F-021

深層畳み込みネットワークの初期化アルゴリズム
Initialization of Convolution Layers in Deep Networks.

遠見 貴彦 *
Takahiko Henmi

加藤 毅 *
Tsuyoshi Kato

1 はじめに

畳み込みニューラルネットワーク（CNN）は，特に画像認識の分野で成果を挙げている機械学習モデルのひとつであり，すでに画像分類・物体検出といった実用的なタスクに応用されている．

CNN の性能向上を目的として，CNN の構成要素である層をより深く積み重ねる深層ネットワークが提案されてきた [1]．しかし，このような深い構造をもつモデルは学習する上で困難が伴う．具体的な問題として，勾配の消失・爆発と呼ばれる，最適化アルゴリズムの計算段階における不安定さが知られている．この問題に対する代表的な対処法のひとつに，学習可能なパラメータの初期値を工夫して勾配を安定させる方法（以下，初期化法）が提案されている．

初期化法の代表的な例は，乱数に基づく初期化法である．この手法では，パラメータの初期値はある分散をもつ分布に従う乱数を用いて生成される．この分散を調整することにより，ネットワークを伝播する信号の分散を一定程度に維持し，結果として勾配を安定させる効果を目的としている．この種の方法で従来から用いられているものでは，Xavier らの方法 [2]，Kaiming らの方法 [3] などが知られている．

これら従来法は CNN に対しても用いられているが，その導出においては CNN を考慮していないか，大幅に単純化して導出されていることから，理論的な CNN のサポートは十分ではないことに注意しなければならない．具体的には，畳み込み演算のオプションに関する解析は限定的で，またプーリング演算に関しては理論的には無視されてしまっている (表 2 参照)．

本研究では，CNN の構造をより精密にモデル化することで，新たな深層 CNN のパラメータの初期化法を提案する．まず，種々の CNN モデルを統一的に表現できるような新たな定式化を導入し，従来の解析において無視されていたプーリング演算等，現在実際に用いられている CNN の構造をより精密に扱うことができるようにする (2.1 節)．次に，この定式化を用いてネットワーク中を伝播する信号の統計的性質を解析する (2.2 節)．さらに，この結果の系として，信号の分散を維持するような条件を導出し，理論的にも CNN に対応した新たな初期化法を示す (2.3 節)．提案する初期化法と従来法を比較するため，実データを用いて CNN の学習する実験を行った．その調査結果を報告する (3 節)．

* 群馬大学理工学府 Gunma University.

2 方法論

2.1 CNN の定式化

CNN では通常，2 次元以上の画像が入力されるため，信号は一般に 3 階以上のテンソルとして表現される．しかし，テンソルを用いると添え字が煩雑となるため，等価な表現として，すべてのテンソルをベクトル化した表現を用いることにする．また，本研究の定式化では，“畳み込み”・“活性化” および “プーリング” の 3 種類の演算を一組みにし，ひとつの “層” として定義する．この定義を用いることで，後述するように CNN の構造をこの 1 種類の層が積層されたものと見做すことができ，様々なモデルを統一的に数式表現できるため，後の解析において扱いやすくなる (図 1 参照)．詳細な定義・導出は文献 [4] を参照されたい．以下，概要を示す．

ネットワークを L 層構造とし，第 ℓ $(1 \leq \ell \leq L)$ 層を式 (1a), (1b) および (1c) で定義する．

$$\begin{cases} u_i^{(\ell)} = \left\langle \boldsymbol{w}_{c(\ell,i),\boldsymbol{a}(\ell,i)}^{(\ell)}, \boldsymbol{z}_{\boldsymbol{s}(\ell,i)}^{(\ell-1)} \right\rangle + b_{c(\ell,i)}^{(\ell)}, & \text{(1a)} \\ v_i^{(\ell)} = f_\ell \left(u_i^{(\ell)} \right), & \text{(1b)} \\ z_i^{(\ell)} = g_\ell \left(\boldsymbol{v}_{\boldsymbol{t}(\ell,i)}^{(\ell)} \right). & \text{(1c)} \end{cases}$$

なお, $\boldsymbol{z}_{\boldsymbol{s}(\ell,i)}^{(\ell-1)}$ 等は新たに導入した記法で, ベクトル $\boldsymbol{z}^{(\ell-1)}$ から添え字集合 $\boldsymbol{s}(\ell,i)$ に含まれる添え字の成分のみを抜き出して並べたベクトルを表す．

