Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-21

I mean, so when I talked about dropout, right, did I give this probability Ps, the probability of selecting a neuron or actually dropping a neuron? What did I write Ps, probability of dropping a neuron, is it? Because when I went back I was just thinking what did I say? Did I say drop a neuron or select a neuron? Drop, okay, then that weight should be 1 - P. Because it is like saying that, I mean, if I drop for sure, that means it is almost like probability 1, okay, that means I am almost certainly going to drop it, then its weight should be 1 - P, it should actually be much less, right, because you are certain to drop, okay. So if it is a probability to actually pick, then it should be P times, the weights, the activation should be multiplied by P. If I had written actually probability as the probability to drop a neuron, it should be 1 - P, okay. None of you mentioned that but I just realized that, you know, I wasn't sure what I meant as P there, okay. Just a small sort of a correction there.

Okay, now let us just go on to this initialization of weights which is the last thing on, you know, which is regards to MLP, so initialization of weights. Similar to the way we initialized or sort of, right, I mean we did something for the inputs which is a preprocessing. Similarly, one should also be careful with respect to how you choose the weights. I mean if you choose your weights arbitrarily, right, I mean, you know, let them go to any range, you know, that they feel like then there could be trouble, I mean when you start the entire process.

So right, I mean, so this, what do you call this, researchers have evolved a systematic way to actually do the initialization. First of all, you know, it is very clear that the weight should be very small and the simplest thing to think about is that W should come from a uniform sort of a distribution between let us say - A and A, okay, where I mean A is some small number, where A is a small sort of a positive number, okay, very small positive number, okay. But then this also works, okay, to some extent. I mean if you start with some very small value like A could be like 0.001 or something, it will still do reasonably well but there is a more formal thing, right, that let us say people have found out, I mean which I will think and I will just mention here and I will also, I will let you figure out, right, that do not choose weights to be a constant or 0, okay, that is something that I want you to figure out why, what will go wrong.

If okay, weights should not be, that means when you start, right, should not be a constant, should not be 0, should not be initialized to 0 or a constant, should not be initialized to 0 or a constant value, okay, this I will leave it you to figure out why you should not do that and what you should ideally do is, you know, pick it to be uniform and this paper, what is that, 2010, it is called Xavier initialization, okay, this is one else, this is an author's name. So the way he does the initialization is as follows. So according to him you should draw from uniformly from - root 6 by root of fan, root of fan in plus fan out to root 6 by root of fan in plus fan out. I will just, I will just, I will just write, indicate to you as an outline as to how they, how we arrived at this so that it does not look like some crazy thing. So for example, right, if you look at the activation, right, and this initialization, so this fan in is like, right, number of, number of, number of inputs into a neuron, again that layer, okay, in some ith layer into a neuron and this initialization that we are talking about will happen at every layer, okay, into a neuron which will be the same as the case for every other neuron and fan out will be the number of, number of outputs going out of the neuron, going out of the neuron in that layer which will be the same for all of them in fact.

All the neuron, any neuron you pick, right, the number of inputs are coming into it and the number of outputs that are going out of it, for all neurons it will be the same, therefore you can just look at it at any one neuron. So fan in is this and fan out is this and so suppose, suppose you look at the activation at any ith neuron and suppose you start with the first layer, right, where the input is being applied and then you have let us say your Ai to be, let me just write this down, so Ai is equal to let us say summation, let us just ignore the bias and all for the time being Wi1 Xi, where Xi is going from 1 to n, that means I have got n inputs and I am kind of looking at the activation in the ith neuron and I am at the first layer, so the input is all Xi's. So as you go forward, right, the inputs will be Ai themselves, right, I mean those will be the input of the next layer and so on. Now if you look at variance of, okay, okay, let us call this, sorry this is not Ai, this is A1, okay, Ai is there, Ai is on the right, okay, this is A1. So if you look at variance of A1, this is at, this is at let us say now at one particular neuron, okay.

So if you look at variance of A1, right, this you can show, I mean assuming that both X and W are random, you can show that, right, this is simple, I mean so you can show that this is n times variance of W into variance of X. I mean you just have to assume the statistical independence between X and W and this is easy, right, I mean you just have to expand it out, take the variance, so all the cross terms will vanish and the only thing that will survive is W square X square, take the variance, so that will be simply a product, right, that is not, they are independent. Now when this goes, now this is at the, this is the first layer but then at the, you know, second layer, right, what will happen? This where X, right, which is what you are taking as the initial input, so this where X when it, so X is

now the output of the first layer, right, that becomes the, you know, input to the second layer. So what will be the variance of the input to the second layer? What will be the variance of the input to the second layer now? N times, n times where W, you know, because, right, that is what, that is what is the variance of the, variance of the, you know, first guy, right, so I mean it is like this, right, so you will be looking at X, right, this X is going in and then you have a first layer and the, the output of the first, of course, you know, we are, we are ignoring a nonlinearity and all that, right, just this again, this is just a sketchy thing but gives you some insight, right, and then, and then, right, I mean, you have the, you know, second layer and so on. So assuming that, so, so, so, right, X is here, correct? Now initially, initially you had, you had a variance for X, right, now the variance that is let us say, let us call this as, this is not really input, right, but then this is an input for the, right, next layer, so let us call that as, say, X dash, okay.

