2U-02

Vision Transformerを用いた画像分類モデルのアテンション機構の軽量化

Lightweight Attention Module of Image Classification Model Using Vision Transformer

川井 智隆† 吉田 明正†
Tomotaka Kawai Akimasa Yoshida

1 はじめに

ディープラーニングの世界では画像識別のモデルで一般的なものはCNNベースのモデルである．しかし，近年Transformerモデルの隆盛により画像識別分野においてもVision TransformerというTransformerベースの画像識別モデルが現れた[1]．実際にVision Transformerは高性能であり，また，物体検出，セマンティックセグメンテーションといったタスクなどにも流用されている．しかし，その性能を発揮するためには計算機に対し，高い計算能力やランタイムメモリを要求する．実運用の際，上記のようなハードルを改善するためにモデル圧縮技術がある．

本論文ではVisionTransformerモデルの各層におけるチャネルの重要性を二つの観点から求め，より教師モデルのチャネルの分布の特性を引き継ぎながらチャネルの減量するモデル圧縮技術を提案する．提案手法の手順は以下の通りである．(1) 各チャネルの重要度スコアを2種類の知識蒸留の手法を用いて計算する．(2) 計算した2種類の重要度スコアを元に，各チャネルの安定性を求める．(3) 求めた安定性を基に教師モデルを刈り込む．提案手法の有効性を示すために，CIFAR-10データセットで学習，推論し，削減されたパラメータと精度，推論時間の関係を示す．

2 VisionTransformer

本稿で利用する画像識別モデルと，一般的なモデル圧縮技術について説明する．

2.1 Vision Transformer モデル

Vision Transformer[1]はTransformerベースの画像分類モデルのことである．元々，Transformerは自然言語処理のタスクにおいてデファクトスタンダードの存在であり，そのTransformerを画像分類タスク用にしたものである．具体的には入力画像をいくつかのバッチに分割し，さらに位置情報(元の画像でどこにいたか)を追加したものを1つのトークンとし，TransformerのEncoderに与え，クラス分類を行う．Transformerベースのモデルは画像分類に限った話ではなく，物体検出タスク[2]，セマンティックセグメンテーションタスク[3]など様々なタスクに応用されている．また，Encoder部分のアーキテクチャはTransformerと同じく，Attention機構というモジュールを何層も重ねている．本稿で利用するVisionTransformerモデルは文献[4]を参考にし，512次元のチャネルを6層重ねるAttention機構を採用している．

†明治大学大学院先端数理科学ネットワークデザイン専攻
Department of Network Design, Graduate School of Advanced Mathematical Sciences, Meiji University

2.2 モデル圧縮技術

機械学習のモデルを実社会運用する際に，モデルを実行するデバイスの性能に制限があることも多い．特にエッジデバイスでモデルを実行する際には，ランタイムメモリや計算資源の制限から特定のモデルが使用できない，十分な性能を発揮できない場面も考えられる．その際には利用するモデルを圧縮することで計算量やメモリ使用量を削減しエッジデバイスでの複雑なモデルの実行を実現する．

代表的なモデル圧縮技術として，枝刈り，知識蒸留，量子化がある．枝刈りはモデルのノードや重みを削除することでモデル内のパラメータ数を削除する手法である．知識蒸留は，より利用したいモデルのパラメータの確率分布を模倣したより小さいモデルを生成することを目的とした手法である．量子化はモデルのパラメータの精度を落とすことでメモリ使用量を下げる手法である．本稿では枝刈りをメインの圧縮技術として用いる．

3 重みと重要度スコアを考慮した枝刈り手法

本節ではVisionTransformerモデル[4]を枝刈りする手法について述べる．

3.1 重みによる枝刈り

重みによる枝刈りは，VisionTransformerの各Attention機構の最終層のチャネルの重みの絶対値を小さいチャネルから刈り取る[5]．具体的には図1において，Attention機構1層目の各チャネルの重みをx軸に，後述する重要度スコアをy軸にし，重みによる枝刈りで50%のチャネルを刈り取った分布図である．青点が枝刈りによって削られたチャネルである．重みの大小を基準にするため，左側半分が削られていることがわかる．

