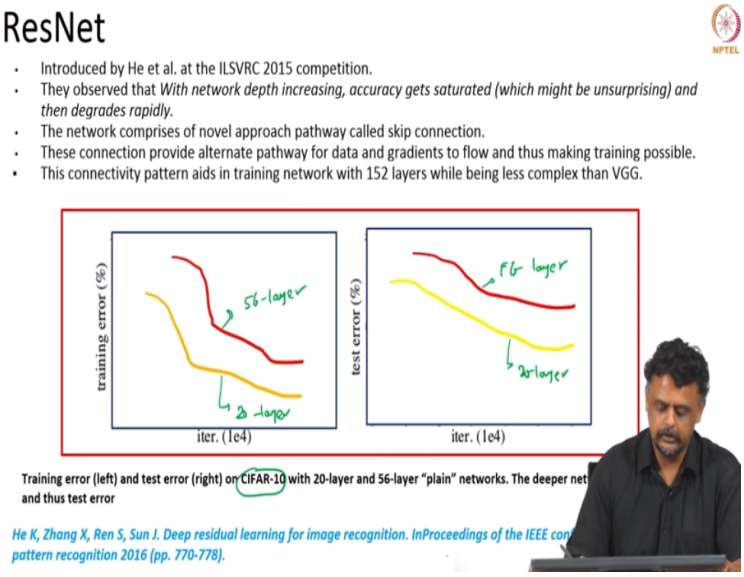
Machine Learning for Engineering and Science Application
Professor Dr. Ganapathy Krishnamurthi
Department of Engineering Design
Indian institute of Technology Madras
CNN Architecture (Google Net)
Part 4

(Refer Slide Time: 00:14)



In this video we will look at ResNet, so the previous video we looked at AlexNet as well as VGG now those layers for instance, AlexNet that at seven layers VGG had sixteen weight layers and if you include max pooling about three or four max pooling layers in it as well as and if you look at inception at 22 layers so they are much deeper they as you see that they depth seems to increase as if you look at the progression so Alex net for seven VGG was 16 or 19 inception was 22 they had other version which are much deeper as well as, now we come to this part rest net, where rest net was different from the networks we have seen, so far is that the number of layers she have just increased dramatically so these network had like 30 50 are up to 152 layers and there are reports of thousand layer rest nets being trained.

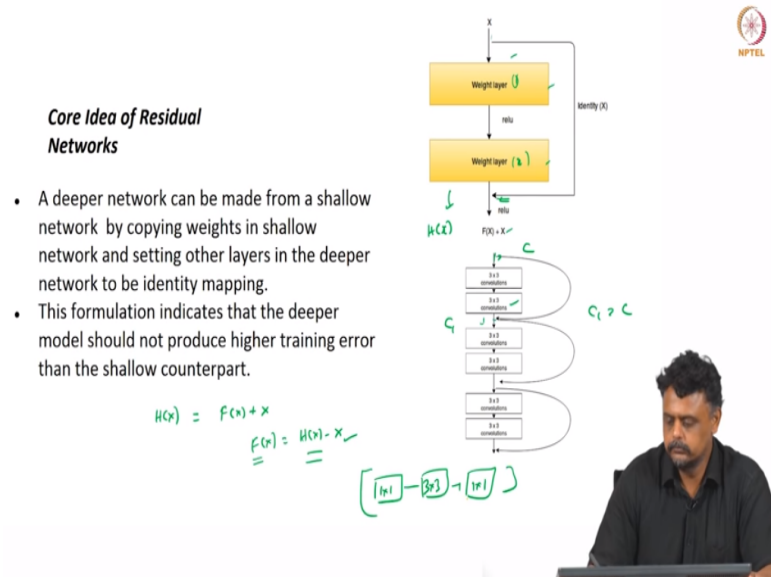
So what is the principal and of course the rest net was the winning entry to images recognition challenge 215 giving raised to error top fire of less than 4 percent about 3.6 percent better than humans, human error rate the top five, 5.1 percent about approximately so resonates where were the networks which we really tied out with very-very deep network in term of number layers and

then even really deep as far as 150 one two layers like I said about thousand as well for several application, now what is the motivation when this network in terms of what is the observation that they have had before they went to create this model, so the paper reports the following so the trained deep networks on 410 data so there is a 20 layer network so this is the training error for the 20 layer network and the yellow is that training for the 56 layer network, similarly for the test error so the test error for the so what is showed here two plots, so this plots correspond in training and test error for network strain with the say for out 10 database.

So what are shown here are two plots the training and test error for networks strain with say for 10 database and if you see the one in the yellow is a one corresponds to the 20 layer network the curve in the yellow and the red one corresponds to a 56 layer network this for training similarly for the testing lower test correspond to 20 layer this one correspond the 56 layer the problem with this picture is that general wisdom says that as you go deeper your error training and tests error should improve because your representations are supposed separate ability among classes so on some word because the non-linear is also increased as you go deeper however practically this seems to be a problem because generally there this is does not seem to work that way as you go deeper there are issues with the air training and testing.

So now what is the reason behind this the reason behind so that there is a problem with the gradient flow so the weights as you go deeper and deeper vanishing gradient problem is there vanishing on exploding you inference there and so that is not alleviated and so it is an optimization problem basically and so as you go deeper optimizing a larger net deeper network becomes harder, so to get over this problem they rest net paper introduces something called skip connections, so these connection are basically identity mapping so then since the network just a replication we will see what that means in the next slide, so they provide an alternate path for the gradient to flow and make explain possible and for the imaginary challenge they managed to clean up to 152 layer network of course the number of parameter is also sometime increase with the kind of approach as you go deeper you have more layers.

