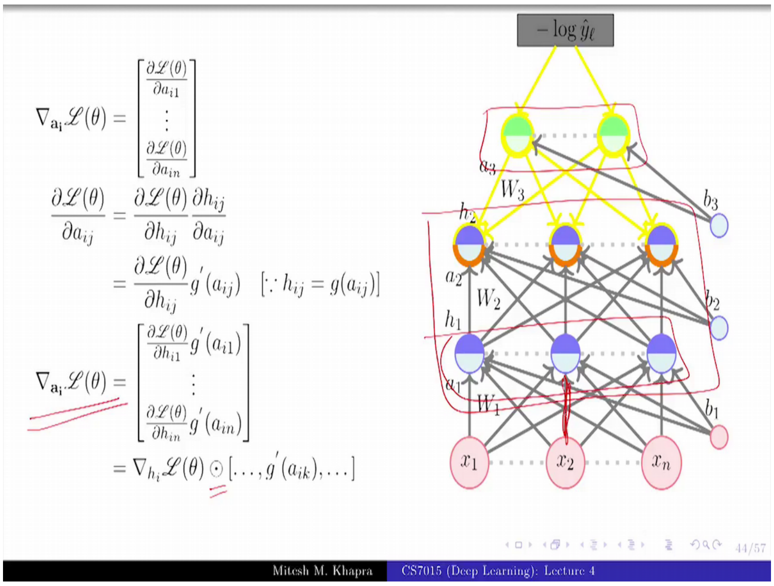
Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 4.7
Lecture - 04
Back Propagation: Computing Gradients w.r.t. Parameters

(Refer Slide Time: 00:16)



$$\nabla_{\mathbf{a_i}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{i1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{in}} \end{bmatrix}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial a_{ij}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial h_{ij}} g'(a_{ij}) \quad [\because h_{ij} = g(a_{ij})]$$

$$\nabla_{\mathbf{a_i}}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial h_{i1}} g'(a_{i1}) \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial h_{in}} g'(a_{in}) \end{bmatrix}$$

$$= \nabla_{h_i}\mathscr{L}(\theta) \odot [\ldots, g'(a_{ik}), \ldots]$$

44/57

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

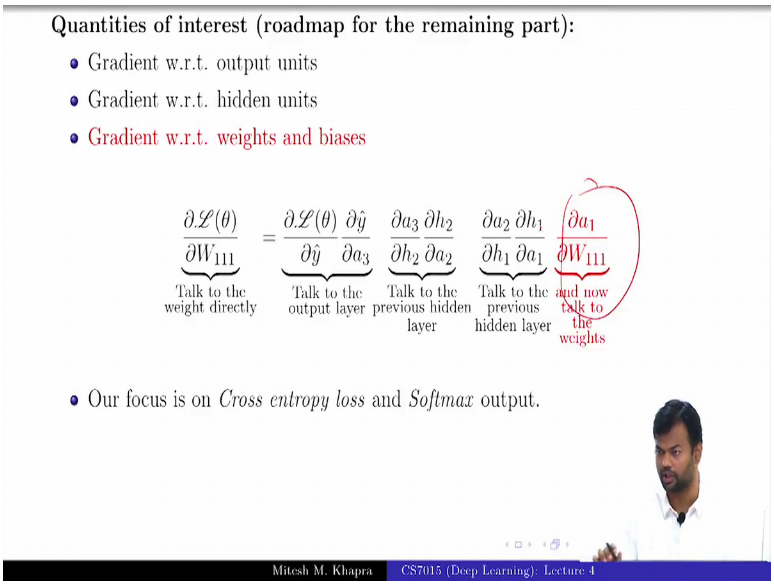
Before we move on to the next moduler, a quick summary of what we have done so far. So, we introduced feed forward neural networks, and we wanted to learn the parameters right from the last layer to the first layer. And we figured out that what we can do is that we can just use the gradient descent algorithm as it is; except that, we have this small problem that we have so many parameters now, and located at differ different points in the network, right some at the initial layer some at the final year, and you want to compute the derivatives or the partial derivatives with respect to all of these.

