深層学習における教師なし特徴抽出手法の比較

Comparison of Unsupervised Feature Learning Approaches with Deep Learning

立花　亮介 †　松原　崇 ‡　上原　邦昭 ‡
Ryosuke Tachibana　Takashi Matsubara　Kuniaki Uehara

1. はじめに

深層学習を用いた手法が画像分類の分野において，高い精度を上げており，その有用性が注目を浴びている．近年高い精度を上げている画像分類における深層学習の手法の多くは，画像認識コンペティション ILSVRC2012 で優勝した際に使用された Convolutional Neural Network（CNN）を利用した手法 [1] を含め，ラベルありデータを用いた教師あり学習のアルゴリズムを用いている．

ラベルありデータは，データに別途ラベルを付与するという作業を行わなければならないため多大な労力がかかる上，ラベルの設計方法によっては，より最適な分類方法があるにも関わらず，ラベルに依存した分類しか行えない恐れもある．一方，ラベルなしデータはラベル付けする労力が不要であり，大量に入手することが可能な上，データに内在する構造を抽出できる可能性がある [2]．このことから，ラベルなしデータを用いた教師なし学習アルゴリズムが重要となる．

深層学習手法において，既存の教師なし学習アルゴリズムとしては，autoencoder（AE），Restricted Boltzmann Machine（RBM），AE を多層にした stacked autoencoder（SAE），RBM を多層にした deep Boltzmann machine（DBM） [3, 4] などが存在する．しかし，AE は他の学習アルゴリズムのための特徴抽出で利用されるか，ディープニューラルネットワーク（DNN）を使用した分類器の事前学習 [5] に利用されることが多く，それ単体で画像分類問題への分類器として利用されることは少ない [6]．SAE は，AE で各層を教師なし学習で事前学習した後，誤差逆伝播法に基づく勾配降下法で教師あり学習を行う，半教師あり学習であるが，教師なし学習と教師あり学習では異なるグラフ構造を用いる．RBM や DBM は生成モデルと呼ばれる．生成モデルは，ネットワークの状態を確率 p に従って生成する．データが未知の確率分布 q に従っているという仮定のもとで，データに基づいてモデル分布 p をその確率分布 q に近づける．そのため，データによって学習されたモデル分布 p は，サンプリングによりデータを生成することができる．RBM や DBM は，データだけを与えればその特徴を抽出する教師なし学習になり，ラベルを与えればデータとラベルの関係を抽出する教師あり学習になる．よって，教師なし学習と教師あり学習で共通の学習アルゴリズムが使用でき，同一のグラフ構造を用いることができる．しかし，モデルを学習させる際や，学習したモデルによりデータの一部からデータを復元する際，サンプリングを多用するため計算量が大きくなる．

本研究では，stacked what-where auto-encoders（SWWAE）[7] を用いて，教師なし学習または半教師あり学習による画像の分類を試みる．SWWAE は，本来，教師あり学習である CNN に，AE の符号化，復号化の概念を取り入れることによって，教師なし学習を行う手法である．また，AE と異なり，SWWAE は同じ学習アルゴリズムで教師あり学習，半教師あり学習，教師なし学習を切り替えることが出来る．さらに，データを決定論的に復元することが可能であり，サンプリングを行わないため，計算量も少ない．

2. 手法

本章では，まず深層学習の基本的な手法である，CNN，AE について述べる．その後，SWWAE について述べる．

2.1 Convolutional Neural Network

CNN の入力は $N \times M$ 画素の画像である．ただし，一画素ごとに複数の値をもつ多チャネル画像にも対応しており，チャネル数を K とすると，入力は $N \times M \times K$ の大きさとなる．例えば RGB からなるカラー画像では，チャネル数は $K = 3$ である．中間層は convolution 層と pooling 層の 2 層を基本モジュールし，これが複数連続して接続している．最終層には，分類クラス数と同数のユニットを置いて，これと直前の pooling 層との間が全結合層で結ばれている．

2.1.1 convolution 層

convolution 層の役割は，入力画像から，特定方向の線分や特定の太さを持つ線分などの特徴的な濃淡構造を抽出することである．convolution 層では，入力画像にフィルタを畳み込むことによって，フィルタの濃淡パターンと類似したパターンが抽出される．抽出した複数の特徴

† 神戸大学　工学部
Faculty of Engineering, Kobe University
‡ 神戸大学　大学院　システム情報学研究科
Graduate School of System Informatics, Kobe University

