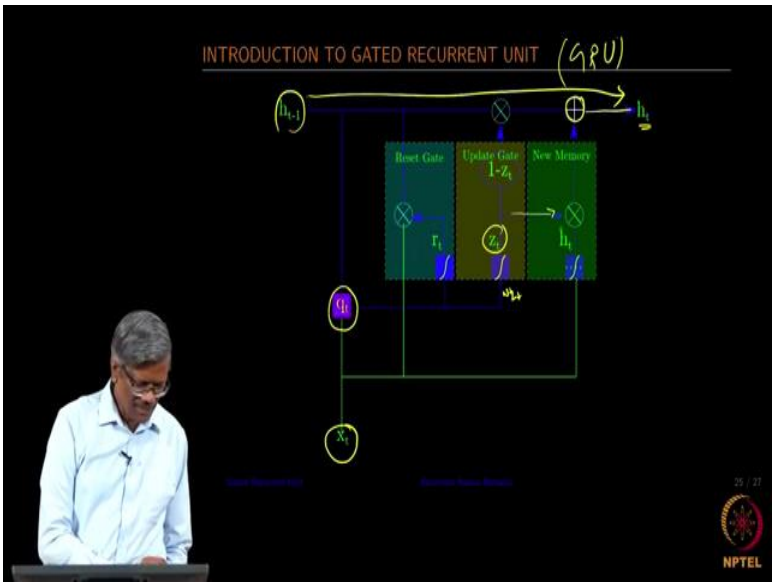
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture - 62
GRU

(Refer Slide Time: 00:15)



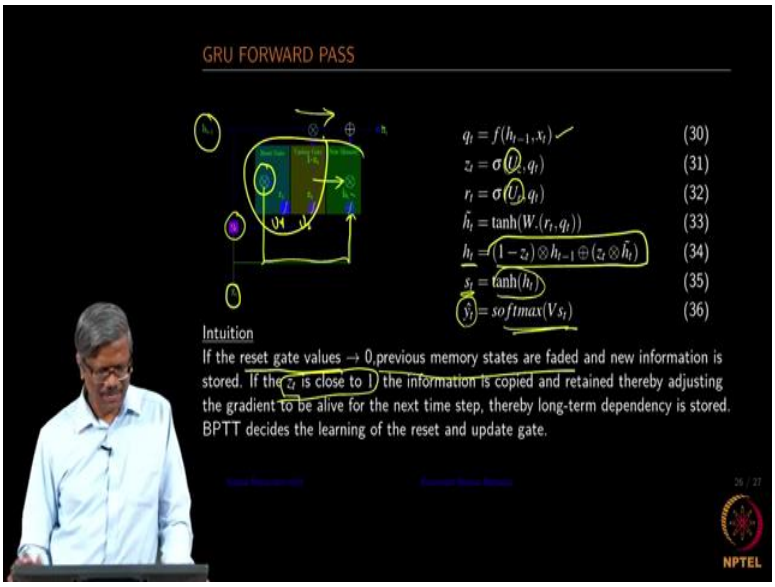
Let us look at another interesting variation to the recurrent neural network, this is called a gated recurrent neural network or gated recurrent unit, we usually call it as a GRU ok. So, there is a slight variation in what we saw earlier with respect to LSDM. Here we have only about two states rather are two gates one is the reset gate, the second one is the update gate and then you would also notice that there is no memory cell that is passing through this ok. It is making use of the same h t that we had computed in the normal RNN or the vanilla or the RNN with a slight change in terms of or how we want to allow the values to be passed on to the next state.

So, in this case, we have a line passing through that becomes h there we have an input then we combine the h t 1 and x there and then call it as q t and then q t is connected to a sigmoid function and then we have our reset gate. So, in this case the reset value is used by other cases. So, we have the reset gate and then we have the update gate that connects the q t through a sigmoidal function rather we have the update gate that connects q t and the weight here which you can call it as w z t, it is a sigmoid function and then what it

does here is, it takes the value that is computed here using the sigmoid and it is 1 minus of that we will come to that later ok.

And then we have another new memory that is computed which takes the value from z t, and then it takes the value from the reset gate and finally, does an element-wise addition to creating an h t there right. So, in this case it looks a little simpler but again this is not a very simple implementation.

(Refer Slide Time: 02:58)



Let look at the forward pass equation, I am not going to be going into the details of how the backward propagation is computed it is very similar to what we had seen in the recurrent neural network. The only thing is we need to worry about the backpropagation in the gates as well ok. So, in this case, as I mentioned $q_t = f(h_{t-1}, x_1)$ the that is that is how we will keep it to make it simple and then we compute the update gate using a sigmoid function and there is a weight that connects this since we are using the hidden layer the weights are of U right.

