Lecture – 82
Variants of Gradient Descend

(Refer Slide Time: 00:16)

Another one that you would have seen in the neural net models, they would be using stochastic gradient descent model or batch model, or mini-batch model, right. They are not very difficult to understand, very simple things; a stochastic model is something like you know if you have a pair of sentences.

Let us say this is the encoder and this is your decoder, your input, and then you output certain sentences, certain words right. In this stochastic gradient model and you present one sentence x_1, y_1; once it is presented you start the learning process ok. You, so you, capture $J\theta$, $J_1\theta$, $J_2\theta$ and so on.

And then combine this to minimize your error function. The error is minimized during the BPTT and then once it reaches a certain stage, you stop the training and then input x_2, y_2. And then you again capture the error values for x_2, y_2 do the BPTT and keep doing it until it reaches an equilibrium state. And then produce, and then provide x_3, y_3 were and do the training process and so on.

So, it is about for every pair you minimize the error function ok. So, this is the stochastic gradient model, which is very expensive right. So, if you have about 15 million sentences, you keep doing this every time for every pair of words. What is the batch? Batch is very similar to this, but only one change; instead of training for every pair you take the entire batch right, feed this en. You feed this once, once it is done feed the second one, feed the third one and then all of those.

Suppose if you have about x m y m you keep doing it until you finish the last pair and then start the training process ok. You just keep feeding one after the other, and then finally start doing the training process, once you finished feeding the x m, y m. So, this is the batch process. And there is a mini-batch. So, in mini-batch what we do is. So, let me rewrite this differently x_1 , y_1.

So, suppose if you have about let us say 1000 pairs of sentences. So, in batch what you do is you feed all the 1000 pairs one after the other and then after a thousand pairs are fed start the training process to minimize the error condition right. So, in this case what happens is, instead of feeding all 1000 at a time, you just feed let us say 10 if your beam size is 100.

You feed 100 at a time and then do the gradient descent algorithm to train, to adjust the weights. So, for every 100 pairs you do the adjustment of weights and then you keep going. This is efficient, this particular model and you will find many of the models use mini-batch for training. So, these are three different types you will find in various research papers as well. So, with this, I conclude all important aspects of machine learning, especially in the translation side of machine learning.