畳み込みニューラルネットワークにおける順伝播・誤差逆伝播計算

宮本　崇

2017/4/7

1　はじめに

畳み込みニューラルネットワークにおける順伝播・逆伝播計算を，定義に基づいたテンソルの要素間の関係式として定式化する．

ここでは，空間方向に 2 次元，チャンネル方向に 1 次元の，3 次元の入力データを仮定する．そのような入力データの例として，RGB の 3 チャンネルを有した 2 次元画像がある．

以下に，畳み込みニューラルネットワークに関して本稿で用いる記号の定義を表 1 にまとめる．

Symbols	Meaning	size
n	ミニバッチに関する添字	N
c	層への入力データのチャンネル数に関する添字	C
h	層への入力データの高さに関する添字	H
w	層への入力データの幅に関する添字	W
k	層からの出力データのチャンネル数に関する添字	K
p	層からの出力データの幅に関する添字	$P = \lceil \frac{H - R + 1 + 2p_h}{u} \rceil$
q	層からの出力データの高さに関する添字	$Q = \lceil \frac{W - S + 1 + 2p_w}{v} \rceil$
r	カーネルの高さに関する添字	R
s	カーネルの幅に関する添字	S
u	高さ方向のストライド	$-$
v	幅方向のストライド	$-$
p_h	高さ方向のパディング	$-$
p_w	幅方向のパディング	$-$
$\boldsymbol{Z}^{(l)} = \{Z^{(l)}_{nchw}\}$	l 層への入力テンソル	$N \times C \times H \times W$
$\boldsymbol{F}^{(l)} = \{F^{(l)}_{kcrs}\}$	l 層におけるカーネルテンソル	$K \times C \times R \times S$
$\boldsymbol{\Delta}^{(l)} = \{\Delta^{(l)}_{nchw} = \frac{\partial E_n}{\partial Z^{(l)}_{nchw}}\}$	l 層におけるデルタテンソル	$N \times C \times H \times W$

表 1　畳み込みニューラルネットワークに関する記号の定義.

2　畳み込みニューラルネットワーク

2.1　ニューラルネットワークにおける順伝播・逆伝播

多層パーセプトロンや畳み込みニューラルネットワークの順伝播・逆伝播の処理は，次のように考えることができる．

l 層への入力を $\boldsymbol{Z}^{(l)}$ とおく．$\boldsymbol{Z}^{(l)}$ は，ミニバッチや空間上の位置，特徴量次元を指定するための添字を有する，高階のテンソルとなる (図 1)．l 層での順伝播計算とは，このテンソルに対して何らかの操作を行い，$\boldsymbol{Z}^{(l+1)}$ を出力して次の層へ渡す処理を意味する．このとき，l 層でのテンソルへの操作に関連するパラメタをまとめたテンソルを $\boldsymbol{W}^{(l)}$ とすると，l 層での順伝播計算は形式的に次のように書くことができる．

$$\boldsymbol{Z}^{(l+1)} = f(\boldsymbol{Z}^{(l)}; \boldsymbol{W}^{(l)}) \tag{1}$$

次に, n 番目のミニバッチに対する誤差関数 E_n を, $\boldsymbol{Z}^{(l)}$ のある要素 $Z^{(l)}_{n(\cdot)}$ で偏微分した値を $\Delta^{(l)}_{n(\cdot)} = \dfrac{\partial E_n}{\partial Z^{(l)}_{n(\cdot)}}$ とおき，$\Delta^{(l)}_{n(\cdot)}$ を要素に持つテンソルを $\boldsymbol{\Delta}^{(l)}$ とする．$\boldsymbol{\Delta}^{(l)}$ は，$\boldsymbol{Z}^{(l)}$ と同じ階数，サイズのテンソルである．また，$E = \dfrac{1}{N}\sum_{n=1}^{N} E_n$ を $\boldsymbol{W}^{(l)}$ のある要素 $W^{(l)}_{(\cdot)}$ で偏微分した値を $\partial W^{(l)}_{(\cdot)}$ とし，$\partial W^{(l)}_{(\cdot)}$ を要素に持つテンソルを $\boldsymbol{\partial W}^{(l)}$ とする．l 層における逆伝播計算とは，これらのテンソルを一つ先の層の $\boldsymbol{\Delta}^{(l+1)}$ から求めることであり，形式的に次のように書くことができる．

