Demystifying Deep Learning (Especially ChatGPT)

Prof Richard Xu

Department of Mathematics, HKBU

2023-02-27

AI history up to 2017

HKBU :: MATH :: Prof Xu

Antiquity

In Greek mythology, Talos was a giant automaton made of bronze to protect Europa


Three Laws of Robotics

1950

Asimov introduced the Three Laws of Robotics

1956

Arthur Samuel (IBM) wrote the first game-playing program for checkers

John McCarthy coins "Artificial Intelligence" for the first time

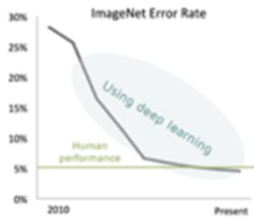
1980

Generation of entrepreneurs believed AI was the wave of the future

1985

Computer Scientists pioneer highly effective ML method

'AI Winter'



2010

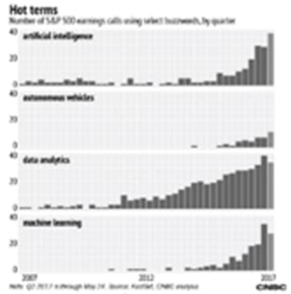
ImageNet Challenge

2015

Google DeepMind's AlphaGo defeats European Go champion

2016

Microsoft's AI can now understand speech better than humans



2017

AI 'everywhere' and accelerating

Image credit: VMware Xplorer Group

Deep Learning Recent Development timeline

HKBU :: MATH :: Prof Xu

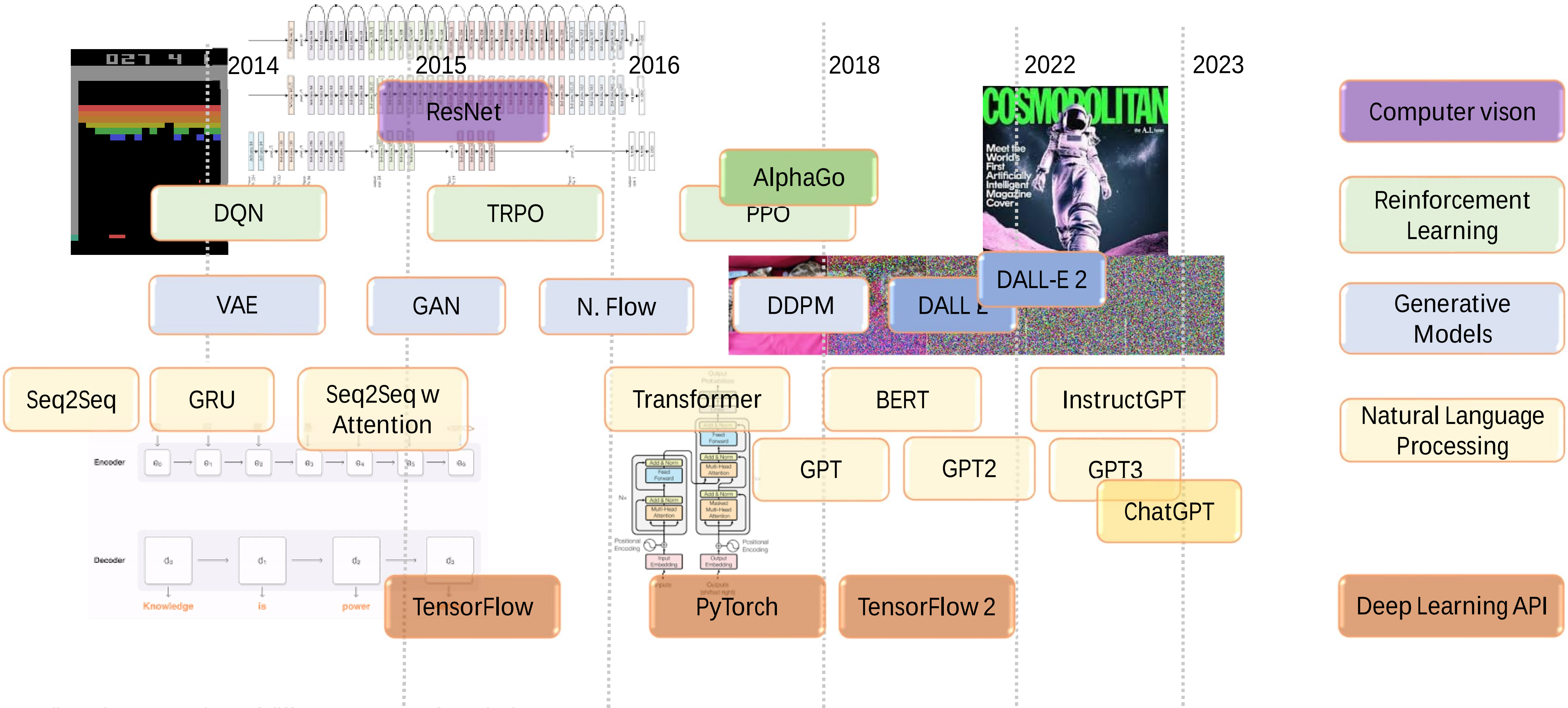
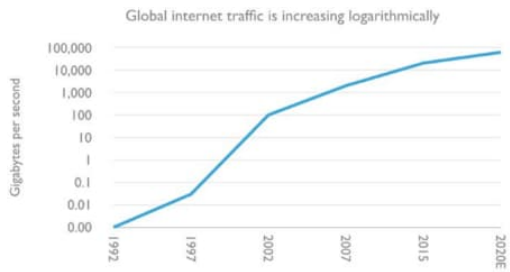