If you can do that, put them all in this large matrix, then we can just use gradient descent as it is wait so, that is what we figured out. And then we wanted to find out the gradients with respect to or the partial derivatives with respect to all these parameters. So, then we realize that this can be done using chain rule, because there is a path from your output which is the loss function to any of these weights. So we just need to follow that path and apply this smart this chain rule smartly and just some of the derivatives across all the paths that lead to that weight, ok. So, in that process we started from the output layer, we just treated it a bit special, because the output function is special and this is the last layer.

So, we just first computed the gradient with respect to the output layers, then we figured out how to compute the gradients with respect to any of the hidden layers. And now if you are at a particular hidden layer, now the weights that feed into this layer we could or we have not reached there right, oh sorry.

So now, the next thing that we need to do is that we have computed the gradients with respect to any of these hidden layers, and now we want to find the gradients with respect to the parameters which is the weights and the biases. So, it is the do you all remember this, or it is all long history or the story is back right, fine. So now, we are at the last point which is computing gradients with respect to parameters, clear?

(Refer Slide Time: 02:04)



Quantities of interest (roadmap for the remaining part):

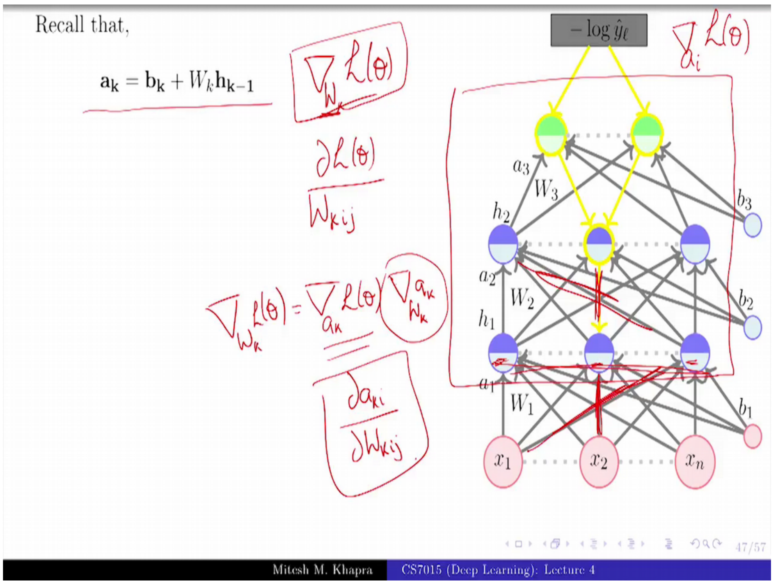
- Gradient w.r.t. output units
- Gradient w.r.t. hidden units
- Gradient w.r.t. weights and biases

$$\underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}}}_{\text{Talk to the weight directly}} = \underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\text{Talk to the output layer}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\text{and now talk to the weights}}$$

- Our focus is on *Cross entropy loss* and *Softmax* output.

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

(Refer Slide Time: 02:14)



Recall that,

$a_k = b_k + W_k h_{k-1}$

$\nabla_{W_k} \mathscr{L}(\theta)$

$\frac{\partial \mathscr{L}(\theta)}{W_{kij}}$

$\nabla_{W_k} \mathscr{L}(\theta) = \nabla_{a_k} \mathscr{L}(\theta) \nabla_{W_k} a_k$

$\frac{\partial a_{ki}}{\partial W_{kij}}$

$\nabla_{a_i} \mathscr{L}(\theta)$

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

47/57

So, again this is the overall picture, we were in this chain rule, and we have come all the way to the last point where we are ready to now compute these quantities. Ok this should ok, ok. So now start by recalling that a k is equal to b k plus w k h k minus 1, right, this is our activation formula, pre activation formula right. So, I am talking about these light blue guys, ok, which is clear in image.

And now I what have I done so far? I have been able to come up with a formula to write the gradient of the loss function with respect to any of these light green guys, right, that is what where we ended last time right, where we are able to compute the gradients with respect to the, sorry light blue guys ok,. And now I want to compute the gradient with respect to any of these parameters or any of these parameters, right.

