Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-82

Right, I mean a deconvolution network is again right, I mean except that except that because right that idea of trying to try to sort of write you know draw the draw the outputs of the previous pooling layers then you know in fusing in order to be able to get the output was not really a smart idea. So, it is more like a formal approach. So, a deconvolution network usually will use you will use what is called a transposed transposed convolution. You guys know about transposed convolution right. So, transposed convolution is like this. So, what this means is right see if I have how to explain this.

So, if I have A x equal to y, if you do a convolution operation right think of think of x as a let us say right this is your forward path that means you are doing a convolution with some stride and which means that x is of a higher this one dimension and then you are going down doing a strided convolution. So, what will happen? So, when you actually construct this matrix. So, suppose x is a 4 X 4 sort of a feature map suppose it is 4 X 4 right then this will be this you can stack it up as a 16 X 1 vector I mean you can write it in a vector matrix vector form and y right the output suppose let us say you are doing a strided convolution such that the output is actually a 2 X 2 output suppose right then it means that if you stack this y up it will be a 4 X 1 vector is what you want to get and therefore this A will be something like 4 X 16 right that will be the matrix that you will have to build. See there is an easy way to just say you know in a sort of illustrative way how to actually arrive at this, but then the equation part is important ok that is just a graphical illustration which I will tell in a minute.

By the way right I know that I have exceeded I just wanted to tell you that you know I will just continue until I finish this topic today and let me also tell you one more thing right this Wednesday again right again I am not available this Wednesday unfortunately and Tuesday is a holiday tomorrow right tomorrow is a holiday correct. So, the next class is only on Friday and since today is a last instruction day after this right I do not even expect you to come I mean right I would be I will encourage you to come, but then it is not a must that you have to come I have actually 2 more classes to go I am taking one on Friday and one on Monday I will be here I will record the session right I mean you are of course right you know encouraged to come, but I am saying that I will not take say

attendance ok. So, the last 2 classes I will not take attendance, but of course right I will be happy to see you guys here if you do not come it is ok ok. So, there is still because the high level vision I want to talk about RCNN and all the YOLO and all that right. So, I thought I will take and there I may even exceed my time because I talked to Mahesh and he said it is perfectly ok it can be done because that once I upload the video you guys can then watch it later ok, but I want to finish off everything by 14 that is what my that was my initial target this thing was unexpected the last week one day I lost and then this Wednesday I am losing again this came due to some unexpected some something else right.

So, I could not beyond my control. So, I am I wanted to finish by this Friday well that is not going to happen. So, Friday may Monday and then maybe one more day ok. So, those 3 classes right you are of course you know absolutely absolutely right about you know you know it will be great if you can come if you can attend if you cannot attend it is ok right I will record it here and I will upload you guys just want to tell you that and now also right I may take another 5 minutes if you have something you can go if you have a class or something ok.

. There a segmentation no this is all hand annotated this is all annotated there is there is there is this there are these people right that are paid for it. So,. . No no no it is just a cow that is all it can be of any color cow is a cow ok.

So, this up sampling thing right. So, it is called. So, down sampling of course, is typically a strided cow. So, what I was trying to say was this kind of this kind of a transpose convolution right. So, what this means is you see what will happen is you know you are doing kind of a down sampling then you have then you are going you know in an up sampling path right.

Now, if you have if you have done a certain sort of a strided convolution at the let us say fourth layer here right then after you reach a bottle neck then you go back then in that fourth in the in the up sampling path when you come to the fourth block right you want to be able to use the same A transpose right. So, that whatever size was here you actually meet that size back again here. Now, that is where, but then the features would have changed right I mean you do not get the same x now you will get some other feature. The idea is that you take you take you take A transpose you can show this ok this can be mathematically shown. You take A transpose which is actually 16 X 4 and then you operate it on that on that on that let me let us call that z does not have to be y right it will be some 2 X 2 feature map there right which you want to get a blow up to 4 X 4.

Understand this right somewhere you had a 4 X 4 you you and then and then right here

now it came out to 2 X 2 on the up sampling path you have a 2 X 2 it needs to be blown to 4 X 4. So, so, so the A that you used here now you you know you need an A transpose. So, whatever you did here right you want to sort of undo it but it is not strictly speaking A inverse ok. So, please do not think of deconvolution as some inverse of the convolution at all it is simply an operation. So, so it is it is simply spreads values I mean that is a way to kind of right think about it ok.

I mean so sitting at a place it spreads the values right in a certain sort of you know a proportion. So, A transpose will act on z which is actually 4 X 1 in order to give you let us say some other whatever p or something which is actually 16 X 1 which is 4 X 4 right that is how you go. So, that is why it is called a transpose convolution and if you and if you think about it right this is how it will look sorry this is max pooling. Let us say transpose convolution actually ok I mean think of this kernel as or change some values I mean I do not know why they have given the same thing right. So, let us say you got A B C D right as the input.

