ロスレスAI：軽量化前後の推論結果同一性を担保した組込みAI

Lossless AI: Toward Guaranteeing Consistency between Inferences before and after Compression

奥野 智行 Tomoyuki Okuno
中田 洋平 Yohei Nakata
石井 育規 Yasunori Ishii
築澤 宗太郎 Sotaro Tsukizawa

要　旨

計算リソースの限られるエッジデバイスを用いて実時間で認識処理を行うためには，ディープラーニング認識モデルの軽量化が必要である．しかし，従来の軽量化手法は認識精度劣化の抑制にのみ着目しており，個々のサンプル単位で見ると，たとえ精度劣化が小さくても軽量化前後で推論結果が変化することがある．こうした変化は，軽量化前に想定していない挙動を招き，製品の品質保証にとって重大な課題となりうる．そこで筆者らは，軽量化前後で推論結果の同一性を担保する組込みAIである「ロスレスAI」を提案する．本稿では，軽量化手法としてパラメータのビット数を減らす量子化を対象とし，教師モデルから生徒モデルへの知識蒸留の枠組みに基づいた学習手法を提案する．提案手法により，画像分類問題における量子化前後の各モデルの推論結果の一致率を2.6ポイント改善（不一致を約60 %抑制）できることを確認した．

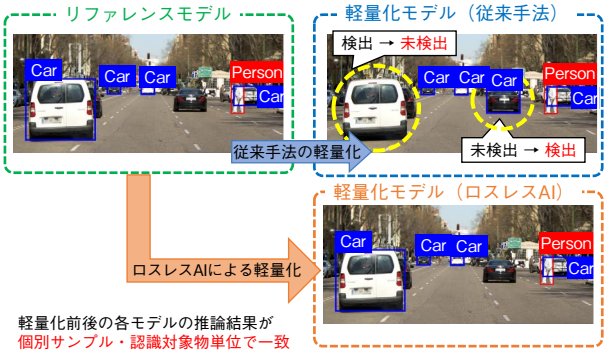
Abstract

Deep learning model compression is necessary for real-time inference on edge devices, which have limited hardware resources. Conventional methods have only focused on suppressing degradation in terms of accuracy, and the inference results may change when we focus on individual samples or objects. Such a change may be a crucial challenge for the quality assurance of products because of unexpected behavior on edge devices. Therefore, we propose a concept called "Lossless AI" to guarantee consistency between the inference results of reference and compressed models. In this paper, we propose a training method using a knowledge distillation framework, which transfers knowledge from the teacher to the student. We demonstrate our method suppresses inferred class mismatching between reference and quantized models by about 60 %.

1. はじめに

ディープラーニングは，物体認識や検出，顔認証，音声認識，翻訳などさまざまなアプリケーションに広く用いられている[1]．入出力やネットワークの遅延も鑑みて，実時間で処理するためには，クラウドではなく，監視カメラやマイクシステムなどに搭載されるエッジデバイスで認識処理を行うこと（組込みAI）が有効である．エッジデバイスはメモリや処理能力などの計算リソースが限られるため，認識モデルの軽量化が必要である．軽量化後に認識精度を維持するため，高性能な計算機で学習された浮動小数点モデルをリファレンスモデルとすることが多い．

軽量化手法は，モデルのパラメータ数を減らす枝刈りや，パラメータのビット数を減らす量子化などがある[2]．これらの手法は軽量化前後の認識精度劣化を抑えることを目指して研究が行われている．しかし，たとえ軽量化モデルがリファレンスモデルと同等の認識精度であっても，個々のサンプル単位の推論結果は軽量化前後で変化する場合がある．例えば**第1図**に示す車と人の検出イメージにおいて，従来手法による軽量化モデルはリファレンスモデルと同じく画像中で4台の車を検出できているが，検出できている車はリファレンスモデルと異なる．すなわち，個々の車に対する推論結果は変化しており，結果的にこの画像サンプルに対する軽量化前後のモデルの挙動は異なる．



第1図　従来手法とロスレスAIの車と人の検出イメージ
Fig. 1　Conceptual examples of conventional compression methods and lossless AI

