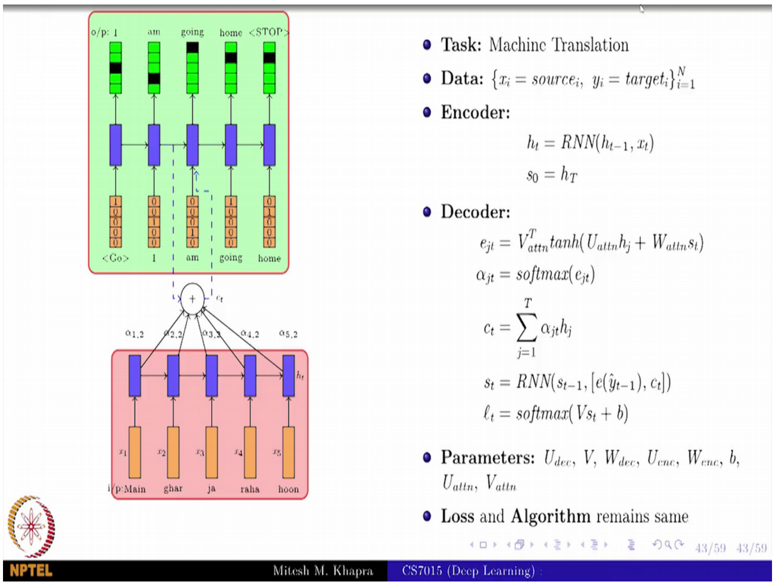
Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 115
Attention Mechanism

(Refer Slide Time: 00:11)



- **Task:** Machine Translation
- **Data:** $\{x_i = source_i,\ y_i = target_i\}_{i=1}^{N}$
- **Encoder:**

$$h_t = RNN(h_{t-1}, x_t)$$
$$s_0 = h_T$$

- **Decoder:**

$$e_{jt} = V_{attn}^{T} tanh(U_{attn}h_j + W_{attn}s_t)$$
$$\alpha_{jt} = softmax(e_{jt})$$
$$c_t = \sum_{j=1}^{T} \alpha_{jt}h_j$$
$$s_t = RNN(s_{t-1}, [e(\hat{y}_{t-1}), c_t])$$
$$\ell_t = softmax(Vs_t + b)$$

- **Parameters:** U_{dec}, V, W_{dec}, U_{enc}, W_{enc}, b, U_{attn}, V_{attn}
- **Loss** and **Algorithm** remains same

43/59 43/59

Mitesh M. Khapra CS7015 (Deep Learning) :

So, let us start. Last lecture we are looking at encoder decoder models and we saw that a bunch of problems from different domains and different modalities images, text, videos and so on. And, even this cross modal or multi modal applications the where they are taking a video and trying to describe it; so video is one modality, description texts is another modality and so on.

We were able to propose modals for all of these using this encoder – decoder architecture. And, then we motivated this attention mechanism where we said that encoder decoder is trying to do this silly thing where it tries to encode the entire input once and that is what how humans do it. He do this back and forth thing where at every time step if we are trying to produce a translation or a single word in the translation we just focus on certain words in the input sentence and kind of ignore the other.

So, the attention mechanism which is this bunch of equations that you see here that allowed you a neural way of modelling attention and the key thing to note here is a there was a supervision for the attention. No one actually tells us that this is the portion of the text which is important at time step t, but they still works better because this is the better modelling choice and I give you that bicycle analogy and also it is a better modelling choice we are able to no one has given you these supervisions but, you are still have more parameters in the model to learn this kind of a behaviour.

(Refer Slide Time: 01:37)

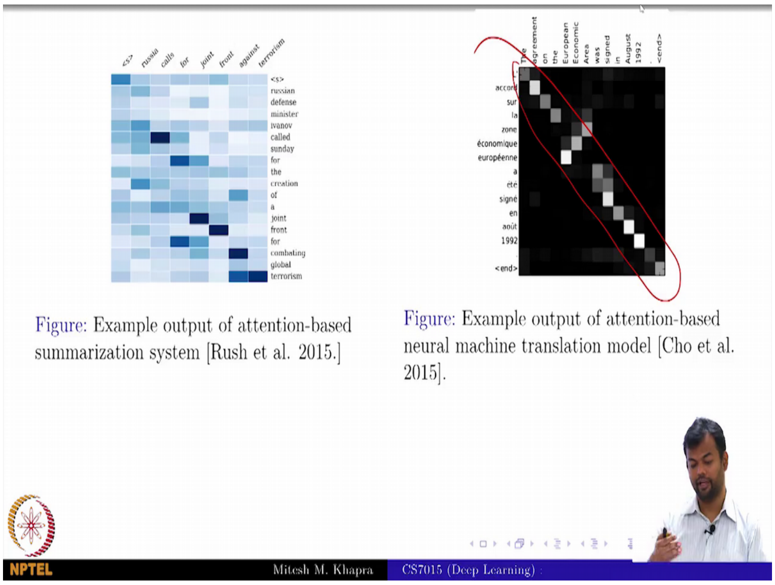


Figure: Example output of attention-based summarization system [Rush et al. 2015.]

Figure: Example output of attention-based neural machine translation model [Cho et al. 2015].

NPTEL Mitesh M. Khapra CS7015 (Deep Learning)

And, then we also saw that we could actually visualise, these attention based and from some experiments on some papers we saw that actually learn some meaningful attentions. In the particular case, on the figure on the on the right hand side: so, the one that clearly shows that for a monotonic kind of a translation scenario between English and French. Most of the attentions based are along a diagram and that is exactly what you would expect, right.

So, that is where we end it.