ISSN 2186-5647

2-29

畳み込みニューラルネットワークを用いたViT

日大生産工院　○張 宏毅　　日大生産工　山内　ゆかり

1. まえがき

畳み込みニューラルネットワーク（CNN）[1]は、画像認識において非常に効果的なモデルであり、主に画像の局所的な特徴を捉えるために設計されている。CNNは平行移動に対して強い不変性を持ち、各層で画像の部分的な情報を集約することで、効率的に画像を理解することができる。

一方、Vision Transformer（ViT）[2]は、トランスフォーマーモデルを画像認識に応用したものであり、グローバルな情報を捉えるのに優れている。しかし、CNNのような平行移動不変性や局所的な受容野を持たないため、より多くのデータが必要であり、データ効率が低いという欠点がある。

この問題を解決するために、ViTの前段にCNNを追加する。CNNの特徴抽出能力を活用することで、ViTのデータ効率を向上させることを提案する。

2. 従来研究

2.1 CNN

CNNは、特に画像処理において優れたパフォーマンスを発揮するディープラーニングモデルである。CNNの基本的な構成要素は以下の通りである：畳み込み層、プーリング層、活性化関数と全結合層である。これらの層が順次配置され、画像から抽出された特徴を基に最終的な分類を行う。

2.2 ViT

2.2.1 概論

ViTは、自然言語処理（NLP）分野で成功を収めたトランスフォーマーアーキテクチャを画像認識に応用したモデルである。ViTは、画像を固定サイズのパッチに分割し、各パッチを線形埋め込みすることで、トークンとして扱う。これにより、従来のCNNとは異なるアプローチで画像の特徴を捉えることが可能となる。Figure 1にViTのイメージを示す。

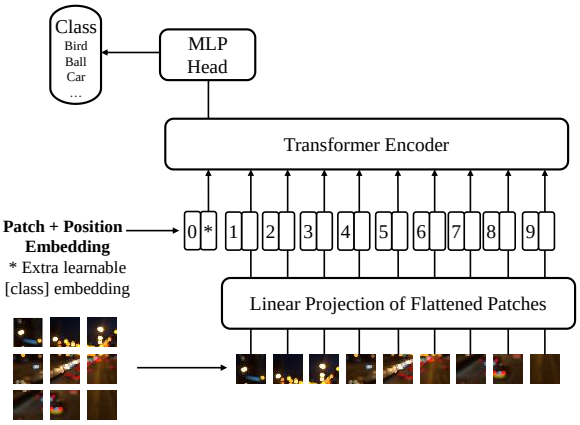

Figure 1　ViT の基本構造

2.2.2 Embedding層

標準的なTransformerモジュールでは、入力としてトークン（ベクトル）のシーケンス、つまり二次元行列[num_token, token_dim]が要求される。Figure 1のように、トークン0-9に対応するのはすべてベクトルである。

画像データに関して、そのデータフォーマットは[H, W, C]という三次元行列である。そのため、Embedding層を通じてデータを変換する必要がある。Figure 1のように、まず画像を指定されたサイズに基づいて複数のパッチに分割する。その後、各パッチを線形変換により一つの一次元ベクトルにマッピングする。

Transformer Encoderに入力する前に、[class]トークンとPosition Embeddingを加える必要がある。これは線形変換で得られた一連のトークンに分類専用の[class]トークンを挿入する。この[class]トークンは学習可能なパラメータであり、他のトークンと同様の形式であり、画像から生成されたトークンと連結される。Figure 1の0番トークンは[class]トークンである。

その後、Position Embeddingについては、Transformerで説明されたPositional Encodingのことで、ここでは学習可能なパラメータ（1D Pos. Emb.）が使用されており、トークンに直接加算される。そのため、形状も同じでなければならない。

