CS7015 (Deep Learning) : Lecture 1

(Partial/Brief) History of Deep Learning

Mitesh M. Khapra

Department of Computer Science and Engineering
Indian Institute of Technology Madras

Acknowledgements

- Most of this material is based on the article "Deep Learning in Neural Networks: An Overview" by J. Schmidhuber[?]
- The errors, if any, are due to me and I apologize for them
- Feel free to write me if you think certain portions need to be corrected (please provide appropriate references)

Chapter 1: Biological Neurons

Reticular Theory

Joseph von Gerlach proposed that the nervous system is a single continuous network as opposed to a network of many discrete cells!

Staining Technique

Camillo Golgi discovered a chemical reaction that allowed him to examine nervous tissue in much greater detail than ever before

He was a proponent of Reticular theory.

Neuron Doctrine

Santiago Ramón y Cajal used Golgi's technique to study the nervous system and proposed that it is actually made up of discrete individual cells formimg a network (as opposed to a single continuous network)

The Term Neuron

The term neuron was coined by Heinrich Wilhelm Gottfried von Waldeyer-Hartz around 1891.

He further consolidated the Neuron Doctrine.

Nobel Prize

Both Golgi (reticular theory) and Cajal (neuron doctrine) were jointly awarded the 1906 Nobel Prize for Physiology or Medicine, that resulted in lasting conflicting ideas and controversies between the two scientists.

The Final Word

In 1950s electron microscopy finally confirmed the neuron doctrine by unambiguously demonstrated that nerve cells were individual cells interconnected through synapses (a network of many individual neurons).

Chapter 2: From Spring to Winter of AI

McCulloch Pitts Neuron

McCulloch (neuroscientist) and Pitts (logician) proposed a highly simplified model of the neuron (1943)[?]

Perceptron

"the perceptron may eventually be able to learn, make decisions, and translate languages" -Frank Rosenblatt

Perceptron

"the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence." -New York Times

First generation Multilayer Perceptrons

Ivakhnenko et. al.[?]

Perceptron Limitations

In their now famous book "Perceptrons", Minsky and Papert outlined the limits of what perceptrons could do[?]

AI Winter of connectionism

Almost lead to the abandonment of connectionist AI

Backpropagation

- Discovered and rediscovered several times throughout 1960's and 1970's
- Werbos(1982)[?] first used it in the context of artificial neural networks
- Eventually popularized by the work of Rumelhart et. al. in 1986[?]

Gradient Descent

Cauchy discovered Gradient Descent motivated by the need to compute the orbit of heavenly bodies

Universal Approximation Theorem

A multilayered network of neurons with a single hidden layer can be used to approximate any continuous function to any desired precision [?]

Chapter 3: The Deep Revival

Unsupervised Pre-Training

Hinton and Salakhutdinov described an effective way of initializing the weights that allows deep autoencoder networks to learn a low-dimensional representation of data. [?]

Unsupervised Pre-Training

The idea of unsupervised pre-training actually dates back to 1991-1993 (J. Schmidhuber) when it was used to train a "Very Deep Learner"

More insights (2007-2009)

Further Investigations into the effectiveness of Unsupervised Pre-training

Greedy Layer-Wise Training of Deep Networks

Why Does Unsupervised Pre-training Help Deep Learning?

Exploring Strategies for Training Deep Neural Networks

Success in Handwriting Recognition

Graves et. al. outperformed all entries in an international Arabic recognition competition[?]

Success in Speech Recognition

Dahl et. al. showed relative error reduction of 16.0% and 23.2% over a state of the art system [?]

New record on MNIST

Ciresan et. al. set a new record on the MNIST dataset using good old backpropagation on GPUs (GPUs enter the scene)[?]

First Superhuman Visual Pattern Recognition

D. C. Ciresan et. al. achieved 0.56% error rate in the IJCNN Traffic Sign Recognition Competition[?]

Winning more visual recognition challenges

Network	Error	Layers
AlexNet[?]	16.0%	8

Winning more visual recognition challenges

Network	Error	Layers
AlexNet [?]	16.0%	8
ZFNet [?]	11.2%	8

Winning more visual recognition challenges

Network	Error	Layers
AlexNet [?]	16.0%	8
ZFNet [?]	11.2%	8
VGGNet [?]	7.3%	19

Winning more visual recognition challenges

Network	Error	Layers
AlexNet [?]	16.0%	8
ZFNet [?]	11.2%	8
VGGNet [?]	7.3%	19
GoogLeNet [?]	6.7%	22

Winning more visual recognition challenges

Network	Error	Layers
AlexNet [?]	16.0%	8
ZFNet [?]	11.2%	8
VGGNet [?]	7.3%	19
GoogLeNet [?]	6.7%	22
MS ResNet [?]	3.6%	152!!

Chapter 4: Cats

Hubel and Wiesel Experiment

Experimentally showed that each neuron has a fixed receptive field - i.e. a neuron will fire only in response to a visual stimuli in a specific region in the visual space[?]

Neocognitron

Used for Handwritten character recognition and pattern recognition (Fukushima et. al.)[?]