畳み込み

式 (1a) は畳み込み演算を表現している．畳み込みは，入力（すなわち直前層の出力）$\boldsymbol{z}^{(\ell-1)}$ の一部分を取り出して，畳み込みカーネル $\boldsymbol{w}_i^{(\ell)}$ と積和を取ったものとして与えられる．ここで，計算にかかわる成分を指定する添え字集合 $\boldsymbol{a}(\ell,i), \boldsymbol{s}(\ell,i)$ などは，パディングやストライドといった畳み込みのオプションに依存して具体的に与えることができる．

活性化関数

式 (1b) は ReLU 関数など活性化関数と呼ばれる非線形な関数の適用を表している．活性化関数を適用しない場合も，f_ℓ を恒等関数とすれば今回の定式化に含まれる．

プーリング

式 (1c) はプーリング演算を表現している．プーリングは入力の一部領域を取り出して，あるスカラー値へ集約するような処理である．この参照する領域を $\boldsymbol{t}(\ell,i)$ で表し，プーリング関数 g_ℓ を適用する．よく使われる g_ℓ

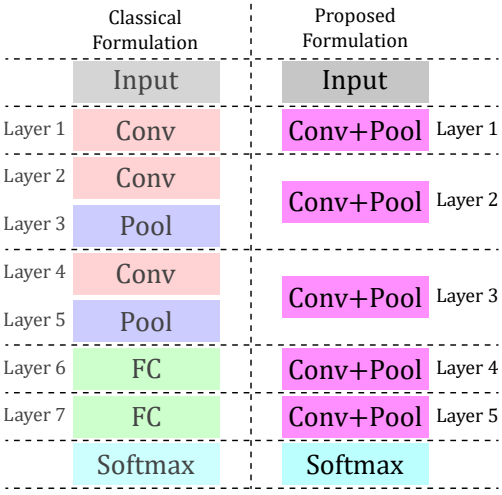

図 1: CNN の統一的な表現による定式化．本研究では CNN に現れる畳み込み・活性化・プーリングの 3 種類の演算を組み合わせたものを新たに “層” と定義し，ネットワークをこの種の “層” の積み重ねとして表現した．

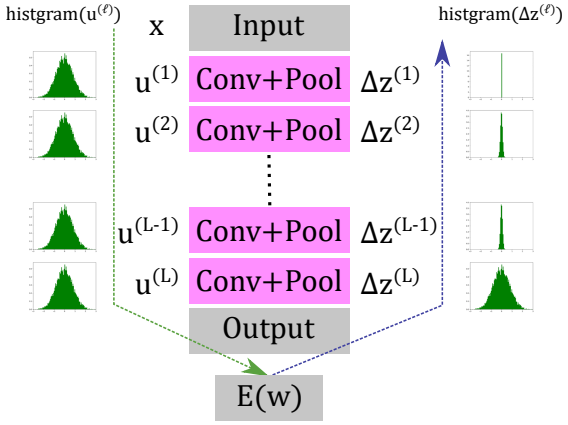

図 2: 分散の解析における勾配消失状態の信号．中心 0 の乱数でパラメータを初期化すると，$u_i^{(\ell)}, \Delta z_i^{(\ell)}$ は平均 0 の分布に従う．このとき，分散が小さければその絶対値も小さくなり，パラメータの更新が停滞する．

としては，最大値を返す Max Pooling，平均値を返す Average Pooling などが知られている．

全結合（アフィン変換）

全結合層の演算は，畳み込みの特殊な場合と見做せる．具体的には，入力のすべての要素を参照するような畳み込みは全結合に一致する．この上で，プーリングなしの “層” は，いわゆる全結合層と等価である．

逆伝播

式 (1a),(1b),(1c) に基づいて表現されたネットワークの逆伝播は，次式で与えられる:

$$\begin{cases} \Delta z_i^{(\ell-1)} = \left\langle \tilde{\boldsymbol{w}}^{(\ell)}_{\tilde{c}(\ell,i),\boldsymbol{h}(\ell,i)}, \Delta \boldsymbol{u}^{(\ell)}_{\boldsymbol{j}(\ell,i)} \right\rangle, & \text{(2a)} \\ \Delta u_i^{(\ell)} = \Delta v_i^{(\ell)} \dfrac{\partial v_i^{(\ell)}}{\partial u_i^{(\ell)}}, & \text{(2b)} \\ \Delta v_i^{(\ell)} = \Delta z^{(\ell)}_{d(\ell,i)} \dfrac{\partial z^{(\ell)}_{d(\ell,i)}}{\partial v_i^{(\ell)}}. & \text{(2c)} \end{cases}$$

ただし，目的関数 E に対する勾配を次のように略記した:

