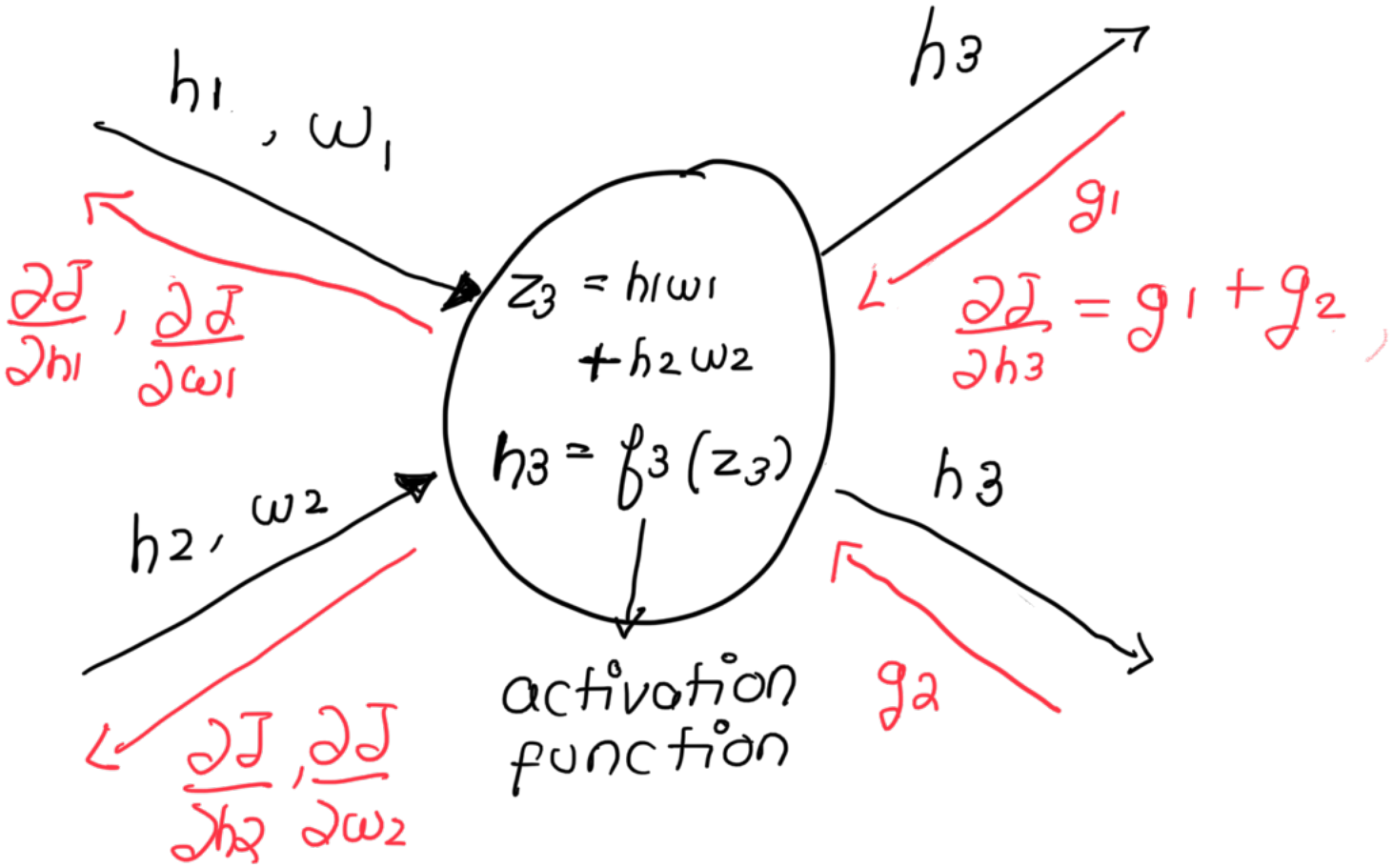
Backpropagation

The goal of backpropagation is to compute $\frac{\partial J}{\partial w}$ for every weight w in the network and perform the update

$$w = w - \ell \frac{\partial J}{\partial w}$$

so that on the next forward pass, the loss J will reduce.

