大規模画像言語モデルを用いた領域埋め込みによる画像分類手法に関する研究

1X20C051-5　櫻井洸介
指導教員　後藤正幸

1. 研究背景と目的

近年，大規模なデータセットを用いて事前学習したモデルを活用し，特定のタスクに対して再度学習（ファインチューニング）を行うことで，高精度なモデルを構築する手法が注目されている．特に画像分類タスクにおいては，自然言語データとの大規模マルチモーダルモデルである Contrastive Language-Image Pre-training（CLIP）[1] が事前学習モデルとして様々な問題に応用されている．CLIP を利用したファインチューニング手法の 1 つである Latent Augmentation using Domain descriptionS（LADS）[2] では，ターゲットドメイン（未観測ドメイン）の画像データを説明した文章（言語説明）を CLIP に入力することで，ターゲットドメインの画像潜在表現（CLIP で学習された潜在空間内の一点）を得る．そして，この拡張したデータをファインチューニングに用いることで，1 つのターゲットドメインに特化した画像分類モデルを学習する．しかし，モデルの汎化性能を向上させる上で，LADS による潜在空間内の一点をサンプリングする単純な拡張手法では，学習データには含まれない様々なドメイン（背景や物体の数が異なる場合など）から生じるデータの多様性が考慮できない．

そこで，本研究では各画像の潜在表現を潜在空間上の領域で表現し，その領域内からのサンプリングによるデータ拡張を通じて，様々なドメインに適応可能でより頑健な画像分類手法 Latent Augmentation using Regional Embedding（LARE）を提案する．また，実データを用いた評価実験により提案手法の有効性を示す．

2. 関連研究

2.1. 大規模画像言語モデル

CLIP や Contrastive Captioner（CoCa）[3] に代表される大規模画像言語モデルは，大規模データセットに含まれる画像とテキストを同一空間に射影し，得られた潜在空間を活用することで様々なタスクに応用可能な事前学習モデルである．CLIP，CoCa の概略図を図 1 左，中央に示す．CoCa は CLIP にテキストデコーダを追加することで，CLIP より高い画像分類を実現可能な画像言語モデルである．

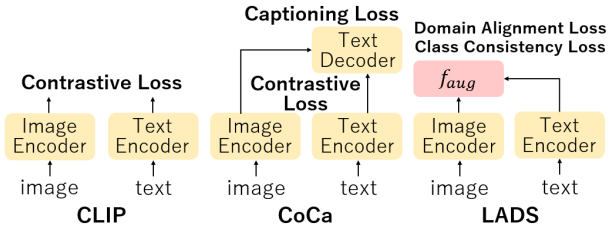

図 1: CLIP，CoCa，LADS の概略図

2.2. LADS

LADS の概略図を図 1 右に示す．LADS は，学習データとして入手することが難しいターゲットドメインにモデルを適応させるために，ソースドメインデータ（学習データ）とターゲットドメインの言語説明を CLIP に入力して得られる潜在表現を利用し，ターゲットドメインの画像潜在表現をデータ拡張する手法である．この際，生成された拡張データは CLIP の潜在空間における一点として表現される．例えば，学習データとして入手困難な「絵」の画像（ターゲットドメイン）の識別精度を上げるために,「写真」の画像（ソースドメイン）と「A painting of a [label]」という言語説明を用いて,「絵」の画像（潜在表現）を疑似的に生成する．生成したターゲットドメインデータを用いて CLIP をファインチューニングすることで，ターゲットドメインに対する精度が向上する．

3. 提案手法（LARE）

3.1. 提案手法（LARE）の概要

本研究では，大規模画像言語モデルから出力される画像の潜在表現（一点）から領域を指定するエンコーダを学習し，得られた領域内からのサンプリングによるデータ拡張を通じて，様々なドメインに対して頑健な画像分類を実現する手法 LARE を提案する．領域を活かしたデータ拡張により，(学習データには含まれないような）多様なドメインを特定の言語説明を介さずにデータ拡張することが可能となる．

