Gradient Calculation I

- For convolutional layers, recall we had

$$
\begin{aligned}
& \text{vec}(S^{m,i}) \\
= & \left(\phi(\text{pad}(Z^{m,i}))^T \otimes \mathcal{I}_{d^{m+1}}\right) \text{vec}(W^m) + \\
& (\mathbb{1}_{a^m_{\text{conv}} b^m_{\text{conv}}} \otimes \mathcal{I}_{d^{m+1}}) \boldsymbol{b}^m
\end{aligned}
$$

Gradient Calculation II

Thus

$$\begin{aligned}
\frac{\partial \xi_i}{\partial \text{vec}(W^m)^T} &= \frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \frac{\partial \text{vec}(S^{m,i})}{\partial \text{vec}(W^m)^T} \\
=& \frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \left(\phi(\text{pad}(Z^{m,i}))^T \otimes \mathcal{I}_{d^{m+1}} \right) \\
=& \text{vec} \left(\frac{\partial \xi_i}{\partial S^{m,i}} \phi(\text{pad}(Z^{m,i}))^T \right)^T \qquad (1)
\end{aligned}$$

Gradient Calculation III

where (1) is from

$$\begin{aligned} \text{vec}(AB)^T &= \text{vec}(B)^T(\mathcal{I} \otimes A^T) & (2) \\ &= \text{vec}(A)^T(B \otimes \mathcal{I}) & (3) \end{aligned}$$

- We applied chain rule here
- Note that we define

$$\frac{\partial \boldsymbol{y}}{\partial (\boldsymbol{x})^T} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_{|x|}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{|y|}}{\partial x_1} & \cdots & \frac{\partial y_{|y|}}{\partial x_{|x|}} \end{bmatrix}, \qquad (4)$$

Gradient Calculation IV

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are column vectors, and $|\boldsymbol{x}|$, $|\boldsymbol{y}|$ are their lengths.

- Thus if

$$\boldsymbol{y} = A\boldsymbol{x}$$

then

$$y_1 = A_{11}x_1 + \cdots + A_{1|\boldsymbol{x}|}x_{|\boldsymbol{x}|}$$

and

$$\frac{\partial \boldsymbol{y}}{\partial (\boldsymbol{x})^T} = \begin{bmatrix} A_{11} & A_{12} & \cdots \\ A_{21} & & \\ \vdots & & \end{bmatrix} = A$$

Gradient Calculation V

- Similarly

$$\begin{aligned}\frac{\partial \xi_i}{\partial (\boldsymbol{b}^m)^T} &= \frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \frac{\partial \text{vec}(S^{m,i})}{\partial (\boldsymbol{b}^m)^T} \\ &= \frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \left(\mathbb{1}_{a^m_{\text{conv}} b^m_{\text{conv}}} \otimes \mathcal{I}_{d^{m+1}} \right) \\ &= \text{vec}\left(\frac{\partial \xi_i}{\partial S^{m,i}} \mathbb{1}_{a^m_{\text{conv}} b^m_{\text{conv}}} \right)^T, \qquad (5)\end{aligned}$$

where (5) is from (3).

Gradient Calculation VI

- To calculate (1), $\phi(\text{pad}(Z^{m,i}))$ has been available from the forward process of calculating the function value.
- In (1) and (5), $\partial \xi_i / \partial S^{m,i}$ is also needed
- We will show that it can be obtained by a backward process.

Calculation of $\partial \xi_i / \partial S^{m,i}$ I

- What we will do is to assume that

$$\frac{\partial \xi_i}{\partial Z^{m+1,i}}$$

  is available
- Then we show details of calculating

$$\frac{\partial \xi_i}{\partial S^{m,i}} \text{ and } \frac{\partial \xi_i}{\partial Z^{m,i}}$$

  for layer m.
- Thus a back propagation process

Calculation of $\partial \xi_i / \partial S^{m,i}$ II

- We have the following workflow.

$$
\begin{aligned}
Z^{m,i} &\leftarrow \text{padding} \leftarrow \text{convolution} \leftarrow \sigma(S^{m,i}) \\
&\leftarrow \text{pooling} \leftarrow Z^{m+1,i}.
\end{aligned} \tag{6}
$$

- From chain rule,

$$
\frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} = \frac{\partial \xi_i}{\partial \text{vec}(\sigma(S^{m,i}))^T} \frac{\partial \text{vec}(\sigma(S^{m,i}))}{\partial \text{vec}(S^{m,i})^T}
$$

