Vol.2021-MPS-136 No.1
2021/12/13

CNN 初期化のためのプーリング層のモデリング

邉見 貴彦[1]　加藤 毅[1,2]

概要：画像処理タスクにおいて大きな成功を収めている畳み込みニューラルネットワーク (CNN) は，畳み込み層やプーリング層など，さまざまな種類の層を組み合わせて構成される深層ニューラルネットワークの一種である．CNN の学習の安定性は，重みの初期値に依存することが知られている．現在，初期化で標準的に用いられているのは，He らによって開発された Kaiming 法である．この手法は CNN の構造を単純化して導出されており，特にプーリング演算は完全に無視されていた．本研究では，プーリング演算を含む CNN の構造を改めて定式化しなおし，このモデルに基づいて初期化法を導出した．また，この新しい初期化法の性能を，従来の初期化法と比較して検証した．ソースコード: https://github.com/hecwitane/ASV-pub/

キーワード：畳み込みニューラルネットワーク, 初期化法, プーリング, 深層ニューラルネットワーク, 確率的ネットワーク

1. はじめに

深層畳み込みニューラルネットワーク (CNN) は，画像の分類 [1], [2], [3]，物体の検出 [4], [5]，セグメンテーション [6]，画像検索 [7] など，画像認識の分野で数多く利用されており，非常に高い評価を得ている．

深層 CNN は，層を深くしてモデルの容量を増やすことによりその性能を向上させてきたが，一方で学習における数値的安定性は現在も主要な技術的課題の一つである．学習を安定させる手法も従来から研究されてきた．双曲線正接関数やシグモイド関数と比較して，ReLU 活性化関数を使用することで，**勾配消失問題**を回避できる可能性がある [8], [9]．また，確率的勾配降下法は，通常の勾配降下法と比較して，その確率的なふるまいにより悪い局所解や鞍点を回避することができる．しかし，安定した収束を実現するには，この 2 つの手法では十分ではない．

深層 CNN の最適化における目的関数は非凸性が高いため，重みやバイアスなどのモデルパラメータの初期化は，最適化の安定性に影響を与えることが知られている．Glorot らは，ネットワークの構造に適応的な初期値の設定方法である **Xavier 法**を提案した [10]．この手法では，ネットワークのパラメータを確率変数とみなし，確率的ネットワークとしてネットワーク中の値を解析することで，初期化方法を導出した．しかしながら，彼らが開発したモデルでは，すべての層の活性化関数に双曲線正接関数が使われていると仮定しており，現在使われている CNN 構造において主流となっている ReLU 関数やその変種を用いる状況とは理論的に齟齬が生じている．現在最も標準的な初期化法として用いられているのが，He らの開発した **Kaiming 法** [11] である．彼らは活性化関数として ReLU 関数を採用し，Glorot らの解析と同様の方法で確率的ネットワークを解析した．Kaiming 法により，Xavier 法では最適化できないより深い構造を安定して学習できるようになった．ただし，Kaiming 法は，より進歩した現在の CNN を一部単純化して導出している．具体的には，彼らのモデルではプーリング演算，パディング，ストライドといった重要な構成要素が理論的には無視されてしまっていた．

本研究では，従来法では無視されていたプーリング演算等の構成要素を考慮した，新しい CNN の初期化法を提案する [12]．提案法は，こうした従来無視されていた構造を取り入れて表現するため，新たに CNN を定式化し直し，従来よりも精密なモデルに基づいて導出しており，プーリング層，パディング，ストライドなど現在使われている構成要素を理論的にサポートするものである (表 1 を参照)．また，従来法と比較して，提案法がどのように機能するかを実データを用いた数値実験により調査した．

なお，理論の証明の詳細と付加的な実験結果については文献 [13] を参照されたい．

2. CNN の定式化

本稿で議論する CNN を明確にするために，本節では

[1] 群馬大学 理工学府, 〒 376–8515 群馬県桐生市天神町 1 丁目 5–1
[2] 群馬大学 次世代モビリティ社会実装研究センター, 〒 371–8510 群馬県前橋市荒牧町 4–2

