Modern Computer Vision

Prof. A.N. Rajagopalan

Department of Electrical Engineering

IIT Madras

Lecture-89

Now faster RCNN that is like you know one level higher, now that wanted to get of it do away with this block itself. So the idea was that right can you actually generate a network that can actually output these region proposals. I mean that is the ideal thing you would have like to have right, have a network that can actually do that, how would you do that? That is the goal of faster RCNN. It wants to do, it wants to build a network that can actually throw up these boxes that can tell likely regions of interest which are your region proposals. How would they have done it? Have you heard of anchor boxes? If you have not okay then in that case right you, okay what is your idea anyway you tell. It is somewhat like that okay, this is somewhat like that but now right in this case what we need really are these boxes right, we need bounding boxes.

Now in order to get bounding boxes right what they do is you know they actually imagine see for example I know I have an input image okay let us say that you know right let us say that I have an image and then I know that I know that there is an object here and I also know that there is a ground truth box there right and then there is another object here I know that I know that I know that right there is a ground truth box corresponding to that and so on. But then during inference time right I mean I cannot use this knowledge because in inference time I do not know where the boxes are right. Therefore you have to have some way to kind of use this information to flag. Now if you look at fast RCNN right it came from a proposal to the feature map right, feature map volume that is how it came right, it took the proposal and then found out where exactly is this guy.

Now faster RCNN does exactly the ulta it goes the other way. So what it does is it assumes that in the image you have at every pixel at every pixel this is again another this is a hyper parameter at every pixel you have 9 anchor boxes at every pixel that means right it is trying to say that imagine that you have anchor boxes all around what is this anchor boxes these are exactly are RPs right. So this is like a region proposal and by anchor boxes what you mean is you know what you mean is rectangles of 3 shapes square and then whatever it elongated and kind of tall and then the other 3 are like the scale okay you can have like 1, 2, 4 right. So you have kind of 9 anchor boxes you can change this it is a hyper parameter. So what it so it kind of right looks like this right if you actually already imagine so it is saying that at every pixel there are these 9 guys right waiting to flag something okay.

Now what happens is when you have this feature volume see as far as faster RCNN is concerned that this part is exactly identical to fast RCNN okay the only thing that really is new is here okay and as I said earlier right when what you want I mean if you had you know this one a network to actually do the job of flagging this one proposals let us call this RPN right RP network. So let us call this region this one a proposal this one a network right let us call this RPN right. Now this network its job is to actually do some kind of let us say that we also wanted to be a classifier in this case but then what kind of a classifier it should be I mean we should all that we need is it should be a binary classifier see right imagine the one below is going to do a lot more see the region proposal that is what I said right the region proposal the selective search when it flagged know that you could have you know bounding boxes here bounding boxes there these RP is read proposals when it threw up it never said what it contained right it only said that this region could be potentially interesting it does not have knowledge about what it contains it is like saying that here is a likely region of interest. So that is all you want from RPN you want to say whether it is 1 or 0 is it background or probably read is it foreground is it some foreground object we do not know which object and all that we will figure out later right because we have the class label for that and all we will do through the lower network but this but this RPN this network all that we wanted to do is do a binary classification and simply write tell us a 1 or 0 for let us say each of these each of these anchor boxes but then but then right the point is we do not want to send all the anchor boxes in right. Now we want to be able to we want to be able to write do it in a smart way what is the what is the smart way of actually doing it so like I said you take this feature volume and then and then the fact is when you have been actually go back right so what they do is okay here is where they kind of do something like a sliding window so I told right in the start that sliding window is actually not really a great idea when you when you kind of do it in the image domain but then I do it in the feature domain when it when it is when it is of a smaller dimension okay it is still right it is not it is not so bad but then that is what they do.

