F-020

Teacher Assistant 及び中間層を模倣する Distillation による ニューラルネットワークのモデル圧縮
Neural Network Model Compression by using Teacher Assistant and Distillation with Hint Learning

森川拓海 † Takumi Morikawa
亀山啓輔 ‡ Keisuke Kameyama

1. まえがき

近年、深層学習は人工知能分野で優れた成果を上げている。特に、畳み込み層を多く持つ畳み込みニューラルネットワーク (CNN) の学習を行うことで、優れた識別性能を獲得することに成功している。また学習手法の開発 [4] や、GPGPU などのハードウェアの性能向上により、規模の大きいニューラルネットワークを学習させ、用いることが容易になってきている。しかしながら、このような規模の大きい CNN は高い識別性能を持つが、スマートフォンなどのデバイスでの使用には向かない。なぜなら、小型のデバイスは計算資源に乏しく規模の大きい CNN を実装することが難しいからである。従って、スマートフォンのような小型のデバイスにも実装することが容易であり、識別性能が高い CNN を実現することが必要とされている。

このような問題を解決する手法の 1 つに Distillation がある [3]。Distillation では、転移学習によって優れた規模の小さい CNN モデルを獲得することができるが、問題点も存在する。それは、教師モデルと生徒モデルを構成するパラメータ数に大きな差がある場合、Distillation による識別性能の改善が行われないことである。この問題の解決策として、教師モデルと生徒モデルの中間の規模である Teacher Assistant(TA) モデルを用いる手法 [6] が存在する。しかし、同手法では出力層の模倣による Distillation しか行っておらず、中間層を模倣してより教師モデルの入出力関係を的確に転移する Distillation を行うなどの工夫を施す余地が存在する。本研究の目的は、このような教師モデルと生徒モデルを構成するパラメータ数に大きな差がある場合でも、モデル圧縮した CNN の識別性能の改善を行うことである。

本研究では、TA モデルを用いて、中間層と出力層を模倣する多段階の Distilaltion を提案する。始めに教師モデルを用いて TA モデルに対して中間層と出力層を模倣する Distillaiton を行う。そして、TA モデルを教師モデルとして生徒モデルに対して同様の Distillation を行う。

実験では、CIFAR-10 データセット [5] を用いた画像分類を提案手法と既存手法によって行い、その識別性能を評価する。また実験は複数回繰り返し、最も優れた識別性能、平均の識別性能、性能の標準偏差を用いて評価を行う。

†筑波大学理工情報生命学術院システム情報工学研究群
‡筑波大学システム情報系

2. 関連研究と課題

Caruana らは大規模なアンサンブルモデルのようなモデルの規模が大きく推論時間を多く必要とする高性能なモデルから単一で小型であり、推論時間が短い高性能なモデルを生成するモデル圧縮手法を提案した [2]。

Hinton らは Caruana らのモデル圧縮手法 [2] に改良を加えたモデル圧縮手法である Distillation を提案した [3]。Distillation では学習済みの規模の大きい高性能なモデルを教師モデル、目的の小型のモデルを生徒モデルとして、教師モデルの知識を生徒モデルに転移することで生徒モデルの性能を改善させる。Distillation では、教師モデルと生徒モデルのロジットを u_{T_i} 、u_{S_i}、温度パラメータを T とすると、温度付き softmax 関数

$$y_{T_i} = \frac{\exp(u_{T_i}/T)}{\Sigma_j \exp(u_{T_j}/T)} \quad (i = 1, \ldots, n) \quad (1)$$

$$y_{S_i} = \frac{\exp(u_{S_i}/T)}{\Sigma_j \exp(u_{S_j}/T)} \quad (i = 1, \ldots, n) \quad (2)$$

を用いて教師モデルと生徒モデルのソフトな出力 y_{T_i}, y_{S_i} を得る。Distillation はこのソフトな出力による損失関数 E_{soft} と正解ラベルによるハードな損失関数 E_{hard} を組み合わせた損失関数

$$E_d = \frac{\lambda_1 T^2 E_{soft} + \lambda_2 E_{hard}}{\lambda_1 + \lambda_2} \quad (3)$$

