Concept を用いた CLIP モデルの Zero-shot クラス分類の性能改善

ER21015　井上貴斗　　指導教授：藤吉 弘亘

1. はじめに

CLIP [1] は，画像とテキスト間の共通の特徴表現を学習する手法である．学習済みの CLIP モデルは，高精度な Zero-shot クラス分類が可能である．推論時には，画像とテキストの特徴量の類似度を基に画像分類を行う．そのため，画像に対応するテキストが詳細であるほど類似度が向上し，分類精度も向上する可能性がある．そこで，本研究では，追加学習を行わずに，推論時のプロンプトにクラスに関連する詳細な情報である Concept を用いて，Zero-shot クラス分類の精度の向上を目指す．

2. 画像とテキストのアライメントモデル

画像とテキスト間の特徴を近づける学習法である，CLIP と DEAL [2] について述べる．

2.1 CLIP

CLIP は，画像とテキストを用いたマルチモーダル対照学習法である．画像とテキストの特徴量の類似度を最大化するように学習する．CLIP は，推論時に“A photo of a {CLS}.”という形式でクラス名を含むプロンプトを用いて，類似度を求めることで Zero-shot クラス分類が可能となる．

2.2 DEAL

DEAL は，画像内の局所領域に対する説明能力を向上させる手法である．LLM を用いて各クラスに対応する Concept を生成し，各 Concept に対応する画像内の領域が重ならないように学習を行う．DEAL は，Concept を生成する際に Chain-of-Thought (CoT) を用いて，画像から予測不可能なテキストを排除する．CoT とは，具体的な質問と模範的な回答を事前に提示することで，より詳細なテキストを生成させる手法である．CoT を用いた Concept の例を図 1 に示す．図 1 より，例えば「wingspan 2.5 meters」の排除が確認できる．

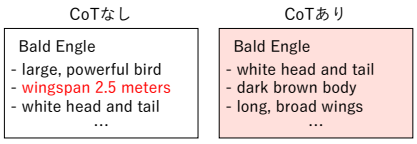

図 1：CoT を用いた Concept の例

3. 提案手法

本研究では，CLIP モデルの Zero-shot クラス分類において，LLM を用いて生成した Concept を用いたプロンプトの改善による分類精度向上を目的とする．Concept は，クラスに関連する詳細な情報であり，画像とテキストの対応関係をより明確にすることで，分類精度向上が期待できる．Concept を用いた Zero-shot クラス分類の流れを図 2 に示す．

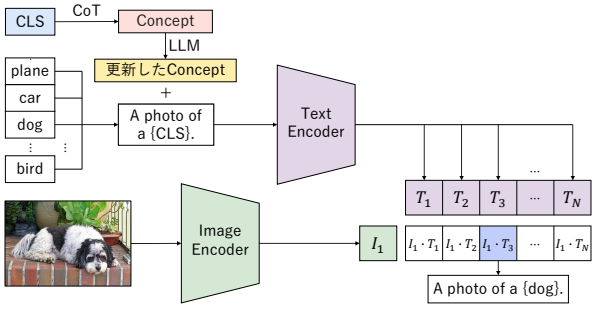

図 2：Concept を用いた Zero-shot クラス分類の流れ

3.1 Concept の作成方法

例として，37 種類の犬と猫の画像を含む OxfordIIITPets データセットにおける Beagle クラスの Concept の作成方法を図 3 に示す．まず，LLM を用いて Concept を生成する．DEAL と同様の CoT を用いて画像から予測不可能なテキストの排除を行う．次に，生成した Concept に対して，LLM を用いて精度向上が期待できる 3 つのアップデートを実施する．1 つ目のアップデートでは，CLIP が持つ上位概念の理解を活用するために，クラスの種類を追加する．2 つ目のアップデートでは，計算コストを減らし，不要な単語が精度に悪影響を与えないようにするために，冗長な単語を削除する．3 つ目のアップデートでは，不自然な文章が精度に悪影響を与えないようにするために，コンマを削除する．

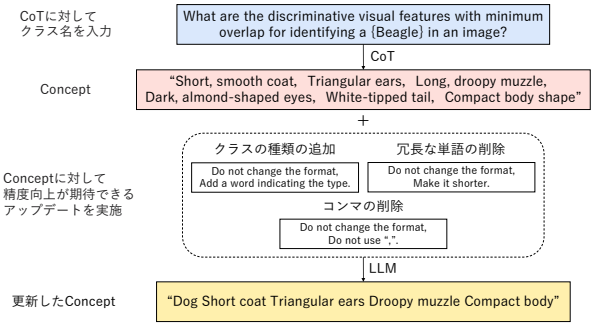

図 3：Beagle クラスの Concept の作成方法

3.2 Concept を用いた Zero-shot クラス分類

まず，クラス名を LLM に入力し，CoT を用いて Concept を生成する．次に，LLM を用いて Concept を更新する．その後，Concept を追加したプロンプトを CLIP モデルの Text Encoder に入力し，対象の画像を CLIP モデルの Image Encoder に入力する．最後に，画像とプロンプトの特徴量の類似度に基づき，Zero-shot クラス分類を実施する．

4. 評価実験

本実験では，LLM を用いて作成した Concept の有無による分類精度への影響を調査する．実験条件として，CLIP の Text Encoder には，Transformer，Image Encoder には，ViT-B/32 を使用する．DEAL の学習データセットには，1000 クラスの画像を含む ImageNet-1K を使用する．CLIP モデルの評価データセットには，OxfordIIITPets を使用する．Concept の有無による分類精度を表 1 に示す．CLIP では，Concept の追加により精度が 5.81 pt 向上した．また，DEAL を学習した CLIP では，Concept の追加により精度がが 9.73 pt 向上した．これらの結果から，CLIP モデルの Zero-shot クラス分類において，Concept の追加による有効性を確認した．

表 1：Concept の有無による分類精度

モデル	DEAL	Concept	分類精度 [%]
CLIP	-	-	76.90
		✓	**82.71**
	✓	-	75.51
		✓	**85.24**

5. まとめ

本稿では，CLIP モデルの Zero-shot クラス分類時に用いるプロンプトに対して，Concept の追加による有効性を確認した．今後は，DEAL の学習を Zero-shot クラス分類に適合させ，精度の向上と汎用性の強化を目指す．

参考文献

[1] A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision”, PMLR, 2021.

[2] Tang Li *et al.*, “DEAL: Disentangle and Localize Concept-level Explanations for VLMs”, ECCV, 2024.