So what they do is you know you take you take a feature volume and then now you take an you take an say n cross n cross n sort of a window and then try to find out where it goes okay from kind of where this region comes in the image and then what you will do is so in that image you have got like I said you got anchor boxes all over right which are kind of which are kind of virtually sitting there right they are not really I mean there is a kind of virtually sitting there now what you do is once you map it back to that region what you do is you find out the IOU of each of those anchor boxes with respect to a ground truth box if it if it exists there right it may not be there no it does not mean that every feature volume if you map it back right it does not mean that there will be a ground truth box if there is a ground truth box there right then you will find that there is one anchor box right which is coming kind of close to that right that is why it is okay right that is in a way telling that here is a here is a region proposal. So you have to fix your IOU somewhere right it is say that some IOU greater than something could be could be actually you know could be a potentially interesting region do you kind of see the idea right so just kind of back project right in a sense go back look at the image look at all the anchor boxes there if you have a ground truth box in that region which you know a priori that information I have right like I said this I have with me right training time. So what you are trying to say is that you are making a network learn so the idea is that right if I you know during a test time if I just give it this image it should sort of immediately figure out that from here and all right I should get these region proposals because it is learned it is trained for that right it just goes back looks at that and then you know does not does not IOU with respect to ground truth finds out that there are a few boxes that could be actually potentially interesting pass them on okay. But then but then right but then you know but then what it will do is so the idea is that right so the idea is like this right what it will do is so here right so here it has a small network it is a small network which does a training in the sense that right out of these out of the proposal boxes right it still tries to find out which one of them should be actually should be like should be considered as a background and which one of them should be considered as a foreground. The foreground does not mean it knows the object simply means that so 1 or a 0 okay so which means which means that all because you see finally right finally okay we realize that we realize that when you want to train this network right we also want to know which is the background and all right because you do not want to put boxes see the problem that will happen is you do not know the background right that then one risk that you run is you know you can because you do not know the background you can end up actually right putting a box there because you do not even know how it looks.

This used to happen with you know if you look at these look at the classical traditional approaches this is always a problem for example right in a whole image you will have 5 faces it will nicely put 5 boxes there and then when you are very just feeling happy it will put one more box somewhere else where you do not even see a face there is no face there but there is things that there is a face there that is because it does not know the background and background is so rich see I mean there is there is not anything like a universal background right for every way the background changes whereas faces are universal I mean in the sense that right all humans look in a certain way objects look in a certain way but look at the background can you ever say that you know in every image they can have a universal background there is no such notion. Therefore it is also important to know what is the background that is why that is why you have a you have a you have a small network here which does that job right because it knows the labels right knows where there is an object where there is not an object therefore it tells that these proposals right I have no so all these so you need you need a positive set of proposals you need a negative set of proposals all that comes from there positive negative that is all right object or not that is all right so right that part comes from there it does a pruning also using some say NMS and all and then after that right this comes here and this this network is exactly the same okay because after this point it everything is the same right this guy will then go back find the feature volume squash it and then right make them all uniform learn end to end okay that is the idea of faster RCNN. So the idea is that you you actually have a network that will also flag those areas right so you want a region proposal network that can actually do it and right or this idea I do not think is that easy to get about the early one I would have thought is more probably right you just have to think a little bit and then you will get it but this I thought is really smart because it is not obvious right I mean how do you how do you get one network to flag proposals automatically right okay. You said that at each pixel there are 9 anchor boxes. At each pixel yeah there are actually 9 anchor boxes correct.

So I am not able to get a proper sensor then what is it what is that because. See what this means see no anchor box is like a really you know a bounding box right if you want to think about it so what you are saying is you know at a pixel rate you are imagining that one box is like this you are imagining another box is like this another box is like this another which is of a bigger area you got like 9 of them sitting at you know at let us say every this one pixel location and all that you are saying is when from this end by end volume when you go and map it right it will map to map to some area in the image. Now if in that area there is a ground truth box, if in the image yeah because I know the input image right I know that right I am I have this image and I know that I have a ground truth box here I have a ground truth box here I have a ground truth box that I know that information I know so when I go back right if it maps here then I know that this is in anything of interest because I know that a ground true box is not there. But then suppose if at some stage when I keep on sliding right what will happen you will encounter this, you will encounter this, you will encounter this right because you are going to sweep the whole volume that is why I said it is sliding. So at a time you can only do locally that is why you cannot go back fully.