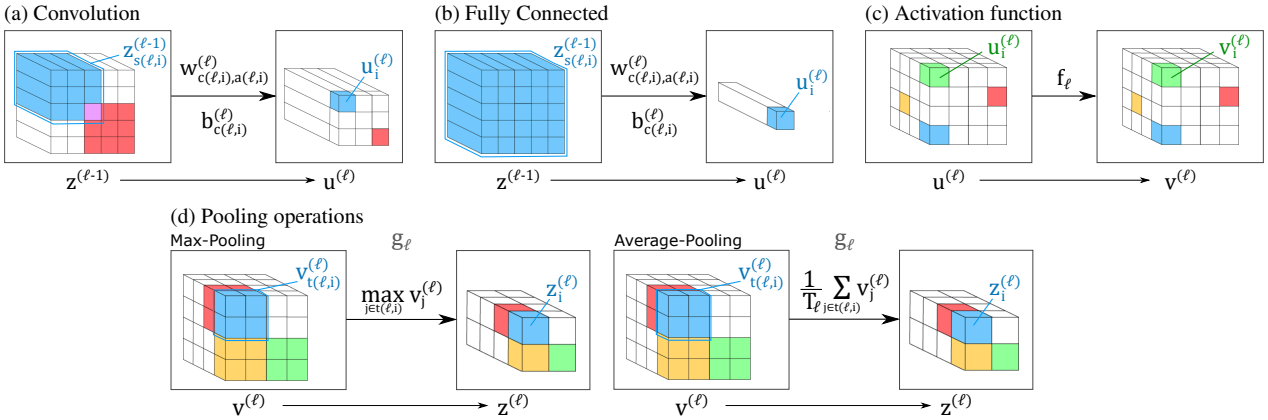


図 1: CNN の定式化．(a) は第 ℓ 層において畳み込み出力の要素 $u_i^{(\ell)}$ を計算する際の例を示している．要素 $u_i^{(\ell)}$ は出力特徴マップのテンソルにおいて $c(i,\ell)$ チャネルに属し，添え字集合 $\boldsymbol{a}(\ell,i), \boldsymbol{s}(\ell,i)$ によって選択された重み $\boldsymbol{w}_{c(i,\ell)}^{(\ell)}$ および $\boldsymbol{z}^{(\ell-1)}$ の部分ベクトルの積和として計算される．(b) は全結合層の計算を表している．全結合層は畳み込み層の特殊な場合とみなせ，$\boldsymbol{s}(\ell,i) = [M_{\ell-1}]$ とした状況に相当する．(c) は活性化関数 f_ℓ の適用を表している．活性化関数は入力に対して要素ごとに非線形変換を行う関数である．(d) はプーリングを表している．プーリングの出力の要素は，添え字集合 $\boldsymbol{t}(\ell,i)$ によって選択される入力 $\boldsymbol{v}^{(\ell)} \in \mathbb{R}^{M'_\ell}$ の部分ベクトルを，ある種の関数 g_ℓ によってスカラーに変換することで計算される．

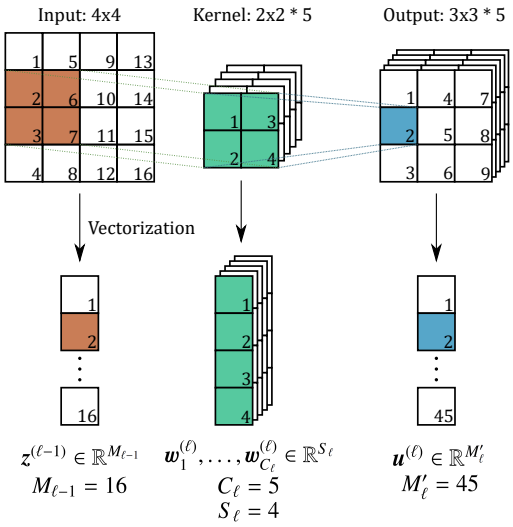


図 2: 畳み込み演算の定式化の具体例．詳細は本文中で説明する．

表 1: 従来の初期化法との比較．FC と Conv は，それぞれ全結合層と畳み込み層を表す．従来法のモデルでは，CNN の構成要素の一部を理論的に扱っていなかった．プーリング演算はいずれの従来法でも無視されているほか，Kaiming 法においては，畳み込み層で用いられるパディング等が考慮されていない．本研究のモデルは，こうした CNN の構成要素を含むよう改めて定式化を行うことで，CNN をより正確に表現している．

Methods	FC	ReLU	Conv	Pooling
Xavier	✓			
Kaiming	✓	✓	✓	
提案法	✓	✓	✓	✓

表 2: 順伝播の議論で用いられる変数．

シンボル	説明
$C_\ell \in \mathbb{N}$	チャネル数．
$S_\ell \in \mathbb{N}$	受容野のサイズ．
$c(\ell,i) \in [C_\ell]$	出力の第 i ユニットを含むチャネル．
$\boldsymbol{a}(\ell,i) \subseteq [S_\ell]$	畳み込みカーネルの要素を選択する添え字集合．
$\boldsymbol{s}(\ell,i) \subseteq [M_{\ell-1}]$	畳み込みにおいて入力の要素を選択する添え字集合．$\lvert\boldsymbol{s}(\ell,i)\rvert = \lvert\boldsymbol{a}(\ell,i)\rvert$ を満たす．
$M'_\ell \in \mathbb{N}$	畳み込み後のユニット数．
$f_\ell : \mathbb{R} \to \mathbb{R}$	活性化関数．
$T_\ell \in \mathbb{N}$	プーリング領域のサイズ．
$g_\ell : \mathbb{R}^{T_\ell} \to \mathbb{R}$	プーリング関数．
$\boldsymbol{t}(\ell,i) \subseteq [M'_\ell]$	プーリング関数が適用される領域を選択する添え字集合．$\lvert\boldsymbol{t}(\ell,i)\rvert = T_\ell$ を満たす．
$M_\ell \in \mathbb{N}$	第 ℓ 層の出力ユニット数．