$$\boldsymbol{\Delta}^{(l)} = g(\boldsymbol{\Delta}^{(l+1)}) \tag{2}$$

$$\boldsymbol{\partial W}^{(l)} = h(\boldsymbol{\Delta}^{(l+1)}) \tag{3}$$

したがって，ニューラルネットワークにおける順伝播・逆伝播計算の定式化とは，式 (1) や (2)，(3) の具体的な形を求めることを意味している．以降では，畳み込み層やプーリング層といった層種ごとに，定式化を行っていく．以上のようなある層での処理を，図 2 に模式的に示す．

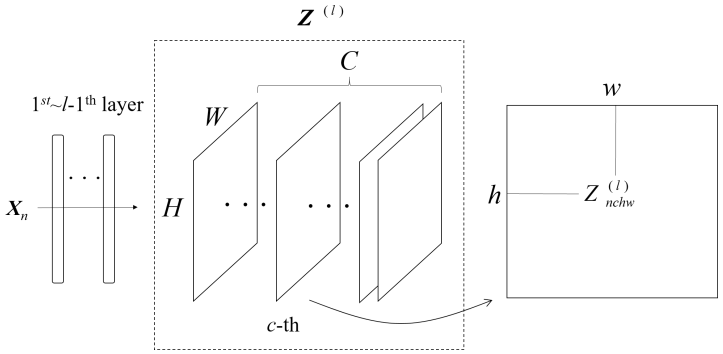

図 1　入力テンソルのイメージ：2 次元画像の場合．図中の記号は表 1 に従う．

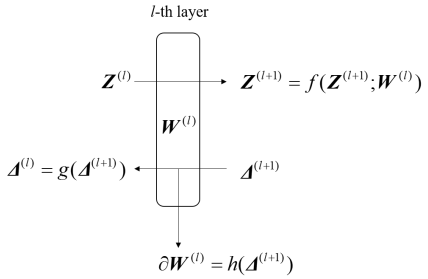

図 2　l 層における順伝播・逆伝播処理.

2.2 畳み込み層

■順伝播 畳み込み層は，入力テンソルに対して，空間サイズの小さなカーネルを移動させながら内積をとる畳み込み操作を行うことによって，出力テンソルを得る層である (図 3)．入力テンソル，カーネル，ストライド，パディングなどの値を表 1 のように定めると，畳み込み層における順伝播計算は以下のように表される．

$$Z_{nkpq}^{(l+1)} = \sum_{c=1}^{C} \sum_{r=1}^{R} \sum_{s=1}^{S} Z_{nchw}^{(l)} \cdot F_{kcrs}^{(l)} + b^{(l)} \tag{4}$$

where

$$h = (p-1) \cdot u + r - p_h$$
$$w = (q-1) \cdot v + s - p_w$$

k は出力テンソルの特徴量の次元，p, q は空間次元でのカーネルの移動量を表す添字である．式 (4) は，特定の p, q に対してカーネルが図 4 の位置に移動し，同じ場所に位置する入力テンソルの要素との内積をとる操作を意味している．

