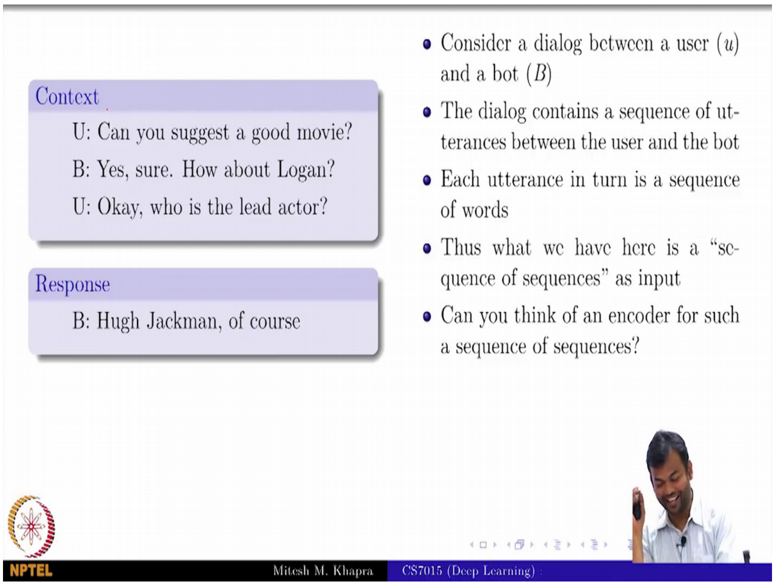
Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 117
Hierarchical Attention

So, we will go on to the next one which is Hierarchical Attention. So, again something very popular in nowadays become very common for various things, so again not very difficult idea to understand.

(Refer Slide Time: 00:24)



Context
U: Can you suggest a good movie?
B: Yes, sure. How about Logan?
U: Okay, who is the lead actor?

Response
B: Hugh Jackman, of course

- Consider a dialog between a user (u) and a bot (B)
- The dialog contains a sequence of utterances between the user and the bot
- Each utterance in turn is a sequence of words
- Thus what we have here is a "sequence of sequences" as input
- Can you think of an encoder for such a sequence of sequences?

So, let us first look at the motivation for this and then we will look at the solution. So, consider a dialog right. Today everyone is interested building chatbots; every second start up wants to build their own chatbot. And every second startup out of that they wants to build for the agriculture domain, or the banking domain, or the healthcare domain. So, here is what a typical dialog looks like right? This is of course, not for any profound purpose this is, but you can see this is an important dialog right very relevant and very important. So, this is what a dialog looks like.

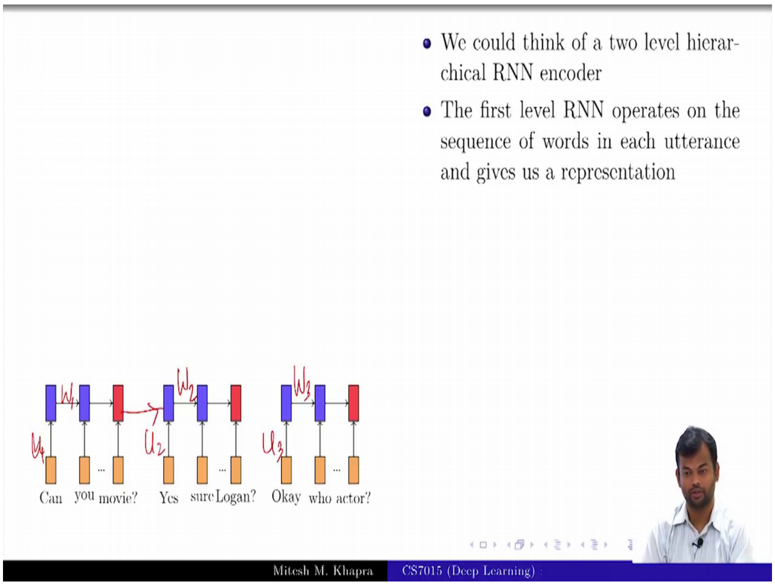
So, let us try to break it down into the kind of entities that we deal with. So, can you tell me about a dialog, what is a dialog? It is a dash; think in terms of things that we have discussed so far.

Student: (Refer Time: 01:07).

Sequence, good right. Again the safest answer is sequence from now on. No, it is only for one lecture. It is a; is it a just a sequence or sequence of sequences right?

So, it contains a sequence of utterances, so each of these lines here is an utterance and each utterance in turn is a sequence of words right ok. So, what we have here is a sequence of sequence as input and this is very common in many applications right. So, can you think of an encoder for such a sequence of sequence? RNN of RNN's good that is the answer right.

