

# 問題への適切性を考慮した 畳み込みニューラルネットワークの初期値決定手法

鈴木 藍雅<sup>1,2,a)</sup> 庄野 逸<sup>3,b)</sup> 坂無 英徳<sup>2,1,c)</sup>

**概要:** ディープラーニングの一モデルとして画像認識の問題に広く用いられる深層畳み込みニューラルネットワーク (DCNN) は、その誤差逆伝搬法による学習において、実現される性能がネットワークの初期値に大きく依存することが知られている。本稿では目的とする問題に応じた DCNN の初期値決定を高速に行うための新たな手法を提案する。提案手法では目的とするデータから特徴抽出に寄与するパラメータを、DCNN 全体の誤差逆伝搬法による繰り返しの学習を行わず、入力空間の特徴点との相関の最大化によって one-shot に初期値を定めることで、学習を効率化し汎化性能の向上を図る手法である。実験においては複数のデータセットに対し、現在主流とされている初期値決定手法との性能比較を行い、提案手法によって有意に DCNN の識別性能が向上することを確認した。

**キーワード:** deep convolutional neural networks, image recognition, weight initialization, one-shot learning

## Efficient Weight Initialization of Convolutional Neural Networks based on the Effectiveness for Target Tasks

AIGA SUZUKI<sup>1,2,a)</sup> HAYARU SHOUNO<sup>3,b)</sup> HIDENORI SAKANASHI<sup>2,1,c)</sup>

**Abstract:** Deep convolutional neural networks (DCNNs), recently known as one of the “deep learning,” have a strong influence on the initial state of parameters to its performance. In our previous study, we have proposed a feature transfer method, which gives an efficient initial state appropriate for the assigned tasks, and have investigated that the effectiveness of the initial state depends on the tasks. In this paper, we propose a novel DCNN’s weight initialization method which regards the efficiency of the given task and works fastly. Our proposed method calculates effective initial weights which represent the salient feature expressions in the input space without training of entire DCNNs in one-shot. In the experiment, we compared model performances between conventional initialization method and our proposed one and investigated significant improvements of the proposed method for some recognition tasks.

### 1. はじめに

深層畳み込みニューラルネットワーク (Deep Convolutional Neural Network; DCNN) は、

Krizhevsky らによる AlexNet[1] が大規模画像分類コンペティションである ILSVRC2012 において成功を収めて以来、コンピュータビジョンの分野におけるデファクトスタンダードの学習モデルになりつつある。DCNN は Fukushima らの哺乳類初期視覚野を模した計算モデルであるネオコグニトロン [2] をルーツに持つ、階層型ニューラルネットワークモデルの一種である。

DCNN は入力画像から本質的な特徴を抽出し次元削減を行う前段の “特徴抽出部” と、抽出された特徴に対して実

<sup>1</sup> 筑波大学システム 情報工学研究科  
茨城県つくば市天王台 1-1-1

<sup>2</sup> 産業技術総合研究所 人工知能研究センター  
茨城県つくば市梅園 1-1-1

<sup>3</sup> 電気通信大学 情報理工学研究科  
東京都調布市調布ヶ丘 1-5-1

a) ai-suzuki@aist.go.jp

b) shouno@uec.ac.jp

c) h.sakanashi@aist.go.jp

際に識別を行う後段の“識別部”に大別される。DCNN の特徴抽出部では、哺乳類の初期視覚野における受容野構造と階層的な情報処理に関する生理学的知見に倣い、局所領域における特徴抽出処理を階層的に積み重ねることによって入力からの特徴抽出を実現している。DCNN の特筆すべき性質として、特徴抽出部と識別部を含めたモデル全体の End-to-End な学習によって、従来のパターン認識のモデルでは手作業により設計されていた、入力データからの特徴抽出を自動的に獲得できることが挙げられる。

