-tuning 前後の特徴量分布を可視化することで，その効果を明示した．

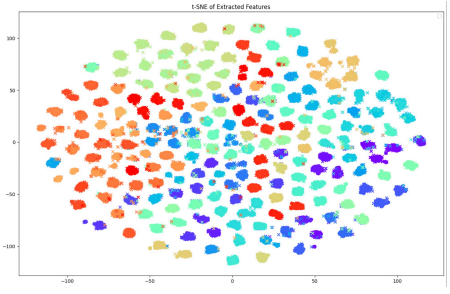
今後の課題としては，他の専門分野やドメイン固有のデータセットへの適用，説明文生成コストの削減，説明文の自動選別・最適化といった点が挙げられる．これらの



(a) Q-Former で変換前



(b) Q-Former で変換後: gpt-3.5-turbo



(c) Q-Former で変換後: gpt-4o-mini

図 7: 生成テキストの特徴量を t-SNE で可視化

課題に取り組むことで，さらなる汎用性と実用性の高い Few-shot Adaptation 手法の構築が期待できる．

謝辞 本研究は JSPS 科研費 JP24K03020 の支援を受けて実施された．

参考文献

[1] 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, *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763 (2021).

[2] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z. and Duerig, T.: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, *Proceedings of the International Conference on Machine Learning*, pp. 4904–4916 (2021).

[3] Li, J., Li, D., Xiong, C. and Hoi, S. C. H.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, *Proceedings of the International Conference on Machine Learning* (2022).

[4] Udandarao, V., Prabhu, A., Ghosh, A., Sharma, Y., Torr, P. H. S., Bibi, A., Albanie, S. and Bethge, M.: No "Zero-Shot" Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance, *Advances in Neural Information processing Systems* (2024).

[5] Maji, S., Rahtu, E., Kannala, J., Blaschko, M. B. and Vedaldi, A.: Fine-Grained Visual Classification of Aircraft, *arXiv preprint arXiv:1306.5151* (2013).

[6] Geiger, A., Lenz, P. and Urtasun, R.: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2012).

[7] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H. and Qiao, Y.: CLIP-Adapter: Better Vision-Language Models with Feature Adapters, *International Journal of Computer Vision*, Vol. 132, No. 2, pp. 581–595 (2023).

[8] Zhou, K., Yang, J., Loy, C. C. and Liu, Z.: Learning to Prompt for Vision-Language Models, *International Journal of Computer Vision*, Vol. 130, No. 9, pp. 2337–2348 (2022).

[9] Gu, S., Clark, C. and Kembhavi, A.: I can't believe there's no images!: Learning Visual Tasks Using Only Language Supervision, *Proceedings of the International Conference on Computer Vision*, pp. 2672–2683 (2022).

[10] Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D. and Parikh, D.: VQA: Visual Question Answering, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2425–2433 (2015).

[11] Xie, N., Lai, F., Doran, D. and Kadav, A.: Visual Entailment: A Novel Task for Fine-Grained Image Understanding, *arXiv preprint arXiv:1901.06706* (2019).

[12] Liu, F., Wang, Y., Wang, T. and Ordonez, V.: Visual News: Benchmark and Challenges in News Image Captioning, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 6761–6771 (2021).

[13] Li, J., Li, D., Savarese, S. and Hoi, S. C. H.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, *Proceedings of the International Conference on Machine Learning* (2023).

[14] Liang, W., Zhang, Y., Kwon, Y., Yeung, S. and Zou, J.: Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning, *Advances in Neural Information Processing Systems* (2022).

[15] Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P. and Zitnick, C. L.: Microsoft COCO Captions: Data Collection and Evaluation Server, *arXiv preprint arXiv:1504.00325* (2015).

[16] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A. and Simonyan, K.: Flamingo: a visual language model for few-shot learning, *Advances in Neural Information processing Systems* (2022).

[17] Maniparambil, M., Vorster, C., Molloy, D., Murphy, N., McGuinness, K. and O'Connor, N. E.: Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts, *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 262–271 (2023).

[18] Saha, O., Van Horn, G. and Maji, S.: Improved Zero-Shot Classification by Adapting VLMs with Text Descriptions, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17542–17552 (2024).

[19] Cui, Y., Song, Y., Sun, C., Howard, A. G. and Belongie, S. J.: Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4109–4118 (2018).

[20] Wah, C., Branson, S., Welinder, P., Perona, P. and Belongie, S. J.: The Caltech-UCSD Birds-200-2011 Dataset, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2011).

[21] Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y. and Li, H.: Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification, *Proceedings of the European Conference on Computer Vision*, Lecture Notes in Computer Science, Vol. 13676, pp. 493–510 (2022).

[22] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. and et al.: Language models are few-shot learners, *Advances in Neural Information processing Systems*, Vol. 30, No. 33, pp. 1877–1901 (2020).

[23] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H. and et al.: GPT-4 Technical Report, *arXiv preprint arXiv:2303.08774* (2024).

[24] Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, *Proceedings of the International Conference on Learning Representations* (2019).

[25] van der Maaten, L. and Hinton, G.: Visualizing Data using t-SNE, *Journal of Machine Learning Research*, Vol. 9, No. 86, pp. 2579–2605 (2008).