(Refer Slide Time: 01:40)



- We could think of a two level hierarchical RNN encoder
- The first level RNN operates on the sequence of words in each utterance and gives us a representation

So, we think of a two level hierarchical RNN encoder. So, first leveller will encode the utterances ok. Let me ask you few questions. Is there is a mistake in a diagram? Should this be connected? Yes, no, maybe, do not care ok. Second question; I will write some parameters here right, what is our notation; w, u. This is u, right and this is w right. Is it fine? If I have a dialog which contains 100 utterances, what will happen? That is a practical problem.

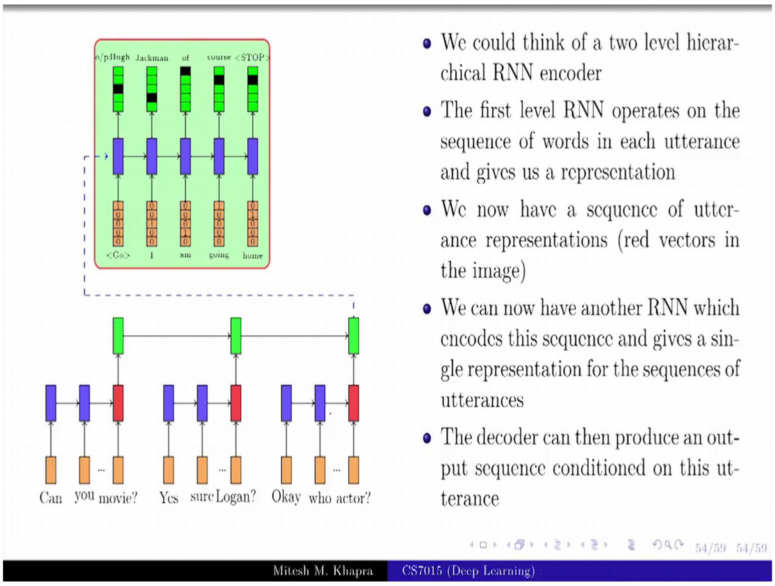
But more conceptually what is wrong here? What is each RNN trying to do? I encode a sentence, encode an utterance. So, why should it be different for the first utterance, second utterance, third utterance and so on right. All these RNN's should be the same; does that make sense? The U 1 is equal to U 2 is equal to U 3 and w 1 is equal to w 2 is equal to w 3 ok. Is that fine, everyone agrees with that? So, now can you tell me if there is a correction should be there, or not?

Conceptually what is each RNN doing? Encoding?

Student: (Refer Time: 02:47).

One sentence right, then why should it be connected to the previous sentence? But then if you do not know all these sentences then how will you predict the utterance, the next response rather, what is missing here? What kind of a what was the title of this module? So, what is missing? Where is the hierarchy right?

(Refer Slide Time: 03:04)



- We could think of a two level hierarchical RNN encoder
- The first level RNN operates on the sequence of words in each utterance and gives us a representation
- We now have a sequence of utterance representations (red vectors in the image)
- We can now have another RNN which encodes this sequence and gives a single representation for the sequences of utterances
- The decoder can then produce an output sequence conditioned on this utterance

54/59 54/59

Mitesh M. Khapra CS7015 (Deep Learning) :

So, what we will have is this right. So, each of these green guys is presentation for one utterance in your input. In fact, red guy sorry, the red guys are the representations for the utterances in your dialog and then the green guy is a sequence of utterances right. So, remember you have the sequence of sequence of words.

So, the red guys are the sequence of words, and the green guys are the sequence of utterances. Does that make sense ok? How many of you get this? Please raise your hands good. So, and now what would the decoder be? You have a hierarchical encoder. What could the decoder be? What is the decoder have to do? It has to produce a everyone?

Student: Sequence of words.

Sequence of words. So, what kind of a decoder will you use? Just an RNN not a hierarchical RNN, the input is hierarchical why should not the output be hierarchical? How it would look unbalanced right? The diagram will not look very neat right if you (Refer Time: 03:57).

