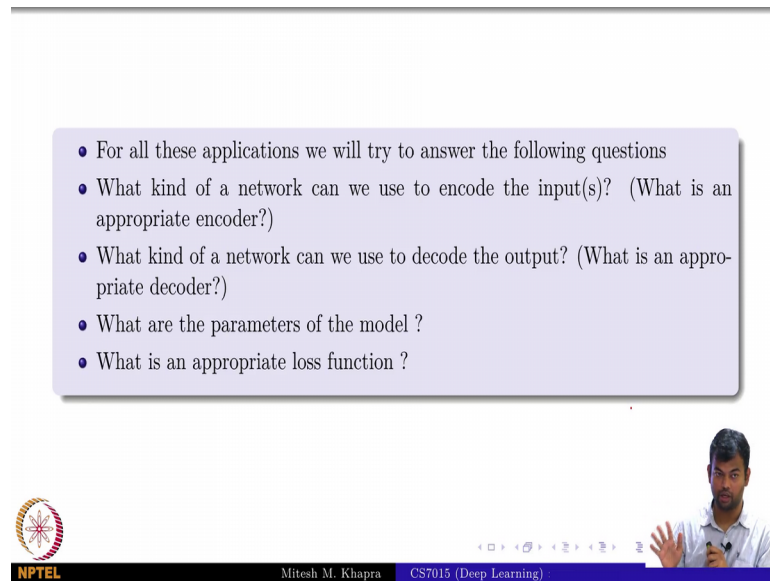


Deep Learning
Department of Computer Science and Engineering
Indian Institute of Technology, Madras



Lecture – 113
Applications of Encoder Decoder models

So, we are going to see a lot of Applications of the Encoder Decoder models.

(Refer Slide Time: 00:17)



- For all these applications we will try to answer the following questions
- What kind of a network can we use to encode the input(s)? (What is an appropriate encoder?)
- What kind of a network can we use to decode the output? (What is an appropriate decoder?)
- What are the parameters of the model ?
- What is an appropriate loss function ?

 
NPTEL Mitesh M. Khapra CS7015 (Deep Learning)

And, for all these applications we are trying to answer the following questions. What kind of network can we use to encode the input? In the previous application what do we use CNN what kind of network can be used to decode the output? What did we use?

Student: RNN.

RNN. What are the parameters of the model? We will see that and what is an appropriate loss function, right.

(Refer Slide Time: 00:37)

Task: Image captioning

Data: $\{x_i = \text{image}_i, y_i = \text{caption}_i\}_{i=1}^N$

Model:

- Encoder:**

$$s_0 = \text{CNN}(x_i)$$
- Decoder:**

$$s_t = \text{RNN}(s_{t-1}, e(\hat{y}_{t-1}))$$

$$P(y_t | y_1^{t-1}, I) = \text{softmax}(Vs_t + b)$$
- Parameters:** $U_{dec}, V, W_{dec}, W_{conv}, b$
- Loss:**

$$\mathcal{L}(\theta) = \sum_{i=1}^T \mathcal{L}_i(\theta) = - \sum_{i=1}^T \log P(y_i = \ell_i | y_1^{i-1}, I)$$
- Algorithm:** Gradient descent with back-propagation

Mitesh M. Khapra CS7015 (Deep Learning) 18/59

So, let us again go back to this task which was image captioning. What is the input what is the training data given to you? What is x? What is y? X is the image, y is the description, right. So, this is what is given to you given N such training pairs where x i is the image and y i is the description and y i in itself is a sequence, right. So, you have y i 1 to y i capital T, everyone gets the input and output.

Now, what is the next thing? Model. Can you write down the model equations? I want an equation which starts from x and goes all the way up to y and since we have several time steps I want an equation for y t, this is generic for every time step right say can you write that equation and feel free to use shortcuts. So, you do not need to write the entire RNN equation just say RNN of something do not even try to write the VGG 16 CNN equations just say CNN ok so, we will go ahead.

The first thing when I am going to do this I am going to write the equation for the encoder. So, the encoder gives me CNN of x i whatever is the input given to me, x i is the i-th training image given to me. So, I will just pass it through CNN I will get a representation for that and I am just being cryptic here it could be the f c 7 representation or the con 5 representation or the max fool 5 representation or whatever you want, right and it is going to denote all of this as CNN of x i, run this CNN take whichever representation you want to take.

