Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-25

So, last class we saw what was an AlexNet, we will also see a few others because these are the ones that are very well known and you know it is important to at least know what they do and what sort of new ideas did those new ones, I mean so as kind of years advance people incorporated new things, so we just want to see what those new aspects are with respect to these deep networks. So the one that we saw last was AlexNet which had a 16.4 sort of a top 5, this is a top 5 accuracy, basically means that the right label that you are looking for is appearing at least in the say top 5, I mean top 1 could be a little lower. Now I mean prior to this VGG which is a very, very important sort of deep net, prior to that in between this AlexNet and VGG there was another network called the ZFnet, that is the one that has this 11.7, I know it is called ZF after Zeller and Fergus, Rob Fergus, it is 2013 network, so this is actually the ZFnet, it goes after the author names Zeller and Fergus and you know just to know as to what they did, they were able to bring down the error from 16.4 to 11.

7 and they actually followed exactly whatever was there in AlexNet except that it was mainly a tuning of the hyperparameters. Hyperparameters in the sense that some of the things that they actually did was if you kind of if you sort of recall in AlexNet the first layer had the filters of size 11 cross 11 with a stride of 4 right, now that they changed, now again you know they just tried a few things and what they did was they took 7 cross 7 stride 2, then you know so this was with respect to the first layer okay, this is the con 11 layer right, so those all with respect to ZFnet that I am saying right now and then with respect to con 3, 4 and 5 okay, if you again say recall we had filters of the size 384 and all that right, instead of that they made it like 512, 1024, 512 again these are number of channels, number of feature maps, so I think last time we had 384, 384, 256 something like that right with respect to AlexNet. Now these people kind of played around with those numbers and found out that if you use more feature maps instead of using what AlexNet had done okay, but architecture is still the same but architecture in the sense that the basic sort of skeleton is still the same, but in terms of the number of feature maps and sort of the filters that you are applying right those were changed. Well, no I think you know this has about 1.

5 million more, so if you look at I mean that was about 60 million parameters right, so this

comes down 61.5, this is in million parameters, it is just a little bit above, but then for that 1.5 million extra that you know that you are incurring, but then you have see drop is quite significant right, I mean that is the reason why it should also be mentioned that is why I am not skipping it okay. So top 5 error is 11.7% and yeah, so I think compare this with 60 million from say AlexNet okay, but yeah but then other than that right they did not change anything, there were no new ideas that were actually brought in just that so it is mainly you know hyper parameters playing around with the hyper parameters.

As I said hyper this in parameters not just includes weighting the cost function and so on, but also includes how many channels you have per layer, how many layers you want all those are hyper parameters okay. So okay then we will move on to the next one which is actually VGG okay, which okay which came from the Oxford group okay. So I will write this as VGG net okay, it is called VGG net, this is visual geometry group okay, that is the expansion for this group from Oxford. The authors and all right it is easy for you to find it, I mean it is actually a very well-known person that Azizaman and a student okay and a student, Simonian okay that is his student's name. Now this VGG net let us see okay what it looks like, let us just go back and see that what it looks like.

So one of the things that you see is that it is actually a deeper network because Alex net as well as ZF net they had about 8 layers right at that time when we talked about, now this is about 16 to actually you know 19 layers. In fact I mean you also have you know sometimes they write this as VGG 11, VGG 13, VGG 16 and VGG 19 okay. So you have all kinds of, each one of them means that 11 layers, 13 layers, 16 layers, 19 layers and so on. And this figure is not exactly correct okay because somewhere I think some layers are missed out okay. So for example if you look at VGG 16 if I count 1, 2 then 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 right there is one 16th layer that is I think you know that should have been somewhere here that is actually a convolution, one more convolution that is been left out but that is okay I mean you know so I think you know those are minor things.

So if you really try to count I mean the 16 does it match or it will match I mean so there is a small error okay in these figures but then the idea is to just see as to what they have done right. Now one of the things so as compared to 8 layers in AlexNet I mean now you are kind of see you know it is a kind of a more deep net you know it is a deeper network you know with 16 you get to see 19 layers and this is actually become an important sort of a network because you know when trained on this image net okay of course all these were actually you know done with respect to a challenge right that is that first challenge that bar chart that I showed you that is that you know large scale image sort of see recognition challenge which means that all these had to be trained on that you know image net which is a big database and you have to show that you have to show how well you are doing on that. Now it turns out that you know that original sort of a database which is

image net okay when you train these this VGG net or for that matter AlexNet and already trained net what happens is the output features that you get but one is that you simply take the network you do a classification but then the you know interesting thing is that before so you could also use the feature representation right that is emerging from this kind of from these architecture for example right you can even look at the con features right that are appearing here or for example right prior to the prior to the last layer right I mean the FC 4096 right at that point of time right what kind of features you are having it could be that you know some of those features are so very useful for let us say other tasks when you want to do a kind of a representation. So now it is for example you know people have used you know a VGG network to do style transfer and all this turns out that you know people have figured out that those features are even important for certain other applications okay. So whereas you know when they were actually built they were primarily built for a classification challenge as well as there is a detection segmentation things like that so these are all built for such challenges but then it turned out that such networks after training that means you do not have to retrain them you always have the option to retrain if you have enough number of examples okay you can always retrain but then even without retraining if you simply had the network with pre-trained weights right that is what it is called it is called pre-trained weights that means somebody for example these authors have themselves trained the network the weights and weights and all are available you do not have to retrain them so if you actually push an image inside it you will get feature maps at all levels.