一般に DCNN の教師あり学習は、誤差逆伝搬法と呼ばれる勾配法に基づく連続最適化問題として実現される。そのため勾配法の特性に伴って、学習によって実現される汎化性能はパラメータの初期値により大きく左右されることが知られている [3]。特に学習に用いることのできるデータが十分でない条件下において、適切でない初期値から学習が行われた場合、パラメータが局所解に停滞するなどして良好な汎化性能が得られない場合がある。

こうしたニューラルネットワークの初期値決定問題に対する先行研究のほとんどは以下の 3 つに大別される。

- (1) 良い性質を持つランダム値を生成するもの [3], [4]
- (2) パラメータを直交制約の元で生成・学習するもの [5]
- (3) 予め目的の問題に適合したパラメータを与えるもの [6], [7]

筆者は (3) の視点に立った転移学習の考えに基づき、目的のタスクに類似した収集が容易なデータによる事前学習を行うことで、目的タスクにおける汎化性能を向上させる初期値決定手法を提案した [7]。この手法は (1), (2) に挙げた手法に比べ高い汎化性能を達成できる一方で、その処理に膨大な計算コストを要する問題があり、計算資源の限られた応用先への可用性に欠ける。さらに事前学習の際に適切でない転移元データを選択した場合、事前学習を行わない場合と比較して汎化性能が悪化することが報告されている [8]。このように DCNN の初期値の適切性は目的とする問題に強く依存しており、事前学習に用いるデータの選定に緻密な調整を要し、適用のコストが高いことも欠点として挙げられる。

本稿ではこれらの知見を元に目的とする問題への適切性を考慮しながら、転移学習に基づく手法の欠点であった計算コストの問題と、事前学習に用いるデータの選択の困難を解消した新たな DCNN の初期値決定手法を提案する。提案手法は転移学習のメカニズムの解析に関する先行研究 [8] で得られた、“DCNN の初期状態は入力されるデータの特徴を予め捉えられていることが好ましい” という知見を元に、ネオコグニトロンにおける競合型学習手法 [9] から着想を得たものである。提案手法では DCNN が初期状態において学習データの特徴を表現できるよう、誤差逆伝搬法とは異なる one-shot なアルゴリズムを用いて、目的とするタスクのデータのみから適切なパラメータの初期化

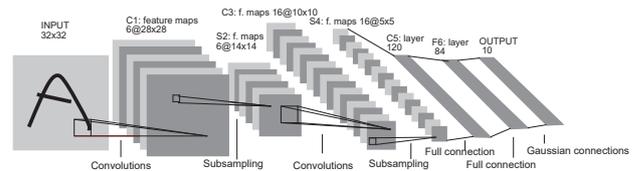


図 1 LeNet-5 の概要図。[10] より引用。畳み込み・空間プーリング (Subsampling layer) からなる特徴抽出部 (C1, S2, C3, S4) と、全結合層からなる識別部 (C5, F6) に二分できる。

を行い汎化性能の向上を図る。

実験では単純な構造を持つ古典的な DCNN である LeNet-5 を用いて、複数の小規模データセットを分類する問題に対して、提案手法を適用した。さらに、近年主流となっている He らによる初期値決定手法である、MSRA[3] と汎化性能を比較し、提案手法の有効性を示した。

## 2. 深層畳み込みニューラルネットワーク (DCNN)

DCNN はコンピュータビジョンにおいて広く用いられる階層型ニューラルネットワークの一種であり、近年ではディープラーニング (deep learning) のモデルの一つとして広く用いられている。DCNN の構造は、基本的に“畳み込み”と“空間プーリング”からなる特徴抽出部と、全結合層の積み重ねからなる識別部に大別される。図 1 に実験で用いる、最初期の単純な DCNN である LeNet-5 の概要図を示す。