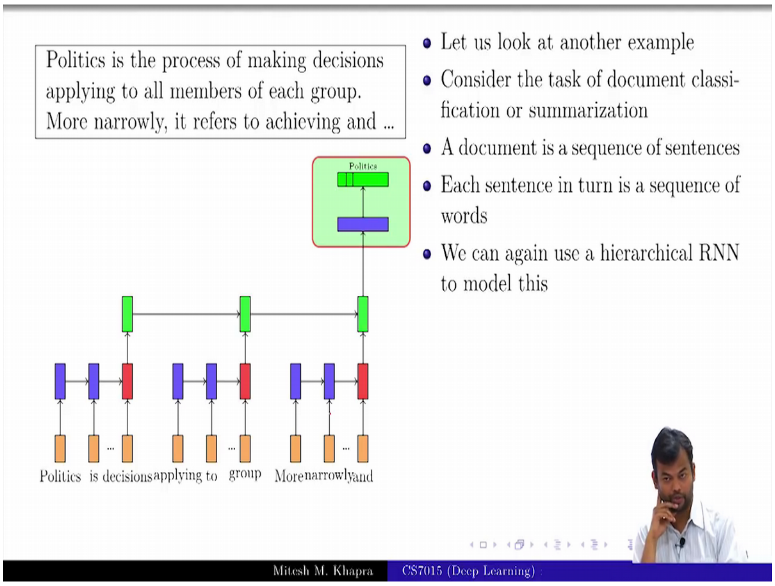
Do you need a hierarchical and decoder? No right, I mean just a simple decoder. Because the decoders has to produce a sequence of words, so it take something from the encoder, what will it take from the encoder in the normal encoder decoder paradigm not the attention based paradigm? What will be the input to the decoder?

The last dash vector, your options are state that is very safe answer here, the last state vector what will that means? Your options are orange, blue, green and red.

Student: Green.

Green; the last green vector that is what the input is going to be, right. So, this is what is going to look like. This is option 1. What is option 2? Feed it to every time step that is what option 2 is going to be ok. So, that is you have your hierarchical encoder decoder network.

(Refer Slide Time: 04:36)



Politics is the process of making decisions applying to all members of each group. More narrowly, it refers to achieving and ...

- Let us look at another example
- Consider the task of document classification or summarization
- A document is a sequence of sentences
- Each sentence in turn is a sequence of words
- We can again use a hierarchical RNN to model this

What is missing here? Utterance, ok. So let us look at another example. Consider the task of document classification, or summarisation. What is the difference between two? In classification what the what were the decoder be? Feed forward neural network with the softmax. What would be decoder be in the case of summarisation?

Students: RNN.

RNN good. What is the document? Sequence of sequence, not sequence of sequence of sequence. It can be a sequence of sequence of sequence also right. I did not think of that then it could be a sequence of sequence of sequence of sequence ok. Let us look at the not so funny case which is sequence of sequence. What is the sequence of sequence of?

It is the sequence of sentences, which in turn is a sequence of words, which in turn is a sequence of character. We will not go there; we will just keep it till words. So it is a sequence of sequence. So, again you need some kind of a hierarchical encoder here right. So, will encode each sentence, then you will treat the sentence sequence as a sequence, encode that and then you will pass it to the classifier.

Now think of this problem right. Now if we want to do document classification, how would you go about it actually? You want to classify whether this is a politics or sports or health or whatever (Refer Time: 05:57). I think in terms of attention, what would you do actually? First we will find the important words in the sentence, to find the important words you will have to read all the.

Students: (Refer Time: 06:10).

If you want to find the important words you will have to read all the sentences. So, what will you do? First find the?

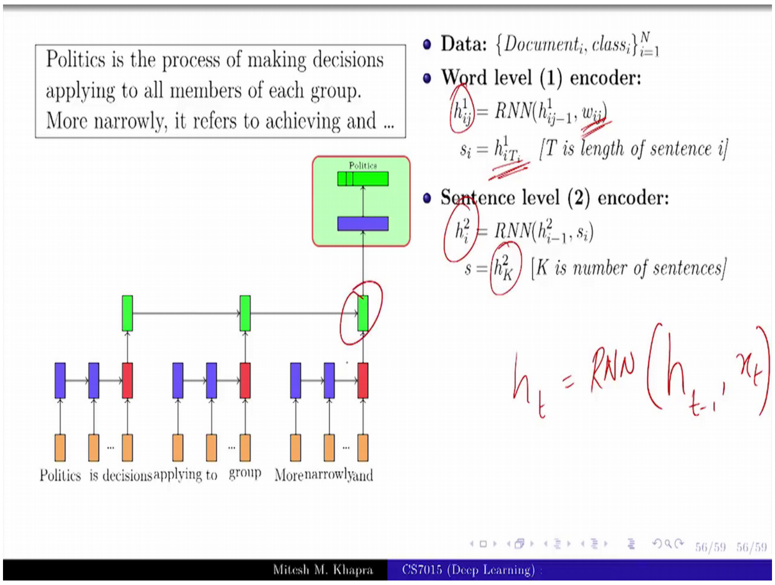
Students: Word in sentences.