Yes. Okay you have to do locally that is why I said that the sliding see I mean right people do not like people frown upon any kind of sliding thing because it means that it takes time but then in this case right there I do not think I mean I do not think that anybody has anybody has said anything bad about this sliding idea. It is not even emphasized that much actually we have to do a careful reading of the paper then you understand that that is a completely local operation because you know if you just take the whole thing you know then where is the locality it will go to the whole image right it does not really help. So it is sort of because you see finally region proposals have to come locally right during inference time what do you have you have only an image you do not have anything else therefore it has to say from this region should something be shot out or not right I mean you know as interesting right I do not know right some of you can still kind of write you know think about something like an extension which would not need a sliding window but in this case it is certainly a sliding window okay and then it goes over the volume and then for every sliding window you have these proposals that are coming out that are already sitting there and then if there is a ground truth box it is finding the IOE with respect to each of these boxes and somewhere you get an IOE which is reusable it could be more than one box all that is fine those are all trained as 1 or 0. That that anchor box or maybe more than 1 that pixel can also get flagged. How do we? How do we generate a yeah so I said no you have a sibling layer here so after this is exactly the same as the fast R-C-N-N you have a 4 dimensional bounding box regression then you have a classification this part is identical only what is coming in as a region proposal is coming from a network instead of coming from some selective search module or something and of course the way the training is done right there are actually 2-3 ways of training this but I think the most common thing is what is called alternating training in the sense that they train RPN actually you know the details you have to kind of write go through the vapour that is why I said I am not going to go through the details but really the way they do is there is some sharing of convolutional layers and all okay between these 2 between the fast R-C-N-N which is sitting down and then the RPN which is sitting up so they train the RPN come back train the fast R-C-N-N go back train the RPN and then in between there is some sharing of convolutional layers weights of convolutional layers and all so little complex training part okay.

It is not like the first one is trained done and then you come to fast R-C-N-N they do what is called alternating training okay so anchor boxes yeah so right you asked right so I think so if you think of a anchor box right that is how that is how they would kind of look like and then right here are some faster R-C-N-N outputs okay. Now finally right there is this not finally I mean as far as object detection is concerned right there is this YOLO right what is YOLO stand for you only look once right so you only look once but then the idea is like this right so when you say that you only look once then what it means is that the YOLO network right as it is called right it kind of takes one look at the entire image okay so that is why we call it as you only look once so where is the region proposal and all it is like you know you take a proposal from that try a proposal so the context information sometimes you may lose there even though you can argue that oh you have an AlexNet right which is the backbone that has actually seen the entire image while training therefore it must be having context and all but that is all implicit but then here it is actually it is straight away it is also in a way implicit but then here right you are not you are not good of looking at doing it that way right you just take an image you take one look at it and then say right what should be the what should be the right where should I have these boxes. Now and it is very fast by the way right I mean you know the other ones are also fast faster are seen and I think it is also reasonably fast but then this guy is like you know 45 frames per second so it is lunch in real time and there are various versions of this now the right only thing right that I thought that I thought actually it is a simpler network right so the way this works is as follows okay now so right let me just straight give you a hang of how this works the YOLO network right so what it does is right so the idea is like this right so it actually so if you have an image right which you would like to okay for which right you know you would like to have these bounding boxes right around objects of interest then what it does is you know it actually splits this into an automatically right splits this into an S cross S grid this S is a hyper parameter typically it can go from 2 to 19 and 19 is very common that means you really have like fine boxes there so each is called a cell a grid cell okay so it is like this right so what it does is see for example okay let me just okay now what can happen is okay let me just build this up a little bit more okay then let us say we have some things here okay so we have an S cross S grid right. So the way it works is that for example what it will do is see I mean if the center I mean suppose I have you know a ground truth this on a bounding box right which is here okay so for example right I mean now the idea was that I mean I wanted to draw a box right which is not exactly centered in this grid okay imagine that let us say right that this box is centered there okay even though that is not the exact center okay so some box right which is sitting there I have an object there now this grid cell right whose center so this grid cell as long as it contains the center of some sort of ground truth box I mean so for the time being right we will actually assume that we have only one object okay per this one right grid cell a grid cell may contain a center may not but if it contains it contains only one center so what this means is that then we say that this grid cell is actually responsible for that object. No this grid cell this grid cell wherein the center is falling this ground truth box is there no I am saying that it has a center here let us say it is a little skewed but let us say the center is there but that is falling in some grid cell right now that grid cell is supposed to be responsible for that object okay so for example which means that right which means that this grid cell is not responsible but then maybe some other object comes and sits right over know whose center is let us say there okay then this grid cell will become responsible for that object okay so what it means is the center of the ground truth box sits in a grid cell then that particular grid cell becomes responsible for that object only that grid cell then what it does is just like you had anchor boxes there right now for each grid cell you can actually think about different shapes of bounding boxes let us say right let us say the number of bounding boxes is B per cell this is again a hyperparameter okay you can choose it as phi or whatever right you can choose some hyperparameter.

