Matt Mazur

A Step by Step Backpropagation Example

Background

Backpropagation is a common method for training a neural network. There is no shortage of papers online that attempt to explain how backpropagation works, but few that include an example with actual numbers. This post is my attempt to explain how it works with a concrete example that folks can compare their own calculations to in order to ensure they understand backpropagation correctly.

If this kind of thing interests you, you should sign up for my newsletter where I post about AI-related projects that I'm working on.

Backpropagation in Python

You can play around with a Python script that I wrote that implements the backpropagation algorithm in this Github repo.

Backpropagation Visualization

For an interactive visualization showing a neural network as it learns, check out my Neural Network visualization.

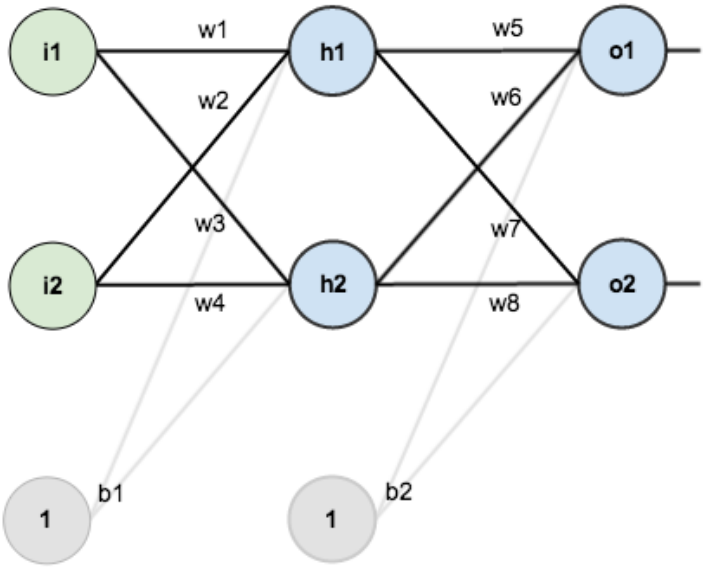
Additional Resources

If you find this tutorial useful and want to continue learning about neural networks and their applications, I highly recommend checking out Adrian Rosebrock's excellent tutorial on Getting Started with Deep Learning and Python.

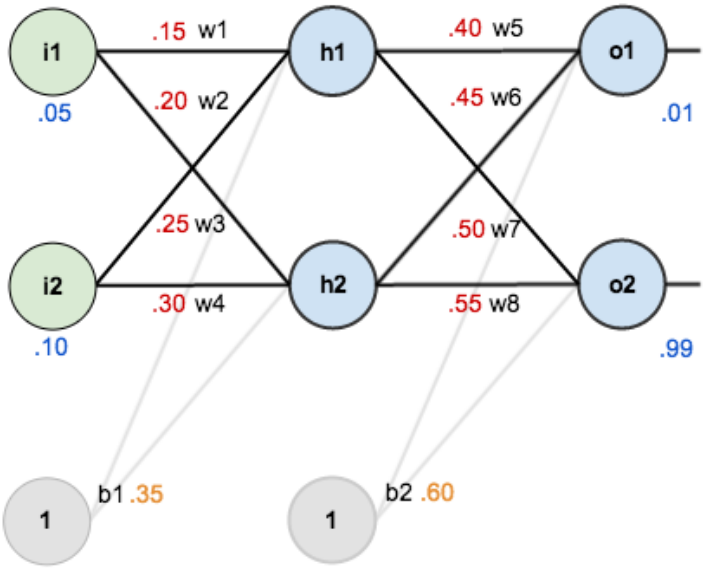
Overview

For this tutorial, we're going to use a neural network with two inputs, two hidden neurons, two output neurons. Additionally, the hidden and output neurons will include a bias.

Here's the basic structure:



In order to have some numbers to work with, here are the initial weights, the biases, and training inputs/outputs:



The goal of backpropagation is to optimize the weights so that the neural network can learn how to correctly map arbitrary inputs to outputs.

For the rest of this tutorial we're going to work with a single training set: given inputs 0.05 and 0.10, we want the neural network to output 0.01 and 0.99.

The Forward Pass

To begin, lets see what the neural network currently predicts given the weights and biases above and inputs of 0.05 and 0.10. To do this we'll feed those inputs forward though the network.

