4U-02

DNNの自己蒸留における学習時間の短縮

高木 純平† 服部 元信‡

山梨大学 大学院医工農学総合教育部† 山梨大学 大学院総合研究部‡

1. はじめに

蒸留は, 教師と呼ばれるディープニューラルネット(DNN)の学習によって獲得した知識をより小さな生徒ニューラルネットの学習に用いることでより高精度な生徒を作成する手法である[1]. また, 作成した生徒を教師としてさらに蒸留を行うことでより高精度な生徒を作成できることもわかっている[2]. しかし, そのような複数世代の蒸留は学習に時間がかかる問題がある. そこで本研究では, 蒸留に必要な世代数を削減し, 学習時間を短縮することを目的とし, 学習途中に得られた最高精度のネットワークを教師として1世代内で蒸留を行う自己蒸留を提案する. 画像分類タスクにおいて, 自己蒸留が従来の蒸留に比べて少ない世代, 短い学習時間で高精度を獲得できることを確認した.

2. 従来手法

本章では, 蒸留と蒸留の単純かつ効果的な改良手法である Born Again Networks(BAN)[2], DNN の性能向上に広く用いられている Data Augmentation (DA)について説明する.

2.1. 蒸留

蒸留は, 教師の学習によって得られた知識を生徒の学習に利用する手法である. 蒸留によって学習した生徒は単純に生徒を学習するよりも良い性能であることがわかっている[1]. 画像分類タスクにおいて, 教師から生徒へ蒸留される知識は DNN の出力である分類確率を用いることが多い. 一般に画像分類タスクでは, 学習データとして画像とその画像の正解クラスを示すラベルのペア集合が与えられる. そのような学習データからは, 画像がどのクラスに属するかの情報しか与えられない. 一方, 教師から蒸留される知識は, 各クラスへの分類確率の違いが示す, 画像が正解クラス以外のどのクラスに似ているかなどの豊富な情報を含んでいる. 正解ラベルに加え, このような情報を知識として生徒の学習に利用することで, より高精度な生徒を作成できる. また, 蒸留を用いた学習は正解ラベルのみの学習に比べて過学習を抑制することもできる.

2.2. Born Again Networks : BAN

蒸留の多くは, 推論時の計算コストやDNNの保持に必要となるメモリ量を削減するために利用される. そのために, 生徒は教師より小さく簡単な DNN が一般に利用される. Furlanello らは, 教師と生徒の大きさが同程度の場合でも蒸留は有効であると考え, 実験によってその効果を証明した[2]. また, Furlanello らは蒸留によって得られた生徒を次世代の教師として複数世代に渡って蒸留を行う(BAN)ことで生徒の性能が向上することを示した[2].

2.3. Data Augmentation : DA

DNN の性能は学習データの量に大きく左右される. しかし, 大量のデータを集めることは困難なことが多い. DA は, 収集したデータに加工を施すことで, データ量の水増しを行う手法である.

3. 研究目的

BAN は複数世代に渡って蒸留を行う必要があるため, 学習に時間がかかる. そこで本研究では, BAN の学習効率を高め, 高精度を得るまでに必要な世代数の削減と学習時間の短縮を研究目的とする.

4. 提案手法

本章では, DA を考慮した自己蒸留手法と BAN と同様に複数世代に渡って自己蒸留を行う手法について説明する.

4.1. 自己蒸留

初めに, 学習データの一部を検証データとして分けて確保する. 確保した検証データはDNNの学習に利用せず, 汎化性能の測定に利用する. 自己蒸留は, 学習時に得られた汎化性能が最も高い時点(汎化 DNN)の出力を教師信号に利用し, 蒸留する手法である. 自己蒸留の目的関数を式(1)に示す.

$$\min_{\boldsymbol{\theta}} L\left(0.5\boldsymbol{y} + 0.5\boldsymbol{y}_v, f(\boldsymbol{x}_{DA}, \boldsymbol{\theta})\right) \quad (1)$$

$\boldsymbol{\theta}$は現在の DNN の重み, $\boldsymbol{y} \in \{0,1\}^n$は正解ラベル, nはクラス数, $\boldsymbol{y}_v$は汎化 DNN の出力, $\boldsymbol{x}_{DA}$は DA によって加工された学習データ, $f(\boldsymbol{x}_{DA}, \boldsymbol{\theta})$は$\boldsymbol{x}_{DA}$を入力した際の DNN の出力, Lは教師信号と DNN の出力のクロスエントロピーをそれぞれ示す.