LARE は 2 つのステップからなる．ステップ 1 では，大規模画像言語モデルから得られる潜在表現を利用し，各画像に対する潜在空間上の領域を指定するエンコーダを学習する．ステップ 2 では，ステップ 1 で得た領域内からランダムサンプリングしたデータと元データを用いて，CoCa にファインチューニングを行う．なお，ファインチューニングにはいくつかの方法が考えられるが，本研究では CoCa の画像エンコーダに追加した 1 層の全結合層のみを学習するリニアプロービングによって実現する．ステップ 1 で拡張したデータを用いることで，ターゲットドメインに適応し，従来よりも汎用性の高い画像分類モデルを構築する．

3.2. ステップ1（領域生成）の詳細

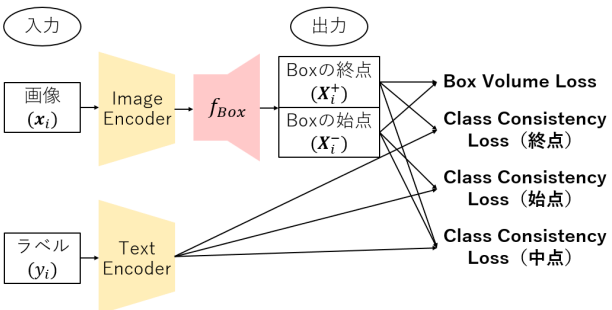

図 2: 提案手法（ステップ 1）の概略図

ステップ 1 の概略図を図 2 に示す．本研究では，各画像の領域を超直方体（Box [4]）として表現し，超直方体の最も小さい座標（始点）$\boldsymbol{X}_i^- \in \mathbb{R}^d$ と最も大きい座標（終点）

$\boldsymbol{X}_i^+ \in \mathbb{R}^d$ を出力するような $f_{Box} : \mathbb{R}^d \to \mathbb{R}^{2d}$ を学習する．なお，ステップ 1 における入力を i 番目の画像 $\boldsymbol{x}_i$ とラベル y_i，出力を i 番目の画像に対する領域（Box）の始点 $\boldsymbol{X}_i^- \in \mathbb{R}^d$ と終点 $\boldsymbol{X}_i^+ \in \mathbb{R}^d$ と定義する．また，LADS と異なり，CLIP ではなく CoCa の画像エンコーダとテキストエンコーダを用いることで，より精度を高める．

ステップ 1 における損失関数を以下に示す．式 (1) は各画像の Box を大きくする（Box の始点と終点を離す）役割を持ち，Box Volume Loss と定義する．

$$L_{BV}(f_{Box}) = \sum_{i=1}^{n} \left(\boldsymbol{X}_i^+ \cdot \boldsymbol{X}_i^- \right) \tag{1}$$

式 (2) は終点，式 (3) は始点，式 (4) は中点に対し，それぞれクラスの情報を保持するための Class Consistency 損失である．なお，式中の $CE(a, b)$ は予測ラベル a と正解ラベル b に対する Cross-Entropy 損失，$S[\cdot]$ は Softmax 関数，$T_\theta(\cdot) \in \mathbb{R}^d$ はテキストエンコーダの出力を表す．

$$L_{CC}^+(f_{Box}) = \sum_{i=1}^{n} CE\left(S\left[\boldsymbol{X}_i^+ \cdot T_\theta(y_i) \right], y_i \right) \tag{2}$$

$$L_{CC}^-(f_{Box}) = \sum_{i=1}^{n} CE\left(S\left[\boldsymbol{X}_i^- \cdot T_\theta(y_i) \right], y_i \right) \tag{3}$$

$$L_{CC}(f_{Box}) = \sum_{i=1}^{n} CE\left(S\left[\frac{\boldsymbol{X}_i^+ + \boldsymbol{X}_i^-}{2} \cdot T_\theta(y_i) \right], y_i \right) \tag{4}$$

以上を踏まえて，ステップ 1 では式 (5) の損失関数 L_{LARE} を最小化するように，図 2 に示す f_{Box} を学習する．なお，f_{Box} は 2 層のニューラルネットワーク，α は各損失の重みを決定するハイパーパラメータである．

$$\begin{aligned} L_{LARE}(f_{Box}) &= (1-\alpha) L_{BV}(f_{Box}) \\ &+ \frac{\alpha}{3} L_{CC}^+(f_{Box}) + \frac{\alpha}{3} L_{CC}^-(f_{Box}) + \frac{\alpha}{3} L_{CC}(f_{Box}) \end{aligned} \tag{5}$$