We figure out the *total net input* to each hidden layer neuron, *squash* the total net input using an *activation function* (here we use the *logistic function*), then repeat

the process with the output layer neurons.

Total net input is also referred to as just *net input* by some sources.

Here's how we calculate the total net input for h_1:

$$net_{h1} = w_1 * i_1 + w_2 * i_2 + b_1 * 1$$

$$net_{h1} = 0.15 * 0.05 + 0.2 * 0.1 + 0.35 * 1 = 0.3775$$

We then squash it using the logistic function to get the output of h_1:

$$out_{h1} = \frac{1}{1+e^{-net_{h1}}} = \frac{1}{1+e^{-0.3775}} = 0.593269992$$

Carrying out the same process for h_2 we get:

$$out_{h2} = 0.596884378$$

We repeat this process for the output layer neurons, using the output from the hidden layer neurons as inputs.

Here's the output for o_1:

$$net_{o1} = w_5 * out_{h1} + w_6 * out_{h2} + b_2 * 1$$

$$net_{o1} = 0.4 * 0.593269992 + 0.45 * 0.596884378 + 0.6 * 1 = 1.105905967$$

$$out_{o1} = \frac{1}{1+e^{-net_{o1}}} = \frac{1}{1+e^{-1.105905967}} = 0.75136507$$

And carrying out the same process for o_2 we get:

$$out_{o2} = 0.772928465$$

Calculating the Total Error

We can now calculate the error for each output neuron using the squared error function and sum them to get the total error:

$$E_{total} = \sum \frac{1}{2}(target - output)^2$$

Some sources refer to the target as the *ideal* and the output as the *actual*.

The $\frac{1}{2}$ is included so that exponent is cancelled when we differentiate later on. The result is eventually multiplied by a learning rate anyway so it doesn't matter that we introduce a constant here [1].

For example, the target output for o_1 is 0.01 but the neural network output 0.75136507, therefore its error is:

$$E_{o1} = \frac{1}{2}(target_{o1} - out_{o1})^2 = \frac{1}{2}(0.01 - 0.75136507)^2 = 0.274811083$$

Repeating this process for o_2 (remembering that the target is 0.99) we get:

$$E_{o2} = 0.023560026$$

The total error for the neural network is the sum of these errors:

$$E_{total} = E_{o1} + E_{o2} = 0.274811083 + 0.023560026 = 0.298371109$$

The Backwards Pass

Our goal with backpropagation is to update each of the weights in the network so that they cause the actual output to be closer the target output, thereby minimizing the error for each output neuron and the network as a whole.

Output Layer

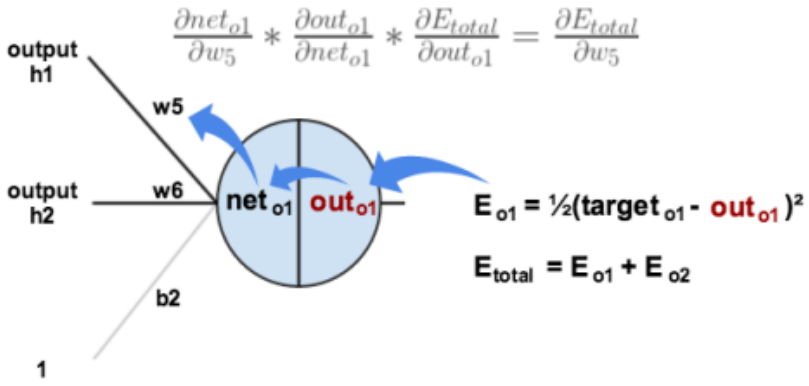
Consider w_5. We want to know how much a change in w_5 affects the total error, aka $\frac{\partial E_{total}}{\partial w_5}$.

> $\frac{\partial E_{total}}{\partial w_5}$ is read as "the partial derivative of E_{total} with respect to w_5". You can also say "the gradient with respect to w_5".

By applying the chain rule we know that:

$$\frac{\partial E_{total}}{\partial w_5} = \frac{\partial E_{total}}{\partial out_{o1}} * \frac{\partial out_{o1}}{\partial net_{o1}} * \frac{\partial net_{o1}}{\partial w_5}$$

Visually, here's what we're doing:



We need to figure out each piece in this equation.

