大規模事前学習済みモデルの知識維持と下流タスク適応を両立した枝刈りに関する研究

TP23014　新田 常顧　　指導教授：山下 隆義

1. はじめに

事前学習済みの大規模モデルは，モデルパラメータを下流タスクに適応させるための微調整を行うことで，様々な下流タスクで高い性能を発揮する．大規模モデルはライブラリ等から簡単に入手できるが，微調整を行うには大規模な計算資源を必要とする．そのため，大規模モデルを微調整する前に軽量化する枝刈りが注目されている．枝刈り手法として代表的な SNIP [1] は，損失関数に対する各重みの勾配を基に評価値を求め，枝刈り対象となる重みを決定する．しかし，SNIP による枝刈り後のモデルは下流タスクへの適応性を考慮できるものの，事前学習によって既に収束した重みの評価値が低くなる．そのため，枝刈り後の性能は劣化する．

そこで本研究では，事前学習済みモデルが持つ下流タスクに有用な重みを維持する枝刈り手法を提案する．具体的には，事前学習済みモデルの各層の出力特徴量を維持しながら，下流タスクの損失に対する勾配を考慮する．これにより，事前学習で得た知識の維持と，下流タスクへの適応性の考慮を両立する．実験により，様々な下流タスクのデータセットに対して高い性能を発揮することを示す．

2. 関連研究

枝刈りを行う時間的，計算的コストを削減するために，初期状態のモデルを学習する前に枝刈りを行う Single-Shot Network Pruning (SNIP) [1] が提案されている．SNIP は，損失関数 $\mathcal{L}$ に対する i 番目の重み w_i の勾配を用いて重みの評価値を決定する手法であり，式 (1) のように重みの評価値 $S_{\mathrm{SNIP}}(w_i)$ を算出する．

$$S_{\mathrm{SNIP}}\left(w_i\right)=\left|\frac{\partial \mathcal{L}}{\partial w_i} w_i\right| \tag{1}$$

SNIP は，損失関数に対する重みの勾配の大きさをその重みの評価値として定義するため，損失を下げるために貢献しやすい重みを高く評価し，貢献しにくい重みを低く評価する．これにより，タスクへの適応力を維持しながら枝刈りすることが可能である．

しかし，SNIP は事前学習済みモデルに適用することを想定していない．例えば，豊富な知識を持った事前学習済みモデルは，下流タスクに対して更新する必要のない，既に収束している重みが多くあると考えられる．収束している重みに対する損失関数の勾配は小さくなるため，SNIP はこれらを低く評価し，枝刈りしてしまう可能性がある．これは，事前学習済みモデルを下流タスクで微調整する際に，その近傍で最適化を促すことが有効であるという従来研究の知見 [2] に反することになる．

3. 提案手法

本研究では，事前学習で得た知識の維持を目的とした枝刈り手法 Retaining Feature Representation (ReFer) を提案する．さらに，ReFer による知識の維持と下流タスクへの適応力の維持を目的とした枝刈り手法 Adaptive Feature Retention (AFR) を提案する．

3.1 Retaining Feature Representation

SNIP を事前学習済みモデルに適用すると，既に下流タスクに対して収束している重みを過小評価する可能性がある．そこで，枝刈り後に事前学習で得た知識を維持することを目的とした Retaining Feature Representation (ReFer) を提案する．例えば，ある層 l が出力する N 次元の特徴量 $\boldsymbol{F}^l$ を事前学習済みモデルが持つ知識として定義し，その変化を抑制するような枝刈りを考える．

まず，事前学習済みモデルの各層から出力される特徴量がどれほど知識を保持し，またその特徴量の分散がどれほど大きいかを評価し，枝刈り後にそれを維持することを考える．例えば，特徴量行列 $\boldsymbol{F}$ がサンプル数 N を行，特徴量次元 D を列として持つ場合，その情報量や分散を求める方法として特異値分解がある．特異値分解では，行列 $\boldsymbol{F}$ を $\boldsymbol{F}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}}$ の形で分解し，特異値 $\boldsymbol{\Sigma}$ の大きさがデータの分散や情報量を示す．このことから，特異値が大きいほど，対応する特徴量はデータにおいて重要な意味を持つと考えられる．実際に，学習済みモデルの出力特徴量の広がりを評価する方法として，特異値分解による特異値を使用する研究も行われている [3]．そこで，式 (2) に示すような特異値 $\boldsymbol{\Sigma}=\{\sigma_1,\sigma_2,...,\sigma_N\}$ の平均を維持するような重みの評価基準 S_{ReFer} を導出する．

$$\begin{aligned}\boldsymbol{F}_{\mathrm{svd}}&=\frac{1}{N}\sum_{i=1}^{N}\sigma_i\\ S_{\mathrm{ReFer}}(w_i)&=\left|\frac{\partial\sum_{l=1}^{L}\boldsymbol{F}^l_{\mathrm{svd}}}{\partial w_i}w_i\right|\end{aligned} \tag{2}$$