このような軽量化によるモデルの挙動の変化は，開発コストの観点で重大な課題となりうる．例えば，リファレンスモデルの学習時に想定していた挙動が軽量化によって変化し，想定外の劣化課題が発生することで，エッジデバイスで要求仕様を満たさなくなる可能性がある．また，こうした劣化課題への対策として追加の学習や評価などの手戻りが発生する．さらに，対策によって劣化課題を改善できたとしても，モデルの挙動が変化することで別のサンプルの推論結果に影響を及ぼす可能性があるなど，実用化に向

けた品質保証が困難となる．以上の問題は，当社事業のうち先進運転支援システム（ADAS）などの車載機器や，製造現場向けをはじめとするロボティクスにおける認識といった，特に厳しい品質水準が求められる応用において問題となると考えられる．

そこで筆者らは，軽量化前後で推論結果の同一性を担保する組込みAIである「ロスレスAI」を開発している．ここで言う推論結果の同一性とは，画像分類問題においては推論クラス同士の一致率，物体検出問題においてはバウンディングボックスの重なり率など，タスクごとにさまざまな指標で評価できる．こうした指標で見たとき，ロスレスAIでは個々のサンプル単位の推論結果までリファレンスモデルに近い軽量化モデルを実現でき，品質保証に必要な開発コストの大幅な削減が期待できる．本稿では，画像分類問題を対象とし，軽量化手法としてパラメータのビット数を減らす量子化を行う．教師モデルから生徒モデルへの知識蒸留の枠組みに基づき，モデル学習を2段階に分けた知識蒸留手法を提案し，その評価結果を述べる．

2. 画像分類問題における量子化前後の各モデルの挙動の違い

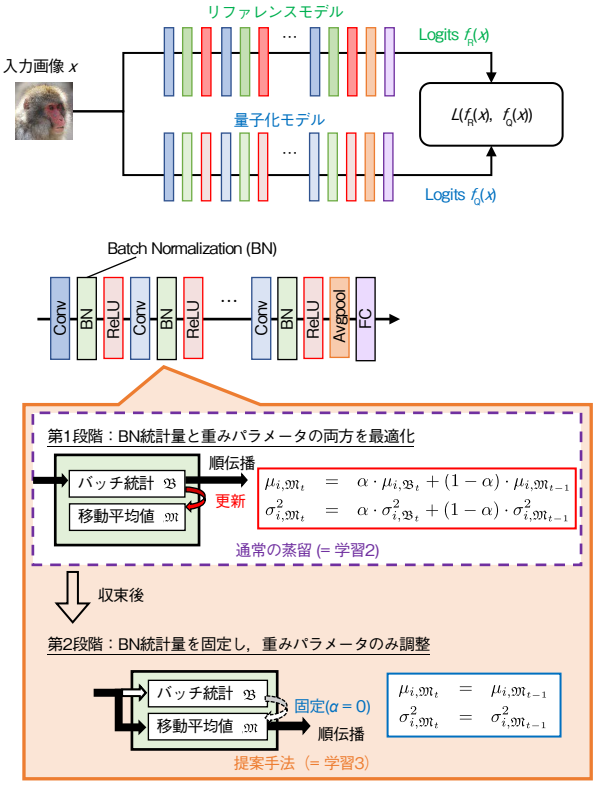
量子化モデルの認識精度を向上させるための学習手法として，例えば量子化を考慮した学習（Quantization-aware training: QAT）[3] や，蒸留（Knowledge distillation）[4] といった手法が挙げられる．QATは，勾配をstraight-through estimator[5]で近似して逆伝播（でんぱ）し，量子化を浮動小数点精度でモデル化することで，認識モデルを量子化しながら学習する手法である．また，蒸留は，大規模な教師モデルの出力(soft target)と小規模な生徒モデルの出力の差を損失関数へ含めることで，教師モデルの知識を生徒モデルへ伝達させる手法である．一般的には，ネットワークの深さや幅などが小さいモデル，もしくは枝刈りや量子化で軽量化されたモデルなどを生徒モデルとして，蒸留を適用することが多い．画像分類問題は，評価データセットの正解ラベル（Ground truth: GT）に対する正答率（認識精度）で評価する．以上の手法はいずれも認識精度を向上させる学習手法であり，量子化前後で認識精度が維持される手法も提案されている[6].

