Deep Learning

00 Introduction and Course logistics

Dr. Konda Reddy Mopuri
Dept. of AI, IIT Hyderabad
Jan-May 2024

Time slot

- B slot

Time slot

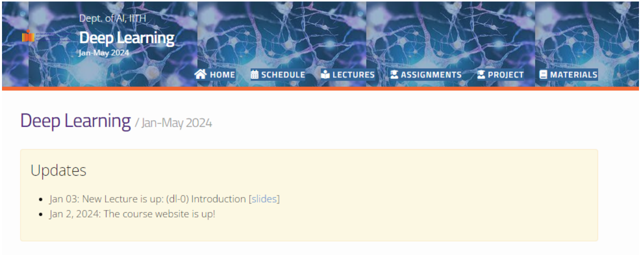
- B slot
- Monday 10 - 10:55 AM
- Wednesday 9 - 9:55 AM
- Thursday 11 - 11:55 AM

Time slot

- B slot
- Monday 10 - 10:55 AM
- Wednesday 9 - 9:55 AM
- Thursday 11 - 11:55 AM
- ALH-1

Logistics

- Course website: `https://krmopuri.github.io/dl24/`



Evaluation

- Programming Assignments - 40% (best 4 of 5; 1 for each of the first 5 segments)
- Project - 20%
- Viva - 20%
- Written exams (best 4 of 5 surprise tests) - 20%

TAs

- Susmit Agrawal (ai22mtech12002@iith.ac.in)
- Rupa Kumari (ai22mtech11002@iith.ac.in)
- Deepika Vemuri (ai22resch11001@iith.ac.in)
- Savarana Datta Reddy (ai20btech11008@iith.ac.in)
- Some more coming up!

Contents

- Broadly: Building blocks of the Deep Learning based solutions

Contents

- Broadly: Building blocks of the Deep Learning based solutions
- Artificial Neuron $\rightarrow$ Generative AI

Contents

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Deep Learning (AI5100) Course Contents

Starting from an artificial neuron model, the aim of this course is to understand feed-forward, recurrent architectures of Artificial Neural Networks, all the way to the latest Generative AI models driven by Deep Neural Networks. Specifically, we will discuss the basic Neuron models (McCulloch Pitts, Perceptron), Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN, LSTM and GRU). We will understand these models' representational ability and how to train them using the Gradient Descent technique using the Backpropagation algorithm. We will then discuss the encoder-decoder architecture, attention mechanism and its variants. That will be followed by self-attention and Transformers. The next part of the course will be on Generative AI, wherein we will discuss Variational Autoencoders, GANs, Diffusion Models, GPT, BERT, etc. We will briefly discuss multi-modal representation learning (e.g., CLIP). Towards the end, students will be briefly exposed to some of the advanced topics and/or recent trends in deep learning.

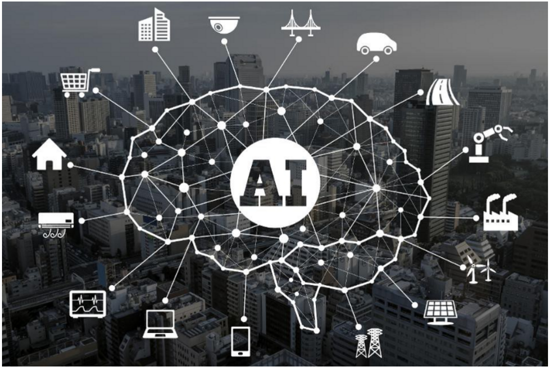
Prerequisites

- Programming in Python (Primer on PyTorch on 13 January, 10 AM - 1 PM in ALH-1)

Prerequisites

- Programming in Python (Primer on PyTorch on 13 January, 10 AM - 1 PM in ALH-1)
- A course on Machine Learning

Why Deep Learning?