Convolutional Neural Network

Handwriting digit recognition using back-propagation over a Convolutional Neural Network (LeCun et. al.)[?]

LeNet-5

Introduced the (now famous) MNIST dataset (LeCun et. al.)[?]

An algorithm inspired by an experiment on cats is today used to detect cats in videos :-)

Chapter 5: Faster, higher, stronger

Better Optimization Methods

Faster convergence, better accuracies

Better Optimization Methods

Faster convergence, better accuracies

Better Optimization Methods

Faster convergence, better accuracies

Better Optimization Methods

Faster convergence, better accuracies

Better Optimization Methods

Faster convergence, better accuracies

Better Optimization Methods

Faster convergence, better accuracies

Better Optimization Methods

Faster convergence, better accuracies

Chapter 6: The Curious Case of Sequences

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

Hopfield Network

Content-addressable memory systems for storing and retrieving patterns[?]

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

Jordan

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

Elman

Drawbacks of RNNs

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

Long Short Term Memory

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

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

RL for Attention

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

Beating humans at their own game (literally)

Playing Atari Games

- Human-level control through deep reinforcement learning for playing Atari Games

Let's GO

- Alpha Go Zero - Best Go player ever, surpassing human players
- GO is more complex than chess because of number of possible moves.
- No brute force backtracking unlike previous chess agents

Taking a shot at Poker

DeepStack defeated 11 professional poker players with only one outside the margin of statistical significance.

Defense of the Ancients

- Widely popular game, with complex strategies, large visual space
- Bot was undefeated against many top professional players .

Chapter 8: The Madness (2013-)

He sat on a <u>chair</u>.

Language Modeling

- Mikolov et al. (2010)[?]
- Li et al. (2015)
- Kiros et al. (2015)[?]
- Kim et al. (2015)[?]

Speech Recognition

- Hinton et al. (2012)[?]
- Graves et al. (2013)[?]
- Chorowski et al. (2015)[?]
- Sak et al. (2015)[?]

Machine Translation

- Kalchbrenner et al. (2013) [?]
- Cho et al. (2014) [?]
- Bahdanau et al. (2015) [?]
- Jean et al. (2015) [?]
- Gulcehre et al. (2015) [?]
- Sutskever et al. (2014) [?]
- Luong et al. (2015) [?]
- Zheng et al. (2017) [?]
- Cheng et al. (2016) [?]
- Chen et al. (2017) [?]
- Firat et al. (2016) [?]

Time	User	Utterance
03:44	Old	I dont run graphical ubuntu, I run ubuntu server.
03:45	kuja	Taru: Haha sucker.
03:45	Taru	Kuja: ?
03:45	bur[n]er	Old: you can use "ps ax" and "kill (PID#)"
03:45	kuja	Taru: Anyways, you made the changes right?
03:45	Taru	Kuja: Yes.
03:45	LiveCD	or killall speedlink
03:45	kuja	Taru: Then from the terminal type: sudo apt-get update
03:46	_pm	if i install the beta version, how can i update it when the final version comes out?
03:46	Taru	Kuja: I did.

Sender	Recipient	Utterance
Old		I dont run graphical ubuntu, I run ubuntu server.
bur[n]er	Old	you can use "ps ax" and "kill (PID#)"

Conversation Modeling

- Shang et al. (2015) [?]
- Vinyals et al. (2015) [?]
- Lowe et al. (2015) [?]
- Dodge et al. (2015) [?]
- Weston et al. (2016) [?]
- Serban et al. (2016) [?]
- Bordes et al. (2017) [?]
- He et al. (2017)
- Serban et al. (2017) [?]
- Lewis et al. (2017)

Task 1: Single Supporting Fact

Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts

John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts

John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations

The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Question Answering

- Hermann et al. (2015)[?]
- Chen et al. (2016)[?]
- Xiong et al. (2016)[?]
- Seo et al. (2016)[?]
- Dhingra et al. (2017)[?]
- Wang et al. (2017)[?]
- Hu et al. (2017)[?]

Object Detection/Recognition

- Semantic Segmentation (Long et al, 2015)[?]
- Recurrent CNNs (Liang et al., 2015)[?]
- Faster RCNN (Ren et al., 2015)[?]
- Inside-Outside Net (Bell et al., 2015)[?]
- YOLO9000 (Redmon et al., 2016)[?]
- R-FCN (Dai et al., 2016)[?]
- Mask R-CNN (He at al., 2017)[?]
- Video Object segmentation (Caelles et al., 2017)[?]