First, how much does the total error change with respect to the output?

$$E_{total} = \frac{1}{2}(target_{o1} - out_{o1})^2 + \frac{1}{2}(target_{o2} - out_{o2})^2$$

$$\frac{\partial E_{total}}{\partial out_{o1}} = 2 * \frac{1}{2}(target_{o1} - out_{o1})^{2-1} * -1 + 0$$

$$\frac{\partial E_{total}}{\partial out_{o1}} = -(target_{o1} - out_{o1}) = -(0.01 - 0.75136507) = 0.74136507$$

$-(target - out)$ is sometimes expressed as $out - target$

When we take the partial derivative of the total error with respect to out_{o1}, the quantity $\frac{1}{2}(target_{o2} - out_{o2})^2$ becomes zero because out_{o1} does not affect it which means we're taking the derivative of a constant which is zero.

Next, how much does the output of o_1 change with respect to its total net input?

The partial derivative of the logistic function is the output multiplied by 1 minus the output:

$$out_{o1} = \frac{1}{1+e^{-net_{o1}}}$$

$$\frac{\partial out_{o1}}{\partial net_{o1}} = out_{o1}(1 - out_{o1}) = 0.75136507(1 - 0.75136507) = 0.186815602$$

Finally, how much does the total net input of $o1$ change with respect to w_5?

$$net_{o1} = w_5 * out_{h1} + w_6 * out_{h2} + b_2 * 1$$

$$\frac{\partial net_{o1}}{\partial w_5} = 1 * out_{h1} * w_5^{(1-1)} + 0 + 0 = out_{h1} = 0.593269992$$

Putting it all together:

$$\frac{\partial E_{total}}{\partial w_5} = \frac{\partial E_{total}}{\partial out_{o1}} * \frac{\partial out_{o1}}{\partial net_{o1}} * \frac{\partial net_{o1}}{\partial w_5}$$

$$\frac{\partial E_{total}}{\partial w_5} = 0.74136507 * 0.186815602 * 0.593269992 = 0.082167041$$

You'll often see this calculation combined in the form of the delta rule:

$$\frac{\partial E_{total}}{\partial w_5} = -(target_{o1} - out_{o1}) * out_{o1}(1 - out_{o1}) * out_{h1}$$

Alternatively, we have $\frac{\partial E_{total}}{\partial out_{o1}}$ and $\frac{\partial out_{o1}}{\partial net_{o1}}$ which can be written as $\frac{\partial E_{total}}{\partial net_{o1}}$, aka δ_{o1} (the Greek letter delta) aka the *node delta*. We can use this to rewrite the calculation above:

$$\delta_{o1} = \frac{\partial E_{total}}{\partial out_{o1}} * \frac{\partial out_{o1}}{\partial net_{o1}} = \frac{\partial E_{total}}{\partial net_{o1}}$$

$$\delta_{o1} = -(target_{o1} - out_{o1}) * out_{o1}(1 - out_{o1})$$

Therefore:

$$\frac{\partial E_{total}}{\partial w_5} = \delta_{o1} out_{h1}$$

Some sources extract the negative sign from δ so it would be written as:

$$\frac{\partial E_{total}}{\partial w_5} = -\delta_{o1} out_{h1}$$

To decrease the error, we then subtract this value from the current weight (optionally multiplied by some learning rate, eta, which we'll set to 0.5):

$$w_5^+ = w_5 - \eta * \frac{\partial E_{total}}{\partial w_5} = 0.4 - 0.5 * 0.082167041 = 0.35891648$$

> Some sources use α (alpha) to represent the learning rate, others use η (eta), and others even use ϵ (epsilon).

We can repeat this process to get the new weights w_6, w_7, and w_8:

$$w_6^+ = 0.408666186$$

$$w_7^+ = 0.511301270$$

$$w_8^+ = 0.561370121$$

We perform the actual updates in the neural network *after* we have the new weights leading into the hidden layer neurons (ie, we use the original weights, not the updated weights, when we continue the backpropagation algorithm below).