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

2.2 信号の統計的性質

まず，従来法を踏襲して，各層のパラメータ $\boldsymbol{w}_i^{(\ell)}, \boldsymbol{b}^{(\ell)}$ は，層ごとに固有の分散パラメータ $\sigma^2_{w^{(\ell)}}, \sigma^2_{b^{(\ell)}}$ を仮定して，それぞれ独立な正規乱数で初期化されていることを前提とする：

$$w_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma^2_{w^{(\ell)}}), \quad b_i^{(\ell)} \sim \mathcal{N}(0, \sigma^2_{b^{(\ell)}}). \qquad (3)$$

ネットワーク中の信号の分散は層を経るごとに変化するが，このとき信号の分散が小さくなると，パラメータの更新に用いられる勾配の絶対値が小さくなり，結果として最適化が停滞する (図 2 参照)．この現象は**勾配消失問題**として知られている．他方で，この分散が大きくなりすぎると数値的な安定性を欠き最適化に失敗する．こちらは**勾配爆発問題**として知られている．モデルパラメータを初期化する分散パラメータを用いて信号の分散をある程度一定に維持することで勾配の消失や爆発を防ぐのが，乱数による初期化法のアプローチである．しかしながら，ネットワーク中の分散を成分ごとに厳密に求めて初期化の分散パラメータを決定するのは困難であるので，従来法では導出にあたって信号の統計的独立性などを仮定して，分散を近似する手法をとっている [2], [3]．本稿でもこれに倣い，一定の仮定をおいたうえで各層間の分散の変化を近似的に記述し，これに基づいて初期化法を導出することとした．導出の詳細は文献 [4] を参照されたい．以下，概要を示す．

順伝播の場合

第 $\ell-1$ $(1 \leq \ell < L)$ 層について，$\boldsymbol{u}^{(\ell-1)}$ の各要素が統計的に独立で，同一の分布 $\mathcal{N}(0, q^{(\ell-1)})$ に従うと仮定する．第 ℓ 層に関して分散 $q_i^{(\ell)} = \mathrm{Var}\left[u_i^{(\ell)}\right]$ を考える．この量はユニット番号 i に依存する．ここで，第 ℓ 層の分散 $q^{(\ell)}$ を $q_i^{(\ell)}$ の層内での平均で近似することにすると，第 $\ell-1$ 層と ℓ 層の分散に関して次の関係が成り立つ:

$$q^{(\ell)} = \sigma^2_{b^{(\ell)}} + \sigma^2_{w^{(\ell)}} q^{(\ell-1)} \tau_{\ell-1} \frac{1}{M'_\ell} \varepsilon_\ell. \qquad (4)$$

ただし，$\tau_{\ell-1}$ は第 $\ell-1$ 層のプーリングの種類に依存する定数 (表 1 参照)，M'_ℓ は第 ℓ 層の畳み込み後のベクトル $\boldsymbol{u}^{(\ell)}$ の次元，$\varepsilon_\ell = \sum_{i=1}^{M'_\ell} |\boldsymbol{s}(\ell, i)|$ は層間の接続の総数

である．以上の議論を深い層へ向かって繰り返し適用すれば，式 (4) の関係が各層に関して帰納的に成り立つ．

逆伝播の場合

順伝播の場合とほぼ同様の議論が成り立つ．第 ℓ $(1 < \ell \leq L)$ 層について，$\Delta \boldsymbol{z}^{(\ell)}$ の各要素が統計的に独立で，同一の分布 $\mathcal{N}(0, r^{(\ell)})$ に従うと仮定する．第 $\ell - 1$ 層の分散 $r^{(\ell-1)}$ を近似して

$$r^{(\ell-1)} = \sigma^2_{w^{(\ell)}} r^{(\ell)} \gamma_\ell \frac{1}{M_{\ell-1}} \varepsilon_\ell \tag{5}$$

が成り立つ．ただし，γ_ℓ は第 ℓ 層のプーリングの種類に依存する定数 (表 1 参照)，$M_{\ell-1}$ は第 $\ell - 1$ 層の出力 $\boldsymbol{z}^{(\ell-1)}$ の次元である．

表 1: プーリングに関する定数 τ_ℓ, γ_ℓ．ここで，T_ℓ はプーリング領域のサイズ，ϕ, Φ はそれぞれ標準正規分布の密度関数，累積分布関数を表す．

プーリング	τ_ℓ	γ_ℓ
Max	$T_\ell \int_0^\infty s^2 \phi(s) \Phi(s)^{T_\ell - 1} \mathrm{d}s$	$\frac{2^{T_\ell} - 1}{T_\ell 2^{T_\ell}}$
Average	$\frac{1}{2T_\ell}\left(1 + \frac{T_\ell - 1}{\pi}\right)$	$\frac{1}{2T_\ell^2}$