Let's zoom in and observe what BP does at a particular neuron in the network:



where

$$\frac{\partial J}{\partial h_1} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_1}$$

$$\frac{\partial J}{\partial w_1} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial w_1}$$

$$\frac{\partial J}{\partial h_2} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_2}$$

$$\frac{\partial J}{\partial w_2} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial w_2}$$

Also keep in mind that

$\frac{\partial h_i}{\partial a}$ for some i and variable a

$$= \frac{\partial h_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial a}$$

because of activations, but for the sake of keeping the chains shorter, I won't always expand

Now, we've setup a procedure we can follow at every neuron that computes $h' = f'(z')$ where f' is the activation and z'

f' [illegible]

$= w_1 h_1 + \dots + w_k h_k$ (the
weighted sum of outputs from the
previous layer)

1. If h' reaches the final output through multiple paths, it will accumulate multiple 'sub gradients' $g_1, \dots, g_j$. Add these up, this gives

$$g_1 + \dots + g_j = \frac{\partial J}{\partial h'}$$

(You can think of each g_i as one of the summands when you compute $\frac{\partial J}{\partial h'}$ using the chain rule)

2. Compute the partial derivative of h' w.r.t every input $w_1, h_1, \dots, w_k, h_k$, the gradients

(we also need the gradients w.r.t $h_1, \ldots, h_K$ because they are required to calculate $\frac{\partial J}{\partial w}$ for weights w in earlier layers)

3. Apply the chain rule for every input!

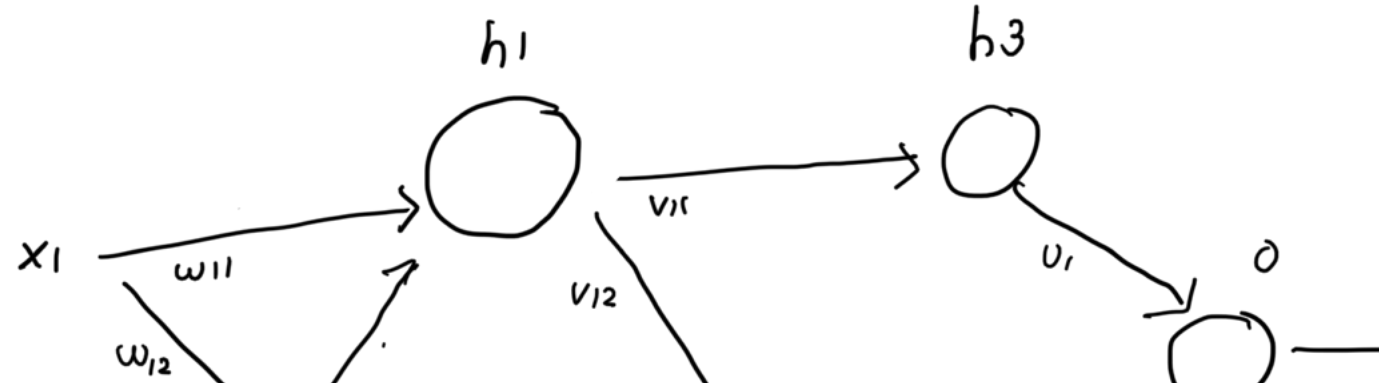
Ex: $\frac{\partial J}{\partial h_1} = \frac{\partial J}{\partial h'} \cdot \frac{\partial h'}{\partial h_1}$

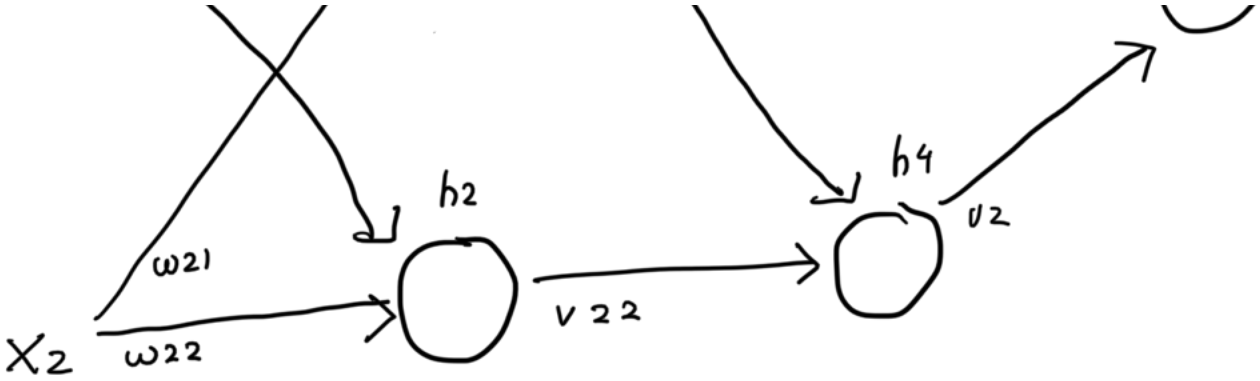
- $\frac{\partial J}{\partial h_1}$: outgoing gradient
- $\frac{\partial J}{\partial h'}$: incoming, precomputed gradient
- $\frac{\partial h'}{\partial h_1}$: local gradient computed in step 2