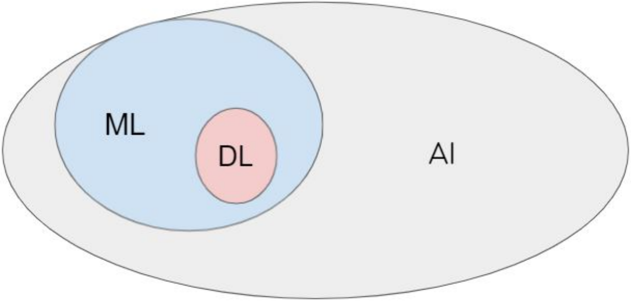
Deep Learning drives the recent AI boom.
Image Source: Artificial Intelligence Magazine

Textbooks and References

- Lot of online resources
  - Michael Nielsen's text book on NN & DL
  - NPTEL course on Deep Learning by Prof. Mitesh Khapra, IITM
  - DL course by François Fleuret, Uni. of Geneva
  - Deep Learning textbook by Ian Goodfellow *et al.*
  - PyTorch - https://pytorch.org/
  - Many more that I could not list and am not aware of...

What is DL?

What is DL?



Subset of ML that is essentially Artificial Neural Networks with more layers

What is DL?

- Crude attempt to imitate the human brain in learning

Classical ML vs. DL

- Classical ML: Handcrafted features + learnable model
- Need strong domain expertise

Classical ML vs. DL

- Classical ML: Handcrafted features + learnable model
- Need strong domain expertise

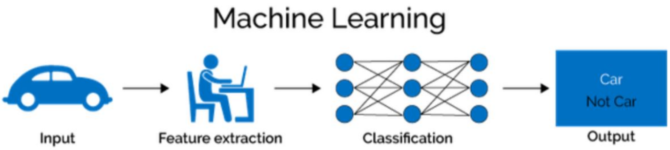


Figure credits: taken from Jay Shaw in Quora (not sure of authenticity)

Classical ML vs. DL

- Deep Learning: Deep stack of parameterized processing
- End-to-End learning

Classical ML vs. DL

- Deep Learning: Deep stack of parameterized processing
- End-to-End learning

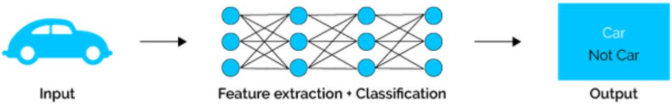


Figure credits: taken from Jay Shaw in Quora (not sure of authenticity)

Classical ML vs. DL

- ANNs predate some of the classical ML techniques

Classical ML vs. DL

- ANNs predate some of the classical ML techniques
- We are now dealing with a new generation ANNs

The Biological Neuron

- About 100 billion neurons in human brain

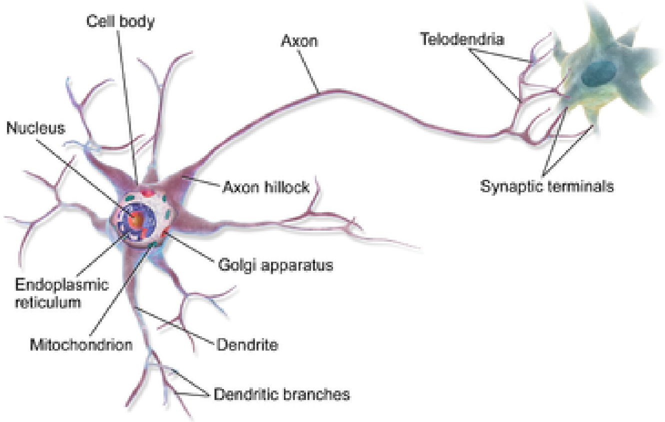


Figure credits: Wikipedia

History of Neural Networks

1. McCulloch Pitts neuron (1943) - Threshold Logic Unit

History of Neural Networks

1. McCulloch Pitts neuron (1943) - Threshold Logic Unit
2. Donald Hebb (1949) - Hebbian Learning Principle

History of Neural Networks

1. McCulloch Pitts neuron (1943) - Threshold Logic Unit
2. Donald Hebb (1949) - Hebbian Learning Principle
3. Marvin Minsky (1951) - created the first ANN (Hebbian Learning, 40 neurons)

History of Neural Networks

1. McCulloch Pitts neuron (1943) - Threshold Logic Unit
2. Donald Hebb (1949) - Hebbian Learning Principle
3. Marvin Minsky (1951) - created the first ANN (Hebbian Learning, 40 neurons)
4. Frank Rosenblatt (1958) - created perceptron to classify 20X20 images

