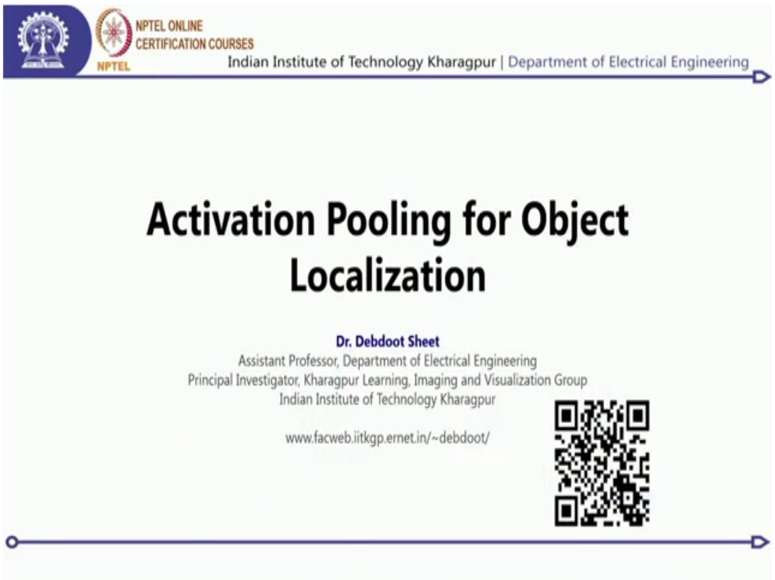
Deep Learning for Visual Computing
Prof. Debdoot Sheet
Department of Electrical Engineering
Indian Institute of Technology, Kharagpur

Lecture - 46
Activation Pooling for Object Localization

Welcome everybody to today's lecture. So, at the start of this week we are going to focus on another important aspect about

(Refer Slide Time: 00:22)



NPTEL ONLINE CERTIFICATION COURSES

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Activation Pooling for Object Localization

Dr. Debdoot Sheet
Assistant Professor, Department of Electrical Engineering
Principal Investigator, Kharagpur Learning, Imaging and Visualization Group
Indian Institute of Technology Kharagpur

www.facweb.iitkgp.ernet.in/~debdoot/

Visual computing and that is about object localization. So, all the our different topics which we have covered till now, where more of dealing with something which is called as object classification and at brief points of time, when we were doing, down about our different kinds of labs and theories about auto encoders you did learn about another aspect of regression over there. But nonetheless either in classification or regression what you were trying to do, is that there was a whole image and there were then you were trying to give an, inference about what's there in the whole image as such?

So, if it was a regression problem in general you were trying to solve down say a denoising auto encoder or just reconstruction mechanism with an auto encoder over there. In that case, it was just a pixel to pixel inferencing which was done and in the other case, was when trying to use a cnn and do any of these problems with your siphon net. Then, it was just a ten class classification problem which you are solving.

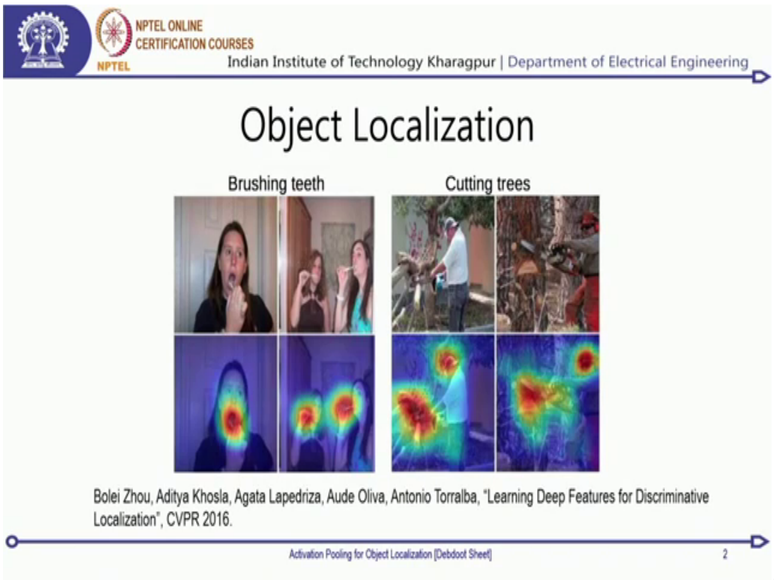
Now, the other this is while, this is one side of the whole gambit of problems and then issues which we deal with, in these kind of problems the other side is, about object localization and tracking. So, here what happens?

