Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-14

So, let us go back to the back propagation, right. So, today we have to do the back prop and as I told you, right, this is the figure that you have, okay and there is something called a delta rule which we will see based upon this delta that is sitting there, okay. So, let us start with, so the idea is that, right, the back is a propagation, the reason why it is called back propagation is because you actually propagate the error back from the output layer back to the input layer, okay, so as to be able to improve upon the weights, right. The idea is that you want to calculate the unknowns, right, the weights and the biases and the way you do it is by actually back propagating the error, okay and you want to make the error as small as possible. So, you use a gradient descent, you sort of update, right, you update such that the overall loss will actually decrease that is that we had a gradient descent equation, right, except that this is main. So, this is a way to solve the optimization problem, okay.

Now, so I would just like you to write, keep this figure in mind, so L equal to 1, L equal to 2, L equal to 3 and L equal to 4, this is just a very, very simplistic sort of a picture in a real situation to be far more complex, okay and then out here, right, the weights are like Wij for the Lth layer, WijL bil and you know that takes you from the Lth to the L plus 1th layer and so, right, that is why the weights are named as such and i typically refers to the ith node, okay. Of course, in between there could be little changes but generally, right, i will refer to the ith neuron and then Y1 hat and Y2 hat are the actual output, right, out of this network and let us say that, let us say that, right, you have a cost and eventually, right, what is the goal, okay. So, the goal is let us say that, let us say that we have, you know, L theta where this theta is weights and the biases and these weights by which we mean all over, okay, not just in one layer to next or something, the entire set of weights, the entire set of biases. So, let us say, right, we have a regression problem, let us just make it simple, okay, it does not matter, you can change the cost if you want but let us say that we have got 2 nodes, right, so let us just scale it by half and let us say that we have something like i equal to 1 to 2 Y i hat minus Y i square, okay.

