Let's consider the following two layer architecture:

$$h_1 = \text{Relu}(W_1 X + b_1)$$

$$z = W_2 h_1 + b_2$$

$$\mathcal{L} = CE(z)$$

Where,

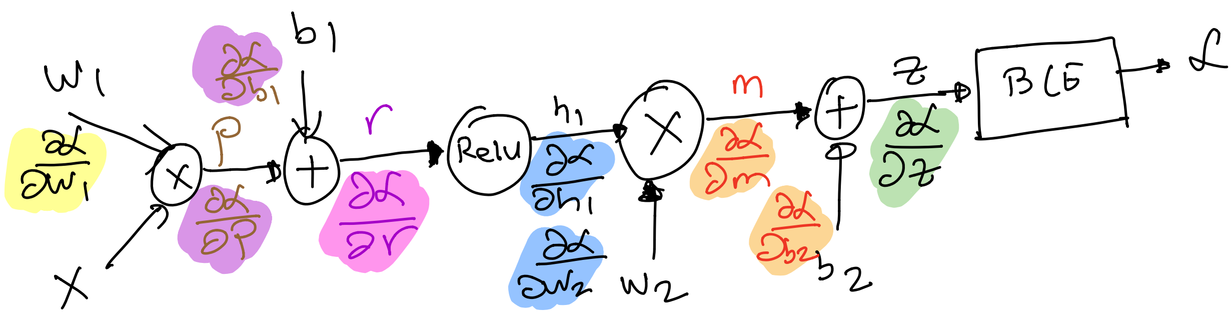
$X \in \mathbb{R}^D$, $h_1 \in \mathbb{R}^H$, $b_1 \in \mathbb{R}^H$

$W_1 \in \mathbb{R}^{H \times D}$, $z \in \mathbb{R}^C$, $b_2 \in \mathbb{R}^C$

$W_2 \in \mathbb{R}^{C \times H}$, $\mathcal{L} \in \mathbb{R}$

and CE stands for cross entropy loss.

The first step in backprop is to draw the computational graph



we know, $\frac{\partial \mathcal{L}}{\partial z}$ from HW2, so

starting with $\frac{\partial \mathcal{L}}{\partial z}$ and backpropagating:

- Since $+$ gate distributes the gradient so,

$$\frac{\partial \mathcal{L}}{\partial m} = \frac{\partial \mathcal{L}}{\partial z}$$

$$\frac{\partial \mathcal{L}}{\partial b_2} = \frac{\partial \mathcal{L}}{\partial z}$$

- Now using the tensor derivative derived in class,

$$\frac{\partial \mathcal{L}}{\partial h_1} = W_2^T \frac{\partial \mathcal{L}}{\partial m}$$

$$\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial \mathcal{L}}{\partial m} h_1^T$$

- Since relu gate routes the gradient, so

$$\frac{\partial \mathcal{L}}{\partial r} = \frac{\partial \mathcal{L}}{\partial h_1} \odot \mathbb{I}(r > 0)$$

- Since $\oplus$ gate distributes the gradient,

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{\partial \mathcal{L}}{\partial r}$$

$$\frac{\partial \mathcal{L}}{\partial b_1} = \frac{\partial \mathcal{L}}{\partial r}$$

- Now using the tensor derivative, derived in class,

$$\frac{\partial \mathcal{L}}{\partial W_1} = \frac{\partial \mathcal{L}}{\partial p} \, X^T$$