F-008

CNN に対するフィルタ刈り込みと欠損した情報の補償による性能低下の抑制

Suppression of performance degradation in CNNs by filter pruning and compensating for missing information

三浦 隼 *　　亀山 啓輔 †
Shun Miura　　Keisuke Kameyama

1 初めに

畳み込みニューラルネットワーク (CNN) による機械学習は、画像識別課題や画像生成課題において最先端の結果を達成しており非常に注目されている技術である。また近年、CNN を用いた計算モデルは、より深い層と、素子数の多い構造を持つようになることで、推論のための計算コストが大幅に増加しており、計算資源の乏しいような環境では運用が難しいという問題がある。そのような問題を解決する手法として、CNN の構造的フィルタ刈り込みが存在する。多くの刈り込みに関する先行研究は刈り込み後に再学習を行うことでモデルの性能を回復する手法をとっているが、再度同じようにモデルの学習を行うことは、時間的なコストの大きさや、学習時のデータセットの依存性等の問題が指摘されている [4]。特に敵対的生成ネットワーク（GAN）のように学習の収束が不安定で学習コストの大きいモデルでは、刈り込みによる性能低下が大きいと再学習による回復が見込めない可能性がある [3]。

そこで本稿では、刈り込み後のモデルの性能低下を抑制することを目的として、刈り込み後に残されたフィルタに対して刈り込みによって欠損した情報を補う手法を提案する。特に CNN は、全体の計算コストのうち畳み込み層での計算コストが大きく占めていることから、畳み込み層を対象にしてフィルタ刈り込みを行う。実験では、基本的な基準による刈り込みに加えて提案手法を適用した場合のモデルの性能変化および出力の変化について報告する。

2 提案手法

2.1 基本方針

本手法は、基本的な刈り込み基準に加えてフィルタパラメータの補償を行うことにより刈り込みによるモデルの性能変化の抑制を目的とする。一般に学習を終えた CNN の特定の層内にはそのはたらきの類似度の高いフィルタが存在しており、そのようなフィルタの組はモデルにとって冗長であると考える。したがって刈り込み時に、刈り込まれたフィルタとモデル内に保持されるフィルタの類似性に基づいて、補償を行うことでモデルの性能低下を抑制することができると考える。この考えに基づいて 2 つの手法を提案する。

2.2 提案手法 1: 余弦類似度による補償

はじめに所定の刈り込み基準を用いて、ある層において刈り込まれるフィルタ集合 J_{keep}と保持されるフィルタ集合 J_{prune}を得る。次に、層の i 番目のフィルタ F_i を刈り込むとき、補償の相手としてその組となるフィルタ F_jは以下のように求める。

$$j = \underset{m}{argmax}\ S(i,m) \quad (i \in J_{prune}, m \in J_{keep}) \tag{1}$$

$$S(i,m) = \left| \frac{< \widehat{F}_i, \widehat{F}_m >}{||\widehat{F}_i||_1 \cdot ||\widehat{F}_m||_1} \right| \tag{2}$$

$S(i,m)$ は i 番目のフィルタと m 番目のフィルタの類似度を表す関数であり、ここではフィルタの係数をベクトル化したもの $(\widehat{F})$ の余弦値によって類似度を計測する。またこのときフィルタ j は刈り込まれるフィルタ i と最も類似度の近いフィルタと言うことができる。最後に、刈り込まれるフィルタ i のパラメータの値をフィルタ j のパラメータに足し合わせることで刈り込みによって欠損した情報の補償を行う。

$$F_j = F_j + F_i \qquad (i \in J_{prune}, j \in J_{keep}) \tag{3}$$

2.3 提案手法 2: フィルタ値の倍加による補償

提案手法 1 同様に、所定の刈込基準を用いて、ある層において刈り込まれるフィルタ集合 J_{keep} を得る。次にフィルタ集合 J_{keep} のすべての要素フィルタ i の値を 2 倍にする。提案手法 2 では、保持されるフィルタは刈り込まれるフィルタとの類似度の高い組が必ず存在するという仮定のもとで、組の一方のフィルタが刈り込まれるとき、もう一方のフィルタの情報を倍にすることでそれを補うことが可能であるという考えに基づいている。

実際に刈り込まれたモデルで推論を行う場合は、刈り込まれた構造とパラメータを反映したモデルを構築し直す必要がある。

* 筑波大学大学院情報理工学位プログラム
† 筑波大学システム情報系

3　実験と分析

3.1　概要

実験では、提案手法の効果を評価することを目的として、画像識別課題を行う CNN および、画像生成課題を行う転置 CNN に対して提案手法を適用した。どちらの課題でも訓練および検証のためのデータセットとして、手書き数字画像の MNIST を用いた。本実験では基本的な刈り込み方法として、層内のフィルタの L1 ノルムの小さい順（L1 ノルム）および大きい順（最大 L1 ノルム）に基づく刈り込みを行った。