Let us say that, let us say that is our cost which I want, which we want to be kind of minimized, Y i is like a sort of a ground truth value, okay, this is a ground truth value which we know and we want a network to actually produce a Y i hat such that for every i, Y i hat is as close as possible to Y i, right, that is the goal. And as we know, right, we want a derivative with respect to whatever, right, W and with respect to B and so on but then before we calculate that let us first write down these equations, okay. So, let me first write down the Z, okay, now if you notice here, right, Z is, red is prior to the activation, A is after the activation, right, so we will first write down Z which is prior to the activation which is simply a linear combination and A i will be after you apply the activation, right. So now I will start from Z i 4, okay or for that matter, yeah, we can start from anywhere, okay, so let us start with Z i 2, okay, I will start from this end, okay, just to be in tandem with this, so Z i 2, okay, Z i 2 if you notice, right, that will be summation, okay, this is i, okay, Z i 2 will be summation, now if you see, right, it is, no, so Z i gets inputs from, you know, X 1, so it gets inputs from X 1 and X 2 and therefore, right, this would be like summation W ij 1, right, because that is the weight for the layer going from 1 to 2 and j going from 1 to 2 X j because that is the input plus you have Bi of 2 and i going from, see I mean you have to see as to how many neurons are sitting there, right, so i going from 1 to 4, so i going from 1, 2, 3 and 4 and then you can write down A i 2 and yeah, and then this is applicable to any neuron, right, in that layer and A i 2 will be simply some activation function applied on Z i 2, right and again i going from 1, 2 up to 4 and similarly, right, we can write down Z i 3 perhaps, right, so this is the layer coming up next, so if you see, right, if you see what is coming into Z i 3, what is coming in are the weights W ij 2 and it is not square, right, that you understand, it is W ij L, so it is not like squared or something, right, W ij 2, so superscript is just a superscript, it is not a square or something and it is getting inputs from where, it is getting inputs from A i, right, that is what is coming in as input to the layer L equal to 3 and therefore, when you try to write down Z i 3, right, what will you have, so you will have summation W ij and then 2, right, because that is, that is bunch of weights and then what will you write here, A j, what is this, A j, A j 2, right, so A j 2, j going from but j will go from where to where now because it has receiving from 4, 4 neurons, right, you guys are following, right, see, I am trying to write down this Z i, okay, as a function of these weights, the biases of course associated with this which will be like, you know, B i 2, right, so here what will you have, you will have like B i 2, where i goes from 1, 2, 3 and j goes from 1 to 4 because that is where the inputs are coming to Z i, right, okay, I mean, right, if I actually write something wrongly alert me, okay, W ij 2, so j going from 1 to 4 plus what will you write here, B i, oh, we will see, here it should have been B i 1, okay, not B i 2, okay, that is why I am saying, right, so you guys should be careful, okay, the first one is B i 1 by the way here, okay and this will be B i 2 plus B i 2 and now i goes from, what will i go from, 1 to 3 because you only see, right, 3 neurons there, so i goes from, sorry, is that correct, 1 to 3, right, so i goes from 1, 2 and then 3 and then you will have A i 3 which is simply F of Z i 3 and then finally, right, we can write down Z i 4, okay, Z i 4 is this last guy, right and Z i 4 receives inputs from, so the inputs to it are A 1 3, A 2 3 and all and then you will have W 1 1 3 and then what will you write here, you will write this as, well, I mean, so B i 3, no, B i, yeah, B 1 for the first neuron and B 2 for the second neuron, so I will just write it as B i 3 and therefore, what you will have here, so Z i 4 will be like summation W i j 3, right and then A j 3 plus B i 3 except that j will now go from, j will go from 1 to 3, right and i goes from 1 to 2 because you only see 2 neurons at the output, right and A i 4 which is the final output that is F of Z i 4, so like I said, right, this F could be there, need not be there but we will assume that it is there, okay and this will be equal to your Y i hat by the way and i equal to 1, 2, right, let me just check whether we made any mistakes or, this looks okay I think because if we make a mistake then it will go wrong elsewhere, I think it looks okay. So now the point is this, right, so now we have these, we have this output relations at every layer and the idea is this, right, finally what do you want to do, you want to be able to see, the idea is this, right, you start with a set of weights and biases which could be randomly initialized, we will see what is the best procedure to initialize and so on but for the time being assume it says randomly initialized then what will happen is for this input, right, X 1 and X 2 you will get some output, right, I mean, if you simply do a forward sort of a propagation, if you forward propagate you have the weights, you have the input therefore you will get outputs at every point, right, every neuron and that will fire the next neuron and so on and eventually you will get some Y 1 hat and Y 2 hat but that need not be equal to Y, that need not be equal to Y 1 and Y 2 which is a target sort of, you know, target value.

So what do you do, you find the error and then you do a back propagation of the error, so you try to go back and see, right, how this error, how the weight should be adjusted so that the error will come down, weights and bias always, when I say weights I mean the bias also, ok, that is the idea of this one back propagation. And that is why it is clean because, you know, if once you write this down for a small this one that you at least know how it works and then one does not even write it in a real, when you do a real implementation there are packages that do this automatically, you just define the cost and then you define the network then they do it automatically but I thought once at least you should do it so that, right, you understand what the intricacies are, right. Then let us now, now the point is right, now let us kind of go and look at the, look at what to say, dou L by let me say, let us start with the first one which is dou L by dou W ij 3, ok, dou W ij 3, ok. So if you see here, right, go back and see this diagram, so you have like, you know, dou L by say dou W ij 3 which is here, right, W ij 3 is here therefore, I mean and it is all an application of chain rule, ok, the whole process is simply an application of chain rule, ok. So this you can, so right, so this I am going to write it, I mean, right, you will, I mean, you can verify this for yourself, so this will be, right, so Zi 4, right, so this will be like dou L by dou Zi and now and we write it in a particular form because this involves what is called a delta rule and that involves doing it in a certain way, ok, not that this is the only way to do it but Zi 4 by dou Zi 4 by dou W ij 3, ok.