Image credit: various screenshots of different system are from the internet

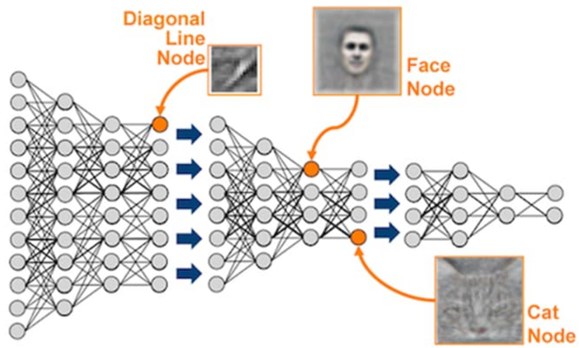
AI and Big Data: other ingredients

HKBU :: MATH :: Prof Xu

Cloud accessibly for Big Data



Better algorithms



Much powerful GPU

Image credit: David Kelnar at Medium.com and NVIDIA

Buzzwords: AI, ML & DL

HKBU :: MATH :: Prof Xu

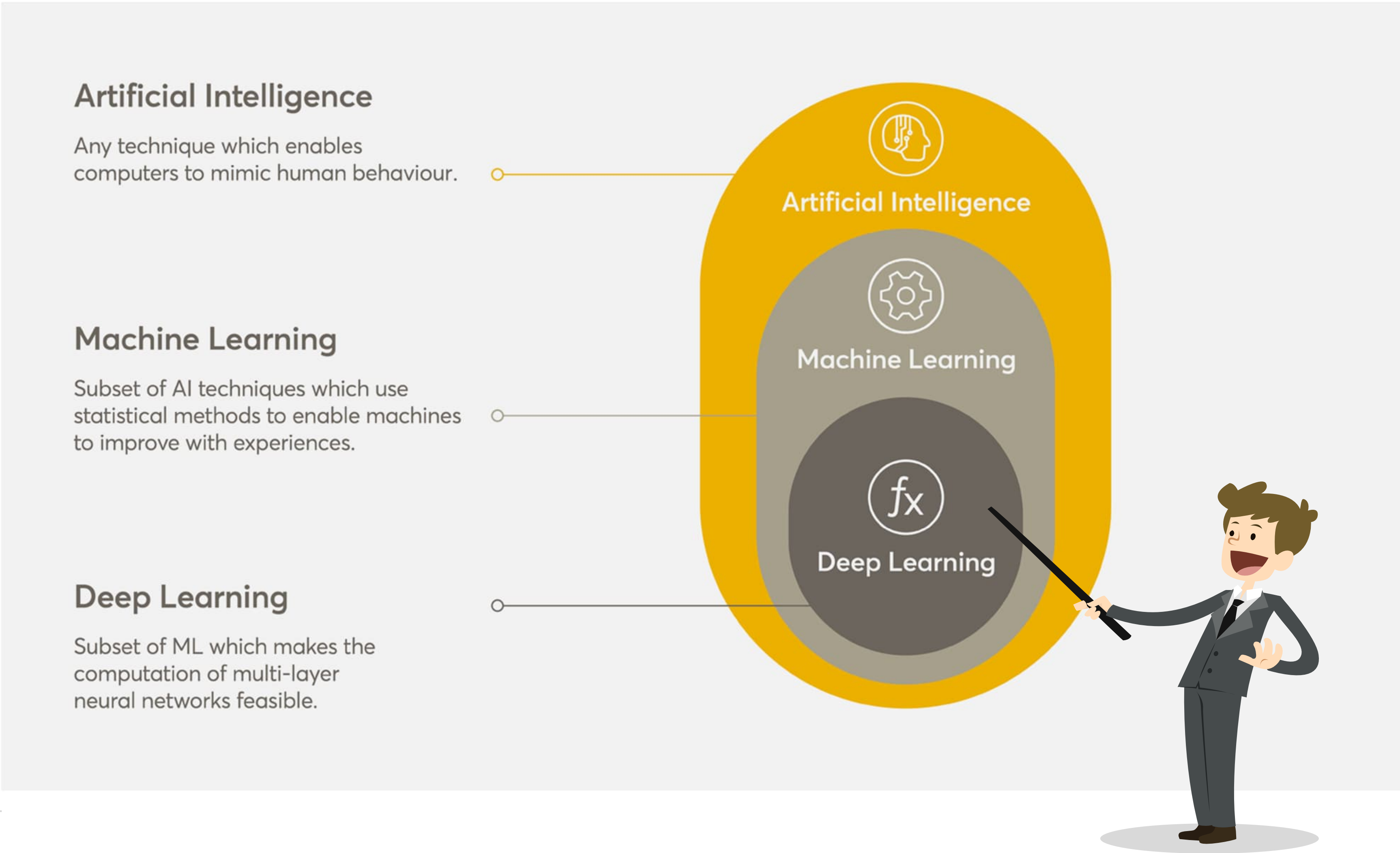
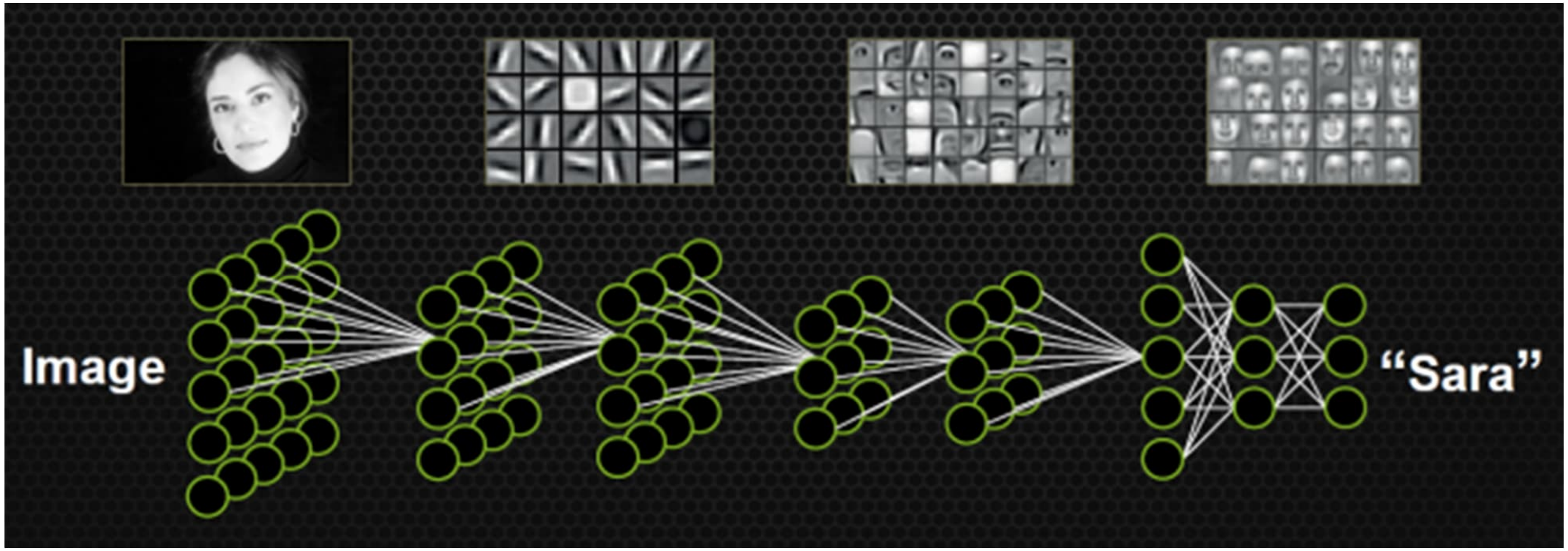