3.2 重要度スコアによる枝刈り

モデル圧縮技術である知識蒸留を利用し，VisionTransformerモデル[4]の各Attention機構の各チャネルの重要度を算出し，その値を基に刈り込む．重要度は教師モデルと生徒モデルの蒸留損失を利用する．損失蒸留を式(1)に定義する[6]．

$$L = L_{CE}(y, p) + \alpha L_{KL}(q, p) \tag{1}$$

L_{CE}はクロスエントロピー誤差，yは正解ラベル，pは生徒モデルの出力確率を表している．また，L_{KL}はカルバック・ライブラー情報量を表しており，qは教師モデルの出力確率を表す．また，αの値は文献[6]を参照している．

重要度スコアによる枝刈りは具体的には図2のようになる．図2はAttention機構1層目の各チャネルの重みをx軸に，後述する重要度スコアをy軸にし，重要度スコアによる枝刈りで50%のチャネルを刈り取った分布図である．青点が枝刈りによって削られたチャネルである．重要度による枝刈りの手法は，教師モデルのチャネルの分布と生徒モデルのチャネルの分布との差を最小にする

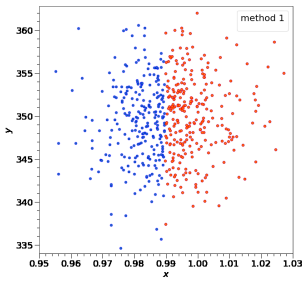

図 1　重みによる枝刈り (50%).

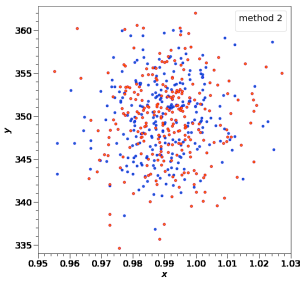

図 2　重要度スコアによる枝刈り (50%).

ように式 (1) で定義した重要度スコアを利用して，刈り取るチャネルを選択する．そのため，図 1 とは違い，刈り取る対象の青点は刈り取る前のチャネルの分布に近くなるように選択されている．

3.3　刈り込みチャネル対象の範囲

Transformer のアーキテクチャ上，複数の Attention 層が存在するため，各 Attention 層毎に刈り込むチャネルを選択するか，まとめて刈り込むチャネルを選択するかの 2 通りが考えられる．

例えば，本稿で利用する 512 チャネルを持つ Attention 機構が 6 層に重なった Transformer モジュールを刈り込む場合，512×6 の 3072 種類のスコアで刈り込むチャネルを選択する．仮に 40%のチャネルを刈り込む場合，3072 個全部のスコアを同時に比較すると，各層の Attention 機構のチャネル数は刈り込み後は必ずしも一致しない．このチャネル範囲の選択方法を均一と呼ぶ．

対照的に，各 Attention 機構の層毎，つまり 512 チャネル毎に 40%の刈り込みを行う場合，各層の Attention 機構のチャネル数は等しく 40%ずつ刈り込まれる．このチャネル範囲の選択方法を全体と呼ぶ．上述の二つの刈り込みチャネル対象の選択範囲の方法も重みと重要度スコアによる枝刈りに対して適用する．