And of course, right, we will have to, we will have to find all these values independently but now, but let us just kind of look at, look at this equation that we had, ok, see Zi 3, ok, right, Zi 3 is in this form, right, this is what you have, ok and let us go to this next one. So you have Zi, so see Zi 4, sorry, you have to look at Zi 4, so Zi 4 is here and, and if you do dou Zi 4 by dou, write W ij 3, what will you get in this equation? Aj, Aj 3 is what you will get, right and therefore that is what, that is what will be the second term there, sorry, this is going to be Aj 3 and let us call this as delta i 4, ok, this is just, just a notation, ok, this quantity, you know, we will, we will, we will explore this a little, you know, in more detail, for the time being just call this delta i 4, ok and then this is Aj 3. Then, then let us look at, what did delta j by dou W ij 2, ok, this again, right, we will write it as again, ok, dou L, sorry, no, not dou j, dou L and this will be like dou L by, I mean, again, right, I mean, you just have to follow whatever, whatever we did just now. So now we are looking at dou L by dou, say W ij 2 and that you can write as dou L by dou Zi 3 and then into dou Zi 3 by dou W ij 2, right. Therefore, you can write this as dou Zi 3 and then dou Zi 3 by dou W ij 2 and this, now just by a notation, right, this will become delta i 3 now and Zi 3, right, if you just watch this equation, Zi 3 is here and if you take with respect to W ij 2, you will get Aj 2.

Therefore, right, this becomes Aj 2, ok, this becomes Aj 2 and then we will have finally, dou L by dou W ij 1 and that will be dou L by dou Zi 2, right, I do not even have to see the diagram, right, this is what it should be like and dou Zi 2 by dou W ij 1 and I just expect this to be Aj 1, right, which you can verify Aj 1 and this is going to be delta i 2, right, oh sorry, ok, now, yeah, this is one change, right, because of the input, right, what you have is X j coming in, right. Therefore, it is not Aj, so Aj becomes, I mean, there is no Aj there, right, I mean, if you just notice, I mean, when you talk about Zi 2, the input is X j, right, the rest of the places you had Aj, Aj and all, at the first layer you have X j, right and therefore, this becomes, so I will just put this as X j, the last one alone, right, so this, let us not call it Aj, let us call this X j, there is no 1 and all there, earlier ones had X j 1, Aj 2, this is just X j, ok. Now, right, this is on the one hand, right, which we have. Now, let us actually examine this one, right, delta i 4, ok, let us examine delta i 4, right, delta, because see, the idea is that you want to do a back propagation, so you have to start from the output layer and come all the way to the input layer. So, if you can find out delta i 4, then the idea is that we want to be able to find out delta i 3 in a sort of a, you know, a recursive manner using delta i 4, then delta i 2 using delta i 3 and we want to kind of go like that, we can show that it can be done, ok, that is called a delta rule, delta learning rule, that is what it is called.

So, let us just look at delta i 4, right, so delta i 4 is dou L by dou z i 4 and L, right, we wrote down it, L is simply half whatever, right, I mean that half is because we have 2 outputs there, i equal to 1 to 2 and then we had y i hat minus y i square, right. And therefore, and this y i hat, ok, now this is, so if you look at this, right, so if you do dou L by dou z i 4, right, that will become, 2 will cancel off, right, you will have like y i hat minus y i and then you will have dou y i hat by dou z i 4. But y i hat, right, if you observe y i hat is what, F of z i 4, right, and therefore, this is simply F dash of z i 4, right, this is simply F dash of z i 4, right. And we know that y hat, y hat we know the value actually, we can actually compute it because y i hat is coming through the input and then the weights, right, that we have initialized with. Therefore, actually, so in that sense delta i 4 you know now, so this is the forward pass in a sense, right, you do a forward propagation with a certain set of weights that you randomly initialize, you will get a y i, you will get the output, I mean you can compute delta i 4 because y i hat is known to you, y i is the ground truth which you know what you want it to be, then y i, right, F dash of z i 4, right, everything you know.

