G-18

画像分類における CNN と Vision Transformer の精度比較

Accuracy Comparison between CNN and ViT for Medical Image Classification

吉次 研二 †　　中本 幸一 †
Kenji Yoshitugu　　Yukikazu Nakamoto

1. はじめに

これまで画像の多クラス分類には畳み込みニューラルネットワーク（CNN）が使われてきた．それが最近その他の方法で CNN の分類精度と競うものが提案されるようになっている．この発表で取り上げた Vision Transformer（ViT）も，これまで自然言語処理に利用されてきた Transformer を画像の多クラス分類に応用した例である．

本発表では，糖尿病性網膜症の医用画像診断コンペで提供された教師データを用いて，CNN に対して ViT の分類精度が具体的にどこまでの結果を出すかの検証として、分類精度の比較を行った．

2. 背景

畳み込みニューラルネットワーク（Convolutional Neural Network: CNN または ConvNet）とは，何段もの層を持つニューラルネットワークで，特に画像認識の分野で優れた性能を発揮しているネットワークである．このネットワークは，図 1 に示すようにいくつかの特徴的な機能を持った層を多段に積み上げることで構成されたネットワークで，現在幅広い分野で活用されている．

特徴的な機能を持った層に「畳み込み層」と「プーリング層」があり，入力画像全体に対して畳み込み層でフィルター処理を行ってプーリング層に渡し，プーリング層で画像の解像度を下げる処理を行う．

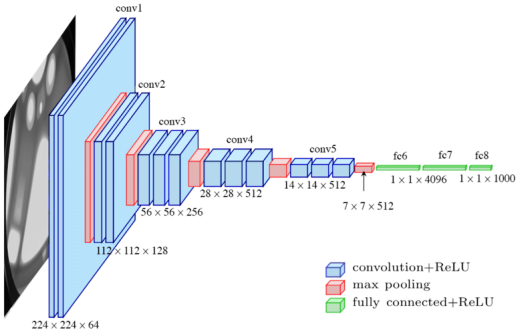

図 1　CNN の多層化の例（VGG-16）[1]

この畳み込み層とプーリング層の組み合わせを多層化することで効果的に画像認識を行うことで，期待する精度を得るためにはどのように各層を組み合わせるべきかは実装側に委ねられている．

† 兵庫県立大学，University of Hyogo

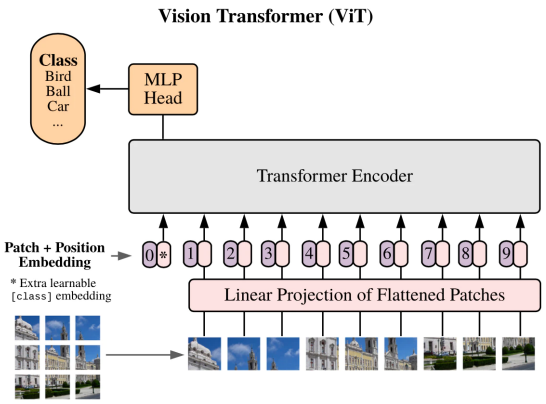

図 2　Vision Transformer[2]

この課題を汎用的な多層化モデルの実装として，現在広く利用されているのが転移学習（Transfer Learning）である．転移学習とは，ある分類で学習済みのモデルを別の分類に転用して，効率的に学習させる方法である．従来，深層学習においては先述の多層化の構造を検討する労力に加えて，学習を完了させるまでにかなりの時間がかかっていたが，学習済みのモデルを利用することで，学習にかかる時間と労力を低減することが可能になり，現在多数の転移学習モデルが公開されて広く利用されている．

Vision Transformer（ViT）は，Google によって初めて提案された画像分類モデル [2] で，一般的な CNN を使わずにこれまで自然言語処理で使われてきた Transformer を使うモデルである．ViT は，図 2 に示すように，自然言語処理におけるトークン（単語）のように，画像を「パッチが連なったシーケンスデータ」として扱うことで，画像処理に Transformer を適用している．

ViT は，大量のデータで事前学習を行い，ImageNet や CIFAR-100 のような画像認識ベンチマークに転移学習を行ってテストした場合に，従来の CNN 型モデルと同等かそれ以上の結果を出す事例が報告されている [3]．しかしその一方で，CNN に比べて計算負荷が高い点や，Transformer はデータセットが小さい場合に CNN がもつ位置等価性や局所性といった帰納的バイアスを欠くため，汎化性能が低

下してしまうという点や，得られた特徴がユニークな特性を示さない傾向がある [2].