2.3 提案する初期化法

前節の結果から，信号の分散を維持する条件が導出できる．ここでは，バイアスは定数 0 で初期化する ($\sigma^2_{b^{(\ell)}} = 0$ に相当）こととし，さらに $q^{(0)} = 1, r^{(L)} = 1$ と仮定して分散を維持する（すなわち，分散の比に関して $q^{(\ell)}/q^{(\ell-1)} = r^{(\ell)}/r^{(\ell-1)} = 1$ とする）条件から，式 (6),(7) のような分散パラメータの設定が導かれる．これが本稿で提案する初期化法である．

順伝播信号の分散を維持する初期化法

$$\sigma^2_{w^{(\ell)}} = \frac{M'_\ell}{\tau_{\ell-1}\varepsilon_\ell}. \tag{6}$$

逆伝播信号の分散を維持する初期化法

$$\sigma^2_{w^{(\ell)}} = \frac{M_{\ell-1}}{\gamma_\ell \varepsilon_\ell}. \tag{7}$$

従来法との比較

簡単な比較は表 2 にまとめた．提案法は従来法である Kaiming らの方法を拡張したものと見ることができる．順伝播条件において，畳み込み演算でパディングなし・カーネルのサイズを $k \times k \times d$ とするとき，$|\boldsymbol{s}(\ell, i)| = k^2 d$ である．また，Kaiming らの方法ではプーリングを無視しているので，プーリングに関する部分を恒等変換とするため $T_\ell = 1$ とおくと $\tau_\ell = 1/2$ となるので，式 (6) は

$$\sigma^2_{w^{(\ell)}} = \frac{M'_\ell}{\tau_{\ell-1}\varepsilon_\ell} = \frac{2M'_\ell}{M'_\ell \cdot k^2 d} = \frac{2}{k^2 d}$$

と表せる．これは Kaiming らの方法の順伝播条件における分散パラメータと一致する．また，逆伝播条件においても同様の設定をすると，畳み込みの出力チャネルを d' とおけば $|\boldsymbol{j}(\ell, i)| = k^2 d', \gamma_\ell = 1/2$ となり，さらに層間の接続数は順伝播・逆伝播で同じであることから $\varepsilon_\ell = \sum_{i=1}^{M_{\ell-1}} |\boldsymbol{j}(\ell, i)|$ が成り立つ．したがって，式 (7) より

$$\sigma^2_{w^{(\ell)}} = \frac{M_{\ell-1}}{\gamma_\ell \varepsilon_\ell} = \frac{2M_{\ell-1}}{M_{\ell-1} k^2 d'} = \frac{2}{k^2 d'}$$

と表せる．これは Kaiming らの方法の逆伝播条件における分散パラメータと一致する．

提案法は次のように振る舞う:

- 畳み込みでパディングをおこなう場合，パディングされた要素を含む領域への畳み込みでは，実質的に参照される入力の要素の数が減少する．これは定式化において $\boldsymbol{s}(\ell, i)$ の要素数が減少するケースとして表現され，ε_ℓ がパディングなしの場合と比較してより小さくなり，結果として $\sigma^2_{w^{(\ell)}}$ は大きくなる．これは，定数であるパディングされた要素を参照する畳み込み部分において分散が相対的に減少することを補う効果がある．
- 順伝播条件において Max Pooling を用いるとき，τ_ℓ は T_ℓ に関して単調増加する．すなわち，$\sigma^2_{w^{(\ell)}}$ は T_ℓ に関して単調減少である．max 演算を経ると値はより大きくなるので，分散も大きくなると考えられる．このような傾向はプーリング領域 T_ℓ が大きければ大きいほど顕著になるため，提案法はこれを補って分散を減衰させる効果がある．
- 順伝播条件において Average Pooling を用いるとき，τ_ℓ は T_ℓ に関して単調減少する．すなわち，$\sigma^2_{w^{(\ell)}}$ は T_ℓ に関して単調増加である．一般に平均をとる操作は分散を減少させる．提案法ではこれを補って分散を増幅する効果がある．
- 逆伝播条件において Max Pooling を用いるとき，γ_ℓ は T_ℓ に関して単調減少である．すなわち，$\sigma^2_{w^{(\ell)}}$ は T_ℓ に関して単調増加である．Max Pooling の場合，最大値を与えた（多くの場合）単一の要素に信号を逆伝播させるので，信号の分散を維持するには単一要素により大きな分散の信号を与える必要がある．提案法はこのような性質を考慮して分散を与える効果がある．
- 逆伝播条件において Average Pooling を用いるとき，γ_ℓ は T_ℓ に関して単調減少である．すなわち，$\sigma^2_{w^{(\ell)}}$ は T_ℓ に関して単調増加である．Average Pooling の場合，逆伝播では入力の各要素に分散が等分されて分配されるので，分散が減少する．提案法はこれを補うため分散を増幅する効果がある．

