L’Intelligence artificielle et ses enjeux sociétaux

Fondation UBS

Prof. Dr. Ashwin ITTOO
ULiège
22 septembre 2023

About Myself – *Ashwin Ittoo*

Full Professor @ HEC Liege, ULiege

Japan Advanced Inst. Of Science & Tech

Machine Learning and Language Understanding Lab

Associate Editor, Elsevier

Computers in Industry

Industry Projects/External Roles

Agenda

- AI Overview
  - Methods, GPT
  - Understand how machines learn
- Bias & Privacy
- Use-cases & Implementation

Artificial Intelligence

How Machines Learn -

An Overview of Principles & Methods

How Do Machines Learn?

- Learn how to
  - Offer personalized recommendations? E.g. Netflix, Amazon
  - Translate text & recognize speech? E.g. Google Translate, Alexa
  - Generate texts? E.g. GPT-based models, Llama,...

How Do Machines Learn?

- Learning occurs from past data --> *training data*
  - Browsing history, comments posted online
  - Huge text collections
  - Images

How to Learn?

- Learning methods
  - Supervised
  - Unsupervised
  - Self-supervised
  - (Reinforcement)

Supervised

- Prediction
  - Credit risk
  - Sentiment

- Labelled data required for training ML method
  - Customer reviews + *associated sentiments* (POS, NEG)
  - Customer records + *net worth scores*
  - Criminals' records + *recidivism risk scores*
  - Product images + *defective regions*

- Training
  - ML system input: labelled examples, e.g. *customer records + risk scores*

Credit Risk Dataset Example

person_age	person_inc	person_home_own	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_in	cb_person_default_	cb_person_cred_hist_length
22	59000	RENT	123.0	PERSONAL	D	35000	16.02	1	0.59	Y	3
21	9600	OWN	5.0	EDUCATION	B	1000	11.14	0	0.1	N	2
25	9600	MORTGAGE	1.0	MEDICAL	C	5500	12.87	1	0.57	N	3
23	65500	RENT	4.0	MEDICAL	C	35000	15.23	1	0.53	N	2
24	54400	RENT	8.0	MEDICAL	C	35000	14.27	1	0.55	Y	4
21	9900	OWN	2.0	VENTURE	A	2500	7.14	1	0.25	N	2
26	77100	RENT	8.0	EDUCATION	B	35000	12.42	1	0.45	N	3
24	78956	RENT	5.0	MEDICAL	B	35000	11.11	1	0.44	N	4
24	83000	RENT	8.0	PERSONAL	A	35000	8.9	1	0.42	N	2
21	10000	OWN	6.0	VENTURE	D	1600	14.74	1	0.16	N	3
22	85000	RENT	6.0	VENTURE	B	35000	10.37	1	0.41	N	4
21	10000	OWN	2.0	HOMEIMPROVEMENT	A	4500	8.63	1	0.45	N	2
23	95000	RENT	2.0	VENTURE	A	35000	7.9	1	0.37	N	2
26	108160	RENT	4.0	EDUCATION	E	35000	18.39	1	0.32	N	4
23	115000	RENT	2.0	EDUCATION	A	35000	7.9	0	0.3	N	4
23	500000	MORTGAGE	7.0	DEBTCONSOLIDATION	B	30000	10.65	0	0.06	N	3
23	120000	RENT	0.0	EDUCATION	A	35000	7.9	0	0.29	N	4
23	92111	RENT	7.0	MEDICAL	F	35000	20.25	1	0.32	N	4
23	113000	RENT	8.0	DEBTCONSOLIDATION	D	35000	18.25	1	0.31	N	4
24	10800	MORTGAGE	8.0	EDUCATION	B	1750	10.99	1	0.16	N	2
25	162500	RENT	2.0	VENTURE	A	35000	7.49	0	0.22	N	4
25	137000	RENT	9.0	PERSONAL	E	34800	16.77	0	0.25	Y	2

