Backprop through Softmax based on Input $\frac{\partial L}{\partial a^{[L]}}$

$$\frac{\partial L}{\partial a_k^{[L]}} = -\frac{1}{a_k^{[L]}}\delta_{k,y}$$

$$\frac{\partial L}{\partial z_j^{[L]}} = \sum_k \frac{\partial L}{\partial a_k^{[L]}} \frac{\partial a_k^{[L]}}{\partial z_j^{[L]}} \qquad a_k^{[L]} = \frac{e^{z_k^{[L]}}}{\sum_\ell e^{z_\ell^{[L]}}}$$

$$\frac{\partial a_k^{[L]}}{\partial z_j^{[L]}} = a_k^{[L]}\delta_{kj} - a_k^{[L]} a_j^{[L]} = a_k^{[L]}\left(\delta_{kj} - a_j^{[L]}\right)$$

$$\frac{\partial L}{\partial z_j^{[L]}} = -\sum_k a_k^{[L]}\left(\delta_{kj} - a_j^{[L]}\right) \frac{1}{a_k^{[L]}}\delta_{ky} = -\left(\delta_{jy} - a_j^{[L]}\right) = \left(a_j^{[L]} - \delta_{jy}\right)$$

$$\frac{\partial L}{\partial z_j^{[L]}} = \sum_k \left(a_k^{[L]}\delta_{kj} - a_k^{[L]} a_j^{[L]}\right) \frac{\partial L}{\partial a_k^{[L]}} = a_j^{[L]} \frac{\partial L}{\partial a_j^{[L]}} - \sum_k \left(a_k^{[L]} \frac{\partial L}{\partial a_k^{[L]}}\right) a_j^{[L]}$$

$$\boxed{\frac{\partial L}{\partial z^{[L]}} = a^{[L]} * \frac{\partial L}{\partial a^{[L]}} - \left(a^{[L]} \cdot \frac{\partial L}{\partial a^{[L]}}\right) a^{[L]}}$$

multiply elementwise

sum over the units in layer, not over samples in minibatch

$\frac{\partial L}{\partial a^{[L]}}, a^{[L]} \sim (n,m)$

$a^{[L]} \cdot \frac{\partial L}{\partial a^{[L]}} \sim (1,m)$

$\left(\text{sum}\left(a^{[L]} * \frac{\partial L}{\partial a^{[L]}}, \text{axis}=0\right)\right)$