3.2　CNN 画像識別モデル

本実験では 4 層の畳み込み層と 2 層の全結合層を隠れ層として持つモデルを用いた。モデルに対する刈り込みと補償を行う場合の精度の変化を図 1 に示す。両方の刈り込み基準において、提案手法 1 による補償を行った場合では識別精度低下の抑制の効果は見られず、刈り込み割合 70% 以下の条件では補償を行わない条件よりも精度がわずかに低くなる結果となった。一方で、保持されたフィルタの値を 2 倍とする条件では、識別精度の低下を大きく抑制できていることがわかる。

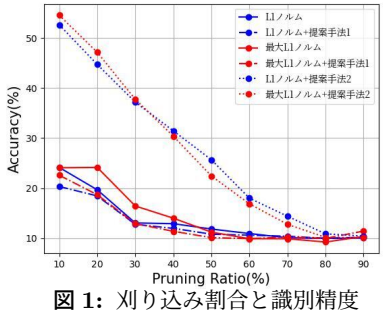


図 1: 刈り込み割合と識別精度

3.3　転置 CNN 画像生成モデル

ここでは画像生成を行う転置 CNN に対して提案手法 1 を適用した結果を示す。特に今回は GAN[1] の一種である DCGAN[2] のアルゴリズムで学習したモデルを使用した。また同じノイズから生成された刈り込み前の生成画像と刈り込み後の生成画像の平均二乗誤差を用いて、刈り込み前後でモデルがどの程度変化したかを評価する。図 2 に示すように、どちらの刈り込み基準においても提案手法 1 による補償を行うと MSE が増加し、生成画像がより大きく変化していることがわかる。また、実際に生成された画像の例を図 3 に示す。補償を行う条件だと、数字の輪郭の消失が大きいといえる。

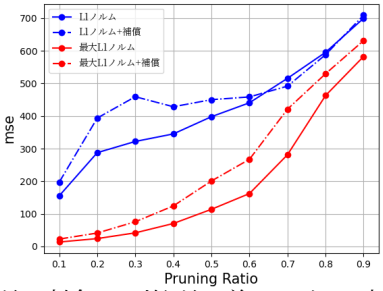
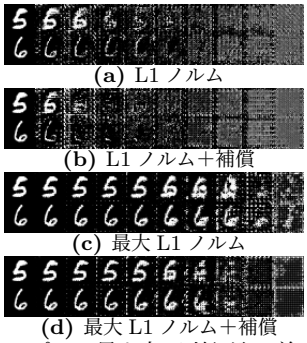


図 2: 刈り込み割合と、刈り込み前のモデルの出力と刈り込み後の出力の平均二乗誤差 (MSE)

ここで本実験で使用した転置 CNN モデルは、画像生成時にもバッチ正規化の処理を必要としている。そのようなモデルでは提案手法 2 による補償は正規化により打ち消されてしまい効果を得られないことに留意されたい。



(a) L1 ノルム
(b) L1 ノルム＋補償
(c) 最大 L1 ノルム
(d) 最大 L1 ノルム＋補償

図 3: 生成画像サンプル。最も左が刈り込み前モデルの生成画像で最も右が 90% 刈り込み後モデルの生成画像

4　結論

提案手法 1 で行った補償を伴うフィルタの刈り込み法は、CNN の画像識別モデルおよび画像生成モデルにおいて、モデルの精度低下を抑制することについて明確な効果を確認することはできなかった。一方で提案手法 2 において、フィルタの値を単に 2 倍にするだけで精度低下の抑制効果が見られたことから、フィルタの値を操作することによる補償の方法について、改善することでより効果的な方法が見つかることが期待できる。さらに今後の課題として、補償の処理によって畳み込みフィルタによる特徴抽出および特徴生成のはたらきがどのように変化しているかを分析し、より効果的な補償の方法について検討する必要がある。また図 2,3 から見られるように、転置 CNN において L1 ノルムの大きい順に刈り込みを行った条件では小さい順に行った条件よりもモデルの変化が少ないという興味深い結果が得られた。これは値の大きいフィルタはモデルに対する影響が大きいという一般的な考えと反するものであり、さらなる検証が求められる。

参考文献

[1] I. J. Goodfellow et al. Generative adversarial nets. In *Proceedings of the 27th International Conference on NeurIPS-Volume 2*, pages 2672–2680, 2014.

[2] A. Radford et al. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[3] H. Wang et al. GAN slimming: All-in-one GAN compression by a unified optimization framework. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 54–73. Springer, 2020.

[4] Z. Xie et al. Pruning with compensation: Efficient channel pruning for deep convolutional neural networks. *arXiv preprint arXiv:2108.13728*, 2021.