軽量 CNN と Vision Transformer の公平比較に向けた実験的評価

著者太郎
所属機関
連絡先: author@example.com

***Abstract*—本稿では，軽量な畳み込みニューラルネットワーク（CNN）と Vision Transformer（ViT）の画像分類性能を，CIFAR-10 を対象に公平に比較する実験プロトコルを提示する。学習率スケジューリング，データ拡張，Test-Time Augmentation（TTA）を統一的に適用し，再現性を重視した評価を実施した。実験の結果，ViT-Ti は TTA により最大で +1.8pt の精度向上を得た一方，ResNet-18 では RandAugment を中心としたデータ拡張で +2.1pt の改善が確認された。これらの知見を通じて，限られた計算資源でも堅牢なベースラインを構築するための指針を示す。**

***Index Terms*—画像分類, 深層学習, Vision Transformer, Test-Time Augmentation, CIFAR-10**

I. はじめに

画像分類タスクは，監視システムや医用画像解析，製造ライン検査など多様な応用領域で中核技術として利用されている。AlexNet によるブレイクスルー以降，深層学習モデルは大規模データセットでの性能向上を牽引し続けており，その性能差を適切に評価する手法が求められている。[1]

しかし，研究コミュニティにおける再現性の欠如や評価条件のばらつきは依然として課題である。特に，軽量な畳み込みニューラルネットワーク（CNN）と Vision Transformer (ViT) 系モデルの比較では，ハイパーパラメータや学習スケジュールが統一されないまま結果が報告されるケースが多く，公平な比較を阻害している。[2]

本研究では, ResNet-18 と DeiT-Ti を対象に, 軽量構成かつ短時間学習の条件下で性能を公平に評価するプロトコルを構築する。統一したデータ前処理，学習率スケジューリング，Test-Time Augmentation (TTA) を適用し，少数エポックでも再現性の高いベースラインを提供することが本稿の目的である。論文の構成は以下の通りである。第II章で関連研究を概観し，第III章で実験設定とプロトコルを示す。第IV章では実験結果を報告し，第V章で考察を述べた後，第VI章で結論と今後の課題をまとめる。

II. 関連研究

CNN 系モデルは，AlexNet による深層畳み込みの導入以降，VGG による深層化，ResNet による残差接続の導入などを通じて性能と学習安定性を向上させてきた。[1], [3], [4] ResNet-18 はモバイル環境でも利用可能な軽量モデルとして広く採用されており，本研究でも比較対象として採用する。

一方，自己注意に基づく Vision Transformer (ViT) は，画像をパッチに分割して系列として処理するアプローチであり，大規模事前学習を前提として高い性能を達成している。[5] ViT の軽量化を目的とした DeiT は，データ効率の良い学習戦略と蒸留手法を導入することで，小規模データセットでも競争力を示す。[6]

学習率スケジューリングや正則化の工夫は，モデルの汎化性能を支える重要な要素である。Cosine Annealing による学習率減衰は CNN で効率的な訓練を可能にし，[7] Label Smoothing や RandAugment は Transformer 系モデルでも精度向上に寄与する。[8], [9] しかし，これらのテクニックを一貫した条件で比較する研究は限られており，複数のモデルを同一基盤で評価するベースラインの整備が求められている。

III. 手法と実験設定

本章では, 評価対象モデル, データセットと前処理, 学習プロトコル，評価指標について説明する。実験環境は単一 GPU (NVIDIA RTX 3090) と 64GB RAM を想定し，すべての実験を同一ハードウェアで実施した。

A. 評価対象モデル

ResNet-18 は 11.7M パラメータを持つ残差ネットワークであり，Batch Normalization により勾配消失を抑制する。[4] Vision Transformer 系モデルとしては DeiT-Ti を採用し，5.7M パラメータで自己注意機構を実現する。[6] いずれのモデルもパラメータ初期化には He 初期化を使用し，最終分類層のみ学習率を 10 倍にスケーリングした。

B. データセットと前処理

CIFAR-10 は 32×32 ピクセルのカラー画像 60,000 枚で構成され，10 クラスを含む。全実験でトレーニング／検証比率は 45,000:5,000 とし，検証セットはクラス分布が均等になるよう分割した。前処理は以下の通りである。

- 標準化: 各チャネルに対しデータセット全体の平均と標準偏差で正規化。
- データ拡張: RandAugment (N=2, M=9) を適用し，Cutout (size=8) を併用。[9]
- Test-Time Augmentation: 水平反転と 10-crop 評価を実施。

C. 学習プロトコル

CNN では SGD + Momentum (0.9) を用い，初期学習率 0.1 を Cosine Annealing によりエポック 90 で 0.0001 まで減衰させた。ViT では AdamW を採用し，学習率 5e-4 と Weight Decay 0.05 を設定，5 エポックの線形ウォームアップ後に Cosine Annealing を適用した。バッチサイズは両モデルとも 128，トレーニングは 90 エポックで早期終了は行わない。Label Smoothing ($\varepsilon = 0.1$) を Transformer のみで有効化した。[8]

