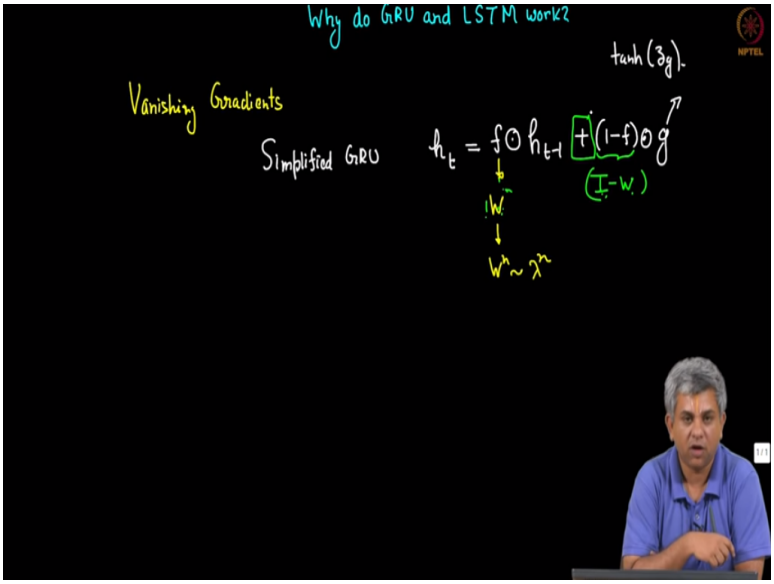
Machine learning for Engineering and Science Application
Professor Balaji srinivasan
Department of Mechanical Engineering
Indian Institute of Technology Madras
Why LSTM Works

(Refer Slide Time: 00:13)



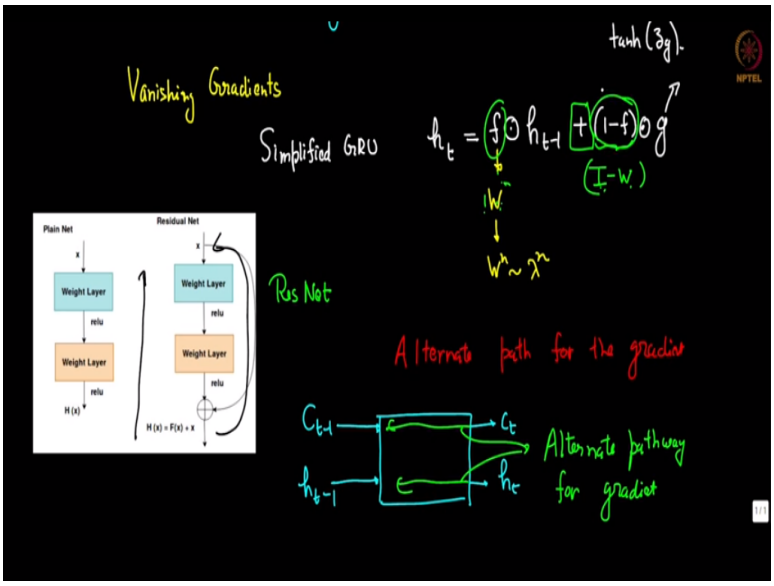
Welcome back in the last video you saw LSTM and it is architecture we also saw in the previous video simplified GRU and JRU, now remember are all these were meant to handle the vanishing gradients issue, now the question is why is it that GRU and LSTM work, so in this video we will give you a very-very short and very heuristic explanation the mathematics of this as far as I understand has not yet been totally worked out this, so this is basically guess work initial guess behind LSTM so it was base more on cognition rather than any direct mathematical reason but I will try and give you a short heuristic of this work.

So remember that when we had simplified GRU our expression was HT is some non-linearity let me use on alternate expression it was actually H times HT minus one plus one minus F times G where G was tan H of ZG, now these was the expression that we used now how does this help the vanishing gradient issue, now remember why was it that the gradient was vanishing in there first place remember you can sort of think of this as of this is of wait matrix and if this wait matrix is multiplies itself multiple times through multiple layers there eigenvalue when it raise to the power N and if it is less than one it can actually go to zero that was the basic problem, when

this goes to WN it went like lambda power N as I explain in the gradients video now how does this term help notice that when this is W this vector or this matrix can be approximated as if it is the identity matrix minus W, if this goes as W again remember all this is very heuristic, if those goes as W that becomes I minus W.

So if this number is small this becomes correspondingly large, if this is point zero one that becomes point nine-nine, so in some sense this term and that term balance out more importantly this plus is what makes things work why it is plus makes thing work because you can now visualize this.

(Refer Slide Time: 02:55)



As if you might recall this from Dr. Ganapati video this is nothing but the architecture of Res not and there whether it was Res not or whether it was Alex net and several other cases you actually saw that there is an alternate path way for the gradient that is when you are doing back prop it can either go directly through this or it can go through this.

So similarly when you offer remember when we were doing LSTM, we had one pathway through HT minus one we had another pathway and this was the reason why we draw the figure through CT, so this alternate pathway for the gradient actually helps you again this is a heuristic explanation whenever you actually provided alternate pathway especially jumps from the end to the beginning if you actually jump a few layers one way or the other or you provide different

paths as so it is provided in Alex net through different GPU's when you do that it typically sort of mitigate gradient problems.

So this is a general theme that you will see across this course, so this is a good lesson you learn you know sort of a heuristic lesson to learn whenever you have training problems try and provides alternate pathway try and provide some skips connections try and provide some different way to actually trained and that is really what as we understand it what happens even within simplified JRU or within LSTM because of alternate raise or mathematically because this F sort of balance is out the one minus F or the I gate in terms of LSTM it actually give you different ways to train the gradient, the gradient goes longer before vanishing thank you.