

順行伝播する教師信号によるディープネットの新しい学習法

A novel learning method with feedforward supervisory signal

篠崎隆志

Takashi SHINOZAKI

成瀬康

Yasushi NARUSE

独立行政法人 情報通信研究機構 脳情報通信融合研究センター 脳機能計測研究室
Brain Imaging Technology Laboratory, CiNet, National Institute of Information and Communications Technology

We propose a novel learning method for multilayered neural networks which uses feedforward supervisory signal and associates the classification of a new input with that of pre-trained input. A multilayered deep neural network is one of the most powerful methods for human-like recognition tasks, image, and speech recognition. Some previous studies demonstrated great performances with supervised learning in signal classification tasks. Supervised learning for multilayered neural network generally uses gradient-based learning rules, in particular, back-propagation learning. However, the amount of supervisory information in the last layer is insufficient to supervise the whole deep neural network because the information is selected and reduced from layer to layer. This tendency is more serious in signal classification tasks which have limited discrete output. The proposed method effectively uses rich input information in the earlier layer for robust learning and revising of the internal representation in a multilayer neural network.

1. はじめに

深層神経回路は画像認識や [Fukushima 80, Le 12], 音声認識など [Dahl 12] においてヒトと同様の知覚判別を実現するための最も強力な手法の一つである。実際、先行研究では教師あり学習を用いた信号判別課題において非常に高い性能を示している [LeCun 89, Krizhevsky 12]。多層神経回路における教師あり学習は一般的には誤差に基づいた学習、特に逆誤差伝播学習法が用いられている [Rumelhart 86]。しかしながら、フィードフォワード神経回路においては伝播する信号に含まれる情報は層を経るに従って少なくなっていくことから、最終層の誤差のみから生成される学習信号は大規模なネットワーク全体を学習するためには必ずしも十分であるとは言えない。さらに単一の判別課題においては最終層の出力は基本的にスパースな形態を示すことから、学習のベースとなる誤差情報についても必然的にスパースなものとなり学習情報の不足はさらに顕著になると考えられる。近年用いられている自己符号化器においてはこれを解決するために層ごとの学習 (layer-wised learning) を繰り返す積層自己符号化器と呼ばれる手法が用いられているが [Bengio 07], 追加学習が困難になるという問題が存在する。

本研究はディープネットにおいて順行伝播する教師信号から、事前学習によって構築された入力信号に対する反応の関連性を学習するための手法である先行伝播学習法を提案する。画像認識課題における検証の結果から、提案する手法は入力された順行伝播する信号の豊富な情報量を利用することによって、多層神経回路における安定で高速な学習を実現する。さらに学習の過程において内部表現が特定の入力に対応した形態から、より一般性の高い形態へと適合していくことが確認された。

2. モデル

2.1 ネットワーク構造

ネットワークの構造はヒト視覚系の皮質上の構造を模したものをを用いた。ネットワークは Kohonen の自己組織化マップ (Self Organizing Map: SOM) [Kohonen 82] をモジュールの基本単位として構成された。個々の SOM モジュールは 100 個のニューロンより構成され、モジュールごとに一つ前の層から受容野 (Receptive Field: RF) に基づいた部分入力を受け取る。図 3.(a) に用いたネットワークの概要図を、表 1 に構造のパラメータの値を示す。例として、1 層目のネットワークは 49 個の SOM、すなわち 4900 個のニューロンから構成され、個々のニューロンは 28×28 ピクセルの入力画像のうちの受容野に相当する 6×6 ピクセルからの入力を受け、層全体で 176,400 個のシナプスを構成する。シナプスの重みの初期値は一樣乱数によって生成し、重みベクトルをニューロンごとに 2 ノルムで規格化した。

個々のニューロンは入力と重みとの内積を計算し、それぞれが含まれる SOM モジュール内で最も大きな値を持つニューロンの出力を 1.0 とし、その近隣のニューロンが距離に依存して減衰する値を出力する Winners Share All (WSA) として働く。距離による減衰にはガウス関数 $G(d) = \exp(-d^2/2\sigma^2)$ を用いた。 d はあるニューロンと最も大きな値を持つニューロンからの SOM 内での距離を表し、 σ は 0.4 とした。

