F-019

潜在空間での最適化を用いた VAE のプルーニングによる圧縮と分類精度向上
Compression and Improved Classification Accuracy by Pruning VAE with Latent Space Optimization

山田 康晴† 徳山 豪†
Yasuharu Yamada Takeshi Tokuyama

1 はじめに

NN（ニューラルネットワーク）は，脳を構成するニューロン（神経細胞）の結合を模した数理的モデルである．それを用いる深層学習では，多層構造にニューロンに対応する頂点を配置し，次の層のニューロンに与える影響力，すなわち辺の重みを自動的に調節することで，人間や動物による思考や判別の学習プロセスを疑似的に再現している．深層学習モデルの実装において，モデルの精度向上を達成するため，モデル内部の NN は冗長性を高く保っている．これによる要求リソースと計算量の肥大化という問題を解決するために，プルーニング [1] と呼ばれる，ネットワークの辺（一般的に重みの絶対値が小さいもの）を除去するモデル圧縮の手法が提案されている．この手法は，人間の脳の成長過程における，ニューロン結合の減少を模している．

本研究では，データの圧縮及び生成が可能な深層学習モデル VAE[2] を扱い，学習後の辺の重みとは異なる尺度として，潜在空間の最適化を目的とした各辺の評価値解析を行う手法を提案する．具体的な結果として，その潜在空間を用いるクラス分類の際，評価値に基づくプルーニングによるモデル規模の縮小化と，分類精度の向上を示す．

本研究によって与えられる知見と意義は 3 つである．

1. ブラックボックスである深層学習の各部位が持つ役割解析．
2. モデル圧縮手法の開発、その効果の測定と可視化．
3. 脳の仕組みに対する，ニューラルネットワークを用いた実験手法の提案．

人間や動物の脳の機能に対する侵襲を伴う実験は，倫理的に行うべきではなく，特にその機能を調べるために一部を改変することは困難である．本研究では，これを倫理的に問題のない人工知能モデルにおいて実験する．具体的には，人の物体認識における形の情報（抽象化された概念）を，特徴抽出した潜在空間の分布と捉え，その最適化を基準にプルーニングを行うことで，圧縮を伴う部分的な変化における，脳機能の認識を VAE で実験する．

2 Variational Autoencoder（VAE）

VAE は，深層学習による教師なし学習可能なデータ生成モデルである．データを圧縮する符号化器の Encoder と，圧縮されたデータを元に戻す復号器である Decoder により構成される．

訓練データ集合を X とすると，Encoder E は各データ $x \in X$ を入力として，ϕ をパラメータとする確率密度関数 $q_\phi(z|x)$ に従い，潜在変数 z を出力する．Decoder D は潜在変数 z を入力として，θ をパラメータとする確率密度関数 $p_\theta(x|z)$ に従い，推定データを出力する．これにより，VAE の学習では以下の式 (1) を目的関数として定式化できる．

$$\arg\max_{\theta,\phi} \sum_{x \in X} L(\theta, \phi, x) \tag{1}$$

$$\begin{aligned}\mathcal{L}(\theta, \phi, x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \\ - D_{KL}(q_\phi(z|x)||p_\theta(z))\end{aligned} \tag{2}$$

式 (2) 第 1 項は，$q_\phi(z|x)$ におけるデータ x の対数尤度 $\log p_\theta(x|z)$ の期待値であり，x が VAE によって再構成された時の誤差を表している．第 2 項は正則化の役割を担い，潜在変数 z の分布が標準正規分布に従うようにカルバック・ライブラー情報量を最小化する．これにより，潜在空間ではその距離に応じて類似性を保持するようになり，連続的なデータ生成を可能にする．VAE を用いるクラス分類では，潜在空間のこの特徴を利用し，その分布を k-means 法などのクラスタリング手法で分類する．