So, what will happen is when you have kernel when you when you actually do this transpose convolution I mean mathematically we do it that is how you should do. Graphically how they show it is so, so first case right you will have 0 0 0 0 because this guy is 0 and then for the second case right when you have 1 ok. So, what you will have is like you know A B C D here ok and then when you have 2 right you will have like 2 A 2 B 2 C 2 D ok. And then then when you have 3 right then you will have like 3 A 3 B 3 C 3 D right and all this you have to add. Spatially right there will be some common entries also right for example, this 2 B and then C and then whatever 3 A they will all add up right yeah that is fine like I told you can go ok.

So, once you sum it up then you can get your output this is exactly the same operation right if you did A transpose on this you see on this 2 X 2 which is this input and that kernel this is exactly what will happen ok. But then you know this is how graphically people already illustrated, but it can be mathematically done properly ok. Now, the other the max pooling max unpooling if people know right how it is done. So, that is that is typically like this right when you want to do max unpooling. So, what you do is if you want to do an unpooling.

So, you kind of so, in this case when you are doing max pooling what did you do you pick the so, here is kind of you know a pooling 2 X 2 pooling with a stride of 2 right or 2 X 2 filter with a stride of 2 that is what it is. So, here you get 5 here you get the max value is 6 max value is 7 max value is 8. Now, when you go back it you may have you may have some other feature which is 1 2 3 4 now you want to you want to do a max unpooling. So, the 1 will go and sit exactly. So, this you have to store somewhere when you do max

unpooling.

So, you have to know that location from where the max was picked. So, that is something that you have to carry with you and then you place your 1 there you place a 2 exactly right in this location where 6 was 3 will go wherever 7 was and 4 will go wherever 8 was. And now after this you followed up with a bunch of convolution kernels which are actually learnable and then because you do not want so, many 0s sitting there right. So, basically they will start filling up all these 0s. Because when you do want to do the back prop you need to you need to get a right back prop through those through those points because those are the ones from where you pick the max value.

This is for the this is for the back prop. Sorry. No, no the point is point is this is simplest way in which in which in which way you can do you can do max unpooling I mean if you had this information you can do the unpooling. If you do not have how will you do it because you should know from where the max was picked you would not know you have to have it. Max unpooling will will require that you know the location in fact, it is always done with standard thing you know from where you are coming.

And then then this then this transpose convolution rate that is how it works then ok. So, so this deconvolution network and here are some examples. So, so that is the that is the only any sort of a takeaway right. So, instead of doing it in that arbitrary manner right you do a proper sort of a transpose convolution here ok. And then and then they they show some output results right.

So, here is a cycle and you can see that it is picked up the wheels correctly and then it is actually segmenting the the the cycle well. Of course, there could be some sort of a confusion here and there, but then then the segment is exactly the same way ok. So, you know we do not talk about segment you know in separation it is basically the same same same logic and in fact, these are all papers that came around the same time. Therefore, right I mean you know it is not like one guy took the idea of the other man or something. So, it is actually it is a contemporary kind of work right.

And then and then deep lab is the last one that I wanted to talk about. So, here so the idea is that right. So, deep lab is nice because it does not it does not even want to kind of reduce your spatial resolution. So, it wants to operate exactly at the at the same sort of resolution at which the input is. All the all the other methods that you saw that are initially down sampling then up sampling right which means that you are already losing some information which you are trying to recreate later and then then then there is every possibility that you have lost something along the way.

Therefore, the way they kind of do it is in order to not not kind of lose resolution is they use what is called what is called as at risk convolution or what is called really a dilated convolution. So, this dilated convolution again kind of see why did you why did you do this down sampling in the first place because you wanted a receptive field right which was higher or it would have meant that meant that you would have used multiple layers. So, that a kernel right will actually instead of looking at 3 X 3 it will effectively look at 5 X 5 or 7 X 7 that is what that is what you ideally want to do. But at the same time they did not want to want to have these computations going up right. So, they wanted wanted a receptive field which is higher and then the computations should also not go up.

So, which is the reason why they kind of came up with this say at risk convolution was already there. So, and then there is a pyramid pooling and they just put the 2 together. So, at risk convolution looks like this. So, you have a 3 X 3 kernel. So, if you have a rate 2 at risk convolution what that means is that it simply you fill 0s everywhere else and then and then you simply place these like you know every third.

So, it is like 1 and then you know 2 third you again pick the pixel. So, it is so, so or sorry second. So, 1 2 that is why it is called a rate 2 1 2 then you know 1 2 1 2 right I mean that is how you play I mean every alternate row and column right you you kind of take it is the same value. So, you do not have to learn anything new it is simply still 3 X 3, but now your receptive field is over 7 X 7 right in a way. And then the fact is the in between there are 0s because they do not want to increase the number of unknowns right I mean then then it would not make sense right.

