Course Name: Machine Learning and Deep learning - Fundamentals and Applications

Professor Name: Prof. M. K. Bhuyan

Department Name: Electronics and Electrical Engineering

Institute Name: Indian Institute of Technology, Guwahati

Week-11

Lecture-39

Welcome to NPTEL online course on machine learning and deep learning fundamentals and applications. In my first class of this week, I explained the concept of convolutional neural network and I compare the CNN with the MLP the multi-layer perceptron. So from the parameter point of view that is the number of parameters, convolutional neural network is better than the multi-layer perceptron. So in case of the CNN the objective is to reduce the number of parameters. So by considering the concept of the convolution, I can reduce the number of parameters. After this, we are considering the concept of the pooling, maybe I can consider max pooling or average pooling.

So with the help of the pooling, I can get the local neighborhood statistics. So with the help of the pooling, I can reduce the number of parameters. That means I can reduce the size of the feature map. And after this, I considered the concept of the fully connected layers and also the parameter stride.

So with the help of the parameter stride, I can adjust the size of the feature map. So if I consider stride is equal to 1 or stride is equal to 2. So based on this parameter, I can adjust the size of the feature map. So this is the concept of the convolutional neural network. So we can reduce the number of parameters by considering the convolutional layer, the pooling layer, maybe the max pooling or the average pooling and by considering the parameter stride.

So that is the concept of the convolutional neural network. Today I am going to discuss about some popular deep architectures. So one is the Lenet 5 and another one is the AlexNet. So understanding of these architectures is quite important to develop or to design a new architecture for different applications. In this class, the objective is to gain some knowledge so that we can design or we can develop some new architecture for different applications, maybe in computer vision applications or maybe in speech applications.

So there are many applications. So based on this understanding, I can design my own convolutional neural network or maybe the deep architecture. So let us discuss about the concept of the Lenet and after this I will discuss the concept of the AlexNet. So now I will discuss some popular CNN models, the convolutional neural network models. So what is the objective of this discussion? The first objective is gaining intuitions for building CNNs for different applications and also the reusing CNN architectures.

So we know the CNN architectures and we can reuse CNN architectures for different applications. So this is the objective of understanding of some popular deep architectures that is the convolutional neural networks. So in this class, I will be discussing the concept of the Lenet 5 and also the concept of the AlexNet. So in my last class, I explained the concept of the CNN. So if you see this figure, so input is the image.

And we are considering multiple convolution and the pooling operations. So first I am doing the convolution to get the feature map and after this we are applying the pooling operation to get local neighborhood statistics. So in a particular window, we are applying the max pooling or the average pooling to get the local neighborhood statistics. So this convolution and the pooling operation I can do repeatedly and I am getting the features. And finally this feature map, it is converted into a vector that is by flattening and finally for the classification, we are considering a fully connected feed forward network.

So this is the structure of the convolutional neural network. So based on this structure, I will be discussing these two popular CNN architecture. One is the Lenet 5 and another one is the AlexNet that concept I am going to discuss today. So this Lenet 5 that was actually proposed in 1998. This was a very popular network and it was proposed during 1998 and mainly it is used for banking automation.

So used by many banks for recognition of handwritten numbers on cheques. So that is the application of the Lenet 5. So it is used for the banking automation. So accuracy should be high. So that is one requirement because it is used in the banks so accuracy should be high.

And it replace the manual feature extraction and you can see the error rate. So error rate is very low. So it is almost 0.95% on the test data. So I can say it is a very low error rate considering the complexity of the problem.

So problem is the recognition of the handwritten numbers on the cheques. So in this case, I have shown the architecture of the Lenet 5. You can see the architecture. So the input image is 32 x 32 and after this we are applying the convolution and the max pooling and again the convolution and the pooling. So like this we are applying all the operations and finally we are considering the fully connected layer and we are considering the output layer.

So for output layer we are considering the softmax activation function so that the numbers are converted into probabilities. So this is the structure of the Lenet 5. So you see the input image is 32 x 32. So if the image size is more than this so we have to scale down and if the image size is less than this we have to scale up to 32 x 32. So the input image size is 32 x 32 and after this we are considering the convolutional layer.