Image credit: NASTEL

Early Deep Learning: Convolutional Neural Network:



More details are found on my Machine Learning notes:
https://github.com/roboticcam/machine-learning-notes/blob/master/cnn_beyond.pdf

Image credit: NVIDIA

Early Deep Learning: Recurrent Neural Network

HKBU :: MATH :: Prof Xu

Elaborate at chatGPT

Image credit: Google Research

02 | Current Deep Learning technologies

- Computer vison
- Reinforcement Learning
- Generative Models
- Natural Language Processing
- Deep Learning API

DL technologies: before we start…

Take home Message

- ML is not that far away from University research as you might think
- Still a lot of research needs to be done
- I select ML applications that I have worked: so you can relate to it easily
- Too many DL areas, I only concentrate on a few

Let’s start

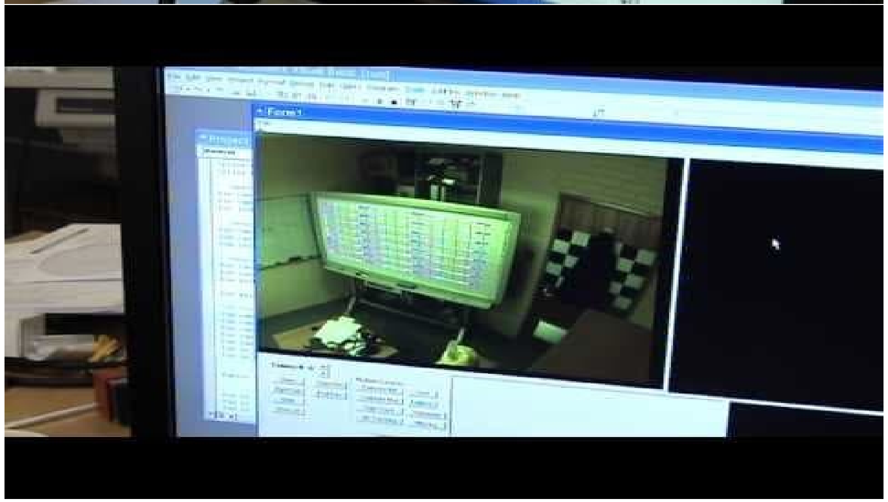
DL technologies: 3D computer vision

HKBU :: MATH :: Prof Xu

Computer vison

Previous work without DL

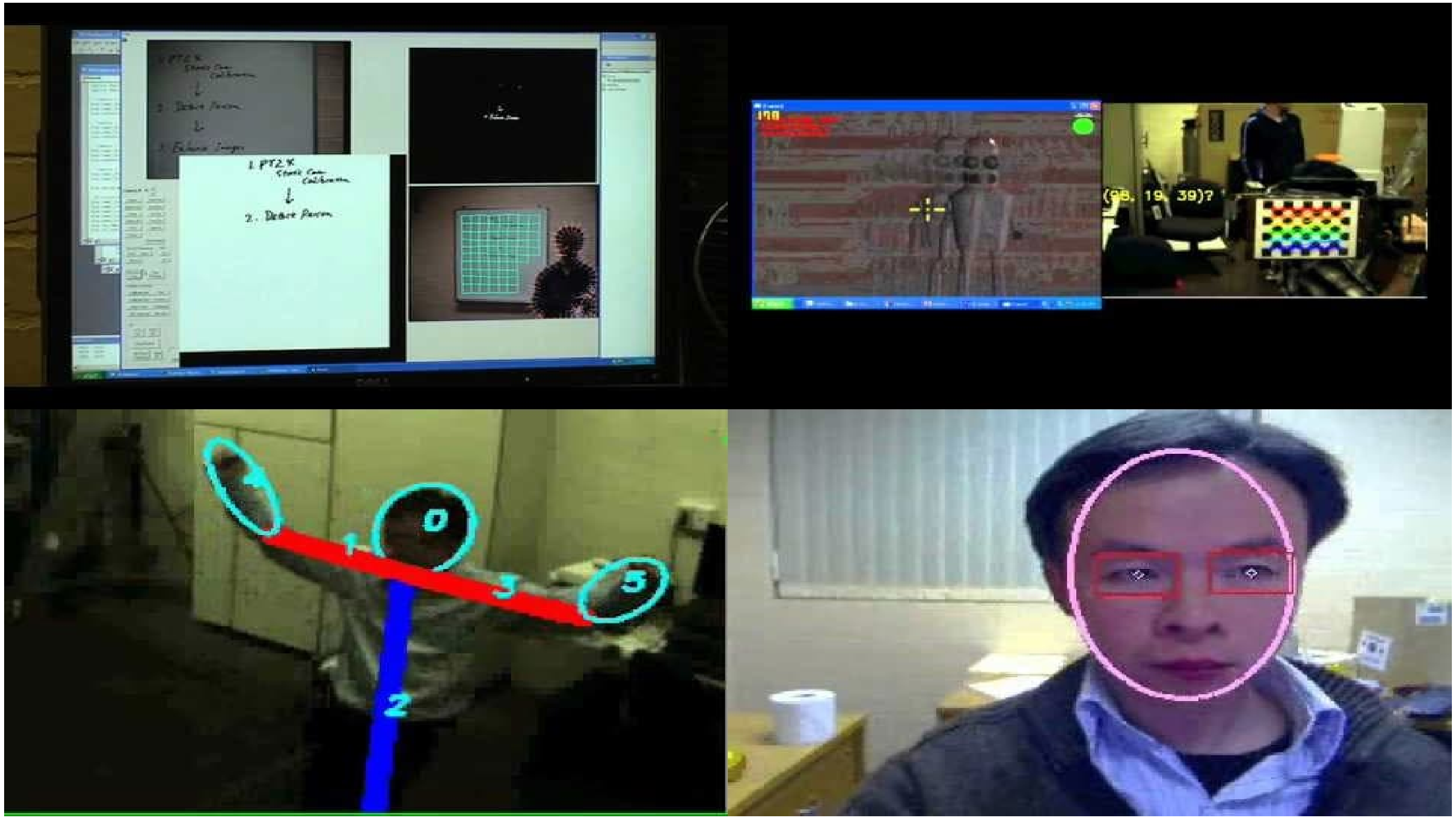
How do I coordinate cameras?