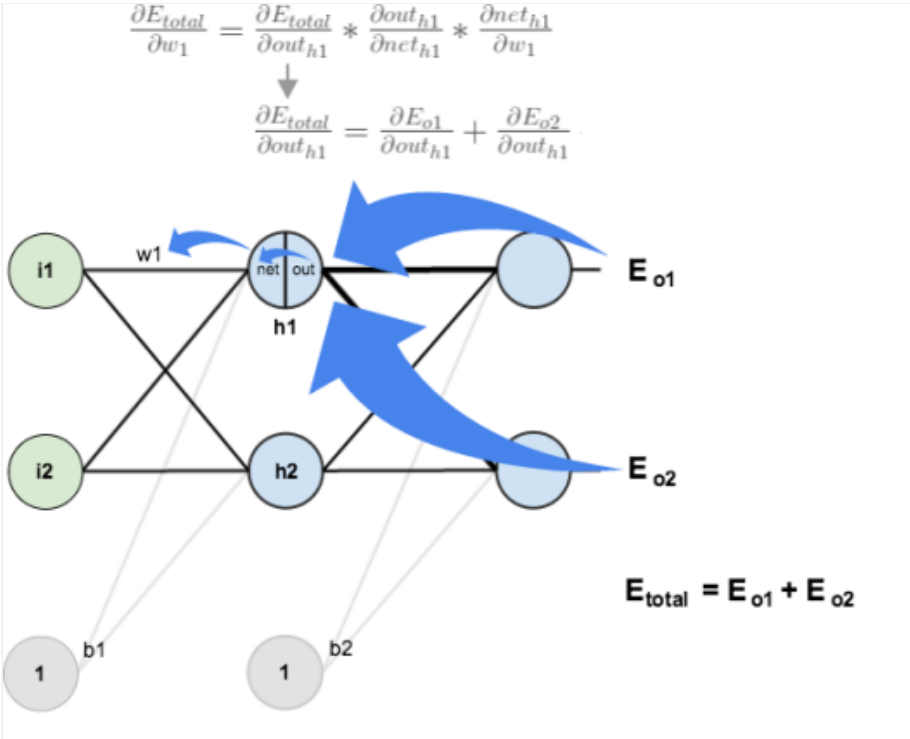
Hidden Layer

Next, we'll continue the backwards pass by calculating new values for w_1, w_2, w_3, and w_4.

Big picture, here's what we need to figure out:

$$\frac{\partial E_{total}}{\partial w_1} = \frac{\partial E_{total}}{\partial out_{h1}} * \frac{\partial out_{h1}}{\partial net_{h1}} * \frac{\partial net_{h1}}{\partial w_1}$$

Visually:



We're going to use a similar process as we did for the output layer, but slightly different to account for the fact that the output of each hidden layer neuron contributes to the output (and therefore error) of multiple output neurons. We know that out_{h1} affects both out_{o1} and out_{o2} therefore the $\frac{\partial E_{total}}{\partial out_{h1}}$ needs to take into consideration its effect on the both output neurons:

$$\frac{\partial E_{total}}{\partial out_{h1}} = \frac{\partial E_{o1}}{\partial out_{h1}} + \frac{\partial E_{o2}}{\partial out_{h1}}$$

Starting with $\frac{\partial E_{o1}}{\partial out_{h1}}$:

$$\frac{\partial E_{o1}}{\partial out_{h1}} = \frac{\partial E_{o1}}{\partial net_{o1}} * \frac{\partial net_{o1}}{\partial out_{h1}}$$

We can calculate $\frac{\partial E_{o1}}{\partial net_{o1}}$ using values we calculated earlier:

$$\frac{\partial E_{o1}}{\partial net_{o1}} = \frac{\partial E_{o1}}{\partial out_{o1}} * \frac{\partial out_{o1}}{\partial net_{o1}} = 0.74136507 * 0.186815602 = 0.138498562$$

And $\frac{\partial net_{o1}}{\partial out_{h1}}$ is equal to w_5:

$$net_{o1} = w_5 * out_{h1} + w_6 * out_{h2} + b_2 * 1$$

$$\frac{\partial net_{o1}}{\partial out_{h1}} = w_5 = 0.40$$

Plugging them in:

$$\frac{\partial E_{o1}}{\partial out_{h1}} = \frac{\partial E_{o1}}{\partial net_{o1}} * \frac{\partial net_{o1}}{\partial out_{h1}} = 0.138498562 * 0.40 = 0.055399425$$

