Vision Transformer の相互学習による高精度化

ER18038 鈴木渉起　　指導教授：藤吉弘亘

1. はじめに

共同学習は，複数のネットワークを用いて学習中に互いの知識を伝達して，ネットワークの性能を向上させる学習法である [1, 2]．これまでの共同学習では，畳み込みニューラルネットワーク (CNN) を対象としていたが，Transformer を対象とした手法も提案されつつある [3]．Vision Transformer (ViT) は，Transformer の構造を直接適用しているため，幅広い応用性がある．しかし，高性能なモデルを得るには膨大な学習データが必要になる．

本研究では，ViT の学習に膨大なデータセットが必要になるという問題を解決するために，相互学習による高精度化するアプローチを提案する．

2. Data-efficient image Transformers (DeiT)

DeiT [3] は，CNN を Teacher，ViT を Student とした Knowledge Distillation (KD) である．DeiT では，CNN の知識を ViT に伝達することで，少ない学習データで高性能な ViT モデルとなる．図 1 にネットワーク構造を示す．DeiT は，ViT に Teacher からの知識を伝達するための Distillation Token が追加されている．学習時，Class Token は Student の推定クラスと教師ラベルとの損失を計算する．Distillation Token では，Teacher の推定クラスを教師ラベルとして，Student の推定クラスとの損失を計算する．ネットワークの出力は 2 つあるため，評価時には，2 つの出力の平均を最終的な推論結果とする．

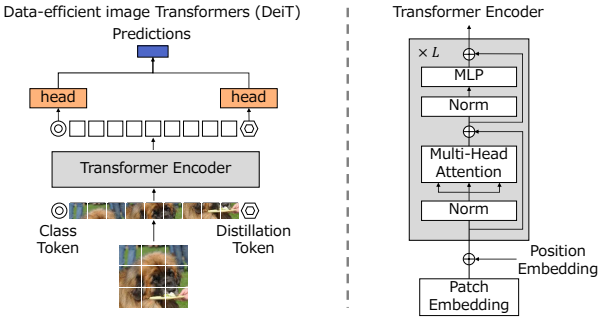


図 1: DeiT のネットワーク構造

3. 提案手法

ViT を少ないデータで高精度化することを目的として，Deep Mutual Learning (DML) によるアプローチを提案する．図 2 に学習方法の概略図を示す．2 つの Student ネットワークを用いた相互学習では，両方のネットワークに同じ画像を入力し，各ネットワークの損失を算出する．式 (1) に Student1 の損失 L_1 を示す．ここで，L_{CE_1} は Student1 の確率分布 p_1 と教師ラベルの Cross Entropy，$D_{KL}(p_2\|p_1)$ は Student2 の確率分布 p_2 との KL divergence である．Student2 も式 (2) により同様に損失 L_2 を算出する．

$$L_1 = L_{CE_1} + D_{KL}(p_2\|p_1) \tag{1}$$

$$L_2 = L_{CE_2} + D_{KL}(p_1\|p_2) \tag{2}$$

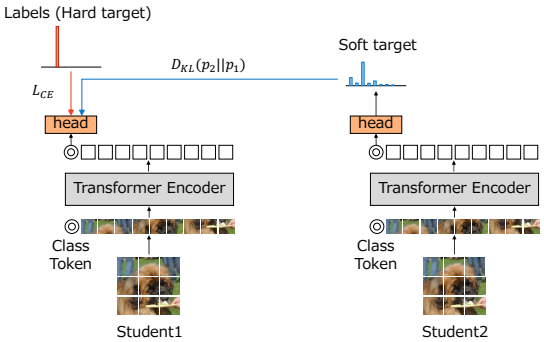


図 2: ViT の相互学習 (Student1 を更新)

4. 評価実験

DML で学習した ViT の性能を評価するために，教師ラベルのみで単体学習した ViT との精度を比較する．また，KD で学習している DeiT は Distillation Token を導入している．そこで，DML における Distillation Token の必要性を調査するために，DeiT 同士による DML の精度比較も行う．この時，DeiT の推論結果は Distillation Token の出力に対して損失を求める．

4.1. 実験条件

ネットワークは，Tiny モデルと Small モデルを用いる．データセットには ImageNet-1k を用いる．エポック数は 300，バッチサイズは 512，最適化手法は AdamW，重み減衰は 0.05 である．学習率は cosine decay によりスケジューリングする．このとき最大学習率は 0.005，Warm-up epochs は 5 である．実験の試行回数は，各条件に対して 1 回である．評価指標は Top-1 Accuracy である．

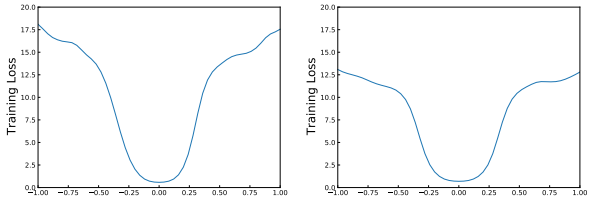
4.2. 実験結果

DML により学習した ViT と DeiT の精度を表 1 に示す．単体学習と DML の精度を比較すると，ViT と DeiT ともに精度向上していることが確認できる．また DML で学習した ViT と DeiT を比較すると，ViT の方が高精度である．この結果から，Distillation Token を使用せず，1 つの Token に対して損失を求めることが精度向上に寄与すると考えられる．

表 1: ViT による DML の精度 [%]

学習方法	Tiny/Tiny	Tiny/Small	Small/Small
Independent	65.8/65.8	65.8/66.0	66.0/66.0
DML (ViT)	**71.6/71.8**	**69.2/73.0**	**70.5/70.5**
DML (DeiT)	69.1/69.2	67.8/70.7	68.9/69.0

図 3 に，学習方法による ViT の誤差局面の比較を示す．誤差局面は，正規分布からサンプリングした行列を学習済みモデルの重みに加えて求める．横軸はノイズの大きさに相当し，縦軸は学習データに対する誤差である．図 3(a) の単体で学習した ViT よりも図 3(b) の DML で学習した ViT の方が，ノイズの大きさに対する誤差の上昇幅が小さいことがわかる．従って，DML で学習した ViT は，広い谷に収束することでより高い汎化能力を獲得したといえる．



(a) 単体で学習した ViT　　(b) DML で学習した ViT

図 3: 学習方法による ViT の誤差局面の比較

5. おわりに

本研究では，Vision Transformer による相互学習を行った．評価実験では，ViT 同士の相互学習により高精度化できることを確認した．また誤差局面を求めることで，DML で学習した ViT が単体で学習した ViT に比べて精度向上している要因を示した．今後は，知識転移グラフへの応用や Transformer が獲得した特徴量を用いた共同学習方法の考案を目指す．

参考文献

[1] G. Hinton, *et al.*, “Distilling the Knowledge in a Neural Network”, NeurIPS Workshop, 2015.

[2] Y. Zhang, *et al.*, “Deep Mutual Learning”, CVPR, 2018.

[3] H. Touvron, *et al.*, “Training data-efficient image transformers & distillation through attention”, ICML, 2021.