DL technologies: 3D computer vision

Computer vison

Previous pre-Deep Learning research



DL technologies: 3D computer vision

HKBU :: MATH :: Prof Xu

Computer vison

Newer Research with Deep Learning

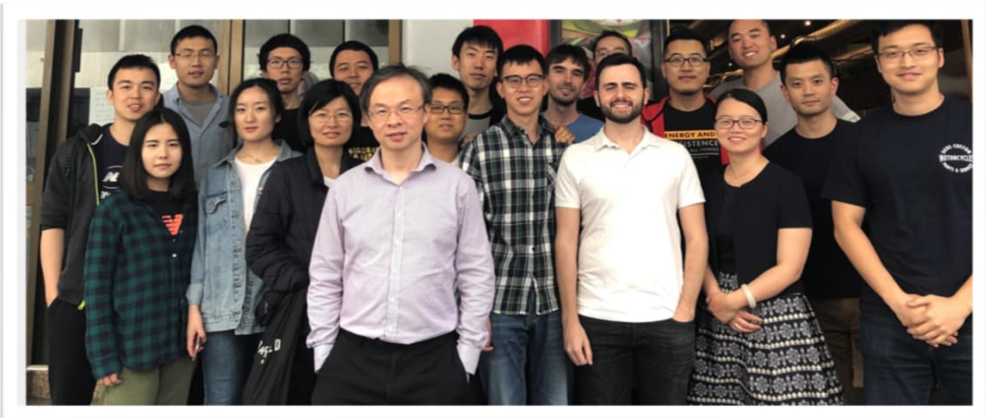
DL technologies: 3D computer vision

Computer vison

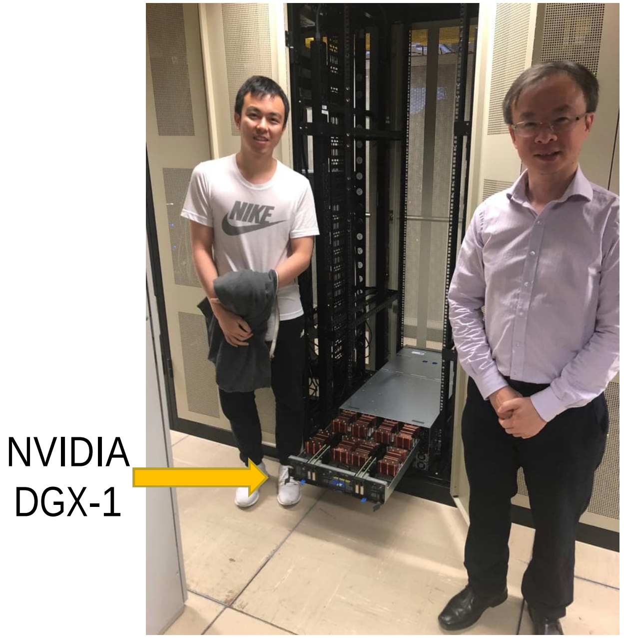
Just DL technology is not enough

We also need:

- Many talented PhD students and engineers



- Lots of high-end GPUs



DL technologies: language model

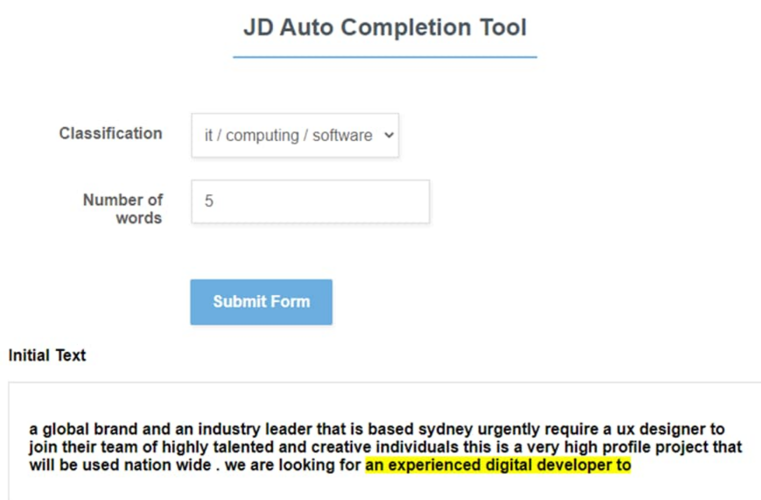
HKBU :: MATH :: Prof Xu

Natural Language Processing

Present in NLP:

- ChatGPT is dominating the news!
- In essence, it is a huge language model

What we have done



- Anyone can train their own (tailored) language model

Future in NLP:

- Produce the world's best selling (in cryptocurrency maybe) novel

DL technologies: other NLP

Natural Language Processing

Chatbot:

- Google Duplex

Image credit: Buzzfeed

What we have done

- My former PhD student's research in Aspect-based sentiment analysis

"The décor (negative) is not special at all but their amazing food (positive) makes up for it"

Future AI-chatbot:

- Chat naturally with a super-knowledgeable human – chatGPT is just the start!