Now, what is the decoder going to be? Decoder is the following RNN remember the equation of RNN was s_{t-1} comma x_t , what is the input of the t -th time step? Whatever we are predicted at the previous time step just the embedding of that. So, e means embedding, if you want take one naught embedding if you want take (Refer Time: 02:29) embedding is that fine.

And then what is the output? It is the soft max function of the following. How many of you get this now please raise your hands, how many of you can say that why you can be written as a function of x , is that pretty straight forward ok. So, you have an encoder, you have a decoder and remember that this final y is a composite function of the original input x ok. Just that you are doing too many computations along the way, but there is a path which exists ok.

What is the loss function everyone at this point should be able to say it?

Student: sum of cross entropies.

Sum of cross entropies. They just wait for me to say two more sentences. What are the parameters U V W B C ?

Student: (Refer Time: 03:12).

A B C D E F [laughter] everything, right. What is that? All the parameters of the?

Student: (Refer Time: 03:17).

Convolutional neural network; that means, all the filters that you have, all the parameters of the RNN which is W , U and the parameters of the output layer which is V , right, all of these is that fine, I am I may have missed some biases but ok. The objective function as you said is a sum of cross entropies, where l_t is the true character at time step true word at time step t . And what is the algorithm that you are going use back propagation through time and with back propagate all the way through the CNN also which is an end to end thing. In practice of course, you do not do that yes you could just said both to be the same. Do you get that question that ok?

(Refer Slide Time: 04:00)

o/p : The ground is wet

i/p : It is raining outside

- **Task:** Textual entailment
- **Data:** $\{x_i = \text{premise}_i, y_i = \text{hypothesis}_i\}_{i=1}^N$
- **Model (Option 1):**
 - **Encoder:**

$$h_t = \text{RNN}(h_{t-1}, x_{it})$$
 - **Decoder:**

$$s_0 = h_T \quad (T \text{ is length of input})$$

$$s_t = \text{RNN}(s_{t-1}, e(y_{t-1}))$$

$$P(y_t | y_1^{t-1}, x) = \text{softmax}(Vs_t + b)$$
- **Parameters:** $U_{dec}, V, W_{dec}, U_{enc}, W_{enc}, b$
- **Loss:**

$$\mathcal{L}(\theta) = \sum_{i=1}^T \mathcal{L}_i(\theta) = - \sum_{t=1}^T \log P(y_t = \ell_t | y_1^{t-1}, x)$$
- **Algorithm:** Gradient descent with back-propagation

Mitesh M. Khapra CS7015 (Deep Learning) 19/59 19/59

Now, let us look at another task we look at the task of textual entailment. What textual entailment does is that I give you input or premise the premises that it is raining outside and you need to tell me a hypothesis. The hypothesis that the ground it is wet with basically means that it is raining outside implies that the ground is wet ok. Now, what is the encoder decoder architecture that you will use for this problem? What is the input here?

Student: A sequence.

A sequence. What is the output? Sequence. It is all the hint that I am going to give you. So, what will you do what is the encoder equation going to be?

Student: RNN.

RNN. What is the decoder equation going to be?

Student: RNN.

RNN. How will it become end to end by setting what to what?

Student: (Refer Time: 04:49).

SO of the decoder to what of the encoder?

Student: (Refer Time: 04:49).

Last time step of the encoder. How many if you get that really we are on the same page first time in I do not know how many lectures by finally, it happened. So, here is what training data is, right it is a collection of premises and hypothesis and you have N of these. There are two options for the model the first option is that you encode the input using an RNN feel free to replace and by an LSTM if you want then you have the decoder where and you set the zero time step to whatever you got from the encoder.

Then every time step you computed using the RNN where remember the input at every time step is whatever you predicted at the previous time step and then the output is just the soft max function, is that fine? And what is the loss function going to be loss function?

Student: (Refer Time: 05:44).

Sum of loss entropies. Training algorithm?

Student: Back propagation.

Back propagation through time and really it is through time right all the way back ok so, we will see that. Let me see if I had any other question ha ok I will ask it parameters I am not going to bother about ok. Now, this was option 1 I have just clearly written, what is option 2 what is the set of equations look like for option 2? Which of these equations will change and how? Remember option 2 was maybe pass the input at every time step which equation will change?

St what will it become, but st can take only I mean RNN take only two inputs, right. St minus when you need to gave embedding you need to gave so, how will you fit in the third input, this animation has it is own mind. So, this is how back propagation will happen, right so, let us it is finish that so, will actually back propagate all the way back through time, fine really all the way back through time.