History of Neural Networks

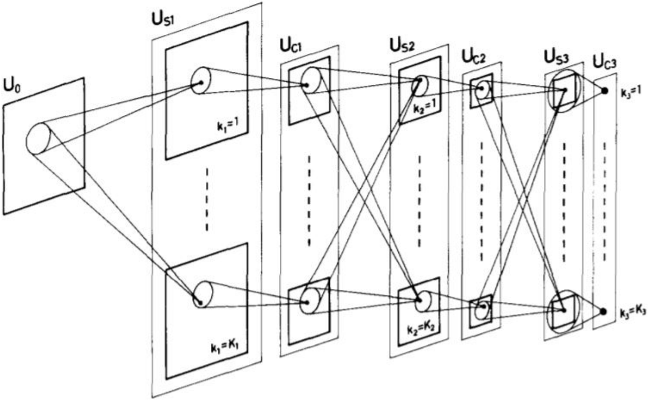
1. McCulloch Pitts neuron (1943) - Threshold Logic Unit
2. Donald Hebb (1949) - Hebbian Learning Principle
3. Marvin Minsky (1951) - created the first ANN (Hebbian Learning, 40 neurons)
4. Frank Rosenblatt (1958) - created perceptron to classify 20X20 images
5. David H Hubel and Torsten Wiesel (1959) demonstrated orientation selectivity and columnar organization in cats visual cortex

Backpropagation

- Paul Werbos (1982) proposed back-propagation for ANNs

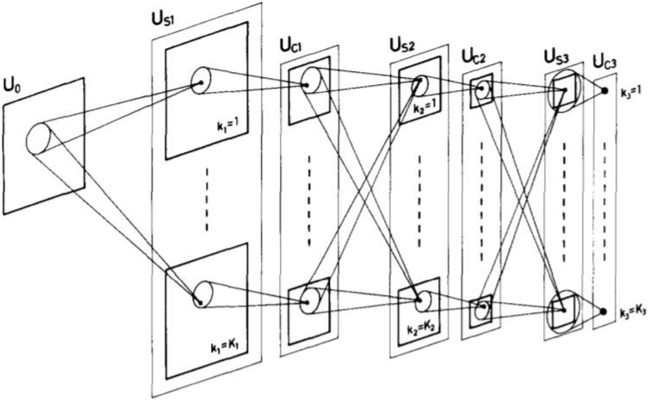
History (contd.)

1. Neocognitron by Fukushima (1980)



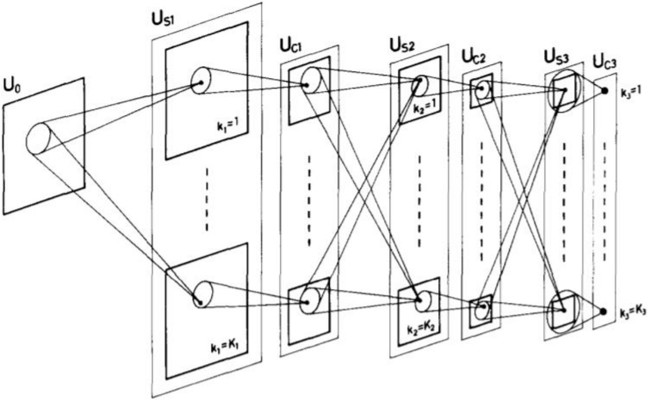
History (contd.)

1. Neocognitron by Fukushima (1980)
2. Implements the Hubel and Wiesel's principles



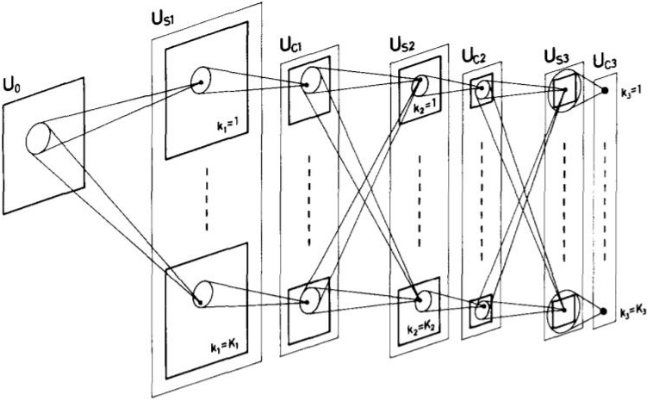
History (contd.)