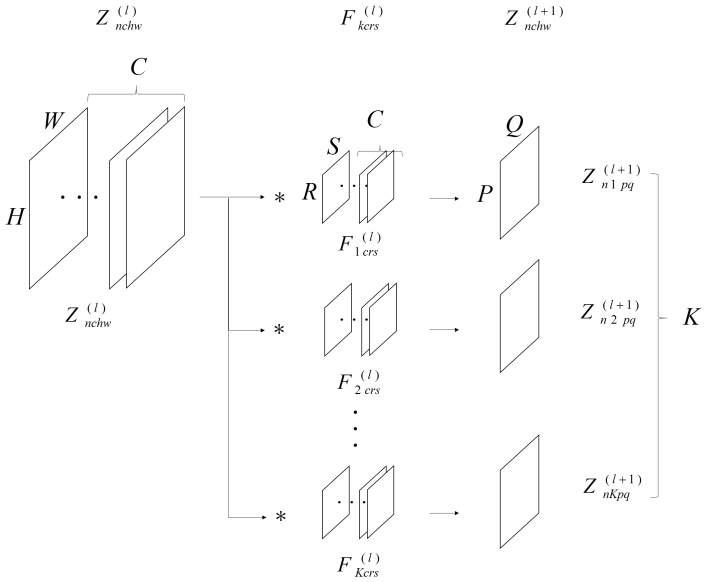

図 3　畳み込み層の概要.

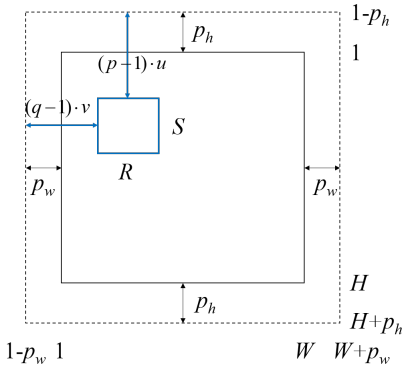

図 4　p, q に対するカーネルの移動位置.

■逆伝播　次に，畳み込み層の逆伝播計算を定式化する．まず，カーネルの要素およびバイアスによる誤差関数の偏微分は，定義に従って次のように計算される．

$$\frac{\partial E}{\partial F_{kcrs}^{(l)}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial E_n}{\partial F_{kcrs}^{(l)}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{p=1}^{P}\sum_{q=1}^{Q}\frac{\partial E_n}{\partial Z_{nkpq}^{(l+1)}}\cdot\frac{\partial Z_{nkpq}^{(l+1)}}{\partial F_{kcrs}^{(l)}}$$
$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{p=1}^{P}\sum_{q=1}^{Q}\Delta_{nkpq}^{(l+1)}\cdot Z_{nchw}^{(l)} \tag{5}$$

where
$$h = (p-1)\cdot u + r - p_h$$
$$w = (q-1)\cdot v + s - p_w$$

$$\frac{\partial E}{\partial b^{(l)}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial E_n}{\partial b^{(l)}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{p=1}^{P}\sum_{q=1}^{Q}\frac{\partial E_n}{\partial Z_{nkpq}^{(l+1)}}\cdot\frac{\partial Z_{nkpq}^{(l+1)}}{\partial b^{(l)}}$$
$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{p=1}^{P}\sum_{q=1}^{Q}\Delta_{nkpq}^{(l+1)}\cdot 1 \tag{6}$$

式 (5)，(6) において微分のチェインルール $\dfrac{\partial E_n}{\partial(\cdot)} = \displaystyle\sum_p\sum_q\frac{\partial E_n}{\partial Z_{nkpq}^{(l+1)}}\cdot\frac{\partial Z_{nkpq}^{(l+1)}}{\partial(\cdot)}$ を適用する際は，$F_{kcrs}^{(l)}$ や $b^{(l)}$ が全ての p,q に対する $Z_{nkpq}^{(l+1)}$ の計算に用いられていることを考慮し，$p=1...P, q=1...Q$ に関する総和を取っている．