Is that you have a whole image given down over there and then, you have a certain object, which is present within the image and you would like to infer only about this object present within the image or say there is a small boy who is a young boy who is playing down with a bat and a ball and then, your objective is just to localize on, where this ball is located? Where the head of the boy is located?

So, these are kind of problems which we are going to deal with today. And where comes the interesting factors at the first method which I am going to discuss about and that is what I discussed today is, where you do not need to make use of any explicit information about where the object is located? Or you do not even need to train the whole classifier oh explicitly to be on this object itself. So, the object say if there is a ball present in the whole image and, it is just a ball classification problem.

It should give a yes. So, that is what does he ended was, otherwise doing always. Now, the whole aspect was that it was able to tell you that there is a ball present in the image, but where is the ball present in the image? Still, a question, which was not solved by these kind of cnn models. So, that is what we are going to solve today, over here. So, the people, which I am going to refer down is called as a learning d features for discriminative localization.

(Refer Slide Time: 02:35)



NPTEL ONLINE CERTIFICATION COURSES

NPTEL

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Object Localization

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba, "Learning Deep Features for Discriminative Localization", CVPR 2016.

Activation Pooling for Object Localization [Debdoot Sheet] 2

So, this is from cvpr 2016 and, from the mit and here, as, as you can see in this examples over here. So, there are two images over here, in one image has one person brushing teeth the other image has two people who are brushing their teeth.

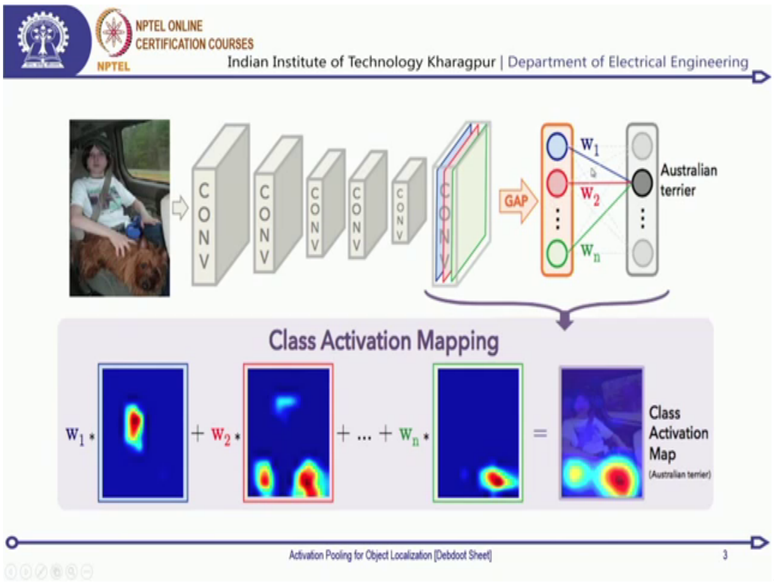
And then on the down side over here, you see some sort of heat map generated out over there, which more or less localizes out where is that brushing teeth action and then what! in the image actually triggered this network to say that, this image had somebody who was brushing the teeth.

