# 枝刈り率を用いた適応的層選択と MPO 分解によるモデル軽量化

EP22038 **大原琉生**　　**指導教授：山下隆義**

## 1. はじめに

Transformer をベースとした深層学習モデルが注目されている．画像認識分野においても，画像エンコーダに Transformer を用いたモデルである Vision Transformer (ViT) が高精度を示している．一方で，ViT は重みパラメータ数が多いため，計算資源の限られたエッジデバイスへの展開が困難である．この課題に対し，モデルの重みパラメータを削減する手法として行列積演算子 (MPO) 分解による低ランク近似がある [2]．MPO 分解はモデルのパラメータ削減と推論の高速化が可能であるが，全ての層に一律で適用するとモデル性能が著しく低下する課題がある．そこで本研究では，層ごとにパラメータの冗長性が異なる点に着目し，各層の重要度に応じて動的に低ランク近似する手法を提案する．これにより，精度維持と高速化を両立した軽量化手法の実現を目指す．

## 2. 従来手法

本章では，提案手法の基盤となる非構造枝刈り手法である Adaptive Feature Retaining (AFR) と，低ランク近似手法である Matrix Product Operator (MPO) 分解について述べる．

### 2.1. Adaptive Feature Retaining

Adaptive Feature Retaining (AFR)[1] は，事前学習モデルの知識維持と下流タスク適用を両立した非構造枝刈り手法である．AFR では知識維持とタスク適応の両観点から重みの重要度を評価する指標を導入している．具体的には，$i$ 番目の重みパラメータ $w_i$ に対する評価値 $S_{\mathrm{AFR}}(w_i)$ を式 (1) のように定義する．

$$S_{\mathrm{AFR}}(w_i) = \mathcal{Z}\left(\left|\frac{\partial \sum_{l=1}^{L} F_{\mathrm{SVD}}^{l}}{\partial w_i} w_i\right|\right) + \mathcal{Z}\left(\left|\frac{\partial \mathcal{L}}{\partial w_i} w_i\right|\right) \quad (1)$$

ここで，$\mathcal{Z}(\cdot)$ は標準化を表す．第 1 項は知識維持の観点に基づいた指標であり，$l$ 層目の出力特徴量に対する特異値の平均 $F_{\mathrm{SVD}}^{l}$ を用いて，$w_i$ が特徴空間の情報量や分布にどの程度寄与しているかを評価する．第 2 項はタスク適応の観点に基づく項であり，損失関数 $\mathcal{L}(\cdot)$ に対する勾配を用いて，下流タスクの精度に対する重みの影響を評価している．これにより，事前学習で獲得した知識を保持しながら，下流タスクに対する適応性も考慮した効果的な枝刈りが実現される．

### 2.2. 行列積演算子によるモデル圧縮

行列積演算子 (MPO) 分解によるモデル圧縮 [2] は，重み行列を低ランクテンソルネットワークに近似することで深層学習モデルを効率的に圧縮する手法である．MPO 分解では入力次元 $N$，出力次元 $M$ を持つ重み行列 $\boldsymbol{W} \in \mathbb{R}^{M \times N}$ に対し，それぞれの次元を $N = \prod_{k=1}^{n} I_k$, $M = \prod_{k=1}^{n} J_k$ となる $n$ 個の因子の積に分解し，高階テンソル $\boldsymbol{W}_{j_1 \ldots j_n, i_1 \ldots i_n}$ に変形する．このとき，$\boldsymbol{W}_{j_1 \ldots j_n, i_1 \ldots i_n}$ は式 (2) のように，$n$ 個のコアテンソル $\boldsymbol{w}^{(k)}$ の縮約として近似できる．

$$\boldsymbol{W}_{j_1 \ldots j_n, i_1 \ldots i_n} \approx \mathrm{Tr}\left(\boldsymbol{w}^{(1)}[j_1, i_1] \cdots \boldsymbol{w}^{(n)}[j_n, i_n]\right) \quad (2)$$

ここで，各コアテンソル $\boldsymbol{w}^{(k)}$ は隣接するコアテンソルと接続するための次元であるボンドインデックス（ボンドランク）を持ち，その大きさを調整することで表現力とパラメータ数のトレードオフを調整できる．