によって学習される。ここで、λ_1, λ_2 は任意の重みパラメータである。また、式 3 のソフトターゲットである E_{soft} は識別するクラス数を n とすると、

$$E_{soft} = \sum_{i=1}^{n} y_{T_i} \log y_{S_i} \quad (4)$$

あるいは、

$$E_{soft} = \sum_{i=1}^{n} y_{T_i} \log \frac{y_{T_i}}{y_{S_i}} \quad (5)$$

が使用されることがある。Hinton らは式 4 の E_{soft} を用いた。また、式 5 のような Kullback-Leibler 情報量を用いて Distillation を行っている研究も存在する [7]。E_{hard} には訓練データの正解ラベルを t とすると、

$$E_{hard} = \sum_{i=1}^{n} t_i \log y_{S_i} \quad (6)$$

のクロスエントロピー誤差の式で表される。

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

Romero らは Hinton らの出力層の模倣 [3] に加えて、中間層の出力を模倣することで多くの層を持つ生徒モデルの性能を改善する手法を提案した [1]。この手法では、層数が少なく規模の大きいモデルを教師モデル、層数が多く規模が小さいモデルを生徒モデルとして扱う。

学習では、始めに、教師モデルのある中間層を Hint 層、生徒モデルのある中間層を Guided 層に設定し、Hint 層と Guided 層の出力の次元を合致させるための regressor と呼ばれる畳み込み層を Guided 層のすぐ後に追加する。Hint 層の出力を z_h、Guided 層の出力を z_g、regressor の出力を r とする。また Hint 層、Guided 層、regressor のパラメータをそれぞれ $\boldsymbol{W}_{hint}$、$\boldsymbol{W}_{guided}$、$\boldsymbol{W}_r$ とすると、

$$E_h(\boldsymbol{W}_{guided}, \boldsymbol{W}_r) = \frac{1}{2}||z_h(s; \boldsymbol{W}_{hint}) - r(z_g(x; \boldsymbol{W}_{guided}); \boldsymbol{W}_r)||^2 \quad (7)$$

のような損失関数で生徒モデルの Guided 層までを学習する。このような中間層までの事前学習を Hint 学習と呼ぶ。Hint 学習終了後、Guided 層までのパラメータの値を初期値として、モデル全体に対して、Hinton の Distillation による学習を行う。

Distillation は優れたモデル圧縮手法である一方で、課題も存在する。例えば、教師モデルと生徒モデルを構成するパラメータ数に大きな差がある場合に、Distillation は生徒モデルの性能向上に効果的に作用しない。この問題を解決するために、Mirzadeh はらは Teacher Assistant Knowledge Distillation(TAKD) と呼ばれる新しい Distillation のフレームワークを提案した [6]。TAKD では、教師モデルと生徒モデルの中間規模であるモデルを TA モデルとして 2 段階の Distillation を行う。始めに、TA モデルを生徒モデルとして教師モデルを使用して Hinton の Distillation[3] を行う。学習終了後、TA モデルを教師モデルとして目的の生徒モデルに対して Hinton の Distillation を行う。

2.1. 既存手法の課題

Distillation は教師モデルの入出力関係を模倣することで、教師モデルの持つ知識を生徒モデルに転移することができる。しかし、教師モデルと生徒モデルを構成するパラメータ数に大きな差がある場合、Distillation による性能向上は見込めない [6]。そのようなモデルのパラメータ数に差がある場合、Mirzedeh らが提唱した TAKD[6] で Distillation を機能させることができるが、この手法では中間層の出力を模倣する Distillation は行っていない。このため、パラメータ数が少ないが層数の多い生徒モデルを学習させる場合には、その効率に課題を残している。

