段階的な蒸留を導入した宝くじ仮説による物体検出モデルの枝刈りに関する研究

TP22017 早川 和希　　指導教授：山下 隆義

1. はじめに

YOLO をはじめとする深層学習による物体検出手法は，高精度な検出が可能である．しかしながら，物体検出のモデルサイズは大きいため，搭載メモリの小さいエッジデバイスへの展開は困難である．そのため，大規模な物体検出モデルを圧縮し，パラメータ数を削減することが求められている．ネットワークの圧縮方法の一つに枝刈り [1] がある．枝刈りは，モデルに含まれる冗長なパラメータを削除し，モデルサイズを削減する．しかし，高い枝刈り率で枝刈りしたモデルは，性能が低下することが確認されている．そこで，本研究では，性能低下を抑制するために枝刈り後に段階的な知識蒸留を行う手法を提案する．具体的には，教師モデルとして，枝刈り率の高いモデルから低いモデルに段階的に切り替えることで，生徒モデルと教師モデル間における特徴表現のギャップを軽減させる．

2. 関連研究

大規模なニューラルネットワークのモデルを圧縮する方法として，枝刈りと知識蒸留がある．

2.1 枝刈り

枝刈りは，モデルのパラメータ数を削除する手法である．推論結果に影響しないパラメータを削除することで，大規模モデルを圧縮する．以下に代表的な枝刈り手法について示す．

Magnitude Pruning

Magnitude pruning(Magnitude) は，学習したニューラルネットワークの重みの中で，絶対値の小さいものから順番に削除する手法である．この時，枝刈りする割合を小さくしておき，繰り返し枝刈りを行う Iterative Magnitude Pruning(IMP) 法がある．Frankle らは IMP 法を用いて大規模なネットワークを枝刈りし，元のネットワークの精度と同等な小規模ネットワーク (サブネットワーク) を見つける宝くじ仮説を提案した [3]．しかしながら，高い枝刈り率で宝くじ仮説に基づく枝刈りを行った場合，急激に精度が低下することが確認されている．

Single-shot network Pruning

学習を必要としない枝刈り手法として，Single-shot Network Pruning(SNIP) がある [4]．SNIP は式 (1) のように，パラメータに摂動 δ を与えた際に，ネットワークの誤差 L がどの程度影響を受けるのかを定量的に評価し影響の少ないパラメータから削除する．

$$\begin{aligned}\Delta L &= |L(\theta_q) - L(\delta\theta_q)| \\ &= |L(\theta_q) - L(\theta_q) - \frac{\partial L}{\partial \theta_q}(\delta\theta_q - \theta_q) - O(||\delta\theta_q||^2)| \\ &\approx \left|\frac{\partial L}{\partial \theta_q}\theta_q\right|\end{aligned}$$

ここで，θ_q はネットワークに含まれる q 番目のパラメータの重み，$O(||\delta\theta_q||^2)$ は二次以上のオーダー表記である．ΔL が小さいパラメータは，削除した際に認識への影響が小さいと考えることができるため，効率的に枝刈りを行うことができる．

2.2 知識蒸留

知識蒸留 [5] は，大規模なネットワークを教師モデル，小規模なネットワークを生徒モデルとして，教師モデルの出力に生徒モデルの出力を近づけるように学習する手法である．しかし，教師モデルと生徒モデル間のモデルサイズの差が大きくなるにつれて，生徒モデルの精度が低下する．そこで段階的な蒸留を行う．Teacher Assistant(TA) が提案されている [6]．TA は，教師モデルと生徒モデルの中間規模のモデルを用意する．そして，教師モデルと中間モデルで蒸留した後，中間モデルと生徒モデルで蒸留を行う．これにより，教師モデルと生徒モデル間の精度のギャップを埋める．

2.3 知識蒸留付き宝くじ仮説

Liu らは，宝くじ仮説と知識蒸留を組み合わせる One-stage 物体検出手法を提案した [7]．枝刈りする前のモデルを教師モデル，宝くじ仮説によって枝刈りしたモデルを生徒モデルとして，目標パラメータ率に達するまで蒸留学習を行う．この手法により，宝くじ仮説による枝刈りのみで行う手法よりも精度の改善を確認した．

3. 提案手法

本研究では，宝くじ仮説による枝刈りを行ったモデルを段階的に蒸留学習を行うことで，知識蒸留付き宝くじ仮説よりも更に精度低下を抑制する手法を提案する．

3.1 段階的な蒸留を導入した宝くじ仮説による枝刈り

提案手法は，宝くじ仮説による枝刈りと段階的な知識蒸留を繰り返すことで，教師と生徒間のモデルサイズの差による特徴表現のギャップを軽減させる．提案手法の流れを図 1 と以下に示す．