を並べたものを特徴マップと呼ぶ．$N \times M \times K$ の大きさの画像の中の位置 (i, j, k) の画素値を x_{ijk}，フィルタの種類を R 種類，フィルタのサイズを $H \times H$ とし，r 種類目のフィルタの位置 (p, q, k) の画素値を h_{pqkr} とすると，畳み込みは

$$u_{ijr} = \sum_{k=0}^{K-1} \sum_{p=0}^{H-1} \sum_{q=0}^{H-1} x_{i+p,j+q,k} h_{pqkr} + b_r \qquad (1)$$

として計算でき，特徴マップ数は R となる．ここで，$b_r (r = 0, \ldots, R)$ はバイアス項であり，通常，バイアスはフィルタごとに共通とされる．u_{ijr} に活性化関数を適用した $z_{ijr} = f(u_{ijr})$ が convolution 層の出力となる．

2.1.2 pooling 層

pooling 層では，convolution 層の出力層の小領域の中から代表値がサブサンプリングされる．pooling 層の役割は，convolution 層で抽出されたデータの特徴の若干の位置ずれを吸収したり，データ量を削減することである．これによって，入力画像に位置ずれや多少の変形が存在しても，抽出される特徴はある程度不変になることが期待できる．代表値として最大値がよく用いられ，このとき特に max-pooling と呼ばれる．convolution 層から出力された特徴マップを，幾つかの小領域に分割する．ただし，小領域に重なりがあってもよい．特徴マップにおける小領域の位置を座標 (i, j)，座標 (i, j) の小領域を P_{ij} とすると，max-pooling は

$$\hat{z}_{ijr} = \max_{(p,q) \in P_{ij}} z_{pqr} \qquad (2)$$

と表される．このように，max-pooling では，少領域 P_{ij} 内で代表値の位置がずれた場合でも出力が同じになるので，convolution 層で抽出された特徴の位置ずれをある程度吸収することができる．

2.1.3 全結合層

全結合層の役割は，convolution 層と pooling 層によって抽出された入力画像の特徴をもとに，入力画像の分類を行うことである．全結合層では，pooling 層の出力をもとにクラス分類が行われる．クラスは，入力を $x_i (i = 1, \ldots, n)$ とすると，softmax 関数を用いて

$$\hat{j} = \mathrm{argmax}_j \frac{\exp(x_j)}{\sum_{k=1}^{n} \exp(x_k)} \qquad (3)$$

と推定される．CNN の学習は，Multi-Layer Perceptron と同様に，教師あり学習であり，誤差逆伝播法によってフィルタ係数 h_{pqks} とバイアス b_s を推定することで行われる [8]．

2.2 Auto-Encoder

AE は，入力データのみを使用する教師なし学習により次元の縮約を行い，データの特徴を抽出することを目的としたニューラルネットワークである．DNN の事前学習手法の一つとして，DNN を層ごとに分割し，入力層から出力層まで 2 層ずつ AE を適応していく方法がある [9]．入力 $\boldsymbol{x}$，出力 $\boldsymbol{y}$，活性化関数を f とすると，

$$\boldsymbol{y} = f(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \qquad (4)$$

と決定される feed-forward ネットワークを考える．ここで，$\boldsymbol{W}$，$\boldsymbol{b}$ はモデルのパラメータであり，それぞれ，重み行列，出力層のバイアス項を表す．次に出力層を折り返す．つまり，出力層に入力層と同数のユニットをもつ層をつなげる．このとき，新たな出力層からの出力を $\hat{\boldsymbol{x}}$ とし，追加した層の活性化関数を $\tilde{f}$ とすると，

$$\hat{\boldsymbol{x}} = \tilde{f}(\tilde{\boldsymbol{W}}\boldsymbol{y} + \tilde{\boldsymbol{b}}) \qquad (5)$$

となる．ここで，$\tilde{\boldsymbol{W}}$，$\tilde{\boldsymbol{b}}$ はモデルのパラメータであり，それぞれ，追加した層の重み行列，追加した出力層のバイアス項を表す．$\tilde{\boldsymbol{W}}$ は，$\tilde{\boldsymbol{W}} = \boldsymbol{W}^{\mathrm{T}}$ とすることが多い．式 (4) の変換を符号化，式 (5) の変換を復号化という．

出力を入力と比較した際，情報が欠落していなければ，およそ適切な変換であると考える．変換が適切であるとき，式 (4) の符号化によって得られた出力 $\boldsymbol{y}$ には，入力 $\boldsymbol{x}$ の情報を縮約した特徴が現れるとされる．ただし，出力 $\boldsymbol{y}$ の次元数は入力 $\boldsymbol{x}$ の次元数よりも少ないとする．また，f，$\tilde{f}$ が恒等関数であれば，主成分分析と等価になる．誤差関数には，$\boldsymbol{x}$ と $\hat{\boldsymbol{x}}$ の近さの尺度である $C(\boldsymbol{x}, \hat{\boldsymbol{x}})$ の総和

$$E(\boldsymbol{\theta}) = \sum_{n=1}^{N} C(\boldsymbol{x}_n, \hat{\boldsymbol{x}}_n) \qquad (6)$$