Now, the next one is about cutting trees. Now, these images are what localized out end and point down as to different areas, where you can see like it if this is an action or that is one certain object present in the image which is going to classify this action and say that this is a tree cutting action going down over here. You can see a chainsaw and then there is a portion operating it. Over here you do not see the portion over the you see some part of the portion and that chainsaw over here,

Now these two examples which clearly say, over here you see a hand and a toothbrush held over there and then you are a mouth and that is what's triggering the action that this is about brushing a teeth. Now, nonetheless the point is that, one way you could solve it out was; obviously, using a regression problem if you had these kind of ground truth. So, somewhere that this region was marked that this is the region, where the which signifies that the portion is brushing a teeth or this is the region which signifies that somebody is cutting a tree, then it becomes much more easier; however, here you do not have that.

So, here what we tell is that the image is something which specifies an action? You are not necessarily saying, what part of the image is going to specify an action? ; however, the challenge is that, you need to train a network in some way, with this limited amount of information which specifies certain action is going down in this image, in order to find out where exactly this action is going on in the image?

(Refer Slide Time: 04:37)



NPTEL ONLINE CERTIFICATION COURSES

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Class Activation Mapping

Activation Pooling for Object Localization [Debdoot Sheet]

3

So, that is where we start down. So, at the starting point over here, we have done multiple of these convolutional neural networks. So, you have studied about Alex net, lee nets, Google nets, vgg, nets rest nets, dense nets, any you can take basically, any of these networks over here.

Now, the primary goal of these networks, were that there were a stack of convolutional layers. Now, once you have these stack of convolutional layers, you see that the layer thickness would be increasing because, the channel width over there keeps on increasing as you keep on going along the depth over there and the spatial size typically keeps on reducing size that the whole idea is that, somewhere near end over there, when you have a classification. Then you are boiling down to much lesser number of features. Say some 1024, or 4096, or 512, these kind of lesser number of features. So, typically for a vgg net kind of a problem, you were solving going down to about 4096 six features over there.

Now, at the essence of it, how it was coming down to those 4096 features was that, it kept on convolving convolving and then at some point of time, you had an tenser, which was really fat it had a lot of channels over there and the spatial span was really small. Now, as the spatial span is really small and the thickness is much fatter over there now, you can use this as something which is splitting out all the information across the channel. So, that is where we come down over here.

So, here the point is, that I have this kind of a convolution given down and it keeps on coming down till a point where the spatial span is really small. Now, once you get down a spatial span, which is really small, what I can do is that? I can chunk out each channel over here and again, going back to our discussion, as you would be going along the depth and you come down to much lesser number of spatial span and more number of channels the whole idea was that, in order to compress down the spatial information, into a channel information over there.

So, that the spechial resolution of each channel is low. But, the total information content, with respect to where an object is present? And, whether an object is present in the image or not? That is what gets consolidated within these channels over here. Now, each channel is in some way represent is, is in some way representing and is responsible for finding out which attribute is responsible for categorizing a certain objects?

Now, that we have this part of it. So, what we end up doing is, we initiate a certain procedure, which is called as global activation pooling. Now, the whole idea is that one channel which comes down over here, can be represented in terms of the mean activation of that channel or just the average value of that channel.

It can be an average value, it can be a sum of all the values nonetheless, it does not matter much over there because, anyways this, average value is just a scaled down version of sum of all the values in this channel. So, all the particular pixel locations in one given channel this what is going to be averaged out and represented in terms of one single neuron over here. Ok?

So, as we look over here, this is something which has just three channels over here. So, we can consider just three channels coming down or n number of channels and they will appropriately be mapped down over here. Now, using these activation maps over here, I am going to retrain and find out my classification engine and in the whole process, I am going to obtain my weights over here. So, what this necessarily means is? that in the earlier case with a cnn what we were having, is as we go along the depth you get down denser number of channels and lesser on the spatial span over there and then you try to linearize them into some linear neural networks and after these fully connected networks, you get down your classification coming out of it.