2.2. 研究の目的

本研究では規模の小さなニューラルネットワークモデルが規模が大きく高性能なニューラルネットワークと同等の性能を持つための学習方式を開発することを目的としている。一般的に規模の小さなモデルは重みパラメータの初期値を乱数によってランダムに生成して学習を開始した場合、学習の効率が悪く、優れた性能を獲得することができない。特に規模の小さなニューラルネットワークモデルでも，層数の多いモデルを用いる場合、誤差逆伝播法による学習を行っても複雑な入出力関係を表現する適切な重みパラメータを獲得することは難しい。このようなパラメータ数は少ないが、層数の多いネットワークを効率的に学習させる手法の開発および、その手法を用いてモデル圧縮を行うことが本研究の目的である。

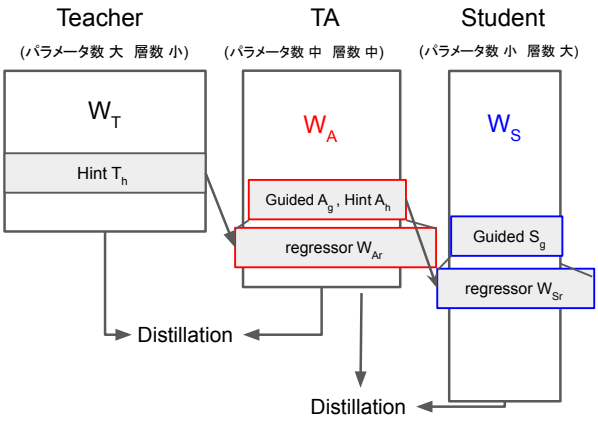

図 1: 提案手法

3. 提案手法

上記の目的を実現させるために、TAKD 手法 [6] と、中間層および出力層を模倣する Distillation[1] を組み合わせて用いることを提案する。教師モデルにはパラメータ数が大きく、層数の少ない高性能なモデルを使用し、反対に、生徒モデルにはパラメータ数が小さく、層数の多いモデルを対象とする。そして TA モデルには教師モデルと生徒モデルの中間規模のモデルを使用する。始めに TA モデルを生徒モデルとして扱い、教師モデルを用いて Hint 学習と Hinton の Distilaltion を TA モデルに対して適用する。こうして最適化された TA モデルを教師モデルとして扱い、目的の生徒モデルに対して同様に Hint 学習と Hinton の Distillation を適用する。図 1 は提案手法の模式図である。

以下に提案手法アルゴリズムの疑似コードを示す。提案手法を説明するために, 以下の変数を使用する。

- 教師モデルの重みパラメータ:$\boldsymbol{W}_T$
- TA モデルの重みパラメータ:$\boldsymbol{W}_A$
- 生徒モデルの重みパラメータ:$\boldsymbol{W}_S$
- TA モデルや生徒モデルの Hint 学習後の重み:$\boldsymbol{W}^*$
- TA モデルの regressor の重みパラメータ:$\boldsymbol{W}_{A_r}$
- 生徒モデルの regressor の重みパラメータ:$\boldsymbol{W}_{S_r}$
- 教師モデルの Hint 層:T_h
- TA モデルの Hint 層:A_h
- 生徒モデル Guided 層:S_g
- TA モデルの Guided 層:A_g
- Hint 学習に用いる損失関数 (7):E_h

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

- Distillation に用いる損失関数 (3):E_d

Algorithm 1 Algorithm for Joint Teacher Assistant and Hint Learning

Input: $\boldsymbol{W}_T, \boldsymbol{W}_A \boldsymbol{W}_S, \boldsymbol{W}_{Ar}, \boldsymbol{W}_{Sr}, T_h, A_h, A_g, S_g$
Output: $\boldsymbol{W}_S^*$

