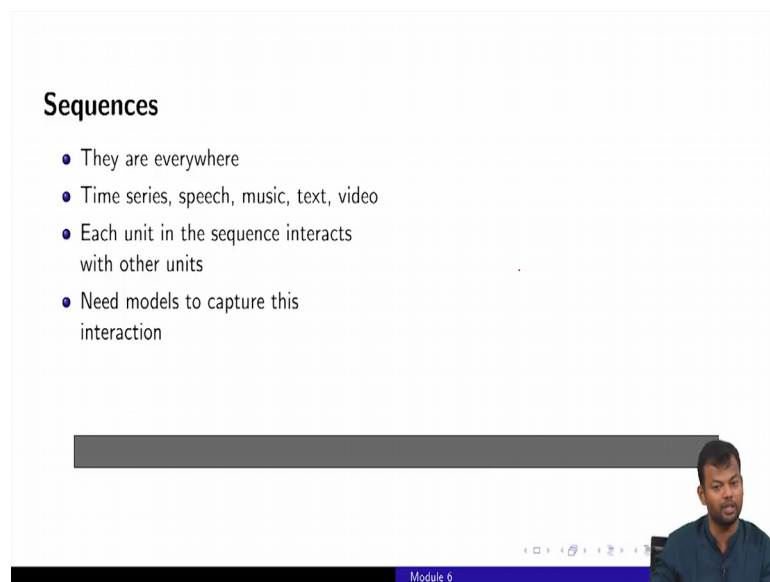


Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 01
Chapter 6: The Curious Case of Sequences

So, I was talking about successes in image, speech, pattern recognition even natural language processing and so on. So, the interesting thing here is about sequences, right. So, I will talk about sequences now.

(Refer Slide Time: 00:27)



Sequences

- They are everywhere
- Time series, speech, music, text, video
- Each unit in the sequence interacts with other units
- Need models to capture this interaction

Module 6

Sequences are everywhere when you are dealing with data. So, you have time series which is like say the stock market trends or any other kind of a series, time series, then you have speech which is again a series of phonemes or you have music. You have text which is a series of words, you could even have videos which are the series of images, right, one frame, each image, each frame can be considered to be an image and so on right.

So, in speech data one peculiar characteristic of speech data is that every unit in the sequence interacts with other units, right. So, words on their own may not mean much, but when you put them together into a sentence, they all interact with each other and give meaning to the sentence, right and the same can be said about music or speech or any

kind of sequence data, right. So, all these elements of the sequence actually interact with each other.

So, there was a need for models to capture this interaction and this is very important for natural language processing because in natural language processing, you deal with sequence of words or all your texts or sentences or documents or all sequences of words, right. So, that is very important and the same in the case of speech also.

So, if you take up any deep learning paper, nowadays it is very likely that you will come across the term Recurrent Neural Network or LSTMS which are long short term memory cells and so on, right.

(Refer Slide Time: 01:47)

Hopfield Network
Content-addressable memory systems for storing and retrieving patterns^[22]

1982

Hopfield

Original 'T'

Half of image corrupted by noise

Module 6

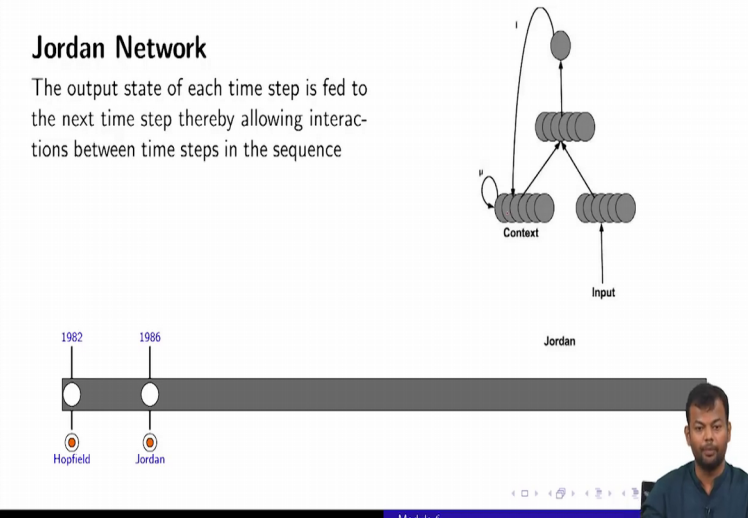
The slide features a diagram of a Hopfield network with four nodes and their interconnections. Below the diagram are two binary images of the letter 'T': one labeled 'Original T' and another labeled 'Half of image corrupted by noise'. A vertical timeline on the left shows the year 1982 with a circular marker and the name 'Hopfield' below it. A video feed of a presenter is visible in the bottom right corner, and the text 'Module 6' is at the bottom center.