まず，初期モデル $\boldsymbol{\theta}_0$ に対して k 回学習し，その重みを $\boldsymbol{\theta}_k$ とする．モデル $\boldsymbol{\theta}_k$ を収束するまで t 回学習し，その時の重みを $\boldsymbol{\theta}_{k+t}$ とする．枝刈り手法に従ってモデル $\boldsymbol{\theta}_{k+t}$ を $p\%$ 枝刈りした後，重みを $\boldsymbol{\theta}_k$ に戻したモデルを $\boldsymbol{m}_1 \odot \boldsymbol{\theta}_k$ とする．ここで，$\boldsymbol{m}_1$ は 1 回目の枝刈り時のマスクである．次に，$\boldsymbol{m}_1 \odot \boldsymbol{\theta}_k$ を学習し，$\boldsymbol{\theta}_{k+t}$ を教師モデル $\boldsymbol{\theta}^{t1}$ として知識蒸留を行い，知識蒸留モデル $\boldsymbol{m}_1 \odot \boldsymbol{\theta}_{k+t}$ とする．$\boldsymbol{m}_1 \odot \boldsymbol{\theta}_{k+t}$ を $p\%$ 枝刈りした後，重みを $\boldsymbol{\theta}_k$ に戻したモデルを $\boldsymbol{m}_2 \odot \boldsymbol{\theta}_k$ とする．$\boldsymbol{m}_2 \odot \boldsymbol{\theta}_k$ を学習し，$\boldsymbol{m}_1 \odot \boldsymbol{\theta}_{k+t}$ を教師モデル $\boldsymbol{\theta}^{t2}$ として 1 回目の蒸留を行った後，$\boldsymbol{\theta}^{t1}$ と 2 回目の蒸留を行う．このように 1 つ前の枝刈りモデルから枝刈り前のモデルまで，段階的に蒸留を行うことで，教師モデルとの精度のギャップを埋める．

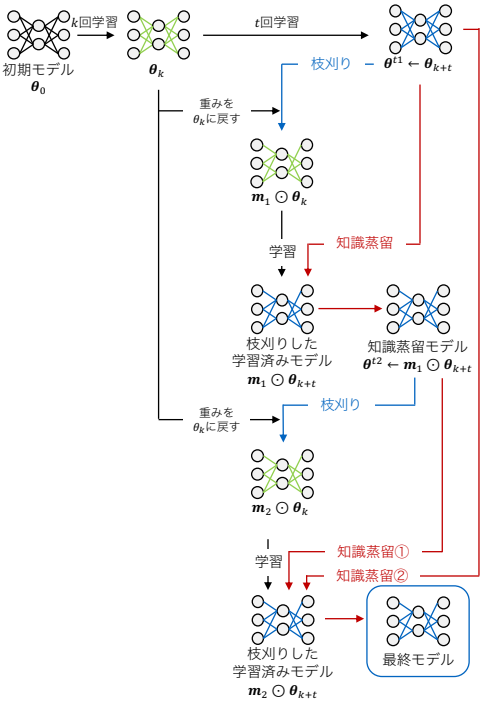
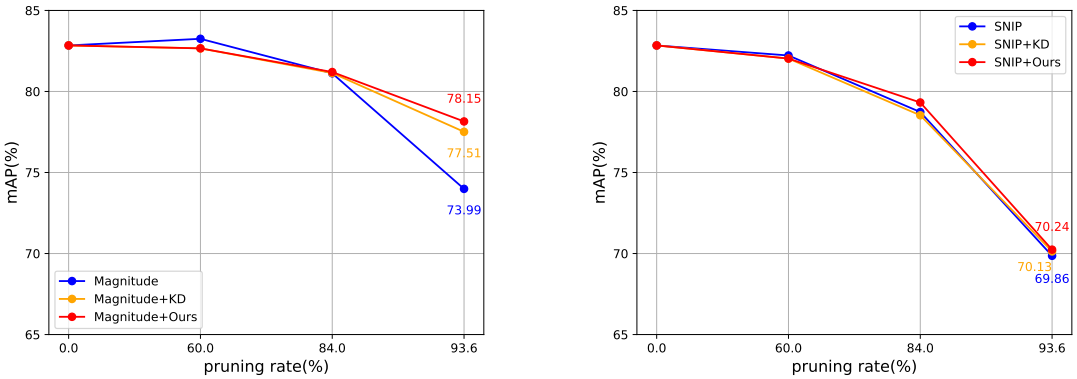


図 1：提案手法の学習流れ．

3.2 損失関数

本研究では物体検出モデルとして YOLOv8 を使用する．YOLOv8 は，損失関数としてクラス分類損失，2 種類のバウンディングボックス回帰損失を使用している．ここでは，冗長なバウンディングボックスの情報を削除するため，クラス分類損失のみを蒸留学習に用いる．損失関数 L を式 (1) に示す．



(a) Magnitude　　(b) SNIP

図 2：枝刈り率 60%の mAP 精度の比較

$$L = L_{hard}(s, G) + \alpha L_{soft}(s, t) \quad (1)$$

L_{hard} は生徒モデル出力と正解ラベルの損失, L_{soft} は生徒モデル出力と教師モデル出力の損失，s は生徒モデル θ^s の出力，G は正解データラベル，t は教師モデル θ^t の出力である．L_{hard} と L_{soft} の損失の割合をハイパーパラメータ α によって調整する．L_{hard} 及び L_{soft} は以下の式となる．

$$L_{hard}(s, G) = L_{cls}(s, G) + L_{box1}(s, G) + L_{box2}(s, G) \quad (2)$$

$$L_{soft}(s, t) = L_{cls_{kd}}(s, t) \quad (3)$$

L_{cls} は生徒モデルと正解ラベルのクラス分類損失, L_{box1},L_{box2} は生徒モデルと正解ラベルのバウンディングボックス回帰の損失である．$L_{cls_{kd}}$ は生徒モデルと教師モデルの分類損失である．ここで $L_{cls_{kd}}$ は平均二乗誤差で算出する．