// first hint learning step
$\boldsymbol{W}_{hint} \leftarrow \{\boldsymbol{W}_T^{(1)}, \ldots, \boldsymbol{W}_T^{(T_h)}\}$
$\boldsymbol{W}_{guided} \leftarrow \{\boldsymbol{W}_A^{(1)}, \ldots, \boldsymbol{W}_A^{(A_g)}\}$
Initialize $\boldsymbol{W}_{A_r}$ to small random values
$\boldsymbol{W}_{guided}^* \leftarrow \arg\min_{\boldsymbol{W}_{guided}} E_h(\boldsymbol{W}_{guided}, \boldsymbol{W}_{A_r})$
$\boldsymbol{W}_A^* \leftarrow \arg\min_{\boldsymbol{W}_A} E_d(\boldsymbol{W}_A)$
// second hint learning step
// $\boldsymbol{W}_A^*$ is new teacher
$\boldsymbol{W}_{hint} \leftarrow \{\boldsymbol{W}_A^{*(1)}, \ldots, \boldsymbol{W}_A^{*(A_h)}\}$
$\boldsymbol{W}_{guided} \leftarrow \{\boldsymbol{W}_S^{(1)}, \ldots, \boldsymbol{W}_S^{(S_g)}\}$
Initialize $\boldsymbol{W}_{S_r}$ to small random values
$\boldsymbol{W}_{guided}^* \leftarrow \arg\min_{\boldsymbol{W}_{guided}} E_h(\boldsymbol{W}_{guided}, \boldsymbol{W}_{S_r})$
$\boldsymbol{W}_S^* \leftarrow \arg\min_{\boldsymbol{W}_S} E_d(\boldsymbol{W}_S)$

このようにして導き出された $\boldsymbol{W}_S^*$ を用いて画像分類などの識別を行うことで規模の小さいモデルでも教師モデルと類似した入出力関係を獲得することを目的としている.

3.1. 提案手法から期待される効果

この手法から期待される効果として、問題として挙げたパラメータ数が少ないが層数の多いモデルの学習効率の改善がある。 TAKD の手法 [6] を用いることで教師モデルと生徒モデルを構成するパラメータ数に大きな差がある場合でも Distillation による性能の改善が見込めることが示されている。また Romero らの中間層と出力層に着目した Distillation 手法 [1] では、パラメータ数を小さくするが層数を多くすることで教師モデルと同等の性能改善を見込めることが示されている. したがって、これらの手法を組み合わせた本研究の提案手法は、段階的に Distillation を行うことでパラメータ数の差に影響を受けず、層数を多くすることでパラメータ数を小さくした場合でも性能の向上と学習効率の改善を期待することができると考えられる。

4. 実験

4.1. 目的

実験の目的は 2 つあり、第 1 に提案手法が既存手法と比較して、生徒モデルの識別性能をどの程度向上させるかを検証することである。第 2 に、モデル圧縮を行う場合、既存手法と提案手法を比較して、生徒モデルに対して安定した識別性能を発揮させることができるかを検証することである。ここで比較対象とする既存手法は Hinton らの Distillation[3]、Romero らの Hint 学習を用いた中間層を模倣する Distillation[1]、Mezadeh らの Teacher Assistant モデルを用いた Distillaion[6] の 3 つの既存手法と提案手法を比較する。以上 2 つの検証を行うために教師モデルと生徒モデル、TA(Teacher Assistant) モデルを用いて提案手法と既存手法の Distillation によるモデル圧縮を行う。

4.2. 評価基準

実験の評価基準として、提案手法と既存手法による Distillation を用いた生徒モデルの学習を複数回行った時のテストデータに対する最良の識別性能、平均識別性能、標準偏差の 3 種類の評価を行う。これらの評価から、提案手法が既存手法に比べて、性能の改善を見込めるか、平均的に優れた性能を小さいモデルで実現することができるかを検証する。

4.3. 実験条件

本実験は CIFAR-10 データセット [5] のテストデータに対する識別性能を前述した評価基準で検証する。学習データとして 50,000 枚の訓練データから 10,000 枚を検証データセットとして扱う。識別性能としてテストデータに対するニューラルネットワークの推論の正解率を用いる。