So, any parameter, it does not matter am at some ith activation layer, pre activation layer, and I just want to compute the gradients with respect to the weights which feed into this layer, ok, and that is what we are interested in. So we are just taking any layer k, and you want to find the gradient with respect to the weights there, ok, now can you tell me? So, can you tell me what is what is the thing that am going to do here? Or what is the recipe that we have been following?

I need to move, what is the recipe that we have been following? Apart from yelling at people who come late, we find the element wise partial derivatives first and then put them all together to get the gradient ok, what is the element here? What is what am I looking for right now? I want to compute this fill this blank, what goes here?

Student: W.

W any of these W is right? And in particular say W k that is what I am looking for. So what is the first thing that am going to attack?

Student: Wkij.

Good. W k i j and once I have this for one of these guys i just know a generic formula with respect to i j and k and I can just put it into a gradient vector ok, is that fine? Ok so now can you, ok, now from here to here if I want to reach from here to here. So, this is what i am interested in, right? Now how is the chain rule going to look on, look like based on whatever you have already seen? Till where have you already reached? You already know this quantity, right? Now if I want this how am I going to write it?

Student: (Refer Time: 04:59).

I will find up to the light blue guys; which is this I already know how to compute it, and then from the light blue guys I will go to the this; is fine right? So, this is the quantity that i am looking for, ok. Now what is one element of this guy? Dou a k by; is it fine? Ok what is the dimension of this actually? Is it a scalar, a vector, a matrix, matrix or a tensor, what is the tensor? What is it? Is it a matrix? What are the dimensions? What does this derivative mean? Or this gradient mean? I change one element of W k how much does one element of a k change? How many elements are there in ak? n, how many elements are there in W k? n cross n. So, how many partial derivatives which I have? n cross n cross n, what is this?

Student: Tensor.

A tensor, right? So, this is going to be a tensor, ok. So, when I say one element of this, I mean this ok. So, this is one element of this gradient, ok. Now can you tell me the formula for this? What is this quantity? Hk minus.

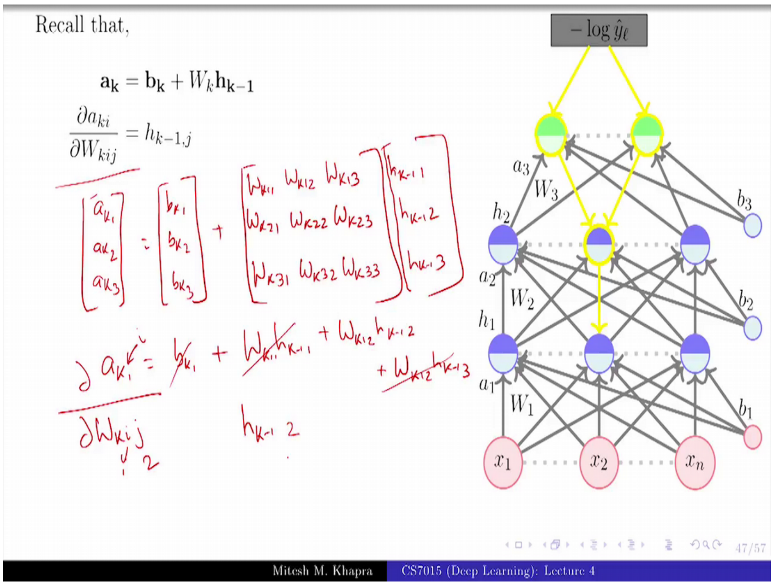
Student: 1 (Refer Time: 06:27).

Hk minus 1 or hk minus 1 j or?

Student: (Refer Time: 06:31).

Everyone gets this hk minus 1i. How many of you get this? Ok.

