**Modern Computer Vision**

**Prof. A.N. Rajagopalan**

**Department of Electrical Engineering**

**IIT Madras**

**Lecture-90**

Okay, now finally right I just wanted to sort of quickly go through because go through this last thing right which is actually visual captioning. You would have seen it, you would have read about it but then I thought right we will just we will just talk about it briefly. That is again a high level task right to make. So all these are parts of high level vision right what we are doing now which was the final part and really right classical approaches and all there is no point right talking too much about them and most of them you can do now because they are all based on like I said right in the right in the right in the you can see beginning they are SIFT based or they are you know Haar wavelet based or they are you know hog based all those things that you know right. But then thing is they are not at all robust, they are not fast, they cannot run real time therefore the whole focus is on actually deep networks that is why for high level vision right I thought there was no point spending time again going back doing classical things and all when things have come so far. It is always better that you learn latest things right which are all deep network based.

So that is because even whatever you do will also be all typically deep network based you know because classical have of course you know they are good but then beyond a point right I mean they can only do so much by the way right. Now image this one captioning right so I am going to be very kind of brief about this okay. So what is image captioning right as we all know okay what we want to do right I mean it is basically you know and kind of you know interaction between what we will call as a visual feature and actually textual feature right like a language. Now until now right what we saw was only you know one sort of this one right this one a modality right in the sense that it was just image all the time or perhaps it was text sometimes right.

But then right here is the case where of course when we did RNN we talked about it right at the time we did flag this thing that you know you could have features coming from visual feature coming and then you could have you know this one you know textual feature coming and then you have to combine them and so on right. So now so this image captioning right what does image captioning mean? It basically means that if I give an image then you should have a network that looks at that image and actually produces a caption for it right which should actually best explain what is happening in that image. But then right what happens is even if I ask let us say individuals to do it right it is not it

would not be the case that what he says is exactly what I will say or let us say what she says right it will not be the same because she will think about it in a certain manner she will write about it in a certain manner I will write about it in a certain manner that is why when they actually develop a data set right what they do is they ask multiple options they give like for example right for each image there could be at least 5 to 10 different sentences all of them are valid. So as long as you produce one of them it is okay it is not like you have to produce that sentence and all and sometimes you cannot even produce exactly that sentence and all it will produce a meaningful thing hopefully where it conveys what is going on but then right what it conveys it could be right any of those any of those say 10 things and again when you give a new image it will just pick up from whatever it has learned it will throw out something okay. So until now right what okay all this we have done right so now so what we want to do is for example in this case right it is not simply that we are interested in object kind of detection or something right all that could also be potentially important for this problem where we want to explain what is going on because maybe it is useful to know that right that there is a person there it is useful to know perhaps that there is water there it is useful to know maybe to actually detect that there is an object there which is a surf board so object detection could still play a role it does play a role in these kinds of things but then right this talk right is going to be at a much lower level okay.

So what this means is that we want to have you know a description and that kind of description will perhaps involve surf will involve the ocean will involve surf board will involve water perhaps right so you could think of multiple options here maybe it is about the car it is about the trees it is about the road right all these you expect all of them to sort of come up when you when you want to say something about that image and as opposed to fixed set of visual categories right. So if you kind of look at what we did previously right when we did object detection we are always talking about n classes now whatever right n number of classes always a fixed set notion that we have in our head maximum thousand we do not even talk beyond that but then vocabulary is not like that right vocabulary is like you know whatever right millions of words whatever right so many words there tons and tons of words and then and there you cannot say that I can only have you know these many classes and all so therefore right this fixed level of classes and all that we have to kind of we have to we have to we have to we have to we have to walk out of that kind of a mindset and in this case right it is kind of producing ideally right we want to say that a person is riding surf board in the ocean and then for the next one we might say red truck is parked on the street lined with trees and so on right. So we need we need a description that should be rich now again right as we said before if you have if you have image right we know that finally from an image we can capture features similarly if you have a sentence which we know is made up of words so we want to know right what kind of features can be associated with words and that is what we talked about what is called word2vec right when we did RNN right if you guys recollect I talked about something