表 3: 逆伝播の議論で用いられる変数．

シンボル	説明
$J_\ell \in \mathbb{N}$	逆伝播における第 ℓ 層の重みの要素数．
$\tilde{\boldsymbol{w}}_c^{(\ell)} \in \mathbb{R}^{J_\ell}$	逆伝播における第 ℓ 層の重み．
$\tilde{C}_\ell \in \mathbb{N}$	逆伝播における第 ℓ 層の重みベクトルの個数．
$\tilde{c}(\ell,i) \in [\tilde{C}_\ell]$	逆伝播で重みを選択する添え字．
$\boldsymbol{h}(\ell,i) \subseteq [J_\ell]$	逆伝播で参照する重みの要素を選択する添え字集合．
$\boldsymbol{j}(\ell,i) \subseteq [M'_\ell]$	逆伝播で参照する勾配の要素を選択する添え字集合．$\lvert\boldsymbol{j}(\ell,i)\rvert = \lvert\boldsymbol{h}(\ell,i)\rvert$ を満たす．
$d(\ell,i) \in [M_\ell]$	入力の第 i 要素に依存関係をもつプーリング出力の要素の添え字．

(a) 順伝播信号

$$\begin{aligned} \boldsymbol{z}^{(0)} \in \mathbb{R}^{M_0} \overset{\text{Conv}}{\mapsto} \boldsymbol{u}^{(1)} \in \mathbb{R}^{M'_1} \overset{\text{Act}}{\mapsto} \boldsymbol{v}^{(1)} \in \mathbb{R}^{M'_1} \overset{\text{Pool}}{\mapsto} \boldsymbol{z}^{(1)} \in \mathbb{R}^{M_1} \overset{\text{Conv}}{\mapsto} \cdots \\ \overset{\text{Conv}}{\mapsto} \boldsymbol{u}^{(L-1)} \in \mathbb{R}^{M'_{L-1}} \overset{\text{Act}}{\mapsto} \boldsymbol{v}^{(L-1)} \in \mathbb{R}^{M'_{L-1}} \overset{\text{Pool}}{\mapsto} \boldsymbol{z}^{(L-1)} \in \mathbb{R}^{M_{L-1}} \overset{\text{Conv}}{\mapsto} \boldsymbol{u}^{(L)} \in \mathbb{R}^{M'_L}. \end{aligned} \tag{1}$$

(b) 逆伝播信号

$$\begin{aligned} \Delta\boldsymbol{u}^{(L)} \in \mathbb{R}^{M'_L} \mapsto \Delta\boldsymbol{z}^{(L-1)} \in \mathbb{R}^{M_{L-1}} \mapsto \Delta\boldsymbol{v}^{(L-1)} \in \mathbb{R}^{M'_{L-1}} \mapsto \dots \\ \mapsto \Delta\boldsymbol{z}^{(1)} \in \mathbb{R}^{M_1} \mapsto \Delta\boldsymbol{v}^{(1)} \in \mathbb{R}^{M'_1} \mapsto \Delta\boldsymbol{u}^{(1)} \in \mathbb{R}^{M'_1}. \end{aligned} \tag{2}$$

図 3: 順伝播信号と逆伝播信号の計算過程．$\boldsymbol{z}^{(0)}$ は画像などのネットワークに対する入力である．例えば，RGB 画像を入力する場合，入力信号の要素数 M_0 は画素数 ×3 として与えられる．演算子 $\overset{\text{Conv}}{\mapsto}$, $\overset{\text{Act}}{\mapsto}$, $\overset{\text{Pool}}{\mapsto}$ は，それぞれ，畳み込み演算，活性化関数，プーリング演算から生じる写像を表す．多クラス分類の場合，出力信号 $\boldsymbol{z}^{(L)}$ の要素数 M_L はクラス数に一致する．

CNN 構造を改めて定式化する．通常 CNN は，畳み込み層，活性化関数，プーリング層，全結合層など，多くの異なるタイプの構成要素を積み重ねたものとして表現される．本研究では，これらを統一的に扱うため，畳み込み，活性化関数，プーリングの演算を 1 つにまとめた **Conv+Pool 層**を導入する．後述するように，Conv+Pool 層は，プーリング演算を行わない畳み込み，大域的平均プーリング，全結合層をも表すことができる．

CNN を画像処理タスクに適用する場合，通常 2 次元の畳み込みを行う．音声認識の場合，畳み込み演算の次元数は 1 となる [14]．一般に CNN を通過する入力信号と中間信号はテンソルの形で表現される．信号を表すテンソルの階数は適用対象によって異なる．表記を簡略化し，テンソルの設計に関わらず定式化を統一するため，本定式化ではすべての信号をベクトル化して表現する．

