3H4-OS-05c-3

視覚情報処理モデルに基づいて改良した畳込みニューラルネットワーク文字認識

Improvement of Convolutional Neural Network Character Recognition Based on Visual Information Processing Model

関野 雅則* Masanori Sekino　木村 俊一* Shunichi Kimura　越 裕* Yutaka Koshi

* 富士ゼロックス株式会社
Fuji Xerox Co., Ltd.

Convolutional neural networks (CNNs) have convolution, rectification, normalization and pooling layers in their feature extraction units. Recently, a lot of varieties of configuration of the normalization and pooling layers have been proposed and increasing their performance. However, the best configuration is still unknown. To clarify the best configuration, we composed various CNNs with a different feature extraction unit configuration by combing three normalization types (none point-wise or pooled), three pooling types (L1, L2 or L∞), two pooling area types (within a map or across maps), and two processing orders. We evaluated them using the MNIST handwritten digit database and found that the best is the configuration recently proposed by us, which is an improved CNN based on human visual information processing model. It achieved the lowest error rate on the undistorted, unprocessed MNIST dataset (0.527%).

1. はじめに

畳込みニューラルネットワーク（CNN）は，文字や自然画像の認識に広く用いられている[LeCun 1989, LeCun 2010]. 近年のCNN は，畳込み，整流，正規化，プーリング処理から成り立っているが [Jarrett 2009]，正規化およびプーリング処理には，さまざまなバリエーションがある.

たとえば正規化において，Local Contrast Normalization（LCN）[Jarrett 2009] では，ある特徴量は，同じ特徴の空間的に異なる位置から除算的な抑制を受けるが，Local Response Normalization（LRN）[Krizhevsky 2012] では，異なる特徴からも除算的な抑制を受ける.

プーリング処理に関しては，L1 プーリング（平均値）[Jarrett 2009] と L∞プーリング（最大値）[Krizhevsky 2012, Cireşan 2011] が広く用いられているほか，さまざまなプーリング（Lp: p=1,2,4,8,12,16,32,∞）を用いる試みもあり，p=2,4,12 で良好な結果が得られたと報告されている [Sermanet 2012]. プーリングの範囲については，一般には空間的なプーリングのみが行われているが，特徴をまたいだプーリングも試みられている [Kavukcuoglu 2009].

プーリングと正規化の順序に関しては，一般に正規化の後にプーリングが行われている.

近年われわれは，初期視覚野の細胞が行っている視覚情報処理を参考にした CNN の改善を提案した. そこでは，二つの特徴にまたがる L2 プーリング（視覚のエネルギーモデル [Adelson 1984] に相当），二つの特徴から抑制を受ける正規化処理に加え，プーリングの後に正規化を行う処理順の変更が導入され，日本語手書き文字の認識で高い認識率を得た [Sekino 2012].

しかしながら，一般的に用いられる CNN の構成に対し，どの処理に対する変更が認識率の向上に寄与したか，明らかではなかった. そこで本論文では，プーリングと正規化の種別と，プーリングと正規化の順序について，さまざまなネットワークを構成し，MNIST 手書き数字データベースの認識率で認識性能を評価して，ネットワークの構成と認識性能の関係を明らかとする.

2 章では，まずネットワークの構成要素として，さまざまな処理層を定義し，3 章では処理層を組み合わせて特徴抽出ユニットを構成する. 4 章では評価条件を示し，5 章では各特徴抽出ユニットを用いたネットワークの認識性能を評価する. 最後に 6 章では結論を述べ考察を行う.

〒259-0157 神奈川県足柄上郡中井町境 430 グリーンテクなかい
E-mail: Masanori.Sekino@fujixerox.co.jp

2. ネットワークの構成要素

本章では，さまざまな識別ネットワークを構成するために，ネットワークを構成する要素として複数の処理層を定義する. 複数の処理層を組み合わせることで特徴抽出ユニットが構成される. 特徴抽出ユニットおよび識別層は図 1 のように接続され，ネットワーク全体が構成される. 特徴抽出ユニット1～3にはすべて同種の特徴抽出ユニットを使用する.