今回の検証には、機械学習コンペティションサイト Kaggle で過去 2 回行われた糖尿病性網膜症（Diabetic Retinopathy：DR）画像診断コンペ [4][5] の画像を利用した。図 3 に眼底画像データの例を示す。

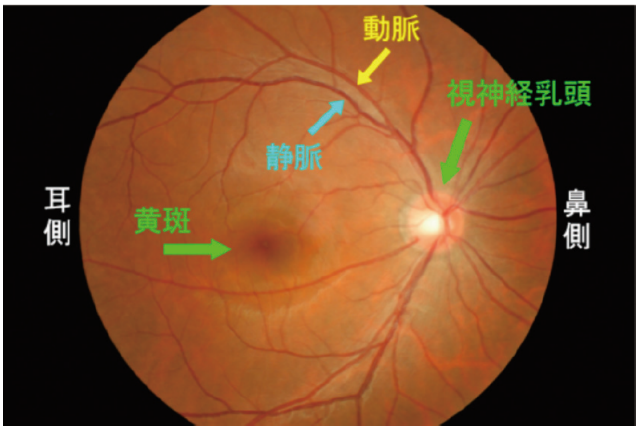

図 3　眼底画像の例（正常な右目）[6]

糖尿病性網膜症の画像診断は、症状の重症度別に表 1 の 5 つに分類することを目的としている。

表 1　糖尿病性網膜症の重症度分類

クラス	重症度
0	正常
1	非増殖性 DR（軽度）
2	非増殖性 DR（中度）
3	非増殖性 DR（重度）
4	増殖性 DR

今回の教師データで特徴的なのは、分類目的は同じでありながら、２つのデータ群の大きさが一桁異なることである。それぞれのデータ群のデータ構成を表 2 と表 3 に示す。

表 2　APTOS[4] の教師データ（全 3662 件）

クラス	データ数
0	1805 件
1	370 件
2	999 件
3	193 件
4	295 件

表 3　EyePACS[3] の教師データ（全 35126 件）

クラス	データ数
0	25810 件
1	2443 件
2	5292 件
3	873 件
4	708 件

これら２つのコンペの教師データに対して CNN と ViT で深層学習を行うことで、分類テーマは同じでもデータ数に差がある場合について，処理速度と診断精度の比較検証を行うことが可能になる．

3.　検証

検証方法を以下に示す：

(1) 使用した PC と OS は以下の通り。
OS:Ubuntu20.04LTS，CPU:Intel Core i7-8700K，64GB メモリ，GPU:NVIDIA RTX3060，Python/PyTorch.

(2) CNN 環境には転移学習モデルとして広く使われている Resnet50[7] を利用し，ViT 環境には専用の事前学習モデルとして用意されている ViT-B_16[2] を利用した．学習は 30 エポックとし、これが終了するまでの時間を所要時間として計測した。

(3) CNN と ViT に一般的なデータオーグメンテーションとして，共通で以下の 6 つの画像処理を行った．前処理は行っていない．
(a) 384 ピクセル× 384 ピクセルにリサイズ
(b) 1/2 の確率で水平に反転
(c) 1/2 の確率で垂直に反転
(d) 1/2 の確率で回転
(e) 1/2 の確率で明るさとコントラストを変更
(f) 正規化

4.　結果と考察

それぞれの結果を順に示す．

(1) CNN で EyePACS を学習／診断した結果（図 4 と図 5）．損失と精度の両方に共通して，学習時に比べて診断時に値が上下に大きくぶれながら推移している（所要時間：167 分 08 秒）．