2.1 順伝播の場合

CNN を構成する Conv+Pool 層の数を L とする．第 ℓ 層のモデルパラメータは重み $\boldsymbol{W}^{(\ell)} := \left[\boldsymbol{w}_1^{(\ell)}, \dots, \boldsymbol{w}_{C_\ell}^{(\ell)}\right]^\top \in \mathbb{R}^{C_\ell \times S_\ell}$ とバイアス $\boldsymbol{b}^{(\ell)} := \left[b_1^{(\ell)}, \dots, b_{C_\ell}^{(\ell)}\right]^\top \in \mathbb{R}^{C_\ell}$ である．Conv+Pool 層は，この層に入力された信号 $\boldsymbol{z}^{(\ell-1)} \in \mathbb{R}^{M_{\ell-1}}$ を変換し，次の層に伝える（図 1 を参照）．

全体の入力信号 $\boldsymbol{z}^{(0)}$ は第 1 層の入力となり，最終層の出力 $\boldsymbol{z}^{(L)} \in \mathbb{R}^{M'_L}$ は図 3a のように計算される．表記の便宜上，最終層に関して $\boldsymbol{v}^{(L)} := \boldsymbol{u}^{(L)}$ および $\boldsymbol{z}^{(L)} := \boldsymbol{v}^{(L)}$ と定義しておく．式 (3),(4),(5) はそれぞれ，畳み込み，活性化，プーリングの演算に対応する:

$$u_i^{(\ell)} = \left\langle \boldsymbol{w}_{c(\ell,i),\boldsymbol{a}(\ell,i)}^{(\ell)}, \boldsymbol{z}_{\boldsymbol{s}(\ell,i)}^{(\ell-1)} \right\rangle + b_{c(\ell,i)}^{(\ell)}, \tag{3}$$

$$v_i^{(\ell)} = f_\ell\left(u_i^{(\ell)}\right), \tag{4}$$

$$z_i^{(\ell)} = g_\ell\left(\boldsymbol{v}_{\boldsymbol{t}(\ell,i)}^{(\ell)}\right). \tag{5}$$

ここで用いた各変数は表 2 に定義した．また，$\boldsymbol{w}_{c(\ell,i),\boldsymbol{a}(\ell,i)}^{(\ell)}$ は添え字集合 $\boldsymbol{a}(\ell,i)$ によって選択された要素からなる $\boldsymbol{w}_{c(\ell,i)}^{(\ell)} \in \mathbb{R}^{S_\ell}$ の部分ベクトルを表す．同様に，$\boldsymbol{z}_{\boldsymbol{s}(\ell,i)}^{(\ell-1)}$ は添え字集合 $\boldsymbol{s}(\ell,i)$ によって選択された要素からなる $\boldsymbol{z}^{(\ell-1)}$ の部分ベクトルを表す．

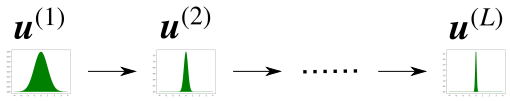
例: 図 2 に小さな畳み込み層を使って定式化 (3) の例を示す．図に示す例は，以下のようにおくことで，定式化 (3) の枠組みに入る．

- 第 $(\ell-1)$ 層のテンソルの大きさは，$4\times4\times1$ である．すなわち，第 $(\ell-1)$ 層には，1 チャネルの 4×4 の特徴マップが存在する．この特徴マップを表すテンソルをベクトル化したものが $\boldsymbol{z}^{(\ell-1)}$ である．$\boldsymbol{z}^{(\ell-1)}$ の要素数は $M_{\ell-1} = 4\cdot4\cdot1 = 16$ であり，これが第 ℓ 層の入力となる．
- 第 ℓ 層のテンソルの大きさは，$3\times3\times5$ である．すなわち，第 ℓ 層には，3×3 の特徴マップが $C_\ell = 5$ チャネル分存在する．この特徴マップを表すテンソルをベクトル化したものが $\boldsymbol{u}^{(\ell)}$ である．$\boldsymbol{u}^{(\ell)}$ の要素数は $M'_\ell = 3\cdot3\cdot5 = 45$ である．
- $C_\ell = 5$ 個の出力チャネルにそれぞれ 2×2 の畳み込みカーネルを用意する．本研究の定式化では，カーネルそれぞれをベクトル化したものを $\boldsymbol{w}_1^{(\ell)}, \dots, \boldsymbol{w}_5^{(\ell)}$ としている．これらのベクトルの要素数はそれぞれ $S_\ell = 2\cdot2 = 4$ である．
- 畳み込み出力の要素 $u_2^{(\ell)}$ を計算する場合（すなわち，式 (3) において $i = 2$ とするとき），参照されるベクトルの要素を選択する添え字および添え字集合は $c(\ell,i) = 1, \boldsymbol{a}(\ell,i) = \{1,2,3,4\}, \boldsymbol{s}(\ell,i) = \{2,3,6,7\}$ のように与えられる．（例はここまで）