DCNN の前段における特徴抽出部の畳み込み、空間プーリングの各層は一般に多チャンネルの 2 次元画像の 3 階テンソルを入力とし、非線形変換を通して特徴マップと呼ばれる 3 階テンソルを出力する。いま DCNN の  $i$  層目に入力される  $c$  チャンネルからなる  $(h, w)$  のサイズの画像を  $f_i(l, \mathbf{x}) \in \mathbb{R}^{c \times h \times w}$  とする。ここで  $l$  は画像のチャンネル、 $\mathbf{x} \in \mathbb{Z}^2$  は離散画像中の平面位置を表す。このとき畳み込み層の活性は

$$f_i^{\text{conv}}(k, \mathbf{x}) = \phi \left( \sum_{l, \mathbf{u}} \mathbf{g}_i(k, l, \mathbf{u}) f_{i-1}(l, \mathbf{x} - \mathbf{u}) \right) \quad (1)$$

と表される。ここで  $k$  は出力のチャンネル、 $\mathbf{g}_i(k, l, \mathbf{u}) \in \mathbb{R}^{c \times s \times s}$  は畳み込み層の振る舞いを決定するフィルタテンソルで、 $\mathbf{u} \in \mathbb{Z}^2$  はフィルタテンソルの添字である。また、 $\phi(y)$  は畳み込み層の非線形活性化関数であり、 $\phi(y) := \max(0, y)$  によって定義される Rectified Linear Unit (ReLU) が一般に用いられる。この畳み込み層の処理は入力中の空間局所領域 (受容野) におけるフィルタテンソルとの相関を求める処理であり、この処理により空間の局所特徴が抽出される。

空間プーリング処理は畳み込み層で得られた特徴マップに対して、入力を空間的にぼかす処理を施すことによって、その次元数を削減するとともに微小変形に対する活性の頑健性を与える処理である。特徴マップ  $f_i(l, \mathbf{x})$  に対し

て、プーリング層の出力は

$$f_i^{\text{pool}}(k, \mathbf{x}) = \max_{\mathbf{r} \in N(\mathbf{x})} (0, f_i(k, \mathbf{r})) \quad (2)$$

と表される。ここで  $N(\mathbf{x})$  は、特徴マップの位置  $\mathbf{x}$  におけるチャンネルにまたがる空間近傍である。式 2 のような、ある領域における最大値を代表値とするプーリング処理は最大値プーリング (max-pooling) と呼ばれる。

これらの特徴抽出部の処理において、その振る舞いを決める調整可能パラメータはフィルタテンソル  $\mathbf{g}_i$  のみである。DCNN の特徴抽出部では、この  $\mathbf{g}_i$  を学習により決定することで入力画像からの特徴表現を獲得する。

また DCNN 後段の識別部は、特徴抽出部で得られた特徴表現を入力とした分類を行う多層パーセプトロンとして働く。特徴抽出部を含めた DCNN 全体を勾配降下法により学習することで、特徴抽出と識別のプロセスを同時に最適化できる。

### 3. 提案手法

提案手法では、学習の初期状態において DCNN の特徴抽出部が予め入力データの特徴を表現するよう、適切なフィルタテンソル  $g_i$  の状態を定める。このフィルタの生成手法は、ネオコグニトロンにおける競合型学習 [9], [11] と呼ばれる学習アルゴリズムから着想を得ている。競合型学習では、入力空間の局所特徴を表現するニューロンを生成しようとする作用 (興奮性作用) と、すでに生成されているものと似た作用を持つニューロンが生成されるのを抑制する作用 (抑制性作用) を競合させることによって、疎な分散表現の学習を行う。提案手法ではこの競合型学習の考え方に基づき、入力データのうち特徴的な部分に相関するフィルタ (シードフィルタ\*) を生成しつつ、似たフィルタ同士を問引く処理を行うことによって、入力空間に現れる特徴を表現した多様なフィルタを one-shot に生成する。