1. Neocognitron by Fukushima (1980)
2. Implements the Hubel and Wiesel's principles
3. Used for hand-written digit recognition



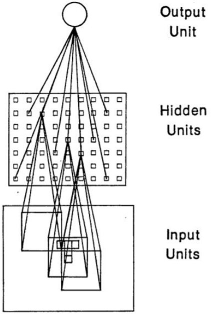
History (contd.)

1. Neocognitron by Fukushima (1980)
2. Implements the Hubel and Wiesel's principles
3. Used for hand-written digit recognition
4. Viewed as precursor for the modern CNNs



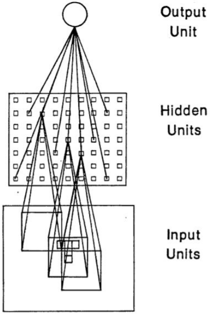
History (contd.)

1. Rumelhart (1986) trained with backprop



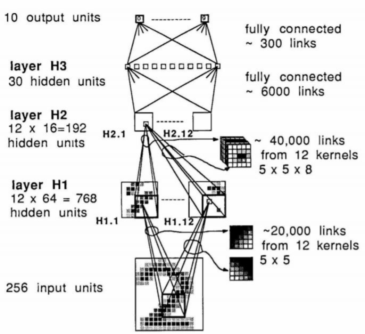
History (contd.)

1. Rumelhart (1986) trained with backprop
2. Showed that hidden units learn meaningful representations



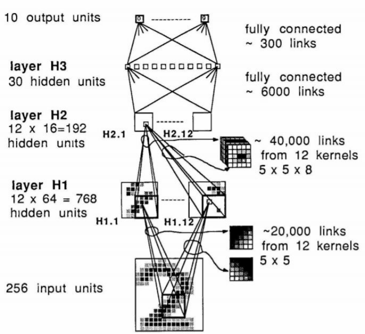
History (contd.)

1. LeNet family (Lecun et al. 1989) is a “convent”



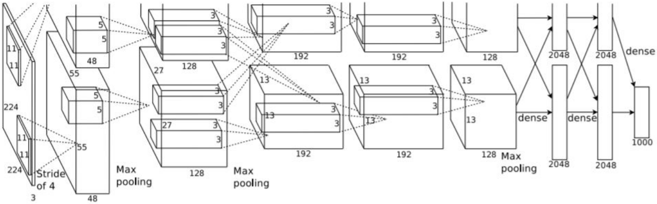
History (contd.)

1. LeNet family (Lecun et al. 1989) is a “convent”
2. Very similar to modern architectures



History (contd.)

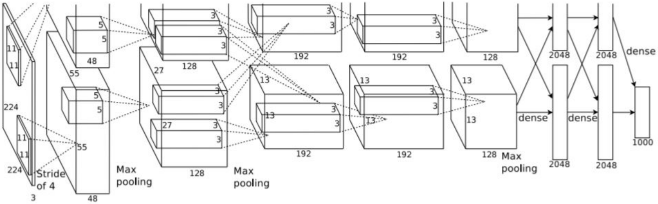
1. AlexNet (2012)



History (contd.)

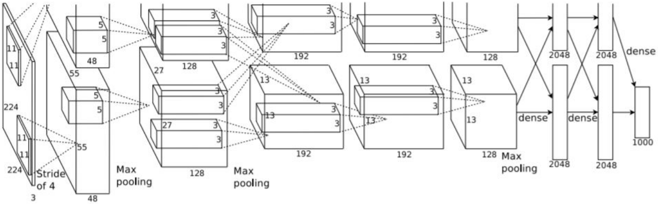
భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

1. AlexNet (2012)
2. Network similar to LeNet-5, but of far greater size



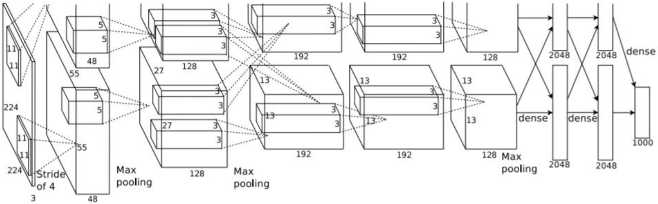
History (contd.)