So, this is also something which was proposed way back in 1986, right.

(Refer Slide Time: 01:49)

Jordan Network

The output state of each time step is fed to the next time step thereby allowing interactions between time steps in the sequence



1982 1986

Hopfield Jordan

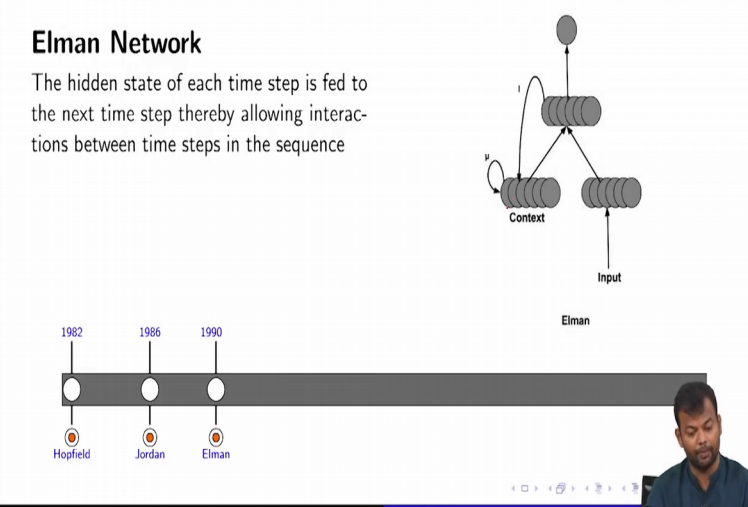
Module 6

So, a recurrent neural network is something which allows you to capture the interactions between the elements of your sequence. I had said at a very layman level, but of course, you are going to see this in much more detail in the course. And this was also not something new even though you hear about it a lot in the past 3 to 4 years. The first recurrent neural network and what you see here is exactly a very similar to what we are going to cover in the course was proposed way back in Jordan by Jordan in 1986.

(Refer Slide Time: 02:23)

Elman Network

The hidden state of each time step is fed to the next time step thereby allowing interactions between time steps in the sequence



1982 1986 1990

Hopfield Jordan Elman

Module 6

Its variant was proposed by Elman in 1990, right. So, this is again not a very new idea. This has existed for some time, but now there are various factors because of which it has been possible to now start using them for a lot of practical applications. As I said one, you have a lot of compute time and the other you have a lot of data and the third is now the training has stabilized a lot because of these advances which I was talking about in terms of better optimization algorithms, better regularization, better weight initialization and so on.

So, it has become very easy to train these networks for real world problems at a large scale, right. So, that is why they have become very popular and hear about them on a regular basis, but it is again something which was done way back.

(Refer Slide Time: 03:04)

Drawbacks of RNNs

Hochreiter et. al. and Bengio et. al. showed the difficulty in training RNNs (the problem of exploding and vanishing gradients)

Timeline:

- 1982: Hopfield
- 1986: Jordan
- 1990: Elman
- 1991-1994: RNN drawbacks

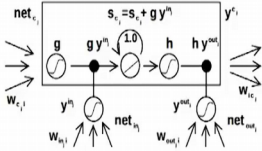
Module 6

So, from 1999 to 1994, actually people also looking at various problems will be training neural networks and recurrent neural networks, so that this problem which is known as exploding and the vanishing gradient problem which is again something that we will see in the course Unreasonable Detail. We have this problem and it is very difficult. You train recurrent neural networks for longer sequences, right. So, if you have a very long sequence or a time series, you cannot really train a recurrent neural network to learn something from that.

