2T-04

複数蒸留モデルのパラメータ探索と最適なモデルの提示

船田毅　櫻井義尚

明治大学大学院先端数理科学研究科　総合数理学部

要旨

蒸留技術はモデル圧縮や知識の抽出において広く用いられ，様々な手法が提案されている．そして，複数の蒸留手法を組み合わせる事で精度が向上することが示されているが，その精度は組み合わせる際のパラメータに大きく依存しており，その設定方針も明らかにされていない．本研究では，複数の蒸留手法を組合せたモデルを用いる際の組み合わせパラメータについて，比較実験を行い，その傾向について分析する．具体的には既存手法である三つの蒸留手法を用いた深層学習モデルにおいて，様々なパラメータでの精度の比較を行う．

1.　はじめに

機械学習における深層学習は画像，言語に限らず，様々な分野において活用され最先端の性能を達成しています．深層学習は複数の層を利用した大規模なネットワークを構築しています．そのような中，モバイル機器や組込みシステムにおいて計算機やストレージが限られた箇所では大規模なネットワークは展開できません．そのためモデルの圧縮化などにはネットワークの刈り込みやモデルの量子化，知識抽出など軽量化した深層学習モデルの学習を目的とした高速化手法が提案されています．知識抽出は能力の高いネットワークから能力の低いネットワークに知識を伝達するもので近年注目され，特に知識蒸留は知識の種類・蒸留戦略・モデル間の構造を用いたモデル圧縮の技術です．そして，近年はモデル圧縮だけでなく知識抽出によるモデル精度向上効果が注目され，蒸留手法を複数掛け合わせて異なる側面から知識抽出を行う手法が提案され始めています．蒸留を複数組み合わせる手法はいくつかのパラメータを用いて各種効果が制御され，タスクに応じたパフォーマンスを行えるようになっています．しかしながら，これらパラメータは精度に影響するにも関わらず，設定方針も明確にされていません．本研究ではパラメータの設定方針を明らかにしタスクに応じたモデルを設定するため，パラメータの分析を行いパラメータの傾向を明かにすると共にタスクに応じた高精度なモデルを追究することを目的としている．

Parameter exploration of multiple distillation models and presentation of the best model

†TakeruFunada Graduate School of Mathematical Sciences, Meiji University

‡YoshitakaSakurai School of Integrated Mathematical Sciences, Meiji University

2.　関連研究

2.1　蒸留

Knowledge Distillation[1]は 2015 年に提案された論文であり，モデル規模が小規模な学習モデルが大規模な学習モデルから知識を獲得する

ことを目的とする手法です．予測精度の良いモデルから得た知識を規模が小規模モデルの学習に利用することで，小規模でありながら大規模モデルに匹敵する精度のモデルを得ることができます．実際の継承方法は，オーソドックスなものにおいて本研究においては深層学習モデルの間で知識伝達する際は最終層の出力を用いる．大規模モデルの最終層ではデータから学習した知識があるためその出力と近づくように小規模モデルの出力を学習させる．そのようにして小規模モデルの挙動を大規模モデルに類似させていく．本研究においては，知識継承の際の手法として用いて実験を行っている．

2.2.　自己蒸留

Kyungyul らが発表した自己蒸留[2]は2019年に発表された蒸留手法で知識継承の際のデメリットであった効率の悪さや適切な大規模モデルを用意する困難さに対して同モデルを用意することで解決する手段を提案した論文です．本研究においてはモデル構築の際に同形状のモデルを用意する際に用いている．

2.3.　複数蒸留

CollaborativeTeacher-StudentLearningviaMultipleKnowledgeTransfer[3]は 2021 年に提案された複数蒸留を組み合わせるアルゴリズムとそれによる複数蒸留による組合せの検証精度と汎用性を検証した論文です．オフライン蒸留の問題点であった，事前学習用の大規模ネットワークによる容量のギャップ問題と蒸留の単一の知識からの継承やインスタンス一貫性の低さに対して，オンライン蒸留と自

己蒸留を組み合わせたモデルアーキテクチャを提案しました．単一の知識からではなく様々な側面から知識継承を行えるアルゴリズムであり，本研究においては蒸留の組み合わせモデル作成の際に用いている