1. AlexNet (2012)
2. Network similar to LeNet-5, but of far greater size
3. Implemented using GPUs



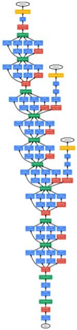
History (contd.)

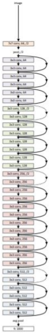
1. AlexNet (2012)
2. Network similar to LeNet-5, but of far greater size
3. Implemented using GPUs
4. Could beat the SoTA image classification methods by a large margin



History (contd.)

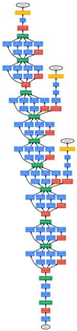
1. AlexNet initiated a trend of more complex and bigger architectures

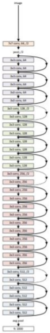




History (contd.)

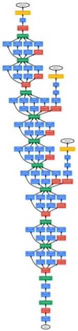
1. AlexNet initiated a trend of more complex and bigger architectures
2. GoogLeNet (2015) contains "inception" modules

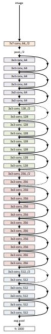




History (contd.)

1. AlexNet initiated a trend of more complex and bigger architectures
2. GoogLeNet (2015) contains "inception" modules
3. ResNet (2015) introduced "skip connections" that facilitate training deeper architectures





History (contd.)

1. Transformers (2017) are attention-based architectures

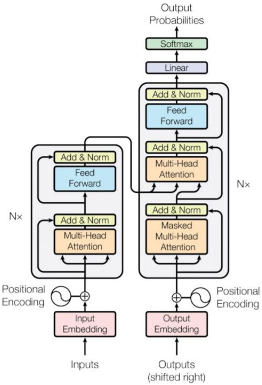


Figure credits: Vaswani et al., 2017

History (contd.)

1. Transformers (2017) are attention-based architectures
2. Very popular in NLP, and CV

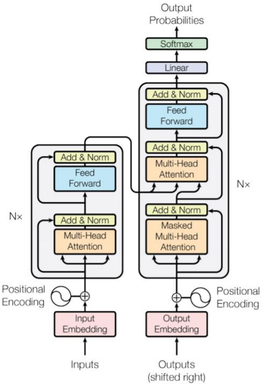


Figure credits: Vaswani et al., 2017

History (contd.)

1. Transformers (2017) are attention-based architectures
2. Very popular in NLP, and CV
3. Some of these models are extremely large (e.g., GPT-3 has 175B, PaLM has 540B parameters, etc.)

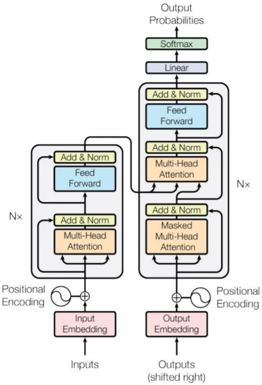


Figure credits: Vaswani et al., 2017

Deep Learning

1. Natural generalization to ANNs - Doesn't differ much from the 90s NNs

Deep Learning

1. Natural generalization to ANNs - Doesn't differ much from the 90s NNs
2. Computational graph of tensor operations that take advantage of
    - Chain rule (back-propagation)
    - SGD
    - GPUs
    - Huge datasets
    - Convolutions, attention, self-attention, etc.

