ViT と CNN を融合した画像分類モデルの提案

魏 鑫　　指導教員：小林 邦和

1　はじめに

近年，コンピュータビジョン (Computer Vision，以下 CV と略記) の分野では，畳み込みニューラルネットワーク (Convolutional Neural Networks，以下 CNN と略記) が主流を占めてきた．しかし，自然言語処理 (Natural Language Processing，以下 NLP と略記) の分野の Transformer[1] を CV 分野に適用し，かなり良い効果を実現している研究が続出している [2][3][4]．ビジョントランスフォーマー (Vision Transformer，以下 ViT と略記)[2] は Transformer を画像分野に適用できるように変更し，複数の画像認識基準上で現在の SOTA(state of the art) 手法に近い性能を実現している．

ViT は訓練時間が掛かる CNN の畳み込み層を使わないのがメリットだが，ViT の CV 分野への応用は，計算量，メモリ占有量が非常に大きいなど多くの問題を解決する必要がある．CNN は下位層の特徴と視覚構造を抽出する上で大きなメリットがある．これらの下位層の特徴は，明らかな幾何学的特性を有し，平行移動，回転などの変換に対する不変性，または共変性に注目することが多い．CNN は，このような共変性を扱う際に自然な選択である．しかし，これらの基本的な視覚要素を検出すると，上位層の視覚特徴は，これらの要素間がどのように関連して 1 つの物体を構成するか，物体と物体との空間的位置関係がどのようにシーンを構成するかに注目する．現在，Transformer はこれらの要素間の関係を扱う上で有効である．

既存の研究からも，両者の融合はより良い結果をもたらすことが確実である．たとえば，A. Hassani ら [5] より提案された Compact Convolutional Transformer(以下 CCT と略記) は正確なサイズおよび符号化により，小規模データセット上で最先端の CNN に匹敵し，より高い正解率をより少ないパラメータで実現することができる．従って，ViT と CNN の融合が良いと考えられる．

本研究では，CV 分野で，ViT と CNN を融合した小規模データセットのための高速で高精度な画像分類モデルを提案する．提案したモデルは CIFAR-10 データセット [6] を用いて，訓練時間と正解率を従来モデル ViT と類似モデル CCT に対して，比較する．同一 Epoch の条件で，提案モデルは訓練時間を短縮するとともに正解率を向上させることを目標とする．

2　提案モデル

本研究では，Transformer の動的注意機構と大域モデリング能力，CNN の局所特徴捕捉能力を兼ね備え，局所と大局の情報を結合するモデリング能力を備えた ViT と CNN を融合した画像分類モデルを提案する．

提案モデルは階層構造であり，階層ごとに 2D の画像または Tokens が Convolutional Embedding によって特徴ベクトルを生成または更新する．各層には N 個の典型的な Convolutional Transformer Block が含まれており，線形変換を畳み込み変換に置き換えてマルチヘッダ Attention 機構に入力し，Layer Norm を行う．

Convolutional Projection は，提案モデルネットワークが画像信号の空間構造情報を維持することを可能にし，Tokens が画像情報の局所情報相関を適切に利用するとともに，Attention を用いて大局的情報をモデリングすることを可能にしている．

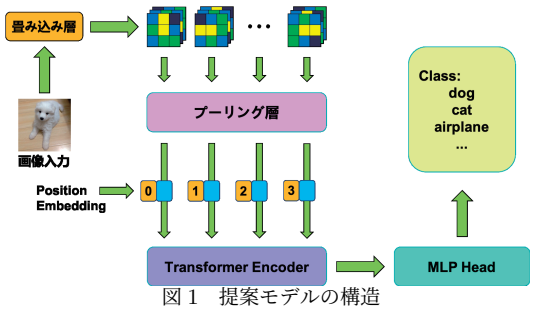

図 1　提案モデルの構造

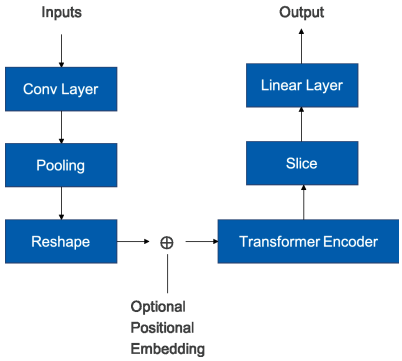

図 2　提案モデルの流れ図

提案モデルは CCT モデルと類似している．CCT は，ローカル情報を保持し，パッチ間の関係を符号化することができる畳み込みに基づくパッチ手法を導入している．一方，提案モデルは Transformer Encoder の入力から ViT モデルとほぼ同様であり，CIFAR-10 などの小規模データセットのより小さなパッチサイズを扱うのに適している．

2.1　構造

提案モデルの構造を図 1 に示す．

提案モデルに誘導バイアスを導入するために，ViT の「画像修復レイヤー」と「埋め込みレイヤー」を単純な畳み込みブロックに置き換える．畳み込みブロックは従来の設計に従い，単一の畳み込みレイヤー，正規化線形ユニット層，およびプーリング層で構成される．

モデルデザインについては，ViT モデルに似た特徴を持つ．位置の埋め込みを追加した後，レイヤーノルムを適用する．Transformer Encoder は，レイヤーの正規化，GELU(Gaussian Error Linear Unit) のアクティブ化，およびドロップアウトを使用する．

