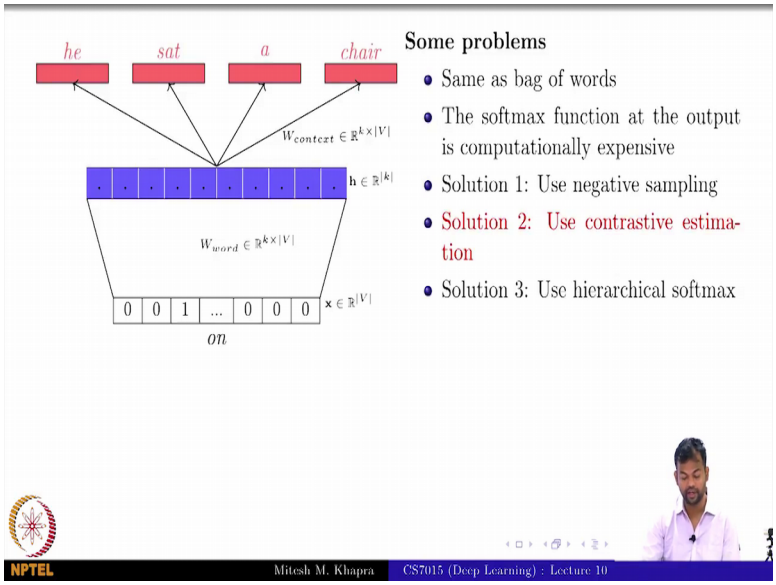
Deep Learning
Prof. Mitesh M Khapra
Department of Computer Science Engineering
Indian Institute of Technology, Madras

Module – 10.6
Lecture – 10
Contrastive estimation

So, we will move on to the next way of dealing with the expensive softmax. So, remember that; so, this is known as Contrastive Estimation.

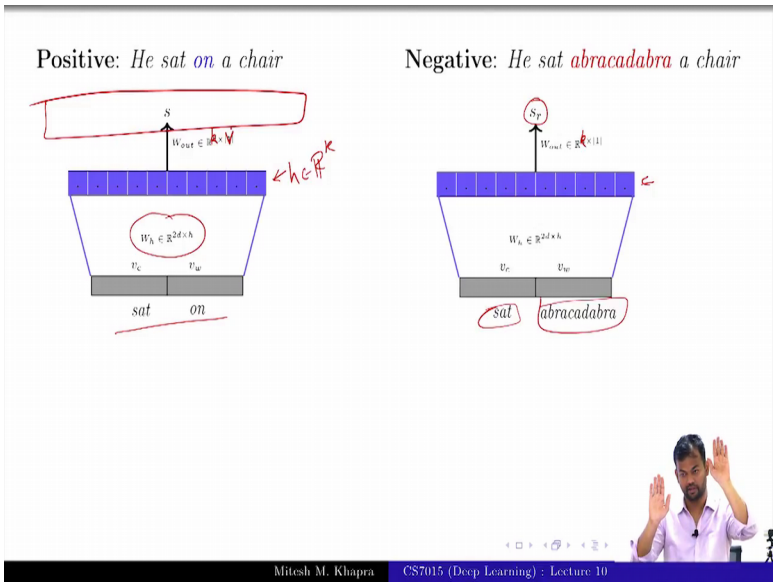
(Refer Slide Time: 00:24)



So, remember that this is where we are in the story that, we saw the bag of words model we saw the skip gram model. And we saw that both of them have this expensive softmax computation at the end, and that is the problem we are trying to deal with. So, we saw one way of dealing with which was negative sampling. So, you I hope you saw that there was no expensive computation there.

The only computation there was the dot product between the 2 words, which appear together or which do not appear together. Now let us see what happens in contrastive estimation.

(Refer Slide Time: 00:51)



So, here again you use a same idea. So, you have a positive sentence or a positive example he sat on a chair. You create a negative sentence which you replace the word by some random word. Now you construct a feed forward network like this which takes these 2 one hot representations. Basically uses your word context matrix, to give you the summation of these 2 representations right. That is exactly what we have done in the skip gram model. Now you have this hidden representation which is the sum of the 2 word representations.