畳み込み演算，活性化関数，プーリング演算それぞれの詳細については文献 [13] の付録 A を参照されたい．以下ではその概要を示す．

- **畳み込み:** 要素 $u_i^{(\ell)}$ は出力特徴マップのテンソルにおいて $c(i,\ell)$ チャネルに属し，添え字集合 $\boldsymbol{a}(\ell,i), \boldsymbol{s}(\ell,i)$ によって選択された重み $\boldsymbol{w}_{c(i,\ell)}^{(\ell)}$ および $\boldsymbol{z}^{(\ell-1)}$ の部分ベクトルの積和として計算される．ほとんどの計算において，添え字集合 $\boldsymbol{a}(\ell,i)$ は $[S_\ell]$ に等しいが，パディングがある場合には $\boldsymbol{a}(\ell,i) \subsetneq [S_\ell]$ となるケースが現れる．一般に，畳み込みのオプションはこれら添え字集合を適切に設定することによって表現可能である（図 1a を参照）．
- **活性化関数:** 活性化関数 $f_\ell : \mathbb{R} \to \mathbb{R}$ として典型的には ReLU 関数が知られている．f_ℓ を恒等関数とおくことにより，活性化関数を適用しないような層も表現する

(a) 分散が減衰し，学習が停滞.



(b) 分散を等しく調整; $\mathrm{Var}\left[u^{(1)}\right] = \cdots = \mathrm{Var}\left[u^{(L)}\right]$.

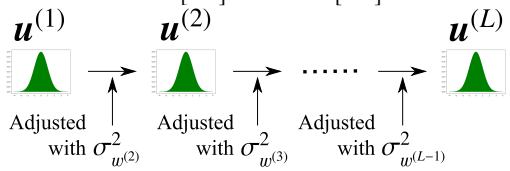


図 4: 順伝播・逆伝播において，分散の減衰または爆発が起こる可能性がある．(a) では，層が深くなるにつれて分散が減衰していく例を示した．提案法では，(b) に示すように，初期化に用いる乱数の分散パラメータ $\sigma^2_{w^{(\ell)}}$ を調整し，信号の分散を等しくする．

ことができる（図 1c を参照）.

- **プーリング:** 活性化関数によって非線形変換された信号 $\boldsymbol{v}^{(\ell)} \in \mathbb{R}^{M'_\ell}$ から，プーリング関数 g_ℓ によって一部領域の最大値もしくは平均値が計算される．最大値もしくは平均値として集約される領域は添え字集合 $\boldsymbol{t}(\ell,i) \subset [M'_\ell]$ によって選択する（図 1d を参照）.

全結合層は Conv+Pool 層の特殊なケースである．すなわち，$\boldsymbol{a}(\ell,i) := [S_\ell], \boldsymbol{s}(\ell,i) := [M_{\ell-1}]$ とし，f_ℓ, g_ℓ を恒等関数としたときに相当する（図 1b を参照）.

2.2 逆伝播の場合

CNN の学習は**誤差関数** E を重みとバイアス $\boldsymbol{\theta} = (\boldsymbol{W}^{(\ell)}, \boldsymbol{b}^{(\ell)})_{\ell\in[L]}$ に関して最小化することで行われる．勾配降下法をニューラルネットワークに適用したものは**誤差逆伝播法**と呼ばれ，信号 $\Delta z_i^{(\ell)}$ を順伝播のときとは逆方向に伝達し計算する．次の勾配を考える:

$$\Delta \boldsymbol{z}^{(\ell-1)} = \frac{\partial E}{\partial \boldsymbol{z}^{(\ell-1)}}, \ \Delta \boldsymbol{u}^{(\ell)} = \frac{\partial E}{\partial \boldsymbol{u}^{(\ell)}}, \ \Delta \boldsymbol{v}^{(\ell)} = \frac{\partial E}{\partial \boldsymbol{v}^{(\ell)}}. \tag{6}$$

すると，逆伝播信号は再帰的に次のように表現できる:

$$\Delta z_i^{(\ell-1)} = \left\langle \tilde{\boldsymbol{w}}^{(\ell)}_{\tilde{c}(\ell,i),\boldsymbol{h}(\ell,i)}, \Delta \boldsymbol{u}^{(\ell)}_{\boldsymbol{j}(\ell,i)} \right\rangle, \tag{7}$$

$$\Delta u_i^{(\ell)} = \Delta v_i^{(\ell)} \frac{\partial v_i^{(\ell)}}{\partial u_i^{(\ell)}}, \tag{8}$$

$$\Delta v_i^{(\ell)} = \Delta z^{(\ell)}_{d(\ell,i)} \frac{\partial z^{(\ell)}_{d(\ell,i)}}{\partial v_i^{(\ell)}}. \tag{9}$$

ここで用いた各変数は表 3 に定義した．式 (7),(8),(9) の導出の詳細は，文献 [13] の付録 B を参照されたい.

