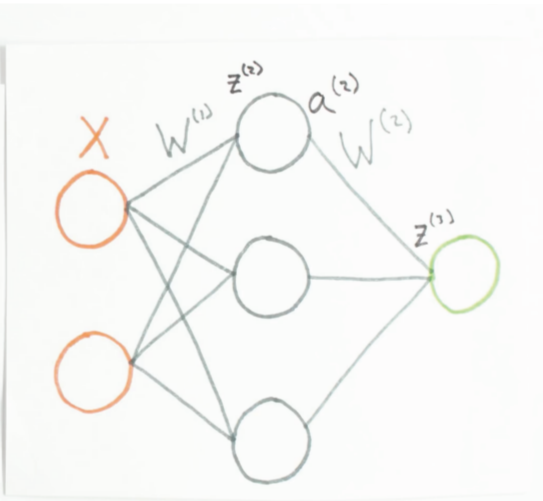
Backpropagation Guided Notes

Backpropagation 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

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

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

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

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

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

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

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

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:

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

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 $a^{(2)}$:
(Hint: to compute the gradients with repsect to each of our 3 weights in W2, **we need to add across our examples.)**

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

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

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

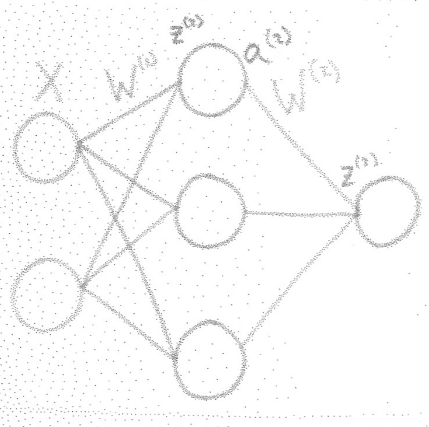
Backpropogation Guided Notes

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

Data

X (Hours Slept, Hours Studied)	y (Score on Test)
(3, 5)	75
(5, 1)	82
(10, 2)	93
(8, 3)	?

Architecture



Forward Equations

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

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

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

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

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

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) (3,1
z2	$z^{(2)}$	Layer 2 activation	(numExamples, hiddenLayerSize) (3,3)
a2	$a^{(2)}$	Layer 2 activity	(numExamples, hiddenLayerSize) (3,3)
z3	$z^{(3)}$	Layer 3 activation	(numExamples, outputLayerSize) (3,1)
J	J	Cost	(1, outputLayerSize)
dJdz3	$\frac{\partial J}{\partial z^{(3)}} = \delta^{(3)}$	Partial derivative of cost with respect to $z^{(3)}$	(3x1) ← one for each example.
dJdW2	$\frac{\partial J}{\partial W^{(2)}}$	Partial derivative of cost with respect to $W^{(2)}$	(3x1)
dz3dz2	$\frac{\partial z^{(3)}}{\partial z^{(2)}}$	Partial derivative of $z^{(3)}$ with respect to $z^{(2)}$	(3x3)
dJdW1	$\frac{\partial J}{\partial W^{(1)}}$	Partial derivative of cost with respect to $W^{(1)}$	(2x3)
delta2	$\delta^{(2)}$	Backpropagating Error 2	
delta3	$\delta^{(3)}$	Backpropagating Error 1	(3x1)

numExamples ↑ output layer size.

For you to figure out!

Your Mission

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

1. The dimension of $\frac{\partial J}{\partial W^{(1)}}$ is (2,3) ← SAME AS $W^{(1)}$.

2. The dimension of $\frac{\partial J}{\partial W^{(2)}}$ is (3,1) ← SAME AS $W^{(2)}$.

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:

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

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 $y, \hat{y}, \frac{\partial \hat{y}}{\partial W^{(2)}}$.

$$\frac{\partial J}{\partial W^{(2)}} = -(y-\hat{y}) \cdot \frac{\partial \hat{y}}{\partial W^{(2)}} \quad (y \text{ is constant}).$$

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

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

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

$$\frac{\partial \hat{y}}{\partial z^{(3)}} = f'(z) = \left((1+e^{-z})^{-1}\right)' = -(1+e^{-z})^{-2} \cdot -e^{-z}$$

$$= \frac{e^{-z}}{(1+e^{-z})^2}$$

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 $\delta^{(3)}$?