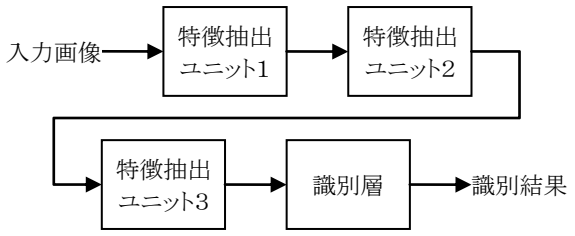

図 1: ネットワークの全体構成

各処理層は，同じサイズをもつ複数の二次元画像からなる特徴マップを入出力とする. 入出力の特徴マップをそれぞれ $x_{n,i,j}$, $y_{n,i,j}$ のように記述する. ここで n は特徴マップのインデックス，i, j はそれぞれ画像内の位置を示すインデックスである. また，サイズが W × H の画像をN 枚もつ特徴マップのサイズを N × W × H と記述する. 特徴マップに含まれる一枚の画像を指

す場合には x_n 特徴マップ全体を指す場合には x のように添え字を省略して記述する.

2.1 畳込み層(C)

入出力の特徴マップの組み合わせごとに異なるフィルタを畳込み演算する.

$$y_n = \sum_m w_{n,m} * x_m \tag{1}$$

ここで $*$ は二次元畳込みであり, w は畳込みの係数である. 入力特徴マップのサイズが $N \times W \times H$ のとき, $K_W \times K_H$ のフィルタを $N \times M$ 個畳込むことで, 出力特徴マップのサイズは $M \times (W - K_W + 1) \times (H - K_H + 1)$ となる.

2.2 整流層(R)

入力特徴マップに, 要素ごとの絶対値処理を行う.

$$y_{n,i,j} = |x_{n,i,j}| \tag{2}$$

2.3 プーリング層(P)

特徴マップの Lp プーリングを行う.

$$y_{n,i,j} = \left(\sum_{n',i',j'} x^p_{n',i',j'} \right)^{1/p} \tag{3}$$

Lp プーリングでは, 入力画像にガウシアンを畳込むものや, プーリング範囲が空間的に重なりをもつものが用いられることもあるが, 簡単のために省略した. 入力特徴マップが $N \times W \times H$ のとき $P_N \times P_W \times P_H$ のプーリングを行うと, 出力画像サイズは $N/P_N \times W/P_W \times H/P_H$ となる. 特徴マップの P_N 画像にまたがる Lp プーリングを $P\langle P_N \rangle, \langle Lp \rangle$ と表記する. たとえば $P_N = 1$ で L2 プーリングを行うプーリング層は P1,L2 と表記する.

2.4 要素ごとの正規化層(NE)

この層は, われわれが先の論文で導入した正規化層である [Sekino 2012]. 二つの特徴のペアを興奮性入力 e, 抑制性入力 s として, 要素ごとに下記の演算を行う.

$$y_{n,i,j} = \frac{e_{n,i,j} - s_{n,i,j}}{1 + e_{n,i,j} + s_{n,i,j}}$$
$$e_n = x_{2n}, s_n = x_{2n+1} \tag{4}$$

2.5 プーリングつき正規化層(NP)

LCN や LRN といった, 一般に用いられている正規化処理では, 正規化の分母でもプーリング処理が行われている. この処理層は, 要素ごとの正規化層(式 3)の分母で LCN や LRN と同様にプーリング処理が行われるよう拡張したものである. 簡単のため, プーリング範囲は, 後段に接続されるプーリング層と一致させ, プーリング層と同様に, ガウシアンの畳込みおよび, プーリング範囲の重なりは省略する.

$$y_{n,i,j} = \frac{e_{n,i,j} - s_{n,i,j}}{1 + E_{n,i,j} + S_{n,i,j}}$$
$$E_{n,i,j} = \sum_{i',j'} e_{n,i',j'}, S_{n,i,j} = \sum_{i',j'} s_{n+1,i',j'} \tag{5}$$

2.6 識別層

入力特徴マップを一次元ベクトルとみなし, 重みおよびバイアスを加えた後, softmax 関数へ入力する. 出力は識別対象のクラス数と同じサイズをもつ一次元ベクトルである.