3.4　提案手法

上述の手法二つの手法を利用し，より効果的な重要度スコアによる枝刈り手法を提案する．式 (1) では係数 α を利用して L_{CE} と L_{KL} の両方を利用している．本手法ではこの点に着目し，二つの蒸留損失のより良い利用方法の観点として安定性を考える．具体的には L_{CE} と L_{KL} は各チャネルごとに計算をしているため，二つの蒸留損失は同じ程度の重要度を示すべきと考える．そこで，どちらかの蒸留損失が示す重要度に差があるほど安定性が低いと考え，これを重要度スコアに反映するために L_{CE} と L_{KL} にて単回帰分析を行い，各チャネルの推定値と実際値の誤差で重要度スコアを調整する．これらを式 (2) として定義する．$\hat{L}_{KL}(q,p)$ は単回帰分析で出力された $L_{KL}(q,p)$ の予測値である．

$$L = L_{CE}(y,p) - |\hat{L}_{KL}(q,p) - L_{KL}(q,p)| \quad (2)$$

4　VisionTransformer モデル軽量化の性能評価

本節では，VisionTransformer モデルの Attention 機構に対する重みによる枝刈りと重要度スコアによる枝刈りを適用し，刈り込み率とモデルの精度，推論時間，残パラメータ量の推移を確認する．

4.1　性能評価環境

データセットには CIFAR10 を利用し，ベースの VisionTransformer モデルには文献 [4] を利用する．提案手法の評価においては，Intel Xeon 2265 3.50GHz 12 コア，128GB メモリ，NVIDIA Quadro RTX6000(学習と枝刈りに使用) を使用した．

表 1　枝刈り手法と刈り込み割合と精度.

枝刈り手法	刈り込み割合			
	30%	40%	50%	60%
重み (均一)	73.73	66.37	56.79	46.95
重み (全体)	73.51	68.52	63.11	51.53
重要度 (均一)	71.43	64.71	56.67	46.28
重要度 (全体)	70.45	64.45	55.89	46.19
提案手法	74.70	68.77	58.13	43.32

(注) 単位は [%].

表 2　枝刈り手法と刈り込み割合と推論時間.

枝刈り手法	刈り込み割合			
	30%	40%	50%	60%
重み (均一)	7588.70	7756.06	7420.79	7233.90
重み (全体)	7427.59	7250.08	7170.44	7016.61
重要度 (均一)	7598.22	7358.15	7192.04	7088.89
重要度 (全体)	7440.94	7193.14	7107.77	6998.88
提案手法	7164.36	6969.38	6809.11	6704.02

(注) 単位は [ms]，元モデルの推論時間は 8253.53[ms].

4.2　精度と推論時間の評価

表 1 と表 2 では文献 [4] のモデルを 30–60%の範囲で各手法と範囲で Attention 機構部分を枝刈りした結果を示す．表 1 の精度は Top-1 Accuracy とする．表 2 の推論時間は画像 1000 枚を CPU で推論するのに要する時間である．結果としては，精度の面では提案手法が刈り込み割合が 50%未満では優位であり，50%を超えると重み (全体) の枝刈り手法が優位となった．推論時間に関しては提案手法が 30–60%の範囲で優位となった．

5　おわりに

本稿では VisionTransformer モデルに対する 2 種類の枝刈り手法を比較した．CIFAR-10 のデータセットにおける学習済モデルの枝刈りの結果，提案手法が精度，推論時間の面で 50%までの枝刈りの場合に優位な結果が確認できた．これらの結果から，提案する枝刈り手法の有効性が確認された．

参考文献

[1] Dosovitskiy, Alexey and Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010. 11929, 2020.

[2] Carion, Nicolas and Massa, et al. End-to-end object detection with transformers, European conference on computer vision Springer, pp. 213–229, 2020

[3] Strudel, Robin and Garcia, et al. Segmenter: Transformer for semantic segmentation, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7262–7272, 2021.

[4] Phil Wang. vit-pytorch, https://github.com/lucidrains/vit-pytorch/blob/main/vit_pytorch/vit.py, 2022.

[5] Zhu, Mingjian and Tang, et al. Vision transformer pruning, arXiv preprint arXiv:2104. 08500, 2021.

[6] Yu, Hao and Wu, Jianxin. A unified pruning framework for vision transformers, arXiv preprint arXiv:2111. 15127, 2021.