Word in sentences and then? Important words within the sentence. So, what kind of an attention mechanism do you need?

Student: Hierarchy.

Hierarchical, right.

(Refer Slide Time: 06:25)



Politics is the process of making decisions applying to all members of each group. More narrowly, it refers to achieving and ...

- **Data:** $\{Document_i, class_i\}_{i=1}^{N}$
- **Word level (1) encoder:**

  $h_{ij}^{1} = RNN(h_{ij-1}^{1}, \underline{w_{ij}})$

  $s_i = \underline{h_{iT_i}^{1}}$ *[T is length of sentence i]*
- **Sentence level (2) encoder:**

  $h_i^{2} = RNN(h_{i-1}^{2}, s_i)$

  $s = h_K^{2}$ *[K is number of sentences]*

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

So, let us look at this; so first let us look at the our data model paradigm. So, what is the data given to you? I given a document and the class table for the document. And your first thing is the word level encoder which looks like this, can I will not explain what this is I will just expect you to know what this is. Why do I have two indices for h, what is i, and what is j? i is the dash id, sentence id and so it is the jth word or the ith sentence right that is what I am encoding. And how many sentences am I encoding? The number of sentences in the document ok.

And what is the second encoder? So, diagram is absolutely clear, but the equations are not; the diagram is absolutely clear right. There is no just need to map the red, blue, orange, green, guys to the equations ok. So, let me ask you this, what is h i j? No, in the diagram? Blue, red, orange?

Student: (Refer Time: 07:23).

What is w i j? How did you write the RNN equation? Time step t is equal to RNN of t minus 1 and the input at time step t right. What is the input at time step t here?

Student: Word.

Word, right. Probably this was not a good choice maybe we should make it x i j; w, I think we might get a confused with the weight’s we should not, but they are. So, fine, so w i j is actually the input word, what is h i j minus 1? Now tell me what colour is w i j? It is like an IQ test at which colour map it, w i j maps to which colour?

Students: Orange.

Orange, good and h i j?

Students: Blue.

Blue, but what about the red? That is which colour I mean sorry not which variable.

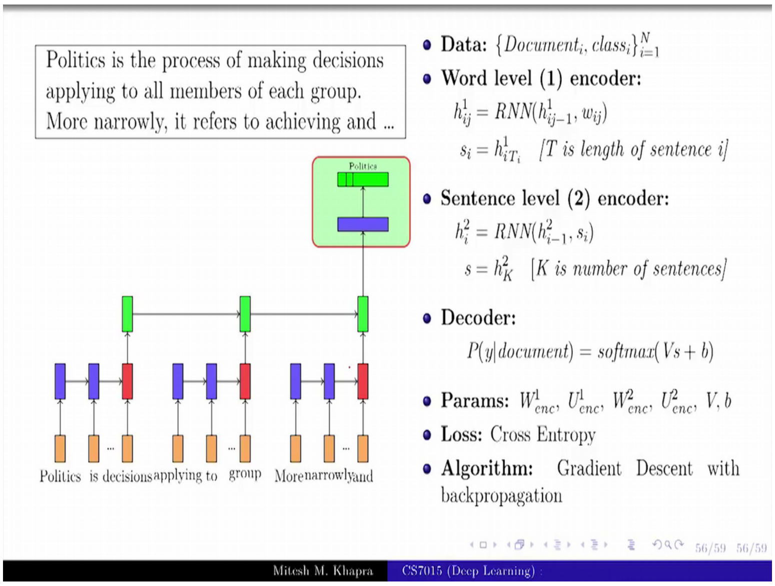
Students: (Refer Time: 08:15).

h i T i that is the last state of every sequence right. T i is a length of the ith sentence right ok. Now, what about h i 2? The green guys right and what is this h 2 k?

Students: (Refer Time: 08:29).

The last green guy, this guy, is it fine?

(Refer Slide Time: 08:33)



Politics is the process of making decisions applying to all members of each group. More narrowly, it refers to achieving and ...

