Fully-connected Layers I

- For convolutional layers, earlier we obtained

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

and

$$\frac{\partial \xi_i}{\partial (\boldsymbol{b}^m)^T} = \text{vec}\left(\frac{\partial \xi_i}{\partial S^{m,i}} \mathbb{1}_{a^m_{\text{conv}} b^m_{\text{conv}}}\right)^T$$

Fully-connected Layers II

- By considering fully-connected layers as a special case, immediately we get

$$\frac{\partial \xi_i}{\partial \text{vec}(W^m)^T} = \text{vec}\left(\frac{\partial \xi_i}{\partial \boldsymbol{s}^{m,i}} (\boldsymbol{z}^{m,i})^T\right)^T \quad (1)$$

$$\frac{\partial \xi_i}{\partial (\boldsymbol{b}^m)^T} = \frac{\partial \xi_i}{\partial (\boldsymbol{s}^{m,i})^T} \quad (2)$$

Fully-connected Layers III

Similarly,

$$\frac{\partial \xi_i}{\partial (\boldsymbol{z}^{m,i})^T} = \left((W^m)^T \frac{\partial \xi_i}{\partial (\boldsymbol{s}^{m,i})} \right)^T \mathcal{I}_{n_m}$$
$$= \left((W^m)^T \frac{\partial \xi_i}{\partial (\boldsymbol{s}^{m,i})} \right)^T, \qquad (3)$$

where

$$\frac{\partial \xi_i}{\partial (\boldsymbol{s}^{m,i})^T} = \frac{\partial \xi_i}{\partial (\boldsymbol{z}^{m+1,i})^T} \odot I[\boldsymbol{s}^{m,i}]^T. \qquad (4)$$

Fully-connected Layers IV

- Finally, we check the initial values of the backward process.
- Assume that the squared loss is used and in the last layer we have an identity activation function
- Then

$$\frac{\partial \xi_i}{\partial \boldsymbol{z}^{L+1,i}} = 2(\boldsymbol{z}^{L+1,i} - \boldsymbol{y}^i), \text{ and } \frac{\partial \xi_i}{\partial \boldsymbol{s}^{L,i}} = \frac{\partial \xi_i}{\partial \boldsymbol{z}^{L+1,i}}.$$

Notes on RELU Activation + Max Pooling I

- Recall we said that in

$$\frac{\partial \xi_i}{\partial W^m} = \frac{\partial \xi_i}{\partial S^{m,i}} \phi(\text{pad}(Z^{m,i}))^T,$$

  $Z^{m,i}$ is available from the forward process

- Therefore

$$Z^{m,i}, \forall m$$

  are stored.

Notes on RELU Activation + Max Pooling II

- But we also need $S^{m,i}$ for

$$\frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} = \left(\frac{\partial \xi_i}{\partial \text{vec}(Z^{m+1,i})^T} P_{\text{pool}}^{m,i} \right) \odot \text{vec}(I[S^{m,i}])^T \tag{5}$$

- Do we need to store both $Z^{m,i}$ and $S^{m,i}$?

Notes on RELU Activation + Max Pooling III

- We can avoid storing $S^{m,i}, \forall m$ by replacing (5) with

$$\frac{\partial \xi_i}{\partial \text{vec}(S^{m,i})^T} = \left(\frac{\partial \xi_i}{\partial \text{vec}(Z^{m+1,i})^T} \odot \text{vec}(I[Z^{m+1,i}])^T \right) P_{\text{pool}}^{m,i}. \tag{6}$$

- Why? Let's look at the relation between $Z^{m+1,i}$ and $S^{m,i}$

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

Notes on RELU Activation + Max Pooling IV

- $Z^{m+1,i}$ is a "smaller matrix" than $S^{m,i}$ after max pooling
- That is, (5) is a "reverse mapping" of the pooling operation
- In (5),

$$\frac{\partial \xi_i}{\partial \text{vec}(Z^{m+1,i})^T} \times P_{\text{pool}}^{m,i} \tag{7}$$

generates a large zero vector and puts values of $\partial \xi_i / \partial \text{vec}(Z^{m+1,i})^T$ into positions selected earlier in the max pooling operation.

Notes on RELU Activation + Max Pooling V

- Then, element-wise multiplications of (7) and $I[S^{m,i}]^T$ are conducted.
- Positions not selected in the max pooling procedure are zeros after (7)
- They are still zeros after the Hadamard product between (7) and $I[S^{m,i}]^T$
- Thus, (5) and (6) give the same results.

Notes on RELU Activation + Max Pooling VI

- An illustration using our earlier example. This illustration was generated with the help of Cheng-Hung Liu in my group
- Recall an earlier max pooling example is

$$\text{image B} \quad \left[\begin{array}{cc|cc} 3 & 2 & 3 & 6 \\ 4 & 5 & 4 & 9 \\ \hline 2 & 1 & 2 & 6 \\ 3 & 4 & 3 & 2 \end{array}\right] \rightarrow \begin{bmatrix} 5 & 9 \\ 4 & 6 \end{bmatrix}$$

Notes on RELU Activation + Max Pooling VII

- The corresponding pooling matrix is

$$
P_{\text{pool}} = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{bmatrix}
$$

Notes on RELU Activation + Max Pooling VIII

- We have that

$$P_{\text{pool}}\text{vec}(\text{image}) = \begin{bmatrix} 5 \\ 4 \\ 9 \\ 6 \end{bmatrix} = \text{vec}(\begin{bmatrix} 5 & 9 \\ 4 & 6 \end{bmatrix})$$

Notes on RELU Activation + Max Pooling IX

- If using (5),

$$
\begin{aligned}
&\boldsymbol{v}^T P_{\text{pool}} \odot \text{vec}(I[S^m])^T \\
= &\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & v_1 & 0 & v_2 & 0 & 0 & 0 & 0 & 0 & v_3 & v_4 & 0 \end{bmatrix} \\
&\odot \\
&\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \\
= &\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & v_1 & 0 & v_2 & 0 & 0 & 0 & 0 & 0 & v_3 & v_4 & 0 \end{bmatrix}
\end{aligned}
$$

Notes on RELU Activation + Max Pooling X

- If using (6),

$$
\begin{aligned}
&(\boldsymbol{v}^T \odot \text{vec}(I[Z^{m+1}])^T)P_{\text{pool}} \\
=&(\boldsymbol{v}^T \odot \begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix})P_{\text{pool}} \\
=&\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & v_1 & 0 & v_2 & 0 & 0 & 0 & 0 & 0 & v_3 & v_4 & 0 \end{bmatrix}
\end{aligned}
$$

- So they are the same
- In the derivation we used the properties of
  - RELU activation function and
  - max pooling

Notes on RELU Activation + Max Pooling XI

to get

$$
\begin{aligned}
&\text{a } Z^{m+1,i} \text{ component} > 0 \text{ or not} \\
\Leftrightarrow &\text{the corresponding } \sigma'(S^{m,i}) = 1 \text{ or } 0
\end{aligned}
$$

- For general cases we might not be able to avoid storing $\sigma'(S^{m,i})$