2R-08

CNN max-pooling 層の初期化法
A new scheme for initialization of CNN with max-pooling layer

佐藤 貴亮 * 廣橋 義寛 † 太田 直哉 * ‡ 加藤 毅 * ‡ §
Takaaki Sato Yoshihiro Hirohashi Naoya Ohta Tsuyoshi Kato

1. はじめに

近年の画像認識の分野においては，**畳み込みニューラルネットワーク（CNN, Convolutional Neural Network）**が広く用いられ，画像認識のコンペティションである ILSVRC において高い成績を収めており，年々深層化が進んでいる．CNN の学習は**誤差逆伝播法**によって行われるが，深層 CNN で誤差逆伝播法を用いる上で，誤差の勾配が消失，発散することによって学習が停滞することがしばしば問題視されてきた．

本論文では，勾配消失，勾配発散を防ぐため，全結合層，畳み込み層，max-pooling 層から構成される CNN に対して，活性化関数に ReLU を用いたときの適切なモデルパラメータの初期化法を提案する．本手法により，max-pooling を含む深層 CNN において，既存法ではほぼ学習不可能であった深層構造でも学習が可能になったことを数値実験により示す．また，畳み込みの逆伝播について，既存法よりも詳細に分布を解析することで，畳み込み層に対する初期化をも劇的に改善することを示す．

2. 関連研究

学習停滞の解決のために，ReLU の導入や，モデルパラメータの事前学習，特定の確率分布によるランダム初期化などが提案されている．

事前学習によるアプローチ：学習済みモデルの fine-tuning は，既存の公開されているモデルを流用する場合には学習時間の短縮に寄与するが，新たなモデルの学習には大規模なデータセットによる事前学習のやり直しが必要となる．また，事前学習は画像認識性能の観点ではランダムな初期化と同等で，事前学習は必要ないということが報告されており [2]，ランダムな初期化はモデル選択の自由度が高い方法と言える．

ランダム初期化法によるアプローチ：Glorot と Bengio[1] は，tanh 関数や softsign 関数を活性化関数としたときの，一様分布によるモデルパラメータの初期化法を提案した．この初期化法の導出は，ReLU に対応したものではなかった．He ら [3] は，図 1 に示すような CNN において，第 ℓ 層 $(1 \leq \ell \leq L)$ の重み係数 $\boldsymbol{W}^{(\ell)} := \left[\boldsymbol{w}_1^{(\ell)}, \cdots, \boldsymbol{w}_{C_\ell}^{(\ell)}\right]^\top \in \mathbb{R}^{C_\ell \times S_\ell}$ の任意の要素 $w_{i,j}^{(\ell)}$ の初期分布を $w_{i,j}^{(\ell)} \sim \mathcal{N}\left(0, 2C_{\ell-1}/(S_\ell C_\ell)\right)$ とする方法を提案した．ここで，C_ℓ, S_ℓ はそれぞれ，第 ℓ 層におけるチャンネル数，畳み込みのカーネルサイズ (e.g. 3×3 畳み込みにおいて $3 * 3 * C_{\ell-1}$) である．これらの初期化方法は，tensorflow や chainer などの主要な深層学習の枠組みにもモジュールが実装され，広く普及している方法である．

max-pooling 層を含む構造：ところで，ILSVRC において高成績を獲得した CNN の多くは，**max-pooling** 層を取り入れている．しかし，前述のランダム初期化法 [1, 3] の導出には max-pooling 層におけるユニットの分布の変化を無視しているため，max-pooling 層を含む CNN に対して用いるのは不適切であった．

3. CNN の統一的表現

本節では，CNN の層を統一的に扱うための再定式化を行う．基本的な CNN は畳み込み層，max-pooling 層，全結合層から構成されている．通常，畳み込み層と max-pooling 層は区別して数えられる．これに対し，本論文では，畳み込み層と max-pooling 層の組み合わせを 1 つの層として纏めたものを考える (図 1)．すると，順伝播，および逆伝播は，それぞれ

$$u_i^{(\ell)} := \left\langle \boldsymbol{w}_{c^{i,\ell}}^{(\ell)}, \boldsymbol{z}_{\boldsymbol{s}^{i,\ell}}^{(\ell-1)} \right\rangle + b_{c^{i,\ell}}^{(\ell)}, z_k^{(\ell)} := \max_{t \in \boldsymbol{t}^{k,\ell}} \max\left(0, u_t^{(\ell)}\right) \tag{1}$$

および

$$\Delta z_k^{(\ell-1)} = \left\langle \boldsymbol{w}_{\boldsymbol{h}^{k,\ell}}^{(\ell)}, \boldsymbol{\Delta u}_{\boldsymbol{j}^{k,\ell}}^{(\ell)} \right\rangle, \quad \Delta u_i^{(\ell)} = \frac{\partial z_{d^{i,\ell}}^{(\ell)}}{\partial u_i^{(\ell)}} \tag{2}$$