$$y_i = \frac{e^{\mathbf{W_i \cdot x + b}}}{\sum_j e^{\mathbf{W_j \cdot x + b}}} \tag{6}$$

3. 特徴抽出ユニットの構成

前記の処理層を組み合わせて, 複数の特徴抽出ユニットを構成する. 各特徴抽出ユニットは, 必ず畳込み層, 整流層を最初に含み, それら後にいくつかの処理層が接続される. そのように構成された特徴抽出ユニットは, 畳込み層, 整流層の後に接続された処理層によって命名される. たとえば, 畳込み層 C, 整流層 R, プーリング層 P2,L2, 正規化層 NE が接続された特徴抽出ユニットは“P2,L2→NE”と命名する.

3.1 プーリングの種別

プーリング層については, $P_N = 1,2$ および $Lp = \infty, 1,2$ の 6 通りの組み合わせを用いる.

3.2 正規化の有無と順序

正規化については, 下記の 4 通りの構成を用いる.

- P(プーリングのみ, 正規化なし)
- NE→P(要素ごとの正規化の後, プーリング)
- NP→P(プーリング付き正規化の後, プーリング)
- P→NE (プーリングの後, 要素ごとの正規化)

プーリング層に関しては, 前述の 6 通りの組み合わせを用いるので, 合計で 6×4=24 通りの特徴抽出ユニットが構成される.

特に, “NP→P1,L1”が広く用いられている CNN の構成に最も近い構成であり, “P2,L2→NE”が以前に初期視覚野の細胞を参考にわれわれが導入した構成である[Sekino 2012].

4. 評価条件

前章で得られた 24 通りの特徴抽出ユニットを, 図1のネットワークの全体構成に適用して得られる 24 通りのネットワークに対して学習を行い, 認識性能を評価する.

4.1 データセット

評価対象として, 画像認識の分野で広く用いられている MNIST 手書き数字データベースを使用する. MNIST 手書き数字データベースは, 28×28 の数字画像と 0〜9 の正解クラスの組みを 50,000 文字含む学習セットと, 10,000 文字含むテストセットを含んでいる. 本論文では, 学習データをさらに 40,000 文字の学習セットと 10,000 文字の検証セットに分割し, 以降, これらを学習セット, 検証セットと呼ぶ.

4.2 ネットワークのサイズ

特徴抽出ユニットおよび識別層への入力特徴マップのサイズには表1の値を用いた. 畳込み層のフィルタの数とサイズ, プーリング層のプーリングサイズは, これらの値から一意に決定される.

表 1: 各ユニットおよび識別層への入力特徴マップのサイズ

特徴抽出ユニット1	特徴抽出ユニット2	特徴抽出ユニット3	識別層
1×28×28	6×12×12	16×6×6	64×1×1

4.3 学習方法

各ネットワークは, 学習セットで 1,000 epoch の学習を行い, 検証セットの認識率が最小である epoch での, テストセットの認識率を評価値とした.

構成ごとに, 評価は異なる初期値で 10 試行され, その平均値, 最小値, 最大値を求めた.

畳込み層のフィルタ係数は, 下式であらわされる normalized initialization [Glorot 2010] で初期化した.

$$\mathrm{U}\left(-\sqrt{\frac{6}{\text{fan-in}+\text{fan-out}}},\sqrt{\frac{6}{\text{fan-in}+\text{fan-out}}}\right) \qquad (7)$$

ここで U は一様分布の乱数, fan-in および fan-out は特徴抽出ユニットの入出力特徴マップの画像数である. 識別層の重みおよびバイアスは 0 で初期化した. これらの値は cross entropy 誤差を最小とするように誤差逆伝搬法で学習し, 学習セットを 500 個ごとに区切ったミニバッチごとに誤差の平均値で更新した. 学習率には 0.1 を用いたが, 発散をふせぐために, 正規化なしの構成で L2,L∞プーリングの場合には 0.01 を, 正規化なしの構成で L1 プーリングの場合には 0.001 を用いた.