(Refer Slide Time: 06:38)



Recall that,

$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

$$\begin{bmatrix} a_{k1} \\ a_{k2} \\ a_{k3} \end{bmatrix} = \begin{bmatrix} b_{k1} \\ b_{k2} \\ b_{k3} \end{bmatrix} + \begin{bmatrix} W_{k11} & W_{k12} & W_{k13} \\ W_{k21} & W_{k22} & W_{k23} \\ W_{k31} & W_{k32} & W_{k33} \end{bmatrix} \begin{bmatrix} h_{k-1\,1} \\ h_{k-1\,2} \\ h_{k-1\,3} \end{bmatrix}$$

$$\frac{\partial \overset{i}{a_{k1}}}{\partial W_{ki\underset{2}{j}}} = \cancel{b_{k1}} + \cancel{W_{k11} h_{k-1\,1}} + W_{k12} h_{k-1\,2} + \cancel{W_{k12} h_{k-1\,3}}$$

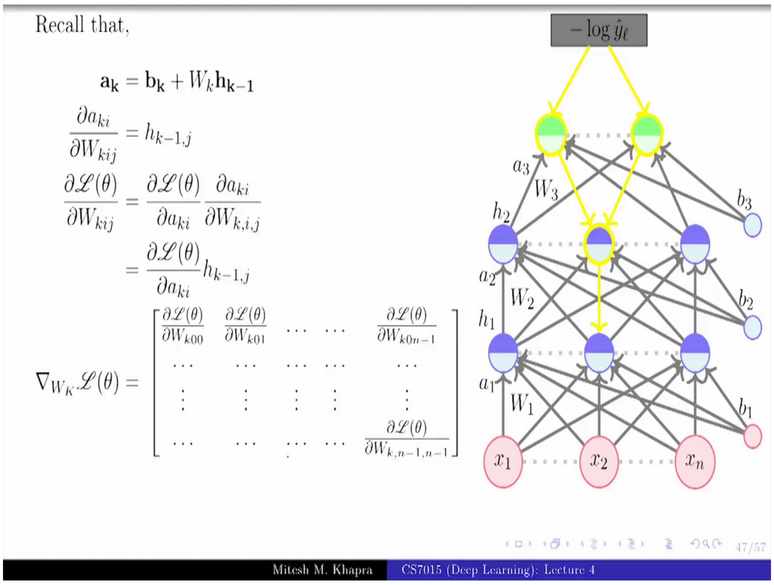
$$h_{k-1\,2}$$

So, let us do it, right? So you have ak1, ak2, ak3 that is your ak vector, ok. You have bk1, bk2, bk3 plus wk1 1, yeah, I know again this is one of those silly things, but if everyone does not raise their hands and compelled to do this; so, h k minus 1 1 hk minus 1 2 hk minus 1 3 ok. So, let us take one of these guys right. So, a k 1 can you tell me the formula for that?

Student: (Refer Time: 07:30).

Plus, first row ok, 1 2 this 1 3, now can you tell me this quantity? So, what is i here? 1 so, i want this by w k i j right so, i is 1 so, i can take any of the j. So, let me take j equal to 2. So, what is it going to be? This will go off this is constant, this is constant only this term remains, and the derivative is hk minus 1 2 which is j right. So, that is what the formula says. So, I have a formula for one of these guys, and that is a generic formula. So, always remember if you cannot figure out what it is just write it down in scalar terms, just add up all the terms and you will get the formula, right. So now this is what the chain rule is going to be? Is this fine? Ok.

(Refer Slide Time: 08:37)



Recall that,

$$\mathbf{a_k} = \mathbf{b_k} + W_k \mathbf{h_{k-1}}$$

$$\frac{\partial a_{ki}}{\partial W_{kij}} = h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} h_{k-1,j}$$

