ソフトテキストを用いた CLIP によるゼロショット分類

EP20052 小出清楓　　指導教授：山下隆義

1. はじめに

CLIP[1] は大規模なデータセットを用いてテキストと画像間で共通する特徴表現を学習する手法である．学習済みの CLIP モデルは画像と固有名詞を予測するテキストを与えることで，ゼロショット分類が可能となる．CLIP モデルは，画像とテキストの類似度をもとにクラスを予測するため，クラスに関する情報がテキストに多く含まれているほど類似度が向上し，分類精度が向上する可能性がある．そこで，本研究では，CLIP モデルの追加学習をすることなく，テキストに類似しているクラス名を含むように作成したソフトテキストによる精度向上を目的とする．

2. CLIP

CLIP モデルは図 1 に示すようにテキストエンコーダと画像エンコーダの 2 つで構成されている．画像エンコーダには画像を入力し，対応する特徴を抽出する．一方，テキストエンコーダには，クラス名を含むテキストを入力し，対応する特徴を抽出する．このとき，全てのクラス名に対する特徴を抽出する．そして，画像の特徴とテキストの特徴の類似度を算出する．最も類似度が高いテキストに含まれるクラスを分類するクラスとする．これにより，ゼロショット分類が可能となる．

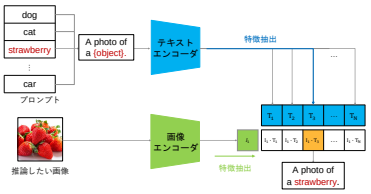

図 1：CLIP モデルの構成

3. 提案手法

CLIP モデルによりゼロショット分類を行う場合，“A photo of a { クラス名 }.” のような 1 つのクラス名のみ含むハードテキストを使用する．ハードテキストには，クラス間の関係性が十分に含まれていないため，類似しているクラスや車の種類などの分類が難しい．そこで，類似しているクラス名を含むように作成したソフトテキストを用いてゼロショット分類を行う手法を提案する．ソフトテキストは CLIP とは別の事前学習したモデルを用いて，クラス予測値の最も高い 2 クラスを使用し，A に似た B（A like B）として画像毎にテキストを作成する．作成方法を図 2 に示す．

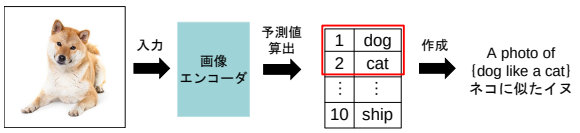

図 2：ソフトテキストの作成方法

4. 評価実験

ソフトテキストの効果を確認するために，ハードテキストとソフトテキストそれぞれに対して同じ画像を用いて CLIP による学習とゼロショット分類を行う．そして，分類精度と特徴量の分布を比較する．

4.1. 実験条件

ソフトテキストを作成するためのモデルとして，ResNet50 を使用する．誤差関数は交差エントロピー誤差とし，バッチサイズを 64，最適化には SGD を用いて 200 エポック学習する．ソフトテキスト作成時に用いる「似ている」という単語は，英語の場合「like」だけではなく「look like」，「resemble」など複数の単語表現があるため，それぞれの単語でソフトテキストを作成する．CLIP のモデルには，ViT-L/14 を用いる．また，特徴量の可視化は UMAP を用いる．使用するデータセットは，CIFAR-10 である．

4.2. 実験結果

ソフトテキストとハードテキストを用いたゼロショット分類精度を表 1 に示す．表 1 より，ソフトテキストを入力とすると，CLIP のテンプレートを用いた場合よりもクラス分類精度が低下した．このことから，従来手法のテキストプロンプトに含まれていない単語を追加したことで，精度が低下したと考える．

表 1：ゼロショット分類時の分類精度の比較 [%]

テキストプロンプト	分類精度 [%]
ハードテキスト	**96.19**
ソフトテキスト (like)	85.56
ソフトテキスト (look like)	95.19
ソフトテキスト (resemble)	91.48

4.3. ハードテキスト特徴量の可視化

CIFAR-10 の画像とハードテキストの特徴量を UMAP で可視化した結果を図 3 に示す．ここで，丸が画像，星がハードテキストを示す．図 3 より，画像とハードテキストの特徴量がクラスごとにまとまっていることが確認できる．そのため，CLIP におけるゼロショット分類ではクラス名を含むテキストを与えると対応するクラスの画像の特徴量に近づくことが分かった．

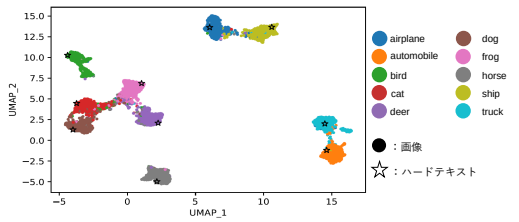

図 3：ハードテキスト特徴量の可視化

4.4. ソフトテキスト特徴量の可視化

精度が最も高い「look like」を用いたソフトテキストと画像の特徴量を UMAP で可視化した結果を図 4 に示す．ここで，丸が画像，星がソフトテキストを示す．

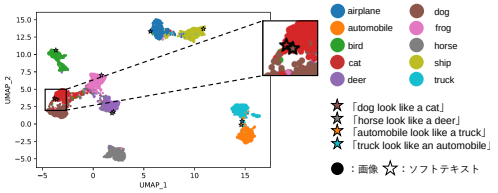

図 4：ソフトテキスト特徴量の可視化

図 4 より，「automobile look like a truck」,「truck look like an automobile」を表すソフトテキストは，automobile と truck の画像特徴の間の位置に分布しており，双方の関係性を表現している．一方で，「dog look like a cat」は誤って cat の画像特徴の位置に分布が確認できる．そのため，追加したクラス名の特徴が強く影響してしまうことが分かった．

5. おわりに

本研究では，ソフトテキストによるゼロショット分類を行い，精度を比較することでソフトテキストを用いることで生じる傾向を調査した．評価実験では，ソフトテキストによる精度の低下を確認できた．今後は，テキストエンコーダの追加学習，複数のソフトテキストを組み合わせることによる精度変化や特徴空間について調査する．

参考文献

[1] A. Radford, *et al.*, Learning Transferable Visual Models From Natural Language Supervision, International Conference on Machine Learning, pp.8748-8763, 2021.