と表現できる．この表現方法では，max-pooling のない畳み込み層は $\boldsymbol{t}^{k,\ell} = \{k\}$ (i.e. 1×1 max-pooling) に相当し，全結合層は $\boldsymbol{t}^{k,\ell} = \{k\}$ かつ $\boldsymbol{s}^{i,\ell} = \{1, 2, \cdots, M_{\ell-1}\}$ に相当する．

4. 提案する初期化法

本研究では，max-pooling を含む CNN の誤差逆伝播における勾配の消失，発散を回避するためのモデルパラメータの初期化法として，

$$w_{i,k}^{(\ell)} \sim \mathcal{N}\left(0, \frac{2^{T_\ell} M_{\ell-1}}{(2^{T_\ell} - 1) S_\ell M_\ell}\right) \tag{3}$$

を開発した．ここで，M_ℓ, T_ℓ はそれぞれ，第 ℓ 層におけるユニット数，max-pooling の窓のサイズである．順伝搬に関しては次の定理が成り立つ．

定理 1. $2M_L \leq M_0$ とする．式 (3) で定義される確率ネットワークにおいて，順伝搬は衰退しない．

* 群馬大学大学院理工学府
† 株式会社デンソー
‡ 群馬大学次世代モビリティ社会実装研究センター (CRANTS)
§ 早稲田大学規範科学総合研究所 (IIRS)

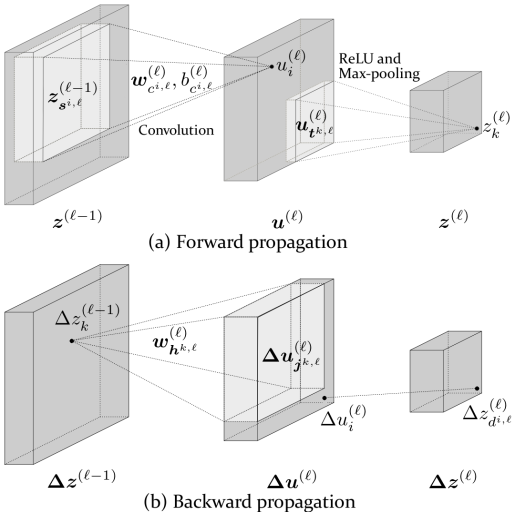

図 1　**CNN の層の統一的表現．** $\boldsymbol{s}^{i,\ell} \in \mathbb{R}^{S_\ell}$ と $\boldsymbol{t}^{k,\ell} \in \mathbb{R}^{T_\ell}$ はそれぞれ畳み込みの走査窓と max-pooling の走査窓を表す添字集合である．畳み込みの逆伝播は，順伝播と似た形式で表現でき，$\boldsymbol{j}^{k,\ell} = \{i \mid k \in \boldsymbol{s}^{i,\ell}\}$ なる走査窓を用いる．$d^{i,\ell}$ は $i \in \boldsymbol{t}^{d^{i,\ell},\ell}$ を満たす添字である．

M_0 は例題の特徴次元数であり，M_L は多クラス分類であればクラス数である．ほとんどの学習問題では $2M_L \leq M_0$ が満たされている．

5. 実験と考察

実験には，公開データセットの Food101 のサブセット (クラス数 10, 例題数 100) を用いた．最適化手法には Adam を用いて，パラメータは論文中の推奨値である $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$ を使用し，ミニバッチサイズは 5 とした．max-pooling を含むモデル (図 2a) において，既存法ではほぼ学習不可能だったモデルを約 50% の割合で学習できるようになった．これは，提案法を少数回繰り返すだけで学習可能になったことを意味する．

畳み込みの逆伝播における既存法との違い: 本研究の貢献は，max-pooling に対応する初期化法の提案だけではない．既存法よりも畳み込みの逆伝播を詳細に解析したことにより，max-pooling のない CNN においても，安定した収束が得られるようになった．逆伝播 (2) において，$\Delta z_k^{(\ell-1)}$ に逆伝播する $\boldsymbol{\Delta u}_{\boldsymbol{j}^{k,\ell}}^{(\ell)}$ の要素数 $J^{k,\ell}$ は順伝播 (1) における S_ℓ のように一定の大きさではなく，ユニット番号 k に依存し，$C_\ell \leq J^{k,\ell} \leq \frac{S_\ell C_\ell}{C_{\ell-1}}$ である．画像端ほど小さな値となり，中心付近で最大値を取る．He らの初期化法は，$T_\ell = 1$ かつ $J^{k,\ell} = \frac{S_\ell C_\ell}{C_{\ell-1}}$ というように，最大値で近似すると導出できるが，この条件を満たすのは全結合層であり，畳み込み層に相応しい初期化法ではなかった．これが，実験に用いた max-pooling を含まない構造でも学習にほぼ失敗した要因と推察される (図 2b)．これに対して，提案法では，$J^{k,\ell}$ をその期待値で近似することで，100% に近い割合で学習できるようになった (図 2b)．