実用化に向けた品質保証の観点では，量子化前後の推論結果を個々のサンプル単位で一致させる必要があり，従来手法は推論結果を一致させる観点では不十分である．例えば，本稿の対象とする画像分類問題においては，量子化モデルの推論結果（Quant.）とリファレンスモデルの推論結果（Ref.）との一致率で推論結果の一致度合いを評価できる．**第2図**に示すように，量子化前後で推論結果が異なる不一致サンプルは，量子化による劣化，改善，誤答クラス間の変化の3パターンに分類できる．従来の量子化では認識精度（正答率）の向上を目指すため，正答・誤答にのみ着目し，量子化前後の一致・不一致は考慮しない．不一致サンプルは，正解ラベルに対する正答・誤答に関わらず，量子化前後でモデルの挙動が異なることで生じる．そこで，ロスレスAIでは，品質保証上課題となる不一致サンプルの数を減らすことで，正答率だけでなく一致率も向上させることを目的とする．

第2図　不一致サンプルの例
Fig. 2　Examples of mismatched patterns

3. 蒸留の枠組みを用いたロスレスAI

本章では，量子化前後の推論結果を揃（そろ）えるために，**第3図**に示すような蒸留を利用した学習手法について述べる．提案手法では，各クラスの推論尤度（ゆうど）の差を小さくし，推論結果を揃えるために，量子化前後の各



第3図　提案手法の概念図
Fig. 3　Illustrations of the proposed training method

モデルのLogits（f_R, f_Q）同士のmean-square error loss（MSELoss）を損失関数として学習する．また，知識を転移しやすくするために，Batch normalization（BN）層[7]が正規化処理に使用するBN統計量（平均と分散）を学習の途中から移動平均値に固定する．

BN層には，入力バッチを正規化することで，深いネットワークでも認識精度や汎化性能を向上させる効果がある．チャネルiに対して，BN層は統計量である平均μ_iと分散σ_i^2を用いた正規化処理とパラメータβ_iとγ_iを用いたアフィン変換からなり，

$$y_i = \gamma_i \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \varepsilon}} + \beta_i \quad \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

x_i：input tensor，y_i：output tensor，ε：constant

と表せる．計算に用いられる統計量として，学習時には，BN層への入力テンソルの平均と分散を計算して得られるバッチ統計量が，推論時には，移動平均統計量が使われる．ここで，移動平均統計量とは，モーメントαをパラメータとする移動平均を利用して，モデル学習時のバッチ統計量で更新された値のことを指す．

Adaptive BN[8]は，推論時に異なる環境（ドメイン）ごとにBN統計量を切り替える方法である．この方法では，分類結果に関する情報は畳み込み層などの重みパラメータに依存し，ドメインに関する情報はBN統計量に依存するという仮説に基づき，推論時にドメインごとに計算されたBN統計量を用いて推論を行う．同様の枠組みは教師なしドメイン適応[9]や異なるカメラ間での人物再照合[10]などへ応用されており，いずれも認識精度の向上に成功している．量子化前後のモデルの違いはドメインの違いの一種と考えられる．そこで，上記の仮説を量子化前後の分類結果を一致させるロスレスAIの観点から着目すると，量子化前後のモデルの違いの影響を考慮しつつ一致率を向上させるためには，BN統計量を固定し，それ以外の重みパラメータのみを調整することが重要だと考えられる．そこで，提案手法は，まず，重みパラメータとBN統計量を同時に最適化する通常の蒸留を行う．その学習が収束した後に，その時点での移動平均統計量をBN統計量とし，以降，BN統計量を更新しない（α= 0に設定する）．このように2段階で学習を行うことで，リファレンスモデルの推論結果に合わせて重みパラメータのみ調整できる．以上の提案手法により，認識精度を維持しつつ，量子化前後で挙動が変わらない量子化モデルを学習できる．