So that X dash will then be n times, n times variance, variance, so, I mean, so at any kth, let us say depth, right, what will you have? You will have like n times where W to the power k into where fx, right, you see this, right, I mean, at the second layer, at the third layer, so, so basically what will happen is every time we will get, right, n times where W times where X and then you go to the kth layer, it will be like n times where W to the power k if you are in the kth depth layer times where fx. So which then means that actually things can either blow up or you know, things can actually, you know, things can actually just shrink because it depends upon what happens to this n times where W sort of quantity because our focus is now on the initialisation of the weights. So this n times where W, right, if it is a number that is actually greater than 1, then your variance of ak will all blow up and if n times where W is something that is actually less than 1, okay, then your variance will actually shrink. But if you want some kind of a balance to be maintained, then probably, right, what you are asking for is something like n times where W to be equal to 1. So that somewhere, so that along the way the variance is not getting either, right, too blown up or too shrunk, right.

So if you want to maintain some sort of order, right, inside it, then one way to kind of look at it is say that n times where W, let me, let me, you know, pick my variance of W such that n times where W is 1. But in this case we are only looking at n which is the, which is the, which is the dimension of the input. But if you look there, right there, it is taken as, you know, both the dimension of the input as well as the dimension of the output, right, is also taken into account. But for the time being, right, again, just to keep matters simple, we can simply argue that if you want your variance of ak's to be somewhat stabilized, then what you need is n times where W, you can ask that or you can say that where of W should be equal to 1 by n. Okay, now if you, if you, if you, but then, then you have say W being, let us say, if it is uniform between - a and a, okay, then what will be the, what will be the, what will be the mean of W? 0.

0, right and then, then if I want to, if I want to write, estimate the variance, right and so, right, this is going to look like this, right. So you have got like - a to a, therefore, it should have an amplitude 1 by 2a, right, in order to make sure that your area under that is 1. So we look at this as your Fw of W, right, that is how it will be, right. So if you, if you actually compute variance of, variance of W, right, that will be integral, yeah, X square, so that is X square, Fx of X, which is 1 by 2a dx or dw, right, whatever it is that you want to use, okay, W if you want, then - a to a, that will be like what, 1 by 2a X cube by 3 - a to a. So you will get what, X cube by, so it is 1 by 2a and then 2 X cube by 3, so you will get X cube by 3a, okay, no, I think, right, let us all put this, a cube by cube, not X cube, a cube by 3a, which is, which is again a square by 3, sorry.

So your variance of W is actually a square by 3, right and that is where, that is what you have, for example, just now we showed, you know, in the other one, so we said that where of W, right, should be, we said equal to 1 by N, right, where N is the, N is the dimension of the input, dimension of the input, okay. Let me write big, of the, of the input, right, but, but then suppose, let us say, right, you want to take both the input and the output into account, then what will happen is this will become 1 by fan in plus fan out. So this will become where of W, right, I mean, and, and of course, and the only thing is, right, you need to, you need to, you need to, you need to scale this by 2, okay, by 2, I mean, so you should take the average and therefore what will happen is that this will kind of go up and then you will get 2 by fan in plus fan out, I mean, if you are just simply taking N, then it would have simply stayed as N, but since you are taking an average of both the input and the output, right, you should, I mean, so again, right, this is something that the, that the author writes in that paper. This is taken as fan in, what is fan in plus fan out by 2 or 2 by fan in. Now where W, right, in terms of, and finally your goal is to find out what is that appropriate A, right, you know that, I mean, you want to be able to draw from, from uniform, okay, uniform - A to A, right, that is what you want to draw from, but your idea is to arrive at that A, right, you do not know what that A should be and therefore, right, what you have now is A square by 3 or A square is equal to 6 by fan in plus fan out, plus fan out or A should be equal to plus - root 6 by root of fan in plus fan out, plus fan out, okay, which is, which is what, which is what is that Xavier initialization, right.

If you go back here, okay, this is what I wrote here as, as actually at the, the, the, right, initialization and in a way, I mean, you can argue, I mean, it is not completely arbitrary, it is not like, you know, somebody just thought that it is a fancy thing to take it like that. So, you can actually go around and show to a reasonable extent that, you know, with some reasonable arguments, of course, you know, we have been involved in nonlinear entities and all, somewhere, right, there is a little bit of hand waving here and there, but that is okay, I mean, just to give, give an insight into why, why this is chosen that way. In fact,

there is a, there is a later paper, I think, you know, that is by Kaiming Hay and that came later and that in fact claims, you know, I would not, I would not go to the details of that, but that came, that claims that draw from, what is that, - 4 by fan, by root fan in plus fan out, 4 by root fan in plus fan out, okay. It says draw from this and you know, both, both work equally well and right, when you, when you implement something right, you will have to, you can state which one of these things that you want to use for your initialization.