を用いる．ここで，$\boldsymbol{\theta}$ はパラメータ $\{\boldsymbol{W}, \boldsymbol{b}, \tilde{\boldsymbol{W}}, \tilde{\boldsymbol{b}}\}$ を並べてベクトルにしたものであり，N は入力データ数である．

2.3 Stacked What-Where Auto-Encoders

SWWAE とは，CNN に AE の符号化，復号化手法を取り入れることで，教師なし学習を可能にしたモデルである．符号化器に convolutional net（Convnet），復号化器に deconvolutional net（Deconvnet）[10] を使用する．Fig. 1 に SWWAE のモデル構造を示す．特徴的なのは，符号化において各 pooling 層から出力する値が “what” と “where” の二種類あり，前者のみ次の層へ送られ，後者は復号化の対応する層（unpooling 層）へ送られ，復号化処理に用いられることである．SWWAE が画期的であるのは，それ単体で，学習アルゴリズムを変えることなく，教師あり学習，半教師あり学習，教師なし学習を切り替えることが出来るという点にある．

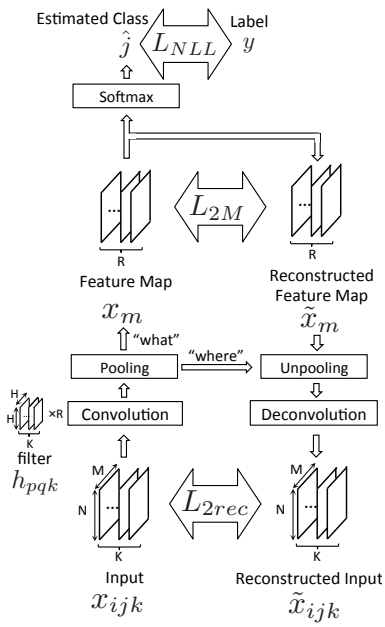

Fig. 1: SWWAE の構造

2.3.1 符号化器

符号化器では，ReLU [11] を用いた convolution 層と max-pooling 層を基本モジュールとし，これが複数連続して接続している．max-pooling 層からは，2 種類のデータが出力される．それは convolution 層の出力層の小領域における max と argmax であり，それぞれ “what”，“where” と呼ぶ．“what” はデータの特徴を表し，“where” は “what” が小領域のどこに位置するかを表す．“what” のみ次の層に送られ，“where” は max-pooling 層に対応する復号化器の unpooling 層に送られる．通常の CNN では，“where” は max-pooling において破棄される情報である．

2.3.2 復号化器

復号化器では，unpooling 層と deconvolution 層を基本モジュールとし，これが複数連続している．unpooling と deconvolution は，それぞれ pooling と convolution の逆の処理である．復号化器への入力には符号化器の最上位の max-pooling 層から出力された “what” を使用する．各 unpooling 層では，前の層から送られる “what” と符号化器の max-pooling 層から送られる “where” を使用して，特徴マップが復元される．これは，“what” を，“where” が示す小領域の元の位置に置くことによって実現される．

2.3.3 学習アルゴリズム

SWWAE は，誤差関数として

$$L = \lambda_{NLL} L_{NLL} + \lambda_{L2rec} L_{L2rec} + \lambda_{L2M} L_{L2M} \quad (7)$$

を用いる．ここで，L_{NLL} は符号化器からの出力と教師信号との誤差である識別誤差を表し，L_{L2rec} は最下層レベルでの誤差，L_{L2M} は各中間層における二乗誤差の和をとったものである．λ は各誤差間の重みを調整する．このような誤差関数を設定することで，教師あり学習を行うときは $\lambda_{L2rec} = \lambda_{L2M} = 0$ とすることによって，単なる CNN とみなせ，教師なし学習を行うときは，$\lambda_{NLL} = 0$ とすることによってこれを実現できる．また，ラベルありデータが入力のとき $\lambda_{L2rec} = \lambda_{L2M} = 0$ とし，ラベルなしデータが入力の時 $\lambda_{NLL} = 0$ とすれば，半教師あり学習を行うことができる．L_{NLL} には，負の対数尤度

$$L_{NLL} = -\sum_i \sum_j I(y_i = j) \log P(y_i | x_i) \quad (8)$$

を用いる．ここで，I(cond) はある条件 cond が成立すれば 1 を，成立しないならば 0 をとる指示関数，x_i は i 番目の入力，y_i は x_i のラベル，$P(y_i|x_i)$ は x_i が観測されたときラベルが y_i である確率を表すベクトルである．L_{L2rec}，L_{L2M} は

$$L_{L2rec} = ||x - \tilde{x}||_2, \qquad L_{L2M} = ||x_m - \tilde{x}_m||_2 \quad (9)$$