Following the same process for $\frac{\partial E_{o2}}{\partial out_{h1}}$, we get:

$$\frac{\partial E_{o2}}{\partial out_{h1}} = -0.019049119$$

Therefore:

$$\frac{\partial E_{total}}{\partial out_{h1}} = \frac{\partial E_{o1}}{\partial out_{h1}} + \frac{\partial E_{o2}}{\partial out_{h1}} = 0.055399425 + -0.019049119 = 0.036350306$$

Now that we have $\frac{\partial E_{total}}{\partial out_{h1}}$, we need to figure out $\frac{\partial out_{h1}}{\partial net_{h1}}$ and then $\frac{\partial net_{h1}}{\partial w}$ for each weight:

$$out_{h1} = \frac{1}{1+e^{-net_{h1}}}$$

$$\frac{\partial out_{h1}}{\partial net_{h1}} = out_{h1}(1 - out_{h1}) = 0.59326999(1 - 0.59326999) = 0.241300709$$

We calculate the partial derivative of the total net input to h_1 with respect to w_1 the same as we did for the output neuron:

$$net_{h1} = w_1 * i_1 + w_2 * i_2 + b_1 * 1$$

$$\frac{\partial net_{h1}}{\partial w_1} = i_1 = 0.05$$

Putting it all together:

$$\frac{\partial E_{total}}{\partial w_1} = \frac{\partial E_{total}}{\partial out_{h1}} * \frac{\partial out_{h1}}{\partial net_{h1}} * \frac{\partial net_{h1}}{\partial w_1}$$

$$\frac{\partial E_{total}}{\partial w_1} = 0.036350306 * 0.241300709 * 0.05 = 0.000438568$$

You might also see this written as:

$$\frac{\partial E_{total}}{\partial w_1} = (\sum_o \frac{\partial E_{total}}{\partial out_o} * \frac{\partial out_o}{\partial net_o} * \frac{\partial net_o}{\partial out_{h1}}) * \frac{\partial out_{h1}}{\partial net_{h1}} * \frac{\partial net_{h1}}{\partial w_1}$$

$$\frac{\partial E_{total}}{\partial w_1} = (\sum_o \delta_o * w_{ho}) * out_{h1}(1 - out_{h1}) * i_1$$

$$\frac{\partial E_{total}}{\partial w_1} = \delta_{h1} i_1$$

We can now update w_1:

$$w_1^+ = w_1 - \eta * \frac{\partial E_{total}}{\partial w_1} = 0.15 - 0.5 * 0.000438568 = 0.149780716$$

Repeating this for w_2, w_3, and w_4

$$w_2^+ = 0.19956143$$

$$w_3^+ = 0.24975114$$

$$w_4^+ = 0.29950229$$

Finally, we've updated all of our weights! When we fed forward the 0.05 and 0.1 inputs originally, the error on the network was 0.298371109. After this first round of backpropagation, the total error is now down to 0.291027924. It might not seem like much, but after repeating this process 10,000 times, for example, the error plummets to 0.000035085. At this point, when we feed forward 0.05 and 0.1, the two outputs neurons generate 0.015912196 (vs 0.01 target) and 0.984065734 (vs 0.99 target).

If you've made it this far and found any errors in any of the above or can think of any ways to make it clearer for future readers, don't hesitate to drop me a note. Thanks!

Related

The State of Emergent Mind
In "Emergent Mind"

Experimenting with a Neural Network-based Poker Bot
In "Poker Bot"

Emergent Mind #10
In "Emergent Mind"

Posted on March 17, 2015 by Mazur. This entry was posted in Machine Learning and tagged ai, backpropagation, machine learning, neural networks. Bookmark the permalink.

← Introducing ABTestCalculator.com, an Open Source A/B Test Significance Calculator

TetriNET Bot Source Code Published on Github →

312 thoughts on “A Step by Step Backpropagation Example”

← Older Comments

Niloufar

— November 12, 2016 at 2:30 pm

thank you very much
it was so good :D

Reply

Caspar

— November 13, 2016 at 7:28 am

Really great tutorial. How would the pattern of back propagation carry on with multiple hidden layers instead of just the one? What would the overall algorithm look like where the hidden layer count, weights, etc, are all variable?

Reply

Raphael Guimarães

— November 15, 2016 at 8:01 pm

OMG you saved my life!

