**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-12**

**Lecture-42**

Welcome to NPTEL online course on machine learning and deep learning fundamentals and applications. In my last class, I explained the concept of GAN that is the generative adversarial network. So, in GAN, there are two models, or I can say there are two networks, and both are competing with each other. And because of this competition, the performance of the models or the performance of the networks will improve. That is the concept of the GAN. Today I am going to discuss the concept of another very popular convolutional neural network that is UNET.

So, UNET is very popular, particularly for the applications like image segmentation. So, image segmentation means the partitioning of an image into connected homogeneous region. That means the separation of the foreground and the background of an image. So, I will explain the concept of the UNET.

And also I will show how it can be applied for image segmentation that is actually the semantic image segmentation. So, let us begin this class, the concept of the UNET and also the applications of the UNET for semantic image segmentation. So, UNET that is the convolutional neural network for image segmentation. So, in my last classes, I explained the concept of CNN the convolutional neural network. In this figure, I have shown one CNN for object detection or object classification.

So, you can see here I have one input image and for feature extraction, we are considering convolutional operation. So, after the convolutional operation, I will be getting the feature maps and also we can apply the pooling operation. In this case, I have shown the max pooling operation and for the CNN, we are considering nonlinear activation function like ReLU activation function we can consider. So, repeatedly I can apply the convolution operation and the pooling operation. And finally, I have fully connected layers and I can do the classifications.

So, you can see the classes different classes after the fully connected layers. So, this is the structure of the CNN the convolutional neural network. So, I have to apply a convolution operation, max pooling operation repeatedly to extract features from the input image. And finally, for the classification, we are considering the fully connected layers that is the fully connected network. So, that is for the final classification.

So, CNN structure for image classification it is like this. So, you can see that is the input image is available. And after this we are considering the convolutional unit cu means the convolutional unit. After this we are doing the down sampling du means the down sampling unit. After this we are considering the convolution again the down sampling the convolution and after this we can consider the global average pooling and for the classification we are considering the fully connected layer FC is the fully connected neural network and we can do the classification with the help of this convolutional neural network that is a CNN.

Now, let us discuss the CNN structure for semantic image segmentation. So, what is segmentation? Segmentation means the partitioning of an image into connected homogeneous region. The homogeneity I can define in terms of the color value or maybe texture information or maybe some other information and based on this homogeneity I can partition the image into connected homogeneous region. So, that means the separation of the foreground and the background that is the definition of the segmentation. But what is semantic segmentation? So, you can understand this concept that is actually the pixel wise semantic segmentation.

So, it is nothing but the labeling of every pixels of the image. So, if you see the input image corresponding to this input image we have done the labeling. So, you can see so, you can see this corresponds to cow the grass. So, these pixels are labeled. So, that is the definition of the semantic segmentation.

Similarly, for the second image, you can see the labeling like this tree, sky, the body, road, grass. So, these are the labeling corresponding to this input image. And similarly for the third image also, you can see the labeling the sky, building, airplane, grass. So, this is the labeling of the input image. So, this is the definition of the semantic segmentation.

That means I can label every pixels of the input image and it is a very popular computer vision problem. So, we will discuss about this semantic segmentation and how the convolutional neural network can be employed for semantic segmentation. So, for this we are considering one very popular architecture that is the unet convolutional neural network. So, here also I have shown the concept of the semantic segmentation. So, corresponding to this input image, you can see the labeling the sky, grass, trees.

So, we can do the labeling and similarly corresponding to this image, you can see the labels of the image the trees, cow, sky, grass. So, these labeling I can do in case of the semantic image segmentation. So, now what is the unet? So, this was proposed in 2015. That is the unet architecture. It was proposed in the IEEE international symposium on biomedical imaging in 2015.

And the application was biomedical image segmentation. So, unet learns segmentation in an end to end setting. So, what is this end to end setting so that I will explain corresponding to this input image if you see this input image, the segmented image will be like this. So, output segmented image. Similarly, if I consider this case, so corresponding to this input image, the unet learns segmentation.

So, corresponding to this output segmentation map I can obtain like this. So, that means I can say the unet learns segmentation. So, input image is available and corresponding output segmented image I will be getting like this. So, let us discuss the structure of the unet. So, that is the unet architecture.

So, why actually it is the unet because if you see the structure it is like U. So, if you see this is like the U. So, that is why it is called the unet. So, in this case, the network has actually two paths. One is the contracting path and another one is the expansive path.

So, this path if you see this path is the contracting path and the second path is the expansive path. So, the contracting path is a typical convolutional network. You can see in the contracting part, it is nothing but the typical convolutional neural network. And in this case, we are applying the convolutional operation and the pooling operation. So, repeatedly we are applying the convolution operation and the pooling operation and the ReLU activation function is considered.

So, repeated application of convolution followed by rectified linear unit that is the ReLU and after this the max pooling operation in case of the contracting path. During the contraction, the spatial information is reduced while feature information is increased. In case of the expansive path, it combines the feature and the spatial information through a sequence of up convolution and the concatenation with high resolution features from the contracting path. If you see the expansive path, it is nothing but the up convolution and also the information is taken from the contracting path. So, you can see this information all this information it is taken from the contracting path.