So, the idea is that with this actually it works very well and then you can have a rate 3 whatever right. So, so idea is that you know with these with these kind of filters right you can have you can have a receptive field that can go up and therefore, you do not have to reduce the image size right you can still have a receptive field that is good. And then alongside came the idea there was this spatial sort of a pyramid pooling which is already there I mean if you kind of remember it in CNN suppose I told you that that I mean you know how what would you do if you had to train with different sizes of inputs can you actually train it. Suppose you had a classification layer somewhere right at the end and then let us say you have CNNs in between and then and then I have I have images right or no which are which I want to feed to this network. So, you saw that all the way AlexNet what was the size it was always fixed right some 224 something X 224 why did it have to be like that right simply because when we were not accommodating for the size somewhere.

So, this spatial sort of a pyramid pooling does exactly that. So, what it will do is irrespective of whatever the feature map size it will always kind of bunch it into some it

will be say 16 zones. So, in this case right for example, right I just divide it. So, so if it is higher than then each block will then be will then be larger if it was smaller image each block will become smaller, but I still get a 16 sort of you know. So, for each of these I will get one value for each of these I get one value one value I get 16 and then if there are 256 feature maps I get 16 X 256, 256 kind of kind of right I mean this one a vector and similarly right I can have 4 instead of dividing into 16 I can have 4 blocks I can have just 1 right.

So, that is the idea of spatial pyramid pooling. So, which means that whatever be the input size right eventually what comes out of this fully convolutional network is finally, one standard size which comes out which means that I can accommodate any size right that is actually a good idea I mean right. So, it lets you operate with images of various sizes and so on. So, that is the spatial pyramid pooling. No, no see what this means is there are 256 feature maps and each one these are all of the same size, but then right these could come from you know a different image size would mean that my feature size is then get us a different because there is a there is a sub sampling ratio that you maintain right even if a convolutional network there is a sub sampling ratio that you maintain, but the size will change right.

So, what this saying is if I have a larger image I will whatever be the image size I will divide it into a certain number of blocks in this case 4 X 4. So, which means that for every block I will I will I will pick one value it will be like a global average pooling for that block I will get one output one output. So, I always get 16 X 1 irrespective of whatever be the size I mean then I can change it I can have 8 X 8 I can have 4 X 4 I can have. So, that is called and then the pooling comes because they do not just just just do it with you know one resolution it is like a resolution of the block you can have a coarser resolution you can have a very coarse resolution then you concatenate all of them that is why it is called a you know pyramid pooling a pyramid because you have like 16 X 16 4 X 4 that is how you saw that you know when you saw Gaussian pyramid I mean what was that. So, similarly a pyramid and then pooling right because all these vectors will then be pooled together in order to create one sort of a flattened out vector which will remain the same dimension irrespective of whatever comes in.

So, that after that you can operate whatever you want right that is the biggest advantage of you see SPP. And now these guys smartly combine both right. So, this parallel pooling idea is this right where you have different rates. So, here the way they pooled they kind of did this they call it spatial pyramid at rest spatial pyramid pooling I think ASPP or something and the idea is that they do they do of course, this at rest convolution, but then because they because they operated at see different rates right. So, for example, rate is equal to whatever 18 24.

So, they operated different rates then they take all these. So, there is a parallel set of filters acting right a parallel set of filters acting right seeing different different parts of the image and then all of them getting fused right in order to be able to get your final output right. So, that is the and deep lab is apparently one of the one of the one of the best that is still around that is why I thought I should cover this and you can see this right, but then you know this by itself does not work. So, well there is still there is still one final thing that they have which is like which is a conditional random field right I mean that by itself gives out sort of a smooth output. So, the whole paper is not just that then there is a post processing after the network this is not part of the network.

So, that is a conditional random field because you do not because you guys have not learned Markov random fields and all right. So, I cannot go into the details, but then what this means is that you know if you think of an image as a graph or you think of the output as a graph right where where you where you think of each kind of pixel as a node and then you have edges right to kind of to you know every other node ok. You do not have to observe to every other node you can have a small neighborhood around which you have nodes sorry edges and then the and then the weight of the edge will tell how important it is that this is the other label for this label what kind of a dependence right it has. Actually there is a whole area called graph based segmentation right which we could not take and because you know these are the ideas that deep networks having come now in full flow right people are no longer going that way. So, there is no point in doing all of that because right eventually you will end up doing only probably one of these things right.

So, I thought classical I wanted to give you a give you a feel for what are the general things there are various methods even there is something called region based growing in segmentation this graph based and all that, but anyway right. So, this comes from somewhere there right. So, where this where the where you have something called a unary term then there is a pair wise term and then how these pair wise terms should should sort of what weights right they should get depending upon they may know how close they are right I mean there is a there is a weight for I mean there is this kind of exponential quantity here right which is in terms of the location in terms of the feature value and so on takes all of that in the account in you know to arrive at something like a like an like an energy minimization. So, the energy minimization is something that will keep coming ok that is all spread all over and typically there are standard ways to kind of do it and that is what right they have done, but what they showed was eventually that using the CRF right things become really better. So, here right if you see right they are and because here if it is actually in the iterative process.

So, with every iteration right things improve until you get down to something like that

right which looks very good right as compared to what they started off with ok and that is what is deep lab and with that right we are done with segmentation.