ILSVRC Error

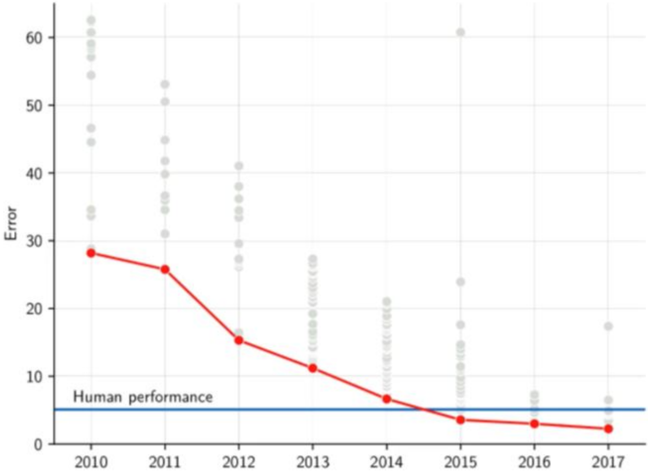


Figure credits: Gershgorn, 2017

LLM performance on the MMLU benchmark

Figure credits: W. Zi, L. El Asri, S. Prince

What makes it work now?

What makes it work now?

1. Huge research and progress in ML

What makes it work now?

1. Huge research and progress in ML
2. Hardware developments - CPUs/GPUs/Storage technologies

What makes it work now?

1. Huge research and progress in ML
2. Hardware developments - CPUs/GPUs/Storage technologies
3. Piles of data over the Internet

What makes it work now?

1. Huge research and progress in ML
2. Hardware developments - CPUs/GPUs/Storage technologies
3. Piles of data over the Internet
4. Collaborative development (open source tools and fora for sharing/discussions, etc.)

What makes it work now?

1. Huge research and progress in ML
2. Hardware developments - CPUs/GPUs/Storage technologies
3. Piles of data over the Internet
4. Collaborative development (open source tools and fora for sharing/discussions, etc.)
5. Collective efforts from large institutions/corporations

What makes it work now?

1. Huge research and progress in ML
2. Hardware developments - CPUs/GPUs/Storage technologies
3. Piles of data over the Internet
4. Collaborative development (open source tools and fora for sharing/discussions, etc.)
5. Collective efforts from large institutions/corporations
6. ...

Deep Learning - practical perspective

1. Doesnt require a deep mathematical grasp

Deep Learning - practical perspective

1. Doesnt require a deep mathematical grasp
2. Makes the design of large models a system/software development task

Deep Learning - practical perspective

1. Doesnt require a deep mathematical grasp
2. Makes the design of large models a system/software development task
3. Leverages modern hardware

Deep Learning - practical perspective

1. Doesnt require a deep mathematical grasp
2. Makes the design of large models a system/software development task
3. Leverages modern hardware
4. Doesnt seem to plateau with more data

Deep Learning - practical perspective

1. Doesnt require a deep mathematical grasp
2. Makes the design of large models a system/software development task
3. Leverages modern hardware
4. Doesnt seem to plateau with more data
5. Makes the trained models a commodity