But now the more interesting part is how does delta i 4 connect to delta i 3, how does that connect to delta i 2 and so on, I mean that is where the power of this whole thing lies in a sense. Sir, there is a summation. Where? In the delta i section. No, you are doing with respect to z i, right, is there a summation? No, because you have y as a subscript i, you know, therefore, it will only work for that i, right. You are doing z i 4, right, it is with respect to particular value of i, right.

So summation would not be there. So delta i 3, let us kind of look at delta i 3. So delta i 3 is here, right. So delta i 3 is dou is dou L by dou is dou z i 3, ok. Let us just go back to that figure, right.

So what you have is, so your L, L is out here, right, that is your cost. Whenever I write L that is at the output, right. So you are talking about dou L by, you see, dou, dou you see, what you call, dou, you know, z i 3, right. So yeah, so one way to, one way to write it again, right, this is not the, this is not a most unique way, but what I can do is I can write this as dou L by dou a i 3 into dou a i 3 by dou z i 3, right. I mean I can come like this, right.

You can come like dou L by dou a i 3 into dou, what is this, dou, no, dou L by dou a i 3 into dou a i 3 by dou z i 3, right. So I will write it in that form. I will write this as dou L by dou 3 into dou a i 3 by dou z i 3. Now if you see, write a i 3, I mean, so if you change z i, right, what gets effect, if you change z i 3, right, what gets affected is a i 3, okay, and in fact a i 3 is f of z i 3. Therefore this is simply f dash of z i 3, right.

This simply f dash of z i 3. But then something, something will happen here. This is a little more tricky because if you change a i, right, if you change a i, okay, it is not just, it is not that it will affect only z i 4, it will also affect z 1 4, z 2 4 if you had further neurons in the output, right, it will affect all of them. See, right, I mean you see the connection, right, I mean if you change z i, it affects only a i and therefore there it is only, it is only one term. Till now, right, we did not encounter a situation where if I change one thing, right, then there are, then there are, see, multiple things, right, you know, right, which get affected.

This is a chain rule which I am sure you are all aware of, right. Maybe you must have done it your 11th and 12th and all, right. So, what this means is that this dou L by dou a i 3, now that will have to be expanded now, right, it will be, it will have to be written in terms of a summation now because changing a i 3 does not affect, you see, you see, you know, because, right, this input is going here also, right, I mean I am just, I am just drawn with one line, but, right, if I had multiple guys down here, it will go, it will go to every one of them, right, and therefore every one of them will get affected, therefore you have to take that into account, right. Therefore what we will have to do, so, right, this term alone, so if you just examine dou L by dou a i 3, this term alone, right, we have to be careful. So, this we will have to say that, right, this will, this will, so where is our a i 3? So, a i 3 is here, right.

So, a i 3 is going to affect all of them and here we have taken a simple example. So, we can write this as, just to, we will keep our notation simple, we will say j equal to 1 to 2 because there are only 2 neurons that get affected by this guy, right, in our case, ok. Again, it depends, I am writing it for the specific case, for the specific simple example, so, right, j equal to 1 to 2, then where is this? Ok, then we have dou L by dou z i 4, where is this? Now we want, yeah, so you want what? You want, what do you want? You want dou L by dou a i 3, right, what is this, what is that we want? You want dou L by dou a i 3, right, and therefore we can write it as dou L by dou z i 4 and dou z i 4 by dou a i 3, ok, as a summation.

So. Z j 4. Z j 4, yes, because the summation is over j, yes, you are right, not i, because I am summing over j. So, dou L by dou z j 4 and dou z j 4 by dou a i 3. Is this ok? Now if you look at, if you look at dou L by dou z j 4, what is that? Delta j 4, right, this guy is delta j 4, not i, j, ok. Now z j, dou z j 4 by dou a i 3, right, so let us again go back to our equation. So let us see z j 4 relation with, I do not know why this goes off.

