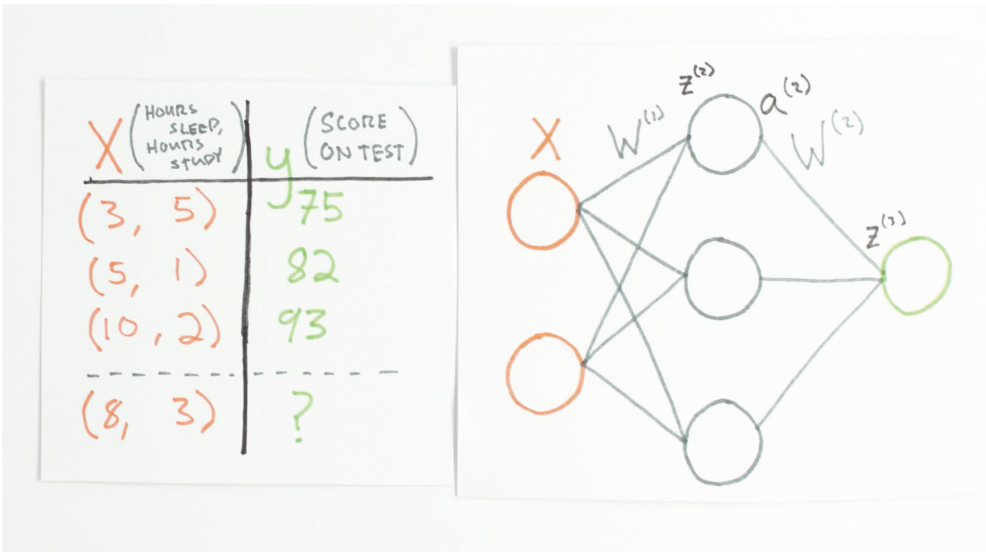
Backpropogation Guided Notes

Backpropogation is a critical piece of modern deep learning. To really get a grasp of how backpropogation works, there's nothing quite like deriving the equations for yourself. Let's do it.

Data **Architecture**



Forward Equations

$$z^{(2)} = XW^{(1)} \tag{1}$$

$$a^{(2)} = f(z^{(2)}) \tag{2}$$

$$z^{(3)} = a^{(2)}W^{(2)} \tag{3}$$

$$\hat{y} = f(z^{(3)}) \tag{4}$$

$$J = \sum \frac{1}{2}(y - \hat{y})^2 \tag{5}$$

Code Symbol	Math Symbol	Definition	Dimensions
X	X	Input Data, each row in an example	(numExamples, inputLayerSize)
y	y	target data	(numExamples, outputLayerSize)
W1	$W^{(1)}$	Layer 1 weights	(inputLayerSize, hiddenLayerSize)
W2	$W^{(2)}$	Layer 2 weights	(hiddenLayerSize, outputLayerSize)
z2	$z^{(2)}$	Layer 2 activation	(numExamples, hiddenLayerSize)
a2	$a^{(2)}$	Layer 2 activity	(numExamples, hiddenLayerSize)
z3	$z^{(3)}$	Layer 3 activation	(numExamples, outputLayerSize)
J	J	Cost	(1, outputLayerSize)
dJdz3	$\frac{\partial J}{\partial z^{(3)}} = \delta^{(3)}$	Partial derivative of cost with respect to $z^{(3)}$	
dJdW2	$\frac{\partial J}{\partial W^{(2)}}$	Partial derivative of cost with respect to $W^{(2)}$	
dz3dz2	$\frac{\partial z^{(3)}}{\partial z^{(2)}}$	Partial derivative of $z^{(3)}$ with respect to $z^{(2)}$	
dJdW1	$\frac{\partial J}{\partial W^{(1)}}$	Partial derivative of cost with respect to $W^{(1)}$	
delta2	$\delta^{(2)}$	Backpropagating Error 2	
delta3	$\delta^{(3)}$	Backpropagating Error 1	

For you to figure out!

Your Mission

$$\frac{\partial J}{\partial W^{(1)}} = ? \quad \frac{\partial J}{\partial W^{(2)}} = ?$$

1. The dimension of $\frac{\partial J}{\partial W^{(1)}}$ is ______________________________.

2. The dimension of $\frac{\partial J}{\partial W^{(2)}}$ is ______________________________.

3. Using (5), we can write $\frac{\partial J}{\partial W^{(2)}} = \frac{\partial \sum \frac{1}{2}(y-\hat{y})^2}{\partial W^{(2)}}$.