Notice how we are performing the chain rule incrementally, one chain 'link' at a time. This is exactly why backprop is efficient - it isn't repeatedly

calculating long chains for every variable because it doesn't need to. The compositional nature of neural networks means that 'sub' chains computed at later layers can be reused to compute the longer chains at earlier layers, meaning HUGE savings on computation overall, especially for large, wide, denser nets

Let's do an example:



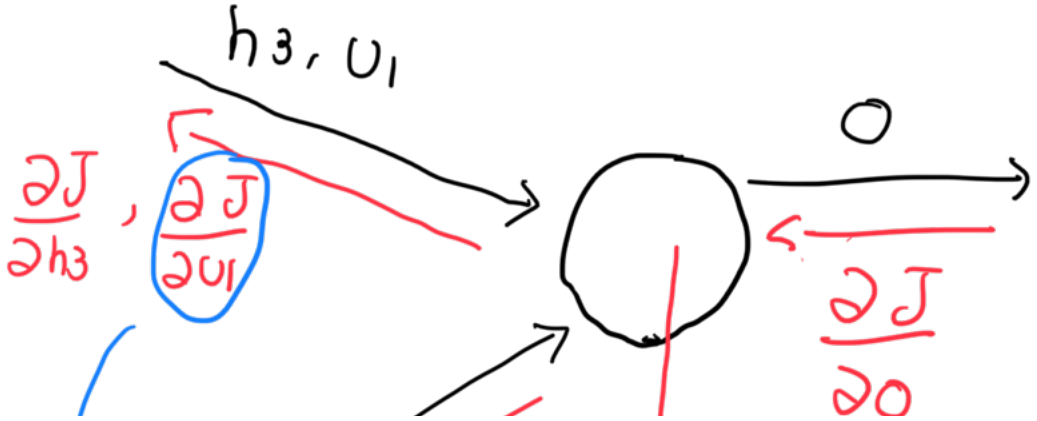


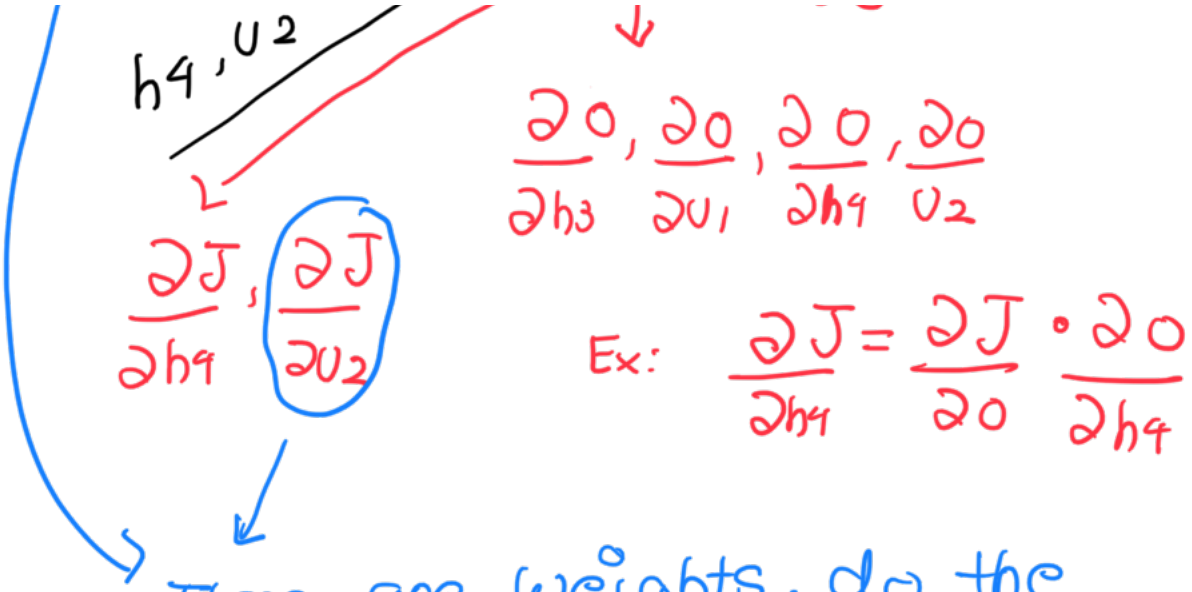
$h_i = f_i(z_i)$ where z_i = weighted sum of inputs

