

スキップ接続によるプーリング層の構成および表現力に基づくそれらのモデルの解析

Construction of pooling layer by skip connection and analysis based on expressive power of these models

長瀬 准平 *1 石渡 哲哉 *2
Jumpei Nagase Tetsuya Ishiwata

*1 芝浦工業大学大学院 理工学研究科
Graduate School of Engineering and Science, Shibaura Institute of Technology

*2 芝浦工業大学 システム理工学部
College of Systems Engineering and Science, Shibaura Institute of Technology

In this research, we consider structures such as pooling layer and skip connection from the viewpoint of expressive power in order to organize design of neural networks models. We showed that widely used these structures can be understood as a composition of affine functions and concatenated activation functions. Moreover, we show the followings: (i) the pooling layer explicitly decreases expressive power, (ii) there is no deference in expressive power between addition and concatenation as skip connection for fully connected neural networks, and (iii) the single activation block has superior expressive power compared to the multiple activation block. These results propose one guideline for design of neural networks models.

1. はじめに

深層学習で用いられるニューラルネットワークモデルの設計は、いくつかのレイヤー（ベクトル関数）を階層的に積み上げることで行われる。ここで、「どのようにレイヤーを積み上げるか」「どのようなレイヤーを積み上げるか」が重要な研究課題である。画像認識の分野では、畳み込み層とプーリング層を積み重ねた畳み込みニューラルネットワーク（Convolution Neural Networks; CNN）と呼ばれるモデルを用いて画像の特徴を効果的に抽出できるとされており、2012年頃から活発に研究が行われている。モデルの構造に関する近年の大きな成果の一つとしてはスキップ接続が挙げられる。スキップ接続はILSVRC2015の優勝モデルであるResNet[?]で用いられたことで注目され、それ以降提案されているモデルの多くにスキップ接続の構造が用いられている。一方で、これらの提案モデルの多くは経験則によるものであり、モデル設計に関する体系的な理論が存在しないことが問題である。

本研究では、ニューラルネットワークモデルの設計を体系化することを目的として、表現力の観点からプーリング層やスキップ接続などの構造について考察する。結果として、広く用いられているこれらの構造は単純なニューラルネットワークモデルに結合のスキップ接続を導入することで構成できることを数学的に示した。表現力に関する解析としては、プーリング層は陽に表現力を低下させること、加算と結合のスキップ接続は全結合ニューラルネットワークにおいて表現力に差異を与えないこと、そして、現在有効とされているsingle activationブロック構造が、multi activationブロック構造と比較して表現力の意味で優位性を持つことが示された。以上の結果より、表現力の観点からモデル設計における一つの指針を与えることができる。

2. ニューラルネットワークモデル

ニューラルネットワークモデルはレイヤーと呼ばれる関数を繰り返し合成することにより設計される。**多層パーセプトロン**

連絡先: 長瀬准平, 芝浦工業大学大学院理工学研究科, mf18061@shibaura-it.ac.jp

ンは**全結合層**と**活性化層**が繰り返し合成されているモデルであり、最も単純なニューラルネットワークモデルの一つとして知られている。画像認識の分野で用いられている**畳み込みニューラルネットワーク**は、特徴的な構造として**畳み込み層**と**プーリング層**をもつ。また、近年では**スキップ接続**と呼ばれる複数のレイヤーをまたぐ接続が導入されることが多く、優れた性能を与えることが知られている。スキップ接続と同様の構造は、入力を複数のレイヤーへ分岐させる場合や、複数の出力を併合させて用いる場合にも扱うことがある。

2.1 多層パーセプトロン

多層パーセプトロンは1970年代頃に神経回路網の数理モデルとして提案されたものが起源となり、現代の深層学習で用いられているニューラルネットワークモデルの基本となっている。その特徴として、全結合層と活性化層と呼ばれる関数の合成で表されることが挙げられる。

定義 1 (L 層多層パーセプトロン) L 層多層パーセプトロンMLPは全結合層 $w_i: \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_{i+1}}$ と活性化層 $f: \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_i}$ の合成によって定義される。

$$\text{MLP}_L(f) = w_L \circ f \circ w_{L-1} \circ \dots \circ f \circ w_0.$$

特に、本稿では活性化層としてReLU関数を用いている多層パーセプトロンを**ReLUパーセプトロン**と呼ぶことにする。ReLU関数などについての詳細は以下で述べる。

