4S-01

進化計算法を用いた深層学習モデルの Pruning 最適化

高梨 雄大† 長尾 智晴‡

†横浜国立大学 大学院環境情報学府 ‡横浜国立大学 大学院環境情報研究院

1 はじめに

近年の深層学習の発展とともに，モデルに必要となる物理メモリや計算量，消費電力は増加している．例えば有名な深層学習のモデルである VGG16 は約 528MB のメモリ容量を必要とし，計算機リソースの少ないスマートフォンなどの端末で直接利用することは困難である．現在までに，モデルの精度を維持したまま圧縮する手法は多数提案されており，その中で畳み込みニューラルネットワーク(Convolutional Neural Network; CNN)をバッチ正則化(Batch Normalization; BN)のパラメータ γ を基準として Pruning する Network Slimming[1]は高い圧縮性能を実現している．しかし，Pruning の対象を決定する基準は，経験則やモデルの近似式から決定されているため，最適化の余地があると考えられる．そこで，本稿では CNN チャネルの Pruning の最適化を進化計算法(Evolutionary Computation; EC)の一種である遺伝的アルゴリズム(Genetic Algorithm; GA)を用いて最適化する手法を提案する．

2 提案手法

本手法では CNN チャネルの Pruning の有無を，01 のビットのマスクで表現する．CNN の各層の出力C_{out}とマスクMのアダマール積を求めることで，Pruning 後の出力C'_{out}を計算することができる．Pruning 後にモデルの精度を維持可能な，マスクMのビット列を求める組み合わせ最適化問題とすることで，GA を適用し最適化する．

2.1 GA の交叉と突然変異

本手法の GA の遺伝子には，全ての個体が等しい圧縮率を実現できるように，0 と 1 の比率を一定にする制約を課す．これによって，全く圧縮を行わない遺伝子の適合度が高くなり，進化が停滞する事態を防止する．全ての遺伝子が

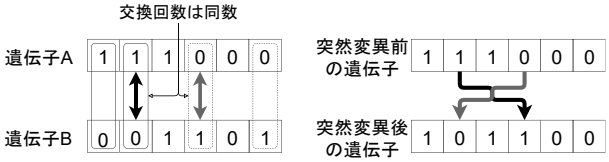

図 1：0 と 1 の比率を維持する一様交叉と転座

初期化時の 0 と 1 の比率を維持するために，図 1 に示すような一様交叉と転座の改変を行う．2 つの遺伝子 A と B の交叉では，遺伝子 A の 0 と遺伝子 B の 1 の交換，遺伝子 A の 1 と遺伝子 B の 0 の交換が同じ回数行われるように，交換の回数を先に二項分布から決定し，その回数だけランダムな交換を行う．また，転座では，1 から 0 の突然変異の回数と 0 から 1 の突然変異の回数が，同一遺伝子内で同じになるように，遺伝子の要素を選択して交換する．

2.2 Surrogate モデル

Pruning を GA で最適化するための適合度関数に，Fine-tuning 後の検証精度を利用すると，各個体の学習に莫大なコストが掛かってしまう．そのため，本手法では適合度関数に Surrogate モデルによる近似を導入する．Surrogate モデルを利用した最適化の流れを図 2 と以下に示す．

1. ランダムなビットマスクMで Pruning し，Fine-tuning した後の検証データで Surrogate モデルの初期学習データの作成
2. Surrogate モデルの学習
3. Surrogate モデルの出力を適合度の近似とする GA の最適化(一定数の世代)
4. 集団の個体の一部を抽出し，初期学習データの生成と同様の手順で追加学習データを生成．Surrogate モデルの学習データに追加
5. 2 に戻る

本手法では Surrogate モデルに LightGBM[2]を選択した．

Pruning Optimization of Deep Learning Model Using Evolutionary Computation

† Yuta Takanashi, Graduate School of Environment and Information Sciences, Yokohama National University

‡ Tomoharu Nagao, Faculty of Environment and Information Sciences, Yokohama National University

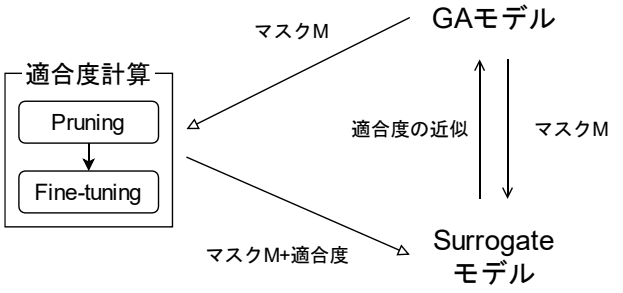

図 2：Surrogate モデルによる GA の最適化

3 実験

3.1 実験設定

MobileNet v1(α=0.25)[3]のチャネル数を 4 分の 1 にする圧縮を行う．Network Slimming，GA による提案手法の 2 種類の Pruning の性能を，Fine-tuning 後の検証データに対する精度で比較をする．

CNN の学習データセットには，画像分類問題で広く用いられている CIFAR-10 を利用する．CIFAR-10 は，学習画像 50,000 枚，検証画像 10,000 枚のデータを含むが，その中の学習画像の 5,000 枚を，GA の適合度関数用のデータとして分離する．また，そのデータは学習画像に 3,000 枚，検証画像に 2,000 枚とした．

モデルの初期学習と Fine-tuning の学習設定を表 1，LightGBM のハイパーパラメータを表 2，GA の実験設定を表 3 に示す．

また，GA の初期個体生成時に Network Slimming で得られるマスクに対応する個体を 1 つ追加し，最適化の高速化を図る．

表 1：初期学習と Fine-tuning の学習設定

パラメータ	値
Epochs	40(初期学習は 160)
Batch size	64
Optimizer	Nesterov's accelerated gradient (momentum=0.9)
Learning rate (value:progress)	0.1:(0%, 50%] 0.01:(50%, 75%] 0.001:(75%, 100%]

表 2：LightGBM のハイパーパラメータ

パラメータ	値
Leaves	32
Steps	100
Objective	リッジ回帰

表 3：GA のハイパーパラメータ

パラメータ	値
集団サイズ	200
遺伝子長	1952
世代数 (GA 世代数 * Surrogate モデルへのデータ追加回数)	100 * 1000
交叉率	0.8
突然変異率	0.02

表 4：実験結果

手法	検証精度(%)
Network Slimming	79.32
提案手法	**79.71**

3.2 実験結果

表 4 に提案手法と比較手法の実験結果を示す．Surrogate モデルを利用した GA による最適化で，チャネルの Pruning の性能が向上した．各層ごとに残ったフィルター数を確認したところ，比較手法と比べて，提案手法では全体のバランスが取れた Pruning をすることができたためだと考えられる．

4 まとめ

本稿では，進化計算法を用いて深層学習モデルの Pruning を最適化する手法を提案し，比較手法に対しての優位性を検証した．今後は，Surrogate モデルの効率化や CNN 以外のネットワークへの適用を目指す．

参考文献

[1] Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., & Zhang, C. (2017). Learning efficient convolutional networks through network slimming. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2736-2744).

[2] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (pp. 3146-3154).

[3] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.