2.2.3 Transformer Encoder層

Transformer Encoder層とは、エンコーダーブロックを複数回スタックしたものであり、一

Vision Transformer Enhanced with Convolutional Neural Networks

Hongyi ZHANG and Yukari YAMAUCHI

般的なCNNに似ており、主にいくつかの部分で構成されている。Figure 2 にTransformer Encoder層のイメージを示す。

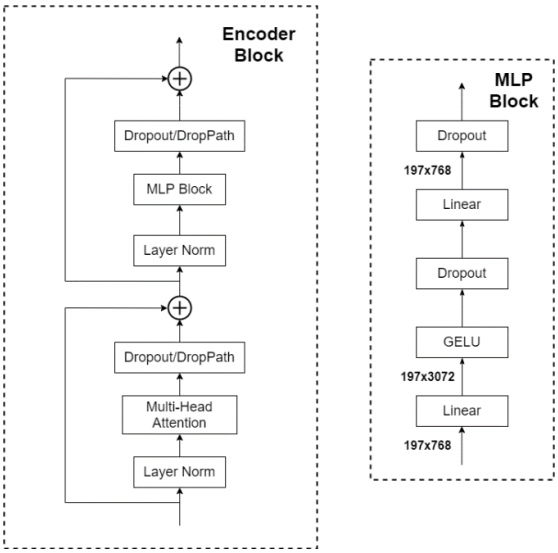

Figure 2　Transformer Encoder 層

Layer Normalization（LN）とは、特定のサンプルに対してそのサンプル内のすべての特徴マップの平均と分散を計算し、正規化する。Dropout は、ニューラルネットワークの訓練時に使用される正則化技術の一つで、主には過学習を防ぐために利用されている。Figure 2 右側に示されているように、MLP Blockは全結合層、GELU活性化関数とDropOutで構成されており、逆ボトルネック構造を採用している。

2.2.4　MLP Head

上記のTransformer Encoderを通過した後、出力のshapeは入力のshapeと変わらない。ここで我々が必要とするのは分類に関する情報だけであるため、[class]トークンに対応するベクトルを抽出すればいい。その後、MLP Headを通して最終的な分類結果を得る。

3.　提案手法

ViTはCNNのような平行移動不変性や局所的な受容野を持たないため、より多くのデータが必要であり、データ効率が低いという欠点がある。本研究では、この欠点を解決するために、画像がTransformerに入る前に、まずCNNを通じて画像全体の特徴を抽出する。言い換えれば、CNNは、元のピクセルを直接処理するのではなく、高次の特徴を抽出するための前置エンコーダーとして使用される。

ViTはグローバルな情報を捉える一方、CNNは局所的な特徴の抽出に優れている。これらの特性を組み合わせた提案モデルは、データ効率の向上に貢献し、より高精度な分類結果を得ることが期待できる。

4.　実験および検討

ViTとCNNの結合を行い、実験および微調整を行う。CNNはベースラインモデルとして実験結果を載せる。

Table 1 実験結果

Datasets	Train Acc	Test Acc
MNIST	0.903	0.899

表 1 は実験結果を示す。この結はMNISTデータセットを使用し、Batchサイズ64、学習率0.001、カーネルサイズ3×3でCNNを10Epoch学習した結果である。CNNは比較的軽量であり、特に小規模なデータセットであるMNISTでは、短時間での収束が期待できる。結果として、訓練時の精度は0.903、テスト時の精度は0.899である。

5.　まとめ

本研究では、ViTのデータ効率が低いという課題に対処するため、CNNを前段に追加するアプローチを提案した。CNNは局所的な特徴を効果的に抽出する能力があり、これを活用することで、ViTのグローバルな情報捕捉能力と補完し合い、データ効率の向上を目指している。これにより、高精度な分類結果が期待される。

参考文献

[1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel., “Backpropagation Applied to Handwritten Zip Code Recognition”, Neural Computation, vol. 1, no. 4, (1989) pp. 541-551.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, arXiv:2010.11929v2 (2020) pp. 1-21.