(Refer Slide Time: 06:40)

o/p : The ground is wet

i/p : It is raining outside

- **Task:** Textual entailment
- **Data:** $\{x_i = \text{premise}_i, y_i = \text{hypothesis}_i\}_{i=1}^N$
- **Model (Option 2):**
 - **Encoder:**

$$h_t = RNN(h_{t-1}, x_{it})$$
 - **Decoder:**

$$s_0 = h_T \quad (T \text{ is length of input})$$

$$s_t = RNN(s_{t-1}, [h_T, e(\hat{y}_{t-1})])$$

$$P(y_t | y_1^{t-1}, x) = \text{softmax}(Vs_t + b)$$
- **Parameters:** $U_{dec}, V, W_{dec}, U_{enc}, W_{enc}, b$
- **Loss:**

$$\mathcal{L}(\theta) = \sum_{i=1}^T \mathcal{L}_i(\theta) = - \sum_{i=1}^T \log P(y_i = \ell_i | y_1^{i-1}, x)$$
- **Algorithm:** Gradient descent with back-propagation

20/59 20/59

Mitesh M. Khapra CS7015 (Deep Learning)

And same task textually entailment I want model 2 option 2 so, this is what will happen. We will just concatenate h_T which is this guy along with the input at every time step, right. How many if you get this the RNN is still taking just two inputs? One is the previous state the other is the concatenation of the current input as well as input that we got from the encoder everyone get this ok. So, this is model 2 I am going forward I am not going to do both model 1 and model 2 it is model 2 is just a very simple variation of model 1, a parameters loss function training algorithm everything remains the same ok.

(Refer Slide Time: 07:20)

o/p : Mein ghar ja rahi hoon

i/p : I am going home

- **Task:** Machine translation
- **Data:** $\{x_i = \text{source}_i, y_i = \text{target}_i\}_{i=1}^N$
- **Model (Option 2):**
 - **Encoder:**

$$h_t = RNN(h_{t-1}, x_{it})$$
 - **Decoder:**

$$s_0 = h_T \quad (T \text{ is length of input})$$

$$s_t = RNN(s_{t-1}, [h_T, e(\hat{y}_{t-1})])$$

$$P(y_t | y_1^{t-1}, x) = \text{softmax}(Vs_t + b)$$
- **Parameters:** $U_{dec}, V, W_{dec}, U_{enc}, W_{enc}, b$
- **Loss:**

$$\mathcal{L}(\theta) = \sum_{i=1}^T \mathcal{L}_i(\theta) = - \sum_{i=1}^T \log P(y_i = \ell_i | y_1^{i-1}, x)$$
- **Algorithm:** Gradient descent with back-propagation

22/59 22/59

Mitesh M. Khapra CS7015 (Deep Learning)

Let us look at machine translation what is the input an English sentence what is the output a Hindi sentence, what is the encoder going to be?

Student: RNN.

RNN. What is the decoder going to be?

Student: RNN6RNN. What is the loss function going to be?

Student: (Refer Time: 07:37).

Soft max. Who said soft max?

Student: (Refer Time: 07:42).

What is the loss function going to be?

Student: sum of.

Sum of cross entropies, training algorithm all the way through time, right ok. So, let us can you draw can you write the equations? Just copy it from the previous slide, right actually copy it from the previous slide, right. If you have the RNN you have the RNN as a decoder again an option when you will set s_0 to h_t you have the loss function the parameters and your training algorithm and for option 2 it is back propagation will, fine and for option 2 what will happen? Option 2 what will happen?

Student: (Refer Time: 08:17).

This will change, right. So, just focus on that we just passed in the last time say belong with that.

(Refer Slide Time: 08:27)

o/p : इ . ड फ व 1

i/p : I N D I A

- Task: Transliteration
- Data: $\{x_i = srcword_i, y_i = tgtword_i\}_{i=1}^N$
- Model (Option 1):
 - Encoder:
$$h_t = RNN(h_{t-1}, x_{it})$$
 - Decoder:
$$s_0 = h_T \quad (T \text{ is length of input})$$
$$s_t = RNN(s_{t-1}, e(y_{t-1}))$$
$$P(y_t | y_1^{t-1}, x) = softmax(Vs_t + b)$$
- Parameters: $U_{dec}, V, W_{dec}, U_{enc}, W_{enc}, b$
- Loss:
$$\mathcal{L}(\theta) = \sum_{i=1}^T \mathcal{L}_i(\theta) = - \sum_{t=1}^T \log P(y_t = \ell_t | y_1^{t-1}, x)$$
- Algorithm: Gradient descent with back-propagation

Mitesh M. Khapra CS7015 (Deep Learning) 23/59 23/59

Now, transliteration, what is transliteration? What is transliteration if you do not know it at least see it from the example what is it?