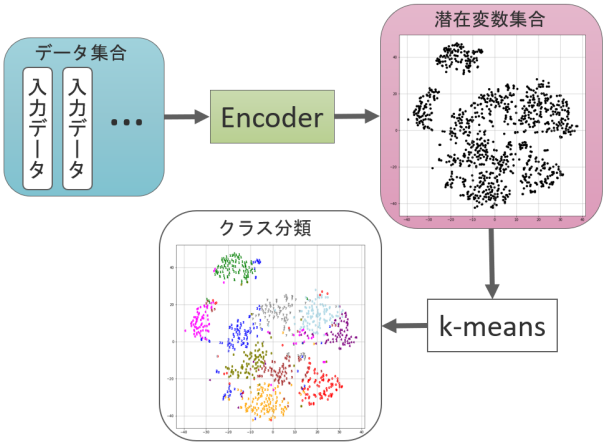

図 1 VAE を用いるクラス分類

3 提案手法

クラス数 C の訓練データ集合 X で学習済みの Encoder E 内で，ネットワークを構成する一つの辺 e_i の評価を求めるために，E から e_i の重みを 0 にした類似 Encoder E_i を作成する．サンプル（ラベル付きデータ集合）$S \subset X$ の入力データを与えた時，E_i が出力する潜在変数集合 Z_i の分布を観測する．そして，Z_i のクラス間分散 σ_b^2 を e_i の評価値とする．

$$\sigma_b^2 = \sum_{j=1}^{C} n_j(\mu_j - \mu_T)^2 \tag{3}$$

ここで，n_j，μ_j はそれぞれ，Z_i におけるクラス j の割合，平均であり，μ_T は Z_i 全体の平均をとする．

評価値に基づくプルーニングでは，この σ_b^2 が最も大きくなる辺を除去していく．

† 関西学院大学大学院理工学研究科 Graduate School of Science and Technology, Kwansei Gakuin University

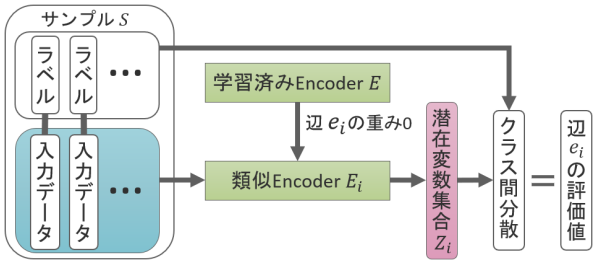

図 2　辺の評価

4　実験

VAE の Encoder（3 層の NN）が出力する潜在変数を用いて，k-means 法による MNIST のクラス分類を行った．その際、層間のネットワーク毎に，提案手法の評価値及び重み絶対値（昇順）に基づくプルーニングを行い，分類精度を比較した（図 3）．また，異なるサンプルを使用した提案手法の評価値に基づくプルーニングの分類精度を比較した（図 4）．

4.1　設定

Encoder は，入力層，隠れ層 1，隠れ層 2，出力層から成る 3 層（全結合層）の NN とし，それぞれのニューロン数は 28×28，400，400，40 とした．また，学習時にのみ用いる Decoder は，Encoder の入力層と出力層のニューロン数を入れ替えた 3 層の NN とする．学習では，訓練データ 6 万，バッチサイズを 1000 としたミニバッチ勾配降下法を行う．最適化は Adam を使用し，学習率は 0.001，反復回数は 100 回に設定した．辺の評価値計算に必要なサンプルは，訓練データ 6 万からランダムに選ばれた 1000 のデータ集合である．k-means 法によるクラス分類では，1 万のテストデータからランダムに選ばれた 1000 のデータ集合を使用した．

4.2　結果及び考察

図 3 より，評価値によるプルーニングでは，重み絶対値とは異なり，層間毎の全てのネットワークで分類精度の向上が確認される．N_2，N_3 では，残ったネットワークが 5 %以下に至るまで，元の Encoder より劣化することなく優れた精度を維持し続けており，高い圧縮率を示した．一方 N_1 においては，3 %程度のプルーニング以降，上昇した精度が大幅に下がっていき，重み絶対値によるプルーニングの方が精度を維持できている．