認識性能を高めるために, 学習時に画像をゆがめて見かけ上の学習データを増加させる affine distortion 法や, erastic distortion 法 [Simard 2003] が広く用いられているが, 本論文では画像をゆがめるいずれの方法も用いていない.

5. 評価結果

各特徴抽出ユニットをもちいたネットワークの評価結果を図 2 に示す. 評価結果の傾向は次のとおりである.

(1) 正規化の有無(NE→P, NP→P, P→NE 対 P)

正規化やプーリングの種類, 処理順によらず, 正規化が誤認識率の改善に有効であることが確認できる.

(2) 正規化・プーリングの順(NE→P, NP→P 対 P→NE)

正規化の後にプーリングした構成(NE→P, NP→P)に対し, プーリングの後に正規化する構成(P→NE)では, 誤認識率の平均値およびばらつきのいずれも改善される傾向がある. ただし, L1 プーリングを用いた場合には, あまり改善が見られない.

(3) 特徴にまたがるプーリングの有無(P2 対 P1)

プーリングが複数特徴にまたがった構成(P2)で誤認識率が改善される傾向があり, 特に L2 プーリングとの組み合わせで顕著である.

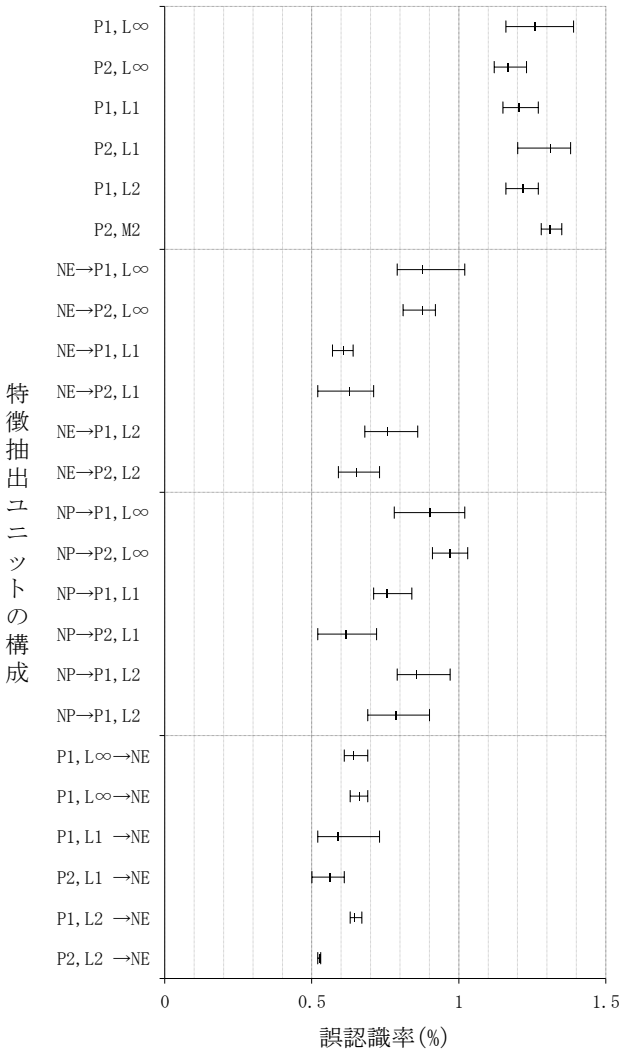

図 2: 各特徴抽出ユニット構成の評価結果

(4) Lp プーリングの選択(L1, L2, L∞)

ほとんどの場合, L1 プーリングで最も低い誤認識率を得られるが, 複数特徴にまたがる L2 プーリングの後に正規化を行う構成(P2,L2→NE)では, 誤認識率の平均値およびばらつきのいずれもが, 特異的に低く抑えられている.

なお, 構成"P2,L2→NE"で得られた平均誤認識率 0.527%は, われわれが知る限り, MNIST データセットを画像のゆがめや前処理なしに学習した場合に得られる最高性能である. これまでの最高性能は 0.53% [Jarrett 2009] であった.

6. 結論と考察

本論文では, 正規化の種類, プーリングの種類および, 正規化とプーリングの順序について, さまざまな組み合わせを用いてそれらの効果を評価した.