So, expansive path is nothing but the up convolution and the concatenation with high resolution features obtained from the contracting path. So, I will explain in detail in contracting path we are considering the convolutional operation, pooling operation and we are considering the ReLU activation function. The objective is to reduce the spatial information, but we are increasing the feature information. In the expansive path, we are

doing the up convolution and we are concatenating the high resolution features from the contracting path. So, that means we are combining the feature information and the spatial information in the expansive path and also we are considering the concatenation of the high resolution features from the contracting part.

So, that concept I will explain in my next slide. So, in this figure I have shown the first one is the contraction phase. So, in the contraction phase, we are increasing the field of view and lose spatial information. So, spatial information is reduced, but we are increasing the feature information. So, already I told you in the contraction phase, we are doing the repeated convolution and also the pooling operation and activation function is the ReLU activation function.

So, in the figure you can see we are considering the input image, the size of the image is 572 x 572. After this we are considering the convolutional operation. So, we have 64 feature maps and the size is 570 x 570 and after this again we are doing the convolution 64 kernels that means 64 channels or the 64 feature maps and size of the feature map is 568 x 568. After this we are doing the pooling operation that is the max pooling operation. So, this is a max pooling operation and after this again we are doing the convolution operation and you can see the size 284 x 284.

So, we are considering the stride 2 and again we are doing the pooling operation, again we are doing the convolution operation. So, repeatedly we are doing the convolution operation and the max pooling operation and activation function is the ReLU activation function. So, during the contraction phase, we are reducing the spatial information, but we are increasing the feature information. So, we are extracting features from the input image. So, all the important features are extracted in this phase.

So, during the contraction phase, I am extracting all the important information from the input image. So, this is about the contraction phase. After this the next phase is the expansion phase. So, this is for the creation of the high resolution mapping. So, you can see we are doing the up convolution.

So, convolution we have done and ReLU function we have considered and also you can see the information is taken from the previous phase that is the contraction phase. So, that means, we are combining this information that is the feature information we are combining and after this we are getting the output segmented map. So, in the expansion phase, we are doing the up convolution and also the feature information is combined with this feature map to get the output segmented image. So, that is the objective of the expansion phase. So, we are just doing the concatenation to consider the feature information that is obtained from the contraction phase.

So, that we have considered. So, we are doing the up convolution and also we are getting the information from the contraction phase that is the feature information and finally, I am getting the output segmented map. So, here again, I am showing the same concept. So, two phases one is the contraction phase and another one is the expansion phase. So, during the contraction phase, I am getting the feature information, but I am decreasing the spatial information and during the expansion phase, I am doing the up convolution and also concatenate with high resolution feature map from the contraction phase. So, this information we are combining to get the output segmented map.

So, that is the objective of these two phases one is the contraction phase another one is the expansion phase. So, now I am showing how we can do the convolution that is for the contraction phase. So, input image size is 572 x 572 after this we are doing the convolution. So, 64 kernels and after doing the convolution I am getting the size of the feature map is 570 x 570. In this case, we are not doing the zero padding.

So, that is why the size of the feature map will be 570 x 570. After this again, we are doing the convolution 64 kernels and size of the feature map will be 568 x 568 because we are not doing the zero padding. So, in the figure you can see this x and y that corresponds to size of the feature map and the feature k that represents the depth of the feature map that is nothing but how many feature maps we are considering. So, x × y that is the size of the feature map and the k means how many features we are considering that is the depth of the feature map. So, you can see we are doing the convolution like this and 3x3 convolution we are considering and we are not considering the border pixels.

So, border pixels are not considering and a ReLU activation function is considered. So, the expression for the output will be like this ReLU is considered and you can see just this is the expression for the convolution. The multiplication of the pixel value with the weight value of the filter or the kernel and just we are doing the multiplication and the sum up and we are doing the convolution for this. The size of the feature map can be obtained by using this formula $\frac{W-F+2P}{S}+1$. So, W means the input volume size F means the filter size P is the zero padding S is for the stride.

So, if I put this value you can see 570 - 3 + 2 x 0 because we are not considering the zero padding divided by 1 plus 1. So, stride is 1. So, the size of the feature map will be 560 x 568. So, the size of the feature map will be 568 x 568. So, that is for the second convolution for the first convolution we are getting the size 570 x 570.

So, by using the same formula we can determine the size of the feature map. So, after the first convolution the size of the feature map will be 570 x 570 and the output size

after the second convolution will be 568 x 568. So, this is the convolution. So, we are doing the convolution like this.

So, we are doing the repeated convolution. So, before doing these convolutions we are doing the max pooling. So, this is the max pooling operation max pooling operation and again we are doing the convolution like this these are the operations during the contraction phase. So, this is the contraction phase. So, after the contraction we have to consider the expansion phase. So, for the expansion phase you can see we are doing the up convolution.