これにより，各層が抽出するべき特徴の情報量とその広がりを，事前学習で得た知識として維持することが可能となる．

3.2 Adaptive Feature Retaining

事前学習で得た知識を維持する ReFer は，下流タスクへの適応力を失う可能性がある．例えば，微調整戦略の分野では，事前学習済みモデルを下流タスクへ適応させる際，更新の必要性が低い重みは凍結し，更新の必要性が高い重みのみを微調整することで，ロバスト性が高まるということが知られている [4] ．しかし，枝刈りは微調整戦略とは異なり，枝刈りする重みを決定する必要がある．そのため，ReFer は事前学習済みモデルが持つ知識の維持のみを目的とすることで，下流タスクへ適応するために更新する必要性が高い重みを過小評価し，枝刈りしてしまう可能性がある．このような場合，枝刈り後のモデルは，事前学習で得た知識として評価された重みのみで下流タスクへの最適化を行うことになる．これは，下流タスクのデータの分布が事前学習で用いたデータの分布と異なるほど，枝刈り後のモデルの最適化を困難にする可能性がある．これらのことから，事前学習で得た知識の維持のみではなく，枝刈り後に下流タスクへ適応する際に，更新が必要とされる重みを同時に高く評価する必要がある．

そこで，事前学習で得た知識の維持と，下流タスク適応を両立する Adaptive Feature Retaining (AFR) を提案する．下流タスクへの適応に必要な重みは，損失関数に対する重みの勾配の大きさで算出することが可能である．これは，勾配の大きさが大きいほど，下流タスク損失を下げるために効果的であるからである．AFR では，事前学習で得た知識を維持可能な ReFer と，下流タスクへの適応を考慮する SNIP を混合した評価基準を導出する．まず，ReFer と SNIP の評価基準を等価に扱えるようにするために，それぞれの算出した評価値を式 (3) に示すように標準化する．ここで，Standardize($S(w_i)$) は，重み w_i に対応する評価値 $S(w_i)$ を標準化する操作を表す．

$$\begin{aligned}\tilde{S}_{\mathrm{ReFer}}(w_i)&=\mathrm{Standardize}\left(S_{\mathrm{ReFer}}(w_i)\right),\\ \tilde{S}_{\mathrm{SNIP}}(w_i)&=\mathrm{Standardize}\left(S_{\mathrm{SNIP}}(w_i)\right)\end{aligned} \tag{3}$$

次に，標準化後の評価値を加算することで $S_{\mathrm{AFR}}(w_i)$ を導出する．

$$S_{\mathrm{AFR}}(w_i)=\tilde{S}_{\mathrm{ReFer}}(w_i)+\tilde{S}_{\mathrm{SNIP}}(w_i) \tag{4}$$

これにより，AFR では，事前学習で得た知識の維持と，下流タスク適応を両立することが可能である．

表 1: 特定物体認識タスクでの分類精度比較 [%]

データセット	(a) 教師あり事前学習			(b) 自己教師あり事前学習		
	SNIP	AFR	精度差	SNIP	AFR	精度差
Diseases	**99.58**	99.57	-0.01	99.49	99.55	**+0.06**
EuroSAT	97.77	**97.92**	+0.15	97.64	97.88	**+0.24**
Caltech-101	76.95	81.10	+4.15	80.01	**86.17**	**+6.16**
ISIC	79.93	80.87	+0.94	80.12	**80.92**	+0.80
Pets	52.60	66.85	+14.35	66.91	**80.18**	+13.27
ChestX	45.45	**47.46**	+2.01	43.59	47.27	**+3.68**
DTD	34.62	39.57	+4.95	41.17	**52.50**	**+11.33**
Cars	29.83	48.92	+19.09	30.92	**51.87**	**+20.95**
SUN397	30.39	36.10	+5.71	35.35	**44.15**	**+8.80**
Aircraft	18.21	26.55	+8.34	24.09	**27.03**	+2.94

表 2: 一般物体認識タスクでの分類精度比較 [%]