See z j 4, where is our z j 4, is here, right, but here we wrote z i here and then a j was coming on the other side, now we have it the other way, right. So what will that be equal to? Now if you do dou z j 4, right, by dou, what is it, a i, that means you have to just replace the summation, I mean j i here, i there, it will become w j i 3, right, it will become w j i 3, we cannot write w i j, because it is reversed, right. So therefore this will become w, so this will become w j i, what is it, 3, right, w j i 3, ok. Then what this means is that I can then go back, right, see the whole idea is to be able to write this, so I will write this as delta i 4 is equal to y i hat minus, where is this, so f dash, ok, no, no, we are writing delta i 3 in terms of delta i 4, so let us write delta i 3, right, is equal to dou L by dou, so this is like summation j equal to 1 to 2 delta j 4 w j i 3, this whole thing in a bracket, right, because this is summed over j and then, ok, right. So basically this term is this and then into f dash of z i 3, this is delta i 3.

Now let us actually look at, ok, now maybe I will do it here itself, let me do delta i 2 because we want to be able to know once and for all, right, what will happen. So I think, right, delta i 2 will then be, right, from here, right, you can see delta i 2 is dou L by dou z i 2, again following the same thing, right, I will write this as dou L by dou a i 2, I will just follow the same thing, right, and then dou a i 2 by dou z i 2, this again will become f dash of z i 2, right, I mean that is straight forward. Now dou L by dou a i 2, right, so if you see a i 2, so here, right, so a i 2 is here, so a i 2 affects how many neurons, I mean three of them, right, so it will affect like z 1 3, z 2 3 and z 3 3, right, if you change any of the a j 2s or a i 2s and therefore, right, therefore what will happen? So you have dou L by dou a i 2, so this I will write this as summation, again I will write this as dou L by, what do you have here? For a i 2, dou L by dou z i 3, right, dou L by dou z i 3 and dou z i 3 by dou a i 2 or a j 2 in this case. So you get dou L by dou z i, no z j, okay, because we have i on the left now, so I cannot put i here, so this is delta i 2, so we will again change this notation to j, we will go like j equal to but j will go from 1 to 3 now because three of them will get affected, so we will get like j equal to 1 to 3 and then you have got like dou L by dou z j 3 and then dou z j 3 by dou, what is this, a i 2, right, alert me, okay, if everything is, if we are making you know a mistake somewhere. Now this is delta 3, right, this is our delta j 3 and z j 3 by, okay, so this again, right, you can go back and check that.

So z j 3 is here and then you are taking with respect to a j 2, therefore it will become w j i 2. So this becomes summation delta j 3 w j i 2 f dash of into, okay, f dash of z i 2, right. So which then means that, which then means that right we can write a recursive sort of a relation, recursion I mean along the path, right, which looks like this, right, so what can we say. So we can say delta i L, right, is equal to, we can write this as summation, let us say j equal to 1 2 and I am going to write this as, let me use some sort of a notation, S L plus 1, okay, what does that mean, S L plus 1, what should it mean, number of neurons in the L plus 1th layer, okay, so S L plus 1 is the number of neurons in the L plus 1th layer, number of neurons in L plus 1th layer, L plus 1th layer. And then we have, what do you have here, then dou j L plus 1, right, see here, you know, dou i 2 is summation dou j 3, so this becomes dou j L plus 1, then w j i L, right, L whole thing multiplied by F dash, then Z i L, right, I mean one, of course, you know, one can also write this in a sort of a matrix vector form to make it look more elegant, but for the time being read it suffices that we have, you know, so we know that, you know, that we start from the output, right, I can get my series, right, delta i all the way back or delta i all the way, okay, and i and all, right, will change accordingly, okay, in whatever layer, sorry.

Will i go from 1 to SL? I will go from 1 to SL then, yeah, I will go from 1 to SL, yeah, if you want to follow the same notation, you say i will go from 1 to SL, right. Then, okay, let us, so what this means is this, right, so it means that, so it means that all of these weights, right, I mean this, the gradient which is what we are ultimately interested in, right, we are trying to, why are we doing all this? It is because we want to apply a gradient descent, right, and to apply a gradient descent you need the gradient of, gradient of the output with respect to the unknowns and the unknowns, right, are for us the weights and the bias. The other thing that we have not done is the bias now, right, so we have only done the weights, so let us just do the bias part now.