Student: (Refer Time: 08:34).

Writing the same word in another language, right. So, this is typically done for named entities very often when you are; when you are translating from one language to another you do not often for Thomas you do not come up with an Indian translation right you just say Thomas in Devanagari, right you just write Thomas in the Devanagari, right. So, for names you typically just do a transliteration; that means, from the English script you just write it in the native language script ok. What is the input? One word the input is the word right. What is it, a sequence of?

Student: Characters.

Characters. What is the output?

Student: Sequence of characters.

A sequence of characters. What will you use for the input?

Student: RNN.

RNN. What for the output?

Student: RNN.

RNN. So, I will becoming too easy, right. Can you write the equations for this yes you will copy it from the previous slide, yes ok. Everything remains the same, right. So, you see why this (Refer Time: 09:28) as become powerful if you do not see it still maybe let us look at something else.

(Refer Slide Time: 09:33)

O/p: White

White

What is the bird's color

CNN

Question: What is the bird's color

- Task: Image Question Answering
- Data: $\{x_i = \{I, q\}_i, y_i = \text{Answer}_i\}_{i=1}^N$
- Model:
 - Encoder:
$$\hat{h}_T = \text{CNN}(I), \tilde{h}_t = \text{RNN}(\tilde{h}_{t-1}, q_{it})$$
$$s = [\tilde{h}_T; \hat{h}_T]$$
 - Decoder:
$$P(y|q, I) = \text{softmax}(Vs + b)$$
- Parameters: $V, b, U_q, W_q, W_{conv}, b$

Mitesh M. Khapra CS7015 (Deep Learning)

Image question answering; tell me what is the data here? What is the input? Image and?

Student: Question.

Question and what is the answer what is the output answer. So, for simplicity we are going to assume that the answer here is a finite vocabulary. We are not generating descriptive answers, we are not being overly dramatic let us going to say one word what is the colour white. We are not going to write I think the colour of the image is white now ok, just white. So, all these outputs are going to be single words and we have v possibilities and we are going to predict one of those v possibilities ok.

Now, give me a model for this, now, things are getting slightly complicated you have one image as the input one sequence as an input and a dash as the output god now think why would you generate the sequence of characters of the answer. I said that the answer is going to be come from a finite vocabulary; that means, you need a?

Student: probability distribution.

A distribution probability distribution is here enough said. Now, tell me what is the model? A model should connect the input to the output you have two inputs here. I see some people doing this [laughter] I do not know what that means, but let us just do it let us make a train. Simple formula simple recipe and whatever input you are given just encode it depending on the type of input you know what is the encoding is going to be for images what is encoding sequences?

Student: (Refer Time: 10:55).

Now what do you do with these two separate things?

Student: (Refer Time: 10:58).

Concatenate them and then?

Student: (Refer Time: 11:01).

After that?

Student: (Refer Time: 11:02).

Can you think of all the equations can imagine all the equations along the way?

Student: Yes.

Of course, yes [laughter] right I mean imagination [laughter] you can always do that [Laughter]. Now, just think about it can you write the output as a function of the input where the input is actually a pair now, it is image comma question what is the model going to look like let us see. So, model will first have an encoder for the image let us call that as h_I , it is going to have an encoding of the question let us call it as h_Q I am going to concatenate these two as someone rightly gestured and then what am I going to do after that pass it through a?

Student: Feed forward network.

Feed forward network and predict of probability distribution. What are the parameters of this network? Parameters of the feed forward network, the parameters of the recurrent

neural networks and the parameters of the CNN, right. So, everything that we have done so fine because ok. How do you train it? Back propagate through time and space also go back to the image also fine is that ok.

