The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

二値化ニューラルネットワークへの変分情報ボトルネックによる正則化

Binarized Variational Information Bottleneck

河野 慎 *1 Makoto Kawano　大屋 優 *2 Yu Oya　八木 哲志 *2 Satoshi Yagi　中澤 仁 *1 Jin Nakazawa

*1慶應義塾大学大学院政策・メディア研究科 Graduate School of Media and Governance, Keio University

*2NTT ソフトウェアイノベーションセンタ NTT Software Innovation Center

Deep neural networks are utilized in various applications in real worlds, thanks to their capabilities. One of the fashions of it is their deployment on edge devices. With edge devices, deep neural networks can be used in the context of IoT. However, the specification of those edge devices is often poor so that deep neural networks cannot be deployed. Binary neural networks, whose weights and activations are binarized, is one of the solutions. There is a well-known issue, the drastic drop in accuracy compared to its full precision networks. We consider that this is because the binary neural networks can only represent a subset of discrete functions so that they become sensitive to the input perturbation: the lack of robustness for inputs. In this paper, we propose a regularization approach that helps to alleviate the over-fitting problem by introducing variational information bottleneck. We show ablation studies on CIFAR-10 that reduce loss value the though accuracy is maintained on AlexNet-like networks with different binary activation functions.

1. はじめに

深層学習は，その性能の高さから様々な分野で応用されている．特に画像認識の分野の発展は著しく，大量の GPU/TPU と数千〜数億枚の画像を学習に用いることで，近年は画像分類タスクだけではなく，高品質・高画質な画像生成 [Brock 18] が可能となっている．一方で，道路の損傷点検 [Kawano 17] のように，高性能計算機ではなく RaspberryPi3 や NVIDIA JetsonTX2 などエッジデバイスでの実行を想定した研究が注目されている．エッジデバイスで深層学習を実行する場合，その性能の低さゆえに深層学習モデルを小さくする必要がある．

深層学習モデルを圧縮することで，エッジデバイスのような低性能計算機でも実行可能にする技術が研究されている．本研究では，圧縮技術の一つである量子化に注目する．量子化は，32 ビットの単精度浮動小数点数で表現されている深層学習モデルのパラメータを 16 ビットや 8 ビットなど低演算精度で表現するものである．特に，重みパラメータと層間の信号 *1 を 1 ビットで表現された binary neural networks（BNNs）[Courbariaux 16] や XNOR-Net[Rastegari 16] は，メモリを最大で 1/32 の節約と速度の向上が期待される．さらに，BNNs は FPGA 上での実行において，その優位性が発揮される．

一方で，BNNs には，単精度浮動小数点数で表現された深層学習モデルと比べて，一段と精度が下がってしまう問題がある．これは，学習の逆伝播時にパラメータの表現力が不足していることが主な原因とされている [Lin 17]. つまり，重みパラメータの量子化（二値化）は情報量の欠落による表現力の低下を引き起こし，重みパラメータ値の急激な変動（$+1 \to -1$ や $-1 \to +1$）によって，入力の摂動に対する頑健性が失われており，過学習しやすくなってしまっている．

過学習を抑制する手法として，重みパラメータへの制約として L1 正則化や L2 正則化が挙げられるが，BNNs において重みは $+1$ もしくは -1 になっているため，その効果を受けることができない．そこで本研究では，確率的にニューラルネットワークを扱う変分情報ボトルネックを BNNs に導入し，BNNs の過学習抑制を試みる．

連絡先: 河野 慎，慶應義塾大学，神奈川県藤沢市遠藤 5322，makora9143@gmail.com

*1 活性化信号．層から出力され，次の層の入力になるもの．

2. 二値化変分情報ボトルネック

2.1 二値化ニューラルネットワーク

BNNs は，重みおよび信号を $+1$ もしくは -1 のいずれかの値に制約することで，モデルのサイズも小さくなり，また推論時に論理演算で行うことが可能となる．重みもしくは信号 x^b の二値化は

$$x^b = \mathrm{Sign}(x) = \begin{cases} +1 & x \geq 0 \\ -1 & \text{otherwise} \end{cases} \tag{1}$$