$o = f_o(z_o)$ where $z_o = v_1 h_3 + v_2 h_4$

o = predictions of the network

With o, we can compute our loss J since we have both the targets and predictions. And then as our first step in backprop, we compute $\frac{\partial J}{\partial o}$



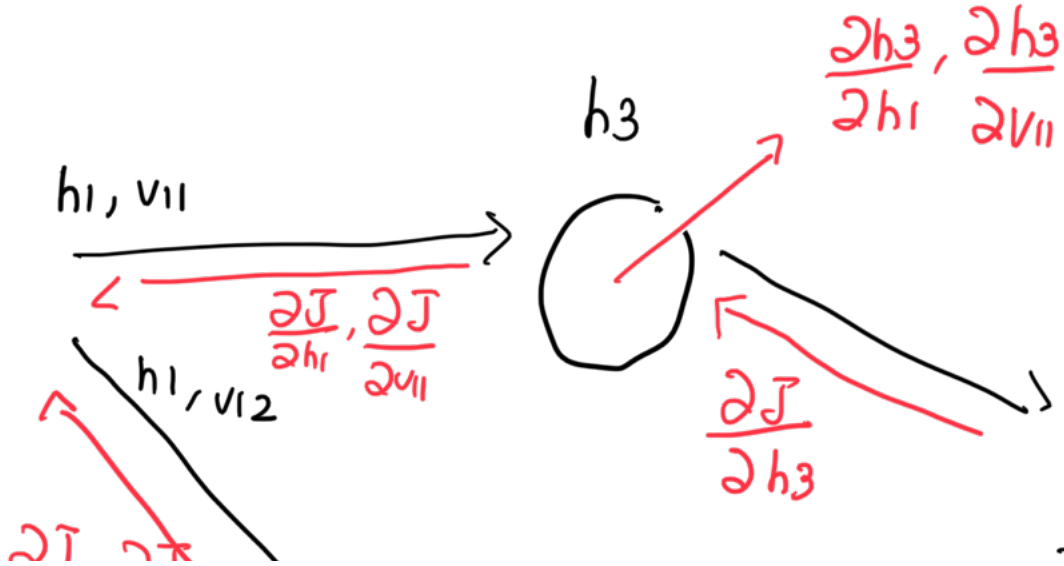


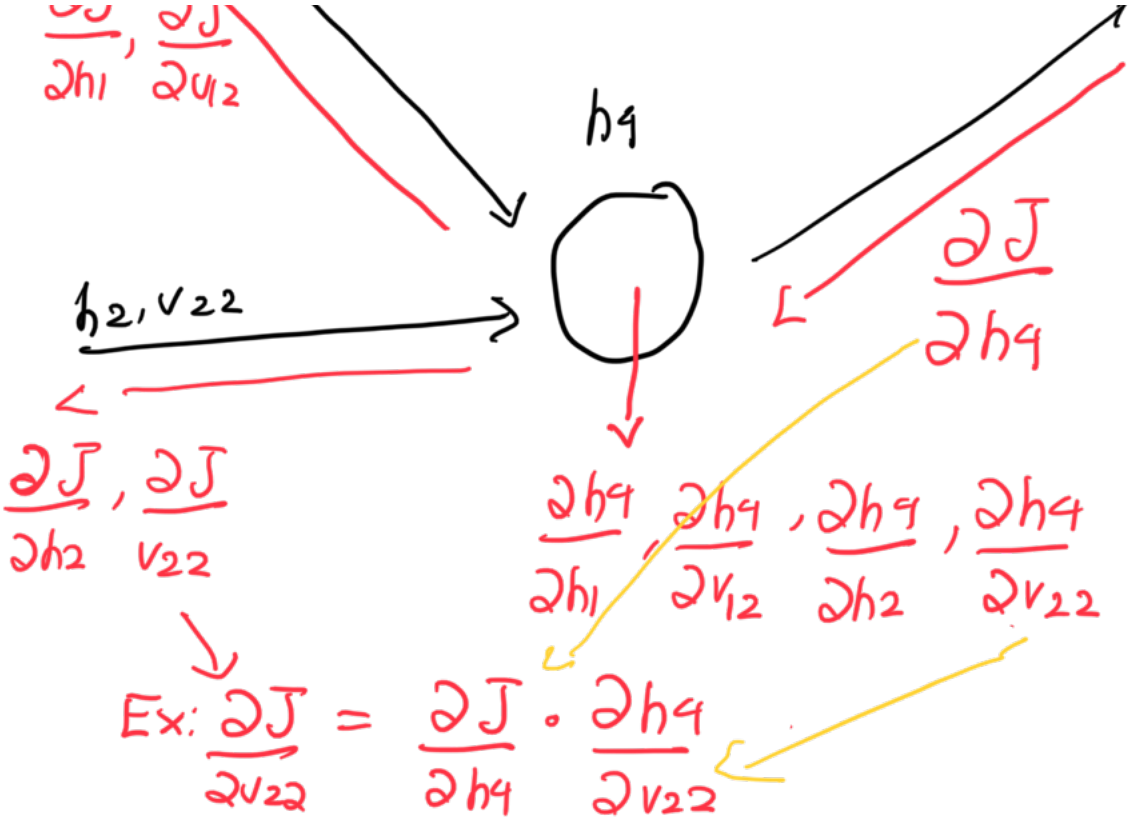
These are weights, do the gradient descent update now!