So one convolution layer we are considering. So in this convolution layer if you see this table so I have a table so we are considering 6 kernels. So that means corresponding to these 6 kernels I am getting 6 feature maps. So the kernel size is 5 x 5. So if you see in the table so first we are considering the convolution operation and we are considering 6 feature maps that means we are considering 6 kernels and size of each kernel is 5 x 5 and the stride is 1.

So stride 1 is considered and the size of the feature map will be 28 x 28. So 28 x 28 so that will be the size of the feature map. So that you can actually determine by the formula that already I have shown. So I can consider this operation so $\frac{F-W}{S}+1$. So floor operation we are considering F is the size of the feature map that is the size of the image and W is the size of the kernel.

So W is the size of the kernel that is the 5 x 5. F is the size of the feature map in this case it is the size of the input image. So it will be 32 x 32 and the stride is 1. So if I consider this operation that is the floor operation. So it is 32 minus 5 divided by stride is 1 plus 1.

So that will be equal to 27 + 1. So it will be equal to 28. So that is the size of the feature map will be 28 x 28. So by considering this equation I can determine the size of the feature map. So this is the first convolutional layer.

So that is the output of the convolution is 6 feature maps and the size of the feature map is 5 x 5. After this we are considering the operation the average pooling operation. So again we are considering 6 feature maps and we are considering the kernel the kernel is 2 x 2. So we are considering the kernel size 2 x 2 that is the size of the window for the average pooling operation. And we can determine the size of the feature map after the average pooling.

So by using the same formula so it will be 14 x 14 after the average pooling. And in this case we are considering the stride is equal to 2. So that is corresponding to this. So after the average pooling I am getting this one. So first one is that this is a convolution operation.

So this is the result of the convolution and after this this is the result of the average pooling operation. After this next we are considering again another convolution. So for this we are considering 16 kernels. So if you see the third row of the table. So we are considering convolution and for this we are considering 16 kernels and the size of the kernel is 5 x 5.

So then I am getting the 16 feature maps and size of the feature map is 10 x 10. So you can see this result. So this result is nothing but the result of the convolution. So we are considering 16 kernels that means I am getting 16 feature maps and the size of the kernel each of the kernel is 5 x 5. So by considering the formula we can determine the size of the feature map that is the 10 x 10 if I consider stride is equal to 1.

So I am getting 16 feature maps. So from the input the input is the 6 feature maps. So from the 6 feature maps I am getting 16 feature maps. So now how to do the connections because the input has 6 feature maps and after the convolution I have 16 feature maps. So for this actually we are doing the connections like this.

So if you see this table. So this is I have the 6 feature maps and I have 16 kernels. So 0 to 15 so 16 kernels. So you can see the kernel 0 takes input from the feature map 0 1 and 2. So that means you can see the first column. So the kernel 0 it takes input from the feature map 0, 1, and 2.

And if you go to the second column so the kernel 1 takes input from the feature map 1 2 and 3. Similarly, the kernel 2 takes input from the feature map 2 3 and 4. So we are considering like this. So 3 consecutive input we are considering. So kernel 0 is taking inputs from 0, 1, and 2 that is we are considering 3 consecutive feature maps kernel 1 it is taking the inputs from the feature map 1 2 3.

So 3 consecutive feature maps we are considering. So up to 5 we are following this principle. After this if you see the kernel 6 it takes the input from 0 1 2 and 3. So that means 4 inputs we are considering. So 4 consecutive inputs we are considering in case of the kernel number 6 kernel number 7.

So it takes input from the feature map 1 2 3 4 that is the 4 consecutive inputs we are considering. So like this we are doing the connections. So if you see this table that is nothing but the asymmetric connection. So we are considering this connection to break the symmetry in the network.

So it is not a symmetric structure. So we are considering the asymmetric structure and this is important to break the symmetry in the network. And also it is important to keep number of connections within a reasonable bounds. So we can keep number of connections within reasonable limit. So that is why we are considering this type of connections.