Reply

Lukas Cerny

— November 16, 2016 at 5:18 am

Thank you an amazing article! Everywhere else the description of the algorithm is extremely theoretically, so I am glad that you got your hands dirty by really describing the actual implementation with numbers and not just by loads of symbols, which particularly in the case of backpropagation was very confusing for me. Cheers from Edinburgh.

Reply

Caspar

— November 16, 2016 at 9:46 am

Having a read now – seems really helpful. Thank you!

Reply

shyamkkhadka

— December 10, 2016 at 12:16 pm

Hello Caspar, did you get information about how to implement it for multiple hidden layer ? I am also looking for this. Thank you.

Reply

Vadym

— November 17, 2016 at 9:45 pm

Thanks for great article!
Question: how do we tune the weights of biases?

Reply

George Pligoropoulos

— December 22, 2016 at 7:26 am

Yes how do we find the updated biases??

Reply

Mehrdad

— January 16, 2017 at 7:27 am

Hi there,
I think those are same with same layer's weight but the output value won't update.(in example suppose 1 for biases weight)

Reply

Mehrdad

— January 16, 2017 at 7:32 am

Sorry Matt please update my post with this:
Hi there,
I think those are same with same layer's weight but the value of biases won't update.(in example supposed 1 for biases weight)

Reply

Viral Parekh

— November 18, 2016 at 12:51 pm

Nicely explained ! :) Thank you.

Reply

Pingl Exploring Neural Networks… – Cluster Chord

ali hesri
— November 20, 2016 at 2:51 pm

great

Reply

Syed
— November 20, 2016 at 2:54 pm

Thanks alot for such a wonderful explanation.

Reply

Nishant
— November 20, 2016 at 9:31 pm

Awesome!

Reply

Tariq
— November 22, 2016 at 9:29 am

This was a pleasure to read with beautiful diagrams and numbers to make it real. Thank you.

Reply

Josef Kerner
— November 22, 2016 at 9:36 am

Best explanation so far, thanks so much!

Reply

C.Zhu
— November 23, 2016 at 12:15 pm

This is probably the best explanation I've seen so far. Thank you so much!

Reply

Lawine
— November 24, 2016 at 8:26 am

Thank you for your great explanation! but i have some questions about it.

in your post, at [The Backwards Pass // Output Layer] part,
(d E_total/ d out_o1) = 0.74136507.

and at [The Backwards Pass // Hidden Layer] part, you said
'We can calculate (d E_o1/d out_o1) using values we calculated earlier.'
(d E_o1/d out_o1) = 0.74136507.

(d E_o1/d out_o1) = (d E_total/ d out_o1) ?
is this two partial deviation values are same? please reply me!

Reply

dogan

— December 9, 2016 at 9:11 am

I wonder this too, is there anybody?

Reply

neyim123

— December 9, 2016 at 9:13 am

Yeah, I wonder this too, anybody?

Reply

Shree Ranga Raju

— December 9, 2016 at 12:43 pm

I guess it is because, E_O2 is independent of Out_O1 and is treated as a constant and when you take the derivative it is zero. Could be wrong.

Reply

Dubois LEE

— December 21, 2016 at 2:32 am

maybe:
E_total = E_o1 + E_o2
d(E_total) = d(E_o1 + E_o2) = d(E_o1) + d(E_o2)
so, with respect to d_out1, the second part will be zero.
As @Shree Ranga Raju said.

Reply

Mehrdad

— January 16, 2017 at 6:02 pm

I think, Yes
those are same:
E_total = E_output1 + E_output2 => E_total = E_output1 + 0 for E_o1 and
E_total = 0 + E_output2 for E_o2.
for calculate E_o1, E_output2 can't effect and is equal 0 then they have same amount.

Reply

franchb
— November 24, 2016 at 3:04 pm

Reblogged this on irusin.

Reply

Li
— November 24, 2016 at 9:49 pm

A very helpful explanation thank you.

Reply

Yair Lopez Poveda (@yalopov)
— November 27, 2016 at 12:20 am

amazing article, those diagrams are very clear. thanks.

i have same question, it was asked before. how do we update biases weights?

Reply

shelton
— November 27, 2016 at 6:54 pm

Thanks! Best explanation I have ever read.
Question: Does it make any difference if I choose bias to be -1?

Reply

richardshandross
— November 28, 2016 at 12:07 am

Hi, thanks for this explanation. What I don't understand is why the weight adjustment is dEtot/dw instead of (dw/dEtot)*(Etot)*(eta).