Now, here we put down another kind of a pooling layer in some way. So, this pooling layers job is to take down an average value of this one channel at a time. Now, if we put down this kind of a pooling layer over here which can be called as a global average pooling. So, it its independent of what is the size of that activation map over there? it's it just going to pull down. So, and it is not even necessary that you have the activation map at that last convolutional layer, you can have these activation maps even pulled out from intermediate convolutional layers ; however, it is not something which is suggested to do the because typically, you get the best granularity on the feature space at the last convolutional layer itself.

So, once you pull it out, then you can train again over here in order to do your classification. Now, once your classification network is trained over here, you have just an update obtained over these weights present over here, and these weights can now be used for weighing down these activations coming out of these convolutional layers.

So, what essentially you do after that is that, say for this green layer over here, is the green channel which got represented in terms of its mean value in this green neuron over here, or say for this blue one which gets represented in terms of its blue neuron, the red one, which gets represented in terms of its red neuron over here. Now, for each of them you have your weights.

Now, what we end up doing is that instead of taking these weights from here and summing it up, we take these weights back onto the layer part over here. So, here what we are going to do is? That each of these weights is going to weigh down this activation map which comes over here. Now, on the activation map side over there what you essentially get down is that, each pixel is going to have a value a very high value or a low value and then these highs and lows where going to vary across the spatial span over there.

Now, as they keep on varying over, the spatial span the whole objective is that wherever. So, since everything is boiling down from this original space over here. Now, you can actually interpolate and map it back onto this original space completely. So, here for this example, it was given down an Australian terrier and then, your objective is that let us find out, where this Australian terrier is located over here?

Now if this is the classification which comes out of this one. So, what we are going to do is we feed down an image over here and as it keeps on going through it we get down these weights corresponding to Australian terrier.

Now, keep in mind, there would be weights of corresponding to, there would be weights corresponding to window or car window, there can be weights corresponding to the seat belt, any of these options over here. Now; however, our only objective is to find out this Australian terrier. So, what we do is we taken these bits which correspond to the Australian terrier only. Ok? And then, use those weights in order to map it up. Now, when you do a weighted summation of these activation maps over there, this is what you get done. Now, one thing what you can see is that there was some activation coming down at this part which corresponds to the boils the chest area over there. Ok? Then, there was some part over here which was corresponding to the region the hair part of this work.

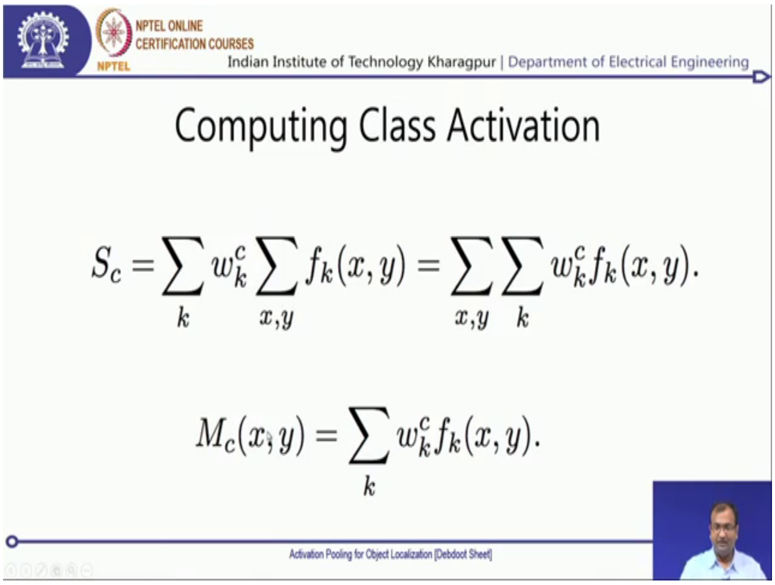
Now, these weights essentially become very less and as a result, what you see over here is that this final activation map which is supposed to find out only the Australian terrier this does not have these kind of spurious activations in any way. And as a result, you see that you can find out each and every object. Now, we basically developed a classification engine over here and used it to retract and get back on to finding our, actual object and where is it located.