2.2　アルゴリズム

提案モデルの流れ図を図 2 に示す．

まず，入力画像を畳み込み層に入力し，特徴を抽出する．その後プーリングを行う．続いて，2 次元配列を 1 次元配列へ変換し，Transformer の Encoder へ入力する．この間，Positional Embedding はオプションである．最後に，Slice と Linear Layer

表 1　ViT，CCT，提案モデルのハイパパラメータ設定

ハイパパラメータ	ViT	CCT	提案モデル
learning rate	0.001	0.001	0.001
weight decay	0.0001	0.0001	0.0001
batch size	256	128	128
num epochs	1000	1000	1000
image size	72	32	32
patch size	6	-	-
conv layers	-	2	2
projection dim	64	128	128
num heads	4	2	2
transformer layers	8	2	2
stochastic depth rate	-	0.1	-
mlp head units	[2048, 1024]	-	[2048, 1024]
positional emb	-	True	True

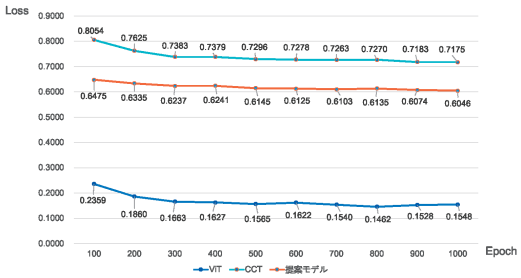

図 3　ViT，CCT，提案モデルにおける Loss の推移

関数を用いて線形層を作成し，分類結果を出力する．

3　計算機シミュレーション

提案モデルの性能評価のため，計算機シミュレーションを行った．

3.1　問題設定

本研究では，提案モデル (ViT と CNN を融合した画像分類モデル) と従来モデル ViT，類似モデル CCT を比較するために，CIFAR-10 データセットで各モデルを訓練し，100 Epoch 毎に，テストデータで汎化性能を求める．同一データセット，同一 Epoch の条件で，各モデルに必要な訓練時間と正解率を比較する．提案モデルは，訓練時間を短縮するとともに，正解率が向上することが期待される．

3.2　ハイパパラメータ設定

ViT，CCT，提案モデルのハイパパラメータ設定を表 1 に示す．ViT のハイパパラメータ設定は ViT の論文 [2] で使用されていた設定値である．CCT のハイパパラメータ設定は CCT の論文 [5] で使用されていた設定値である．提案モデルで用いるハイパパラメータ設定は，ViT と CCT の設定値を参考にしたうえで，試行錯誤で設定した．

3.3　結果と考察

ViT，CCT，提案モデルで 100 Epoch 毎の Loss を図 3 に示す．ViT，CCT，提案モデルで 100 Epoch 毎の正解率を図 4 に示す．

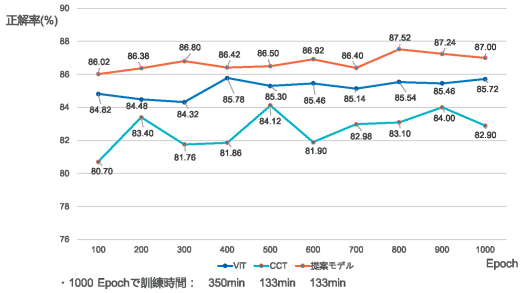

図 4　ViT，CCT，提案モデルにおける正解率の推移

計算機シミュレーションの結果を見ると，CIFAR-10 データセットにおいて，ViT も CCT も提案モデルも満足のいく結果を得ることができなかったが，同一データセット，同一 Epoch の条件で，提案モデルは ViT モデルを比較すると，訓練時間を半分以上減少するとともに，正解率が向上している．類似モデル CCT を比較すると，訓練時間は減少しなかったが，正解率が向上している．

ViT の論文 [2] で指摘されているように，モデルの性能はアーキテクチャ選択だけでなく，学習率計画，オプティマイザ，重み減衰などのハイパパラメータの影響を受ける．

4　おわりに

提案モデルは ViT モデルと CNN モデルの各優位性を融合した新しい基盤ネットワークである．また，計算機シミュレーションを通して，CIFAR-10 のデータセットで画像分類タスクにおける提案モデルの有効性を明らかにした．

今後は，1000 Epoch 以上で訓練すること，さらに多数の Transformer 層を用いること，入力画像サイズとパッチサイズを調整すること，投影次元を増加させることなどにより，事前訓練を行わずにモデルの性能を向上させて行きたい．

参考文献

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin: “Attention Is All You Need”, In Advances in Neural Information Processing Systems (NIPS), (2017)

[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 16 × 16 Words: Transformers for Image Recognition at Scale”, In International Conference on Learning Representations (ICLR), (2021)

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo: “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”, arXiv:2103.14030, (2021)

[4] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou: “Training data-efficient image transformers & distillation through attention”, arXiv:2012.12877, (2021)

[5] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi: “Escaping the Big Data Paradigm with Compact Transformers”, arXiv:2104.05704, (2021)

[6] A. Krizhevsky, V. Nair, and G. Hinton: “Learning Multiple Layers of Features from Tiny Images”, (2009)