So, we have a new z and then we compute z t and then we compute the reset gate using q t right and the weights that connect this, this is your U t or rather U r, this is your U z or U z and the new memory is computed using h there which makes use of the reset gate values and the values coming in from. So, it comes from here ok. So, the connections are little daunting let me try to have another simple diagram at the end of this, and then we

compute the new value or the new state using this right I think it is very clear in the diagram how we are doing this. Is explained through below equation

$$h_1 = \tanh\ (w.(r_1, q_1)$$

$$y_t = \ softmax\ (V_{st}\ )$$

And then the new state of this is computed using the hyperbolic tangent function and then the output or rather the predicted value is computed using the softmax it is done in a very similar fashion as we had done earlier ok. So, what are the institution that you get in this right? If you reset if the reset gate values or you know tend to 0 the previous memories states are faded ok.

So, the values are not going to be 0 if for every element right in this case, there could be some values because we are taking the σ value. So, it is going to be between 0 and 1 and there could be some intermediate values. So, when you do the element-wise multiplication the values would become smaller or it would get boosted up ok.

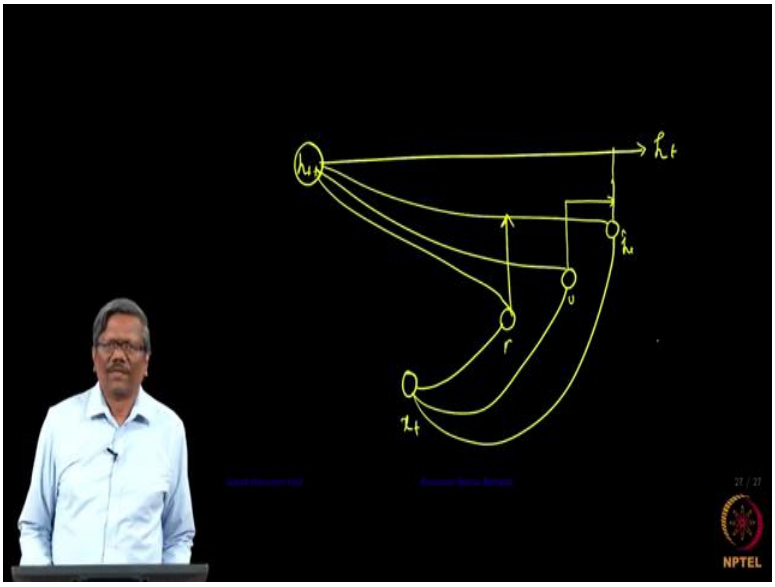
So, that is what happens in the reset gate. And if the z_t or the if the z_t value is close to 1 the information is copied and we are going to be keeping that right and then that what happens when the values are very close to 0 the h_t is going to be maintaining that value. So, we are going to be taking the value that is coming from the reset value and then element-wise multiply the z_t and the new memory,s and finally, is it with the h_{t1} there is coming along in this direction using an element-wise operation to get this right.

So, when you do this operation when is that is close to 1, the gradients are kept alive; that means, I am going to be there for the next operations as well you know in the forward pass.

So, by doing does what happens is the values that are found in the long distance from the target or kept alive so, that when you do the backpropagation you still managed to find that. So, that you are able to really do the gradient design properly and then the gradients are behaving properly because we are taking care of the derivatives right during the forward pass it's using these gates. So, this is the intuition that I have with respect to the gate that we find in the GRU. So, the gates are actually meant for managing the values of your chain that goes across from state 1 to state or t.

So, the intuition here is when you do the backpropagation, the errors that you find between the cold and the predicted values are propagated back and those values really condition the matrices that it comes across. Since the gates already have conditioned the values in such a way that there is going to be a gradient alive when you backpropagate we will not get into the problem of vanishing gradient. So, these gates are extremely important in terms of really keeping the gradients alive during the long term backpropagation ok.

(Refer Slide Time: 09:11)



So, if this looks daunting, let me try to draw this little bit better. I think I found this from one presentation somewhere and then we have the input correct and then we have gates let us say this is the recent gate, this is the update gate and then this is your new memory correct ok.

So, x t connects to all of these rights and then h t connects to the reset gate, h t connects to u, h t connects to the new memory right and then there is a value that comes here right. So, this reset gate connects to the new memory and then you connect in this fashion and so on. So, this is a very simple representation of what we saw in the earlier diagram ok. So, with this I conclude the lecture series on RNN, LSDM, and GRU.