$$u_1 = u_1 - \ell \frac{\partial J}{\partial u_1}$$

$$u_2 = u_2 - \ell \frac{\partial J}{\partial u_2}$$

Now we move backward to the previous layer ...

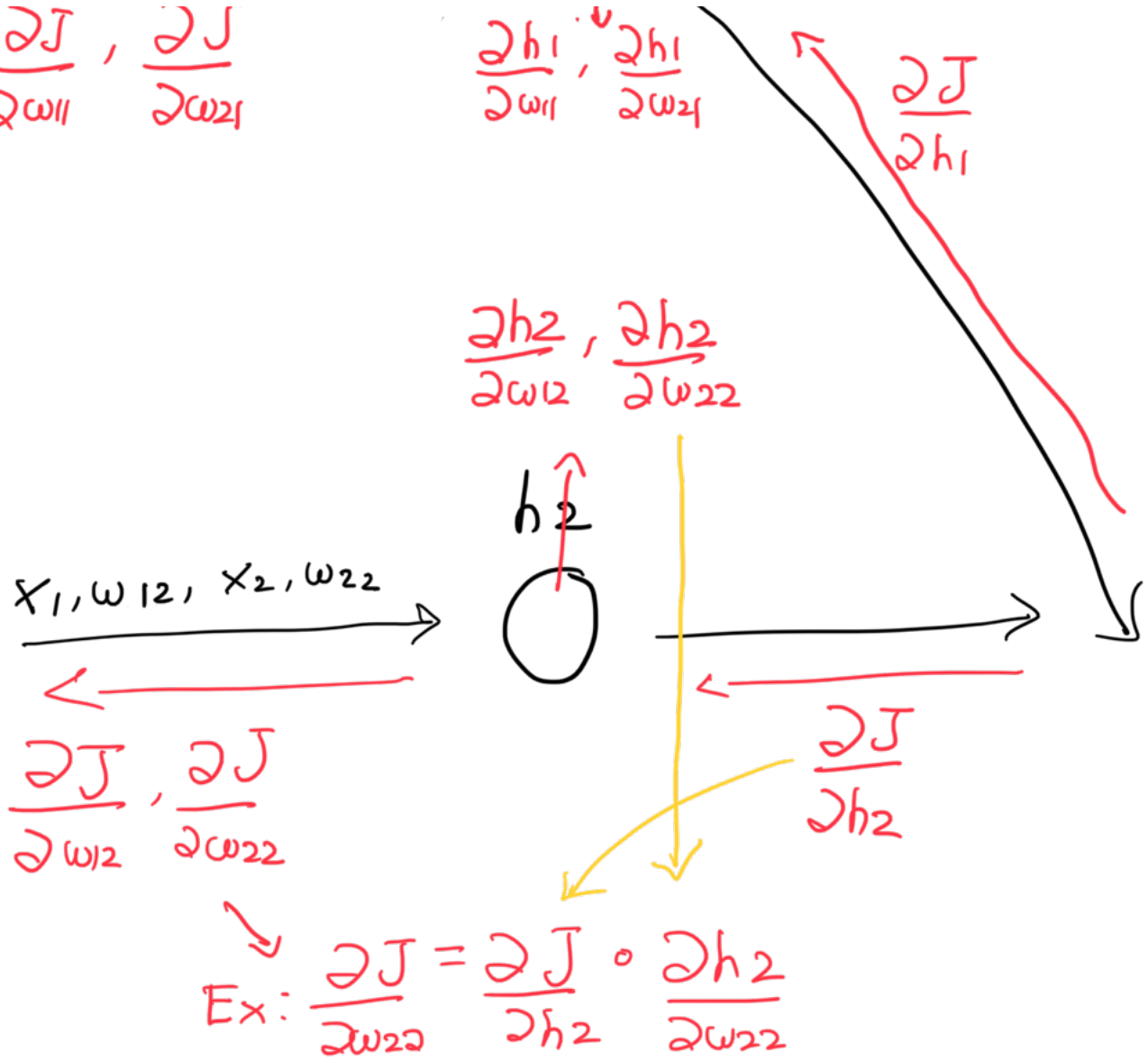




We've calculated $\frac{\partial J}{\partial v_{11}}, \frac{\partial J}{\partial v_{12}}, \frac{\partial J}{v_{22}}$.

Now we can do the gradient descent update for those

Last layer! (combining input arrows into 1)



Notice a few things:

1) There are 2 $\frac{\partial J}{\partial h_1}$'s coming into h_1. These are the 'sub' gradients I was referring to earlier. The actual, official ∂J would be the

$\frac{\partial J}{\partial h_1}$

sum of these and what ultimately gets back propagated /used. Ex: when calculating

$$\frac{\partial J}{\partial w_{11}} = \frac{\partial J}{\partial h_1} \cdot \frac{\partial h_1}{\partial w_{11}}$$

I didn't relabel these sub gradients because it keeps things cleaner and easier to follow. But remember, neurons with more than 1 channel of output accumulate / add multiple subgradients to compute a final gradient that is shared with earlier layers

2) we didn't calculate ∂J

for x_1, x_2. Because we don't have to! They would be useless - there are no earlier layers and it doesn't make sense to adjust inputs (we don't have control over that, duh)