4. 提案手法の評価

本章では，提案手法の評価結果について述べる．本実験では，100クラスの画像分類問題であるCIFAR-100データセット[11]で学習（50000枚）と評価（10000枚）を行う．量子化対象のネットワーク構造として，50層のResNet[12]を用いる．

4.1 学習条件

まず，浮動小数点精度で200 epochの学習を実施し，最も正答率が良いepoch（Best epoch）をリファレンスモデルとした．次に，以下の3通りのQATを40 epoch実施し，正答率と一致率を比較する．

- 学習1（ベースライン）：蒸留を用いず，正解ラベルと推論クラス間のクロスエントロピーを量子化モデルの損失関数とする．
- 学習2（蒸留）：正解ラベルは用いず，リファレンスモデルと量子化モデルのLogitsの平均二乗誤差を量子化モデルの損失関数とする．
- 学習3（提案手法）：リファレンスモデルと量子化モデルのLogitsの平均二乗誤差を損失関数とし，10 epoch学習後，BN統計量をその時点での移動平均値に固定する．

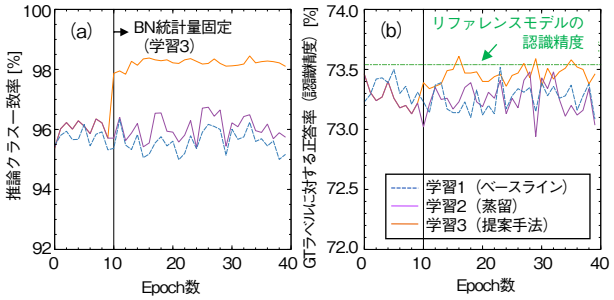
本稿で述べる全ての学習はPyTorchベースで実装されたIntel Corp.のオープンソースであるDistiller[13]を元とした環境下で実施し，QATについては重みと活性化関数に対して8 bitの量子化を実施している．

学習条件は原則Distillerリポジトリ内の初期設定に沿っている．以下では，独自に設定した学習条件やハイパーパラメータについて述べるが，いずれも経験的に決定したものである．バッチサイズは，浮動小数点学習で128，QATで64に設定した．学習率は浮動小数点学習では初期学習率を0.1に設定し，80，120，160 epoch目にそれぞれ0.1倍，0.1倍，0.2倍する．学習率の初期値は，学習1（ベースライン）で1×10-4，学習2（蒸留）と3（提案手法）で5×10-6に設定し，20，30 epoch目にそれぞれ0.1倍する．

4.2 評価結果

学習1〜3の条件で学習した際の，一致率と正答率の学習過程を**第4図**に示す． 学習1（ベースライン）と学習2（蒸留）では一致率が1.35ポイントの変化幅でほぼ一定となった．一方で，提案手法ではBN統計量を固定した時点で顕著に一致率が向上し，以後は高水準のままであった．**第1表**に一致率と正答率についてのBest epochにおける一致率と正答率をまとめる．正答率のBest epoch同士，一致率のBest epoch同士で比較したところ，いずれも提案手法，蒸留，ベースラインの順で一致率が高かった．この結果から，損失関数としてLogits同士のMSELossを導入する蒸留が有効であり，さらに筆者らの新たな着眼である，学習途中からBN層の統計量を移動平均値に固定することにより一致率が向上することを検証できた．また，正答率のBest epoch同士で比較すると，ベースラインの一致率95.79 %に対して，提案

手法は一致率を98.38 %まで2.6ポイント改善できており，量子化前後の各モデルの挙動の差異，すなわち推論結果の不一致を4.21 %から1.62 %まで約61 %抑えられた．



第4図　一致率（a）と認識精度（b）の学習過程

Fig. 4　Training procedures of the (a) match rate and (b) accuracy

第1表　CIFAR-100で学習したResNet-50の評価結果

Table 1　Evaluation results for ResNet-50 trained on CIFAR-100

	Best epoch（正答率）[%]				Best epoch（一致率）[%]		
	Ref.	学習1	学習2	学習3	学習1	学習2	学習3
Epoch	170	23	28	16	11	26	33
正答率	73.54	73.52	73.48	**73.61**	73.14	73.18	**73.48**
一致率	-	95.79	96.65	**98.38**	96.34	96.74	**98.45**