自己蒸留のアルゴリズムをアルゴリズム 1 に示す. このアルゴリズムでは, 学習データを少数のデータ集合(ミニバッチ)に分け, ミニバッチ単位で 1 度重みを更新する学習を行う. 本研究では, 全てのミニバッチによって更新がなされた回数を epoch としてカウントする. また, 汎化性能の測定は学習時間を考慮し, 1epoch ごとに行う. 学習には各 epoch ごとに乱数値によって異なる DA が施された学習データが使用される.

Reducing Training Time for Distillation of DNN by Self Distillation

†Jumpei Takagi, University of Yamanashi

‡Motonobu Hattori, University of Yamanashi

アルゴリズム 1 自己蒸留

Input
$\boldsymbol{x}$:学習データ, $\boldsymbol{y}$:正解クラス, $\boldsymbol{\theta}$:重み
1:初期化: $\boldsymbol{\theta}, \boldsymbol{\theta}_v, (acc_{max\_v} \leftarrow 0), (\boldsymbol{y}_v \leftarrow \boldsymbol{y})$
2:**for** $e \leftarrow 1$**to**学習$epoch$ 数の上限
3: $\boldsymbol{x}$から$\boldsymbol{x}_{DA}^{e}$を作成
4: $\boldsymbol{tmp} \leftarrow \phi$
5: **for** $b \leftarrow 1$**to**(学習データ数/ミニバッチ)
6: $\boldsymbol{x}_b, \boldsymbol{y}_b \leftarrow \boldsymbol{x}_{DA}^{e}, \boldsymbol{y}$から 1 ミニバッチ取り出し
7: $\boldsymbol{tmp} \leftarrow \{\boldsymbol{tmp}, f(\boldsymbol{x}_b, \boldsymbol{\theta})\}$
8: $\boldsymbol{x}_b, \boldsymbol{y}_b, \boldsymbol{\theta}_v$,式(1)を用いて$\boldsymbol{\theta}$を更新
9: $acc_v \leftarrow \boldsymbol{\theta}$の汎化性能
10: **if** $acc_v > acc_{max\_v}$ **then**
11: $acc_{max\_v} \leftarrow acc_v$, $\boldsymbol{\theta}_v \leftarrow \boldsymbol{\theta}$, $\boldsymbol{y}_v \leftarrow \boldsymbol{tmp}$
12: **if** 学習終了条件を満たす **then** Break

自己蒸留はミニバッチ学習時に保持した出力を教師信号に利用するため, $\boldsymbol{y}_v$と$\boldsymbol{x}_{DA}$が正確に対応していないことや正確な汎化 DNN の出力値ではないなどの特徴がある.

4.2. BAN + 自己蒸留

本節では, BAN と自己蒸留を組み合わせた手法を解説する. 4.1 節に示した方法で得られた DNN を 1 世代目($\boldsymbol{\theta}_1$)として BAN と同様に複数世代に渡って蒸留を行う. k世代目の自己蒸留の目的関数を式(2)に示す. 第 1 項は式(1)と同様であり, 第 2 項が BAN の最小化を行う項である. 2 世代目以降は, 式(1)の代わりに式(2)を使用したアルゴリズム 1 を用いて学習を行う.

$$\min_{\boldsymbol{\theta}_k}\{L(0.5\boldsymbol{y} + 0.5\boldsymbol{y}_v, f(\boldsymbol{x}_{DA}, \boldsymbol{\theta}_k)) + L(f(\boldsymbol{x}_{DA}, \boldsymbol{\theta}_{k-1}), f(\boldsymbol{x}_{DA}, \boldsymbol{\theta}_k))\} \quad (2)$$

5. 計算機実験

自己蒸留の有効性を示すために BAN と BAN+自己蒸留の 2 手法の画像分類実験による比較を行った. 本章では, 行った実験の条件と実験によって得られた結果について述べる.