That is one complete backward pass! We could compute $\frac{\partial J}{\partial w}$ for all weights w in the network. Next, we would repeat the cycle of doing a forward pass of predicting with the updated weights - backpropping

the gradients and so on till convergence (weights stabilize)

To show that backprop is correct and efficient, consider what the formula for $\frac{\partial J}{\partial w_{11}}$ would be.

w_1 influences o through 2 paths:

1) $w_1 \longrightarrow h_1 \longrightarrow h_3 \longrightarrow o$
2) $w_1 \longrightarrow h_1 \longrightarrow h_4 \longrightarrow o$

So we'll have 2 summands in the formula for $\frac{\partial J}{\partial w_{11}}$:

$$\frac{\partial J}{} = \frac{\partial J}{} \frac{\partial o}{} \frac{\partial h_3}{} \frac{\partial h_1}{} +$$

$$\frac{\partial w_{11}}{\partial o}\ \frac{\partial o}{\partial h_3}\ \frac{\partial h_3}{\partial h_1}\ \frac{\partial h_1}{\partial w_{11}}$$

$$\frac{\partial J}{\partial o}\ \frac{\partial o}{\partial h_4}\ \frac{\partial h_4}{\partial h_1}\ \frac{\partial h_1}{\partial w_{11}}$$

Now let's see if we get this from our backprop calculation. Continuously unraveling the chain:

$$\frac{\partial J}{\partial w_{11}} = \frac{\partial J}{\partial h_1} \cdot \frac{\partial h_1}{\partial w_{11}}$$

$$\| $$

$$\left(\frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_1} + \frac{\partial J}{\partial h_4} \cdot \frac{\partial h_4}{\partial h_1}\right) \frac{\partial h_1}{\partial w_{11}}$$