DL technologies: cross domain translation

HKBU :: MATH :: Prof Xu

Generative Models

Present:

- DALL-E 2

Future:

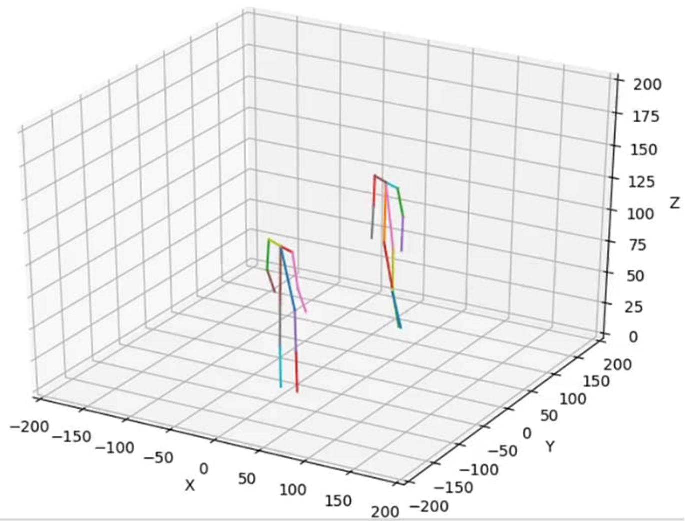
- AI create full-feature film

Image credit: Times https://time.com/collection/best-inventions-2022/6225486/dall-e-2/

What we have done

- My teams' work ins script to animation

I meet a friend, check my watch and suggest we leave



Generative Models

Present AI music:

- SONY CSL Research Lab created first-ever entire songs composed by robot: Daddy's Car
- Google's AI song

What we have done

- Two hand's piano play form one-hand play input
  - One hand
  - Left hand + Right hand

Future AI music:

- AI compose a set of viral songs, i.e., AlanWalkerBot

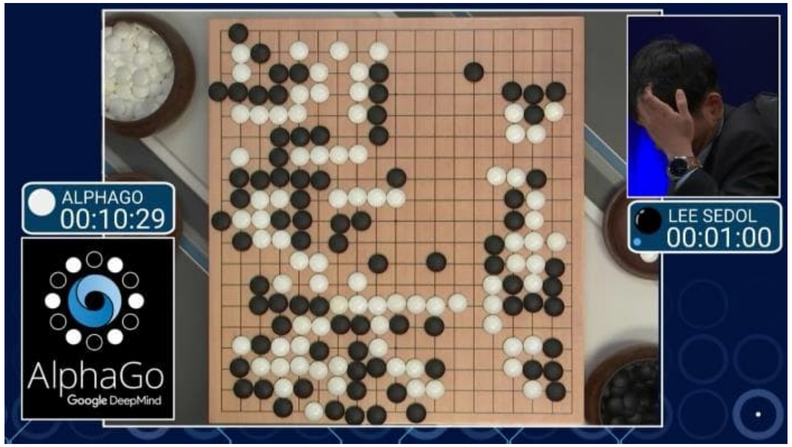
DL technologies: RTS strategy

HKBU :: MATH :: Prof Xu

Reinforcement Learning

Present AI-powered gaming

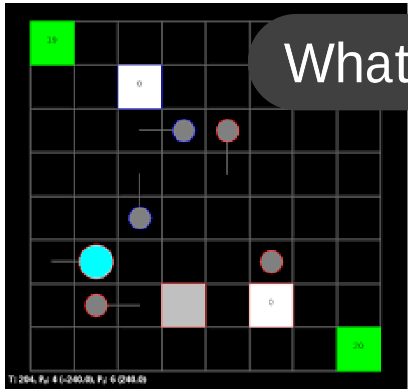
- Google DeepMind AlphaGo



Future AI-powered gaming:

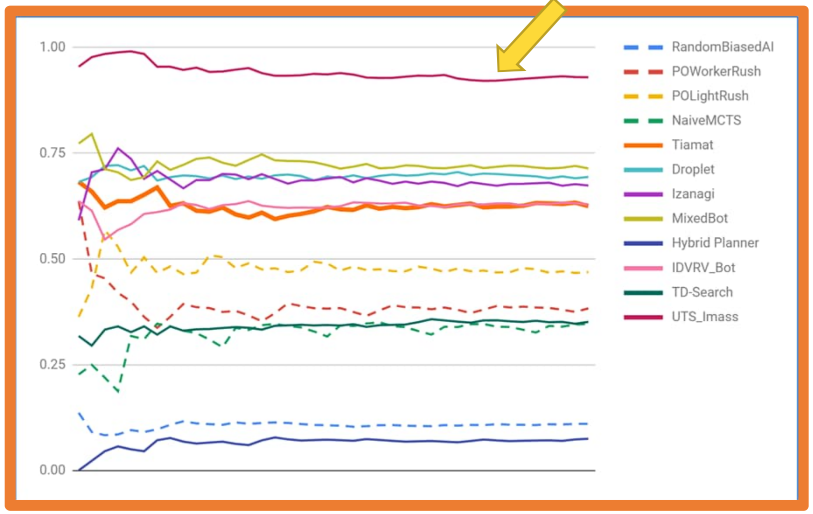
- DeepMind alphaStar beat best human player in StarCraft

What we have done



- My team's work in MicroRTS competition (2019 results)

My team @UTS



DL technologies: some background on RL

HKBU :: MATH :: Prof Xu

Supervised Learning

Human can easily label this image as "elephant"

- Has instructional label
- Data is usually all available for training

Elaborate at chatGPT

Reinforcement Learning

Human cannot easily label this image as "win"

- only reward signal
- feedback is delayed, i.e., not instantaneous
- data are not i.i.d., (consecutive frames are similar)
- agent's actions affect subsequent data it receives
- find the policy most likely to maximize expected future returns by taking current actions

Image credit: https://a-z-animals.com/animals/elephant/

DL & AI in real world:

My industry focused research (Australia)

6 years, 14 companies, 24 individual AI projects developing following technologies:

- Product range modelling
- Customer segmentation
- Predictive inventory/maintenance
- Recommendation system
- Personalized email/campaign
- Product correlations
- Transport passenger prediction
- HR performance indictor prediction
- Sentiment analysis
- ...

- Smart city applications
- CityRail Bus replacement optimization
- Injury prediction
- Peer-to-peer lending ranking
- Process optimization
- Education-to-job alignment
- Real-time strategy for defense
- Network traffic modelling
- ...