And now, what you can also have is? Since, these weights are very much specific to which particular object category you are going to classify?. So, you can actually pick up an object category and say then I want to find out where is this object actually located?

So, as in over here was the Australian terrier. So, it was easy and straightforward way of carrying out. So, if you had seat belt or boy or something over there, you would definitely be getting a different weightage coefficient over here, over all of these activation channels and similarly that would be what is going to impact your final activation coming up?

Now, let’s get into the math,

(Refer Slide Time: 12:10)



NPTEL ONLINE CERTIFICATION COURSES

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Computing Class Activation

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k w_k^c f_k(x,y).$$

$$M_c(x,y) = \sum_k w_k^c f_k(x,y).$$

Activation Pooling for Object Localization [Debdoot Sheet]

of trying to solve it down. So, over here what happens is, quite simple. So, we need to find out this weights wkc. Now k is basically, which is representing down one particular channel over there and c is the coefficient which represents a particular class and sc is supposed to be the total activation for that class.

As a resultant of that final classification neural network; this inner summation over here, which is summation over x comma y of f k x comma y so, this is what is going to sum up all the activations in a given channel, in a particular one. Now k is the channel over here and then, you have a weight associated with a channel and then you have each of these classes coming down over here.

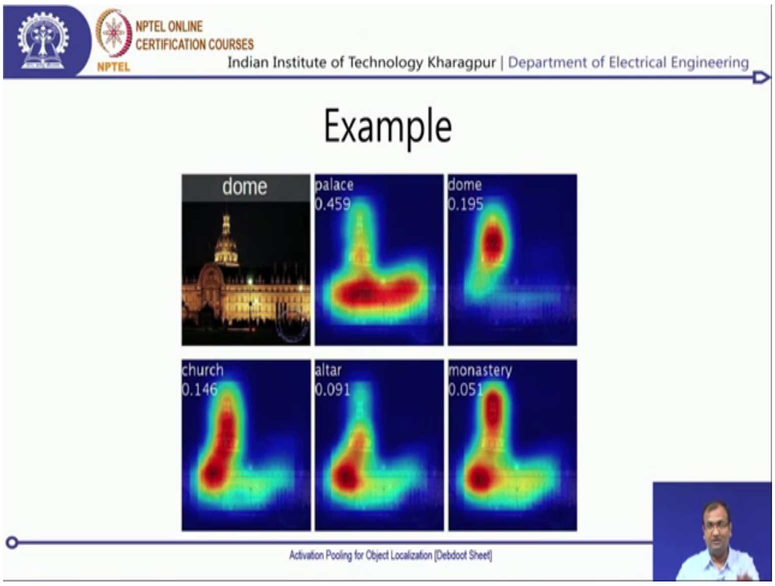
Now, what you are going to take down for your classification class over there is, a summation over this product over all the particular, over all the possible channels over there. Ok?

Now in essence, this whole thing comes down also to a form which looks like this which is? Where you can have each of these channels weighing down the coefficients over there and then take a summation. So, this is essentially what we unwrap and find it out.

So, if we take this part on my second part of this equal to, where I have this inner summation, which is summation over all the channels of wkc fkx comma y then this is going to give me and if I do not take a summation over x comma y then that is going to give me a whole channel magnified view of where my objects are present? which I can obtain from the earlier mechanism and, and this is just by way of splitting down my multipliers over there because, each is independent of the other.

And essentially what that is going to give me, is this part I can write down as some sort of a activity map of at x comma y for a given class c which can be represented as this inner product. So, so this inner part is what gets represented over here and this is my activity map which comes down. Now, that is what I would be using down for finding out different kinds of objects and their location in an image.