を用いる．x は符号化器への入力を表し，$\tilde{x}$ は復号化器に復元された入力を表す．また，x_m は符号化器での特徴マップを表し，$\tilde{x}_m$ は復号化器で復元された，x_m に対応する特徴マップである．

AE を多層にしたモデルである SAE は，各層を分割し AE で層ごとに教師なし学習で事前学習した後，誤差逆伝播法に基づく勾配降下法で教師あり学習を行い，分類を行う．SAE は，非対称な形をしており，SAE に行えるのは多対一の変換，つまり入力に対して分類クラスを出力することであり，分類クラスから入力を復元する一対多の変換ではデータの欠落が無視できない．一般に，これを可能にするには，一対多のマッピングを確率で表現する．これは RBM や DBM で用いられる手法である．RBM や DBM は生成モデルであり，ネットワークの状態変化を確率的に記述する．これらのモデルでは，一対多のマッピングにサンプリングを用いる．つまり，サンプリングを複数回にわたって繰り返すことによってデータ生成を実現する．ネットワークが複雑になると，確率的に表現するための組み合わせが莫大になり，計算量が指数関数的に増加するという問題がある．一方，SWWAE は，決定論的にデータを復元するので，ネットワークが複雑になっても，RBM や DBM ほど計算量が増加することはない．

3. 実験

まず，SWWAE によるクラス分類の性能を，他の教師なし特徴抽出手法 [12] と比較する．次に，ラベルデータから情報を復元する際，SWWAE では “what” と “where” の 2 種類の情報のみが使用されているが，これを他の特徴量情報に変更したり，他の特徴量情報を加える事によって，クラス分類の精度の向上，より多様なデータの復元，データの生成を試みる．

まず，第一の実験について検討する．教師なし学習における比較手法としては，CNN を使用した教師なし特徴抽出手法 [12] がある．この手法は学習サンプルを増やすデータ拡張を利用したもので，まず一枚のラベルなし画像に様々な変換を施すことによって，画像に “ノイズ” を加えた画像を大量に生成する．その大量の画像に一つのサロゲートクラスを与え，これを教師として CNN で学習する．これによって，加工を加えたどの画像に対しても不変な特徴を抽出することが期待できる．CNN に適応するアルゴリズムは教師あり学習であるが，教師データを自ら生成するため教師なし学習とみなせる．

次に，第二の実験について検討する．SWWAE では，通常の CNN において破棄される情報を保存しておき，復号化器での復元に利用している．SWWAE が追加で保存する情報は，最大値の位置を表す “where” のみである．“what” はクラス分類するための情報となり，“where” はデータの特徴を表す情報となる．しかし，“where” 以外にも特徴を表す情報を保存しておく，あるいは “where” に代わる，よりデータの特徴を捉えた情報の保存方法を考えることによって，より正確にデータの特徴の復元を行えると考えられる．

詳細な結果については当日報告する．

参考文献

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.

[2] Quoc V Le, Marc'Aurelio Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S Corrado, Jeff Dean, and Andrew Y Ng. Building High-level Features Using Large Scale Unsupervised Learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[3] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18:1527–1554, 2006.

[4] Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep Boltzmann Machines. In *Proceedings of The 12th International Conference on Artificial Intelligence and Statics*, pages 448–455, 2009.

[5] Geoffrey E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(July):504–507, 2006.

[6] Hugo Larochelle and Yoshua Bengio. Classification using discriminative restricted Boltzmann machines. In *Proceedings of the 25th International Conference on Machine Learning*, pages 536–543, 2008.

[7] Junbo Zhao, Michael Mathieu, Ross Goroshin, and Yann Lecun. Stacked What-Where Auto-encoders. *arXiv*, pages 1–9, 2015.

[8] Yann LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, 1989.

[9] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy Layer-Wise Training of Deep Networks. In *Advances in Neural Information Processing Systems 19*, pages 153–160, 2007.

[10] Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, and Rob Fergus. Deconvolutional networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2528–2535, 2010.

[11] Vinod Nair and Geoffrey E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010.

[12] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative Unsupervised Feature Learning with Convolutional Neural Networks. *arXiv*, pages 1–13, 2014.