Now some of those feature maps people have found out are actually useful for certain applications okay so when we actually look at these architectures we should not be looking at them as simply trying to solve a classification problem though of course that was the main goal but at the same time realize that these architectures are so general in the sense that you know the feature maps coming out of them even out of pre-trained networks are useful you can also of course retrain them if you wish if you had enough number of examples for a particular problem okay that is an aside. Now one of the things that it does is read advocates have no higher depth okay and okay now what it does is so it says so the idea right being that you use smaller small filters but then use a deeper layer so this is like kind of going back to a receptive field kind of thing right for example AlexNet had you see right I mean 11 cross sort of you know 11 filter right at the con1 layer which means that receptive field right is that large but then you know but their argument was instead of trying to have a receptive I mean have a filter that is that large why do not you have let us say multiple layers with smaller filters for example right that is what you are seeing here right 3 cross 3, 3 cross 3, 3 cross 3 so something like that when you have okay multiple layers 3 cross 3, 3 cross 3, 3 cross 3 what will be the if I have 3 such things what will be the effective receptive field that means this is the first guy this is the second guy this is third guy what will be the receptive field of this 7 cross 7 right. So the idea is that you know effectively the third guy will see a 7 cross 7 sort of a region and though of course

you know it is not the same as seeing you know 11 cross level but if you look at ZFNet that had 7 cross 7 filters right so their argument was primarily they were trying to say that instead of trying to use a large filter because then your unknowns even if you keep it a 7 cross 7, 7 cross 7 is still 49 unknowns whereas 3 cross 3 right used to 3 times is still roughly you know even kind of half that number right not exactly yeah but still pretty so what they were trying to say was do not use large filters try to use smaller filters but then add you know sort of depth okay into the this one network that is the reason why it goes up to 16 to say 19 layers and it is not only the it is not just that just that right it is also that in terms of the accuracy final top error rate and already you are able to bring it down I mean that is why the argument becomes stronger it is not simply a receptive field that you are trying to achieve with fewer you know with fewer unknowns but also the fact that all of this leads to an improved accuracy. So they have like you know 3 cross 3 constraints 1 pad 1 and all that right now so and then what else right is there and of course you know you might ask why did they stop at VGG 19 and why did they not go to some VGG 25, 26 and all but what they already found was again for the you know imaginary challenge okay it did not make a big sort of a difference in jumping from 16 to actually 19 so whatever advantages they were right they were able to get out of it you know having a deeper network kind of flattened off right beyond 19. So you do not have something you know more layers than that and again these are all empirical right somebody has to do the studies to find out you know where you sort of where you sort of right hit that hit the sweet spot in a sense so VGG 19 and then you know you do not have more than that.

Then is there anything else that I need to tell so more again right I mean if you see the number of unknowns right actually pretty huge okay why because of the fact that you know you have VGG net right let me see so 120, 122 million right so the number of parameters it is actually 120 million and that is like double that of Alex net right so number of let us see parameters is 122 million okay and that is coming mainly right from this layer in fact I mean you know if you look at from going from here to here right this jump okay if you see right you got like 3 cross 3 con filter and then 512 filters I mean and then if you try to if you try to flatten that out right and then if you link it up with a with a with a 4096 fully connected layer when you when you do the full connection right there itself you incur about you know about roughly 100 million or something right so here itself right you are actually incurring a huge number but that is okay I mean right by then GPUs had come and then you know it was not I mean computation was not such a big issue so you could like you know you could go from 60 million to 120 what is that 122 million and still be able to sell the architecture because of the fact that the computational power was there and then you could actually bring down the error okay significantly because right beyond that I mean every kind of little thing matters you see when you jump from 25 to 16 that is a big jump but then after that right it starts to flatten off 16 to 11 is very hard and 11 to 6 is far more hard because right then once you come closer and closer to what you want to achieve

okay things become more and more hard and therefore any sort of architecture that allows you to get there is okay I mean you know provided it is you know provided you know it is going to be reasonable and it makes sense. So and as I said right so one can use the last FC layer or the common layer for an abstract representation okay you can also use that as an abstract sort of a representation okay which basically means that like I said right you can actually use that as a feature for something else okay so you just push an image inside you get a you get a you get a feature for that and that feature right what you would like to do with that feature is entirely up to you but then that is like just like a succinct sort of explanation of that image right taking into account that you know that this network has seen millions of images and knows how to sort of how to how to arrive at a feature map okay. So I think so 7.3% top error okay in that challenge and so it came down from 11.7 that is what you had for ZFnet and this when this VGG is actually a very very this whenever you know a well-known sort of a network a popular network and you know people use it left and right.

So I think yeah so I think I have explained okay most of this fewer parameters deeper more nonlinearities but fewer parameters in the sense that only in some places but then most of the this one a memory is in early con that means I mean you know the so when you say this one memory that means the number of neurons right that is what we mean last time itself I think I told you but then the this one the unknowns the parameters right I mean here is where when you jump from the fully from the conv layer to the fully connected right there itself we are incurring something like 7 cross 7 into 512 into 4 so you know 100 million 102 million so here it is already incurring a huge huge this one 122 million is what I have here but it says total parameter 138 million okay I think this has to be changed okay. In okay I think second this is all okay I think you know.