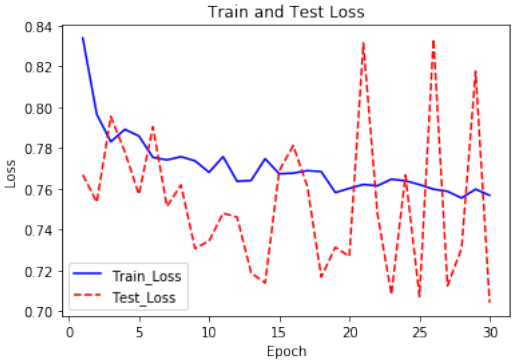

図 4　CNN EyePACS 損失

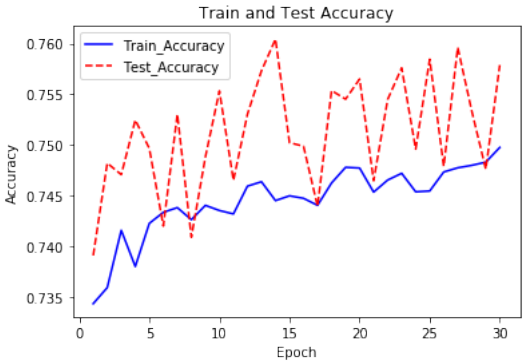

図 5　CNN EyePACS 精度

(2) ViT で EyePACS を学習／診断した結果（図 6 と図 7）．CNN の場合に比べて処理に時間がかかっているが，5 エポック辺りで値が収束し始めている（所要時間：897 分 43 秒）．5 エポックで学習を終了させると、処理にかかった時間は CNN の場合より短くなる．

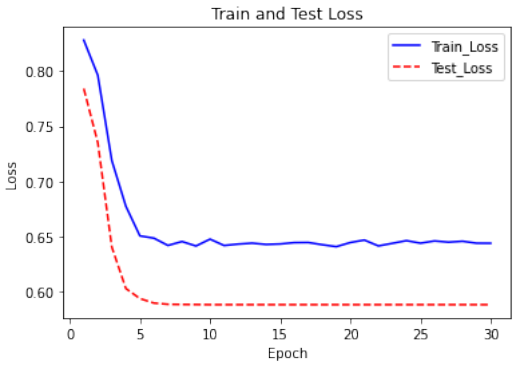

図 6　Vit EyePACS 損失

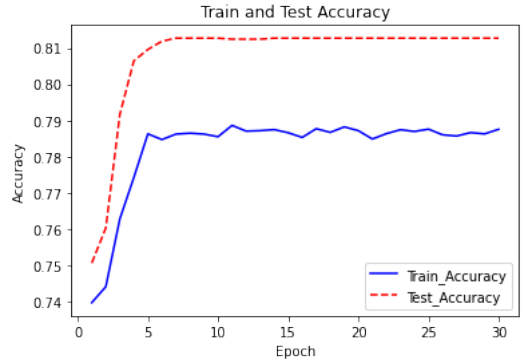

図 7　ViT EyePACS 精度

(3) CNN で APTOS を学習／診断した結果（図 8 と図 9）．損失と精度の両方に共通して，学習時に比べて診断時に値が EyePACS の場合ほどではないが上下にぶれながら推移している（所要時間：135 分 22 秒）．

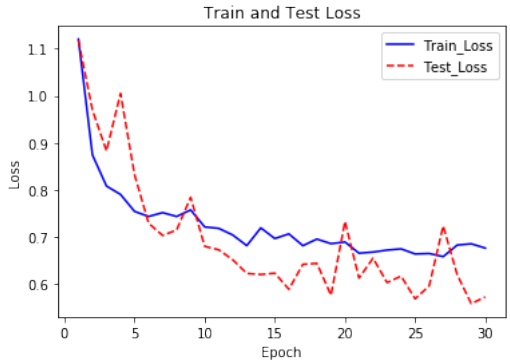

図 8　CNN APTOS 損失

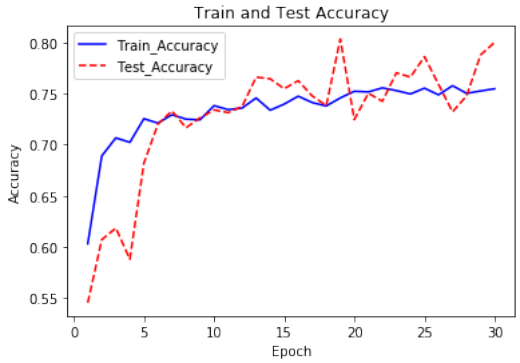

図 9　CNN APTOS 精度

(4) ViT で APTOS を学習／診断した結果（図 10 と図 11）．CNN の場合に比べて処理にかかった時間に EyePACS のときほどの大きな違いはないが、やはり 5 エポック辺りで値が収束し始めている（所要時間：188 分 40 秒）．5 エポックで学習を終了させると、処理にかかった時間は CNN の場合よりかなり短くなる．

