

SSII30周年技術マップ：枝刈り

構造化枝刈り
(Structured Pruning)

非構造化枝刈り
(Unstructured Pruning)

Transformer

枝刈りのタイミング

- 学習前 (初期状態のモデルの重みを評価)
- 学習中 (スパース正則化等でモデルとマスクを学習)
- 学習後 (学習済みモデルの重みを評価)
- 推論時 (入力に応じて使用する重みを動的に選択)

Transformerモデルの登場
Attention Is All You Need [Vaswani+, NIPS'17]

Analyzing Multi-Head Self-Attention [Voita+, ACL'19]
LO正則化を用いてヘッドを削除

他の圧縮手法との併用

DynaBERT [Hou+, NIPS'20]
枝刈り+蒸留で層と幅を圧縮

ROSITA [Liu+, AAAI'21]
枝刈り+低ランク因数分解+知識蒸留の併用

Unified Visual Transformer Compression [Yu+, ICLR'22]
枝刈り+ブロックスキップ+知識蒸留

WDPPruning [Yu+, AAAI'22]
層の深さと幅を同時に削減
説明性を考慮した手法

X-Pruner [Yu+, ICCV'23]
ViTの説明性を考慮した枝刈り

Zero-TPrune [Wang+, CVPR'24]
トークンの重要度と類似度で枝刈り

Jamba [Lieber+, arXiv'24]
MoE的にサブネットワークを切り替え

Are Sixteen Heads Really Better than One? [Michel+, NIPS'19]
ヘッド毎の勾配の大きさを基準に枝刈り

Structured Pruning of Large Language Models [Wang+, ACL'20]
対角行列はHard Concrete分布からサンプリング

Chasing sparsity in vision transformers [Chen+, NIPS'21]
ヘッドの出力特徴をテイラー展開を応用して評価

Vision Transformer Pruning [Zhu+, KDD'21]
各層の前後でトークンの次元数を圧縮するための対角行列を導入

Patch Slimming [Tang+, CVPR'22]
パッチ間の類似度を観察

Deja Vu [Liu+, ICML'23]
入力に応じて動的にチャンネルやヘッドを切り替え

DepGraph [Fang+, CVPR'23]
DNNの各層の依存関係をグラフとして表現

CNN

CNNモデルの登場
LeNet [LeCun+, 98]
CNNモデルの発展
AlexNet [Krizhevsky+, NIPS'12]

Learning Structured Sparsity in Deep Learning [Wen+, NIPS'16]
カーネル、チャンネル、層の深さに対して正則化を適用

Network Slimming [Liu, ICCV'17]
スケールリングファクタの導入と正則化

Rethinking the Smaller-Norm Less-Informative Assumption [Ye+, ICLR'18]
正則化項に計算量の基準を導入

NetAdapt [Yang, ECCV'18]
実行時間を基準に評価

Importance Estimation [Molchanov+, CVPR'20]
カーネルを削除した際の損失の変化量の二乗を評価

Group Fisher Pruning [Liu+, ICML'21]
Fisher informationを用いた重みの評価

ResRep [Ding+, ICCV'21]
冗長な重みにのみ正則化を適用

CPrune [Kim+, ECCV'22]
DNNコンパイラを用いた実行速度ベースの枝刈り

レイヤー単位で枝刈り
Channel Pruning [He, ICCV'17]
層の前後で特徴量の変化が最小になるような重みを再構成

Discrimination-aware Channel Pruning [Zhuang+, NIPS'18]
層毎に識別器を導入

Filter Pruning via Geometric Median [He+, CVPR'19]
カーネル内の重み及びばらつきを考慮

Gate Decorator [You+, NIPS'19]
テイラー展開ベースのスケールリングファクタの評価

枝刈り後の微調整不要
DeepMoE [Wang+, PMLR'20]
学習したゲートネットワークで動的にチャンネルを枝刈り

OTO [Chen+, NIPS'21]
パラメータをグループ化してグループ単位で枝刈り

ATO [Chen+, CVPR'24]
コントローラネットワークでマスクを学習

勾配ベース

算出した勾配を基にして枝刈り

枝刈りの始まり

Optimal Brain Surgeon [Hassibi & Stork, NIPS'92]
ある重み1つに対するテイラー展開を用いてOBDを一般化

Optimal Brain Damage [LeCun+, NIPS'90]
ヘシアンを用いて損失が増加しない重みを特定

学習前に枝刈り可能

SNIP [Lee+, ICLR'19]
損失関数に対する重みの感度を評価

GraSP [Wang+, ICLR'20]
ヘシアンを用いて重み同士の関係性を評価

Synflow [Tanaka+, NIPS'20]
高圧縮率による層の崩壊へ対応(データフリー)