Now what it does is so the way this works is right so the way to kind of understand this right this is actually you know a Google net okay with some small things right some small you know small things that they have changed here and there right more or less a Google net but then it is this kind of a tensor that you have to see at the output now this output tensor it is what they want so this they want this network to actually produce a certain output tensor now until now right I said that right we will not actually worry about the size and all I would not go to but here I have to. Now the 7 cross 7 by actually a 30 right where it actually comes from is because we want this network to actually predict this volume what this volume actually means is that means is that okay is this so it is the following right so the 7 cross 7 is actually let us say for the case when S is 7 okay and this actually and this 30 right is coming because of the fact that I think for 30 right so the way they have it is in the number of classes the bounding boxes are 2 okay I think for this case okay for this particular case I am saying you can choose bounding box whichever you want and then there is another vector right which I will talk about so every bounding box has actually 5 elements in it. See right think about this as some grid where the where the you know as a grid where there are S square cells sitting and in each cell right I have let us say bounding box let us say 5 bounding boxes whatever b number of bounding boxes and for each bounding box right I will have I will have certain elements which I want this is a network to output okay that is what I want this output tensor should be that so what you want is a bounding box for which right it should tell a confidence score which is given as one sort of a parameter S this confidence score will mean that will be in that if you took that bounding box and if you did an IOU with the underlying ground truth box that is sitting there then if you get a high value of IOU that means your confidence is higher that you can do now because you know we know where the actual ground truth boxes and then you have this 5 boxes for that cell so you find out what is the kind of IOU that you get with the ground truth box if S is high it cannot be of course more than 1 right so whatever between 0 and 1 if it is high then we know that okay that is very likely this is a potentially good bounding box which could contain my objective interest S and then it has an offset bx and by which will be the offset which will actually tell where is the ground truth box center relative to the sorry where is the ground truth box center relative to the bounding box center and if you take all bounding box centers to be about the center then that is offset will be the same for all of them. Then you will have bw and then b height what this means is how much should you scale your bounding box along the whatever right along the width and along the height in order to be able to match up to the ground truth box you get this okay all so these 5 are for that bounding box for each bounding box you got like you know 5 things right which you want to find one is a confidence score second is an offset and then third is the scale think of scale factors offsets and this one a confidence score. Then this 30 number is coming because in addition to this right so when we generate a feature vector right so when we generate this output right this is like I mean if you go along the depth right if you take so it is like all your cells are sitting here right you have to see 7 cross 7 these are your cells and for each one of them along the depth you have actually a 30 cross 1 vector that vector is nothing but these 5 elements that I actually put here that is S and then all the way up to whatever bh this 5 okay and yeah this is for a cell right yeah.