2.2 事前学習

事前学習には SOM の最も一般的な学習法である、教師なしの競合学習を用いた [Kohonen 82, Rumelhart 85]。競合学習では入力ベクトルと重みベクトルとの内積を求め、値が SOM 内で最大となるニューロンおよびその近傍のニューロンの重みベクトルを入力ベクトルに近づけるように更新する。重みの更新の規則は以下のように記述される。

$$\Delta \mathbf{w} = \alpha \mathbf{x}_{\text{input}} \exp(-d^2/2\sigma^2) \quad (1)$$

ここで \mathbf{w} は重みベクトル、 $\mathbf{x}_{\text{input}}$ は入力ベクトル、 d はあるニューロンと最も大きな値を持つニューロンからの SOM 内で

層	マップ数	ニューロン数	受容野サイズ	受容野ステップ	シナプス数
1	7×7	4,900	6×6 pixels	4×4 pixels	176,400
2	5×5	2,500	3×3 maps	1×1 map	2,250,000
3	3×3	900	3×3 maps	1×1 map	810,000
4	1×1	100	3×3 maps	1×1 map	90,000

表 1: ネットワーク構造のパラメータ

の距離を表す。学習係数 α は初期値 1.00 から学習セッション終了時の値 0.00 へと単調に減少し、ガウス関数による距離減衰係数 σ も同様に初期値 3.5 から終了時の値 0.0 へと減少する。重みベクトルは以下の式のように値の更新ごとに規格化される。

$$\mathbf{w}_{\text{new}} = (\mathbf{w} + \Delta\mathbf{w}) / |\mathbf{w} + \Delta\mathbf{w}| \quad (2)$$

学習によって SOM モジュール内において似通った信号に対応したニューロン同士が近くに配置される、(Topographic Independent Component Analysis: TICA [Hyvärinen 01]) と同様の二次元空間構造が形成される。

入力には MNIST の手書き数字の画像データセット (0 から 9 まで 10 種類, 28×28 ピクセル, グレースケール) [LeCun 14] を用いた。事前学習では生体における臨界期 (critical period) をヒントに、層ごとに順番に学習を進めていく方式 (layer-wised learning) を用いた。最初に第 1 層目のみが 10,000 サンプルで学習され、次に 2 層目が追加した状態で 10,000 サンプルを学習のように層を増やすごとに 10,000 サンプル学習させる。最終的には第 1 層目が 40,000 サンプル、第 2 層目が 30,000 サンプル、第 3 層目が 20,000 サンプル、第 4 層目が 10,000 サンプルで学習することになる。図 3.(b) に事前学習で第 1 層目に構築されたヒトの一次視覚野の方位選択制カラムに似た 2 次元構造を示す。このように事前学習によって入力画像を認識するための基礎となる内部表現が構築されていることが見て取れる。

2.3 先行伝播学習法

本研究で提案する先行伝播学習法は入力間の信号の関連付けによって学習する教師あり学習法である。提案する学習法は Learning Vector Quantization (LVQ) [Ahalt 90] と呼ばれる SOM の学習法を基にしており、これを多層ネットワークに対応させ順行伝播の教師信号を追加したものである。新規に学習する信号を入力する前に、学習が完了した入力信号を先行して入力し、ネットワークの最終層まで伝播させる。その後、新規に学習する信号を先行して入力された信号の残効を加味した状態で伝播させる。この後、LVQ と同様の出力の分類が成功か否かによって強化か減衰が切り替わる競合学習を行うことによって、新規の入力信号に対して求める分類が行われるような学習が行われる。

提案する学習法は以下の手順で行われる。はじめに学習する目標である入力信号をネットワークで処理し、出力される結果のラベルを確認する。出力されたラベルが正解でなかった場合のみ先行伝播学習法が適用される。先行伝播学習法では事前学習によって求めるラベルを出力することが確認されている順行入力 \mathbf{x}_{adv} を先行伝播信号としてネットワークで処理する。続いて、先ほど正解ラベルを導かなかつた、学習の目標となる入力信号 $\mathbf{x}_{\text{target}}$ を先行伝播信号の残効のもとで処理する。この時、実際のネットワークに処理される入力 $\mathbf{x}_{\text{input}}$ は以下の

ようになる。

$$\mathbf{x}_{\text{input}} = \beta \mathbf{x}_{\text{adv}} + (1 - \beta) \mathbf{x}_{\text{target}} \quad (3)$$