So connections between 6 and 16. So 6 feature maps and 16 kernels. After this you can see the next we are considering the average pooling. So if you see here next we are considering the average pooling corresponding to 16 feature maps. So size of the kernel is 2 cross 2 and we are considering the stride 2 and what will be the size of the feature map.

The size of the feature map will be 5 cross 5. And finally we are considering 3 fully connected layers. So FC FC FC. So these are fully connected layers. So we are considering in the first fully connected layer we have 120 feature maps. And you can see the size is 1 x 1 and kernel size is 5 x 5 and the stride is 1.

And for all these cases you can see we are considering tan hyperbolic as an activation function. So we are considering tan hyperbolic as an activation function in the conventional neural network we use the sigmoidal activation function. And finally we are considering the output layer. So in the output layer we are considering the softmax. So that is actually the number is converted into probabilities.

That is for the classification. So this is the structure of the Lenet-5. So this structure was popular at that time and it was used for banking automation. So summary of this network I can show here again. So if you see the table. So in the table the summary of the network is available.

And you can see we have the convolutional layer the convolution layers are available. And we are considering the average pooling operation and we have the fully connected layers. So this is the structure of the Lenet-5. And same thing I am showing here again.

So you can see the image size is 32 x 32. And after this we are applying the convolution. The size of the kernel is 5 x 5 stride is equal to 1. After this we are applying the average pooling again convolution again average pooling. And after this we are considering the fully connected layers.

So the nodes in the fully connected layer it is 120 nodes. Finally for the output we are considering the softmax activation function. So this is the overview of the Lenet-5 architecture. So next one is we are considering the AlexNet architecture. So this is the AlexNet architecture.

It is the winner of one competition. The name of the competition is ILSVRC 2012. So what is ILSVRC that I am going to explain in my next slide. But this AlexNet was developed in 2012. And it was the winner of the competition.

The competition is ILSVRC. So what is ILSVRC so that I am explaining in my next slide. So this ILSVRC that is actually the ImageNet large scale visual recognition salience. So in this case it is actually the annual Olympics for the computer vision researchers. So teams from across the world compete to see who has the best computer vision model for tasks such as classification, detection, localization.

So these are the computer vision problems. And this is actually the computer vision workshop. So this is the computer vision workshop. So people or the researchers across the world compete to see who has the best computer vision model for the computer vision tasks like classification, localization, detection. So that is the objective of this workshop. So the name of the workshop is ImageNet large scale visual recognition challenge.

So in this challenge there are 1000 image categories that is 1000 classes and total number of images in the database is around 15 million. So in the 2012 marked the first year where the CNN was used to achieve a top 5 test error rate that is around 15.3%. So in the 2012 that is the first year where a CNN was used to achieve a top 5 test error rate that is very less this is 15.3%. The next base entry achieved an error rate of 26.2%. So this AlexNet was the winner of this competition in 2012. So that is the ILSVRC competition that is the workshop of the computer vision and the AlexNet was the winner of this competition in 2012. So this is the simple image from the ImageNet database. So already I told you in this case 1000 classes we are considering I am showing only some of the images not all the images and total number of images are 15 million images and out of all these images some are used for the training and some are used for the testing that is the validation. So you can see different classes you can see one dog image you can see bottles the bird the cat.

So different images are available in this database and all together 1000 image classes are available. So this is a classification problem for this classification problem the AlexNet was developed in 2012 and you can see before 2012 nobody was using the deep architectures for this problem. So this 2012 that is the beginning of the CNN for the computer vision applications. So first CNN based winner that is in 2012 and you can see the AlexNet is the winner and in this AlexNet 8 layers are available and the error rate was 16.4 %. After this next important network was the VGG network that was proposed in 2014. So in the ILSVRC 2014 that model was proposed and VGG was the winner in that year and the error rate is 7.3 that time the error rate was 7.3 and in this network VGG network 19 layers are available. After this in 2014 there is another network so that is the Google net so in this case 22 layers and the error rate was 6.7 % and finally in 2015 so another network was proposed that is the residual network. So in my next class I will be explaining these networks but today I am explaining the concept of the AlexNet. So in the AlexNet it is the first CNN based winner of the competition that is a competition is ILS VRC competition in 2012. So it has 8 layers.