(Refer Slide Time: 14:20)


NPTEL ONLINE
CERTIFICATION COURSES
NPTEL
Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Example

Activation Pooling for Object Localization [Debdoot Sheet]

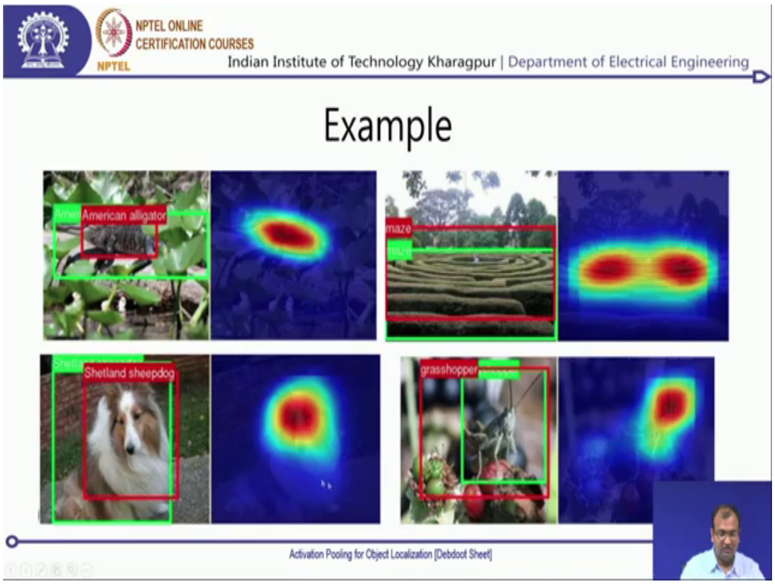
So, as an example say you take one of these images, which is otherwise classified as dome and then, since we were putting different kind of weights for all the activations over there then, you see what are the activations coming down under different objects. So, if I try to classify this as a palace then I find out, what are those weights associated with the palace? And then find out which parts of my image actually say that this is what is corresponding to a palace.

If you look down at a palace, yes it, it takes in almost whole of the structure which looks like a palace. If you say dome then, it just takes out only this topmost part over there, which signifies a dome. If you call this as a church, then it’s going to look only at this part. So, this is somewhat it looks like the entrance of a church and this is the tower of a church and these parts is what’s it’s trying to neglect to the maximum possible way. So, most of the images of churches will not be the side view or image of a church was the front from the altar where you are going to enter.

So, now, if you classify this as an altar, then you see the front entrance porch over there, which looks like an altar and there is something which looks like a monastery. So, they an object can have say sort of different kinds of variations over there. So, if you try to look at a cat and a dog and on an image which has a dog and you try to find out where is my cat located? Yes, it will still give you some certain manifestations which would look as if like a cat. So, a cat also has a tail, a dog has a tail, the cat has four legs over there, a dog also has four legs over there, a head two ears.

So, these are things, which will come down in commonality and you will get down, you will end up getting some sort of activations coming down over there as well. So, that is the whole point why we are ending up getting these kinds of activations, though this is not an exact object. But then these objects look very close to the object which we are trying to find out over there the.

(Refer Slide Time: 16:07)



NPTEL ONLINE CERTIFICATION COURSES

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Example

Activation Pooling for Object Localization [Debdoot Sheet]

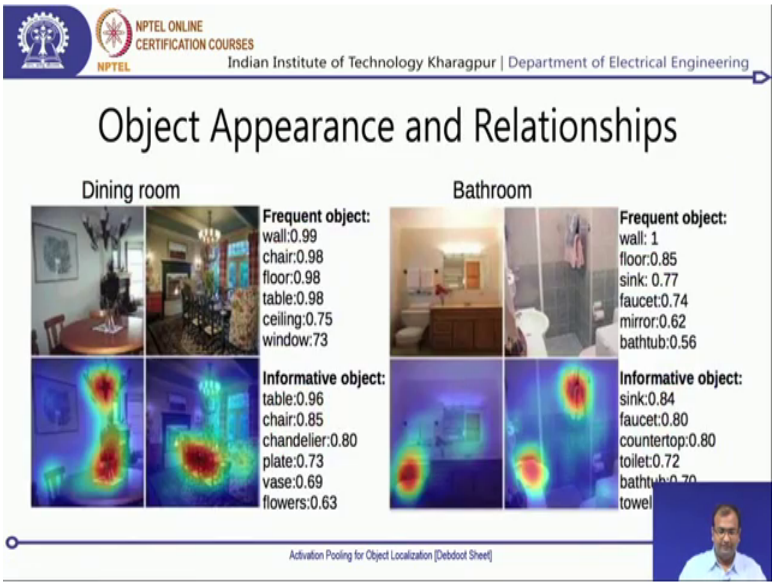
As an another example, that was where you found out the whole object located in across the image in terms of a heat map. So, now, we can actually have multiple objects over there as well and, what that would mean is that, so we need to somehow find out where, where the whole object is located?