図 4 より，使用するサンプルに左右されることなく，元の Encoder より精度向上が実現できることが確認される．また，元の Encoder より N_1 では最大 4 %，N_2 では最大 6.5 %，N_3 では最大 5.7 %の精度向上を示した．サンプルによって精度の上昇率が異なり，*sample*2 においては，他のサンプルより全体的に精度が低い傾向にある．加えて，*sample*1 と *sample*3 は，N_2 と N_3 で精度が逆転しており，層間のネットワークによって評価値に優れたサンプルは異なると推察される．

5　まとめ

本研究では，深層学習モデルである VAE の Encoder が出力する潜在空間を用いて，クラスタ最適化を目的とした辺の評価値解析を行う手法を提案した．そして，評価値に基づくプルーニングによるクラス分類精度の比較を行った．実験により，モデルの性能を劣化させずネットワークのサイズを大幅に削減できること，更に，ネットワークの圧縮に伴う，性能向上が実現できることを示した．この研究は科学研究費基盤研究 B 20H04143 の支援をうけた．

参考文献

[1] Han, Song, Huizi Mao, and William J. Dally. ”Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding.” arXiv preprint arXiv:1510.00149 (2015).

[2] Kingma, Diederik P., and Max Welling. ”Auto-encoding variational bayes.” arXiv preprint arXiv:1312.6114 (2013).

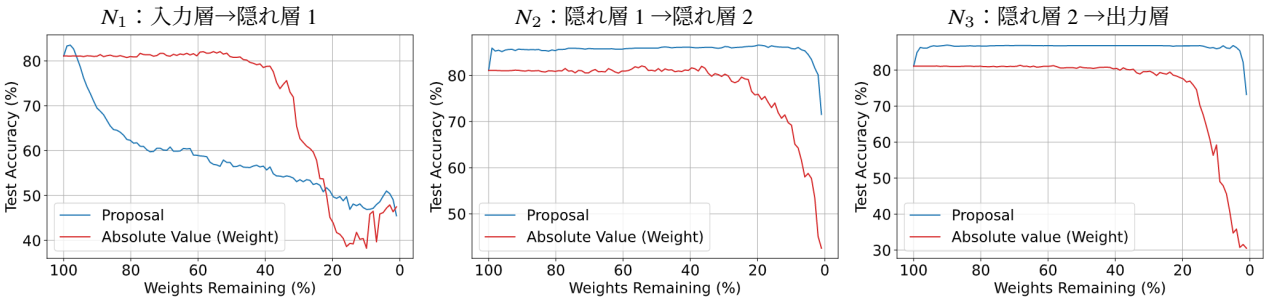

図 3　提案手法の評価値及び重み絶対値（昇順）に基づくプルーニングの比較 (5 回の平均分類精度)

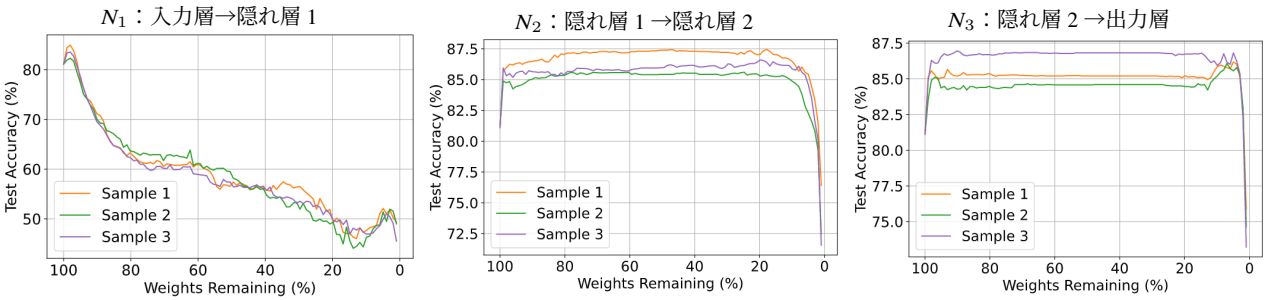

図 4　異なるサンプルを使用した提案手法の評価値に基づくプルーニングの比較 (5 回の平均分類精度)