データセット	枝刈り手法	枝刈り率		
		90%	95%	98%
CIFAR-10	SNIP	93.82	92.19	89.04
	ReFer	94.09	93.02	89.89
	AFR	**94.89**	**93.74**	**90.00**
CIFAR-100	SNIP	74.40	69.55	59.78
	ReFer	74.62	70.09	63.58
	AFR	**76.22**	**72.90**	**64.50**
Tiny-ImageNet	SNIP	61.99	55.32	46.26
	ReFer	60.73	54.78	47.55
	AFR	**64.05**	**58.83**	**50.24**

4. 評価実験

ReFer，AFR の有効性を検証するために，評価実験を行う．

4.1 実験条件

モデルには ImageNet-1k データセットで事前学習された Vision Transformer のベースモデル (ViT-B/16) を使用する．下流タスクには，一般物体認識タスクとして CIFAR-10, CIFAR-100, Tiny-ImageNet データセット，特定物体認識タスクとして Aircraft，Caltech-101 等の 10 個のデータセットを用いて評価を行う．枝刈り率は，一般物体認識タスクで {90%, 95%, 98%} を割り当て，特定物体認識タスクで 80%とする．評価方法は分類精度 [%]，比較手法は SNIP，ReFer，AFR とする．

4.2 一般物体認識タスク

教師あり事前学習した ViT-B/16 に対して，一般物体認識タスクのデータセットで枝刈りした結果を表 2 に示す．ここで，表中の太字は最も精度が高いこと，下線は 2 番目に精度が高いことを意味する．また，精度差の列は，SNIP と AFR の精度の差を意味する．

表 2 より，全ての条件において AFR の精度が高いことが分かる．一方で ReFer は，CIFAR に対して SNIP よりも精度が優れているものの，Tiny-ImageNet ではその効果を充分に発揮できていないことが分かる．これは，Tiny-ImageNet が CIFAR と比較して難しいタスクであり，タスク適応に更新が必要な重みが多く含まれるからだと考える．これは，事前学習で得た知識の維持だけでなく，下流タスクに適応するための重みを考慮する AFR の有効性を裏付ける結果である．

4.3 特定物体認識タスク

教師あり事前学習した ViT-B/16 に対して，特定物体認識タスクのデータセットで枝刈りした結果を表 1(a) に示す．表 1(a) より，Diseases を除き，AFR が最も高精度であることがわかる．特に，Cars データセットにおいては，SNIP と比較して 19.09pt の精度向上を確認した．

4.4 事前学習モデルの違いによる効果

DINO-v1 による自己教師あり事前学習した ViT-B/16 に対して，特定物体認識タスクのデータセットで枝刈りした結果を表 1(b) に示す．ここで，精度差の列の太字は，表 1(a) と比較して精度差が大きいものを意味する．表 1(b) より，全ての条件で AFR が最も高精度であることがわかる．また，SNIP との精度の差は，教師あり事前学習したモデルよりも，自己教師あり事前学習したモデルの方が，ISIC，Pets，Aircraft を除いた 7 つのデータセットで大きくなっていることがわかる．これは，自己教師あり事前学習はラベルに依存しない特徴を学習するため，モデルがよりタスクに依存しない一般的な特徴を学習しているからであると考える．特定物体認識において，一般的な特徴は，対象物の多様な形状やパターンを捉える上で有用であり，事前学習で得た知識の維持を考慮する AFR がさらに効果を発揮したと考えられる．この結果から，大規模事前学習済みモデルの枝刈りにおいて，下流タスクに合わせたモデルの選択も重要であると言える．

5. おわりに

本研究では，事前学習で得た知識を維持することを目的とした枝刈り手法の ReFer と，知識の維持のみではなく，下流タスクへの適応を考慮する枝刈り手法の AFR を提案した．AFR は，モデルの各層の出力特徴に対する特異値の平均を維持する ReFer と，下流タスク損失に対する勾配を考慮した SNIP を混合することで重みの評価をすることが可能である．評価実験より，一般物体認識タスク，特定物体認識タスクでの分類精度の向上を実現した．今後は事前学習モデルの変更や下流タスクの変更を行い，AFR の有効性を調査する．

参考文献

[1] N. Lee *et al.*, “SNIP: Single-shot Network Pruning based on Connection Sensitivity”, ICLR, 2019.

[2] J. Tian, *et al.*, “Trainable Projected Gradient Method for Robust Fine-Tuning.”, CVPR, 2023.

[3] N. Park, *et al.*, “What Do Self-Supervised Vision Transformers Learn?”, ICLR, 2023.

[4] G. Li, *et al.*, “Robustness Preserving Fine-tuning using Neuron Importance”, ECCV, 2024.

研究業績

[1] 新田常顧 等, “事前学習モデルの特徴表現を維持した single-shot Foresight pruning”, 画像の認識・理解シンポジウム, 2024.
(他 4 件)