Compute getting cheaper

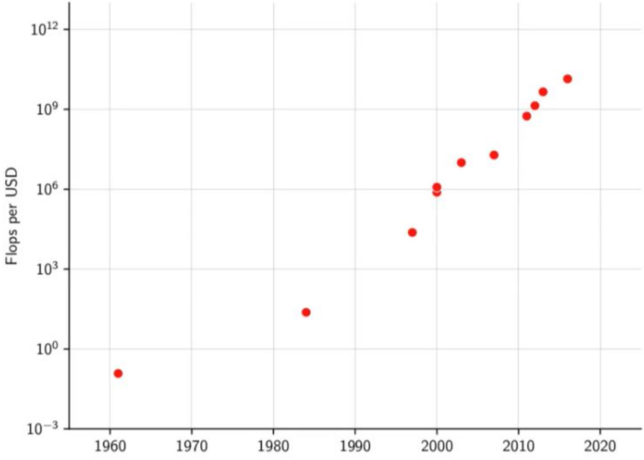


Figure Credits: Wikipedia

Storage getting cheaper

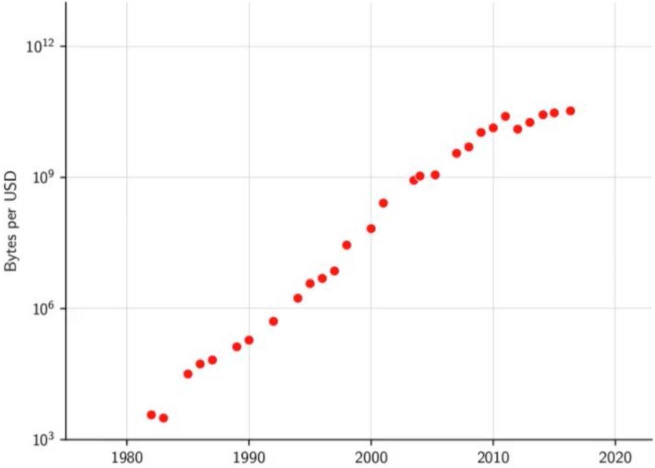


Figure Credits: John C Mccallum

AlexNet to AlphaGo: 300000X increase in compute

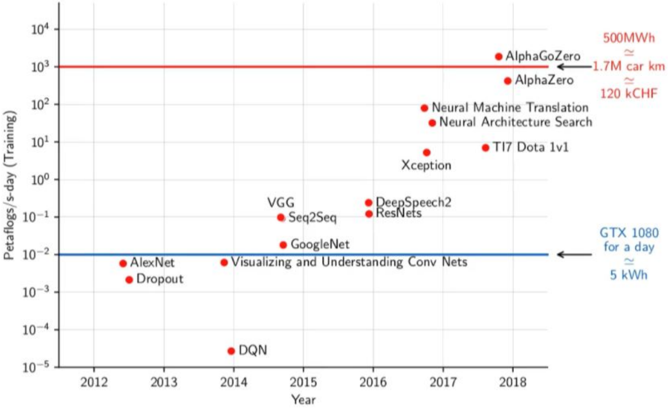


Figure Credits: Radford, 2018. 1 petaflop/s-day ≈ 100 GTX 1080 GPUs for a day, ≈ 500kwh

LLM compute

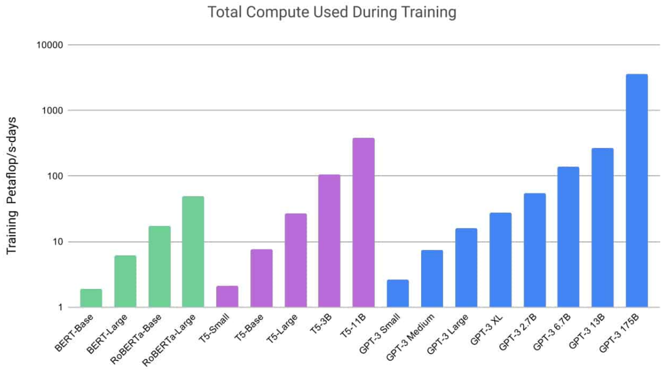


Figure Credits: NVIDIA blog

Datasets

Data-set		Year	Nb. images	Size
MNIST	(classification)	1998	60K	12Mb
Caltech 101	(classification)	2003	9.1K	130Mb
Caltech 256	(classification)	2007	30K	1.2Gb
CIFAR10	(classification)	2009	60K	160Mb
ImageNet	(classification)	2012	1.2M	150Gb
MS-COCO	(segmentation)	2015	200K	32Gb
Cityscape	(segmentation)	2016	25K	60Gb

Data-set		Year	Size
SST2	(sentiment analysis)	2013	20Mb
WMT-18	(translation)	2018	7Gb
OSCAR	(language model)	2020	6Tb

Figure Credits: François Fleuret

Datasets

- GPT-3 uses 45TB of text data for training

Implementation

	Language(s)	License	Main backer
PyTorch	**Python**, C++	BSD	Facebook
TensorFlow	Python, C++	Apache	Google
JAX	Python	Apache	Google
MXNet	Python, C++, R, Scala	Apache	Amazon
CNTK	Python, C++	MIT	Microsoft
Torch	Lua	BSD	Facebook
Theano	Python	BSD	U. of Montreal
Caffe	C++	BSD 2 clauses	U. of CA, Berkeley

Figure Credits: François Fleuret

References

- Please visit lectures tab in the course website for the full list of references
- Please share your comments/suggestions/any errors (technical or references) with the instructor (krmopuri@ai.iith.ac.in)
- Thank You!