表 2: 従来法との比較．従来法は CNN に対して理論的な裏付けが不十分ながらも用いられているのに対して，提案法は今日用いられている CNN の多くの構造を理論的にサポートしている．

初期化法	演算のサポート			
	全結合	ReLU	畳み込み	プーリング
Xavier	○	×	×	×
Kaiming	○	○	△	×
提案法	○	○	○	○

3 実験

本節では，紙面の制約から実験の一部を紹介する．詳細な実験条件および結果は，文献 [4] を参照されたい．

データセット

実験に用いたデータセット `Fungi10` は iNaturalist 2019[*1] データセットのサブセットで，スーパーカテゴリ “Fungi” の先頭 10 カテゴリを取り出した分類タスク向けデータセットである．データ全体のうち 75% を訓練用データに，残り 25% を検証用データとし，データはそれぞれ 224×224 サイズにリサイズし，正規化を行った．データ拡張は行っていない．

モデルアーキテクチャ

実験に用いたモデル `F34` は全 34 層からなる深層 CNN である（表 3 参照）．このモデルには Max Pooling や畳み込みのパディングが含まれており，さらに畳み込みネットワークの最後には Global Average Pooling を適用している．

最適化に関する設定

今回のタスクは分類問題であり，損失関数には Cross Entropy Loss を用いた．また，最適化アルゴリズムには Stochastic Gradient Descent (SGD) を用いた．学習率は 10^{-1} から 10^{-5} まで 1/10 刻みで実験した．SGD のオプションとして，batch size は 64，momentum は 0.9，weight decay は 10^{-4} を設定している．

結果と考察

実験では初期化法として Xavier, Kaiming, 提案法の 3 種類を用いてそれぞれ初期値をランダムに生成しながら 10 回ずつ学習し，検証用データに対する正解率を比較した．

表 4 にその結果を示す．これは，各手法においてそれぞれ最も高い正解率を示した学習率 10^{-3} を用いて，十分収束したと考えられる 300 エポックまで学習したものである．表に示すように，最大値で比較すると，提案法は従来法よりも高い正解率を得られた．これは，提案法で精密に信号の伝播を表現した結果として勾配の消失や爆発が抑制され，従来法を用いたときよりも最適化がうまく進行したためである．また，各四分位点において比較しても，それぞれ従来法と同等かそれ以上の正解率を実現している．このことは，提案法を適用し，初期値を生成しなおして繰り返し学習するだけで，従来法よりも高い正解率を示す学習結果を得られる可能性を示唆している．

表 3: 実験に使用した CNN の構造．なお，(k/ss/pp) はカーネルサイズが $k \times k$, ストライドが s, パディングが p であることを表している．

層の出力サイズ	`F34`
(3,224,224)	Input Image
(64,112,112)	Conv(7/s2/p3)
(64,56,56)	MaxPool(3/s2/p1)
(64,56,56)	Conv(3/p1)×6
(128,28,28)	Conv(3/s2/p1)
(128,28,28)	Conv(3/p1)×7
(256,14,14)	Conv(3/s2/p1)
(256,14,14)	Conv(3/p1)×11
(512,7,7)	Conv(3/s2/p1)
(512,7,7)	Conv(3/p1)×5
(512,1,1)	Global Average Pooling
10	Linear
パラメータ数	2.11×10^7

表 4: 検証用データに対する正解率 (%) の比較．提案法は，10 回試行の最大値，各四分位点いずれにおいても従来法と同等かそれ以上の予測性能を得た．

初期化法	Xavier	Kaiming	提案法
最大値	24.61	69.92	**70.90**
第 1 四分位点	24.61	68.36	**69.19**
中央値	24.61	69.04	**69.73**
第 3 四分位点	24.61	69.48	**70.21**

参考文献

[1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[2] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, PMLR, 2010, pp. 249–256.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[4] T. Henmi, Y. Hirohashi, and T. Kato, *Initialization of convolution layers in deep networks*, to appear on arXiv, Aug. 2020.

[*1] `https://www.kaggle.com/c/inaturalist-2019-fgvc6`