4. 評価実験

提案手法との有効性を比較するため，従来手法である宝くじ仮説，知識蒸留付き宝くじ仮説 (KD) による学習と比較実験を行う．

4.1 実験概要

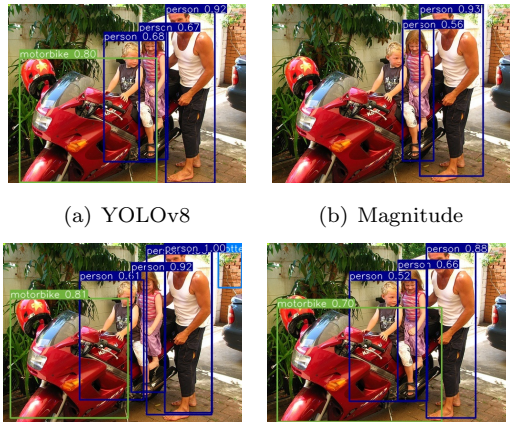
本実験では，Pascal VOC 2007 と 2012 を組み合わせたデータセットを用いる．学習用及び評価用データセットの枚数は，それぞれ 16,551 枚と 4,957 枚であり，クラス数は 20 クラスである．評価指標には mean Average Precision(mAP) を用いる．比較する手法として，通常の宝くじ仮説，Liu らが提案した知識蒸留付き宝くじ仮説，提案手法を用いる．宝くじ仮説による枝刈りの方法は Magnitude, SNIP のそれぞれで比較を行う．学習時のハイパーパラメータとして，学習回数を 100，バッチサイズを 128，式 (1) で示した α を 1 とする．また，60% の枝刈りを 3 回行い，削除したパラメータ率が 93.6% の時の結果を用いる．

4.2 定量的評価

表 2 に枝刈り率 60% で枝刈りした各手法の mAP を示す．Magnitude ベースの枝刈り時に提案手法で学習したモデルは，従来の宝くじ仮説よりも 4.16pt 精度が向上し，精度低下を抑制していることがわかる．また，知識蒸留付き宝くじ仮説と比較した場合でも，0.64pt の精度向上が確認できる．また，SNIP ベースの枝刈り時に提案手法で学習したモデルは，宝くじ仮説よりも 0.38pt 精度が向上し，精度低下を抑制していることがわかる．また，知識蒸留付き宝くじ仮説との比較した場合でも，0.11pt の精度向上が確認できる．

4.3 定性的評価

図 3 に Magnitude ベースによる検出結果例を示す．図 3(b) の Magnitude による枝刈り手法において，‘motorbike’ と ‘person’ を検出できなくなっている．一方で，図 3(c) の Magnitude による枝刈りと知識蒸留を組み合わせた手法では正しく検出されている．しかし，右上の ‘bottle’ が誤検出されている．図 3(d) による提案手法では，‘motorbike’ と ‘person’ が正しく検出されており，‘bottle’ の誤検出もないことが確認できる．



(a) YOLOv8　　(b) Magnitude

(c) Magnitude+KD　　(d) Ours

図 3：枝刈り率 60%の検出結果

5. おわりに

本研究では，段階的な蒸留を導入した宝くじ仮説にる物体検出モデル枝刈りを提案した．また，評価実験において，提案した段階的な蒸留を行うことで，検出精度が向上することを確認した．今後は，MS COCO 2017 や BDD100K など他のデータセットでの検証や，スクラッチモデルから事前学習したモデルの有効性を検討する．

参考文献

[1] S. Han, *et al.*, “Learning both weights and connections for efficient neural networks. ”, 2015.

[2] S. Ren, *et al.*, “Faster r-cnn: Towards realtime object detection with region proposal networks. ”, In CVPR, 2015.

[3] J. Frankle, *et al.*, “The lottery ticket hypothesis: Finding sparse, trainable neural networks.”,arXiv:1803.03635, 2018.

[4] N. Lee, *et al.*, “Snip: Single-shot network pruning based on connection sensitivity”,2019.

[5] G. Hinton, *et al.*, “Distilling the Knowledge in a Neural Network”, arXiv:1503.02531, 2015.

[6] S. Mirzadeh *et al.*, “Improved Knowledge Distillation via Teacher Assistant.”, arXiv:1902.03393, 2020.

[7] L. Liu *et al.*, “Knowledge Distillation と Pruning を用いた物体検出モデルの圧縮に関する研究.”, 修士論文, 2023.

研究業績

木村秋斗, 早川和希, 森巧磨, 長内淳樹, 平川翼, 山下隆義, 藤吉弘亘, “Grid-wise-attention による物体検出の視覚的説明”, 画像センシングシンポジウム, 2022.