ここで β は全体の入力に対する先行入力の比率を表す。この入力ベクトル $\mathbf{x}_{\text{input}}$ は目標の入力ベクトルの方向を、教師信号である先行伝播信号 \mathbf{x}_{adv} の残効によって修正したものとなる。ここで重要なことは、ネットワークのダイナミクスは WSA に表されるように強く非線形であり、先行伝播信号 \mathbf{x}_{adv} の残効のもとでの入力ベクトル $\mathbf{x}_{\text{input}}$ の処理は、単純に 2 つの線形和とはならず、先行伝播信号での SOM の活動パターンにより近づいたものとなることである。このように処理された入力信号は式 1 のように、事前学習で用いたものと同様の競合学習によって学習される。以上に加え、多層化に対応する項と SOM モジュール内での距離減衰の項を加えた学習の式は以下のように記述される。

$$\Delta\mathbf{w} = r^{n-l} \alpha (\beta \mathbf{x}_{\text{adv}} + (1 - \beta) \mathbf{x}_{\text{target}}) \exp(-d^2/2\sigma^2) \quad (4)$$

ここで r は層から層への学習の減衰係数で、 n はネットワークの層の総数、 l は学習を適用する層となる。これによって入力に一番近い層では一番弱い学習が行われ、出力に近づくほど強い学習が行われる。重みベクトルは更新の度に式 2 にしたがって規格化される。パラメータには $(\alpha, \beta, r, n) = (0.05, 0.6, 0.5, 4)$ を用いた。

3. 結果

事前学習が完了した状態のネットワークに対して先行伝播学習法を適用した。事前学習の結果において、それぞれのラベル (「0」から「9」までの数値) に対して最大の値を出力したニューロン (代表ニューロン) を発火させる入力画像を先行伝播学習法のための教師信号として用いた。学習は、ランダムで選択された入力画像に対してまず教師信号なしの状態での出力を確認し、出力のラベルが求めるラベルと異なる場合にのみ先行伝播の教師信号を併せて提示して学習法を行った。教師信号としての入力画像は、学習中も代表ニューロンが発火させる入力があるごとに更新した。

図 3.(c) は先行伝播学習法によるエラー率の変化を示す。最終的に先行伝播学習法は、事前学習直後の $17.2 \pm 1.0\%$ のエラー率 (特に代表ニューロンによるものだけに限定した場合は $80.3 \pm 1.2\%$) を、トレーニングセット (60,000 サンプル) の 20 回の適用後に $5.6 \pm 0.1\%$ にまで低下させた ($n=10$)。エラー率は先行伝播学習初期に事前学習直後よりも悪化したのち、続く学習過程で安定して減少していく。この初期のエラー率の悪化は内部表現構造の再編成によって生じていると考えられる。初期状態ではより同一のラベルを持つ入力に対してより分散して冗長であった内部表現が、先行伝播学習法によって統合され、より効率の良い、ネットワーク活動的にはスパースな状態に移行したと言える (図 3.(d))。これによって代表ニュー