Visual Tracking

- Zhang et al. (2017)
- Choi et al. (2017)[?]
- Yun et al. (2017)[?]
- Luo et al. (2017)
- Alahi et al. (2017)[?]
- Van et al. (2016)

Retr.	1. Top view of the lights of a city at night, with a well-illuminated square in front of a church in the foreground; 2. People on the stairs in front of an illuminated cathedral with two towers at night;	1. Tourists are sitting at a long table with beer bottles on it in a rather dark restaurant and are raising their bierglaeser; 2. Tourists are sitting at a long table with a white table-cloth in a somewhat dark restaurant;
Gen.	A square with burning street lamps and a street in the foreground;	Tourists are sitting at a long table with a white table cloth and are eating;

Image Captioning

- Mao et al. (2014)[?]
- Mao at al. (2015)[?]
- Kiros et al. (2015)[?]
- Donahue et al. (2015)[?]
- Vinyals et al. (2015)[?]
- Karpathy et al. (2015)[?]
- Fang et al. (2015)[?]
- Chen et al. (2015)[?]

A group of young men playing a game of soccer

A man riding a wave on top of a surfboard.

Video Captioning

- Donahue et al. (2014)[?]
- Venugopalan at al. (2014)[?]
- Pan et al. (2015)[?]
- Yao et al. (2015)[?]
- Rohrbach et al. (2015)[?]
- Zhu et al. (2015)[?]
- Cho et al. (2015)[?]
- S. Sha 2017

What is the mustache made of?

AI System

bananas

Visual Question Answering

- Santoro et al. (2017)[?]
- Hu at al. (2017)[?]
- Johnson et al. (2017)[?]
- Ben-younes et al. (2017)[?]
- Malinowski et al. (2017)[?]
- Nam et al. (2016)
- Kazemi et al. (2016)[?]

She _____.

(nods)

She opens the _____.

(door)

Question: What is the cat doing? Answer: playing with a tablet

Video Question Answering

- Tapaswi et. al. 2016[?]
- Zeng et. al. 2016[?]
- Maharaj et. al. 2017[?]
- Zhao et. al. 2017[?]
- Yu Youngjae et. al. 2017[?]
- Xue Hongyang et. al. 2017[?]
- Mazaheri et. al. 2017[?]

Video Summarization

- Chheng 2007[?]
- Ajmal 2012[?]
- Zhang Ke 2016[?]
- Zhong Ji 2017[?]
- Panda 2017[?]

Generating Authentic Photos

- Variational Autoencoders (Kingma et. al., 2013)[?]
- Generative Adversarial Networks (Goodfellow et. al., 2014)[?]
- Plug & Play generative nets (Nguyen et al., 2016)[?]
- Progressive Growing of GANs (Karras et al., 2017)[?]

Generating Raw Audio

- Wavenets (Oord et. al., 2016)[?]

Pixel RNNs

- (Oord et al., 2016)[?]
- (Oord et al., 2016)[?]
- (Salimans et al., 2017)[?]

Chapter 9: (Need for) Sanity

The Paradox of Deep Learning

Why does deep learning work so well despite

[a] https://arxiv.org/pdf/1710.05468.pdf

The Paradox of Deep Learning

Why does deep learning work so well despite

- high capacity (susceptible to overfitting)

[a] https://arxiv.org/pdf/1710.05468.pdf

The Paradox of Deep Learning

Why does deep learning work so well despite

- high capacity (susceptible to overfitting)
- numerical instability (vanishing/exploding gradients)

[a] https://arxiv.org/pdf/1710.05468.pdf

The Paradox of Deep Learning

Why does deep learning work so well despite

- high capacity (susceptible to overfitting)
- numerical instability (vanishing/exploding gradients)
- sharp minima (leading to overfitting)

"panda"
57.7% confidence

"gibbon"
99.3% confidence

[a] https://arxiv.org/pdf/1710.05468.pdf

The Paradox of Deep Learning

Why does deep learning work so well despite

- high capacity (susceptible to overfitting)
- numerical instability (vanishing/exploding gradients)
- sharp minima (leading to overfitting)
- non-robustness (see figure)

[a] https://arxiv.org/pdf/1710.05468.pdf

The Paradox of Deep Learning

Why does deep learning work so well despite

- high capacity (susceptible to overfitting)
- numerical instability (vanishing/exploding gradients)
- sharp minima (leading to overfitting)
- non-robustness (see figure)

No clear answers yet but ...

[a] https://arxiv.org/pdf/1710.05468.pdf

The Paradox of Deep Learning

Why does deep learning work so well despite

- high capacity (susceptible to overfitting)
- numerical instability (vanishing/exploding gradients)
- sharp minima (leading to overfitting)
- non-robustness (see figure)

No clear answers yet but ...

- Slowly but steadily there is increasing emphasis on explainability and theoretical justifications! [a]

[a] https://arxiv.org/pdf/1710.05468.pdf

The Paradox of Deep Learning

Why does deep learning work so well despite

- high capacity (susceptible to overfitting)
- numerical instability (vanishing/exploding gradients)
- sharp minima (leading to overfitting)
- non-robustness (see figure)

No clear answers yet but ...

- Slowly but steadily there is increasing emphasis on explainability and theoretical justifications! [a]
- Hopefully this will bring sanity to the proceedings !

[a] https://arxiv.org/pdf/1710.05468.pdf

https://github.com/kjw0612/awesome-rnn

[i]Source: https://www.cbinsights.com/blog/deep-learning-ai-startups-market-map-company-list/

References I