Talent hunting in Data Mining

HKBU :: MATH :: Prof Xu

03

Finally... chatgpt

intuitive explanation

Natural Language Processing

Difference between

vs *“there is a unicorn standing at the river bank”*

image *piece of natural language*

- Image is comprised by pixels, which are already in numeric form
- Natural language comprised of words; they are nominal, incomparable tokens
- Therefore, first step is convert word tokens to numerical representations.

Natural Language Processing

Easiest encoding

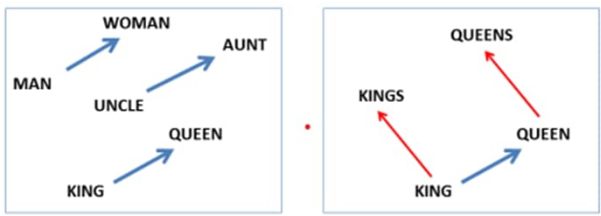
$$
\begin{array}{c}
\text{“}a\text{”} \\
\text{“}abbreviate\text{”} \\
\vdots \\
\text{“}zoology\text{”}
\end{array}
\begin{bmatrix}
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 1
\end{bmatrix}
$$

- structure is huge and sparse.
- as long as total number of words in vocabulary
- every pair of words are exactly $\sqrt{2}$ apart. cannot measure similarity or dissimilarity between them
- Beautiful visualization: https://projector.tensorflow.org/

word2vec encoding

Mikolov et. al., (2013)

- Find numerical representation of each word such that their probability conditional on their neighbourhood is maximized.
- In an effort that words have similar representations if their neighbourhood are similar.
  - a very “*nice*” picture
  - a very “*beautiful*” picture
- Then, possible to perform arithmetic on words:



Deep Natural Language Process word2vec example

HKBU :: MATH :: Prof Xu

" the cat sit on the mat" window = 3

Step 1: find all windows

(a) "the", "cat", "sit", target: cat

(b) "cat", "sit", "on" , target: sit

(c) "sit", "on ", "the", target: on

(d) "on" , "the", "mat", target: the

Step 2: find all context target pairs

$$
\begin{array}{rcl}
\text{"the"} & \leftarrow & \text{"cat"} \\
\text{"sit"} & \leftarrow & \text{"cat"} \\
\text{"cat"} & \leftarrow & \text{"sit"} \\
\text{"on"} & \leftarrow & \text{"sit"} \\
\text{"sit"} & \leftarrow & \text{"on"} \\
\text{"the"} & \leftarrow & \text{"on"} \\
\text{"on"} & \leftarrow & \text{"the"} \\
\text{"mat"} & \leftarrow & \text{"the"}
\end{array}
$$

Step 3: maximize probabilities:

$$
\Pr(\text{"the"}|\text{"cat"}) \times \Pr(\text{"sit"}|\text{"cat"}) \times \Pr(\text{"cat"}|\text{"sit"}) \times \Pr(\text{"on"}|\text{"sit"}) \times \\
\Pr(\text{"sit"}|\text{"on"}) \times \Pr(\text{"the"}|\text{"on"}) \times \Pr(\text{"on"}|\text{"the"}) \times \Pr(\text{"mat"}|\text{"the"})
$$

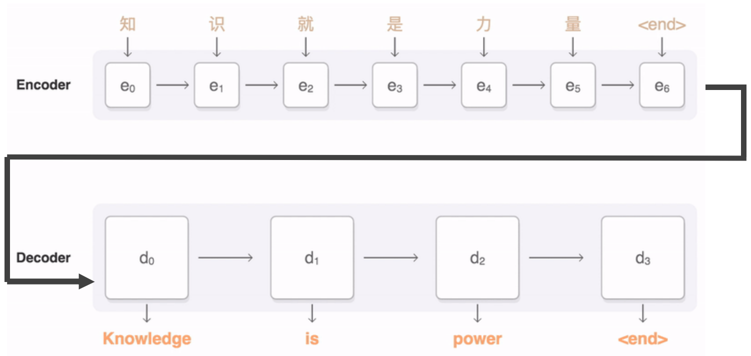
Deep Natural Language Process attention

HKBU :: MATH :: Prof Xu

Natural Language Processing

Pure Seq2Seq

Sutskever et al., (2014)



- Uses bi-directional Recurrent Neural Network
- Translation analogy:
  A person first reads input from left to right (and right to left) to understand entire sentence. He then translate word for word into the output. The translation of each output word depends on the preceding translated words.

Seq2Seq with Attention

Luong et al., (2015)

input sentence

output sentence

- added attention so elements of decoder can access individual elements of encoder directly.
- Translation analogy:
  In addition to fully understanding input, when translating each word for output, the person also considers the amount of contribution (attention) from each individual input word.

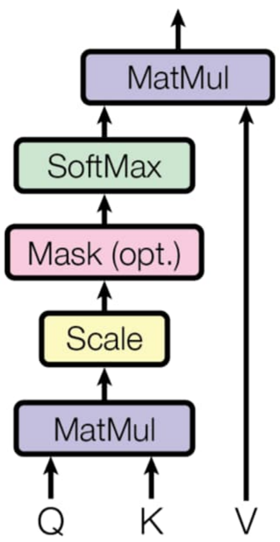
different way to understand sentence

- Idea 1: One read a sentence, finding out how each individual word is related to the rest of words. Thus, he/she gets the deeper meaning of each word.
- Idea 2: With their deeper meanings, the person revisits the relationship between the words so he/she obtains an even deeper meaning for each word.
- Idea 3: The process is repeated many times until the person is satisfied with the depth of understanding
- Idea 4: Even better, we ask many people examine the same sentence using the same mechanism, since everyone may interpret it differently
- Idea 5: Finally, their combined understanding of each word becomes the final word representation

After this process, the understanding is so thorough!