- **Data:** $\{Document_i, class_i\}_{i=1}^{N}$
- **Word level (1) encoder:**

  $h_{ij}^{1} = RNN(h_{ij-1}^{1}, w_{ij})$

  $s_i = h_{iT_i}^{1}$ *[T is length of sentence i]*
- **Sentence level (2) encoder:**

  $h_i^2 = RNN(h_{i-1}^2, s_i)$

  $s = h_K^2$ *[K is number of sentences]*
- **Decoder:**

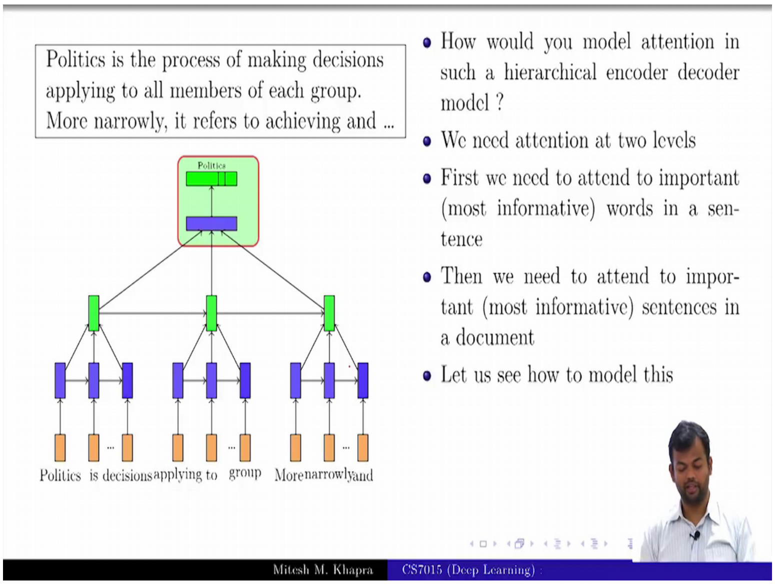
  $P(y|document) = softmax(Vs + b)$
- **Params:** $W_{enc}^1, U_{enc}^1, W_{enc}^2, U_{enc}^2, V, b$
- **Loss:** Cross Entropy
- **Algorithm:** Gradient Descent with backpropagation

56/59 56/59

Mitesh M. Khapra CS7015 (Deep Learning) :

And then the decoder is just a softmax, we do not need to go with that and loss and everything is fine. So, this again whatever it is we should always be comfortable and writing the N to N equations from x to y right and you can write it in this case.

(Refer Slide Time: 08:45)



Politics is the process of making decisions applying to all members of each group. More narrowly, it refers to achieving and ...

- How would you model attention in such a hierarchical encoder decoder model ?
- We need attention at two levels
- First we need to attend to important (most informative) words in a sentence
- Then we need to attend to important (most informative) sentences in a document
- Let us see how to model this

Mitesh M. Khapra CS7015 (Deep Learning)

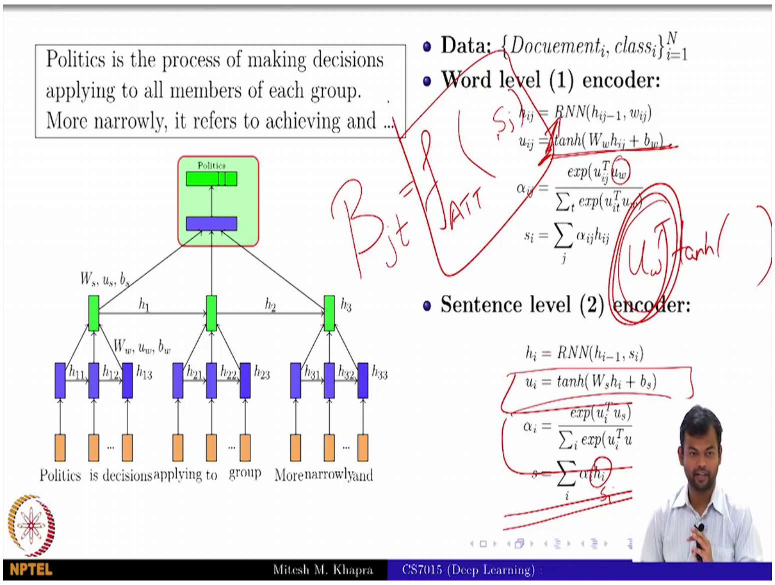
Now, let us make it a bit interesting. How would you model attention in such a hierarchical encoder decoder model? How many attention functions would you need? Two; one for attention over sentences, the other for attention over?

Students: Works.

Works ok. Can you think of these equations? Not a very big stretch from what we have done already right. I mean at level 1 it should be straight forward, at level 2 just ignore level 1.

How many if you can imagine the equations? It is not very hard I am not joking I am I mean just think about it. And the level 1 should be straight forward because that is just the same as ok. So first we need to attend to the most important words in a sentence and then we need to attend to the most important sentence in a document ok.

(Refer Slide Time: 09:35)



Politics is the process of making decisions applying to all members of each group. More narrowly, it refers to achieving and ...