次に，入力テンソルの要素による誤差関数の偏微分は，次のように計算される．

$$\Delta_{nchw}^{(l)} = \frac{\partial E_n}{\partial Z_{nchw}^{(l)}} = \sum_{k=1}^{K}\sum_{p=P_{\min}}^{P_{\max}}\sum_{q=Q_{\min}}^{Q_{\max}}\frac{\partial E_n}{\partial Z_{nkpq}^{(l+1)}}\cdot\frac{\partial Z_{nkpq}^{(l+1)}}{\partial Z_{nchw}^{(l)}}$$
$$= \sum_{k=1}^{K}\sum_{p=P_{\min}}^{P_{\max}}\sum_{q=Q_{\min}}^{Q_{\max}}\Delta_{nkpq}^{(l+1)}\cdot F_{kcrs}^{(l)} \tag{7}$$

where
$$r = h - (p-1)\cdot u + p_h$$
$$s = w - (q-1)\cdot v + p_w$$
$$P_{\min} = \max(1,\ \lceil\frac{h-R+p_h}{u}+1\rceil),\quad P_{\max} = \min(P,\ \lfloor\frac{h-1+p_h}{u}+1\rfloor) \tag{8}$$
$$Q_{\min} = \max(1,\ \lceil\frac{w-S+p_w}{v}+1\rceil),\quad Q_{\max} = \min(Q,\ \lfloor\frac{w-1+p_w}{v}+1\rfloor) \tag{9}$$

式 (7) において微分のチェインルール $\dfrac{\partial E_n}{\partial Z_{nchw}^{(l)}}\displaystyle\sum_{k=1}^{K}\sum_p\sum_q\frac{\partial E_n}{\partial Z_{nkpq}^{(l+1)}}\cdot\frac{\partial Z_{nkpq}^{(l+1)}}{\partial Z_{nchw}^{(l)}}$ を適用する際は，次のことに注意する必要がある．ある特定の入力テンソルの要素 $Z_{nchw}^{(l)}$ は，全ての p,q に対する $Z_{nkpq}^{(l+1)}$ の計算に持ちいられるわけではなく，p,q の値に応じて移動するカーネルがこの要素に重なっている時のみ計算に用いられる．したがって，チェインルールを適用する際は，カーネルとこの要素が重なる p,q の範囲で総和を取る必要がある．式 (8)，(9) はそのような p,q の範囲を示しているものであるが，これは次のようにして求めることができる．

今，ある特定の $Z_{nchw}^{(l)}$ に対して，p が変化しながらカーネル $F_{kcrs}^{(l)}$ が移動する場合を考える．$F_{kcrs}^{(l)}$ は 4 階のテンソルであるが，ここでは k,c を固定し，r,s のみが変化する 2 次元の行列であると考える．カーネルは，p の値に応じて図 5 のように $Z_{nkhw}^{(l)}$ と重なる場合と重ならない場合がある．今，カーネルの 1 列目の座標値

を (row_1, col)，R 列目の座標値を (row_R, col) とおくと，row_1，row_R は図 4 を参照するとそれぞれ次のように表される．

$$row_1 = (p-1) \cdot u + 1 - p_w \tag{10}$$
$$row_R = (p-1) \cdot u + R - p_w \tag{11}$$

このとき，カーネルが $Z^{(l)}_{nchw}$ に重なるためには，図 5 のように $row_1 \leq h \leq row_R$ である必要があり，これを満たす p の条件は次のようになる．