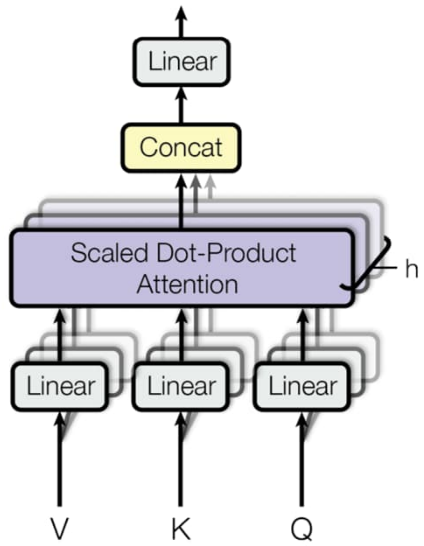
Vaswani, et al., (2017)

Scaled Dot-Product Attention



encapsulate Idea 1, Idea 2, Idea 3

multi-head

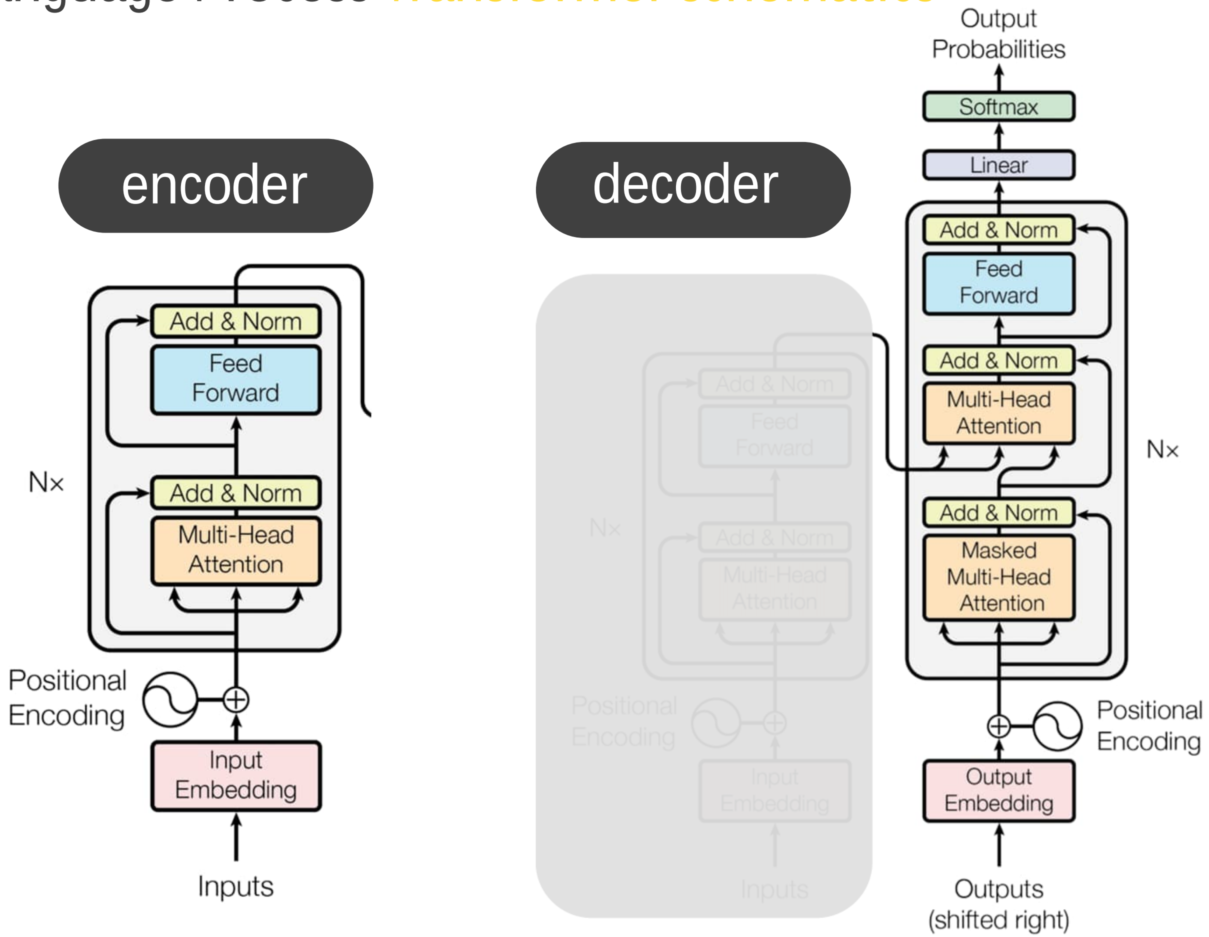


encapsulate Idea 4, Idea 5

- For mathematical explanation, please refer to the accompanying notes in Moodle

Deep Natural Language Process Transformer schematics

HKBU :: MATH :: Prof Xu



Put Ideas into seq2seq setting

Bidirectional Encoder Representations from Transformers

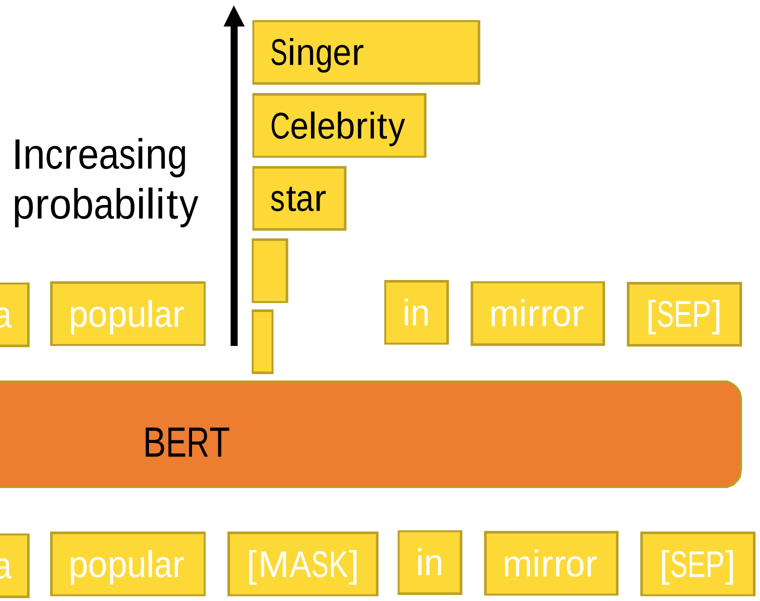
Devlin et al., (2018)