2.1.1 全結合層

全結合層は入力 \mathbf{x} に対して重み行列 W とバイアスベクトル \mathbf{b} を用いて出力 $y = W\mathbf{x} + \mathbf{b}$ を返すレイヤーである。入力の成分 x_i と出力の成分 y_j が全て重み付けられて結合していることから全結合層や重み層と呼ばれる。あるいは、結合が密であることからDense Layer、重み層による変換が数学的にはAffine関数として理解されることからAffine Layerなどとも呼ばれる。本稿では、しばしば重み層とAffine関数を同一視して説明する。ニューラルネットワークモデルの学習では、重み層がもつパラメータが最適値に近づくように更新することが目的となる。

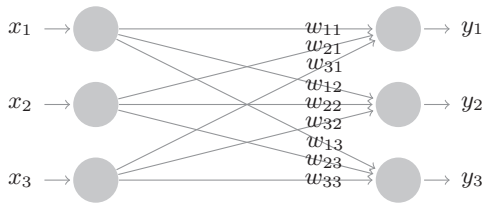


図 1: 全結合層の模式図. 入力 x_i と出力 y_j がパラメータ w_{ij} によって重み付けており, それぞれの出力は入力の線形結合によって表される.

2.1.2 活性化層

活性化層は入力 \mathbf{x} に対して非線形変換を行うレイヤーである. 神経素子のモデルとして神経の発火現象を模倣するために導入されたことから活性化層と呼ばれる. 本稿では成分ごとに同じ変換を行うもの, すなわち出力を $\mathbf{y} = (y_1, \dots, y_d)^T$ として

$$\mathbf{y} = (y_1, \dots, y_d)^T = (f(x_1), \dots, f(x_d))^T =: \mathbf{f}(\mathbf{x})$$

のような変換を行うレイヤーを活性化層として扱い, f を用いた活性化層 f などと呼ぶことにする^{*1}. 活性化層で用いられる変換の例として次のものが挙げられる.

- ReLU

$$\text{ReLU}(x) = \max(x, 0)$$

- シグモイド^{*}

$$\sigma(x) = (1 + e^{-x})^{-1}$$

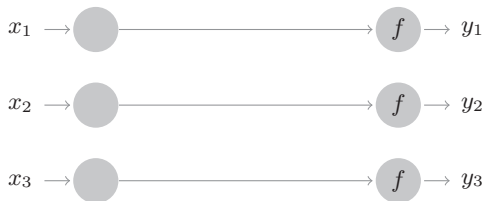


図 2: 活性化層の模式図. 入力の各成分が $f: \mathbb{R} \rightarrow \mathbb{R}$ によって変換される.

2.2 畳み込みニューラルネットワーク

畳み込みニューラルネットワークは主に画像認識の分野で発展したモデルであり, 畳み込み層と呼ばれる局所的な特徴抽出のための重み層と, 特徴量の圧縮のためのプーリング層が用いられることが特徴である. そのモデルの多くは活性化層や全結合層も併用して設計される.

2.2.1 畳み込み層

畳み込み層は入力 \mathbf{x} に対してフィルター処理と呼ばれる操作によって出力 \mathbf{y} を得るレイヤーである. 畳み込み層は学習可能なフィルター処理として考えることができる. 実用上は, RGB 画像などの多チャンネルデータに対して異なるフィルターを多数用いて特徴抽出を行うため, 入力として多次元配列が考えられることが多いが, 本研究では単純な 1 チャンネルの 2 次元画像データを想定して解析を行なっている^{*2}.

^{*1} 成分ごとに異なる変換を行うソフトマックス層などのレイヤーも存在するが, 出力層としてのみ用いられることが多いために本研究では扱わない.

^{*2} 本研究の結果は多チャンネルの場合にも単純に拡張できる.

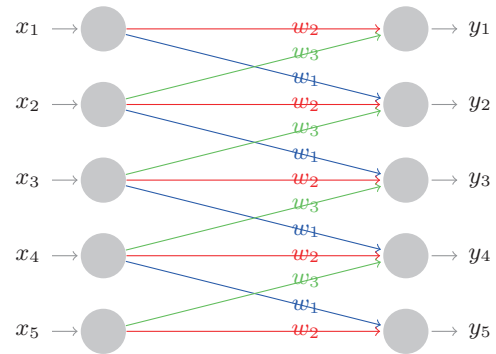


図 3: サイズ 3, 移動幅 1 の 1 次元畳み込み層の模式図. 同一の重み w_1, w_2, w_3 のみを用いてスパースな線形結合が行われる.