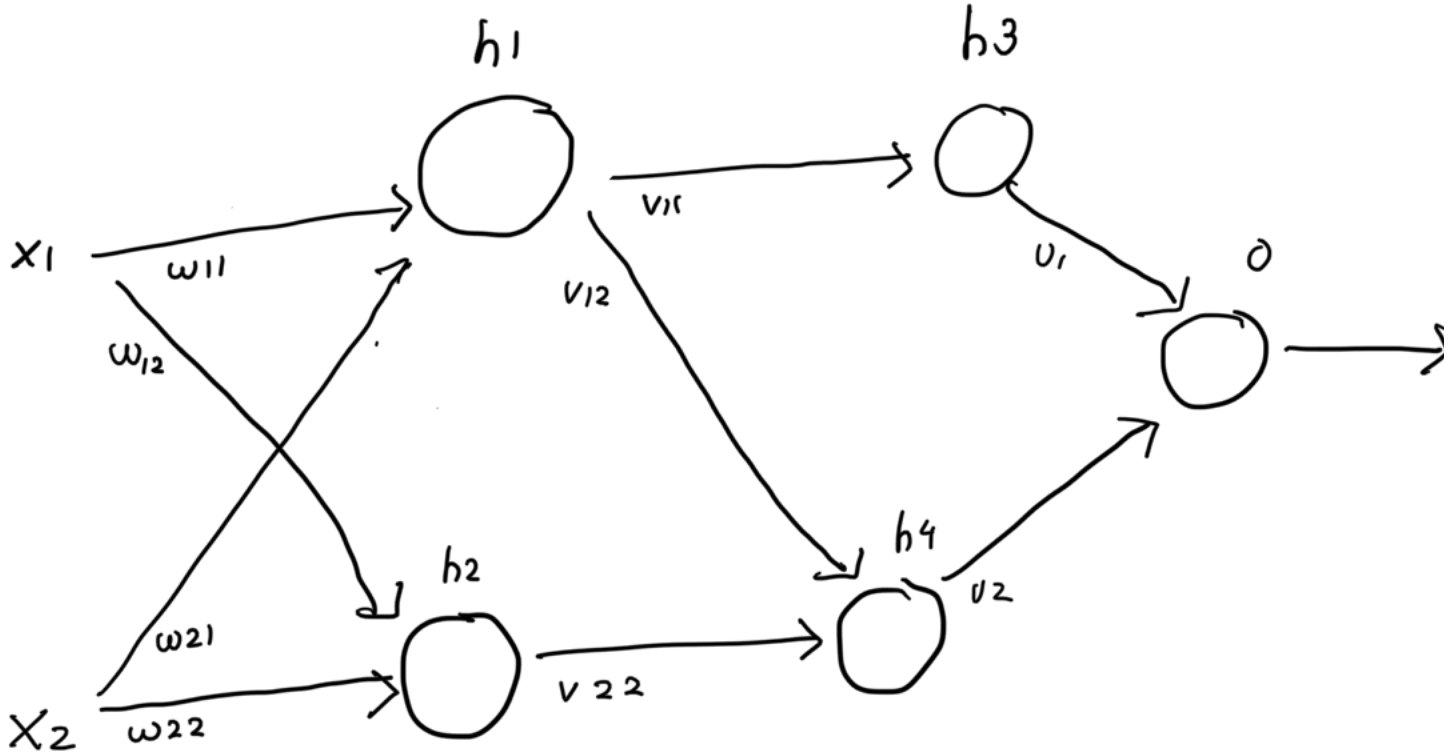
$$= \left(\frac{\partial J}{\partial o}\ \frac{\partial o}{\partial h_3}\ \frac{\partial h_3}{\partial h_1} + \frac{\partial J}{\partial o}\ \frac{\partial o}{\partial h_4}\ \frac{\partial h_4}{\partial h_1}\right) \frac{\partial h_1}{\partial w_{11}}$$

$$= \frac{\partial J}{\partial o}\ \frac{\partial o}{\partial h_3}\ \frac{\partial h_3}{\partial h_1}\ \frac{\partial h_1}{\partial w_{11}} + \frac{\partial J}{\partial o}\ \frac{\partial o}{\partial h_4}\ \frac{\partial h_4}{\partial h_1}\ \frac{\partial h_1}{\partial w_{11}}$$

and voila!

Now let's do an example with actual numbers and functions

Reminder about the network:



$h_i = f_i(z_i)$

z_i = weighted sum of inputs

Define

$h_1 = f_1(x) = x$

$h_2 = f_2(x) = \frac{1}{2}x$

$h_3 = f_3(x) = 2x$

$h_4 = f_4(x) = e^x$

$0 = f_0(x) = \sin(x)$

And assume all weights are initialized to 0, i.e

$w_{11} = w_{12} = w_{21} = w_{22} = v_{11} = v_{12} = v_{22} = v_1 = v_2 = 0$

Let's do one cycle of a forward and backword pass with the training instance

$(x_1, x_2, y) = (0, 1, 0)$

Step 1: forword pass

Calculate:

$h_1 =$

$n_2 =$

$h_3 =$

$h_4 =$

$o =$

Step 2 : Compute loss

Let's define

J as $\frac{1}{2}(o-t)^2$

where t is the target and o is the output of the network

Calculate

$\frac{\partial J}{\partial o} = o - t$

and plug in t, o

Step 3: Backpropagate

Layer 3

(Remember to plug in values you've already computed, including from the forward pass)