Politics
W_s, u_s, b_s
h_1
h_2
h_3
W_w, u_w, b_w
h_{11} h_{12} h_{13} h_{21} h_{22} h_{23} h_{31} h_{32} h_{33}
Politics is decisions applying to group More narrowly and

- Data: $\{Docuement_i, class_i\}_{i=1}^{N}$
- Word level (1) encoder:

$$h_{ij} = RNN(h_{ij-1}, w_{ij})$$
$$u_{ij} = tanh(W_w h_{ij} + b_w)$$
$$\alpha_{ij} = \frac{exp(u_{ij}^T u_w)}{\sum_t exp(u_{it}^T u_w)}$$
$$s_i = \sum_j \alpha_{ij} h_{ij}$$

$\beta_{jt} = f_{ATT}(s_i$

$u_w^T \tanh( )$

- Sentence level (2) encoder:

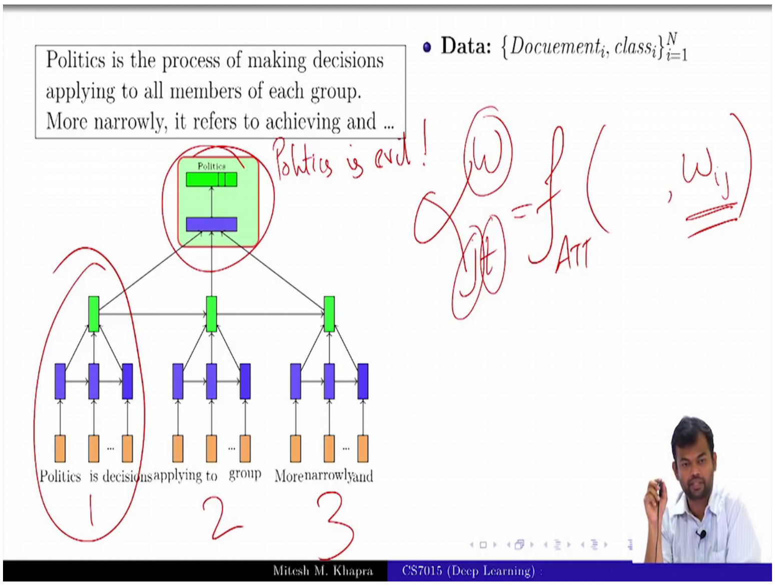
$$h_i = RNN(h_{i-1}, s_i)$$
$$u_i = tanh(W_s h_i + b_s)$$
$$\alpha_i = \frac{exp(u_i^T u_s)}{\sum_i exp(u_i^T u}$$
$$s = \sum_i \alpha_i h_i$$

s_i

Mitesh M. Khapra CS7015 (Deep Learning)

Let us have see how to model this. So, we have document, again the same input, then you have the word level RNN ok. Now what be the word level attention equation look like?

(Refer Slide Time: 09:49)



Politics is the process of making decisions applying to all members of each group. More narrowly, it refers to achieving and ...

- Data: $\{Docuement_i, class_i\}_{i=1}^{N}$

$\alpha_{ijt}^{(w)} = f_{ATT}(, w_{ij})$

Mitesh M. Khapra CS7015 (Deep Learning)

I am looking for the attention equation for words. What are the indices going to be? J, i j t, ok. What is i j t? That is I have put as superscript in w. This is the word level attention. So, at the t-th time step I want the importance of the?

Student: (Refer Time: 10:04).

j-th good, right. What would that equation depend on?

Students: The word.

The word should be straightforward right, it should depend on but sorry this is onlyfor this guy right, ok. So, you have focusing on one of these; so you trying to find the importance of these three words right which are there in the first sentence.

So, you have computed h i j; that means, you have computed all the representation for each of these word. That means, you have computed the first three blue vectors that you see ok. And then you are computing the attention, no, so this is so instead of having it here. This is how we have been writing at right.

Student: (Refer Time: 10:49).

Ok, I so sorry I should have check this. So, read this as u i j is equal to, or let me just explain it. So, remember this is a vector and we wanted to do this operation to make it a scalar right. So, u w is that parameter which was getting multiplied earlier. So, we had this attention equation as u w transpose tan h of something right. So, now, that u w has been removed from here in equation 2 and has been added as exponent to equation to the alpha equation. Does is that ok?

How many of you completely confused? Please raise your hands. How many of you understand this completely? Once can the sum the 1, ok what did I do? Let us just see if I can still salvage this. Let us go one by one. So, what is let me just delete some of these things, let us try to write it on our own right. So, this is what we are trying to do.

So, I will I am ignoring the sentence id rights. So this is sentence 1, 2 and 3; so let us just focus on one sentence and the same equations we hold for the other sentences also ok. So first of all the attention weight would depend on what? It would depend, what are we trying to pay attention to?