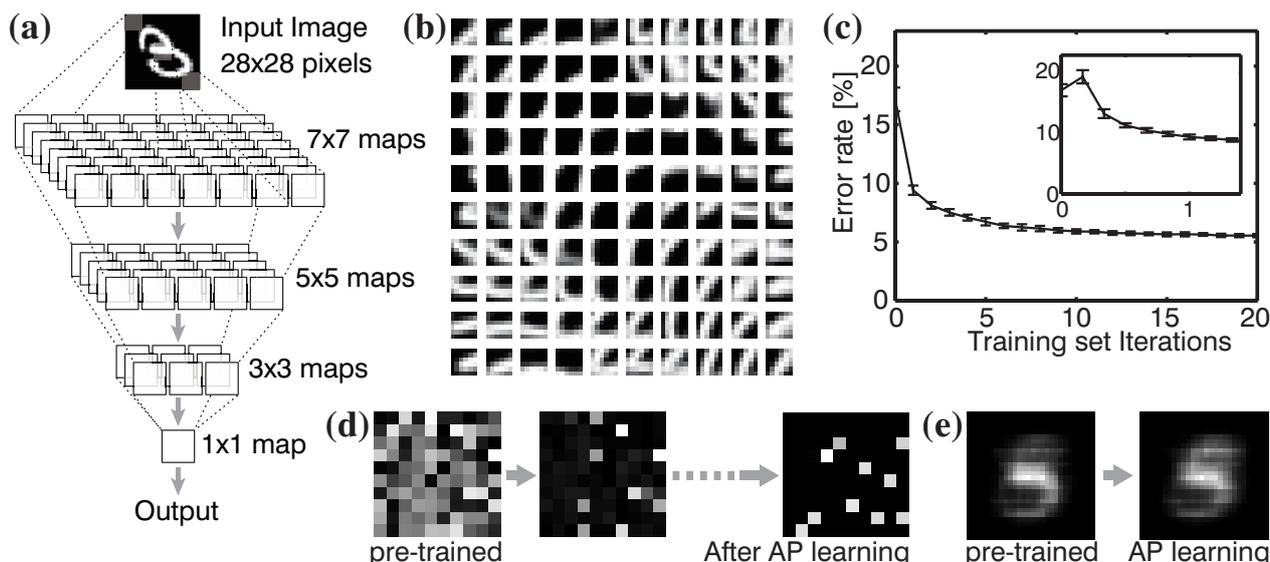


図 1: (a) ネットワーク構造の概要．(b) 事前学習によって第 1 層に構築された機能地図の例．方位選択制が 2 次元空間で形成されている．(c) 学習によるエラー率の変化．内側の図は教師あり学習開始部分の拡大図．エラー率は標準偏差 ($n=10$)．(d) 学習による最終層におけるニューロンの使用状態の変化．主要ニューロンに表現に統合されていく状態が見て取れる．(e) 提案する学習法適用前と後の主要ニューロンの内部表現の一例 (数値の「5」に対応するもの)．教師あり学習により，特徴的な形状のみに対応した状態から，より一般的な形状への適応を獲得している．

判別器	エラー率 [%]
提案する学習法	5.6
2-layer NN, 1000 hidden units (LeCun <i>et al.</i> , 1998)[LeCun 98]	4.5
Convolutional net LeNet-5 (LeCun <i>et al.</i> , 1998)[LeCun 98]	0.95
Committee of 35 conv. net (Ciresan <i>et al.</i> , 2012)[Ciresan 12]	0.23

表 2: 先行研究とのエラー率の比較

ロンが表す内部状態も図 3.(e) のように，初期状態では特異的な形状にのみ対応している状態から，先行伝播学習後はより一般的な形状に対応するように変化しているのが見て取れる．

4. まとめ

本報告では多段のフィードフォワード神経回路における新しい教師あり学習法である先行伝播学習法を提案し，画像認識課題によってその検証を行った．提案する学習法は順行伝播する入力信号の豊富な情報量を利用することによって効率良くかつ安定した教師あり学習を実現する．実際，画像認識課題による検証では，提案する学習法は事前学習された多層神経回路を教師あり学習によって再編成 (図 3.(c)) し，その内部表現についても再構成によってより汎化した表現が獲得させることが確認された (図 3.(d))．画像認識課題における認識性能については，よりチューニングの進んだ先行研究 [LeCun 98, Ciresan 12] と比較すると未だ不十分ではあるが (表 2)，先行研究の手法による初期の性能には到達しており，パラメータやネットワーク構造等のチューニングや手法の改善によってさらなる発展が見込まれる．将来的には局所的な活動の度合いに応じた学習則を適用することによって，より正確で効率の良い学習法の確立を目指す．

従来の教師あり学習では学習を安定させるために層ごとの学習 (layer-wise learning) を繰り返すのが一般的であったが，

提案する学習法はネットワーク全体を一括して学習する．さらにネットワーク全体に伝えられる学習情報としては正解/不正解のみを用い，局所的にはそこを伝播した順行伝播信号そのものを教師あり信号として利用する．このことはハードウェアの実装において誤差情報の入力が不要であることを意味しており，これによって提案する学習法はより大規模化に適した構造を持つと共に，大脳皮質の神経回路をより模したものとなっている．

提案する学習法の優れた特徴の一つとして，他の神経回路の学習手法 (強化学習 [Sutton 98] および競合学習 [Kohonen 82, Rumelhart 85]) との親和性がある．提案する学習法は，先行伝播する教師信号を用いずに，ただ正解と不正解のみを利用した場合には強化学習として振る舞い，さらに正解と不正解の情報を用いない完全な教師なし学習として実行した場合は SOM などに代表される一般的な競合学習を行う．従ってこれらの学習法は同一のハードウェア実装で実現できると考えられ，先行伝播の有無，もしくは正解/不正解の利用の有無によってその学習モードを切り替えうると考えられる．さらに，正解/不正解を適用する層を変化させることによって，関連付けの度合いを調整できるとともに，より深いネットワークやリカレントネットワークでの利用が考えられる．