Neural Tangent Transfer [Liu+, ICML'20]
NTKで特徴づけた学習ダイナミクスを模倣するサブネットワークを探索

Iterative SNIP [De Jorge+, ICLR'21]
SNIPをマルチショットで適用

NTK-SAP [Wang+, ICLR'23]
NTKの変化が小さい重みを評価

マグニチュードベース

重みパラメータの大きさを基にして枝刈り

Learning Both Weights and Connections for Efficient Neural Networks [Han+, ACM'15]
重みの大きさを基準に枝刈り

Dynamic Network Surgery for Efficient DNNs [Guo+, NIPS'16]
枝刈りされた重みを回復させるためのオプションを追加

Exploring Sparsity in Recurrent Neural Networks [Narang+, ICLR'17]
RNNの学習過程で効率的に適用

The Lottery Ticket Hypothesis [Frank+, ICLR'19]
初期ネットには同等性能のサブネットが存在

Early-Bird tickets [You+, ICLR'20]
早い学習段階で当たりくじを発見

Winning Lottery Tickets in Deep Generative Models [Kalibhat+, AAAI'21]
生成モデルに拡張

Dual Lottery Ticket Hypothesis [Bai+, ICLR'22]
ランダムに選択したサブネットは当たりくじに変えることが可能

Strong Lottery Ticket Hypothesis [Ramanujan+, CVPR'20]
大規模なネットには未学習でも同等性能のサブネットが存在

Graph Lottery Ticket Hypothesis [Chen+, ICML'21]
Graph Neural Networkに拡張

Towards Structurally Sparse Lottery Tickets [Chen+, ICML'22]
非ゼロ要素をグループ化して構造的にスパースな当たりくじを獲得

Single-shot Pruning For Pre-trained Models [Kohama+, ICCVW'23]
事前学習済みモデルに対応

CLIP-Q [Tung+, CVPR'18]
マグニチュードによる枝刈りと量子化を併用

Learnable Pruning [Yao+, arXiv'21]
予め設定した枝刈り率に対応した正則化手法

Pruning Deep Neural Networks from a Sparsity Perspective [Diao+, ICLR'23]
経済学の富の分配法則を用いた枝刈り率の動的決定

枝刈りのハードウェア対応

EIE [S.Han+, ISCA'16]
重み共有と疎行列演算に対応

NVIDIA Ampere [NVIDIA, GTC'20]
枝刈り後の推論を高速化する機能を実装

DRP-AI [Renesas, ISSCC'24]
マイクロプロセッサであるRZ/V2Hに搭載されるAIアクセラレータ

1990 2016 2018 2020 2022 2024 年

- 技術マップ (PDF)

- http://mprg.jp/SSII/SSII2024_map_Pruning.pdf

技術マップ



- 文献情報 (Google スプレッドシート)

- <https://docs.google.com/spreadsheets/d/1qi29ksnt1Kejx2wmKWwdngcYmYJsyWStCFghbJy14kg/edit?usp=sharing>

文献情報



3	タイトル	学会	年	種類
4	Learning Structured Sparsity in Deep Neural Networks	NIPS	2016	Struct...
5	Less is More: Data Pruning for Faster Adversarial Training	ECCV	2016	Struct...
6	Fast ConvNets Using Group-wise Brain Damage	CVPR	2016	Struct...
7	Learning Efficient Convolutional Networks through Network Slimming	ICCV	2017	Struct...
8	PRUNING FILTERS FOR EFFICIENT CONVNETS	ICLR	2017	Struct...
9	Structured Pruning of Deep Convolutional Neural Networks		2017	Struct...

- 作成者

- 新田 常顧, 伊藤 天詞, 小濱 大和, 西川 実希, 平川 翼, 山下 隆義, 藤吉 弘亘 (中部大学)
- 関川 雄介, 安倍 満, 佐藤 育郎 (DENSO IT LAB)