2.2.2 プーリング層

プーリング層は畳み込み層の後に用いられることが多く, 入力 (特に画像データ) を圧縮するための変換をもつ. 本稿では入力データのサイズを $1/n$ に圧縮するプーリング層をサイズ n のプーリング層と呼ぶことにする. 平均値を取る平均値プーリングと, 最大値を取る最大値プーリングが一般に用いられている.

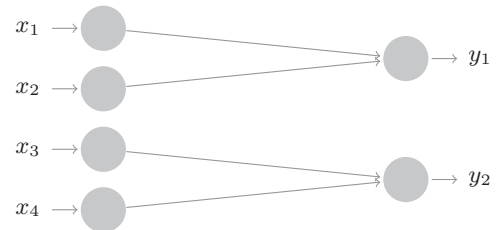


図 4: サイズ 2 の 1 次元プーリング層の模式図. $(x_1, x_2), (x_3, x_4)$ をそれぞれ圧縮したものが y_1, y_2 として出力される.

2.3 スキップ接続

スキップ接続はある層をまたぐ接続として導入される. 広く用いられているスキップ接続は加算と結合の二種類であり, スキップ接続によりバイパスされるいくつかの層をまとめてブロックと呼ぶ. ブロックの変換を $B(\mathbf{x})$ と表すと, 加算と結合のスキップ接続はそれぞれ次のように定義される.

- 加算

$$\mathbf{y} = B(\mathbf{x}) + \mathbf{x}$$

- 結合

$$\mathbf{y} = \begin{pmatrix} B(\mathbf{x}) \\ \mathbf{x} \end{pmatrix}$$

ただし, $(,)$ はベクトルの結合を表す.

その他の接続に関しては, 各成分ごとの差, 積, 最大値, 平均値を取るものや, 入力 \mathbf{x} に線形変換などを施してから接続するものなどがある. また, 二種類の層 L_1 と L_2 への分岐や併合としても加算や結合を考えることができ,

$$L_1(\mathbf{x}) + L_2(\mathbf{x})$$

のような層の併合もスキップ接続と同様に自然に導入される.

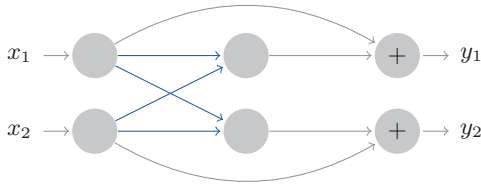


図 5: 加算のスキップ接続の模式図. 青色の変換はバイパスされるレイヤーやブロックを表している.

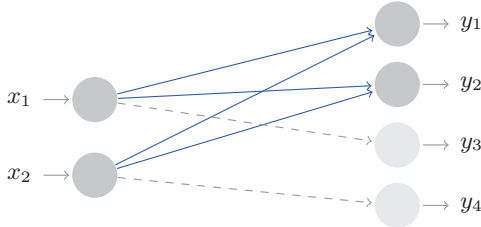


図 6: 結合のスキップ接続の模式図. 青色の変換はバイパスされるレイヤーやブロックを表している.

3. 本論

重み層 (Affine 関数) と活性化層の合成により単純なニューラルネットワークモデルである多層パーセプトロンが設計される. 本研究では, 多層パーセプトロンに結合のスキップ接続を導入したモデルを用いて一般的なプーリング層やスキップ接続を具体的に構成できることを示す. さらに, 活性化関数の種類を追加することで積のスキップ接続も具体的に構成することができる. これらの結果に関連して, プーリング層やスキップ接続の表現力に基づいた解析を行い, いくつかのモデルの解析結果を述べる.

3.1 プーリング層の構成と表現力

平均値プーリングは入力ベクトルに対して部分的に平均値を取る変換であるため, 低ランクの線形変換により実現できる. また, 最大値関数に関する性質

$$\max(x, y) = \text{ReLU}(x - y, 0) + y$$

より, 任意のサイズのベクトルの最大値関数が ReLU 関数と線形変換の合成により実現される. したがって, 入力ベクトルに対して部分的に最大値を取る変換である最大値プーリングも実現することができる. 例として, (a_1, a_2) の最大値プーリングは次のように表せる.

$$\left(f \circ \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \end{pmatrix} \right) \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad \left(= \max(a_1, a_2) \right),$$

ただし, f は ReLU 関数を用いた活性化層である. 以上より, 次の結果を得る.