$$\frac{h - R + p_h}{u} + 1 \leq p \leq \frac{h - 1 + p_h}{u} + 1 \tag{12}$$

p は，この条件を満たす 1 から P までの整数であることから，式 (8) が導かれる．また，q についても同様に考えることにより，式 (9) が得られる．

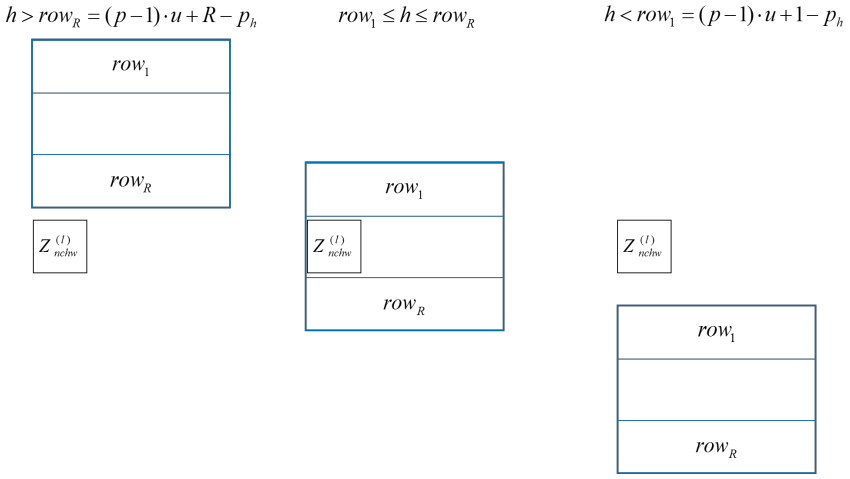

図 5　p に応じたカーネルの移動と $Z^{(l)}_{nchw}$ の重なり方．

2.3　マックスプーリング層

■順伝播　マックスプーリング層は，畳み込み層のようにあるサイズのカーネルを空間方向に移動させながら，カーネルと重なる入力テンソルの要素のうち最大値を出力していく層である．畳み込み層と同様にカーネルサイズやストライド・パディングの記号を定めると，順伝播計算は次のように表される．プーリングにおいては，畳み込み層との操作と異なり特徴量次元 k, c の方向には縮約されないことに注意する．

$$Z^{(l+1)}_{nkpq} = \max_{r,s} \{ Z^{(l)}_{nkhw} \ ; \ 1 \leq r \leq R, 1 \leq s \leq S \} \tag{13}$$

where

$$h = (p-1) \cdot u + r - p_h$$
$$w = (q-1) \cdot v + s - p_v$$

■逆伝播　プーリング層には最適化の対象となるパラメタがないため，逆伝播計算には $\Delta^{(l)}_{nchw}$ のみが定義され，次のように表される．

$$\begin{aligned}\Delta^{(l)}_{nchw} = \frac{\partial E_n}{\partial Z^{(l)}_{nchw}} &= \sum_{p=P_{\min}}^{P_{\max}} \sum_{q=Q_{\min}}^{Q_{\max}} \frac{\partial E_n}{\partial Z^{(l+1)}_{ncpq}} \cdot \frac{\partial Z^{(l+1)}_{ncpq}}{\partial Z^{(l)}_{nchw}} \\ &= \sum_{p=P_{\min}}^{P_{\max}} \sum_{q=Q_{\min}}^{Q_{\max}} \Delta^{(l+1)}_{ncpq} \cdot B^{(l)}_{ncpqhw} \end{aligned} \tag{14}$$

where

$$B^{(l)}_{ncpqhw} = \begin{cases} 1 & (Z^{(l+1)}_{ncpq} = Z^{(l)}_{nchw}) \\ 0 & (otherwise) \end{cases}$$

$$r = h - (p-1) \cdot u + p_h$$

$$s = w - (q-1) \cdot v + p_w$$

$$P_{\min} = \max(1, \ \lceil \frac{h - R + p_h}{u} + 1 \rceil), \quad P_{\max} = \min(P, \ \lfloor \frac{h - 1 + p_h}{u} + 1 \rfloor$$

$$Q_{\min} = \max(1, \ \lceil \frac{w - S + p_w}{v} + 1 \rceil), \quad Q_{\max} = \min(Q, \ \lfloor \frac{w - 1 + p_w}{v} + 1 \rfloor)$$

上式において，チェインルールの適用時における p, q の範囲は，畳み込み層と同様に考えることができる．また，$\dfrac{\partial Z^{(l+1)}_{ncpq}}{\partial Z^{(l)}_{nchw}}$ の値は，$Z^{(l+1)}_{ncpq} = Z^{(l)}_{nchw}$ のときに 1，それ以外の場合は 0 となる．