提案手法の正答率は比較した全てのQAT方式の手法やリファレンスモデル（Ref.）に比べてわずかに良く，正答率が劣化しなかった．さらに，一致率のBest epoch同士での比較についても，提案手法はベースラインに比べて推論結果の不一致を58 %抑えられており，なおかつリファレンスモデルからの正答率の劣化を73.54 %から73.48 %までの0.06ポイントに抑えられた．以上から，提案手法は，量子化によりモデルサイズを削減しつつ，リファレンスモデルからの正答率の劣化と不一致率のいずれも抑制できた．

5. まとめ

軽量化前後の推論結果同一性を担保した組込みAIである「ロスレスAI」の実現に向けた開発の一例として，本稿では，軽量化手法としてパラメータのビット数を減らす量子化を対象とし，教師モデルから生徒モデルへの知識蒸留の枠組みを用いた手法を画像分類問題で評価した．今後は，提案手法で残った不一致サンプルを解析することでさらなる性能向上を図ると同時に，物体検出やセマンティックセグメンテーションのような他のタスクへの応用や，枝刈りやネットワーク構造の圧縮のような他の軽量化手法への拡張など，より汎用的に展開可能な技術となるよう開発を進めていく予定である．

参考文献

[1] J. Gu et al., “Recent Advances in Convolutional Neural Networks,” Pattern Recognition, vol.77, pp.354-377, May 2018.

[2] Y. Cheng et al., “Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges,” IEEE Signal Processing Magazine, vol.35, no.1, pp.126–136, 2018.

[3] B. Jacob et al., “Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.2704-2713, Jun. 2018.

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

[5] Y. Bengio et al., “Estimating or Propagating Gradients through Stochastic Neurons for Conditional Computation,” arXiv:1397.3432, Aug. 2013.

[6] R. Krishnamoorthi, “Quantizing Deep Convolutional Networks for Efficient Inference: A Whitepaper,” arXiv: 1806.08342, Jun. 2018.

[7] S. Ioffe et al., “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” Proceedings of the 32nd International Conference on Machine Learning (ICML), vol.37, pp.448–456, Jul. 2015.

[8] Y. Li, et al., “Revisiting Batch Normalization for Practical Domain Adaptation,” arXiv:1603.04779, Mar. 2016.

[9] W.-G. Chang, et al., “Domain-Specific Batch Normalization for Unsupervised Domain Adaptation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.7354–7362, Jun. 2019.

[10] Z. Zhuang, et al., “Rethinking the Distribution Gap of Person Re-Identification with Camera-Based Batch Normalization,” Proceedings of the European Conference on Computer Vision (ECCV), pp.140–157, Aug. 2020.

[11] A. Krizhevsky, et al., “Learning Multiple Layers of Features from Tiny Images,” CiteSeer, Apr. 2009.

[12] K. He, et al., “Deep Residual Learning for Image Recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770–778, Jun. 2016.

[13] N. Zmora, et al., “Neural Network Distiller: A Python Package for DNN Compression Research,” arXiv:1910.12232, Oct. 2019, https://github.com/IntelLabs/distiller, 参照 Oct. 20, 2021.

執筆者紹介

奥野　智行　Tomoyuki Okuno
テクノロジー本部　デジタル・AI技術センター
Digital & AI Technology Center, Technology Div.

中田　洋平　Yohei Nakata
テクノロジー本部　デジタル・AI技術センター
Digital & AI Technology Center, Technology Div.
博士（工学）

石井　育規　Yasunori Ishii
テクノロジー本部　デジタル・AI技術センター
Digital & AI Technology Center, Technology Div.

築澤　宗太郎　Sotaro Tsukizawa
テクノロジー本部　デジタル・AI技術センター
Digital & AI Technology Center, Technology Div.