定理 1 (プーリング層の表現力) サイズ n の平均値プーリングは線形パーセプトロン (0 層の多層パーセプトロン) により表現することができる. サイズ n の最大値プーリングは, スキップ接続をもつ m 層の ReLU パーセプトロンにより表現することができる. ただし, m は $\log_2(n) \leq m$ をみたす.

3.2 スキップ接続の構成と表現力

スキップ接続やレイヤー併合として用いられる演算の多くは線形演算であるため, 併合したい入力をあらかじめ結合させておくことで実現できる. また, 各成分ごとの最大値を取る場合には, 最大値プーリングと同様の方法により実現することができる. 例として, $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$ と $\mathbf{y} = (y_1, y_2)^T \in \mathbb{R}^2$ の成分ごとの平均値, 最大値を返すスキップ接続はそれぞれ

$$\begin{pmatrix} \frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix} \quad \left(= \begin{pmatrix} \text{mean}(x_1, y_1) \\ \text{mean}(x_2, y_2) \end{pmatrix} \right),$$

$$\left(\sigma \circ \begin{pmatrix} \mathbf{I} & -\mathbf{I} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} + \begin{pmatrix} \mathbf{O} & \mathbf{I} \end{pmatrix} \right) \begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix} \quad \left(= \begin{pmatrix} \max(x_1, y_1) \\ \max(x_2, y_2) \end{pmatrix} \right),$$

として表すことができる. ただし, \mathbf{O}, \mathbf{I} はそれぞれ 2×2 の零行列と恒等行列, σ は ReLU 関数の活性化層である. したがって, 基本的なスキップ接続は結合スキップ接続に対する線形変換と活性化層の合成で表現可能である. また, \log と \exp に関して性質

$$\exp(\log(x) + \log(y)) = xy$$

が成り立つことから, 新たな活性化層として \log と \exp を用いることで成分ごとの積を返すスキップ接続を表現することができる. すなわち, σ_1, σ_2 をそれぞれ \log と \exp を用いた活性化層として, \mathbf{I} を 2×2 の恒等行列とすると,

$$\sigma_2 \circ \begin{pmatrix} \mathbf{I} & \mathbf{I} \end{pmatrix} \circ \sigma_1 \begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix} \quad \left(= \begin{pmatrix} x_1 y_1 \\ x_2 y_2 \end{pmatrix} \right),$$

と表すことができる. 以上より, 次の結果を得る.

定理 2 (スキップ接続の表現力) 加算, 除算, 平均値のスキップ接続はいずれも結合のスキップ接続と線形変換の合成で表現することができる. 最大値のスキップ接続は加算のスキップ接続をもつ 2 層の ReLU パーセプトロンで表現できる. 積のスキップ接続は \log と \exp の二種類の活性化関数をもつ 2 層のパーセプトロンで表現できる.

また, 結合と加算のスキップ接続について次の対応がある.

命題 1 (結合と加算のスキップ接続の対応) w を全結合層とし, 二つの層 L_1, L_2 との合成を考える. このとき, 次の条件をみたす全結合層 u, v が存在する.

$$w \circ \begin{pmatrix} L_1 \\ L_2 \end{pmatrix} = u \circ L_1 + v \circ L_2.$$

ただし, u と v の出力次元は w と等しく, 入力次元は L_1 と L_2 の出力次元とそれぞれ等しい. また, 入力次元の和は w の入力次元と等しい.

この性質から、適切な次元の全結合層を用いることで加算と結合のスキップ接続に対応関係を与えることができ、スキップ接続の種類が表現力に差異を与えないことがわかる。

定理 3 (全結合層におけるスキップ接続の表現力) w_1, w_2, w' を全結合層, σ_1, σ_2 を活性化層とすると, 次の二つのモデルは表現力が等しい.

$$M_a = w_1 \circ \sigma_1 + w_2 \circ \sigma_2, \quad (1)$$

$$M_b = w' \circ \begin{pmatrix} \sigma_1 \\ \sigma_2 \end{pmatrix}. \quad (2)$$

ただし, w' は w_1, w_2 を合わせたサイズをもつ.