学習曲線図をここに挿入（例: `figures/learning_curve.pdf`）

Fig. 1. ResNet-18 と DeiT-Ti の学習曲線（Top-1 精度と損失）。

TABLE I
ベースライン設定での検証精度（Top-1）と推論時間。

モデル	精度 [%]	パラメータ [M]	推論時間 [ms]
ResNet-18	93.4	11.7	3.2
DeiT-Ti	92.7	5.7	5.6

D. 評価指標とログ

主指標は検証セットの Top-1 精度であり，補助指標として Top-5 精度，クロスエントロピー損失，推論時間（CPU/GPU）を記録した。各設定で異なる 5 個の乱数シードを用いて実験を実行し，平均と標準偏差を報告する。学習曲線と TTA の影響を可視化する図表の作成に向けて，エポック毎のログを統一フォーマットで保存した。

IV. 実験結果

本章では，学習曲線と最終精度，TTA の効果，シードごとの統計を順に報告する。全ての結果は検証セットに対する評価であり，基本設定（データ拡張あり，TTA なし）を基準とした差分も合わせて示す。

A. 学習曲線の比較

Figure 1 に CNN と ViT の損失および Top-1 精度の推移を示す。ResNet-18 は初期段階で急速に収束し，エポック 30 以降は安定して推移する。一方，DeiT-Ti は収束が緩やかであるものの，最終段階で精度を押し上げる傾向が確認された。

B. 最終精度と *TTA* の効果

Table I はベースライン設定での最終精度を，Table II は TTA の有無による精度変化と推論時間を比較した結果を示す。ResNet-18 は RandAugment の導入により +2.1 ポイントの改善が得られ，DeiT-Ti は TTA によって +1.8 ポイントの向上が確認された。

C. 再現性と分散の分析

Table III は異なるランダムシード 5 回の実験結果を集約したものである。ViT 系モデルでは分散が相対的に大きい一方，CNN は初期重みによる揺らぎが小さい傾向にある。これらの観察は，将来的な蒸留やデータ拡張の最適化に向けた指針となる。

V. 考察

実験結果を踏まえ，軽量 CNN と ViT の特性を考察する。ResNet-18 は初期収束が速く，短時間で実用的な精度を達成するため，限られた学習時間やオンライン更新が求められる環境に適している。一方，DeiT-Ti は最終精度が高く，TTA の恩恵を強く受けるが，推論時間が増加する点に留意する必要がある。

RandAugment に代表されるデータ拡張は CNN において顕著な性能向上をもたらすが，ViT では効果が飽和しやすい傾向が見られた。これは，自己注意機構が入力の多様性に対して比較的ロバストであるためと考えられる。TTA は両モデルで性能改善に寄与するが，特に ViT で顕著であり，推論コストとのトレードオフを踏まえた活用が求められる。

TABLE II
TTA の有無による精度差と推論時間の比較。

モデル	TTA	精度 [%]	推論時間 [ms]
ResNet-18	なし	93.4	3.2
	あり	93.9	5.8
DeiT-Ti	なし	92.7	5.6
	あり	94.5	10.4

TABLE III
ランダムシード 5 回の平均と標準偏差。

モデル	平均精度 [%]	標準偏差
ResNet-18	93.4	0.21
DeiT-Ti	93.6	0.47

本実験の制約として，CIFAR-10 のみを対象とした点，大規模事前学習や蒸留の効果を評価していない点，ハードウェアコストの詳細な分析が未実施である点が挙げられる。今後はより大規模なデータセット（例: ImageNet-1k）や半教師あり学習との組み合わせを検討し，軽量モデルの性能限界を明らかにしたい。

VI. 結論

本稿では，ResNet-18 と DeiT-Ti を対象に，軽量構成での画像分類性能を公平に比較する実験プロトコルを提案した。統一したデータ前処理，学習率スケジューリング，TTA を組み合わせることで，再現性の高いベースラインを構築できることを示した。ResNet-18 はデータ拡張により大きな性能向上を示し，DeiT-Ti は TTA によって精度を伸ばせることから，利用目的に応じて両者を使い分ける指針を得た。

今後の研究では，大規模データセットや蒸留手法，ハードウェア最適化といった要素を組み合わせ，軽量モデルの性能限界と運用上の最適化を探る予定である。また，公開ベンチマークにおける再現性確保のため，コードと設定ファイルをオープンソースとして提供することを計画している。

謝辞

本研究は，オープンソースコミュニティおよび深層学習フレームワーク（PyTorch など）の開発者の貢献に支えられている。ここに謝意を表する。また，実験環境の整備に協力いただいた研究室メンバーにも感謝する。

付録 A 実験ハイパーパラメータ

Table IV に主要なハイパーパラメータをまとめる。追加の設定や学習ログは公開予定のリポジトリで提供する。

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, 2012.

TABLE IV
主要ハイパーパラメータ一覧。

項目	ResNet-18	DeiT-Ti
初期学習率	0.1	5e-4
最適化手法	SGD+Momentum	AdamW
エポック数	90	90
バッチサイズ	128	128
学習率スケジュール	Cosine	Cosine+Warmup
Label Smoothing	なし	0.1

[2] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, “Deep reinforcement learning that matters,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3207–3214, 2018.

[3] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *International Conference on Learning Representations*, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.

[6] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, “Training data-efficient image transformers (deit),” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1037–1047, 2021.

[7] I. Loshchilov and F. Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” in *International Conference on Learning Representations*, 2017.

[8] R. Müller, S. Kornblith, and G. Hinton, “Does label smoothing really improve model calibration?,” in *Advances in Neural Information Processing Systems*, pp. 4696–4705, 2019.

[9] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “Randaugment: Practical automated data augmentation with a reduced search space,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.