Thanks, Rich

Reply

sridhar reddy
— November 30, 2016 at 11:01 am

Nice explanation. Very helpful article.

Reply

jmufugi
— December 1, 2016 at 1:46 pm

Best explanation I've seen so far on backpropagation. Great job, many thanks!

Reply

Mojeeb

— December 3, 2016 at 12:37 pm

Thanks much…

Reply

Ping! Learning Machine Learning | ebc

Sam

— December 7, 2016 at 12:10 am

Why is the weight for the bias the same for a layer? For instance, for the input layer, the bias going into the hidden layer nodes h1, h1 has weight b2.

Reply

seaofocean

— December 7, 2016 at 8:15 pm

amazing! Thanks!

Reply

Ping! 7 Steps to Understanding Deep Learning – What does a Ph.D need?

shyamkkhadka

— December 10, 2016 at 12:16 pm

How can I implement it for multiple hidden layers (more than one) ?

Reply

Mehrdad

— January 16, 2017 at 6:23 pm

I think, You can do it for any number of hidden layer just you need to repeat process between output layer and hidden layer(Section Hidden Layer) for extra hidden layer you have.

Reply

shyamkkhadkaa

— December 12, 2016 at 2:51 am

Nice tutorial thanks. But how can we make this to work for multiple hidden layers(not a single hidden layer) ?

Reply

Kanchan
— December 14, 2016 at 3:46 am

Nice Explanation. Thank you.

Reply

dfdds
— December 14, 2016 at 8:47 am

Awesome!

Reply

hathal
— December 14, 2016 at 11:29 am

Thanks.

Reply

Jerzy Kaczorek
— December 15, 2016 at 6:16 pm

Very good article, however:
neth1 = 0.15*0.5 + 0.1*0.2 + 1*0.35 is not equal 0.3775 but 0.445
I suggest you examining all the calculations in order to improve your credibility

Reply

Jerzy Kaczorek
— December 16, 2016 at 1:45 pm

Sorry, I was wrong. You have 0.05 not 0.5. I made a mistake by wrong copying data.

Reply

George Pligoropoulos
— December 22, 2016 at 7:21 am

You are explaining that dE/dO1 = 1/2 * 2 * (target1 – output1) * (-1) = 0.7414
But dE/dO2 = 1/2 * 2 * (target2 – output2) * (-1) = -0.217
Above you have considered it as +0.217
By taking it as negative, makes w7+ = 0.48871 and not 0.51130 as you have calculated.
Could you verify it?

Reply

George Pligoropoulos
— December 22, 2016 at 7:39 am

Sorry my bad

Reply

Davide Quaroni
— December 23, 2016 at 5:10 am

Thank you very much, this was a great explanation and I could check the results of my neural network output by confronting them with your calculations. I also managed to improve the error reduction after 10000 cycles by updating the biases of neurons during backpropagation.

Reply

Ping! How I learn Neural Network (and Deep Learning) | Malioboro;

Tinniam V Ganesh
— December 29, 2016 at 5:54 am

Thank you so much! That was brilliant!

Reply

Aditya
— December 31, 2016 at 9:43 pm

Excellent! Thanks for writing this up.

Reply

superbo
— January 2, 2017 at 10:03 am

Thank you. This is easy to follow. This helps me a lot.

Reply

Ognyan Zhelezov
— January 5, 2017 at 8:38 am

As far as I understand, all these calculations are made to get two-dimensional output vector (target) as a result of two-dimensional input vector. I think, one linear system of two equations can give the same result. Am I right?

Reply

Murry
— January 5, 2017 at 9:36 am

It was an amazing tutorial. Succint, yet comprehensive. Thanks a million.

Reply

Adrian
— January 12, 2017 at 5:25 pm

Nicely done! Thanks for doing this – helped me a lot!

Reply

Ping! Backpropagation | Headbirths

Paresh Kamble
— January 14, 2017 at 4:39 am

Very nice explanation with sound examples.

Reply

emersonfarianobreblog
— January 18, 2017 at 7:42 am

Thank you,

Very clear and easy to understand.

It gave me an end-to-end vision.

Other articles only explain math calculus, but not a systematic diagrama.

Reply

← Older Comments

Leave a Reply

Enter your comment here...

Blog at WordPress.com.