Cross-Entropy Loss

$$\mathcal{L}(\vec{y}, \hat{\vec{y}}) = -\sum_i y_i \log \hat{y}_i$$

$$\frac{\partial \mathcal{L}}{\partial \hat{\vec{y}}} = \boxed{\frac{-\vec{y}}{\hat{\vec{y}}}} \quad \text{← elementwise division.}$$

Simple enough. Now we need the gradient of softmax:

Softmax

$$\hat{\vec{y}} = \text{softmax}(\vec{a})$$

$\frac{\partial \hat{\vec{y}}}{\partial \vec{a}}$ is a matrix $= \begin{bmatrix} \frac{\partial y_1}{\partial a_1} & \cdots & \frac{\partial y_1}{\partial a_i} & \cdots & \frac{\partial y_1}{\partial a_n} \\ \vdots & & \vdots & & \vdots \\ \frac{\partial y_i}{\partial a_1} & \cdots & \frac{\partial y_i}{\partial a_j} & \cdots & \frac{\partial y_i}{\partial a_n} \\ \vdots & & \vdots & & \vdots \\ \frac{\partial y_n}{\partial a_1} & \cdots & \frac{\partial y_n}{\partial a_j} & \cdots & \frac{\partial y_n}{\partial a_n} \end{bmatrix}$

There are really only two things to find here. $\frac{\partial y_i}{\partial a_j}$ when $i \neq j$ and $i = j$.

Let $\Sigma = \sum_i e^{a_i}$ so

$$\text{softmax}(a_i) = \frac{e^{a_i}}{\Sigma}$$

Then: $\frac{\partial y_i}{\partial a_j} = \frac{\partial}{\partial a_j} \frac{e^{a_i}}{\Sigma} = \frac{0 \cdot \Sigma - e^{a_i} e^{a_j}}{\Sigma^2} = \boxed{-\hat{y}_i \hat{y}_j}$

(recalling $\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$)

$$\frac{\partial y_i}{\partial a_i} = \frac{\partial}{\partial a_i} \frac{e^{a_i}}{\Sigma} = \frac{e^{a_i}\Sigma - e^{a_i} e^{a_i}}{\Sigma^2} = \frac{e^{a_i}}{\Sigma}\left(1 - \frac{e^{a_i}}{\Sigma}\right)$$

$$\boxed{= \hat{y}_i (1 - \hat{y}_i)}$$

To backprop through softmax, we need

$$\frac{\partial \mathcal{L}}{\partial \vec{a}} = \frac{\partial \mathcal{L}}{\partial \hat{\vec{y}}} \cdot \frac{\partial \hat{\vec{y}}}{\partial \vec{a}}$$

Let $\hat{\vec{y}}, \vec{y}, \vec{a} \in \mathbb{R}^{1 \times N}$. Then $\frac{\partial \mathcal{L}}{\partial \hat{\vec{y}}} \in \mathbb{R}^{1 \times N}$ and $\frac{\partial \hat{\vec{y}}}{\partial \vec{a}} \in \mathbb{R}^{N \times N}$.

$$\frac{\partial \mathcal{L}}{\partial \vec{a}} = \left[\frac{-y_1}{\hat{y}_1} \cdots \frac{y_N}{\hat{y}_N}\right] \left[\frac{\partial \hat{\vec{y}}}{\partial a_1} \cdots \frac{\partial \hat{\vec{y}}}{\partial a_n}\right] = \left[(y_1 - \hat{y}_1) \cdots (y_n - \hat{y}_n)\right]$$

$$= \vec{y} - \hat{\vec{y}}$$

This dot product gives the sum

$$\frac{-y_i}{\hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial a_i} - \sum_{j \neq i} \frac{y_j}{\hat{y}_j} \frac{\partial \hat{y}_j}{\partial a_i}$$

$$= -y_i(1 - \hat{y}_i) - \sum_{j \neq i} (-y_j \hat{y}_i) = y_i - \hat{y}_i$$

We also want to backprop through ReLU and Linear:

ReLU

$$\vec{a} = \text{ReLU}(\vec{h}), \quad \text{ReLU}(\vec{x}) = \begin{cases} x_i & \text{if } x_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Then $\frac{\partial \vec{a}}{\partial \vec{h}} = \begin{cases} 1 & \text{if } h_i > 0 \\ 0 & \text{otherwise} \end{cases}$

Linear

$$\vec{h} = \vec{x}\vec{w} + \vec{b}, \quad \vec{x} \in \mathbb{R}^{1 \times M}, \vec{w} \in \mathbb{R}^{M \times L}, b \in \mathbb{R}^{1 \times L}$$

$$\frac{\partial \vec{h}}{\partial \vec{x}} = \vec{w}^T, \quad \frac{\partial \vec{h}}{\partial \vec{b}} = \vec{1}, \quad \frac{\partial \vec{h}}{\partial \vec{w}} = \vec{x}$$

Linking these together gives us a way to backprop through our network.