その結果, 広く用いられている CNN の構成に最も近い"NP→P1,L1"を含む, L1 プーリングを用いた構成で, 比較的安定して良い認識性能を得られることを確認した. さらに, 特徴をまた

いだプーリングを行うことで，その認識性能を改善できることを確認した．

ほとんどの構成では，L1 プーリングの選択は L2 プーリングよりも良好な認識性能につながったが，われわれが以前に初期視覚野の細胞を参考に導入した“P2,L2→NE”の構成では例外的に，誤認識率の平均およびばらつきのいずれにおいても，最高の認識性能であった．なぜこの構成が例外的にふるまうのかを明らかにすることは，今後の課題である．

CNN は脳の視覚情報処理をまねることで，その認識性能を向上してきた [LeCun 1989, Jarrett 2009, Sekino 2012]．今後も明らかになりつつある脳の処理を取り込むことで，さらなる性能改善が期待できる．

本論文では，プーリング時の重みや範囲の重なりを，かなり簡略化しているので，さらなる改善が可能であると考える．

また，特徴抽出ユニット 1～3 にすべて，同じ構成を用いたが，必ずしも同じ構成を用いる必要はない．入力画像に対しては視覚のエネルギーモデル（二特徴にまたがる L2 プーリング）が意味を持つと期待できるが，位置情報を失ってゆく高次の特徴では，別のプーリング方法や正規化が適切である可能性がある．

参考文献

[Adelson 1984] Edward H. Adelson and James R. Bergen: Spatiotemporal energy models for the perception of motion, The Journal of the Optical Society of America A (JOSA A) Vol. 2, No. 2 February 1985, OSA, 1985.

[Cireşan 2011] Dan C. Cireşan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, Jürgen Schmidhuber: Flexible, High Performance Convolutional Neural Networks for Image Classification, Proc. International Joint Conference on Artificial Intelligence (IJCAI11), AAAI Press, 2011.

[Glorot 2010] Xavier Glorot, Yoshua Bengio: Understanding the difficulty of training deep feedforward neural networks, Proc. International Conference on Artificial Intelligence and Statistics (AISTATS 2010), 2010.

[Jarrett 2009] Kevin Jarrett, Koray Kavukcuoglu, Marc’Aurelio Ranzato and Yann LeCun: What is the Best Multi-Stage Architecture for Object Recognition?, Proc. International Conference on Computer Vision (ICCV'09), IEEE, 2009.

[Kavukcuoglu 2009] Koray Kavukcuoglu, Marc’Aurelio Ranzato, Rob Fergus, Yann LeCun: Learning Invariant Features through Topographic Filter Maps, Proc. International Conference on Computer Vision and Pattern Recognition (CVPR'09), IEEE, 2009.

[Krizhevsky 2012] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton: ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems (NIPS 2012), NIPS, 2012.

[LeCun 1989] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel: Handwritten digit recognition with a back-propagation network, Advances in Neural Information Processing Systems (NIPS'89), NIPS, 1989.

[LeCun 2010] Yann LeCun, Koray Kavukcuoglu and Clément Farabet: Convolutional Networks and Applications in Vision, Proc. International Symposium on Circuits and Systems (ISCAS'10), IEEE, 2010.

[Sekino 2012] 関野 雅則, 木村 俊一, 越 裕: 視覚エネルギーモデルと交差方位抑制モデルに基づく複雑型細胞層を用いた畳み込みニューラルネットワーク文字認識, 第 15 回画像の認識・理解シンポジウム (MIRU2012) 論文集, 電子情報通信学会 PRMU 研究会, 2012.

[Sermanet 2012] Pierre Sermanet, Soumith Chintala and Yann LeCun: Convolutional Neural Networks Applied to House Numbers Digit Classification, International Conference on Pattern Recognition (ICPR 2012), IEEE, 2012.

[Simard 2003] Patrice Y. Simard, Dave Steinkraus, John C. Platt: Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis, Proc. International Conference on Document Analysis and Recognition (ICDAR2003), IEEE, 2003.