So, this is a way where you find out, if say on this image if we try to mention, what is an? Where is an American alligator? Ok? Then, like the ground floor there is something which is on this green block over there, but the found out value is what is marked out in this red part over there.

Now, if you find out what? Where is a ship land shepherd sheepdog over there? Then, this is where it comes out and that is more of based on these activation maps. So, you can actually have some sort of object localization mechanism which you can create down with these kind of activity maps as well.

So, similar goes down with the next example where, where there is maize? So, yes it finds out and, when there is a grasshopper? So, the localization is much more coarse in terms of my actual found out object over here. So, there, there can be different ways of doing that.

(Refer Slide Time: 17:17)



NPTEL ONLINE CERTIFICATION COURSES

NPTEL

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Object Appearance and Relationships

Dining room

Frequent object:
wall:0.99
chair:0.98
floor:0.98
table:0.98
ceiling:0.75
window:73

Informative object:
table:0.96
chair:0.85
chandelier:0.80
plate:0.73
vase:0.69
flowers:0.63

Bathroom

Frequent object:
wall: 1
floor:0.85
sink: 0.77
faucet:0.74
mirror:0.62
bathtub:0.56

Informative object:
sink:0.84
faucet:0.80
countertop:0.80
toilet:0.72
bathtub:0.70
towel

Activation Pooling for Object Localization [Debdoot Sheet]

Now, another interesting fact which say presented in this particular paper was about, the frequency of appearance of certain objects in case of a certain kind of a classification. Now, if I am trying to look into a dining room over there and then, what will be the frequency at which my objects are going to be present?

So, wall has the highest chance then chair, floor, ceiling, window and these of there. So, what this makes out is? That, if there is a certain room or a big object to be classified out, that is made out of smaller and distant kind of attributes over there.

Now over here, the attributes say for a dining room, is what consists of the wall, chair, floor, table, ceiling and a window and that is what you essentially get down over there? Now, if we look down and the next part, which is informative object, which means basically, what is the weight associated with these objects being localized over here? And in terms of, what the neural network is learning over there?

So, these weights are also something which are almost in line with this one; however, except for the fact that, chandelier is something, which really gets a higher weight, which was not even featuring out on this list. So, that is another a newer kind of a thing you can look into an object discovery.

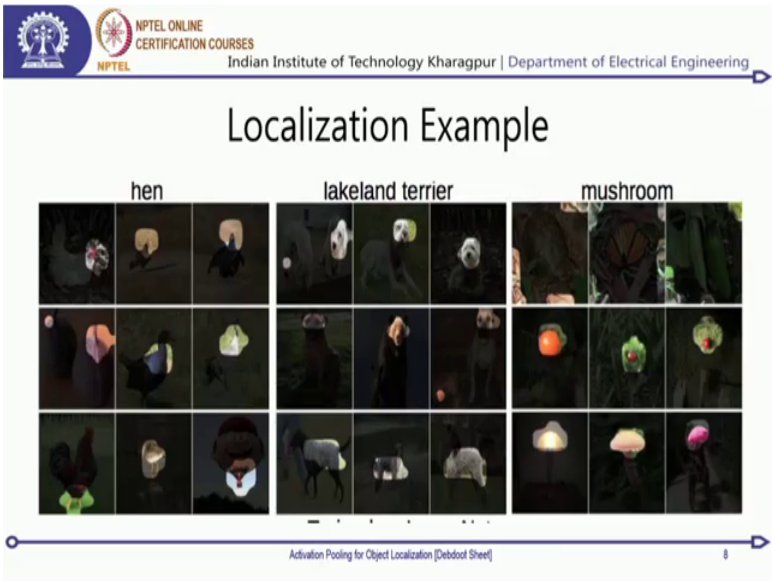
So, something we as humans often tend to ignore in case of understanding certain scenes can now start popping up that has a major significant advantage, if you are trying to do if you are trying to discover your kind of visual attributes, which signify a particular kind of an object or an appearance over there in an image. So, these can help you out in terms of even doing visual discovery of features as well.