So, the concept is very similar only we are doing the up convolution in this case the stride 2 is considered and again the convolution is nothing but that multiplication and the sum up and we are considering the ReLU activation function. So, corresponding to this input feature map we are applying this filter and we are getting the resulting feature map having factor 2 higher resolution. So, that means we are doing the up convolution. So, this is the operation for the expansion phase after this the concatenation with high resolution features from the contraction phase. So, you can see just we are combining the feature information that is obtained from the contraction phase.

So, that is nothing but concatenation with high resolution features from the contracting path. So, this we are considering and we are getting the output segmentation map and in this case we have only two classes one is the background another one is the program. So, we are getting the output segmentation map and it is nothing but a two class problem because we have the background and the program. So, segmentation is nothing but the separation of the program and the background. So, this is the principle of the unet that is the structure of the unet.

So, this phase is the contraction phase. So, in the contraction phase we are increasing the feature information that means we are increasing the what information, but we are reducing the spatial information that means we are reducing the where information. So, during the contraction phase, we are increasing the what information but reducing the where information. That means that this information is increased and the spatial information is reduced.So that is the contraction phase. So, during the expansion we have to create high resolution segmentation map.

So, that is why we are considering the operation the operation is the concatenation operation with high resolution features which are obtained from the contracting path because the objective is to create the high resolution segmentation map. So, that is the objective of the expansion. So, in the expansion phase we are doing the up convolution because we have to create the high resolution segmentation map. So, this is the concept of the unet architecture.

So, this concept I can explain like this. So, please move to the next slide. So, it is nothing but the encoder decoder concept. So, you can see here one phase is the encoding phase and the second phase is the decoding phase. So, input image we are considering. So, first we are doing the encoding to get the encoded representation of the input image.

So, that is the compressed representation of the input image. After this we are doing the decoding to reconstruct the original image. So, decoding is done to reconstruct the original image. So, this principle is encoding and the decoding principle. And because of the encoder, I am getting the encoded representation that is the compressed representation of the input image and from the compressed representation of the input image.

I am reconstructing the original image that is the decoding. This encoder decoder structure is actually the concept of the unet architecture. So, I can explain this concept in my next slide. So, this is the unet. So, you can see the input image and during the contraction phase, we are doing the convolution operation, pooling operation that is the down sampling, convolution operation, down sampling. So, we are considering these repeated operations to increase the feature information and to reduce the spatial information.

So, this is the contraction phase. During the expansion phase, we are doing the up convolution and also we are concatenating the feature information which is obtained from the contraction phase with the expansion phase. So, that means we are taking the information from the previous phase to this expansion phase. So, that is nothing but the concatenation of the feature information. We are doing the up convolution and also we are taking the feature information that is obtained in the previous phase that is the contraction phase.

And finally, we are getting the segmented output. So, this is the segmented output. So, this structure is very similar to encoder decoder structure. So, the summary of the unet. So, already I told you there are two phases, one is the contraction phase, another one is the expansion phase. So, in the contraction phase, we reduce the spatial dimension, but increases what that means we are increasing the feature information, we are reducing the spatial dimension.

So, in the contraction phase, we are reducing the spatial dimension, but increases what information that is the feature information. In the expansion phase, we are recovering object details and the dimensions. So, that is why we are considering the up convolution because we have to recover the dimension. So, that is why we are considering the up convolution and also we are considering the operation that is the operation is the concatenation operation to recover object details.

That means, the where information is also considered. So, what information is considered in the contraction phase and the where information is also considered in the expansion phase. So, I am repeating this in case of the contraction phase, we are reducing the spatial dimension, but we are increasing the feature information that is the what information. In case of the expansion phase, we are recovering object details that is the feature information and also the dimensions. So, that means, the where information. So, concatenating feature maps from the contraction phase helps the expansion phase which recovering the where information.

So, that is the objective of the concatenation. So, I am repeating this, the concatenating feature maps from the contraction phase helps the expansion phase which recovering the where information. So, that means, we are considering the concatenation operation to recover where information. So, that is the summary of the unit architecture. And these are the outputs corresponding to the biomedical image segmentation. So, you can see corresponding to this input image, these are the output image that is the segmented image and corresponding to this ISBI cell tracking cell in 2015 that was proposed in 2015.

So, corresponding to this input image, you can see the segmented output image. So, these are the results of the unet architecture. So, in this class, I explained the concept of the unet architecture, which can be employed for image segmentation. So, in the unit architecture already I told you, there are two phases one is the contraction phase and another one is the expansion phase. In the contraction phase, I am reducing the spatial information, but I am increasing the feature information.

In case of the expansion phase, I am increasing the feature information as well as I am also increasing the spatial information because I am considering the concatenation operation. So with the help of this concatenation operation I am increasing the feature information as well as I am increasing the spatial information with the help of the up convolution operation. So, So that is the summary of the unet architecture. So, two important phases, one is the contraction phase and another is the expansion phase.

And I have shown how it can be employed for image segmentation. So let me stop here today. Thank you.