SVD を用いた LoRA と段階的知識蒸留による効率的なモデル圧縮

EP22061 熊澤綱佑　　指導教授：山下隆義

1. はじめに

Vision Transformer（ViT）は画像認識分野を代表する深層学習モデルの 1 つであり，高い認識性能を示す．その一方で，ViT はパラメータ数が多く，学習・推論時の計算量やメモリ消費が大きいという課題がある．この課題に対し，LoRA による学習パラメータ削減と知識蒸留によるモデル圧縮を組み合わせた手法として WeCoLoRA が提案されている．WeCoLoRA は，生徒モデル構築時に中間層を間引くため，中間層が持つ重要な特徴表現が失われる可能性がある．そこで本研究では，間引かれた層の知識を LoRA の初期化時に効果的に活用する．さらに，段階的な知識蒸留により，教師モデルの間引かれた層の特徴表現を継承した生徒モデルを構築する手法を提案する．

2. WeCoLoRA

WeCoLoRA[1] は，図 1 に示すように，事前学習済み ViT を教師モデルとし，一部の層を間引いて生徒モデルを構築する知識蒸留手法である．WeCoLoRA では，間引き率 r に基づいて教師モデルの層を等間隔に選択し，その重みをコピーして生徒モデルを構築する．構築後，生徒モデルの Transformer block 全体に LoRA を適用し，教師モデルとの知識蒸留を行う．LoRA の低ランク行列はランダムに初期化され，学習を通じて更新される．損失関数を式 (1) に示す．

$$\mathcal{L}\left(E_i^T, E_i^S\right) = \left|E_i^T - E_i^S\right| \tag{1}$$

ここで，E_i^T は教師の特徴ベクトル，E_i^S は生徒の特徴ベクトルを表す．

WeCoLoRA は，間引いた層の重みを蒸留や初期化に利用しないため，教師モデルが中間層で獲得した特徴表現を生徒モデルに十分に継承できないという課題がある．

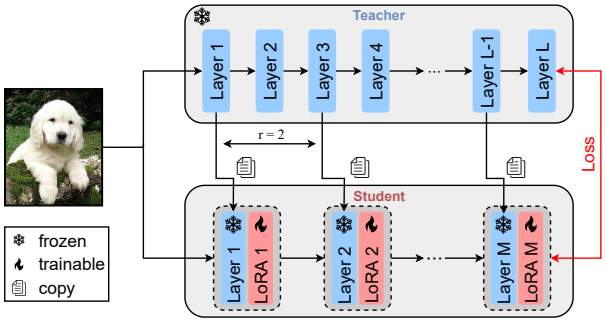

図 1：WeCoLoRA の概要

3. 提案手法

提案手法の概要を図 2 に示す．本研究では，間引いた層の知識を活用するために，その重み行列に SVD を適用して LoRA のパラメータの初期値に活用する．また，LoRA を浅い層から深い層へ段階的に適用することで，教師モデルの中間出力を再現する段階的蒸留手法を提案する．

3.1. SVD を用いた LoRA の初期化

生徒モデル構築時に，間引く層の重み行列に SVD を適用し，主要成分を LoRA の初期値として利用する．重み行列 W は式 (2) のように分解できる．

$$W = U\Sigma V^\top \tag{2}$$

ここで，U は出力空間，V は入力空間における基底ベクトルを表す．特異値の大きい成分が主要情報を担うことから，Σ の上位 k 成分を用いて低ランク近似を行う．低ランク行列 A, B を式 (3) に示す．

$$A = \sqrt{\Sigma_k}\, V_k^\top, \qquad B = U_k \sqrt{\Sigma_k} \tag{3}$$

3.2. 知識蒸留を用いた段階的 LoRA チューニング

間引いて構築した生徒モデルに対し，教師モデルの中間出力を用いた段階的蒸留を行う．段階的蒸留では，生徒モデルの浅い層から順に蒸留対象層を追加し，対応する教師モデルの中間層出力を用いて学習することで，各層が段階的に教師モデルの表現に近づくよう促す．

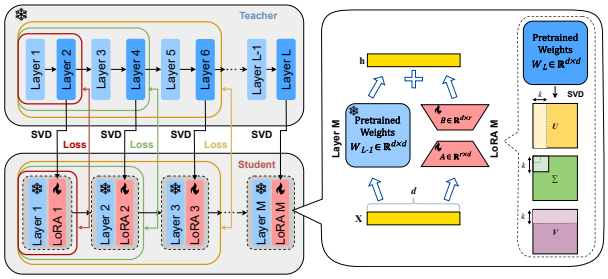

図 2：提案手法

4. 評価実験

事前学習済みの教師モデルから間隔 2 で間引いて生徒モデルを構築する．構築した生徒モデルに対して，提案手法および WeCoLoRA をそれぞれ適用し，精度を定量的に比較する．

4.1. 実験概要

本実験では，事前学習済み ViT-B を教師モデルとする．蒸留には ImageNet-1k の 1% のデータを用いる．WeCoLoRA は 15 エポック蒸留を行うのに対し，提案手法では 5 層分の段階的蒸留後，残りのエポック数で全体の蒸留を行う．総蒸留エポック数は WeCoLoRA と同じである．下流タスクの評価には ImageNet-1k または CIFAR-100 を用いて，50 エポック学習する．

4.2. 精度比較

各手法における精度を表 1 に示す．表 1 より，ImageNet-1k データセットにおける Top-1 accuracy は，WeCoLoRA が 64.50% であるのに対し，提案手法は最大 68.87% を達成し，4.37pt の精度向上が確認された．また，CIFAR-100 データセットでは，WeCoLoRA の 62.46% に対して，提案手法は最大 68.13% を達成し，5.67pt の精度向上が得られた．さらに，学習時間は，WeCoLoRA と比較して約 15.8% 短縮した．これにより，精度の向上と効率的なモデル圧縮の両立という観点から，本手法の有効性を確認した．

表 1：各手法の精度比較

手法	LoRA の初期値		学習数		学習時間	Top-1 accuracy	
	ランダム	SVD	段階的蒸留	蒸留		ImageNet-1k	CIFAR-100
教師モデル	-	-	-	-	-	81.37	80.57
WeCoLoRA	✓		-	15	0:20:33	64.50	62.46
提案手法	✓		1×5	10	0:17:17	66.27	67.39
		✓				**68.87**	**68.13**

5. おわりに

本研究では，間引く層の知識を活用し，生徒モデルを教師モデルの中間出力により近づける蒸留手法を提案した．今後は，様々なデータセットやモデルサイズを変更して，より汎化性能の高い生徒モデルの構築を目指す．

参考文献

[1] D. Grigore, *et al.*, “Weight Copy and Low-Rank Adaptation for Few-Shot Distillation of Vision Transformers,” arXiv preprint arXiv:2404.09326, 2024.