4. 実験

4.1. 実験条件

本研究では，提案手法におけるデータ拡張の有効性を確認するため，CUB+CUB-Painting と CIFAR-100 を用いて実験を行う．CUB と CUB-Painting は共に 200 クラスからなる鳥の画像データセットであり，それぞれ鳥の写真，鳥の絵画データが含まれる．CIFAR-100 は 100 クラスからなる物体カラー画像データセットである．実験は各 5 回ずつ行い，その分類精度（正解率）の平均と標準偏差を記載する．

4.2. 実験結果と考察

データセット CUB+CUB-Painting を用いて，CUB の一部を学習データ，残りの CUB と CUB-Painting をテストデータとした実験結果を表 1 に示す．表 1 における In-domain はソースドメイン（CUB）に対する精度，Out-of-domain はターゲットドメイン（CUB-Painting）に対する精度を表す．

表 1 より，In-domain，Out-of-domain 共に提案手法 LARE が最も良い精度を示していることがわかる．このことから，提案手法の有効性が示唆される．また，Out-of-domain に関して，LARE の精度は CoCa(fine-tuning) の精度を上回ることから，未観測ドメインに対する提案手法の有効性が示唆される．

表 1: CUB（CUB-Painting）に対する分類精度比較

手法	正解率 [%]	
	In-domain	**Out-of-domain**
CLIP(zero-shot)	63.27	54.71
CoCa(zero-shot)	73.63	65.05
CLIP(fine-tuning)	86.39(± 0.03)	64.40(± 0.16)
CoCa(fine-tuning)	86.98(± 0.16)	71.56(± 0.12)
LADS	86.31(± 0.15)	65.90(± 0.15)
LARE	**87.01(± 0.04)**	**72.98(± 0.16)**

次に，CIFAR-100 を用いて，少量のデータに対する提案手法の有効性を確認するため，few-shot 学習の結果を図 3 に示す．図 3 における zero-shot は，学習データを用いず大規模画像言語モデルの潜在表現のみで分類した精度を指す．図 3 より，少量のデータに対する学習精度は CoCa のファインチューニング精度を大幅に上回ることが確認された．このことから，学習データが少ない場合やクラスごとのデータ数に偏りがある場合に，LARE は有効であると考えられる．

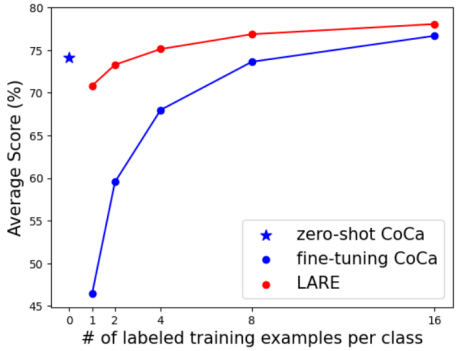

図 3: CoCa と LARE の few-shot 精度比較

5. まとめと今後の課題

本研究では，画像言語モデルに領域埋め込みを適用したデータ拡張・画像分類手法を提案した．未観測ドメインのデータや少量のデータに対する実験結果より，従来の画像言語モデルと同等以上の精度を示したことから，提案手法は有効なデータ拡張手法であることが示唆される．今後の課題としては，領域サイズの決定方法や最適な結合パラメータを決定する方法の確立などが挙げられる．

参考文献

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

[2] Dunlap, L., et al. “Using Language to Extend to Unseen Domains,” In *Proceedings of ICLR*, 2023.

[3] Yu, J., et al. “CoCa: Contrastive Captioners are Image-Text Foundation Models,” In *Proceedings of CVPR*, 2022.

[4] Dasgupta, S. S., et al. “Word2Box: Capturing Set-Theoretic Semantics of Words using Box Embeddings,” In *Proceedings of ACL*, 2022.