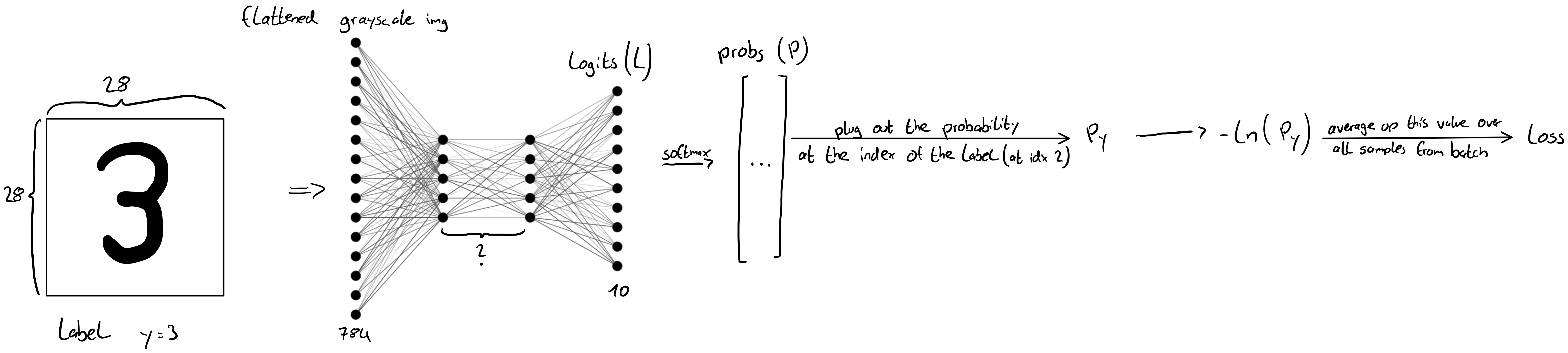
Manual Backpropagation through Cross Entropy Loss

Montag, 2. Oktober 2023 23:13



What we are looking for:

$$\frac{d\,Loss}{d\,Logits}$$

Just with one sample:

$$Loss = -\ln(P_y) \qquad \Big| \; P_i = \underbrace{\frac{e^{L_i}}{\sum_{j=1} e^{L_j}}}_{\text{softmax}}$$

$$= -\ln\left(\frac{e^{L_y}}{\sum_{j=1} e^{L_j}}\right)$$

$$\frac{dLoss}{dL_i} = \frac{d}{dL_i} -\ln\left(\frac{e^{L_y}}{\sum_{j=1} e^{L_j}}\right) \qquad \Big| \text{ chain rule}$$

$$= \frac{d}{dL_i}\left(\frac{e^{L_y}}{\sum_{j=1} e^{L_j}}\right) \cdot \left(-\frac{1}{\frac{e^{L_y}}{\sum_{j=1} e^{L_j}}}\right) \qquad \Big| \text{ simplify}$$

$$= -\frac{\sum_{j=1} e^{L_j}}{e^{L_y}} \cdot \frac{d}{dL_i}\left(e^{L_y} \cdot \frac{1}{\sum_{j=1} e^{L_j}}\right)$$

if y=i:

$$\frac{dLoss}{dL_i} = -\frac{\sum_{j=1} e^{L_j}}{e^{L_i}} \cdot \frac{d}{dL_i}\left(e^{L_i} \cdot \frac{1}{\sum_{j=1} e^{L_j}}\right) \qquad \Big| \text{ product rule}$$

$$= -\frac{\sum_{j=1} e^{L_j}}{e^{L_i}} \cdot \left(\frac{d}{dL_i}\left(e^{L_i}\right) \cdot \frac{1}{\sum_{j=1} e^{L_j}} + e^{L_i} \cdot \frac{d}{dL_i}\left(\frac{1}{\sum_{j=1} e^{L_j}}\right)\right) \qquad \Big| \text{ simplify}$$

$$= -\frac{\sum_{j=1} e^{L_j}}{e^{L_i}} \cdot \left(\frac{e^{L_i}}{\sum_{j=1} e^{L_j}} + e^{L_i} \cdot \frac{d}{dL_i}\left(\frac{1}{\sum_{j=1} e^{L_j}}\right)\right) \qquad \Big| \text{ factor out } e^{L_i}$$

$$= -\sum_{j=1} e^{L_j} \cdot \left(\frac{1}{\sum_{j=1} e^{L_j}} + \frac{d}{dL_i}\left(\frac{1}{\sum_{j=1} e^{L_j}}\right)\right) \qquad \Big| \text{ chain rule}$$

$$= -\sum_{j=1} e^{L_j} \cdot \left(\frac{1}{\sum_{j=1} e^{L_j}} + \frac{d}{dL_i}\left(\sum_{j=1} e^{L_j}\right) \cdot \left(-\frac{1}{\left(\sum_{j=1} e^{L_j}\right)^2}\right)\right) \qquad \Big| \; \frac{d}{dx_3}\left(e^{x_1}+e^{x_2}+e^{x_3}\right) = e^{x_3}$$

$$= -\sum_{j=1} e^{L_j} \cdot \left(\frac{1}{\sum_{j=1} e^{L_j}} - \frac{1}{\left(\sum_{j=1} e^{L_j}\right)^2} \cdot e^{L_i}\right) \qquad \Big| \text{ multiply out}$$

$$= \frac{e^{L_i}}{\sum_{j=1} e^{L_j}} - 1 = P_i - 1$$

if y≠i:

$$\frac{dLoss}{dL_i} = -\frac{\sum_{j=1} e^{L_j}}{e^{L_y}} \cdot \frac{d}{dL_i}\left(e^{L_y} \cdot \frac{1}{\sum_{j=1} e^{L_j}}\right) \qquad \Big| \; L_i \text{ is only inside } \sum e^{L_j} \Rightarrow \text{product rule}$$

$$= \cancel{-}\frac{\sum_{j=1} e^{L_j}}{\cancel{e^{L_y}}} \cdot \cancel{e^{L_y}} \cdot \frac{d}{dL_i}\left(\sum_{j=1} e^{L_j}\right) \cdot \left(\cancel{-}\frac{1}{\left(\sum_{j=1} e^{L_j}\right)^2}\right) \qquad \Big| \text{ simplify}$$

$$= \frac{1}{\sum_{j=1} e^{L_j}} \cdot \frac{d}{dL_i}\left(\sum_{j=1} e^{L_j}\right) \qquad \Big| \; \frac{d}{dx_3}\left(e^{x_1}+e^{x_2}+e^{x_3}\right) = e^{x_3}$$

$$= \frac{e^{L_i}}{\sum_{j=1} e^{L_j}} = P_i$$