(Refer Slide Time: 04:55)



So what is the principle behind which the residual network work so if my construction you can do the following we know that the shallow network seems to work well so we use the weights of the shallow network to construct a much deeper network and wherever we have gaps we just do identity connections okay we have identity mappings so this is just by construction but for of course for the real networks constructed from scratch by back problem but so the general idea is that you know if you have a deeper network its training and testing should not be higher than that of the shallower network that is a intuition behind this so the way they are approach this is to have skip layer, so we just it is illustrated here so for instance these are two convolution layers that say it take as input bunch of feature maps X and the output is the output we will denote as H of X, so in addition to the output what is also done in a restaurant is that the input is copied here to the output, so we have here skipping two layers of two convolution layers and the input to this particular convolution layer one is copied to the output of the convolution layer too.

So your output H of X is actually of X plus X , so how does this help in any way so instead of learning H of X you would learn the residual so basically you are trying to learn so the worst case scenario if you don not learn if the weights are not updated at all it is everything is a small or zero value you will at least learn X, so that is the idea so just at least or the gradient flow happens so that is the end the residual in this case this way it is called_learning this is the residual in this case in most cases would expect to be small the idea is that every layer only slightly perturbed your input so whatever you have to learn is a very small perturbation of you input so

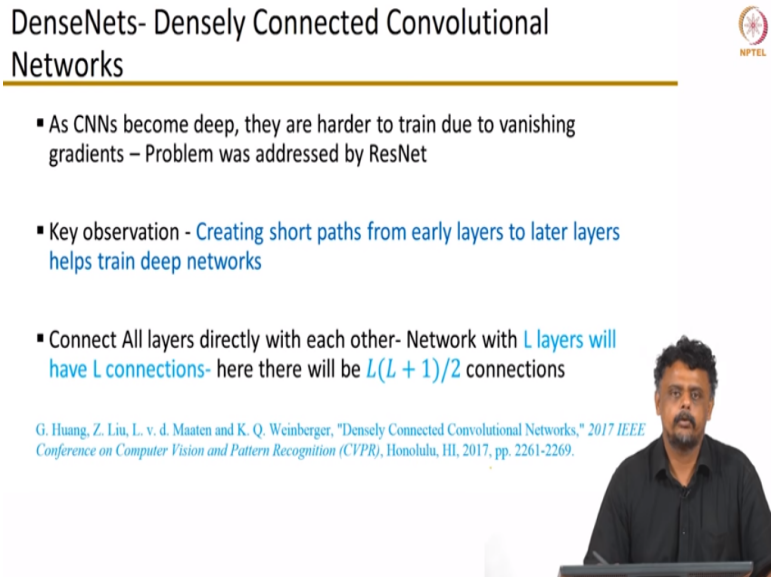
your residual will be much smaller so this is the principle behind doing the skip connections the authors also show in their paper there are presentations available on land which show that the gradient updates step is additive and it is not multiplying so that it already in step leads to actual updates to your weights and so it does not there no vanishing gradient problem that is also explain with the authors, so for the rest net used in the image net challenge there is no explanation given in the paper but typically by skipping as we saw skipping two layers at a time the skip connection are two layers at a time you are able to get pretty good results or very sad of the artisans as less than your percent error rate on the image and database using the residual networks.

So the networks itself using the residual networks the network itself varied from 30 yard layers to 152 layers so basically you would alternate this again very similar to VGG they used only three by three convolution across all the layers and you would subsampling was done by either max pooling layer at the beginning or at the end and by just by striding convolutions and for very deep networks they also introduced the bottleneck concept in some cases so you would have a so in this you would have for some of the networks they add bottleneck wherein they had a one cross one you have at a three cross three and followed by a one cross one, so that was for every instead of everything class three layer you would have this one so a three cross three layer would be replaced by this kind of a bottleneck module for some of the deeper networks so that the number of computations kept sustainable.

So I say as I mentioned earlier this showed about 3 three of our than four percent error rate on the minion challenge and this skipping layers the skip layer connection is used now in many in most modern implementation of all CNN just to clarify it is actually an addition so you would take feature maps as input from as I say in this case you take the input from here are see feature maps and you take them and add them to the output of this three by three additions this particular convolution now it is possible that the output convolutions here the number of feature map here are different let us say this s C1 feature maps and typically C1 greater than C so you would have to choose C feature maps to which to ass so that is again a heuristic left to the person making a neural network and also you can of course you can ask can we skip more can you skip three or four the rest net people who developed rest net seem to think that two layers at a time work best, it seems like a heuristic at this point.

(Refer Time Slide: 10:39)

Name of Network	Year	Developed by	Top 5 %	No. of Parameters
LeNet-5	1998	Lecun		60 thousand
AlexNet	2012	Alex Krizhevsky	15.3	60 million
VGG-16	2014	Simonyan	7.3	138 million
ResNet	2015	Kaiming He	3.6	65 Million



So far this is summary of what have seen, so far LeNet-5 we saw with basic one of the earlier networks which I mentioned earlier had a sequence of convolution pooling layers just still followed today AlexNet of error rate of imagine it is 15.3 percent VG16 at 7.3 percent error again all of them are on, not one network ensemble of results a rest had top 5 percent of 3.6 percent like it is better than human raters of course inception Google inception has better one than better than VGG or the same order of magnitude this is a number of parameters they are very high number of parameters these two are comparable but of course this rest net first much-

much deeper you see AlexNet had 60 million parameters for seven layers for rest net what 52 layers or 115 a very-very deep network and many are multiple times the number of layers in Alex net so that leave 52 to 152, let us still had only 65 million parameters a very deep network but with compare number of parameters to relatively I mean comparatively shallow network in Alex 9 so this is the progression, so far in terms of the results on the image net challenge so we will look at one more network called dense nets in the subsequent video.