called word2vec which can actually give a word embedding and a word embedding is needed because you cannot use a one shot vector for a one hot vector for a word because it does not make sense right because words occur together in a certain way and therefore right in that space right if two words frequently occur in a sentence then there then there should be then there should be in that feature space they should occur occur right somewhere right near to each other however it does not make sense you do a if you do a one shot one hot right then one guy will be here one guy will be somewhere else whereas you do not even maintain that kind of you do not you do not actually I think all this I talked about right when we when we did word2vec. So the idea is that you want a you want an embedding which will actually capture the fact that certain words occur together or certain works typically come together in a sentence all of that should be captured as a word embedding right that is what that is what that is what that is what word2vec does right when you can do it using GloVe whatever right now as far as an image is concerned we know we can do we can use Alex net you can use VGG net any of those in order to get our feature map representation right. So the feature map again right again what you want to pick whether you want to pick the final convolution layer feature map or whether you want to convolute sorry the fully corrected feature map or whether you want to pick the convolutional feature map depends upon what you want to do see for example if you wanted to pay attention and attention is something that you have heard right deep networks with attention, attention is there everywhere now.

So what does it mean that means what attention means what that means when you are generating something you want this network to look at that region you want you to pay attention to something okay which means that that network should be able to go and locate something there if you give a fully corrected feature map it cannot do that at all right that is why when you want a convolutional feature map and when you want a fully connected feature map it depends upon whether you want attention or not if you want attention then typically it is a convolutional feature map because that gives a handle for this network to go back and so it says that I am going to pay attention to a feature in that convolutional feature map it means that in the image there is a corresponding region where it is paying attention if it is a fully connected thing nothing like that if you pick something it is the whole image right. So again attention and all when it comes because captioning with attention is a very common thing because you do not want to just throw captions like that you should also pay attention to what you are looking at that time when that word so for example if a certain word has already emerged and I want to write you know pick the next word up I should pay attention to what is there already with me and then I should also pay attention to the new word right I mean which I mean during you know when you of course train right when I train I am also going to give the next word therefore you want to pay attention where is that coming from right what is there in that image in that region so whether you pick convolutional feature map or whether you pick a fully connected feature

map depends upon various things including attention okay and of course this is that word to vector so what really do you want right you want a sentence to come out right which normally when you train you will train with lots of sentences that is how training happens right so you have an image you have a so you should have kind of paired data should have an image you should have of course you know for that image you could have a bunch of captions not just one caption you could have a bunch of captions and then you push this image you give one caption you push the same image again you say maybe the other caption is also equally okay right any one of them is okay but then something like this not see for example cats heart on mat right this we know is very likely to occur therefore such a sentence should get a probability which is high so this we are likely to see that anywhere you read probably cats heart on mat is more common mat's heart on cat is far less common right so you want a probability for such sentences to be low but then the point is right you know training a network with such a joint probability is not you know is not you know is not actually a good thing to do so that is why what you can do is you know you can actually you know give this conditional right you can arrive at this one you know a conditional sort of a representation because you know that you have you know a causal nature here right for example after I have at all have produced the first word and actually pick the second word this the second word right depends upon what came earlier the third word depends upon what came as second word and what came as first word of course you know beyond a point probably it does not even matter whether you need to know what the first word is but sometimes you can have you know a memory that can go really far back right sometimes when we do this summarization of movies and all right somebody says rotten movie he begins like that and then say something else at the end but then you want to make sure that that rotten thing what he said should still be factored in when you actually give some kind of a rating right so this connection how far should I retain the memory at all that is a totally sort of a different issue but I think this kind of you know dependence allows you to split this probability right this joint right you can actually write it like this and this is what exactly an RNN will do okay. So for example so what we will do is instead of predicting a joint probability of the sentence we will give kind of one word at a time and then and then and then at each of these probabilities we would like to maximize right so we like to we like to maximize the probability that word S2 comes out when I have seen S1 and word S3 comes out when I have seen S1 and S2 already right so that is exactly what I what I what I what I will do right so here is where what is this oh yeah so basically right so here is where here is how it will look so for example the first H0 right I mean okay so so right this is an RNN right it could be an LSTM or whatever and then you have an H0 here you have an H1 here you have an H2 here and so on. So what is H0 will do is it will actually take the features from your CNN in this case right because it is an image so this image goes in so so its feature map comes and then that is fed and then you get a give a start token so start token is something right which you have to give so which means that okay take this image take this start token and start something and during training of course