(Refer Slide Time: 03:34)

Long Short Term Memory

Showed that LSTMs can solve complex long time lag tasks that could never be solved before



The diagram illustrates the internal structure of an LSTM cell. It shows an input layer with nodes y^m and y^{md} , a hidden state layer with nodes g and h , and an output layer with nodes y^r . The hidden state is updated according to the equation $s_t = s_{t-1} + g y^m$. The input y^m is processed by a gate g , and the hidden state h is processed by another gate h . The output y^r is produced by a third gate. The diagram also shows the weights w_{x_i} and w_{h_i} connecting the input and hidden state layers.

Timeline of key events in RNN development:

- 1982: Hopfield
- 1986: Jordan
- 1990: Elman
- 1991-1994: RNN drawbacks
- 1997: LSTMs

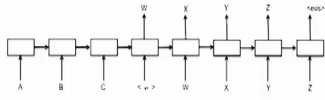
Module 6

To overcome these problems around 1997, Long Short Term Memory cells were proposed and this is again something that we will cover in the course and this is now almost the factor standard used for training for a lot of work. LSTM are used as one of the building blocks and another variants of LSTMs which are known as gated recurrent units and some other variants.

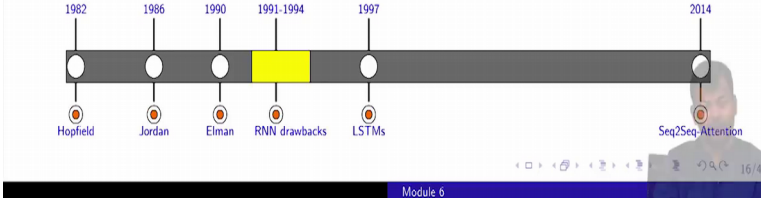
So, this is also not something new even though they have become very popular nowadays like almost any article that you pick about to talk about, any article on deep learning. The topic about to talk about recurrent neural networks or LSTMs or gated recurrent units, this is not something which is new.

(Refer Slide Time: 04:23)

Sequence To Sequence Learning



- Initial success in using RNNs/LSTMs for large scale Sequence To Sequence Learning Problems
- Introduction of Attention which inspired a lot of research over the next two years



Timeline of research milestones:

Year	Researcher/Model
1982	Hopfield
1986	Jordan
1990	Elman
1991-1994	RNN drawbacks
1997	LSTMs
2014	Seq2Seq Attention

Module 6

LSTMs had come way back in 1997 ah, but again due to various compute and other issues which I said at that time, it is not so easy to use them, but by two 2014 because of these parallel progresses which I mentioned in terms of optimization regularization and so on, people are now able to use RNNs LSTMs for large scale sequence to sequence problems. And in particular a very important discovery at this time are very important model which was proposed at this time which is Attention Mechanism which is used in a lot of deep neural networks nowadays which enabled to deal with a lot of sequence prediction problems.

For example, translation where you have given one sequence in one language and you want to generate the equivalent sequence in another language. So, this is known as a sequence to sequence translation problem. So, for that people proposed a sequence to sequence attention network and this was one of the key discoveries which then led to a lot of adaptation of adoption of deep neural networks for NLP.

A lot of research in NLP happened which was then driven by deep neutral networks. So, a lot of existing algorithms which are non neural network based algorithms which are traditionally used for NLP was slowly replaced by these deep neural network based algorithms, ok.

(Refer Slide Time: 05:33)

RL for Attention

Schmidhuber & Huber proposed RNNs that use reinforcement learning to decide where to look

The diagram features a horizontal timeline with a dark grey bar. Above the bar, vertical lines mark the years 1982, 1986, 1990, 1991, 1991-1994, 1997, and 2014. Below the bar, corresponding icons and labels are placed: Hopfield (1982), Jordan (1986), Elman (1990), RNN drawbacks (1991), RL-Attention (1991-1994), and LSTMs (1997). The 'RL-Attention' segment is highlighted in yellow. A small inset image of a man is visible in the bottom right corner of the slide.

1982 1986 1990 1991 1991-1994 1997 2014

Hopfield Jordan Elman RNN drawbacks RL-Attention LSTMs

Module 6

Again this idea of attention itself is something that was explored earlier also somewhere around 1991 or so and it was something known as Reinforcement Learning which was used for learning this attention mechanism. What attention basically tells you is that if you have a large sequence and if you want to do something with this sequence, what are the important entities of this sequence or elements of this sequence that you need to focus on, right. So, this is again something that we will look at in detail in the course.