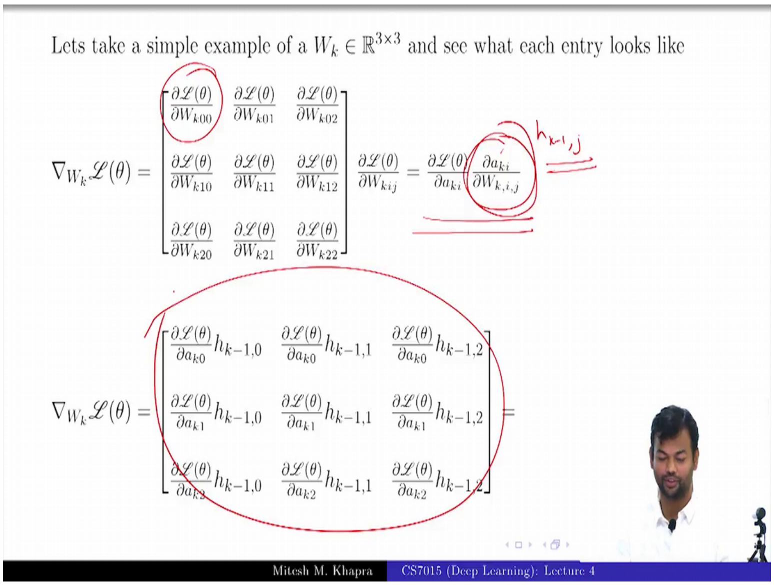
$$\nabla_{W_K} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \dots & \dots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k0n-1}} \\ \dots & \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \dots & \dots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k,n-1,n-1}} \end{bmatrix}$$

47/57

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

So, this is what it is going to be. This is one element of that tensor ok, is this fine? This is how that entire thing is going to look, I have just flattened it out and put it here.

(Refer Slide Time: 09:05)



Lets take a simple example of a $W_k \in \mathbb{R}^{3\times3}$ and see what each entry looks like

$$\nabla_{W_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k02}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k10}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k20}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} \end{bmatrix} \quad \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

$h_{k-1,j}$

$$\nabla_{W_k} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} \end{bmatrix} =$$

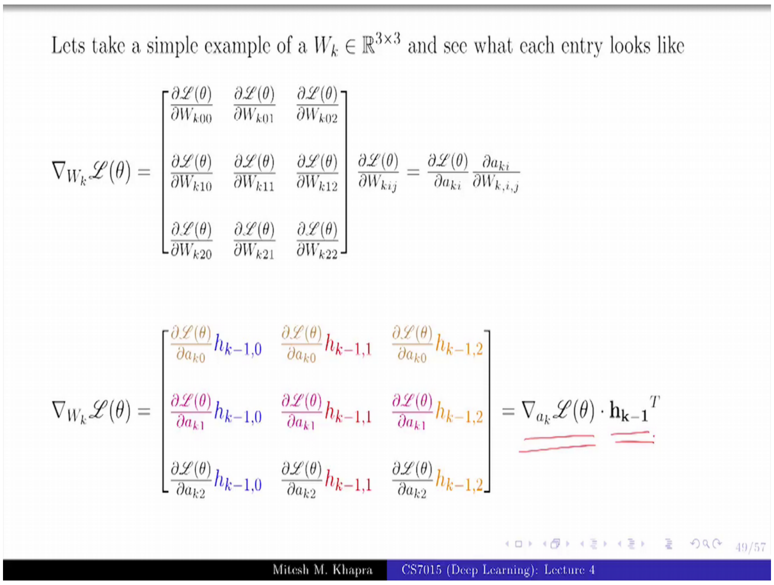
Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

Now let us take a Simple example of wk belonging to r cross ah; 3 cross 3 everyone is fine so far right or anyone who everyone is fine please raise your hands. i mean fine i mean not in life, but with the lecture fine , ok. So, this is what it looks like right for a 3 cross 3 matrix, fine.

Now, let us see we already found out that this guy is equal to hk minus 1 comma j right. So, this is what this matrix looks like, nothing rocket science here right. So, each of these quantities is actually can be written in this form; where i appropriately substitute I k and j. And I know that this quantity can be further written as this quantity, right? That this is our clear rate so, I have written it as this.

Now can you simplify this, I do use a lot of this ok, can you simplify it? Is it looks similar to something that you did on the assignment. Does this look like matrix which has some very regular patterns? Yeah I can see someone doing this, and this everyone gets it, ok.