we know exactly what we want to say we want to say the caption is straw hat right that is our caption so we say straw should come out which means that which means that this this this H0 right will get you know you will get an H0 based upon this feature map these weights and this input which is start token and whatever weights that are sitting there then then you will output hopefully you will output straw right that is your whole idea and then there of course you have a you have a cross entropy loss here because because you have this vocabulary that you expect to come out and then that is right if that vocabulary does not come out then there is a cross entropy loss and then whatever is that word that you that that you that this that this that came out of this its word embedding will go next here right and then and now that it knows that the the earlier earlier earlier word was straw now of course you know so the earlier information is coming from here and then and then no straw okay and then okay the next one that you want is what is actually hat right which is what you which is what you kind of which is what you want as output and one straw and hat have come right you wanted to get an end so you want to get right put in an end token and that is why the end token will come right. So in a sense this is what you this is what this what is going on and this is exactly an implementation of this right where you where you are trying to write maximize the where the probability of a certain word given that given that you know a certain number of words have already occurred.

Now when you when you do this so so right so the way you can see it is for example right you would not take of course the last thing right you would you just take the fully connected feature map coming from here so right that is what you would you would kind of right push into H0 and then you give given the start token then out comes Y0 and then write Y0 its embedding will go here and then and then right and then it kind of see goes on right and then of course this and all you have already done when we did RNN. Now now after training right during kind of a testing time right what will happen so a testing time right of course I mean so so right it will it will it will start with a start token and then it will output something and then when does it know how to end when does it know how to end so for example in this case we know it is a straw hat that should come out right so so then what it will take the start token it will take the image of course right because that is the that is the whole thing it is going to it is going to take the image features and then and then now for this it needs to produce a caption right. So it will kind of see kind of produce something and then you know and then it that will go in again hopefully it is a straw and then hopefully it is a hat and typically what happens is right there is because it is an RNN right there is there is nothing like a length fixed length right you know in RNN there is nothing like a fixed length vectors right so what it will do is it will keep on producing words until it will automatically hits a stop token and that comes from whatever it has learnt. So wherever it is seen many times straw hat and there is a full stop then it will know that after hat it is over straw hat it should be a full stop okay sometimes that is why it is sometimes when it produces captions you find that something is funny about that caption

it does not seem to end at the right point because it does not know where to end it is not like a human right where you know that are precisely end here right it does not know that. So it has to go by based upon whatever it has learnt and then whereas if you have a transformer I think you can do something better you can fix the length and all but here you cannot do any of that if it is an RNN it will keep on producing until it itself encounters a stop token and stops there.

So when we show these examples it does not mean that exactly right that is the way it will happen right it does not mean that straw hat and full stop will happen okay. So for example so if you look at a training data set which is a Microsoft Coco here so right so somebody has done this work right 120,000 images for which captions 5 sentences each so these are all humans that have written the captions so for example you can write you know 5 different things about this guy a man riding a bike on a dirt path through a forest someone else says bicyclist raises his fist so all of these are correct equally correct right so any of these captions is okay that is what we mean by there is no single caption and finally the test caption when you give this image or something similar not this image when you give something similar it does not mean that you will get one of these it can still be a mixture of this right whatever it believes is correct and as long as you are okay with that right you are okay with that I mean that is why right even to get a proper score and all it is not easy in language but you see it produces a caption you have to know whether the caption is good or bad right. So for example so here it is done a fairly sort of write a good job you know a group of people standing around a room with you know with remotes which is actually pretty correct right it actually figure out that there are remotes here then similarly a young boy is holding a baseball bat very good right during inference time and then a cow I think this is from seen from India cow is standing in the middle of the street right what to do. Now here is where it went very bad a young boy is holding a baseball that is a toothbrush right went totally wrong and then a man standing next to a clock on a wall clock is correct man standing is correct but on a wall is really not correct right. So again so write such failures can happen and one of the reasons for these failures is that it is not paying attention perhaps right attention to attention to things that it should pay attention to right which is why image captioning with attention right is important and with the moment you say attention that then you are saying that you should also look at certain places in the image you know where from where right you should actually pick whatever information is needed and attention cannot be supervised.