So now let us discuss about the structure of the AlexNet. So image-based classification with deep convolution neural network that is the name of the research paper it was proposed in 2012. So they use GPUs the implementation was done in GPUs and highly optimized convolutional implementation and a large data set that is the image net data set they considered. This network has 16 million parameters compared to 60k parameters of the linear pipe. So training is one issue. So for training of the AlexNet approximately one week is required because it is a huge data set around 15 million images are available.

So for training of the 16 million parameters almost one week is required for this AlexNet. So this is the structure of the AlexNet. So the image size is 227 x 227 that is the input image size.

So in this case we are considering the RGB image. So that is why it is the 3. So for the R channel G channel and the B channel we have to consider. So all the inputs are basically RGB images. So 3 channels we have to consider and we are considering the kernel size the kernel size is 11 x 11 that is the filter size and we are considering 96 kernels. So 96 kernels are considered in case of the AlexNet and the stride is 4. So by considering that formula that already I have explained I can determine the size of the feature map.

So the size of the feature map is 55 x 55. So 55 x 55 and the depth of the feature map is 96 that means I have 96 feature maps after the convolution. So again I am repeating the size of the feature map is 55 x 55 and the depth of the feature map is 96 that means I have 96 feature maps corresponding to 96 kernels. After this we are considering the overlapping max pooling operation. So size of the window is 3 x 3 and the stride is equal to 2. So we are applying this overlapping max pooling operation and corresponding to this we are getting the feature map.

The feature map is 27 x 27 and number of feature maps will be same. So it is 96. So same number of feature maps. So 96 and the size of the feature map is 27 x 27. And in this case we are considering the overlapping max pooling operation. So the next step of the AlexNet is again the convolution. So if you see again we are considering the convolution and in this case the size of the kernel is 5 x 5 and 256 kernels are used and the zero padding is used.

So because of the zero padding you can see the size of this feature map and this feature map will remain same. So if you see the input size the input size is 27 x 27 and because of the zero padding I am getting the same size after the convolution that is the 27 x 27. So we are considering 256 kernels and the size of the kernels are 5 x 5 and we are considering the zero padding pad is equal to 2 and based on this we are doing the convolution and we are getting the feature map. So 256 feature maps and size of each feature map is 27 x 27 that is after the convolution.

After this we are doing the overlapping max pooling again. So again we are doing the overlapping max pooling operation. So size of the window is 3 x 3 and stride is equal to 2. So if I do the max pooling operation then I will be getting the feature maps the size of the feature map is 13 x 13 and again we are getting the 256 feature maps that means 256 channels that means we are getting 256 channels after the pooling operation. After this again we are doing the convolution and in this case we are considering 384 kernels and we are doing the padding. So because of the padding the size of the feature map after the convolution will remain same so input size is 13 x 13 and after the convolution we are getting 13 x 13 because of the zero padding.

So the size of the feature map after the convolution will be 13 x 13 and we have 384 channels. That means 384 feature maps. After this again the convolution is done so again we are doing the convolution and we are considering the 384 kernels that is the 384 channels feature maps and again the zero padding is done. So after this convolution the size of the feature map will be 13 x 13 and in this case 384 channels. After this again the convolution is done and the size of the kernel is 3 x 3 so we have 256 kernels and each of the size is 3 x 3 and zero padding is done so pad is equal to 1 and after this convolution we are getting the feature map so we have 256 channels that means 256 kernels and size of the feature map is 13 x 13.

