General rules for backprop:

input — module — output

* How does output change wrt module $\times$ signal from output $\leftarrow$ how module changes wrt signal

* How does output change wrt input $\times$ signal from output $\leftarrow$ how signal propagates backwards

* Note that the "multiplication" rule is just a rule for general interaction. It's a personal responsibility to determine the order of multiplication for shape compatibility

- Rules for shape

Let s be a scalar, $\vec{v} \in \mathbb{R}^{k\times 1}$, $\vec{y} \in \mathbb{R}^{m\times 1}$, $\vec{x} \in \mathbb{R}^{n\times 1}$, and $W \in \mathbb{R}^{m\times n}$

then $\frac{\partial s}{\partial \vec{v}} \in \mathbb{R}^{1\times k}$, $\frac{\partial \vec{y}}{\partial \vec{x}} \in \mathbb{R}^{m\times n}$, and $\frac{\partial s}{\partial W} \in \mathbb{R}^{m\times n}$

- Dealing with pure linear modules

$$\vec{h}^1 \xrightarrow{\quad W^2 \quad} \vec{h}^2 \leftarrow \partial L$$

$\vec{h}^2 = W^2 \vec{h}^1$

$h^2_i = W^2_{ij} h^1_j$ $\quad \frac{\partial L}{\partial \vec{h}^2}$

$$\frac{\partial L}{\partial W^2} = \frac{\partial \vec{h}^2}{\partial W^2}\frac{\partial L}{\partial \vec{h}^2} \rightarrow \frac{\partial L}{\partial W^2_{ij}} = \frac{\partial L}{\partial h^2_i} \cdot \frac{\partial h^2_i}{\partial W^2_{ij}} = \frac{\partial L}{\partial h^2_i} \cdot h^1_j = \left(\frac{\partial L}{\partial \vec{h}^2}\right)^T \left(\vec{h}^1\right)^T$$

$$\frac{\partial L}{\partial \vec{h}_1} = \frac{\partial \vec{h}_2}{\partial \vec{h}_1}\frac{\partial L}{\partial \vec{h}_2} = \frac{\partial L}{\partial \vec{h}_2} W^2$$ *switch for compatability

Dealing with functional and linear modules

$$\vec{h}^1 \xrightarrow{\quad W^2 \quad} \vec{v} \xrightarrow{\quad ReLu(\vec{v}) \quad} \vec{h}^2 \leftarrow \partial L$$

* Use $\vec{v}$ as an intermediary variable

$\vec{v} = W^2 \vec{h}^1$

$\vec{h}^2 = ReLu(\vec{v})$

$v_i = W^2_{ij} h^1_j$

$h^2_i = ReLu(W^2_{ij} h^1_j)$

$h^2_i = ReLu(v_i)$

$\frac{\partial h^2_i}{\partial v_j} = \delta_{ij} \mathbb{1}(v_i > 0)$

$\frac{\partial L}{\partial \vec{h}^2}$

$$\frac{\partial L}{\partial \vec{v}} = \frac{\partial \vec{h}^2}{\partial \vec{v}}\frac{\partial L}{\partial \vec{h}^2} = ReLu'(\vec{v})\frac{\partial L}{\partial \vec{h}^2}$$

ensure row vector shape $\rightarrow \frac{\partial L}{\partial v_k} = \frac{\partial \vec{h}^2_j}{\partial v_k}\frac{\partial L}{\partial h^2_j} = \frac{\partial L}{\partial h^2_j}\frac{\partial h^2_j}{\partial v_k} = \frac{\partial L}{\partial \vec{h}^2} ReLu'(\vec{v})$

$$\frac{\partial L}{\partial W^2} = \frac{\partial L}{\partial \vec{v}}\frac{\partial \vec{v}}{\partial W^2} \rightarrow \frac{\partial L}{\partial v_i}\frac{\partial v_i}{\partial W^2_{ij}} \rightarrow \frac{\partial L}{\partial v_i} \cdot h^1_j$$ ensure matrix shape

$$= \left(\frac{\partial L}{\partial \vec{v}}\right)^T \left(\vec{h}^1\right)^T = \left(\frac{\partial L}{\partial \vec{h}^2} \cdot \mathrm{diag}(\mathbb{1}(\vec{v} > 0))\right)^T \left(\vec{h}^1\right)^T = \mathrm{diag}(\mathbb{1}(\vec{v} > 0))\left(\frac{\partial L}{\partial \vec{h}^2}\right)^T \left(\vec{h}^1\right)^T$$

$$\frac{\partial L}{\partial \vec{h}^1} = \frac{\partial L}{\partial \vec{v}}\frac{\partial \vec{v}}{\partial \vec{h}^1} = \frac{\partial L}{\partial v_i}\frac{\partial v_i}{\partial h^1_j} = \frac{\partial L}{\partial h^2_k}\frac{\partial h^2_k}{\partial v_i}\frac{\partial v_i}{\partial h^1_j} \rightarrow (1\times k)\times(k\times i)\times(i\times j) \rightarrow (1\times j)$$

$$= \frac{\partial L}{\partial \vec{h}^2} \cdot ReLu'(\vec{v})\, W^2 = \frac{\partial L}{\partial \vec{h}^2} \cdot \mathrm{diag}(\mathbb{1}(\vec{v} > 0)) \cdot W^2$$

For linear modules (and functional modules)

signal from output * how output changes wrt input = how signal propagates back

how output changes wrt module * signal from output = how module changes wrt module

Bonus: updating bias

$\vec{x}$ o———o $\vec{y}$ $\leftarrow \partial L$

$\vec{y} = W\vec{x} + \vec{b}$

$y_i = w_{ij} x_j + b_i$

$$\frac{\partial L}{\partial \vec{y}}$$

$$\frac{\partial L}{\partial \vec{b}} = \frac{\partial L}{\partial \vec{y}} \frac{\partial \vec{y}}{\partial \vec{b}} \rightarrow \frac{\partial L}{\partial b_k} = \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial b_k} = \frac{\partial L}{\partial y_i} \delta_{ik}$$

* Note that bias follows a different multiplication ordering
* Best to use general rule and use shape compatability to understand ordering