Variables

- $X_{ij}, \quad 0 \leq i < N, \quad 0 \leq j < K^{(0)}$
  Inputs: N row of input samples with $K+1$ features. The first feature is always 1 for bias.

- $W_{ij}^{(l)}, \quad 0 \leq l < L, \quad 0 \leq i < K^{(l)}, \quad 0 \leq j < K^{(l+1)}$
  Network Weights: Where l is the layer in the network. j is the number of neurons in that layer, i is number of inputs from previous layer. Note $K_0 = K+1$ from the input.

- $A^l = Z^l W^l, \quad 0 \leq l < L$
  Activation of the l-th layer. Note that $Z^0 = X$

- $Z^l = f^{l-1}(A^{l-1}), \quad 1 \leq l < L$
  Output for layer l.

- $f^l(A^l) = f^l(Z^l W^l)$
  Filter function for layer l

- $\tilde{Y}_{ij}, \quad 0 \leq i < N, \quad 0 \leq j < K^{(L)}$
  Note $Y = Z^L$

- $T_{ij}, \quad 0 \leq i < N, \quad 0 \leq j < K^{(L)}$
  Target value used to compared with Y

- $E(\tilde{Y}, T) = E(Z^{(L)})$
  Error function

Note that Z goes from $Z^{(0)} \dots Z^{(L)}$, f goes from $f^{(0)} \dots f^{(L-1)}$, A goes from $A^{(0)} \dots A^{(L-1)}$ and W goes from $W^{(0)} \dots W^{(L-1)}$

For each layer l by chain rule:

$$\frac{\partial E}{\partial W_{ij}^{l-1}} = \sum_{s,t} \frac{\partial E}{\partial Z_{st}^l} \frac{\partial Z_{st}^l}{\partial W_{ij}^{l-1}}$$

Since

$$\begin{aligned}
Z_{st}^l &= f^{l-1}(A_{st}^{l-1}) = f^{l-1}\left(\sum_k Z_{sk}^{l-1} W_{kt}^{l-1}\right) \\
\frac{\partial Z_{st}^l}{\partial W_{ij}^{l-1}} &= \frac{\partial f^{l-1}}{\partial W}\left(\sum_k Z_{sk}^{l-1} W_{kt}^{l-1}\right) \frac{\partial \sum_k Z_{sk}^{l-1} W_{kt}^{l-1}}{\partial W_{ij}^{l-1}} \\
&= \frac{\partial f^{l-1}}{\partial W}\left(\sum_k Z_{sk}^{l-1} W_{kt}^{l-1}\right) Z_{sk}^{l-1} \delta_j^t \delta_i^k \\
&= \frac{\partial f^{l-1}(A_{st}^{l-1})}{\partial W} Z_{sk}^{l-1} \delta_j^t \delta_i^k
\end{aligned}$$

Substitute the 2nd term, we have

$$\begin{aligned}
\frac{\partial E}{\partial W_{ij}^{l-1}} &= \sum_{s,t} \frac{\partial E}{\partial Z_{st}^{l}} \frac{\partial f^{l-1}(A_{st}^{l-1})}{\partial W} Z_{sk}^{l-1} \delta_j^t \delta_i^k \\
&= \sum_{s} \frac{\partial E}{\partial Z_{sj}^{l}} \left[\partial_W f^{l-1}(A^{l-1})\right]_{sj} Z_{si}^{l-1}
\end{aligned}$$

For the other compnent

$$\begin{aligned}
\frac{\partial E}{\partial Z_{sj}^{l}} &= \sum_{u,v} \frac{\partial E}{\partial Z_{uv}^{l+1}} \frac{\partial Z_{uv}^{l+1}}{\partial Z_{sj}^{l}} \\
&= \sum_{u,v} \frac{\partial E}{\partial Z_{uv}^{l+1}} \frac{\partial f^{l}\left(\sum_p Z_{up}^{l} W_{pv}^{l}\right)}{\partial Z_{sj}^{l}} \\
&= \sum_{v} \frac{\partial E}{\partial Z_{sv}^{l+1}} \left[\partial_Z f^{l}(A^{l})\right]_{sv} W_{jv}^{l}
\end{aligned}$$