(1, 1) ← NUMBER OF EXAMPLES.
↑ OUTPUT LAYER SIZE

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.

$$z^{(3)} = a^{(2)}W^{(2)} \Rightarrow z^{(3)} = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} \end{bmatrix} \begin{bmatrix} W_{11}^{(2)} \\ W_{21}^{(2)} \\ W_{31}^{(2)} \end{bmatrix}$$

9. Now, using your calculus skills:

$$\frac{\partial z^{(3)}}{\partial W^{(2)}} = \begin{bmatrix} \frac{\partial z^{(3)}}{\partial W_{11}^{(2)}} \\ \frac{\partial z^{(3)}}{\partial W_{21}^{(2)}} \\ \frac{\partial z^{(3)}}{\partial W_{31}^{(2)}} \end{bmatrix} = \begin{bmatrix} a_{11}^{(2)} \\ a_{12}^{(2)} \\ a_{13}^{(2)} \end{bmatrix}$$

This is just like:
$f(x) = ax \Rightarrow f'(x) = a$
(We're assuming a is constant, which is true here b/c our training data is fixed).

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

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

11. $W^{(1)} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ $W^{(2)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ let $y = 0.75$ $X = [0.3\ 1.0]$ $\frac{\partial J}{\partial W^{(2)}} = ?$

FORWARD PASS FIRST:

$z^{(2)} = [0.3\ 1.0] \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ (1 → 2)

$= [1.3\ 1.3\ 1.3]$

$a^{(2)} = f([1.3, 1.3, 1.3])$

$= [0.79, 0.79, 0.79]$

$z^{(3)} = a^{(2)}W^{(2)} = 2.37$

$\hat{y} = f(2.37) = 0.92$

$J = \frac{1}{2}(0.75 - 0.92)^2$

$J = 0.015$ ← Don't need this by computer still

$\frac{\partial J}{\partial W^{(2)}} = (\hat{y} - y)^T = \begin{bmatrix} 0.79 \\ 0.79 \\ 0.79 \end{bmatrix}$

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.

$$\begin{bmatrix} z_{11}^{(3)} \\ z_{21}^{(3)} \\ z_{31}^{(3)} \end{bmatrix} = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} & a_{23}^{(2)} \\ a_{31}^{(2)} & a_{32}^{(2)} & a_{33}^{(2)} \end{bmatrix} \begin{bmatrix} W_{11}^{(2)} \\ W_{21}^{(2)} \\ W_{31}^{(2)} \end{bmatrix}$$

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

Rows = examples
cols = hidden units/neurons.

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

3×1

$$\begin{bmatrix} \delta_{11}^{(3)} \\ \delta_{21}^{(3)} \\ \delta_{31}^{(3)} \end{bmatrix} = \begin{bmatrix} -(y_{11} - \hat{y}_{11}) f'(z_{11}^{(3)}) \\ -(y_{21} - \hat{y}_{21}) f'(z_{21}^{(3)}) \\ -(y_{31} - \hat{y}_{31}) f'(z_{31}^{(3)}) \end{bmatrix}$$

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

SUMMING ACCROSS EXAMPLES.

2 HIDDEN NEURONS

$$\frac{\partial J}{\partial W^{(2)}} = \begin{bmatrix} \frac{\partial J}{\partial W_{11}^{(2)}} \\ \frac{\partial J}{\partial W_{21}^{(2)}} \\ \frac{\partial J}{\partial W_{31}^{(2)}} \end{bmatrix} = \begin{bmatrix} a_{11}^{(2)} \delta_{11}^{(3)} + a_{21}^{(2)} \delta_{21}^{(3)} + a_{31}^{(2)} \delta_{31}^{(3)} \\ a_{12}^{(2)} \delta_{11}^{(3)} + a_{22}^{(2)} \delta_{21}^{(3)} + a_{31}^{(2)} \delta_{31}^{(3)} \\ a_{13}^{(2)} \delta_{11}^{(3)} + a_{23}^{(2)} \delta_{21}^{(3)} + a_{33}^{(2)} \delta_{31}^{(3)} \end{bmatrix}$$

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

$$\frac{\partial J}{\partial W^{(2)}} = (a^{(2)})^T \delta^{(3)} = -(a^{(2)})^T (y - \hat{y}) \frac{\partial z^{(3)}}{\partial W^{(2)}}$$