0

Calculate:

$$\frac{\partial o}{\partial h_3} = \frac{\partial f_o}{\partial z_o} \cdot \frac{\partial z_o}{\partial h_3} = \cos(z_o)\, u_1$$

$$\frac{\partial o}{\partial h_4} = \frac{\partial f_o}{\partial z_o} \cdot \frac{\partial z_o}{\partial h_4} = \cos(z_o)\, u_2$$

$$\frac{\partial o}{\partial u_1} = \frac{\partial f_o}{\partial z_o} \cdot \frac{\partial z_o}{\partial u_1} = \cos(z_o)\, h_3$$

$$\frac{\partial o}{\partial v_2} = \frac{\partial f_o}{\partial z_o} \cdot \frac{\partial z_o}{\partial v_2} = (os(z_o)\, h4$$

$$\frac{\partial J}{\partial h3} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial h3} = (t-o)\cos(z_o)\, v_1$$

$$\frac{\partial J}{\partial h4} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial h4} = (t-o)\cos(z_o)\, v_2$$

$$\frac{\partial J}{\partial v_1} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial v_1} = (t-o)\cos(z_o)\, h3$$

$$\frac{\partial J}{\partial v_2} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial v_2} = (t-o)\cos(z_o)\, h4$$

Layer 2

h3

Calculate:

$$\partial h_3 = \partial f_3 \cdot \partial z_3 = 2 \cdot v_{11}$$

$$\frac{\partial h_3}{\partial h_1} \quad \frac{\partial f_3}{\partial z_3} \quad \frac{\partial z_3}{\partial h_1}$$

$$\frac{\partial h_3}{\partial v_{11}} = \frac{\partial f_3}{\partial z_3} \cdot \frac{\partial z_3}{\partial v_{11}} = 2 \cdot h_1$$

$$\frac{\partial J}{\partial h_1} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_1} = 2(t-o)\cos(z_0) u_1 v_{11}$$

$$\frac{\partial J}{\partial v_{11}} = \frac{\partial J}{\partial h_3} \cdot \frac{\partial h_3}{\partial v_{11}} = 2 \cdot h_1 (t-o)\cos(z_0) u_1$$

h_4

Calculate:

$$\frac{\partial J}{\partial h_1} = \frac{\partial J}{\partial h_4} \cdot \frac{\partial h_4}{\partial h_1} = \frac{\partial J}{\partial h_4} \cdot \frac{\partial f_4}{\partial z_4} \cdot \frac{\partial z_4}{\partial h_1}$$

$$\frac{\partial J}{\partial h_2} = \frac{\partial J}{\partial h_4} \cdot \frac{\partial h_4}{\partial h_2} = \frac{\partial J}{\partial h_4} \cdot \frac{\partial f_4}{\partial z_4} \cdot \frac{\partial z_4}{\partial h_2}$$

$$\frac{\partial J}{\partial} = \frac{\partial J}{\partial} \cdot \frac{\partial h_4}{\partial} = \frac{\partial J}{\partial} \cdot \frac{\partial f_4}{\partial} \cdot \frac{\partial z_4}{\partial}$$

$\frac{\partial v_{12}}{\partial v_{12}} \quad \frac{\partial h_4}{\partial v_{12}} \quad \frac{\partial h_4}{\partial z_4} \quad \frac{\partial v_{12}}{\partial v_{22}}$

$$\frac{\partial J}{\partial v_{22}} = \frac{\partial J}{\partial h_4} \cdot \frac{\partial h_4}{\partial v_{22}} = \frac{\partial J}{\partial h_4} \cdot \frac{\partial f_4}{\partial z_4} \cdot \frac{\partial z_4}{\partial v_{22}}$$

Add up the 2 $\frac{\partial J}{\partial h_1}$'s you calculated and treat this as your new $\frac{\partial J}{\partial h_1}$

Layer 3

h_1

Calculate

$$\frac{\partial J}{\partial w_{11}} =$$

$\frac{\partial J}{\partial w_{12}} =$

h_2

Calculate

$\frac{\partial J}{\partial w_{21}} =$

$\frac{\partial J}{\partial w_{22}} =$

Finally, do all the weight updates using

weight

$$w = w - \ell \frac{\partial J}{\partial w}$$

what are the new values of

$w_{11} =$

$w_{12} =$

$w_{21} =$

$w_{22} =$

$v_{11} =$

$v_{12} =$

$v_{22} =$

$v_1 =$

$v_2 =$

02