(Refer Slide Time: 12:09)

o/p : India won
the world cup

i/p : India beats Srilanka to win ICC WC 2011.
Dhoni and Gambhir's half centuries help beat SL

- **Task:** Document Summarization
- **Data:** $\{x_i = \text{Document}_i, y_i = \text{Summary}_i\}_{i=1}^N$
- **Model:**
 - **Encoder:**

$$h_t = \text{RNN}(h_{t-1}, x_{it})$$
 - **Decoder:**

$$s_0 = h_T$$

$$s_t = \text{RNN}(s_{t-1}, e(\hat{y}_{t-1}))$$

$$P(y_t | y_1^{t-1}, x) = \text{softmax}(Vs_t + b)$$
- **Parameters:** $U_{dec}, V, W_{dec}, U_{enc}, W_{enc}, b$
- **Loss:**

$$\mathcal{L}(\theta) = \sum_{i=1}^T \mathcal{L}_i(\theta) = - \sum_{t=1}^T \log P(y_t = \ell_t | y_1^{t-1}, x)$$
- **Algorithm:** Gradient descent with backpropagation

26/59

Document summarisation; what is the input? Sequence. What is the output? RNN, RNN everywhere, fine. I will not even bother to ask you

(Refer Slide Time: 12:18)

A man walking on a rope

- **Task:** Video Captioning
- **Data:** $\{x_i = \text{video}_i, y_i = \text{desc}_i\}_{i=1}^N$
- **Model:**
 - **Encoder:**

$$h_t = \text{RNN}(h_{t-1}, \text{CNN}(x_{it}))$$
 - **Decoder:**

$$s_0 = h_T$$

$$s_t = \text{RNN}(s_{t-1}, e(\hat{y}_{t-1}))$$

$$P(y_t | y_1^{t-1}, x) = \text{softmax}(Vs_t + b)$$
- **Parameters:** $U_{dec}, W_{dec}, V, b, W_{conv}, U_{enc}, W_{enc}, b$
- **Loss:**

$$\mathcal{L}(\theta) = \sum_{i=1}^T \mathcal{L}_i(\theta) = - \sum_{t=1}^T \log P(y_t = \ell_t | y_1^{t-1}, x)$$
- **Algorithm:** Gradient descent with backpropagation

27/59

Video captioning; sequence of images I want to hear the choice of phrasing that you use I just want to hear that I love hearing that every time.

Student: (Refer Time: 12:30).

RNN of CNNs whatever that means, what RNN of CNN. Every time I do this everyone says RNN of CNN [laughter] I do not know what that means, but it is the right answer. What does it mean what is the video? It is a sequence. What is the sequence of?

Student: Images.

Images. So, what will you do, encode every image and then pass it through a?

Student: RNN.

RNN can you imagine the equations, let us see and in this case what is the output again? The sequence. So, what is the decoder going to be?

Student: RNN.

So here is the model. So, first what you do is for every time step you compute the CNN encoding of the frame, then you pass it through an RNN to get the final time step t and then you feed it to a decoder and generate one word at a time that fine. So, this thing apparently is called RNN CNNs ok. And so, that is and loss function would again be the same sum of cross entropies and back propagation through time and space good. Please do not quote me on this thing also this is getting a recorded but ok.

(Refer Slide Time: 13:41)

The image shows a presentation slide with the following elements:

- Top left: "o/p: Surya Namaskar"
- Top right: "Task: Video Classification"
- Center: A sequence of three frames showing a person performing a yoga pose (Surya Namaskar), with an ellipsis between the second and third frames.
- Bottom right: A small video feed of a man speaking.
- Bottom: A navigation bar with icons for back, forward, and search, and a footer with "Mitesh M. Khapra" and "CS7015 (Deep Learning)".

The next one video classification what is the decoder? Decoder is probability distribution, what is the decoder?

Student: feed forward neural network.

Feed forward neural network.

(Refer Slide Time: 13:54)

- Task: Dialog
- Data: $\{x_i = \text{Utterance}_i, y_i = \text{Response}_i\}_{i=1}^N$
- Model:
 - Encoder:

$$h_t = \text{RNN}(h_{t-1}, x_{it})$$
 - Decoder:

$$s_0 = h_T \quad (T \text{ is length of input})$$

$$s_t = \text{RNN}(s_{t-1}, e(\hat{y}_{t-1}))$$

$$P(y_t | y_1^{t-1}, x) = \text{softmax}(V s_t + b)$$
- Parameters: $U_{dec}, V, W_{dec}, U_{enc}, W_{enc}, b$
- Loss:

$$\mathcal{L}(\theta) = \sum_{i=1}^T \mathcal{L}_i(\theta) = - \sum_{i=1}^T \log P$$

Mitesh M. Khapra CS7015 (Deep Learning)

This one, dialogue how are you I am not fine, input.