2.4　活性化層

多層パーセプトロンの定式化では，全結合層において前の層からの出力のアフィン変換と活性化関数の作用をまとめて一つの層の計算としていたが，ここではこれらの計算をそれぞれアフィン変換層，活性化層として個別の層に分離し，定式化を行う．

■順伝播　活性化層は，入力テンソルの個々の要素に活性化関数を作用させる働きを持つ層であり，順伝播計算は次のように表される．

$$Z^{(l+1)}_{nkpq} = \phi\{Z^{(l)}_{nkpq}\} \tag{15}$$

■逆伝播　活性化層もまた最適化の対象となるパラメタがないため，逆伝播計算には $\Delta^{(l)}_{nchw}$ のみが次のように定義される．

$$\Delta^{(l)}_{nchw} = \frac{\partial E_n}{\partial Z^{(l)}_{nchw}} = \frac{\partial E_n}{\partial Z^{(l+1)}_{nchw}} \cdot \frac{\partial Z^{(l+1)}_{nchw}}{\partial Z^{(l)}_{nhcw}} = \Delta^{(l+1)}_{nchw} \cdot \phi'(Z^{(l)}_{nchw}) \tag{16}$$

2.5　アフィン変換層

畳み込みニューラルネットワークにおいて，アフィン変換層（全結合層）は通常，畳み込み層やプーリング層の階層によって特徴量が抽出された後，この特徴量に対する多層パーセプトロンとして配置される．前の層までのように，入力テンソル $Z^{(l)}_{nchw}$ の位置 (添字 h や w) やチャンネル数 (添字 c) に関係なく全ての要素が全結合するため，これらの添字を区別しないこととして入力テンソルを $Z^{(l)}_{nj}(1 \leq j \leq H \cdot W \cdot C)$ と表す．

■順伝播　アフィン変換層の順伝播計算は，多層パーセプトロンと同様に行列の線形変換であり，次のように表される．

$$Z^{(l+1)}_{ni} = \sum_{j=1}^{H \cdot W \cdot C} w^{(l)}_{ij} \cdot Z^{(l)}_{nj} + b^{(l)}_i \tag{17}$$

■逆伝播　アフィン変換層では，線形変換時の重み $w^{(l)}_{ij}$ とバイアス $b^{(l)}_i$ が最適化の対象となるパラメタであり，誤差関数に対するこれらのパラメタによる偏微分は次のように表される．

$$\frac{\partial E}{\partial w^{(l)}_{ij}} = \frac{1}{N} \sum_{n=1}^{N} \frac{\partial E_n}{\partial w^{(l)}_{ij}} = \frac{1}{N} \sum_{n=1}^{N} \frac{\partial E_n}{\partial Z^{(l+1)}_{ni}} \cdot \frac{\partial Z^{(l+1)}_{ni}}{\partial w^{(l)}_{ij}} = \frac{1}{N} \sum_{n=1}^{N} \Delta^{(l+1)}_{ni} \cdot Z^{(l)}_{nj} \tag{18}$$

$$\frac{\partial E}{\partial b_i^{(l)}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial E_n}{\partial b_i^{(l)}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial E_n}{\partial Z_{ni}^{(l+1)}}\cdot\frac{\partial Z_{ni}^{(l+1)}}{\partial b_i^{(l)}} = \frac{1}{N}\sum_{n=1}^{N}\Delta_{ni}^{(l+1)}\cdot 1 \tag{19}$$

次に，入力テンソルの要素による誤差関数の偏微分は，次のように計算される．

$$\Delta_{nj}^{(l)} = \frac{\partial E_n}{\partial Z_{nj}^{(l)}} = \sum_{i=1}^{K\cdot P\cdot Q}\frac{\partial E_n}{\partial Z_{ni}^{(l+1)}}\cdot\frac{\partial Z_{ni}^{(l+1)}}{\partial Z_{nj}^{(l)}} = \sum_{i=1}^{K\cdot P\cdot Q}\Delta_{ni}^{(l+1)}\cdot w_{ij}^{(l)} \tag{20}$$