So this 5 into let us say right I mean in this case I have taken b to b2 right okay that means you already have 5 parameters for one bounding box correct 5 parameters for one bounding box into 2 bounding boxes for which I have to find out that is already 10 and then the number of classes number of these classes which is in this case actually 20. So the actual feature right it is like this you can think about this is S into S into b into 5 plus c so c is the number of classes. So what this feature vector right we really will look like it right is as follows if you look at this vector right it will look like this okay it will actually look like all these guys right I mean see for example I mean you know if you have 2 bounding boxes then you will have say 10 elements there if you have 3 bounding boxes you will have you know into 3 right 15 that is why this is 5 okay this is not S this 5 is fixed that is not a hyperparameter that is fixed that is confidence score 4 other values we got yeah confidence offset and scale so that 5 is fixed this is a parameter this is not fixed this actually depends upon what kind of a database you have so c is a number of object classes in your data set b is a hyperparameter which is the number of bounding boxes and S is the grid number of grids that you have that is decided by S. Number of bounding boxes means the number of bounding boxes whose centers are lying in that. In that cell in that cell you could also shift the center if you wish in which case you will have a bx vy that will change for every bounding box that is why I think they have left it they left it free okay now so which is why I am saying it is like b into 5 because you are assuming that everything could have an unknown okay now you have a b into 5 plus c okay c is a number of classes so here right what so now what will happen is right so you have all these guys right which are related to the bounding boxes then actually think about think about think about like classes c1 to let us say c20 and then and then this will be a one hot vector which will say that which will say that which will say that right so what this means is with respect to my ground truth okay see this is what I want as output in my output right I will say that if I have a ground truth box there I will know what should be the offset with respect to each of those boxes what should be the scale with respect to each of those boxes and then what should be the class right that that that that should come out right so that is like one of them will be a 1 everything else will be a 0 right that is my that is my kind of a feature that is my output tensor expected tensor and you are trying to learn this make this and you are trying to train this network such that it outputs that tensor that means for every bounding box right it should know that if I if I move by this offset if I move by this scale if I move if I do this then I will sit on top of that ground truth box and again right you can have again multiple boxes coming then you do an NMS exactly the same way pick the guy which has the highest highest confidence score here S and then and then you do an intersection with the other boxes wherever the intersection is greater than 0.

5 throw them all out okay that is how you do the NMS exactly similar to right how it is done in RCNN but then actually the overall network is actually pretty simple this all this all that is involved this is all right you are just asking for an for a for a for an output tensor and that and there you there you through all of you through all so you see right that that right you do not you are not really talking about a region proposal network this is one shot

right look at the whole image produce that tensor that is YOLO that is why it is simple I mean it is simple once you understand it is how it is doing right then it becomes simple the only catch is that right for example sometimes right you could have you could have you could have for example ride a car here and then and then you could also have let us say this bounding box is our car and then you could have another person standing in front of that car no right I mean if you look at the way right you are doing it right I mean you are only allowing one class per cell so now in this case right in that cell you have centers of both they say the center of the person is also coming because there is a ground truth box for that say for that person then you have a ground truth box for that car let us say both of them are coming somewhere inside right they do not have to be the same place one falls somewhere here one falls but both are falling into that grid now this original thing that we said will not allow that because it will certainly say take one but then that is not a good idea right you want to also say that there is a person sitting there because this grid cell should be responsible for both it cannot say that I am responsible so what they do what will you do so they simply do a concatenation for the other one right what happened I mean in the sense that you had this one vector it is for one object class no you do a concatenation wherein the other label will get flagged as one I mean if this was C5 car then maybe C7 will be one here that is up to you that is what you have to decide how many times such a thing can occur I mean if you underestimate it badly and train with two but if there are four guys coming then there could be a problem yeah that is something that you have to think about but that is all they do. Yeah see one other way to avoid this kind of multiple thing is make S really large that is why I said 19 right so the chances that two centers will come in a box is little so you can take care of that I mean right so just make it really small right so that you know even then it can happen that is why they say right beyond 2 and all probably it is not even needed I mean you do not like get like 4 guys coming in into one grid cell as long as that is why I said 19 is a higher side right you do not take the lower one you do not take 2 you go like 19 so you go like 19 cross 19 you divide the whole image into 19 cross 19 grids and then you flag okay now that is about that is about the detection object detection it is a way so we saw R-CNN we saw faster R-CNN we saw faster R-CNN we saw YOLO. Yeah two bounding boxes will have the same class label whereas if you have two objects then your class labels will change that is why I said then you will have to actually attach you will have to do a concatenation but a concatenate to tell that there are actually 2 object centers sitting there 2 different classes. Two possible boxes out of which you may finally take only one after doing non-maximus oppression and all but when you have 2 object classes center sitting there then it means that you want to flag both right you want to get a bounding box for this as well as that and that you have to flag because then you cannot have the same label for both so the class label will have to change right that is the reason why you need actually one more concatenation where you will say that there is another class label that should be one so it is not the same.