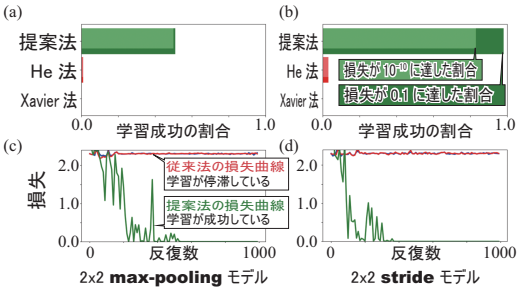

図 2　**学習の安定性に関するランダム初期値法の比較．** (a),(b) 18 層モデルにおいて異なる初期値で 100 試行したうち，損失の最小値が 0.1 および 10^{-10} 以下となった割合を，従来法 (He 法及び Xavier 法 [3])，および提案初期化法に対して比較している．(c),(d) 損失曲線の典型例．

付録 A. 導出と証明

提案法 (3) の導出: まず第 ℓ 層のバイアスを $b_i^{(\ell)} = 0$ と初期化し，重み係数の各要素を独立に $w_{i,k}^{(\ell)} \sim \mathcal{N}\left(0, \sigma_{w^{(\ell)}}^2\right)$ というように正規分布で初期化することを決めておく．すなわち，各層について未決定のパラメータ $\sigma_{w^{(\ell)}}^2$ を決定することが目的となる．ここで，逆伝播について，$\Delta \boldsymbol{z}^{(\ell)}$ の各要素が同分布であることを仮定すると，任意の要素の確率分布は，正規分布の再生性より，$\Delta z_k^{(\ell)} \sim \mathcal{N}\left(0, r^{(\ell)}\right)$ である．逆伝播での勾配の消失，発散を回避するためには，$r^{(\ell-1)} = r^{(\ell)}$ を満たすような $\sigma_{w^{(\ell)}}^2$ を決定する必要がある．

各層の逆伝播の分散の関係を導出すると，

$$r^{(\ell-1)} = r^{(\ell)} \sigma_{w^{(\ell)}}^2 (2^{T_\ell} - 1) S_\ell M_\ell / (2^{T_\ell} M_{\ell-1}) \tag{4}$$

という漸化式が得られる，これより，$r^{(\ell-1)} = r^{(\ell)}$ を満たすような重み係数の分散は，

$$\sigma_{w^{(\ell)}}^2 = 2^{T_\ell} M_{\ell-1} / ((2^{T_\ell} - 1) S_\ell M_\ell) \tag{5}$$

であり，我々の提案する初期化法 (3) が得られる．□

定理 1 の証明: 第 ℓ 層の活性前ユニット $\boldsymbol{u}^{(\ell)}$ の各要素の確率分布は，逆伝播と同様に考えると $u_i^{(\ell)} \sim \mathcal{N}\left(0, q^{(\ell)}\right)$ である．この時の入力層と出力層の順伝播の分散 $q^{(\ell)}$ の関係を導出すると，

$$q^{(L)} = 0.5 q^{(1)} M_0 M_L^{-1} \tau_1 \cdot \cdots \cdot \tau_{L-1}$$
$$\left(\text{where} \quad \tau_\ell = \frac{T_\ell 2^{T_\ell}}{2^{T_\ell} - 1} \int_0^\infty z^2 \phi(z) \Phi(z)^{T_\ell - 1} dz \right) \tag{6}$$

が得られる．$\phi(z), \Phi(z)$ はそれぞれ標準正規分布の確率密度関数と累積分布関数である．また，出力層 L については，活性化関数を恒等関数と考えており，中間層と同様の議論で導出した $\sigma_{w^{(L)}}^2 = M_{L-1}/(S_L M_L)$ を用いている．このとき $\tau_\ell \geq 1$ が成立しており，よって $2M_L \leq M_0$ ならば，$q^{(L)} \geq q^{(1)}$ が満たされ，順伝播における分散の減衰はない．□

参考文献

[1] Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, *PMLR* 9:249-256, 2010.
[2] He, K., Girshick, R. and Dollr, P.: Rethinking ImageNet Pre-training, arXiv:1811.08883v1, 2018.
[3] He, K., Zhang, X., Ren, S. and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, *ICCV15*, 1026–1034, 2015.