3. 提案する初期化法

この節では，重みとバイアス $\boldsymbol{\theta} = (\boldsymbol{W}^{(\ell)}, \boldsymbol{b}^{(\ell)})_{\ell\in[L]}$ の初期値を与えるための 2 種類の方法を提案する．本研究でも，Xavier 法および Kaiming 法のアプローチに従い，重みおよびバイアスの各要素の初期値は，平均 0 の正規乱数で生成されると仮定する:

$$w_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma^2_{w^{(\ell)}}), \quad b_i^{(\ell)} \sim \mathcal{N}(0, \sigma^2_{b^{(\ell)}}). \tag{10}$$

同じ層の中では，重みおよびバイアスのそれぞれ各要素に関して，分布の分散パラメータ $\sigma^2_{w^{(\ell)}}, \sigma^2_{b^{(\ell)}}$ が共通であることに注意されたい.

3.1 ネットワーク中の信号の分散

式 (10) の仮定の下で，順伝播信号と逆伝播信号の分散を解析する．従来法では，勾配の消失や爆発を回避するため，順伝播または逆伝播の信号の分散がほぼ等しくなるように，2 つの初期化パラメータ $\sigma^2_{w^{(\ell)}}$ と $\sigma^2_{b^{(\ell)}}$ を決定する（図 4 を参照）．従来法と比較して，提案法では，こうした信号の分散を解析するにあたって，より正確な CNN の定式化を採用している．以下では，信号分散の再帰的表現を導出し，これに基づいて新たな初期化手法を提案する．より詳細な導出は文献 [13] を参照.

順伝播信号の解析: 第 $(\ell-1)$ 層の信号 $\boldsymbol{u}^{(\ell-1)}$ の各要素が統計的に独立であり，同一の正規分布 $\mathcal{N}(0, q^{(\ell-1)})$ に従うと仮定する．第 ℓ 層について，$u_i^{(\ell)}$ の分散を $q_i^{(\ell)}$ とおく．ここで，第 ℓ 層における分散 $q^{(\ell)}$ を $q_1^{(\ell)}, \ldots, q_{M'_\ell}^{(\ell)}$ の平均で近似することを考えると，次式が成り立つ:

$$q^{(\ell)} = \sigma^2_{b^{(\ell)}} + \sigma^2_{w^{(\ell)}} q^{(\ell-1)} \tau_{\ell-1} \frac{1}{M'_\ell} \varepsilon_\ell. \tag{11}$$

ここで，定数 τ_ℓ は，g_ℓ が最大プーリングを表す場合，次のようにおく:

$$\tau_\ell := T_\ell \int_0^\infty s^2 \phi(s) \Phi(s)^{T_\ell - 1} ds. \tag{12}$$

g_ℓ が平均プーリングを表す場合には，次のようにおく:

$$\tau_\ell := \frac{1}{2T_\ell} \left(1 + \frac{T_\ell - 1}{\pi}\right). \tag{13}$$

また，T_ℓ はプーリング領域の大きさ，ϕ, Φ はそれぞれ標準正規分布の密度関数および累積分布関数，$\varepsilon_\ell := \sum_{i=1}^{M'_\ell} |\boldsymbol{s}(\ell,i)|$ は第 $(\ell-1)$ 層と第 ℓ 層の間の接続の総数を表す．以上の議論を深い層へ向かって再帰的に適用すれば，式 (11) の関係が各層に関して帰納的に成り立つ.

逆伝播信号の解析: 順伝播の場合とほぼ同様の議論が成り立つ．第 ℓ 層において，$\Delta \boldsymbol{z}^{(\ell)}$ の各要素が統計的に独立であり，同一の正規分布 $\mathcal{N}(0, r^{(\ell)})$ に従うと仮定する．さらに，$\Delta z^{(\ell)}_{d(\ell,j)}$ と $\partial z^{(\ell)}_{d(\ell,j)} / \partial u_j^{(\ell)}$ の統計的独立性を仮定する．第 $(\ell-1)$ 層について，$\Delta z_i^{(\ell-1)}$ の分散を $r_i^{(\ell-1)}$ とおく．第 $(\ell-1)$ 層における分散 $r^{(\ell-1)}$ を $r_1^{(\ell-1)}, \ldots, r_{M_{\ell-1}}^{(\ell-1)}$ の平均で近似することを考えると，次式が成り立つ:

$$r^{(\ell-1)} = \sigma^2_{w^{(\ell)}} r^{(\ell)} \gamma_\ell \frac{1}{M_{\ell-1}} \varepsilon_\ell. \tag{14}$$