Now from here on instead of doing this softmax computation which we had earlier; we just predict a single score. We just predict the score for this word pair being of correct word pair we do the same thing with the random pair. So, we take sat we take abracadabra and the add up there word representations, you get this hidden representations and you get a score Sr fine. So, what is the output computation right now? What is the; is it a matrix operation? Is it a scalar operation? Is it a vector operation? What is this h is equal to? We need to change this to k or (Refer Time: 02:11) right. So, what is this product w into h, just a dot product between 2 vector, right?

W is just k cross 1; that means, it is a vector. So, as compared to k cross V earlier, we just have k cross 1, you get that? How many of you get this we have a very simple computation at the end, but now how we set up by loss function? Earlier I could set up the loss function as maximizing the log like it of the correct word, but now I just

predicting 2 scores. So, what is the loss function? What should I try to intuitively do? And today they are going to see a new loss function which we have not seen earlier. So, try to think about this, what would you do? Forget about the math forget about the machine learning all that, what would you actually want? What is your wish list? That should be easy to characterize

Score S score Sr do you want this or this first one right, you want S to be greater than Sr. Can you think of making an objective function out of this you want to maximize.

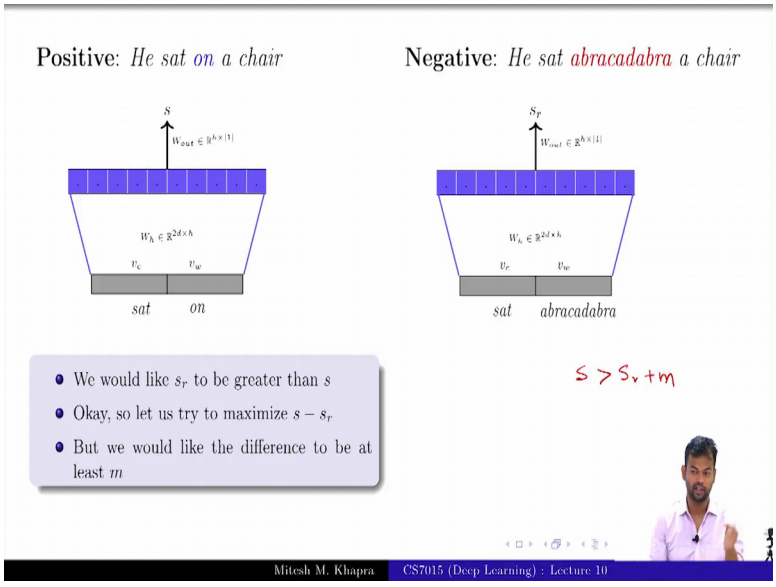
Student: (Refer Time: 03:18).

S minus Sr. Fine that is a good starting point. So, would you be happy with this? What would you want? This or this, both cases S is greater than Sr right, what would you want?

Student: A big margin.

A big margin fine

(Refer Slide Time: 03:39)



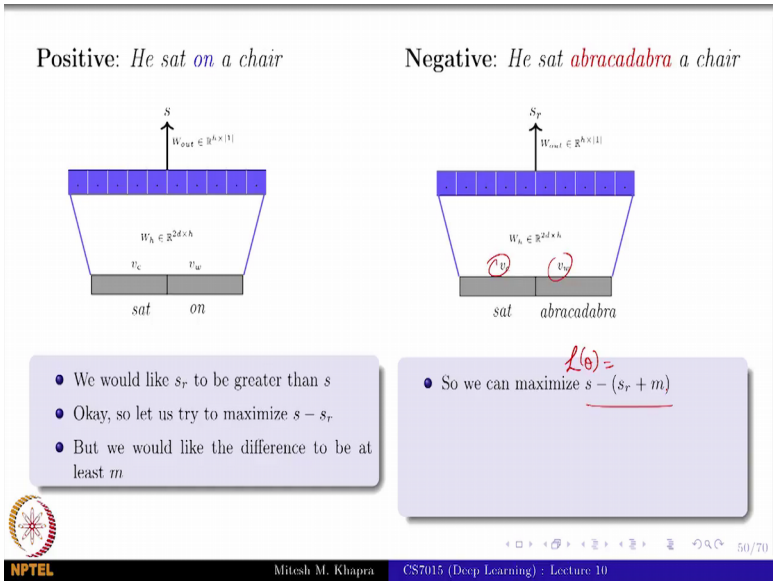
So, we would like Sr to be greater than S and not just. So, we could try to maximize S minus Sr, but we would also like this difference to be a certain margin; that means, I would want S to be greater than Sr, by at least a margin of m; and that m is something I will decide. So, I could say that it should be at least 10 points greater than Sr or 1 point