There is no way to tell where you should pay attention to right when you look at something do we go and do we all pay attention the same way no each one has his own way of paying some attention we gather something important. So that is why supervision can only be on the words that come out you cannot have a supervision when you say attention so what you hope is that the network will automatically pay the right kind of attention to actually

produce the right word and the way you incorporate attention is of course you know now you have changed your layers right see the feature maps you have gone to convolutional feature map now automatically in the image domain because now you want to know where you are right where to pay attention and then what you do is you know on your feature map right so for example so it is like this right what you will do is you know if you kind of think about the feature map is being divided into some grids so what you will say you will attach you will attach an attention for each of these regions okay think of this as V1, V2, V3 and then you have an attention weight which is like A1, A2, A3 and so on. Therefore what happens is so this network you know in addition to producing whatever output it was earlier producing which was a vocabulary it should also produce the attention weights what that means is the attention weights will then be right taken through and then sort of rate given as given as the input to your to a subsequent word it means along with the word it is also telling where to pay the attention Z2 is actually a scalar it is like summation Ai, Vi so which will then mean that if it is a hard attention right what it will do is it will only have one nonzero Ai everything else will be zero that is like a hard attention a soft attention will mean that well you pay some attention there but then pay a lot more attention here so there are kind of different kinds of attention right within image captioning and now the hope is that right so when a bird comes right it will probably look at this area because in that image feature map right I mean again there is something that it should learn on its own we cannot tell there is a bird it should know where the bird is right so that is why right when they train this network in order to show that it is doing the right thing they actually do they show this they show that where is the attention you can go and map it no see for example I mean if I have a weight 0.1 I have 0.1 and I have 0.

7 here everything else is small if I go and map this region right where is the highest intensity for 0.7 hopefully right it should be on the bird at that time. When this bird bird pops up. Yeah when that bird bird pops up so it is like when I when I when I kind of print that caption right the bird at the time I should actually look at the attention where was the attention then it says bird sitting on the tree then when it comes to tree at the time if you look at the attention because that attention map is not this that is gone here no somewhere here it will come when the tree comes at the time you should go and look at the attention map then you will see the roads looking at the tree so that is the way to kind of convince yourself that it is doing the right thing because you are not there is no supervision that is that is why it is not that the wall should automatically come yeah see that is why I write I mean how to remove that bias and there is a lot also to do with these data sets itself the data set has a bias right then you will learn all that bias right you cannot help it. So, you will see if the data set included some image where the clock is on somewhere else then.

No, but then you will see that there will be some other problem then would it be something else you know, but then the hope is that on the whole it does something reasonable on the

whole you should say for example, right so bird right so when this bird came let us see the attention right where the attention is right so for example, flying again probably that has to do with the bird therefore, the attention is still there and then body so now you see that the attention is shifting to the water and then when you come to water right look at I think around the bird it is all fully water only it may not precisely happen, but you can see there is something going on I mean you cannot be you cannot like you got like hang the author saying that hey you know how come there is something on the bird and all right it will do something, but I think you can you right you can you can see the idea right. But then but then it is something is still not there right which you would have like to have which is like which is like you know see one other thing right that you might want to do is you know which are actually extensions and all right I mean after this so for example, so see look at this a large white bird standing in a forest it has gone wrong right I mean this animal right it kind of took it to a bird took it to be a bird despite the attention and all attention and all is beautiful on that object, but now right it thinks that it is a bird. Now if you had I mean I mean if you had you know something like you know prior to that right if you had an object detector which would have done its own job and told that right okay right I mean you know this is a this is a animal right this is actually right a certain animal a giraffe in this case then this guy could have used that and not fallen into some bird trap right I mean it could have figured out that there are only these many objects in this event therefore whatever word I produce whatever sentence I produce cannot pick something else other than the objects in the scene right. So such constraints you can bring in that is what that is what people do so that is why I said when you have when you have you know this one a detection object detection network it does not mean that right that is a that is a you know a standalone problem sometimes you might want to bring that in right when you are doing image captioning because you want to let the captioning network know that hey look this image has these objects you could have missed some in which case there could be trouble it does not mean that object detection is going to be 100 percent perfect right what if there is an object and it missed it because it was too small it could not happen or it was occluded gone right so all kind of things still got but then you know you would not say a bird hopefully right I mean you will realize that you know this is actually a giraffe and so on. So the extensions are like that I mean there are very many so it is a whole interesting area right where you can where you can go there how to marry the two right vision and language in order to be able and then you have video captioning then you have lots of things and you know again in video captioning it is about which frames should you pay attention to because what is the point in paying attention to every frame because probably there is not much what you see in one frame maybe the same thing exists in two frames here and there so right so basically do not just pay attention to every frame again attention to which frames right so there the temporal thing enters now here the attention was on which part of the image now the attention shifts to which frames in the video where should be the attention right so just scales up right like that.

So anyway right so there is a lot more so I think anyway right so I mean we can go on and on and on and on and on right I mean I write endlessly but then we have to kind of stop somewhere so with this right we actually end the course.