ここで，定数 γ_ℓ は，g_ℓ が最大プーリングを表す場合，次のようにおく:

$$\gamma_\ell := \frac{2^{T_\ell}-1}{T_\ell 2^{T_\ell}}. \tag{15}$$

g_ℓ が平均プーリングを表す場合には，次のようにおく:

$$\gamma_\ell := \frac{1}{2T_\ell^2}. \tag{16}$$

3.2 提案法: ASV Forward/Backward Method

以上を踏まえて，信号の分散を維持する初期化法を導出する．ここでは，バイアスの初期値を $\boldsymbol{b}^{(\ell)} = \mathbf{0}$ とし（これは $\sigma^2_{b^{(\ell)}} = 0$ とすることに相当する），さらに $q^{(0)} = r^{(L)} = 1$ とする．これらの条件から以下の 2 つの方法が得られる．

順伝播信号の分散を維持する初期化法 (ASV Forward Method):

$$\sigma^2_{w^{(\ell)}} = \frac{M'_\ell}{\tau_{\ell-1}\varepsilon_\ell}. \tag{17}$$

逆伝播信号の分散を維持する初期化法 (ASV Backward Method):

$$\sigma^2_{w^{(\ell)}} = \frac{M_{\ell-1}}{\gamma_\ell\varepsilon_\ell}. \tag{18}$$

実際にそれぞれ，式 (11),(14) に代入すると，すべての層で $q^{(\ell)} = 1$，$r^{(\ell)} = 1$ が成り立つことが確かめられる．

表 4: 実験に用いた CNN の構造．この CNN では，最大プーリングが第 1 層において，大域的平均プーリングが特徴抽出部の最終層において利用されている．また，特徴マップの解像度を縮小するために，一部の畳み込み演算では 1 より大きいストライドが設定されている．

Component `InputBlock`(c)
Conv(7x7,channels=c,padding=3,stride=2), MaxPool(3x3,padding=1,stride=2)

Component `ConvBlock`(c, s = 1)
Conv(3x3,channels=c,padding=1,stride=s), Conv(3x3,channels=c,padding=1,stride=1)

ℓ-th Layer	Output Shape	34-layer Architecture
	(3,224,224)	Input Image
1	(64,112,112)	`InputBlock`(c = 64)
2–7	(64,56,56)	`ConvBlock`(c = 64)×3
8–15	(128,28,28)	`ConvBlock`(c = 128, s = 2) `ConvBlock`(c = 128)×3
16–27	(256,14,14)	`ConvBlock`(c = 256, s = 2) `ConvBlock`(c = 256)×5
28–33	(512,7,7) (512,1,1)	`ConvBlock`(c = 512, s = 2) `ConvBlock`(c = 512)×2 Global Average Pooling
34	10	Linear
	2.11×10^7	Number of Parameters

表 5: 初期化法ごとの検証用データに対する正解率の比較．表の数値は 1000 反復の学習中に得られた最高値を表している．太字の数字は，表の中で最も高い値であることを表している．

(a) Car

Xavier	Kaiming (forward)	Kaiming (backward)	ASV (forward)	ASV (backward)
71.95	70.52	73.10	72.74	**81.49**

(b) Food

Xavier	Kaiming (forward)	Kaiming (backward)	ASV (forward)	ASV (backward)
72.83	75.72	69.75	76.49	**78.81**

(c) Fungi

Xavier	Kaiming (forward)	Kaiming (backward)	ASV (forward)	ASV (backward)
65.23	68.16	66.99	67.97	**69.73**

4. 実験

初期化法の違いが，層の深い CNN の学習にどのような影響を与えるのか調査するため，実データを用いた数値実験を行った．

実験は公開データセットを用いて用意した Car, Food, Fungi により行った．これらのデータセットはそれぞれ 10 クラスのカテゴリを持ち，CNN はクラス数 10 の多クラス分類問題を学習するよう設計した．データはすべて 3 チャネルのカラー画像として与えられている．3 種類のデータセットに対して同じ構造の CNN で学習できるようにするため，画像を正方形に切り抜き，224×224 のサイズとなるようにリサイズした上で，正規化を施した．データ拡張は用いなかった．1 つ目のデータセット Car は VMMRdb [15] から得た．全体の 75%のデータを訓練用データとして無作為に選択し，残りのデータを検証用データとして用いた．2 つ目のデータセット Food は iFood 2019 [16] のサブセットである．全体の 90%のデータを訓練用データとして選び，残りの 10%を検証用データとした．3 つ目のデータセット Fungi は iNaturalist 2019 [17] に含まれるスーパーカテゴリ Fungi を用いて構成した．データ全体を 75%と 25%に排他的に分割し，それぞれ訓練用および検証用のデータとして用いた．

実験に用いた CNN の構造を表 4 に示す．これは前述の 3 つのデータセットの学習それぞれにおいて共通して用いた構造である．ネットワークはパディングやストライドを持つ畳み込み，最大プーリング，大域的平均プーリング，全結合層を含む，全 34 層で構成した．活性化関数にはすべて ReLU 関数を用いた．多クラス分類問題を学習するため，損失関数はクロスエントロピー損失を用いた．最適化アルゴリズムとして Adam を用いた．ミニバッチサイズは 64 とした．学習率は 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3} をそれぞれ検証した．このような条件を共通に用いて，初期化法ごとに結果