多層化した場合も同様の結果が得られる.

3.3 Single activation ブロックの表現力

w を重み層, σ を活性化層として, 次のようなニューラルネットワークモデルを考える.

$$M_1 = w_2 \circ \begin{pmatrix} \sigma \circ w_1 & \\ & \text{id} \end{pmatrix} \circ \begin{pmatrix} \sigma \circ w_0 & \\ & \text{id} \end{pmatrix},$$

$$M_2 = w'_2 \circ \begin{pmatrix} \sigma \circ w'_1 \circ \sigma \circ w'_0 & \\ & \text{id} \end{pmatrix}.$$

M_1 は single activation block 構造, M_2 は double activation block 構造である. このとき, 任意の w'_0, w'_1, w'_2 に対して,

$$w_0 = w'_0,$$

$$w_1 = \begin{pmatrix} w'_1 & \mathbf{0} \\ \mathbf{0} & w'_0 \end{pmatrix},$$

$$w_2 = \begin{pmatrix} w'_2 & \mathbf{a} & -\mathbf{a} & w'_2 \end{pmatrix}$$

とすることで $M_1 = M_2$ となる. ただし, \mathbf{a} は任意の Affine 関数とし, ${}^*3w'_2 = \begin{pmatrix} w'_2 & v'_2 \end{pmatrix}$ とおいた. 同様の考察は, 活性化層が L 個含まれるモデルについても行うことができ, 以上より, 次の結果を得る.

定理 4 (single activation block の優位性) 単層パーセプトロンに結合のスキップ接続を加えて L 回合成したモデルは, L 層の多層パーセプトロンモデルに結合のスキップ接続を加えたモデルよりも表現力が大きい.

また, 加算のスキップ接続を用いたモデルに関しても, 加算と結合の対応規則 (命題 1) を用いて同様の結果が得られる.

4. 先行研究におけるモデル構造

本稿で取り扱ったプーリング層とスキップ接続の種類はディープラーニングライブラリの一つである「keras」に実装されて

*3 合成等が定義される適切な次元の, 任意の Affine 関数とする.

いるものを参考にした. ResNet[?] では加算のスキップ接続が *residual skip connection* として提案され, DenseNet[?] では結合のスキップ接続が提案された. 結合のスキップ接続はベクトルの次元と同時に計算量やパラメータも増加させる働きがあり, 学習を高速化させるための多数の工夫が DenseNet では取り入れられていた. その他のスキップ接続としては, 平均値スキップ接続が FractalNet[?], 多段階スキップ接続が RoR[?] で用いられている. また, single ReLU ブロック構造は PyramidialResNet[?] や DenseNet などから良い成果が報告されている.

5. 結論・まとめ

本研究では, 結合のスキップ接続を導入した ReLU パーセプトロンや, 活性化層の種類を増やしたモデルを用いることでプーリング層やスキップ接続の一般的な構造が構成できることを示し, これらの構造を表現力の観点で解析した. 結果として, 最大値プーリングが多層の非線形性を持つこと, プーリング層や (結合以外の) スキップ接続が表現力に関して有意性がないこと, そして, single activation ブロックが multi activation ブロックよりも高い表現力を持つことが数学的に導かれた. 以上の結果は, スキップ接続やプーリング層といった構造を用いる場合に, 大きな表現力をもつモデルの設計方法を示していることから, ニューラルネットワークモデルの設計における一つのガイドラインを与えることができたと考えられる. 一方で, 学習に関する効果については触れておらず, 今後, 表現力と学習の両面からモデル設計の理論が構築されていくべきであると考える.

参考文献

- [Han 17] Han, D., Kim, J., and Kim, J.: Deep pyramidal residual networks, in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 6307–6315 (2017)
- [He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016)
- [Huang 17] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q.: Densely connected convolutional networks, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708 (2017)
- [Larsson 16] Larsson, G., Maire, M., and Shakhnarovich, G.: Fractalnet: Ultra-deep neural networks without residuals, *arXiv preprint arXiv:1605.07648* (2016)
- [Zhang 18] Zhang, K., Sun, M., Han, T. X., Yuan, X., Guo, L., and Liu, T.: Residual networks of residual networks: Multilevel residual networks, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 28, No. 6, pp. 1303–1314 (2018)