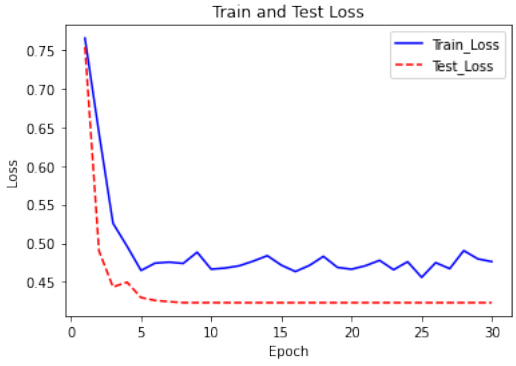

図 10　ViT APTOS 損失

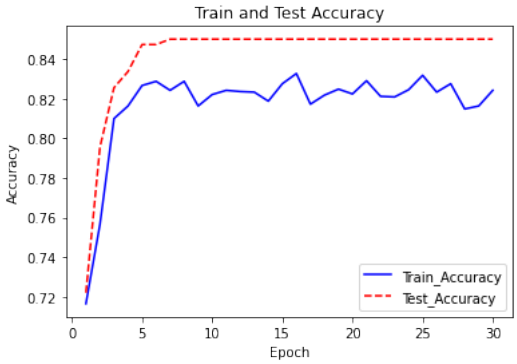

図 11 ViT APTOS 精度

(5) 今回のそれぞれの結果の最大値と所要時間を表 4 に示す.

表 4 結果の最大値と所要時間

	EyePACS (35126 件)		APTOS (3662 件)	
	CNN	ViT	CNN	ViT
学習時精度	0.7498	0.7886	0.7580	0.8328
診断時精度	0.7605	0.8127	0.8038	0.8501
学習時損失	0.7555	0.6409	0.6587	0.4560
診断時損失	0.7042	0.5884	0.5588	0.4233
所要時間	167’ 08”	897’ 43”	135’ 22”	188’ 40”

今回の精度比較で以下のことが示唆される.

- データ数が少ない場合（APTOS）も多い場合（EyePACS）も，CNN よりも ViT の方が損失が低く、精度が高い.
- CNN は処理時間がデータの量にあまり影響しないが、ViT は処理時間がデータの量に大きく影響する。
- 同じエポック数であれば、ViT の方が少ないエポック数で収束し始める.

5. 結論と課題

今回の比較検証だけで，画像の多クラス分類には CNN より ViT の方が有効であると結論づけるのは性急であるが，有効な場合があることは示せたと考える.

今後は，ViT 以外の分類手法との検証比較を継続して，よりよい分類手法を探求したい.

参考文献

[1] Ferguson, M., Ak, R., Lee, Y. and Law, K. H.: Automatic localization of casting defects with convolutional neural networks, *2017 IEEE International Conference on Big Data (Big Data)*, Los Alamitos, CA, USA, IEEE Computer Society, pp. 1726–1735 (online), DOI: 10.1109/BigData.2017.8258115 (2017).

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=YicbFdNTTy⟩ (2021).

[3] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y. and Tao, D.: A Survey on Visual Transformer (2021).

[4] kaggle: Diabetic Retinopathy Detection : Identify signs of diabetic retinopathy in eye images, (online), available from ⟨https://www.kaggle.com/c/diabetic-retinopathy-detection⟩.

[5] kaggle: APTOS 2019 Blindness Detection : Detect diabetic retinopathy to stop blindness before it’s too late, (online), available from ⟨https://www.kaggle.com/c/aptos2019-blindness-detection⟩.

[6] 村松千左子：眼底画像の診断のための画像解析手法，*Medical Imaging Technology*, Vol. 36, No. 4, pp. 181–188（オンライン），DOI: 10.11409/mit.36.181 (2018).

[7] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).