Use the sum rule for differentiation to move the summation outside the gradient:

$$\frac{\partial J}{\partial W^{(2)}} =$$

4. Let's temporarily remove the summation, and consider $\frac{\partial J}{\partial W^{(2)}}$ in terms of just one example (numExamples = 1). Using the chain rule, derive and expression for $\frac{\partial J}{\partial W^{(2)}}$ in terms of y, $\hat{y}$, $\frac{\partial \hat{y}}{\partial W^{(2)}}$.

$$\frac{\partial J}{\partial W^{(2)}} =$$

5. Now, use the chain rule again to express $\frac{\partial J}{\partial W^{(2)}}$ in terms of y, $\hat{y}$, $\frac{\partial \hat{y}}{\partial z^{(3)}}$, $\frac{\partial z^{(3)}}{\partial W^{(2)}}$.

$$\frac{\partial J}{\partial W^{(2)}} =$$

6. $\hat{y}$ and $z^{(3)}$ are connected by our simgoid activation function $f(z) = \frac{1}{1+e^{-z}}$.

$$\frac{\partial \hat{y}}{\partial z^{(3)}} = f'(z) =$$

You should now have an equation that looks something like this:

$$\frac{\partial J}{\partial W^{(2)}} = -(y-\hat{y})f'(z^{(3)})\frac{\partial z^{(3)}}{\partial W^{(2)}}$$

To simplify our equations a little, let's introduce a new term, the "backpropogating error":

$$\delta^{(3)} = -(y-\hat{y})f'(z^{(3)})$$

7. What is the dimension of $\delta^{(3)}$?

8. Now we need to work on $\frac{\partial z^{(3)}}{\partial W^{(2)}}$. To get started, write out the full matrix equation for (3), using numExamples = 1, and inputLayerSize = 2, hiddenLayerSize = 3, and outputLayerSize = 1.

9. Now, using your calculus skills:

$$\frac{\partial z^{(3)}}{\partial W^{(2)}} = \begin{bmatrix} \frac{\partial z^{(3)}}{\partial W_{11}^{(2)}} \\ \frac{\partial z^{(3)}}{\partial W_{21}^{(2)}} \\ \frac{\partial z^{(3)}}{\partial W_{31}^{(2)}} \end{bmatrix} = \begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix}$$

10. Now, write $\frac{\partial z^{(3)}}{\partial W^{(2)}}$ in terms of the vector $a^{(2)}$:

$$\frac{\partial z^{(3)}}{\partial W^{(2)}} =$$

11. $W^{(1)} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ $W^{(2)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ $X = \begin{bmatrix} 0.3 \ 1.0 \end{bmatrix}$ $\frac{\partial J}{\partial W^{(2)}}$ =?

12. Next, let's let's deal with the numExamples > 1 case. Back in question 4 we temporarily took away the summation, we'll figure out how to re-intoduce it now. To get started, write out the full matrix equation for (3), using numExamples = 3, and inputLayerSize = 2, hiddenLayerSize = 3, and outputLayerSize = 1.

What do the rows and columns of your "a" matrix represent?

13. Now that we've let numExamples=3, what is the dimension of $\delta^{(3)}$?

14. Almost there! Now, sum across our examples in terms of the individual elements of $\delta^{(3)}$ and $a^{(2)}$:

$$\frac{\partial J}{\partial W^{(2)}} = \begin{bmatrix} \frac{\partial J}{\partial W_{11}^{(2)}} \\ \frac{\partial J}{\partial W_{21}^{(2)}} \\ \frac{\partial J}{\partial W_{31}^{(2)}} \end{bmatrix} = \begin{bmatrix} \quad \end{bmatrix}$$

15. Now express the above operation in terms of the matrix $a^{(2)}$ and the vector $\delta^{(3)}$.

$$\frac{\partial J}{\partial W^{(2)}} =$$

16. $$W^{(1)} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad W^{(2)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad X = \begin{bmatrix} 0.3 & 1.0 \\ 0.5 & 0.2 \\ 1.0 & 0.4 \end{bmatrix} \quad \frac{\partial J}{\partial W^{(2)}} = ?$$

17. Derive an expression for $\frac{\partial J}{\partial W^{(1)}}$ by continuing to propogate errors backwards through our network.