For convolutional layers, recall we had

\[
\text{vec}(S_{m,i}) = (\phi(\text{pad}(Z_{m,i})))^T \otimes I_{d_{m+1}} \text{vec}(W^m) + \\
(I_{a_{\text{conv}}} b_{\text{conv}}^m \otimes I_{d_{m+1}})b^m
\]
Thus

\[
\frac{\partial \xi_i}{\partial \text{vec}(W^{m})^T} = \frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \cdot \frac{\partial \text{vec}(S^{m,i})}{\partial \text{vec}(W^{m})^T}
\]

\[
= \frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \left( \phi(\text{pad}(Z^{m,i}))^T \otimes \mathcal{I}_{d^{m+1}} \right)
\]

\[
= \text{vec} \left( \frac{\partial \xi_i}{\partial S^{m,i}} \phi(\text{pad}(Z^{m,i}))^T \right)^T
\]

(1)
where (1) is from

\[ \text{vec}(AB)^T = \text{vec}(B)^T (I \otimes A^T) \] (2)

\[ = \text{vec}(A)^T (B \otimes I) \] (3)

- We applied chain rule here
- Note that we define

\[
\frac{\partial y}{\partial (x)^T} = \begin{bmatrix}
\frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_{|x|}} \\
\vdots & \ddots & \vdots \\
\frac{\partial y_{|y|}}{\partial x_1} & \cdots & \frac{\partial y_{|y|}}{\partial x_{|x|}}
\end{bmatrix}, \quad (4)
\]
Gradient Calculation IV

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

Thus if

$$\mathbf{y} = A\mathbf{x}$$
	hen then

$$y_1 = A_{11}x_1 + \cdots + A_1|x| x_1$$

and

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

\[
\frac{\partial \xi_i}{\partial (b^m)^T} = \frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \frac{\partial \text{vec}(S^{m,i})}{\partial (b^m)^T} \\
= \frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} \left( \mathbb{1} a_{\text{conv}}^m b_{\text{conv}}^m \otimes I_{d^m+1} \right) \\
= \text{vec} \left( \frac{\partial \xi_i}{\partial S_{m,i}} \mathbb{1} a_{\text{conv}}^m b_{\text{conv}}^m \right)^T,
\]

where (5) is from (3).
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 $\frac{\partial \xi_i}{\partial S^{m,i}}$

- 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 $\partial \xi_i/\partial S^{m,i}$

- We have the following workflow.

$$Z^{m,i} \leftarrow \text{padding} \leftarrow \text{convolution} \leftarrow \sigma(S^{m,i})$$
$$\leftarrow \text{pooling} \leftarrow Z^{m+1,i}. \quad \text{(6)}$$

- From chain rule,

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

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

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

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

(7)
Calculation of $\partial \xi_i / \partial S^{m,i}$

- Note that (7) is from

$$Z^{m+1,i} = \text{mat}(P^m_i \text{vec}(\sigma(S^{m,i})))_{d_{m+1} \times a_{m+1} b_{m+1}}$$

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

$$\frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} = \left( \frac{\partial \xi_i}{\partial \text{vec}(Z^{m+1,i})^T} P^m_i \right) \odot \text{vec}(\sigma'(S^{m,i}))^T$$
In the end we calculate $\partial \xi_i / \partial S^{m,i}$ and pass it to the previous layer.
Calculation of $\partial \xi_i / \partial S_{m,i}$

\[
\frac{\partial \xi_i}{\partial \text{vec}(Z_{m,i})^T} = \frac{\partial \xi_i}{\partial \text{vec}(S_{m,i})^T} \frac{\partial \text{vec}(S_{m,i})}{\partial \text{vec}(\phi(\text{pad}(Z_{m,i})))^T} \frac{\partial \text{vec}(\phi(\text{pad}(Z_{m,i})))}{\partial \text{vec}(\text{pad}(Z_{m,i}))^T} \frac{\partial \text{vec}(\text{pad}(Z_{m,i}))}{\partial \text{vec}(Z_{m,i})^T} \\
= \frac{\partial \xi_i}{\partial \text{vec}(S_{m,i})^T} \left( \mathcal{I}_{a_{\text{conv}}b_{\text{conv}}} \otimes W^m \right) P^m \phi P^m_{\text{pad}} \\
= \text{vec} \left( (W^m)^T \frac{\partial \xi_i}{\partial S_{m,i}} \right)^T P^m \phi P^m_{\text{pad}},
\]

(8)
Calculation of $\frac{\partial \xi_i}{\partial S^m,i}$ XI

where (8) is from

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

$$= (\mathcal{I}_{a_{\text{conv}}^m b_{\text{conv}}^m} \otimes \mathcal{W}^m) \text{vec}(\phi(\text{pad}(Z^m,i))) + (\mathbf{1}_{a_{\text{conv}}^m b_{\text{conv}}^m} \otimes \mathcal{I}_{d_{m+1}^m}) b^m$$

and (9) is from (2).