Student: Words.

Words; so it should have word in the input right. What else can you at have in the input?

Student: Previous.

Previous state; unfortunately for this problem do we have a previous state?

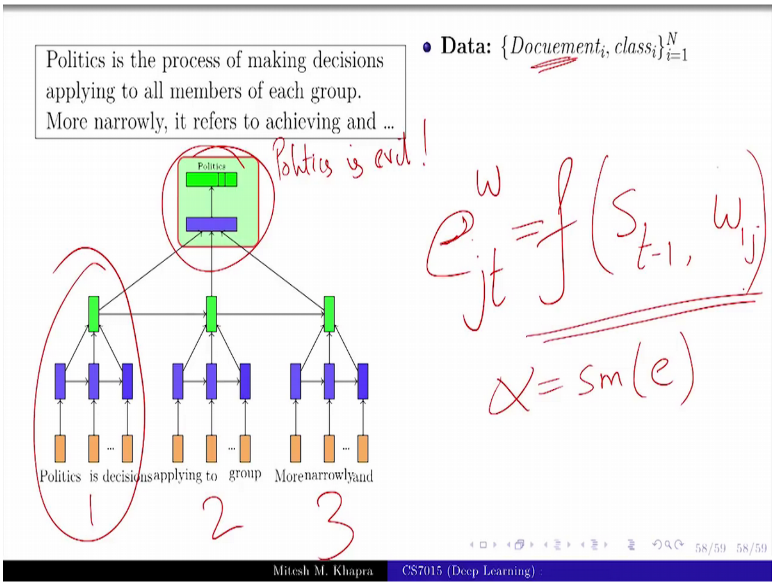
Student: (Refer Time: 12:27).

No, we just doing one prediction right, there is no RNN here, we just the feed forward network. Do you have any s t minus 1 in the output? No, that we have put here. So, this was the importance of the j-th word at the t-th time step. So j belongs to the input and t belongs to the decoder right. In this case does decoder have multiple time steps? There is only one time step of the decoder right that is the problem which we have run into.

But let us assume instead of classification we are trying to do summarisation; that means, we are given this document. And we were trying to generate summary of this and let us say the summary was the following ok. So, this is the summary that you are trying to generate from this document ok. And now this summary has 3 or 4 time steps if you count exclamation as the last time step. Now, how is it going to be? What is the decoder going to be in that case? RNN right and the decoder will have some k time steps ok. Now, at every time step at a given time step t, what am I trying to do?

In next assignment try to develop a better eraser for this, ok.

(Refer Slide Time: 13:41)



Politics is the process of making decisions applying to all members of each group. More narrowly, it refers to achieving and ...

- Data: $\{Docuement_i, class_i\}_{i=1}^{N}$

$e_{jt}^{w} = f(s_{t-1}, w_{ij})$

$\alpha = sm(e)$

Mitesh M. Khapra CS7015 (Deep Learning)

58/59

So, we want to compute, when you compute the attention for a word; the jth word in particular at that t-th time step, right. And we have for a minute understood that we do not have a feed forward network at the decoder. We have a recurrent neural network at the decoder because we are trying to generate a summary, ok, we are trying to generate a sequence at the output. So, at the t-th time step we are interested in understanding which of the document was to pay attention to ok. So, now, that is going to be a function of what and I finded a bit irritating for the want of a better word that at least one input to this function should be straight forward right. What is it?

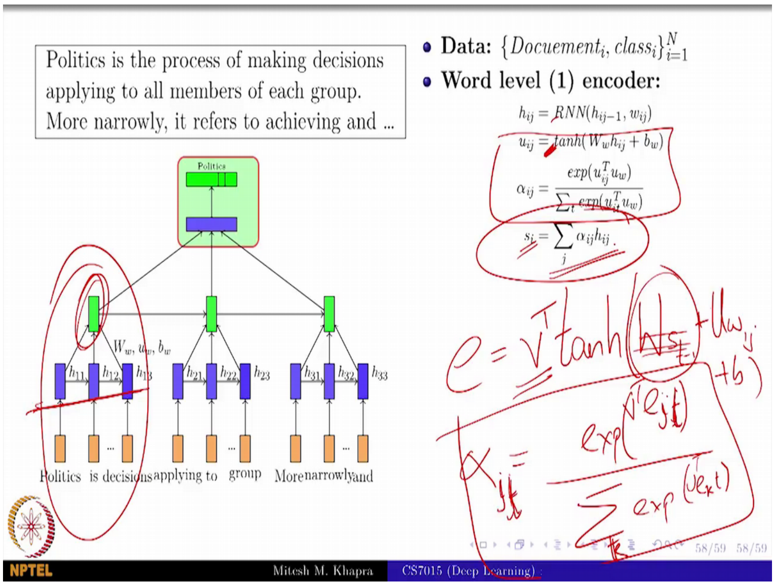
Student: The word.

