Derivations for paper

Backpropagation formulas:

$$\frac{\partial E}{\partial w_{ij}} = \delta_j o_i$$

$$\delta_j = \varphi'(net_j) \sum_{k \in \text{output nodes}} \delta_k w_{jk}$$

where $\delta_j = \frac{\partial E}{\partial o_j} \frac{\partial o_j}{\partial net_j}$

1) $\delta^{i,s} \equiv$ softmax error at node i

x^i → y^i output

$y^i = \text{softmax}(W_s x^i)$

$$\Rightarrow \delta^{i,s} = \underbrace{\varphi'(net_i)}_{f'(x^i)} \otimes \underbrace{\delta^{s+1}}_{y^i - t^i} \underbrace{W_{s+1}}_{\frac{\partial (W_s x^i)}{\partial x_i}}$$

$$\Rightarrow \delta^{i,s} = f'(x^i) \otimes (y^i - t^i) W_s$$

$$= W_s^T (y^i - t^i) \otimes f'(x^i)$$

Derivative with respect to V at node p_2:

$$\text{Using } \frac{\partial E}{\partial w_{ij}} = \delta_j o_i = \delta_j \frac{\partial net_j}{\partial w_{ij}}$$

Let $j = p_2$; $i \in \{q, p_1\}$

Let $w_{ij} = V$

$$\Rightarrow \frac{\partial E^{p_2}}{\partial V_{\alpha\beta p}} = \delta^{p_2, com} \frac{\partial r(a, p_a)}{\partial V_{\alpha\beta p}}$$

where $r(a, p_1) = \begin{bmatrix} a \\ p_1 \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} a \\ p_1 \end{bmatrix} + W \begin{bmatrix} a \\ p_1 \end{bmatrix}$

(Let $v = a \oplus p_1$)

or $r(a, p_1)_\gamma = v^\alpha V_{\alpha\beta\gamma} v^\beta + W_{\gamma\lambda} v^\lambda$

$$\Rightarrow \frac{\partial r(a, p_1)_\gamma}{\partial V_{\alpha\beta\lambda}} = \delta^\lambda_\gamma v^\alpha v^\beta$$

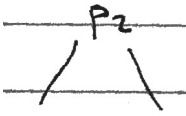
$$\Rightarrow \frac{\partial \xi^{p_2}}{\partial V_{\alpha\beta\lambda}} = \delta^{p_2, com}{}_\gamma \delta^\lambda_\gamma v^\alpha v^\beta$$

$$= \delta^{p_2, com}{}_\lambda v^\alpha v^\beta$$

$(v = a \oplus p_1)$

eg. $v^\alpha = a^\alpha + p_1{}^\alpha$

Error messages for the two children of p_2



By the backpropagation formula,

$$\delta^{a, com} = f'(a) \sum_{k \in \text{output nodes}} \delta_k \frac{\partial net_k}{\partial o_j}$$

$$\Rightarrow \delta^{a, com} = f'(a) \delta^{p_2, com} \frac{\partial r(a, p_1)}{\partial a} + f'(a) \delta^{a, s}$$

Similarly,

$$\delta^{p_1, com} = f'(p_1) \delta^{p_2, com} \frac{\partial r(a, p_1)}{\partial p_1} + f'(p_1) \delta^{p_1, s}$$

$$r(a, p_1)_\gamma = v^\alpha v^\beta V_{\alpha\beta\gamma} + W_{\gamma\lambda} v^\lambda$$

$(v = a \oplus p_1)$

$$= (a^\alpha + p_1{}^\alpha)(a^\beta + p_1{}^\beta) V_{\alpha\beta\gamma} + W_{\gamma\lambda}(a^\lambda + p_1{}^\lambda)$$

$$\Rightarrow \frac{\partial \Gamma(a, p_1 | \gamma}{\partial a_\alpha} = W_{\gamma\alpha} + v^{\beta} V_{\alpha\beta\gamma} + v^{\beta} V_{\beta\alpha\gamma}$$

(same for $\frac{\partial \Gamma(a, p_1|\gamma}{\partial p_1 \alpha}$)

$$= W + (V^{[12]} + V^{[13]})^T \begin{bmatrix} a \\ p_1 \end{bmatrix}$$

(in paper's notation)

Simplifying,

$$\delta^{a,com} = f'(a) \otimes \left\{ \delta^{p_2,com} \frac{\partial \Gamma(a,p_1)}{\partial a} + \delta^{a,s} \right\}$$

$$\delta^{p_2,com} = f'(p_1) \otimes \left\{ \delta^{p_2,com} \frac{\partial \Gamma(a,p_1)}{\partial p_1} + \delta^{p_2,s} \right\}$$

Or in the paper's notation,

$$\delta^{i,s} \longrightarrow \delta f'(k_i) \otimes \delta^{i,s}$$

$$\delta^{a,com} = f'(a) \otimes \delta^{p_2,com} \frac{\partial \Gamma(a,p_1)}{\partial a} + \delta^{a,s}$$

$$\delta^{p_2,com} = f'(p_1) \otimes \delta^{p_2,com} \frac{\partial \Gamma(a,p_1)}{\partial p_1} + \delta^{p_2,s}$$

where $\frac{\partial \Gamma(a,p_1)}{\partial a} = \frac{\partial \Gamma(a,p_2)}{\partial p_1}$

$$= W + (V^{[12]} + V^{[13]})^T \begin{bmatrix} a \\ p_1 \end{bmatrix}$$

Full derivative for ξ w.r.t V

By the chain rule,

$$\frac{\partial \xi}{\partial V} = \sum_{\text{nodes } i} \frac{\partial \xi}{\partial x_i} \frac{\partial x_i}{\partial \text{net}_i} \frac{\partial \text{net}_i}{\partial V}$$

$$= \sum_{\text{nodes } i} \delta_{i,\text{com}} \frac{\partial \text{net}_i}{\partial V}$$

$\frac{\partial \text{net}_i}{\partial V} \neq 0$ only for $i \in \{p_1, p_2\}$

$$\Rightarrow \frac{\partial \xi}{\partial V} = \delta_{p_1,\text{com}} \frac{\partial \text{net}_{p_1}}{\partial V} + \delta_{p_2,\text{com}} \frac{\partial \text{net}_{p_2}}{\partial V}$$

$\frac{\partial \text{net}_{p_1}}{\partial V}$: $\frac{\partial r(b,c)}{\partial V} = \begin{bmatrix} b \\ c \end{bmatrix} \begin{bmatrix} \frac{b}{c} \end{bmatrix}^T$ (computed earlier)

$\frac{\partial \text{net}_{p_2}}{\partial V}$: $\frac{\partial \xi^{p_2}}{\partial V}$ (computed earlier)

$$\Rightarrow \frac{\partial \xi}{\partial V^{\alpha\beta\lambda}} = \delta_{p_1,\text{com}}\, \alpha \left(b^{\beta} + c^{\beta}\right)\left(b^{\lambda} + c^{\lambda}\right) + \frac{\partial \xi^{p_2}}{\partial V^{\alpha\beta\lambda}}$$