(Refer Slide Time: 10:28)



Lets take a simple example of a $W_k \in \mathbb{R}^{3\times 3}$ and see what each entry looks like

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{k00}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k01}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k02}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k10}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k11}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k12}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{k20}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k21}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{k22}} \end{bmatrix} \quad \frac{\partial \mathscr{L}(\theta)}{\partial W_{kij}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial W_{k,i,j}}$$

$$\nabla_{W_k}\mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k0}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k1}} h_{k-1,2} \\ \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,0} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,1} & \frac{\partial \mathscr{L}(\theta)}{\partial a_{k2}} h_{k-1,2} \end{bmatrix} = \underline{\nabla_{a_k}\mathscr{L}(\theta)} \cdot \underline{\mathbf{h_{k-1}}}^T$$

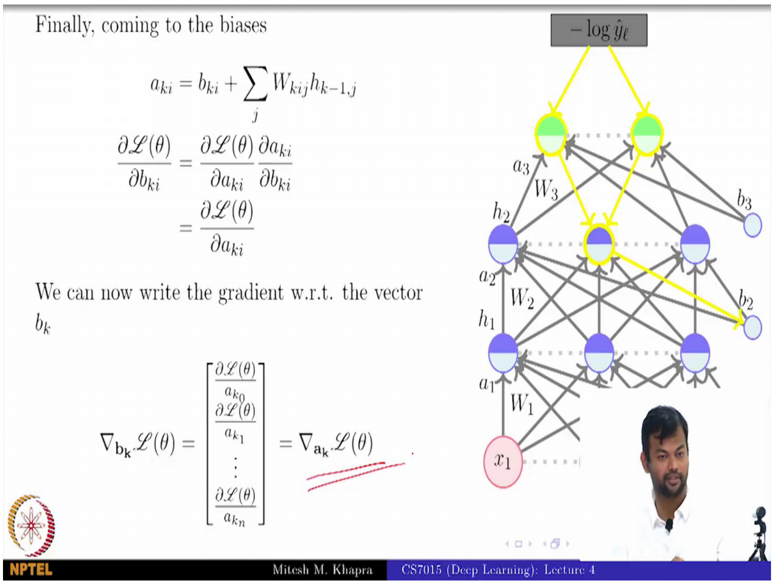
49/57

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

So let us see so this the first column, the second term in the product is all same throughout all the rows, right? What I mean is all these guys are similar, same thing happens in the second row the third row, right ah? That is sorry the second column and the third column. What about the rows? These are all equal right, so what does this look like actually? The outer product of 2 vectors, everyone gets this? Raise your hands? Ok good.

So, I do not need to do an example. So, it is fine right this is an outer product of these 2 vectors, one happens to the quantity to be the quantity that we already knew, right? And the other happens to be a quantity that we can figure out. I mean we already know this, what is we know how to compute the hidden representations, right? The edge case we can compute ok.

(Refer Slide Time: 11:20)



Finally, coming to the biases

$$a_{ki} = b_{ki} + \sum_j W_{kij} h_{k-1,j}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_{ki}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial b_{ki}}$$

$$= \frac{\partial \mathscr{L}(\theta)}{\partial a_{ki}}$$

We can now write the gradient w.r.t. the vector b_k

$$\nabla_{\mathbf{b_k}} \mathscr{L}(\theta) = \begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{a_{k_0}} \\ \frac{\partial \mathscr{L}(\theta)}{a_{k_1}} \\ \vdots \\ \frac{\partial \mathscr{L}(\theta)}{a_{k_n}} \end{bmatrix} = \nabla_{\mathbf{a_k}} \mathscr{L}(\theta)$$

NPTEL Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

So, fine so, finally, we come to the biases. This is what one entry looks like, this is exactly the sum which I had written out, right. Now I take the derivative with respect to b k i of the loss function. So, I could write it into as this chain rule; where the first quantity is something I already know I have computed the gradient with respect to the pre activation layers what about the second quantity? Anonymous roar is what I was expecting.

Student: 1.

1 ok, fine, we can now write the gradient with respect to the bias, what would it be? What is this? What is this? It is just the gradient with respect to the pre activation layer, right simple, fine. So now, we are done with all the gradients that we were interested in, right?