本実験で用いる CNN モデルは全ての実験を通して表 1 のものを用いる。教師モデルを CNN8、TA モデルを CNN12、生徒モデルを CNN モデル 16 と呼ぶ。

表 1: 実験に用いた CNN モデル

CNN8	CNN12	CNN16
Conv $((3\times3)\times64)\times2$ MaxPool (3×3)	Conv $((3\times3)\times32)\times3$ MaxPool (3×3)	Conv $((3\times3)\times16)\times4$ MaxPool (3×3)
Conv $((3\times3)\times128)\times2$ MaxPool (3×3)	Conv $((3\times3)\times64)\times3$ MaxPool (3×3)	Conv $((3\times3)\times32)\times4$ MaxPool (3×3)
Conv $((3\times3)\times256)\times2$ MaxPool (3×3)	Conv $((3\times3)\times128)\times3$ MaxPool (3×3)	Conv $((3\times3)\times64)\times4$ MaxPool (3×3)
Conv $((3\times3)\times512)\times2$ MaxPool (3×3)	Conv $((3\times3)\times256)\times3$ MaxPool (3×3)	Conv $((3\times3)\times128)\times4$ MaxPool (3×3)
FC (128)	FC (128)	FC(128)
FC(10)	FC(10)	FC(10)
パラメータ数:4, 760, 010	パラメータ数:1, 996, 042	パラメータ数:707, 322
Hint 層:第 4 層,Guided 層:なし	Hint 層:第 6 層,Guided 層:第 6 層	Hint 層:第 8 層,Guided 層:第 8 層

表 1 内の Conv$((h\times h)\times C)$ は畳み込み層のフィルタサイズ $(h\times h)$、出力チャンネル数 C を表している。また、MaxPool$(H\times H)$ はマックスプーリングの小正方領域 $(H\times H)$ を表し、FC(n) は全結合層のユニット数を表す。実験では層数が多いモデルを扱うので学習の効率化を行うため、それぞれの畳み込み層の後に Batch Normalization[4] を行う層を追加している。また CNN の最後の層である FC(10) では活性化関数に Softmax 関数を用いる。 その出力をクラス分類の確率値として扱い識別性能として評価する。Hint 学習を行う場合、Hint 層、Guided 層はそれぞれの CNN モデルの入力層と出力層の中間に位置する層を指定する [1]。

4.4. 結果

4.4.1. 識別性能の比較

既存手法と提案手法を生徒モデルに適用した場合の、各エポック時点の検証データセットに対する平均識別性能を図 2 に示す。図 2 の「Baseline BP」は生徒モデル (CNN16) に対して誤差逆伝播法を用いた教師あり学習を行った場合の識別性能を表している。また「Distillation」、「Hint learning & Distillation」、「TAKD」はそれぞれ Hinton らが提案した Distillation 手法 [3]、Romero らが提案した Hint 学習を用い

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

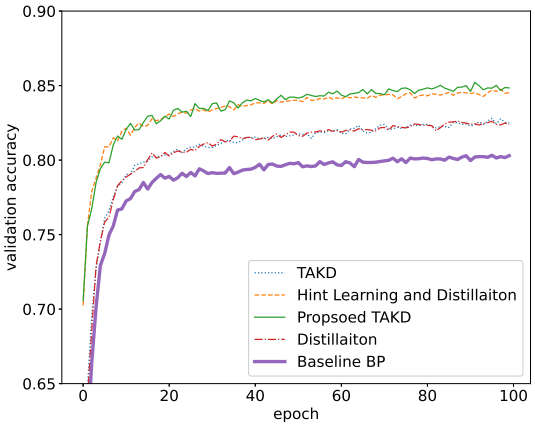

図 2: 提案手法と既存手法による生徒モデル (CNN16) の識別性能

た Distillation[1]、Mirzadeh らが提案した TAKD 手法 [6] を用いた教師あり学習の識別性能を表している。そして「Proposed TAKD」が本論文の提案手法である。本論文の提案手法が最も高い分類能力を生徒モデルに獲得させる結果となった。また、Hint 学習を用いた手法が Hinton の Distillation のみを行う手法と比べて高い識別性能を獲得させる結果となった。