DCNN に入力  $x$  が与えられたとき、 $l$  層目の畳み込みフィルタの初期値決定を行う手続きの流れを Algorithm 1 に示す。提案手法においては入力データが与えられたとき、はじめに初期値決定を行う畳み込み層まで信号を順伝搬し、ある時点での活性を求める (Algorithm 1, Line 3)。この時あるしきい値  $\theta$  以上の活性を示さないならば畳み込み層がその入力に対する特徴表現を獲得できていないとみなし、畳み込み層への入力のうちから特徴点を選び出し、その成分に活性を示すフィルタを生成しシードフィルタとする (Algorithm 1, Line 10–12)。シードフィルタは標準 Gauss ノイズによるランダムフィルタと混合比  $\delta$  による重み付き和によって、畳み込み層の初期値となる。

反対に、ある入力に対して畳み込み層がしきい値  $\theta$  以上の活性を示しているならば、その入力に対する新たなフィ

\*1 ランダムフィルタと混合することによって、畳み込み層の初期値をなすため Algorithm 1 における  $g_x$  をシードフィルタと呼ぶ

#### Algorithm 1 畳み込み層の初期値決定アルゴリズム

**Require:**

入力刺激:  $x$ , 学習閾値:  $\theta$ , 抑制強度:  $\epsilon$   
 許容半径:  $r_t$ , フィルタ混合係数:  $\delta$

**Ensure:**

フィルタテンソル  $g_i \in \mathbb{R}^{c \times s \times s}$  ▷  $c$  チャンネル,  $s \times s$  ピクセル

```

1: function GENERATE_FILTER( $x$ )
2:   Initialize:
3:      $prev :=$  注目畳み込み層の入力  $f_{l-1}(x)$ 
4:      $act :=$  注目畳み込み層の出力  $f_l(x)$ 
5:      $seed \leftarrow \phi(prev - \epsilon \cdot \max(prev))$ 
6:     while  $p \in act$  do ▷  $act$  内の全ての位置を走査
7:       if  $act[p] - \theta > 0$  then
8:          $act[N(p, r_t)] \leftarrow 0$  ▷  $N(\cdot, r)$  は  $\cdot$  の  $r$  近傍
9:       end if
10:    end while
11:     $seed \leftarrow seed \odot act$  ▷  $\odot$  は Hadamard 積
12:     $\zeta^* \leftarrow \arg \max_{\zeta} \|seed[N(\zeta, s)]\|_2^2$ 
13:     $g_x \leftarrow seed[N(\zeta^*, s)] / \|seed[N(\zeta^*, 2)]\|_2^2$ 
14:     $g_r \leftarrow \{x_{n,i,j} \sim \mathcal{N}(0, 1) \mid 1 < n < c, 1 < i, j \leq s\}$ 
15:    return  $(1 - \delta) \cdot g_r + \delta \cdot g_x$  ▷  $\delta$  による重み付き和
16: end function
    
```

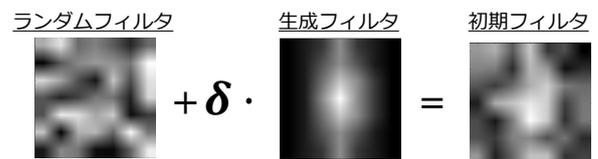


図 2 生成された初期フィルタの例。  $\delta = 0.5$ 。ランダムフィルタと生成されたシードフィルタの重み付き和より初期フィルタが決定される。

ルタ初期値は生成しない。この処理は似たようなフィルタが生成されることで、初期値の多様性が失われないようにする処理であり、競合型学習における抑制性作用と対応している。

実際にモデル初期値を決定する際には、この手続きを全ての学習データに対して 1 度ずつ行い、得られたフィルタ集合の中の活性 (Algorithm 1,  $act$  の Frobenius ノルム) の大きな上位  $k$  個を選択する。この手続きによって生成されるフィルタの例を図 2 に示す。