Calculation of $\partial \xi_i / \partial S^{m,i}$ III

- If σ is a scalar function, then

$$\frac{\partial \text{vec}(\sigma(S^{m,i}))}{\partial \text{vec}(S^{m,i})^T}$$

is a squared diagonal matrix of

$$|\text{vec}(S^{m,i})| \times |\text{vec}(S^{m,i})|$$

Calculation of $\partial \xi_i / \partial S^{m,i}$ IV

- We further assume that the RELU activation function is used. Recall that we assume

$$\sigma'(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

though $\sigma(x)$ is not differentiable at $x = 0$

Calculation of $\partial \xi_i / \partial S^{m,i}$ V

- We can define

$$I[S^{m,i}]_{(p,q)} = \begin{cases} 1 & \text{if } S^{m,i}_{(p,q)} > 0, \\ 0 & \text{otherwise,} \end{cases}$$

and have

$$\frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} = \frac{\partial \xi_i}{\partial \text{vec}(\sigma(S^{m,i}))^T} \odot \text{vec}(I[S^{m,i}])^T$$

where $\odot$ is Hadamard product (i.e., element-wise products)

Calculation of $\partial \xi_i / \partial S^{m,i}$ VI

- Q: can we extend this to other scalar activation functions?
- Yes, the general form is

$$\frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} = \frac{\partial \xi_i}{\partial \text{vec}(\sigma(S^{m,i}))^T} \odot \text{vec}(\sigma'(S^{m,i}))^T$$

- Next,

Calculation of $\partial \xi_i / \partial S^{m,i}$ VII

$$
\begin{aligned}
& \frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \\
= & \frac{\partial \xi_i}{\partial \text{vec}(Z^{m+1,i})^T} \frac{\partial \text{vec}(Z^{m+1,i})}{\partial \text{vec}(\sigma(S^{m,i}))^T} \frac{\partial \text{vec}(\sigma(S^{m,i}))}{\partial \text{vec}(S^{m,i})^T} \\
= & \left(\frac{\partial \xi_i}{\partial \text{vec}(Z^{m+1,i})^T} \frac{\partial \text{vec}(Z^{m+1,i})}{\partial \text{vec}(\sigma(S^{m,i}))^T} \right) \odot \text{vec}(I[S^{m,i}])^T \\
= & \left(\frac{\partial \xi_i}{\partial \text{vec}(Z^{m+1,i})^T} P_{\text{pool}}^{m,i} \right) \odot \text{vec}(I[S^{m,i}])^T \qquad (7)
\end{aligned}
$$

Calculation of $\partial \xi_i / \partial S^{m,i}$ VIII

- Note that (7) is from

$$Z^{m+1,i} = \text{mat}(P_{\text{pool}}^{m,i} \text{vec}(\sigma(S^{m,i})))_{d^{m+1} \times a^{m+1} b^{m+1}}$$

- If a general scalar activation function is considered, (7) is changed to

$$\frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} = \left(\frac{\partial \xi_i}{\partial \text{vec}(Z^{m+1,i})^T} P_{\text{pool}}^{m,i} \right) \odot \text{vec}(\sigma'(S^{m,i}))^T$$

Calculation of $\partial \xi_i / \partial S^{m,i}$ IX

- In the end we calculate $\partial \xi_i / \partial Z^{m,i}$ and pass it to the previous layer.

Calculation of $\partial \xi_i / \partial S^{m,i}$ X

$$
\begin{aligned}
&\frac{\partial \xi_i}{\partial \text{vec}(Z^{m,i})^T} \\
=&\frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \frac{\partial \text{vec}(S^{m,i})}{\partial \text{vec}(\phi(\text{pad}(Z^{m,i})))^T} \frac{\partial \text{vec}(\phi(\text{pad}(Z^{m,i})))}{\partial \text{vec}(\text{pad}(Z^{m,i}))^T} \\
&\frac{\partial \text{vec}(\text{pad}(Z^{m,i}))}{\partial \text{vec}(Z^{m,i})^T} \\
=&\frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \left(\mathcal{I}_{a^m_{\text{conv}} b^m_{\text{conv}}} \otimes W^m \right) P^m_\phi P^m_{\text{pad}} \qquad (8) \\
=&\text{vec}\left((W^m)^T \frac{\partial \xi_i}{\partial S^{m,i}} \right)^T P^m_\phi P^m_{\text{pad}}, \qquad (9)
\end{aligned}
$$

Calculation of $\partial \xi_i / \partial S^{m,i}$ XI

where (8) is from

$$\begin{aligned}
&\text{vec}(S^{m,i}) \\
=&\left(\mathcal{I}_{a^m_{\text{conv}} b^m_{\text{conv}}} \otimes W^m\right) \text{vec}(\phi(\text{pad}(Z^{m,i})))+ \\
&\left(\mathbb{1}_{a^m_{\text{conv}} b^m_{\text{conv}}} \otimes \mathcal{I}_{d^{m+1}}\right) \boldsymbol{b}^m
\end{aligned}$$

and (9) is from (2).