4.4.2. 識別性能と安定性

各モデルの平均識別性能 (Average Accuracy)、最も優れた性能 (Best Accuracy)、性能の標準偏差 (Standard Deviation) の結果を表 2 に示す。既存手法と比較すると提案手法が生徒モデルに対しての Distillation 手法として最も優れた識別性能、平均識別性能ともに、最も高い性能を記録した。また提案手法を適用した場合の生徒モデルの性能は約 7 倍のパラメータ数を持つ教師モデルの性能とほぼ同等になった。標準偏差では、誤差逆伝播法のみを用いた場合と比べて Distillation を行うことで性能のばらつきが小さくなる結果となった。

表 2: CIFAR-10 のテストデータに対する識別性能

Algorithm	Average Accuracy	Best Accuracy	Standard Deviation
Baseline BP Teacher(CNN8)	**85.59%**	**86.65%**	**0.00477**
Baseline BP TA(CNN12)	84.50%	85.22%	0.0038
Baseline BP Student(CNN16)	80.53%	81.18%	0.00419
Distillation(CNN16)	82.89%	83.52%	0.00319
Hint learning & Distillation (CNN16)	84.74%	85.24%	**0.00310**
TAKD(CNN16)	82.92%	83.57%	0.00320
Proposed TAKD(CNN16)	**85.43%**	**85.97%**	0.00317

4.5. 考察

上記のような結果の要因に、提案手法が層数の多い生徒モデルに対してより効率の良い学習を可能にさせたことが考えられる。提案手法では、生徒モデルと教師モデルの層数やパラメータ数に差が少ない状態で Hint 学習と Distillation を行うことができる。これにより、Distillation の問題であるモデルを構成するパラメータ数の差が大きい場合と層数の多いモデルに対する学習の効率化の改善がなされたと考えられる。

5. まとめと今後の課題

5.1. 提案手法と成果

本論文では、層数が多くパラメータ数の少ない CNN の学習効率を改善する方法として、Teacher Assistant(TA) を用いた中間層と出力層の出力を模倣する Distillation 手法を提案した。実験では、提案手法を既存手法と比較した際に、生徒モデルにおける優れた識別性能の改善を実現し、約 1/7 倍のパラメータ数のニューラルネットワークへの圧縮が可能であることを示した。

5.2. 今後の課題

多様なデバイスで深層学習を応用するためには更にパラメータ数を小さくした場合でも教師モデルと同程度の性能を出す必要があると考えている。この圧縮率を高くする上で最適な Distillation の方法は未だに明らかではない。また、Distillation 以外にもモデル圧縮を行う手法は存在し、それらの手法を組み合わせた場合に圧縮率にどのような変化が起こるのかを検証した研究は少ない。これらのことを考慮して、圧縮率の向上を目指すために定量的に、そして定性的にも Distllation を理解することを今後の課題とする。

参考文献

[1] S.E. Kahou A. Chassang C. Gatta A. Romero, N. Ballas and Y. Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[2] R.Caruana C Buciluǎ and A. Niculescu-Mizil. Model compression. In *Proc. the 12th ACM SIGKDD Int'l Conf on Knowledge discovery and data mining*, pp. 535–541, 2006.

[3] O. Vinyals G Hinton and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[4] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc.ICML*, pp. 448–456, 2015.

[5] A. Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

[6] A. Li N. Levine A. Matsukawa S.I. Mirzadeh, M. Farajtabar and H. Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proc. the AAAI Conf. on AI*, Vol. 34, pp. 5191–5198, 2020.

[7] T.M. Hospedales Y. Zhang, T. Xiang and H.Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on CVPR*, pp. 4320–4328, 2018.

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.