である．ただし，x は，実数値を表しており，Sign は符号関数を表している．BNNs の順伝播時，式 (1) によって二値化された重みパラメータ W^b を用いて計算される．層間の活性化関数にも符号関数が用いられ，二値の信号が出力される．しかしながら，符号関数の勾配はほとんど至るところで 0 であるため，逆伝播時に勾配が消失してしまい，学習ができない．この問題を解決するため，BNNs では，Hinton の講義（2012）で紹介された straight through estimator (STE) ：

$$\frac{\partial \mathrm{Sign}}{\partial x} = 1_{|x| \leq 1}$$

が用いられることが多い．なお，STE は hard tanh とみなすことも可能であることが報告されている [Courbariaux 16]．また，符号関数を滑らかな関数として近似した SignSwish ：

$$\mathrm{SS}_\beta(x) = 2\sigma(\beta x)[1 + \beta x\{1 - \sigma(\beta x)\}] - 1$$

関数を用いることで，より学習が安定して行われることが報告されている [Darabi 18]．なお，σ はシグモイド関数を表す．

STE により BNNs の学習が可能になる一方で，依然として過学習が起きやすい．この原因として，BNNs

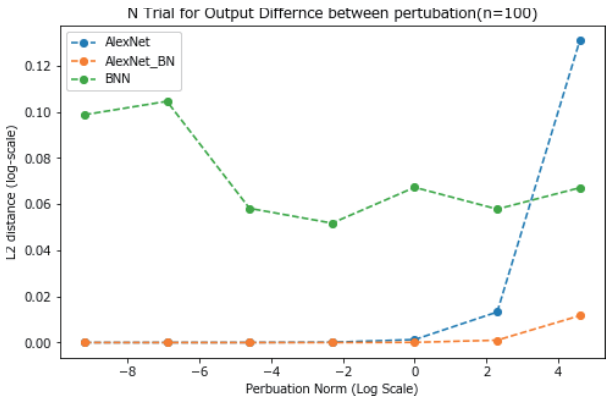


図 1: 入力にノルム γ の摂動を与えた時の出力の誤差.

は摂動に対して弱い問題が挙げられる．摂動に弱い様子を図 1 に示す．学習前の 3 種類のモデル（AlexNet, AlexNet+BatchNormalization[Ioffe 15]，BNN）に MNIST の同一データにノルム γ の摂動を追加した時の追加前との誤差を示している．32 ビット浮動小数点数で表現されている深層学習モデルに比べ，BNN は摂動の大きさに関係なく誤差が生じてしまっている．そこで本研究では，この摂動に対してロバスト性を向上するための正則化項を提案する．正則化項の技術として，変分情報ボトルネック（variational information bottleneck, VIB）を導入する．

2.2 変分情報ボトルネック

変分情報ボトルネック（VIB）[Alemi 16] は，情報ボトルネック [Tishby 15] の変分下界を学習する手法である．情報ボトルネックでは，教師あり学習を表現学習としてみなす．すなわち，入力データ X をラベル Y の予測に利用可能な潜在表現 Z への確率的な写像関数を求める．

二つの確率変数 Z, Y 間における相互情報量を $I(Z;Y)$ とすると，情報ボトルネックは，次のように表現される：

$$\max I(Z;Y) \text{ subject to } I(Z;X) \leq R. \tag{2}$$

ただし，R は，ボトルネック（=定数）である，式 (2) を非制約ラグランジュ最適化問題として書き直すと，

$$\max I(Z;Y) - \eta I(Z;X)$$

となる．情報ボトルネックは一般的に計算困難で知られているため，Alemi ら [Alemi 16] は変分法を適用し，

$$\max_{\theta,\phi,\psi} \mathbb{E}_{p(x,y)e_\theta(z|x)} \left[\log q_\psi(y|z) - \eta \log \frac{e_\theta(z|x)}{m_\phi(z)} \right] \tag{3}$$

とすることで，変分下界を導出している．ただし，q は分類時の尤度，$\frac{e}{m}$ は任意の潜在表現空間 m に関連した潜在表現 e の長さを罰則するレートを表している．また，$e_\theta(z|x)$ は，入力 X を潜在表現 Z に変換する確率的なエンコーダを表している．さらに，$q_\psi(y|z)$ は，潜在表現 Z からラベル Y を予測する変分分類器（もしくはデコーダ）であり，$m_\phi(z)$ は，変分近似分布を示している．変分情報ボトルネックを導入することで，ニューラルネットワークが敵対的事例 [Goodfellow 15] を含めた摂動に対し，頑健になることが報告されている [Alemi 18].