The word, that you are trying to learn how much attention to pay to right. So, that should be very straightforward, so that should be w 1 j because I am considering the first sentence right now ok. Is that fine? For the first sentence I am trying to find out which are the words which are important. What is the second input that you could put in? It should depend on the index t right, so it should be the previous state of the decoder ok. And then of course, I will have a this is again actually not alpha, but e right. And then you get how do you get alpha from there, how do you get the alphas?

Student: Softmax.

Softmax, is that fine ok. So alpha will be some softmax of the e's; is that ok, fine. So, this is for the word level. Now, the equations that are written on the slide are slightly. So, the equation has slices slightly differently written, so let me just go back and write our own equation. And so we want to write an equation for this, what was our equation?

(Refer Slide Time: 15:26)



Politics is the process of making decisions applying to all members of each group. More narrowly, it refers to achieving and ...

- Data: $\{Docuement_i, class_i\}_{i=1}^{N}$
- Word level (1) encoder:

$$h_{ij} = RNN(h_{ij-1}, w_{ij})$$

$$u_{ij} = tanh(W_w h_{ij} + b_w)$$

$$\alpha_{ij} = \frac{exp(u_{ij}^T u_w)}{\sum_t exp(u_{it}^T u_w)}$$

$$s_i = \sum_j \alpha_{ij} h_{ij}$$

$$e = v^T tanh(W s_{t-1} + U w_{ij} + b)$$

$$\alpha_{jt} = \frac{exp(v^T e_{jt})}{\sum_k exp(v^T e_{kt})}$$

Ignore the equations on the slide. it was something like this; v transpose tan h of w.

Student: s t.

s t minus 1 plus u w 1 j plus?

Student: b.

b ok. Now, imagine that your decoder is a feed forward neural network. What will be missing in that case? There is no s t minus 1; there is no previous state of a decoder because we just want to make a prediction once by paying attention to all the important words and sentences in your document right. So, which part will go away? w s t minus 1 ok, fine.

Now, the other thing that you are doing is alpha was actually or other alpha i j is actually exponent of e i j divided by summation of other k sorry; alpha j t e j t, e k t; is that fine? It is just the softmax equation; is that ok right? Now, the only thing that you see different and these two equations here is, first you do not have the w s t minus 1 because the decoder only has one time step. And second we have taken out this v transpose from here and instead we have added it here. Is that ok? Does that make sense? This is just different way of writing it. So, again you write the attention equation for the words now ok. Now what about the sentence? Now first of all, earlier what were we using for the green guy, what was the representation for the green vector? It was the; what was the green vector?

In the absence of attention what was the green vector? h t i right, the last time step of sentence 1. Is it fine everyone get me please raise your hands, if you are with me ok. Now what would it be? It would be a dash sum of w vectors, a weighted sum, attention weighted sum right; so that is exactly what this equation is capturing. So, what did you saying is there is representation of sentence i is a weighted sum of the representations of all the words in that sentence is it ok.

So, that we will get a representation for s 1, s 2 up to s capital K all the sentences that you have. Now what do you want to do for the second level what do we want? We want to compute the importance of that sentence for the t-th time step right. So, let us call that beta. So I am interested in beta if we need to really read out this I said again alpha is being used in both the places right.

So, you want to find out the importance of the j-th sentence at the t-th time step. What is it going to be a function of? 1 is a sentence representation, what is the sentence representation given by?

Student: s j.

s j and what else? The decoder state at the previous time step right. Does the decoder have a previous time step state here? No. So, it will just depend on s j and that is exactly what this equation is capturing right. And again the same trick that I have added this extra parameter to the exponent. Is that fine? And the final representation being fed to the feed forward network is a weighted sum of the sentence representations right. So, this again has to be s i alpha into s i ok.

I really sorry about this, but I am pretty sure that once we correct the slides and then you go back and look at it should be clear right. It just two sets of equation, one set of equation sorry that is correct sorry. So, this the idea is there are two sets of attention mechanisms, for each you will have your own set of equations. The basic form if you can work out what the f attention would depend on. The actual form would depend from would differ from paper to paper or the toolbox to toolbox. That does not matter so much, you just need to know that you have these as the input. You are going to add some parameters to every input that means, you are going to do linear transformation. And then you just need to make sure that alphas eventually turn out to be scalars right. That is why you will have this additional vector getting multiplied at one point.