And finally in this case we are considering the overlapping max pooling so overlapping max pooling we are considering so we are considering 3 x 3 kernel and stride is equal to 2 so number of feature maps will remain same so it is 256 input is 256 so 256 feature maps and size of the feature map is 6 x 6 that is the size of the feature map. After this we are considering two fully connected layers so FC and FC these are the fully connected layers so if you see how many nodes here so how many nodes here so 6 x 6 x 256 so that is equal to 9216 so that means in this case I have 9216 nodes so this 9216 nodes are connected with 4096 nodes so this 9216 nodes are connected with 4096 nodes of the fully connected layer so we have to consider these connections and after this we have another fully connected layer so in this case also I have 4096 nodes so 4096 nodes are connected with 4096 so in this case if you see the connections here so how many connections 9216 into 4096 and in this case here so how many connections 4096 is connected with 4096 so all these connections we have to consider and finally I am considering the output layer so in the output node we are considering 1000 nodes so why actually we are considering 1000 nodes because the problem is the classification problem and we have 1000 category 1000 classes so for each and every classes I have one node so all together I have 1000 nodes in the output layer so if I consider these connections so how many connections so 4096 x 1000 so this number of connections we have to do and we are considering the softmax operation because I am getting the results in the form of the probability so it is a problem of classification and we have 1000 classes so this is the complete architecture of the AlexNet so this architecture is implemented in two parallel channels so if you see in this case we are considering two parallel channels for the implementation of the AlexNet and in this case two GPU cards were used for the implementations so because there are two parallel branches for these two parallel branches two GPU cards were used in the implementation of the AlexNet so half of the network is put in one channel and the remaining half is put in the another channel so there are two channels which are implemented in two GPUs so I am repeating this half of the network is put in one channel and the remaining half is put in the another channel so this is the concept of the AlexNet and you can see some of the features extracted by the AlexNet so these are the image features which can be extracted by the filters of the AlexNet so again I am showing the overall structure so input is the RGB image the size of the image is 227 x 227 x 3 and after this we are considering the convolution layer again the overlapping max pooling concept and after this the convolution again max pooling again convolution convolution convolution and finally max pooling and after this we are considering the fully connected layers these are the fully connected layers of the AlexNet and finally we are considering the softmax for the final classification and in the output layer I have 1000 nodes corresponding to 1000 classes so the summary of the AlexNet is so in the AlexNet the relu function was used instead of the sigmoid activation function 60 million parameters are available in the AlexNet and 6 lakhs 50 thousand neurons so all these parameters we have to train during the training so if I consider the image net dataset so for training of all these parameters the AlexNet needs almost one week also the train on GTX 580 GPUs with only 3 GB memory that time so the network is split into two pipelines and was trained on two GPUs that already I have explained so the network spread in two GPUs and half the neurons on each GPUs and input size already I told you the input size is the 256 x 256 RGB that we are considering but in this case the input to the AlexNet is 227 x 227 so that is why the cropping is important so we have to do the cropping of the image so we have to consider the RGB image so three channels are considered in the AlexNet and the input size to the AlexNet is 227 and 227 so that is why we have to do the cropping of the image the input image and if I consider the grayscale image then what we need to consider the grayscale images to be replicated to obtain three channel RGB image so this network can be used for the grayscale image also but the grayscale images to be replicated to obtain three channel RGB images and for the training we consider the back propagation training algorithm that is nothing but the stochastic gradient descent algorithm so already I have explained the concept of the back propagation training and this is one technique that is the stochastic gradient descent algorithm for the training of the parameters and we are considering the moment optimizer so that is used to make the stochastic gradient descent algorithm more efficient so since we are considering the gradient descent algorithm and this one technique that is the momentum optimizer so one optimization technique is used to make the gradient descent algorithm more efficient so this is the discussion of the AlexNet so AlexNet was the coming out party for CNNs in the computer vision community and this was the first time a model performed so well on a difficult data set that is the image net data set this paper that is the paper on the AlexNet illustrated the benefits of the CNN and back them up with record-breaking performance in the competition so already I told you the AlexNet was the winner of this competition and the performance was so good for this complicated data set that is the image net data set so in this class I explained the concept of two popular CNN architectures the first one is the LeNet-5 that was developed in 1998 for the automation of the banking industry and the second one is the AlexNet that was used for the problem of image classification and for this one big dataset that is the image net dataset was considered and 1000 classes were considered for the classification problem so image net performance was very good for this dataset and it was the winner of the competition in 2012 so this is the brief discussion on these two popular CNN architecture one is the LeNet-5 and another one is the AlexNet so in my next class I will be explaining the concept of another three popular CNN architecture   so let me stop here today thank you.