Cross Entropy Loss Derivative

Roei Bahumi

In a Supervised Learning Classification task, the "Softmax Classifier" computes the cross-entropy loss:

$$H(p,q) = -\sum_x p(x)\log q(x)$$

We use a 1-hot encoded vector for p, where the 1 is at the index of the true label (y):

$$p_i(x) = \begin{cases} 1 & \text{if y=i} \\ 0 & \text{otherwise} \end{cases}$$

and the softmax function over the logits outputs (z) as our q:

$$q_i(z) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

Because the only non-zero element of p is at the y index, the p vector is in practice a selector for the true label's index in the q vector. Therefore, the loss function for a single sample then becomes:

$$Loss = -\log(\frac{e^{z_y}}{\sum_j e^{z_j}}) = -z_y + \log\sum_j e^{z_j}$$

Calculating the derivative for each z_i:

$$\begin{aligned}
\nabla_{z_i} Loss &= \nabla_{z_i}(-z_y + \log\sum_j e^{z_j}) \\
&= \nabla_{z_i}\log\sum_j e^{z_j} - \nabla_{z_i} z_y \\
&= \frac{1}{\sum_j e^{z_j}}\nabla_{z_i}\sum_j e^{z_j} - \nabla_{z_i} z_y \qquad \text{from} \quad \frac{d}{dx}ln[f(x)] = \frac{1}{f(x)}\frac{d}{dx}f(x) \\
&= \frac{e^{z_i}}{\sum_j e^{z_j}} - \nabla_{z_i} z_y \\
&= q_i(z) - \nabla_{z_i} z_y \\
&= q_i(z) - \mathbb{1}(y=i)
\end{aligned}$$

The effect of $p(x)$, the 1-hot label vector on the gradient is therefore intuitive and directed towards the correct classification:

- The gradient for the true label's logit $(q_y(z) - 1)$ will be negative and decrease proportionally in magnitude as $q_y(z)$ increases.
- The rest of the logits gradient $(q_i(z))$ will be positive and increase proportionally as $q_i(z)$ increases.
- In the specific case of perfect classification where $q_y(z) = 1$, the gradient will be $\vec{0}$.