### 4. 実験

実験では古典的で単純な構造を持つ DCNN である、LeNet-5[10] を用いた。LeNet-5 は畳み込み 2 層と多層パーセプトロンからなる DCNN であり、提案手法によって初期化されるパラメータは畳み込み 1 層 (C1) のフィルタ  $\mathbb{R}^{20 \times 3 \times 7 \times 7}$  と畳み込み 2 層 (C3) のフィルタ  $\mathbb{R}^{50 \times 20 \times 5 \times 5}$  である。

対象とする問題は手書き数字認識データセットである MNIST と小規模自然画像データセットである CIFAR-10, CIFAR-100 の 3 つのデータセットに対する識別問題を用いた。表 1 にそれぞれのデータセットの概要を示す。実験において初期値による性能変化の影響を明確にするため、各

表 1 実験に用いたデータセット

	種類	サイズ	学習数	検証数
MNIST	数字	1 × 28 × 28	60000	10000
CIFAR-10	自然画像	1 × 32 × 32	50000	10000
CIFAR-100	自然画像	1 × 32 × 32	50000	10000

学習データの総数をクラスごとの比を保って  $r = 1, 5, 10\%$  に制限し学習に用いた\*2。提案手法の有効性の検証のために、現在主流とされ広く用いられているランダム初期値の決定手法である MSRA[3] と提案手法で DCNN の初期化を行い、学習の結果得られたテストデータに対する識別精度 (Accuracy) によって汎化性能の比較を行う。ネットワークの学習は一貫して、バッチサイズを 128 とし Adam[12] によってパラメータの更新を訓練誤差が十分に収束するまで行った。フィルタ混合係数  $\delta$  はタスクごとに汎化性能が向上する値を探し出した。評価における Accuracy は、極端に良い外れ値の性能の影響を排除するために、学習が十分に収束した後の 75-パーセンタイル値 (Q3/4) を代表値として用い、異なる乱数シードから学習を  $n = 10$  回行った際の平均値を  $\alpha = 0.95$  を有意水準とした t 検定を行い、比較を行った。

## 5. 結果

提案手法と MSRA の比較結果を表 2 に示す。ここで混合比  $\delta$  は予備実験の結果得られた最良の値である。

平均の Accuracy はほとんどの条件において、提案手法が MSRA による性能を上回っている。しかし  $r = 1\%$  の例と CIFAR-100 における例では、平均に有意差は見られなかった。これは  $r = 1\%$  の場合では学習データ数が 5,600 件と極端に少なすぎることに由来する過学習が起きていること、更に CIFAR-100 の場合では用いた LeNet-5 アーキテクチャの表現力の不足から、その学習結果に大きなばらつきが生じているためと考えられる。その他の良い条件下においては汎化性能は有意に向上しており、とくに CIFAR-10 の例においては顕著な差が見られることから、提案手法は MSRA に比べて適切な初期値を与えていると考えられる。

一方で本手法はフィルタ混合比  $\delta$  に対して極めて鋭敏であり  $\delta$  が適切に選ばれない場合、MSRA による汎化性能を下回るばかりか、誤差逆伝搬法による学習が全く進行しなくなる現象が観察された。例えば MNIST の学習において  $r = 10\%$ ,  $\delta = 0.3$  としたときの平均の Accuracy は 20.31% であり、MSRA の 98.71% と比較して、著しく低い。これはフィルタが過学習を起こしたような状況になっていると予測されるが、原因は定かではない。