greater than Sr, depending on the scores that I have. So, all my scores are between 0 to 1 then probably a margin of 0.3 or 0.4 is, from 3 0 right; that means, S could be 0.6 and Sr could be 0.2 does that make sense right?

So, what I am saying is, what I am trying to say is that; this is my Sr, I want S to be greater than Sr. I am not just happy with that I am saying that even if I add a margin to Sr even then this condition should hold right.

And that is the same as saying that S Sr and there should be at least a margin of m between that. That is the difference that I accept. I am not if you tell me that S is 0.99 and Sr is 0.98 where then you are not really distinguishing much. I want at least S to be 0.9 and Sr to be at least less than 0.5 or somewhere.

(Refer Slide Time: 04:55)

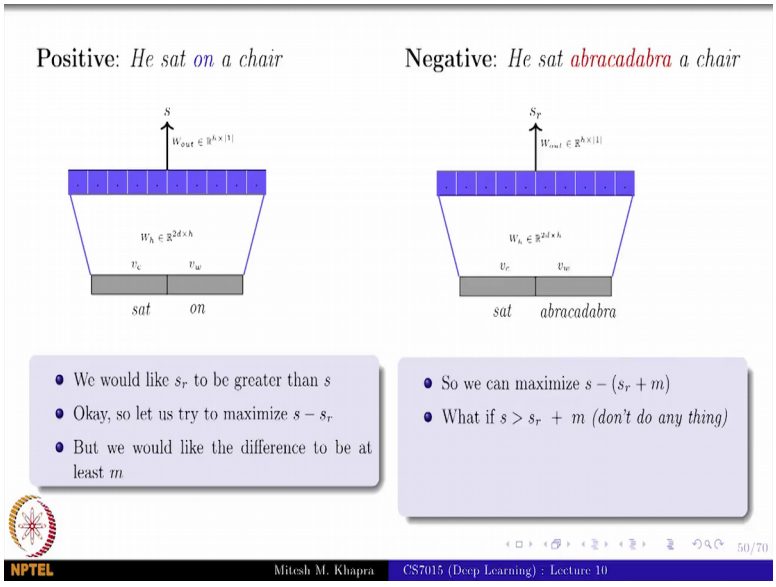


So, there should at least some gap between that and that gap is m. So, instead of maximizing S minus Sr, I am going to maximize S minus Sr plus m, is that fine? Now suppose you are at some point of training, I will have some need some parameter configuration; that means, you have learned some values for Vc and Vw. And you do this forward propagation compute S and Sr. And we actually find that this condition holds right. So, right now my loss function is this at some point you are doing this, and you observe that this condition holds; that means, S is actually greater than Sr plus m. In that case, what do you want a loss to be? How many of you get the question?

I want that S and Sr should be separated from a margin of m, in the favor of S. I am doing my training I am at certain configuration for Uc S and Vw S and so on. I pass it through the feed forward network and I get S and Sr. And I observe that this condition already holds.

Is my network doing anything wrong at this point? It is doing it is job properly? What should be the loss that I back propagate? 0, again gets that there is nothing to correct here, I do not need to back propagate any loss.

(Refer Slide Time: 06:14)



So, then can you give me the full objective function? Maximize this, but at this condition already holds then do not do anything is that fine? So, that is about this so and again observe that we have gotten rid of the expensive softmax computation.