Summary of Operations I

- We show convolutional layers only and the bias term is omitted.
- Also we assume that RELU activation and max pooling are used.
- Operations in order

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

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

Note that after (1), we change a vector \( \frac{\partial \xi_i}{\partial \text{vec}(\mathbf{S}^{m,i})^T} \) to a matrix \( \frac{\partial \xi_i}{\partial \mathbf{S}^{m,i}} \) because in (2) and (3), matrix form is needed.

In (1), information of the next layer is used.
Instead we can do

\[
\frac{\partial \xi_i}{\partial \text{vec}(Z^{m,i})^T} \circ \text{vec}(I[Z^{m,i}])^T
\]

in the end of the current layer

This becomes the information passed to the previous layer

Then only information in the current layer is used
Finally an implementation for one convolutional layer:

\[ \Delta \leftarrow \text{mat(vec(} \Delta \text{)}^T P_{\text{pool}}^{m,i}) \]

\[ \frac{\partial \xi_i}{\partial W^m} = \Delta \cdot \phi(\text{pad}(Z^{m,i}))^T \]

\[ \Delta \leftarrow \text{vec } \left( \left( W^m \right)^T \Delta \right)^T P_{\phi}^m P_{\text{pad}}^m \]

\[ \Delta \leftarrow \Delta \odot I[Z^{m,i}] \]

A sample segment of code in MATLAB
for m = LC : -1 : 1
    dXidS = reshape(vTP(param, model, net, m, dXidS, 'pool_gradient'),
                    model.ch_input(m+1), []);

    phiZ = padding_and_phiZ(model, net, m);
    net.dlossdW{m} = dXidS*phiZ';
    net.dlossdb{m} = dXidS*ones(model.wd_conv(m)*
                                model.ht_conv(m)*S_k, 1);

if m > 1
\[ v = \text{model.weight\{m\}}' * \text{dXidS}; \]
\[ \text{dXidS} = \text{vTP(model, net, m, num\_data, v, } \]
\[ \text{'phi\_gradient')}\];

\% vTP_pad
\[ \text{dXidS} = \text{reshape(dXidS, model.ch\_input(m)}, \]
\[ \text{model.ht\_pad(m), } \]
\[ \text{model.wd\_pad(m), []]); } \]
\[ p = \text{model.wd\_pad\_added(m)}; \]
\[ \text{dXidS} = \text{dXidS(:, p+1:p+model.ht\_input(m)}, \]
\[ p+1:p+\text{model.wd\_input(m), :}) \]
% activation function

\[ dXidS = \text{reshape}(dXidS, \text{model.ch_input}(m), []) \times (\text{net.Z}{m} > 0); \]

end
Storing $\phi(\text{pad}(Z^{m,i}))$

- From the above summary, we see that $\phi(\text{pad}(Z^{m,i}))$ is calculated twice in both forward and backward processes.
- If this expansion is expensive, we can store it.
- But memory is a concern as this is a huge matrix.
- So this setting of storing $\phi(\text{pad}(Z^{m,i}))$ trades space for time. It’s more suitable for CPU environments.
To see where the computational bottleneck is, it’s important to check the complexity of major operations.

Assume \( l \) is the number of data (for the case of calculating the whole gradient).

For stochastic gradient, \( l \) becomes the size of a mini-batch.
Complexity II

• Forward:

\[ S^{m,i} = \mathbf{W}^m \text{mat}(P^m_{\phi} P^m_{\text{pad}} \text{vec}(Z^{m,i})) \]
\[ = \mathbf{W}^m \phi(\text{pad}(Z^{m,i})) \]

\[ \phi(\text{pad}(Z^{m,i})) : \mathcal{O}(l \times h^m h^m d^m a_{\text{conv}}^m b_{\text{conv}}^m) \]  

\[ \mathbf{W}^m \phi(\cdot) : \mathcal{O}(l \times d^{m+1} h^m h^m d^m a_{\text{conv}}^m b_{\text{conv}}^m) \]

\[ Z^{m+1,i} = \text{mat}(P^m_{\text{pool}} \text{vec}(\sigma(S^{m,i}))) \]

\[ \mathcal{O}(l \times d^{m+1} a_{\text{conv}}^m b_{\text{conv}}^m) \]
\[ = \mathcal{O}(l \times h^m h^m d^{m+1} a^{m+1} b^{m+1}) \]
See also (4) as for pooling we also have a $\phi$ to generate sub-images

- Backward:

$$\Delta \leftarrow \text{mat}(\text{vec}(\Delta)^T P_{\text{pool}}^{m,i})$$

Size of $\Delta$ same as $S^{m,i}$ so cost is

$$O(l \times d^{m+1}a_m^{\text{conv}}b_m^{\text{conv}})$$

$$\frac{\partial \xi_i}{\partial W^m} = \Delta \phi(\text{pad}(Z^{m,i}))^T$$
\[ \mathcal{O}(l \times d^{m+1} a_{\text{conv}}^m b_{\text{conv}}^m h^m h^m d^m) \].

\[ \Delta \leftarrow \text{vec} \left( (W^m)^T \Delta \right)^T P_{\phi}^m P_{\text{pad}}^m \]

\[ (W^m)^T \Delta : \mathcal{O}(l \times h^m h^m d^m d^{m+1} a_{\text{conv}}^m b_{\text{conv}}^m) \]

\[ \text{vec}(\cdot) P_{\phi}^m : \mathcal{O}(l \times h^m h^m d^m a_{\text{conv}}^m b_{\text{conv}}^m) \] (5)

For (5) we convert a matrix of

\[ h^m h^m d^m \times a_{\text{conv}}^m b_{\text{conv}}^m \]

to a smaller matrix

\[ d^m \times a_{\text{pad}}^m b_{\text{pad}}^m \]
We see that matrix-matrix products are the bottleneck
If so, why check others?
The issue is that matrix-matrix products may be better optimized
Recall we have

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

We note that

\[ P_{\text{pool}}^{m,i} \]

is not a constant 0/1 matrix

It depends on \( \sigma(S^{m,i}) \) to decide the positions of 0 and 1.
Thus like the RELU activation function, max pooling is another place to cause that $f(\theta)$ is not differentiable.

However, it is almost differentiable around the current point.

Consider

$$f(A) = \max \left( \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right)$$

and

$$A_{11} > A_{12}, A_{21}, A_{22}$$
Then
\[ \nabla f(A) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{ at } A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \]

This explains why we can use $P_{\text{pool}}^{m,i}$ in function and gradient evaluations.