5.1. 実験条件

実験には, 20 層の ResNet[3]を使用し, 100 クラス, 32×32pixel の一般物体画像データセット Cifar100 を使用した. Cifar100 には, 学習画像として 5 万枚, テスト画像として 1 万枚の画像が用意されている. 学習データ内から 5 千枚ランダムに抜き出し, 検証データとして利用した. DA として画像を確率 50%でランダムに鏡像反転, 0~4pixel 上下左右にランダムにずらす加工を行った. 学習率は初期値を 0.1, 最小値を 0.001 として, 汎化性能が 20epoch 向上しなければ重みをその世代の学習中に得られた最も汎化性能が高い重みにリセットし, 0.1 倍した. また, 汎化性能が 40epoch 向上しなければ次世代の学習に遷移させた. テスト性能の測定や BAN における次世代の教師には, その世代の学習中に得られた最も汎化性能の高い重みを利用した. その他の条件は以下に示す通りである. ミニバッチサイズを 100, 慣性項を 0.9, 重み減衰として1×10^{-4}を使用した.

5.2. 実験結果

実験結果を図 1 に示す. 第 1 縦軸はテスト画像の分類精度を, 第 2 縦軸はその世代を学習するまでに必要な累積学習時間を, 横軸は世代を示している. 2 つの学習手法の累積学習時間の有意差を t 検定によって検定した結果, 同じ世代数で有意差は確認できなかった. 同様にテスト精度について検定した結果, 5, 6 世代目間を除いた全ての世代で BAN に比べて自己蒸留のテスト精度が有意に高いことがわかった. また, 自己蒸留 2 世代目のテスト精度は BAN3～7 世代目のテスト精度より高く, BAN3, 4, 7 世代目との間に有意差が確認できた. この結果から, 自己蒸留を用いることで少ない世代, 短い学習時間で高精度を得られることがわかった.

BAN に比べて自己蒸留の汎化性能が高い一因として, 自己蒸留は$\boldsymbol{y}_v$と$\boldsymbol{x}_{DA}$が正確に対応していないため, 教師信号にノイズが入ることが考えられる. 教師信号にノイズを加え, 蒸留することで生徒の汎化性能が向上する[4]ことから, 自己蒸留は BAN に比べて汎化性能が高いと考えられる.

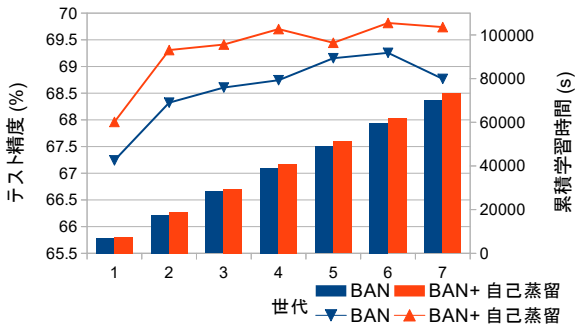

図 1: 10 試行の平均結果, 折れ線グラフ:テスト精度, 棒グラフ:累積学習時間

6. まとめ

本研究では, BAN に必要な世代数, 学習時間の短縮を目的とし, 1 世代内で蒸留を行う自己蒸留を提案した. 計算機実験の結果から, 自己蒸留は高精度な DNN を少ない世代, 短い学習時間で作成でき, BAN の学習時間の削減に貢献していることを確認できた.

参考文献

[1] G.Hinton, O.Vinyals, and J.Dean. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.
[2] T.Furlanello, Z.C.Lipton, M.Tschannen, L.Itti, and A.Anandkumar. Born Again Neural Networks. In Metalearn 2017 NIPS Workshop, pp. 1-5, 2017.
[3] K.He, X.Zhang, S.Ren, and J.Sun. Identity mappings in deep residual networks, In European Conference on Compurter Vision, pp.630-645, 2016.
[4] B.B.Sau, V.N.Balasubramanian. Deep model compression: Distilling knowledge from noisy teachers. arXiv:1610.09650, 2016.