2.3 BNN への変分情報ボトルネックの導入

本研究では，BNN の最終層手前の層を，変分オートエンコーダ [Kingma 13] のようにガウス分布 $\mathcal{N}$ の平均 μ と分散 σ^2 を出力するため，2 種類の出力をするように変更する：

$$\begin{aligned} q_\phi(y|z) &= \mathbb{E}_{z\sim e_\theta(z|x)}[\text{softmax}(W_1 z + b_1)] \\ \log e_\theta(z|x) &= \log \mathcal{N}\left(z; \mu, \sigma^2 I\right) \\ \text{where } \mu &= W_2 h + b_2 \\ \sigma &= \text{softplus}(W_3 h + b_3) \\ h &= \text{NN}(x) \end{aligned} \tag{4}$$

ただし，$I \in \mathbb{R}^{zdim \times zdim}$ は単位行列，W_1, b_1 は，ソフトマックス層の全結合層，W_2, b_2 と W_3, b_3 はそれぞれ，平均と分散を出力する全結合層のパラメータを示し，NN$(\cdot)$ は最終層から 2 層手前までのニューラルネットワークを表す．

Activation	$-\log q_\phi(y\|z) \downarrow$	Accuracy ↑
BNN(HT)	$1.199 \pm 3.93 \times 10^{-3}$	61.592 ± 0.267
BVIB(HT)	$\mathbf{1.186 \pm 9.15 \times 10^{-3}}$	$\mathbf{61.646 \pm 0.178}$
BNN($SS_{\beta=5}$)	$1.302 \pm 5.40 \times 10^{-3}$	$\mathbf{61.270 \pm 0.410}$
BVIB($SS_{\beta=5}$)	$\mathbf{1.290 \pm 10.82 \times 10^{-3}}$	61.134 ± 0.475

表 1: CIFAR10 の実験結果．矢印は，高いスコア/低いスコアのどちらが良いのかを示している．表中の HT は HardTanh, SS は SignSwish を表す．

3. 実験

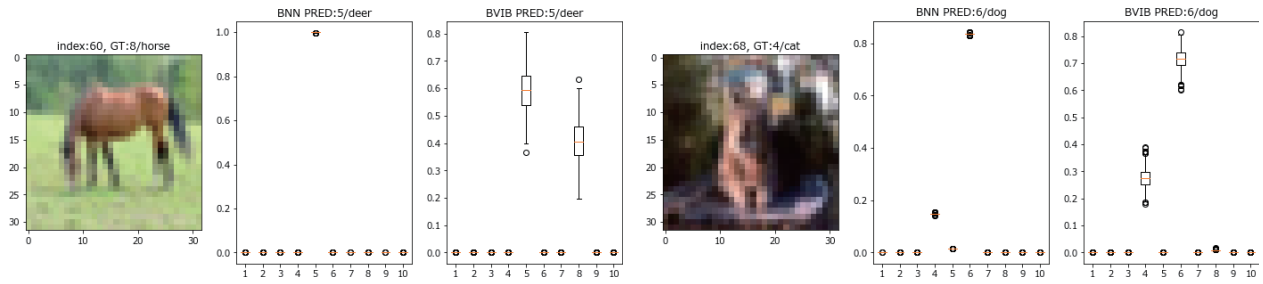
提案手法の有効性を検証するため，本実験では，CIFAR10 を用いた損失関数の値および精度の比較を行う．まず，CIFAR10 のデータセットに対して本研究では，正規化のみを行った．これは，CIFAR10 において，左右反転やランダムクロップなどのデータ拡張処理は，精度に非常に大きな影響があることが報告されているため [Graham 14]，本研究では平均 0，分散 1 への正規化のみを行った．

本実験では，深層学習モデルの基本的な構造として AlexNet[Krizhevsky 12] を用いる．ただし，AlexNet の畳み込み層部を式 (4) の NN とし，全結合層は 3 層ではなく 2 層とし，$\mu, \sigma \in \mathbb{R}^{256}$ とする．さらに BatchNormalization を畳み込み層の直後に導入する．また活性化関数（HardTanh/SignSwish）については，Rastegari ら [Rastegari 16] が指摘するように Max プーリング層の直前に入れてしまうと出力される値がほとんど+1 になってしまうことが懸念されるため，畳み込み層の直前に挿入した．SignSwish のハイパーパラメータは，[Darabi 18] で固定値の中で最も精度が高かった $\beta = 5$ とした．バッチサイズは 128 とし，オプティマイザには Adam[Kingma 14] を採用，学習率を 10^{-3}，エポック数を 50 とし，10 エポックごとに 0.1 倍にした．本実験の結果を表 1 に示す．なお，異なるランダムシードを 5 種類用い，その平均と標準偏差を載せた．

4. 考察・まとめ

本研究では，BNNs が入力の摂動に対して敏感であるために過学習が起きてしまう問題に対し，変分情報ボトルネックを導入することで，正則化を試みた．

表 1 に示す通り，変分情報ボトルネックを入れることで，HardTanh と SignSwish いずれの活性化関数においても，目的関数であった負の対数尤度は下回っている．精度においては



(a) 正解ラベル「馬」の画像を入力した場合.