\*2 実験で用いた LeNet-5 の複雑度が低いため全ての学習サンプルを用いた場合、初期値による汎化性能に有意な差は見られなかった

## 6. おわりに・今後の課題

本稿では DCNN の学習において、最終的な汎化性能を左右する重要な要素である初期値決定問題について新たな手法を提案した。提案手法は現在主流とされている“目的とする問題を考慮しないランダム値によるもの”ではなく、問題を考慮した決定論的な初期値決定を行う手法である。提案手法では入力空間における顕著な局所特徴に対応する成分を予めパラメータに含ませることにより、学習の効率化を図る。実験では提案手法を単純な DCNN による小規模データセットの分類問題に適用し、従来のランダム値による初期化手法と比較して、DCNN の汎化性能が有意に向上することを示した。提案手法は転移学習に基づく事前学習手法とは異なり、初期値決定において DCNN 全体の学習を必要としないため極めて高速で、かつ目的とする問題のデータのみを用いるため、転移学習における転移元データの恣意性を排することができる。

今後は提案手法の実用化に向け、アルゴリズムのさらなる改良を行う。特に 5 節で報告したように、シードフィルタとランダムフィルタの混合係数  $\delta$  は、最終的な汎化性能に極めて大きな影響を及ぼすことがわかっている。この値は実験結果において観察できるように、対象とする問題に応じて変化することが予想されるため、最適な  $\delta$  を人手による調整なしに決定する手法の確立を行う必要がある。また、本稿の実験では小さな DCNN を用いて、比較的単純な分類問題に対する有効性の検証を行ったが、提案手法がより複雑な問題に対しても適用できるかを検証するため、ImageNet[13] など複雑な問題において、大規模な DCNN を用いる際の有効性の検証を行う。

## 参考文献

- [1] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, *Advances in neural information processing systems*, pp. 1097–1105 (2012).
- [2] Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biological cybernetics*, Vol. 36, No. 4, pp. 193–202 (1980).
- [3] He, K., Zhang, X., Ren, S. and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034 (2015).
- [4] Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256 (2010).
- [5] Amari, S.-i., Ozeki, T., Karakida, R., Yoshida, Y. and Okada, M.: Dynamics of Learning in MLP: Natural Gradient and Singularity Revisited, *Neural computation*, Vol. 30, No. 1, pp. 1–33 (2018).
- [6] Shouno, H., Suzuki, S. and Kido, S.: A transfer learning

表 2 実験結果 (値は Accuracy[%])

タスク フィルタ混合比	MNIST			CIFAR-10			CIFAR-100		
	$\delta = 0.1$			$\delta = 0.05$			$\delta = 0.05$		
$r$	MRSA	提案手法	$p$ 値	MRSA	提案手法	$p$ 値	MRSA	提案手法	$p$ 値
1%	95.25	95.04	0.263	27.25	31.94	0.06	6.662	7.102	0.139
5%	97.82	98.07	>0.001	32.57	40.69	>0.05	10.98	11.76	0.06
10%	98.49	98.65	>0.05	37.49	42.88	>0.01	12.15	13.17	0.08

method with deep convolutional neural network for diffuse lung disease classification, *International Conference on Neural Information Processing*, pp. 199–207 (2015).

- [7] Suzuki, A., Suzuki, S., Kido, S. and Shouno, H.: A 2-staged Transfer Learning Method with Deep Convolutional Neural Network for Diffuse Lung Disease Analysis, *Proceedings of the 2017 International Forum on Medical Imaging in Asia*, pp. 95–99 (2017).
- [8] 鈴木藍雅, 坂無英徳, 木戸尚治, 庄野 逸: 2 段階転移学習を用いた深層畳み込みニューラルネットによるびまん性肺疾患の識別と特徴表現の解析, 第 117 回情報処理学会数理モデル化と問題解決 (MPS) 研究会 研究報告, No. 13, 情報処理学会, pp. 1–6 (2018).
- [9] Fukushima, K.: Training multi-layered neural network neocognitron, *Neural Networks*, Vol. 40, pp. 18–31 (2013).
- [10] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324 (1998).
- [11] 和家伸明, 福島邦彦: ネオコグニトロンの新しい学習法, 電子情報通信学会論文誌 D, Vol. 75, pp. 1892–1899 (1992).
- [12] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).
- [13] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database, *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE, pp. 248–255 (2009).