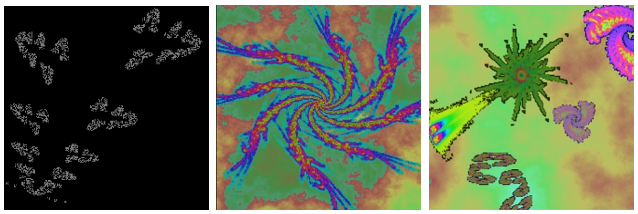
ISS-SP-016

人工生成画像による Transformer の学習

梅川昇也 [1]　鏡川悠介 [1]　前田英作 [1]

[1] 東京電機大学



(a) FractalDB[1]　(b) Single-Instance[2]　(c) Multi-Instance[2]

図 1: 人工生成データセットの比較

1　まえがき

画像分類モデルのパラメータ数増加に伴い，学習のためのデータセットにも大規模な自然画像データセットが必要とされている．しかし，大規模データセットの作成には大量の画像収集や膨大なアノテーションコストを必要とする．また，人手によるアノテーションの一貫性や画像の著作権，コンテンツの偏りなど様々な問題も存在する．こうした問題を解決するため，片岡らは人手を介さず画像と対応するラベルを自動生成する FractalDB を提案し，一部の分類タスクにおいて既存手法に匹敵する正解率を達成した [1]．さらに Anderson らは FractalDB の生成手順を改良した手法を提案し，畳み込みニューラルネットワーク (CNN) モデルの Pre-Training (PT) での有効性を確認した [2]．本研究では，CNN に代わって注目され始めている Transformer アーキテクチャを使用した Data-efficient image Transformers (DeiT) [3] を対象に人工生成データセットで PT を行い，画像分類タスクにおいて評価する．

2　実験

PT データセットには片岡らによる FractalDB(a)，Anderson らによる Single-Instance(b)，Multi-Instance(c) の 3 つの人工生成データセットを用意した (図 1)．また，自然画像データセットである ImageNet2012 [4] も用意した．各データセットは約 100 万枚で構成される．学習モデルには DeiT(small_patch16_224) と ResNet50 [5] を用意し，各 PT データセットを 90 epoch 学習した．

ResNet50，DeiT について，それぞれ 4 種の PT 済みモデルを CIFAR-100(C100)，CIFAR-10(C10) データセットで Fine-Tuning (FT) し，評価を行った．C10，C100 は動植物や車両などを含むそれぞれ 10 クラス，100 クラスの合計 6 万枚の自然画像データセットである．5 万枚を FT に，1 万枚を評価に使用した．各 PT 済みモデルに対して FT を 150 epoch 行った．

使用言語は Python，モデルを作成するフレームワークには PyTorch を使用した．学習には NVIDIA A100(40GB) を 2 枚使用して，64 のバッチサイズで学習した．学習条件は Anderson らのものと統一した[1]．

[1] https://github.com/catalys1/fractal-pretraining

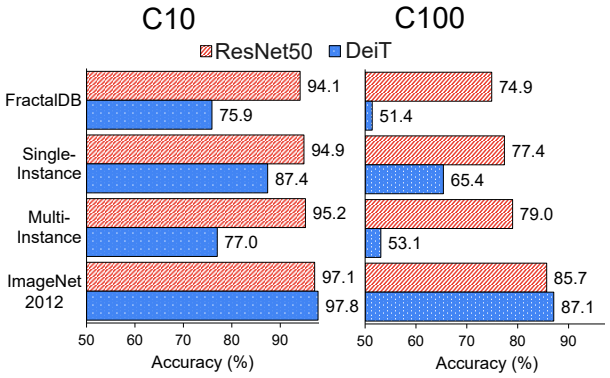


図 2: PT 別 ResNet50，DeiT の Accuracy (C10,C100)

3　結果

4 種のデータセットで PT された ResNet50，DeiT について C10，C100 データセットで FT を行い評価を行った結果，C10，C100 どちらも DeiT，ResNet50 共に Single-Instance，Multi-Instance での PT は，FractalDB での PT より正解率 (Accuracy) が高くなった (図 2)．また 3 つの人工生成データセットの中で最も PT に有効であったものは，ResNet50 では Multi-Instance であるが DeiT では Single-Instance であった．これらの傾向は C10，C100 どちらのデータセットでも同じ傾向であった．

4　考察

ImageNet2012 で PT した場合，DeiT の正解率は ResNet50 を上回った．しかし人工生成データセットで PT した場合，DeiT の Accuracy は ResNet50 を下回った．これは Anderson らが ResNet50 用に使用した学習条件が，フラクタル画像を用いた DeiT の学習に適していないためであると考えられる．また Multi-Instance で PT した場合，ResNet50 は Single-Instance で PT した場合より正解率が高くなるが，DeiT では低くなる．これは DeiT の学習方式である，入力画像のパッチ分割が影響を及ぼしていると考えられる．具体的には，1 枚の中に異なるラベルが配置された Multi-Instance 画像をパッチに分割してしまうと学習が困難になると考えられる．

5　今後の課題

Multi-Instance で PT した DeiT は，ResNet50 の場合と比べて正解率が低かった．Touvron のコードで学習することや，224 サイズの入力画像からパッチに分解するサイズを 2 倍にした DeiT(small_patch32_224) を使用することで検証する．

謝辞

本研究は JSPS 科研費 JP119H01134 の助成を受けた．

参考文献

[1] H. Kataoka *et al.*, ACCV, 2020.

[2] C. Anderson and R. Farrel, WACV, 2022.

[3] H. Touvron *et al.*, ICML, 2021.

[4] J. Deng *et al.*, CVPR, 2009.

[5] K. He *et al.*, CVPR, 2016.