(b) 正解ラベル「猫」の画像を入力した場合.

図 2: 任意の画像を入力した時の出力結果の違い．それぞれ，入力画像，BNN，BVIB の各出力結果となっている．BVIB は変分情報ボトルネックの性質によりサンプリングが可能であるため，出力は箱ひげ図（サンプリング数 1000）となっている．

HardTanh を用いた場合は上回っている一方で，SignSwish を用いた場合は下回ってしまった．目的関数と精度が一致していない理由として，分類結果の出力の仕方があげられる．一般的に，分類結果の出力は，最終層のソフトマックス層出力で最も大きい値を出力している次元のラベルとみなされる．したがって，モデルの学習がより上手く行っているとしても，つまり目的関数がより小さい値になったとしても，最終的に最も大きい値を出力するラベルが変わらなければ精度は変わらない．

BNN と BVIB 共に不正解であった CIFAR10 データおよびモデルの出力の一例を図 2 に示す．BVIB の出力は $y \sim p(y|z)p(z|x)$ であるため，サンプリング数を 1000 とし，結果には箱ひげ図を用いた．図 2(a) において，入力画像の正解ラベルが「馬」なのに対し，BNN の出力は「鹿」が最大となっており，またそのほかのラベルの値はほぼ 0 になっている．一方で，BVIB では BNN と同様に「鹿」が最大となっているが，「馬」のラベルが次に高い値を示しており，依然として不正解ではあるが，BNN よりも正しく認識できていると言える．また，図 2(b) において，入力画像の正解ラベルが「猫」に対して，BNN と BVIB 共に一番大きい値となったラベルは「犬」であり，二番目に大きい値となったラベルは「猫」となっている．しかし，BNN に比べると BVIB の方が「猫」ラベルの値は大きく，より正しく認識できている．

以上のことから，BNN に変分情報ボトルネックの枠組みを取り入れることで，正則化され，過学習が抑えられていることが確認できた．一方で，32 ビット浮動小数点数で表現されたニューラルネットワークに比べると依然として精度は低い．過学習を抑えつつ，精度を高めていくことが今後の課題である．

参考文献

[Alemi 16] Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K.: Deep Variational Information Bottleneck (2016)

[Alemi 18] Alemi, A. A., Fischer, I., and Dillon, J. V.: Uncertainty in the variational information bottleneck, *arXiv preprint arXiv:1807.00906* (2018)

[Brock 18] Brock, A., Donahue, J., and Simonyan, K.: Large scale gan training for high fidelity natural image synthesis, *arXiv preprint arXiv:1809.11096* (2018)

[Courbariaux 16] Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., and Bengio, Y.: Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1, *arXiv preprint arXiv:1602.02830* (2016)

[Darabi 18] Darabi, S., Belbahri, M., Courbariaux, M., and Nia, V. P.: BNN+: Improved binary network training, *arXiv preprint arXiv:1812.11800* (2018)

[Goodfellow 15] Goodfellow, I., Shlens, J., and Szegedy, C.: Explaining and Harnessing Adversarial Examples, in *International Conference on Learning Representations* (2015)

[Graham 14] Graham, B.: Spatially-sparse convolutional neural networks, *arXiv preprint arXiv:1409.6070* (2014)

[Ioffe 15] Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, *arXiv preprint arXiv:1502.03167* (2015)

[Kawano 17] Kawano, M., Mikami, K., Yokoyama, S., Yonezawa, T., and Nakazawa, J.: Road marking blur detection with drive recorder, in *Big Data (Big Data), 2017 IEEE International Conference on*, pp. 4092–4097IEEE (2017)

[Kingma 13] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* (2013)

[Kingma 14] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014)

[Krizhevsky 12] Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, in *Advances in neural information processing systems*, pp. 1097–1105 (2012)

[Lin 17] Lin, X., Zhao, C., and Pan, W.: Towards accurate binary convolutional neural network, in *Advances in Neural Information Processing Systems*, pp. 345–353 (2017)

[Rastegari 16] Rastegari, M., Ordonez, V., Redmon, J., and Farhadi, A.: Xnor-net: Imagenet classification using binary convolutional neural networks, in *European Conference on Computer Vision*, pp. 525–542Springer (2016)

[Tishby 15] Tishby, N. and Zaslavsky, N.: Deep learning and the information bottleneck principle, in *Information Theory Workshop (ITW), 2015 IEEE*, pp. 1–5IEEE (2015)