## 3. 提案手法

本研究では，非構造枝刈りにおける枝刈り率を指標として，MPO 分解による低ランク近似を適用する層を動的に選択する手法を提案する．具体的には，事前学習済みモデルに対する非構造枝刈り手法として有効な AFR を用いて各層の枝刈り率を算出する．そして，枝刈り率が高い上位 $K$ 個の層の集合 $L_{\mathrm{topk}}$ に対して MPO 分解による低ランク近似を適用する．層 $l$ に含まれる重み行列を $\boldsymbol{W}$，提案手法適用後の重みを $\hat{\boldsymbol{W}}$ とすると，動的な層選択は式 (3) のように定式化される．

$$\hat{\boldsymbol{W}} = \begin{cases} \Phi_{\mathrm{MPO}}(\boldsymbol{W}) & (l \in L_{\mathrm{topK}}) \\ \boldsymbol{W} & (\text{otherwise}) \end{cases} \quad (3)$$

ここで，$\Phi_{\mathrm{MPO}}(\cdot)$ は MPO 分解による低ランク近似を表す．これにより，重要層は保持し冗長層のみ圧縮することで，精度維持と高速化を実現する．また，本手法では非構造枝刈りを不要な重みを除去する前処理として利用する．枝刈りによってノイズが低減された行列に対して MPO 分解を行うことで，元の行列を近似する場合と比較して，重要な情報を損なわずに低ランク近似が可能になることを期待する．

## 4. 評価実験

ImageNet-1k で事前学習された ViT-B/16 を baseline とし，提案手法による画像分類精度の変化およびモデル圧縮と高速化の効果を評価する．

### 4.1. 実験概要

MPO 分解を適用する対象は MLP とし，適用する層数は 12 ブロック中 6 ブロックとする．MPO 分解はコア数 2，ランク 4 とする．また，提案手法の有効性を検証するため，以下 2 つの選択方法において比較を行う．

**固定選択** 偶数番目のブロックに MPO 分解を適用する．なお，前処理として AFR による枝刈り率を 70%とする．

**動的選択** 提案手法に基づき，ブロックごとの枝刈り率が高い上位 6 ブロックを選択して MPO 分解を適用する．AFR における枝刈り率を，10%および 70%とする．

評価データセットは CIFAR-10 と CIFAR-100，Stanford Cars を用い，提案手法を適用後に 150 エポックのファインチューニングを行う．

### 4.2. 実験結果

ViT-B/16 に対して提案手法を適用した際の画像分類精度と推論の高速化率を表 1 に示す．表 1 より，固定選択に比べ枝刈り率 70%の動的選択が高い精度を示した．これは，枝刈り率を指標とすることで，圧縮による影響が少ない層を適切に選択できていることを示している．また，枝刈り率 10%と 70%の比較では，全てのデータセットで 70%の方が高い精度を示した．これは，非構造枝刈りにより低ランクでの近似のしやすさが向上していることが考えられ，低ランク表現化における枝刈りの有効性を示している．さらに，推論速度についてはベースラインと比較して最大 1.06 倍の向上が確認されたものの，大幅な改善には至らなかった．

表 1: 圧縮後モデルの分類精度 [%]

| Dataset | baseline | Fixed | Ours (10%) | Ours (70%) |
|---|---|---|---|---|
| CIFAR-10 | 98.39 | 94.34 | 94.06 | **94.51** |
| CIFAR-100 | 89.53 | 76.78 | 77.03 | **80.84** |
| Stanford Cars | 68.37 | 17.91 | 13.05 | **24.39** |
| Speed-up | 1.00x | 1.05x | 1.06x | 1.06x |

## 5. おわりに

本稿では，非構造枝刈りにおける枝刈り率を指標として，MPO 分解を適用する層を動的に選択する手法を提案した．評価実験の結果，固定的な層選択と比較して，提案手法による動的選択が精度の向上に寄与することを確認した．特に，非構造枝刈り率が高い設定においてその効果が顕著であった．今後は，非構造枝刈りで生じた 0 値が低ランク近似に与える影響の調査と，より高度な層の選択指標の検討を行う．

## 参考文献

[1] 新田常顧, et al. “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 Single-shot Foresight Pruning”, 画像の認識・理解シンポジウム, 2025.

[2] Ze-Feng Gao, et al. “Compressing deep neural networks by matrix product operators”, Physical Review Research, 2020.