を比較した．用いた初期化法は，提案する 2 つの手法 **ASV forward 法**および **ASV backward 法**に加えて，**Kaiming 法**（**順伝播の条件**，および**逆伝播の条件**に基づく 2 種類がある），そして **Xavier 法**の，合計 5 つの手法である．性能評価は，検証用データセットに対する正解率を指標とした．

表 5a には，データセット Car を用いたときの結果を示している．それぞれの値は，異なる学習率ごとに学習を行った結果の中で，最も高い正解率である．それぞれの学習率に対する個別の結果は，文献 [13] の表 5a を参照されたい．まず，2 つの提案法 ASV forward 法と ASV backward 法について比較すると，表 5b および表 5c にあるように，今回用いた 3 つのデータセットいずれに関する結果においても ASV backward 法のほうが高い性能を得た．ASV backward 法のほうが性能が高くなった原因は，逆伝播信号の分散を維持する条件から導出されているので，ASV forward 法に比べて，より直接的に勾配消失や爆発を抑制するアプローチとなっているためであると考えられる．他方で，Kaiming 法における順伝播，逆伝播の 2 種類の方法の間には，逆伝播が順伝播より性能が高くなるという現象は確認できなかった．これは，Kaiming 法が CNN を過度に単純化したモデルに基づいて導出されたためであると考えられる．ASV forward 法と ASV backward 法の顕著な違いは，プーリングによって特徴マップの解像度が変化したときに生じる．ASV backward 法では，プーリング演算を含む層において，ASV forward 法よりも初期値の分散を大きく与えている（文献 [13] の図 3 を参照）．一方，Kaiming 法のモデルでは，プーリング演算を無視していた．提案法と従来法の性能に差が生じたのは，Kaiming 法ではプーリング演算に起因する信号の分散の変化を無視していたためと考えられる．

5. まとめ

本研究では，CNN の構造を改めて定式化することにより，CNN のための新たな初期化法を提案した．実験により，提案法は学習の安定性に寄与し認識性能を向上させることを確認した．

謝辞

本研究の一部は，（独）環境再生保全機構の環境研究総合推進費（JPMEERF20205006）により実施された．また，JSPS 科研費 19K04661 の助成を受けた．

参考文献

[1] Kleinberg, B., Li, Y. and Yuan, Y.: An Alternative View: When Does SGD Escape Local Minima?, *Proceedings of the 35th International Conference on Machine Learning* (Dy, J. and Krause, A., eds.), Proceedings of Machine Learning Research, Vol. 80, Stockholmsmässan, Stockholm Sweden, PMLR, pp. 2698–2707 (2018).

[2] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going deeper with convolutions, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE (2015).

[3] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Bengio, Y. and LeCun, Y., eds.) (2015).

[4] Girshick, R.: Fast R-CNN, *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE (2015).

[5] Ito, E., Sato, T., Sano, D., Utagawa, E. and Kato, T.: Virus Particle Detection by Convolutional Neural Network in Transmission Electron Microscopy Images, *Food Environ Virol*, Vol. 10, No. 2, pp. 201–208 (2018).

[6] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE (2016).

[7] Ding, S., Lin, L., Wang, G. and Chao, H.: Deep feature learning with relative distance comparison for person re-identification, *Pattern Recognition*, Vol. 48, No. 10, pp. 2993–3003 (2015).

[8] Hochreiter, S.: The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 06, No. 02, pp. 107–116 (1998).

[9] Nair, V. and Hinton, G. E.: Rectified Linear Units Improve Restricted Boltzmann Machines., *ICML* (Fürnkranz, J. and Joachims, T., eds.), Omnipress, pp. 807–814 (2010).

[10] 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*, PMLR, pp. 249–256 (2010).

[11] He, K., Zhang, X., Ren, S. and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034 (2015).

[12] Henmi, T., Zara, E. R. R., Hirohashi, Y. and Kato, T.: Adaptive Signal Variances: CNN Initialization Through Modern Architectures, *2021 IEEE International Conference on Image Processing* (2021).

[13] Henmi, T., Zara, E. R. R., Hirohashi, Y. and Kato, T.: Adaptive Signal Variances: CNN Initialization Through Modern Architectures (2020). https://arxiv.org/abs/2008.06885.

[14] Kalchbrenner, N., Grefenstette, E. and Blunsom, P.: A Convolutional Neural Network for Modelling Sentences, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics (2014).

[15] Tafazzoli, F., Frigui, H. and Nishiyama, K.: A Large and Diverse Dataset for Improved Vehicle Make and Model Recognition, *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 874–881 (2017).

[16] Kaur, P., , Sikka, K., Wang, W., Belongie, s. and Divakaran, A.: FoodX-251: A Dataset for Fine-grained Food Classification (2019).

[17] Ueda, K.: iNaturalist Research-grade Observations (2020).