- Now that sentence embedding is also context dependant

she walks along the river bank

she works at an investment bank

- BERT trains itself of how to fill in the blanks



Bidirectional Encoder Representations from Transformers

- Now that sentence embedding is also context dependant

she walks along the river bank

she works at an investment bank

- BERT also trains itself of recognize if the adjacent sentence is indeed the next to each other

[CLS] Keung-To is a popular [MASK] in mirror [SEP] so is Yeung-lok-man [SEP]

Label = IsNext

[CLS] Keung-To is a popular [MASK] in mirror [SEP] weather is very nice [SEP]

Label = NotNext

- In both cases, the training labels do not require human

Deep Natural Language Process concept of prompt

Traditional

"I love No. 1 Seed." ⇨ Encode (e.g., BERT) ⇨ Sentiment analysis ⇨ Sentiment [POS] 92% / Sentiment [NEG] 8%

Example of downstream task

Prompt

- So if design a template: "[X] *it was a* [Z] *song."*

  X = *"I love No. 1 Seed."*

  ⇩

  X' = " *I love No. 1 Seed. it was a [Z] song."* ⇨ Z = {"excellent", "great", "wonderful" ...} [POS sentiment]
  Z = {"bad", "awful", "silly" ...} [NEG sentiment]

- Language models are so good that you can just generate word(s), instead of solving an additional problem (sentiment analysis)
- Can perform zero-shot learning

Deep Natural Language Process concept of prompt

HKBU :: MATH :: Prof Xu

with powerful language model

- Language model *understands* prompt

[CLS] how many members in the band mirror [SEP] there are 12 members

- the answer is generated to correspond to input sentence (input)

Deep Natural Language Process GPT

Radford et al., (2018)

HKBU :: MATH :: Prof Xu

Unsupervised pre-training

- Maximize the probabilities of the words in sentences

$$\text{Prob}(2^{nd} \mid 1^{st}, \text{param}) \times \text{Prob}(3^{rd} \mid 2^{nd}, 1^{st}, \text{param}) \times \text{Prob}(4^{th} \mid 3^{rd}, 2^{nd}, 1^{st}, \text{param}) \times \ldots$$

- For each head:

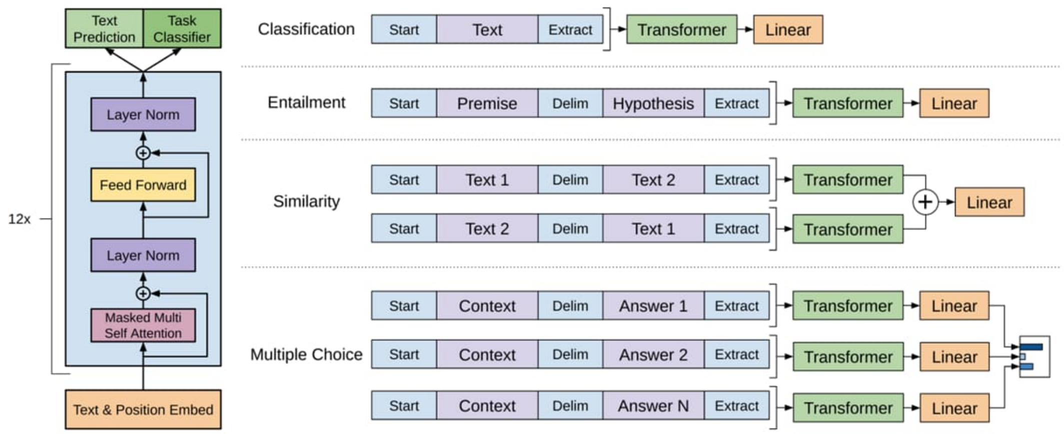
first layer: word tokens, word embedding, position embedding ⇨ h_0

repeat: $h_l = \text{transformer}(h_{l-1})$ last layer h_n

- Can perform zero-shot learning:

(h_n, word embeddings) ⇨ words pprobabilities

Supervised fine-tuning



GPT3 + InstructGPT = chatGPT

- When grab data from the internet, answer may not always be useful.

e.g., "*who sings the song "No. 1 Seed"?*"

Keung-To. ⇐ useful answer

I don't know, please tell me! ⇐ not a useful answer

- Need human (teacher) guidance

"*who sings the song "No. 1 Seed"?*" ⇒ A teacher model

high

⇒ *Keung-To.*

⇒ *I don't know, please tell me!*

low

Reward signal in RL

Where to find more material?

HKBU :: MATH :: Prof Xu

- My never-ending Machine Learning knowledge dissemination notes:

  https://github.com/roboticcam/machine-learning-notes

- All my digital footprint (links to things I talked about today)

  https://github.com/roboticcam/demo_links

- Mathematics supplementary notes on Moodle:

Live Machine Learning Class:

中文机器学习研究线上课

2022年我坚持每周日晚上8:30直播机器学习研究课程系列 (微信二维码在这个链接)- From 2022, I hold regular 8:30pm Sunday Night live (SNL) broadcast on Machine Learning theory.

English version

From April 2022, I started a machine learning research seminar series every 2-3 weeks in English via Zoom. It's at 7pm Hong Kong Time. I will continue to explain machine learning using an intermediate level mathematics. The current topic is: "Gradient Descend Research". You need a solid understanding of linear algebra, calculus, probability and statistics. You can register via meetup https://www.meetup.com/machine-learning-hong-kong/ (Back in Australia, I also conducted research training to all machine learning PhD students at Australian universities, with over 100 students participating via Zoom.)

Learning Theory Classes

- Class 1: Introduction

Mathematics in Modern Natural Language Modeling

Richard Xu

February 27, 2023

1 A few words

To start this topic, I will first discuss the mathematics in modern natural language processing. By the way, Natural Language Processing (NLP) is one of the most important/exciting applications of artificial intelligence, machine learning and data mining. Contrary to computer vision, NLP will influence the work of many people in the future.

Although neural networks play an important role in NLP. However, we will cover neural networks later in this topic. Therefore, we temporarily avoid talking about N-N.

Also note that while this topic is about techniques in NLP, these techniques can also be applied to other machine learning and data science settings.