提案する学習法は新規の入力情報をすでに学習済みの入力情報のラベルと関連付ける学習法である．このような学習の一

番の利点は学習する入力データについては関連性の情報のみが必要であり、その意味について厳密に定義しなくとも適用できるということがある。これはビッグデータにおけるデータマイニングのように情報の大まかな分類はできているものの、その意味付けや分類構造が明確でない場合にも本手法が有効であることを示唆しており、これに図 3.(e) のような内部表現の視覚化を組み合わせることによって、データの持つ隠れた構造を明らかにできる可能性が期待される。一方で自然画像認識などに応用する場合においても、新規の自然画像を、学習済みの人工的でクリーンなデータと関連付けられる可能性が考えられ、これによって自然画像内の文字認識などが可能になると期待される。また先行伝播する教師信号についても、本報告においては単一の信号を用いていたが、これを多重化(例えば「赤く」で「丸い」果物として「りんご」を学習する)することも可能であり、これによって特に自然界のマルチモーダルな情報の統合(例えば画像と音など)が実現できる可能性が考えられ、さらなる発展が見込まれる。

参考文献

- [Fukushima 80] Fukushima, K.: Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biol Cybern.*, **36**(4), pp. 193-202 (1980)
- [Le 12] Le, Q.V., Ranzato, M.A., Monga, R., Devin, M., Chen, K., Corrado, G.S., Dean, J. & Ng, A. Y.: Building high-level features using large scale unsupervised learning, *Proc. 29th Int. Conf. on Machine Learning* (2012)
- [Dahl 12] Dahl, G.E., Yu, D., Deng, L. & Acero, A.: Context-dependent pre-trained deep neural networks for large vocabulary speech recognition, *IEEE Trans. on Audio, Speech, and Lang. Process.*, **20**(1), pp. 30-42 (2012)
- [LeCun 89] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard R.E. & Jackel, L.D.: Backpropagation applied to hand-written zip code recognition, *Neural Computation*, **1**(4), pp. 541-551 (1989)
- [Krizhevsky 12] Krizhevsky, A., Sutskever, I & Hinton, G.E.: ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems 25*, pp. 1106-1114 (2012)
- [Rumelhart 86] Rumelhart, D.D., Hinton, G.E. & Williams, R. J.: Learning representations by back-propagating errors, *Nature*, **323**(9), pp. 533-536 (1986)
- [Bengio 07] Bengio, Y., Lambling, P., Popovici, D. & Larochelle, H.: Greedy layer-wise training of deep networks, *Advances in Neural Information Processing Systems 19*, pp. 153-160 (2007)
- [Kohonen 82] Kohonen, T.: Self-organized formation of topologically correct feature maps, *Biol. Cybern.*, **43**, pp. 59-69 (1982)
- [Hinton 06] Hinton, G.E., & Salakhutdinov, R.: Reducing the dimensionality of data with neural networks, *Science*, **313**, pp. 504-507 (2006)
- [Rumelhart 85] Rumelhart, D.E., & Zipser, D.: Feature discovery by competitive learning, *Cognitive Science*, **9**, pp. 75-112 (1985)
- [Hyvärinen 01] Hyvärinen, A. & Hoyer, P.O.: A two layer sparse coding model learns simple and complex cell receptive fields and topography from natural images, *Vision Research*, **41**(18), pp. 2413-2423 (2001)
- [LeCun 14] LeCun, Y. & Cortes, C. *The MNIST database of handwritten digits*.
- [Ahalt 90] Ahalt, S.C., Krishnamurthy, A.K., Chen, P. & Melton D.E.: Competitive learning algorithms for vector quantization, *Neural Networks*, **3**(3), pp. 277-290 (1990)
- [LeCun 98] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, **86**(11), pp. 2278-2324 (1998)
- [Ciresan 12] Ciresan, D., Meier, U. & Schmidhuber, J.: Multi-column deep neural networks for image classification, *CVPR 2012*, pp. 3642-3649 (2012)
- [Sutton 98] Sutton, R.S. & Barto, A.G.: Reinforcement Learning: An Introduction, The MIT Press (1998) (邦訳: 強化学習, 三上貞芳・皆川雅章 訳, 森北出版 (2000))