Student: RNN.

Output?

Student: RNN.

RNN, right. So, you see this why this has become so, popular we took a wide range of problems different modalities right we took images, we took videos, we took sequences or combination of this right image question answering has a combination of images and sentences and the output you have a probability distribution. All of this could be model by this unified end to end network, all of the components are neural network based components whether it is a convolutional network or a feed forward network or a recurrent neural network, right.

Now, let me stretch this right what if you have video question answering. What is the input going to be sequence of images and sequence of?

Student: Words.

Words. The output is?

Student: (Refer Time: 14:46).

No. Just a word right we will pick from a fixed vocabulary. How are you going to model the input?

Student: (Refer Time: 14:53).

RNN for the question, RNN or CNN for the video, then?

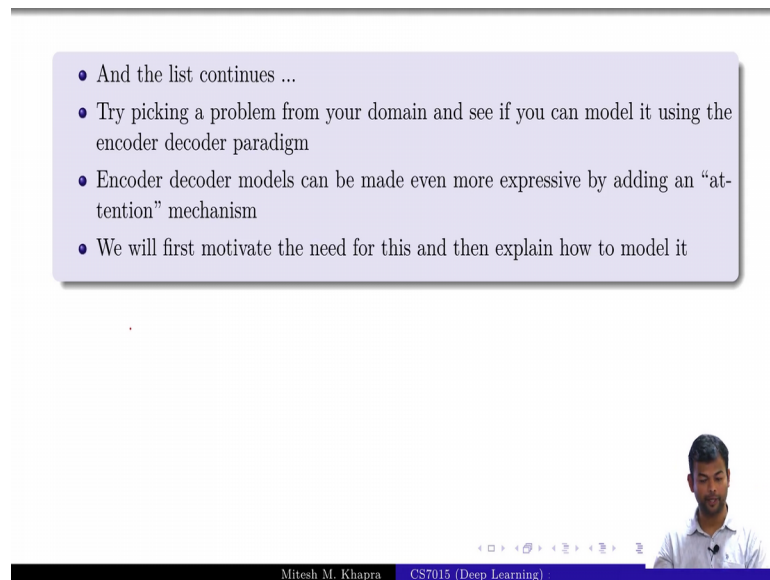
Student: Concatenate.

Concatenate and?

Student: Feed.

Feed it to your feed forward network right. So, all of these become to what extent that work is a separate question, but all it was not even possible to model all of this as an end to end network, right. But, now I just because possible to model it as an end to end network with all the components being neural components, right. And, this is this story still not complete everything is not as easy as it looks they still and very crucial component that we have missing in the architectures that we have seen so far and we will talk about that soon ok.

(Refer Slide Time: 15:33)



The slide contains a list of four bullet points in a light purple box:

- And the list continues ...
- Try picking a problem from your domain and see if you can model it using the encoder decoder paradigm
- Encoder decoder models can be made even more expressive by adding an “attention” mechanism
- We will first motivate the need for this and then explain how to model it

Below the list is a video inset showing a man speaking. At the bottom of the slide, there is a navigation bar with the text "Mitesh M. Khapra" and "CS7015 (Deep Learning)".

Now, let us just continue that I challenge you to do this, pick up any problem there are student from relevant different departments, pickup any problem do not say that I want to design some gear for certain aeroplane and all that and I want to use neural network to do so no. Something which involves machine learning, right not problem is does not involve machine learning and see if you can model it using the encoder decoder frame work. just try to do.

This take problem from by take right for example, they given a sequence of genes and you want to predict whether this person is susceptible to a certain disease, what will you do? Conduct a blood test ok. Do not do not [laughter] do not go and do neural networks for that, but if you had to do that this is what you will do it will take a sequence you will treat the sequence of sequence at the given DNA as a sequence of genes and then you will try to predict something as the output try to predict a probability distribution over disease it a possible, right.

So, you can think of many applications from many domains and all of that you could problems involving machine learning with potentially model than using the neural encoder decoder architecture. But, there is a very important part missing from this whole story which is attention which is a very important idea and we will spend some time on that in the remainder of the lectures. So, we will first motivate why do we need attention

and from there we will see that how do you make how do you integrate attention with all these encoder decoder architectures that you have seen so far.