So, similarly, if you look at the bathroom part over there, you would be seeing similar kind of an examples, where you have say the wall, floor, sink, faucet mirror and bathtub and then, what is the relative frequency in which they keep on appearing over there? So, and then these informative objects or informative frequency is something which is just guided by the weights over there and nothing beyond it.

But nonetheless, this is a very generic way of finding out different objects and its really easy to do; however, the only downside is that you need to be able to have enough number of smaller object, or smaller granule object categories over here in terms of classes so, that you can train the next part of it where do you find out the weights.

So, the initial part of the cnn is easy to train, but that subsequent part where you need to find out these weights is something which is interesting to train and might be complicated at time. So, you need to have sufficient number of a substantial example, so that you can train that one as well.

(Refer Slide Time: 19:43)



NPTEL ONLINE CERTIFICATION COURSES

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Localization Example

Activation Pooling for Object Localization [Debdoot Sheet] 8

Now, as an example what they show in their paper, is a segmentation map over images. So, you have the main image over there and then do some sort of an alpha blending such that, the brightness of the intensity levels of the objects which are other than the main object over here is diminished. So, that you find out where?.

So, if you are looking down for a hen and you want to classify it on an image you see the hen comes out, but then there are even misfits you would get down. So, one misfit is definitely that that there is a parachute and then, people are on this parachute or hot air balloon basically and they are riding that one even that, I do not know in some way the activation maps have a 0 down on that one, looking as if like a hot air balloon also resembles a hen in some way and then, there is a part where the legs of a hand are pointed out well.

That's technically not improper in any way or wrong. So, that is also justified that the legs of a hen can be used as discriminant objects and discriminant attributes for finding out where the hen is? Then say for this dog, which is a lakeland terrier?

Now, in most of the cases it was zeroing down on its face and it was finding it easy to find it out , but then in some of these cases you see that, there is not exactly a dog and then even over here, you do not see a dog in any way. And still, somehow like looking at the body over there for different animals it just find.

So, then say for a mushroom in some of these cases it 0 down on near a butterfly as well, in order to find out and some of these traces regions over there were identified as mushroom, where as these mushrooms, which are pointed out they are perfectly connect over there. So, nonetheless the major challenge still is that it's not a very robust technique or a reliable technique.

You still can get down a lot of false negatives, sorry false positives which I am suddenly, pop up over there, but nonetheless activation pooling does provide a very easy and standard technique for finding out the location of certain objects within an image, which you can use for describing a scene.

So, that ah place a long way into, what we are doing?. So, in the next lecture, I am going to get you introduced onto another kind of mechanism, which is called as a region proposal network. So, what it does as while you are using a network in order to classify you are in this particular example, you use those activation maps in order to localized it out. I

In case of a region proposal network, you are going to use those features over there for creating another network, which is going to give you a region proposal or some sort of a bounding box hints of where the actual object may be present.

So, we just stay tuned for the next lecture, where we I am going to come down with a that example and also, there are newer variants called as fast are cnn and faster are cnn and, how the computational methods were implemented over there?, in order to make this computationally really attractive and first is something which we will be doing in the lecture. So, till then stay tuned and,

Thanks.