3.実験

3.1. 実験目的

本実験においてはパラメータの設定方針を明らかにし，タスクに応じた最適なモデルを提案することを目的とする．そのために，パラメータの挙動を理解するために実験 1 でパラメータの検証による性質理解，パラメータの強弱による変化を調べるために実験 2 ではパラメータの制御の有無を検証する．その後必要に応じて追実験を行う．

3.2. 実験 1 パラメータの傾向探索

蒸留，自己蒸留，関係性蒸留を組み合わせたモデルを用意し，パラメータを各種変更，パラメータの傾向を分析する．具体的にはモデル学習の際における損失関数を制御している九つのパラメータに対し，ランダム探索と先行研究による比較的精度が高いパラメータ値を基準として探索を行いパラメータの傾向を分析する．

3.3. データセット

本研究においてはモデル精度の実験に CIFAR-10 を用いる．CIFAR-10 は画像分類のためのデータセットで十クラスの画像を含んだ五万枚の訓練画像と一万枚のテスト画像からなります．画像サイズは 32×32 ピクセルで 3 チャンネルのカラー画像です．物体カラー写真であり乗り物や動物などの分類が出来るデータセットであり，データは前処理としてパディングやランダムクロップを使用している．

3.4. 実験手順

九つのパラメータはそれぞれ，蒸留による損失関数 loss を調整するパラメータ 3 種(便宜上，調整パラメータ a, 調整パラメータ b, 調整パラメータ c とする，今後は都度割り振る)，各種損失関数，kd_loss 関係性蒸留 rkd_loss, 自己蒸留 skd_loss で調整するパラメータ 6 種を用いる．

$$loss = akd_{loss} + brkd_{loss} + cskd_{loss}$$

そして，既存手法と比較を行いどのパラメータの組み合わせが良いのか検証する．実験条件としてモデルは ResNet18，epoch 数は 10 とし，最適化手法は確率的勾配降下法，パラメータ範囲は先行研究を加味して定義した．

Table1.　研究事例と実験時のパラメータ

パラメータ名	研究事例	探索範囲
調整パラメータ a	0.1	0.1～1.0
調整パラメータ b	0.05	0.1～1.0
調整パラメータ c	0.09	0.1～1.0
関係性蒸留パラメータ d	25.0	25.0, 50.0, 75.0, 100.0
関係性蒸留パラメータ e	50.0	25.0, 50.0, 75.0, 100.0
関係性蒸留パラメータ f	1.0	0.1～1.0
関係性蒸留パラメータ g	1.0	0.1～1.0
自己蒸留パラメータ h	0.1	0.1～5.0
自己蒸留パラメータ i	1e-6	0.1～5.0

3.5. 実験 2 カリキュラム学習

実験 1 に対してパラメータの制御を弱めた状態で学習した場合を調べる．モデル内損失関数の更新をパラメータを減らした状態で行う．具体的には，実験 1 で用いた調整パラメータ abc による蒸留の損失関数の調整を無くし，それぞれの蒸留効果を並列にモデルに適用する．今まではそれぞれを統合することで効果を適用していたが，独立させた状態での精度への影響を調べる．既存手法と精度を比較し検証を行う．実験条件は実験 1 と同条件のモデルを用いて行った．

4.0 おわりに

本論文ではパラメータによる蒸留への影響とそれによるモデルへの影響を調査し，タスクによる最適なモデルを提案した．蒸留の性質と組み合わせることによる効果の相乗効果を今回の提案で確認し，別なモデルやタスクに対しても適用できるように検証していきたい．

参考文献

[1]GeoffreyHintonVinyals,andJeffDeanOriol:"Distilling the Knowledge in a Neural Network", International Conference on Neural Information Processing Systems,pp.1047-1055 (2015)

[2] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, “Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation,” Int. Conf. Comput. Vis., (2019)

[3]LiyuanSunGou,LanDu,DachengTaoJianping:"Collaborative Teacher-Student Learning via Multiple Knowledge Transfer",(2021)