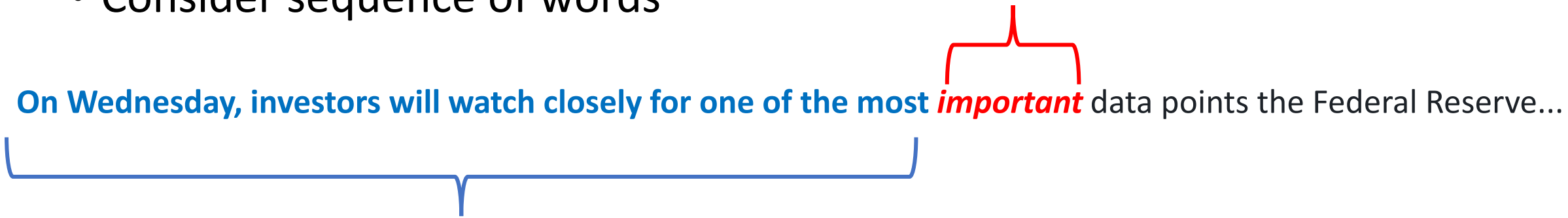
https://tinyurl.com/vykew837

Unsupervised Learning

- Descriptive methods (vs. Predictive for Supervised Learning)
- No requirements for labelled data
- Algorithms concerned with analysis of data
  - Discover hidden patterns
  - Group data items based on patterns similarity
- Suitable for various applications
  - Customer segmentation for targeted marketing
  - Recommender systems
  - Understanding large amounts of data
  - Pre-processing prior to supervised learning

Self-supervised Learning

- Part of input data serves as labeled output
- Consider sequence of words



- Context (features) to learn predict next word
  - Next word serves as label for context
  - Unlabeled data reconsidered as labeled data
- Reminds us of...?

Some Methods

Classical Methods

- Decision Trees
- Random Forests
- Support Vector Machines
- ...
- **Artificial Neural Networks**

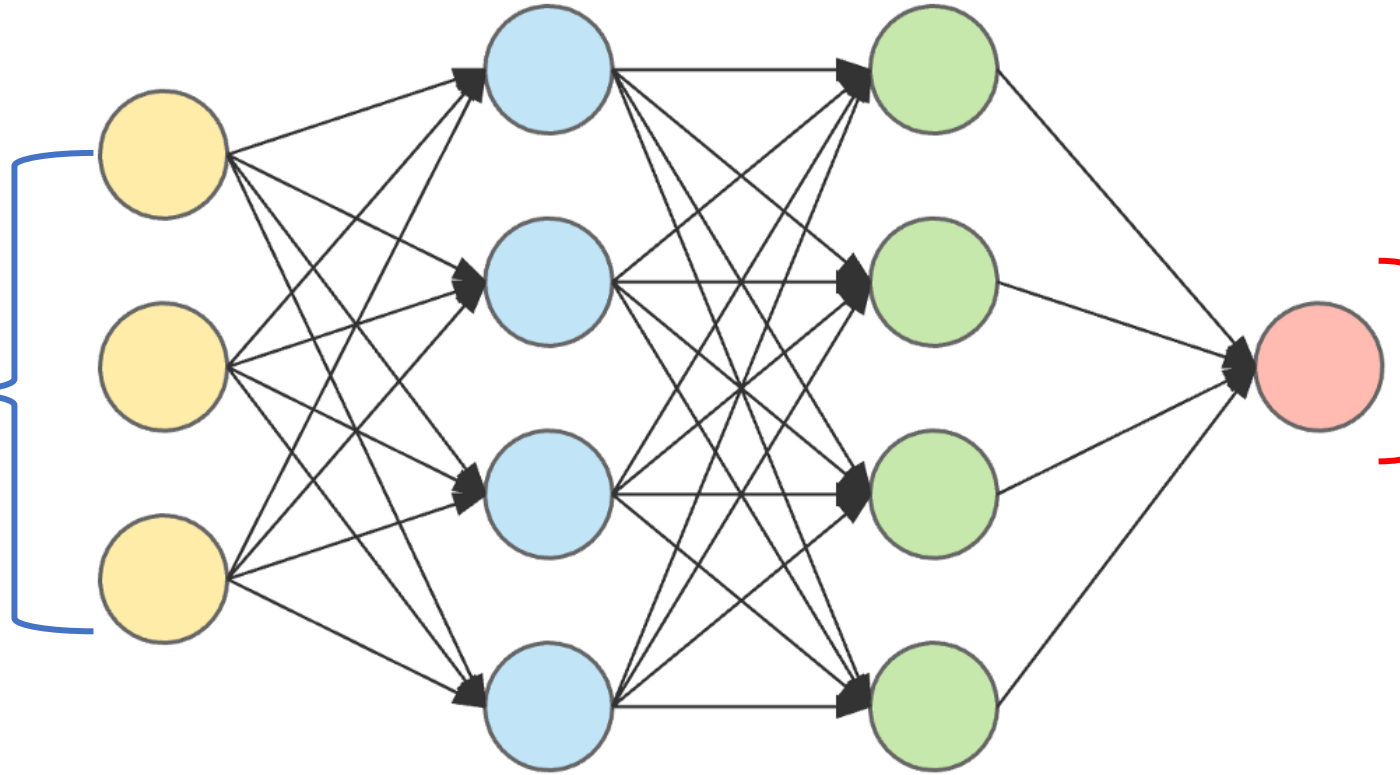
Artificial Neural Networks

- Models inspired by structure and function of biological neural networks

- Versatile, lends itself to
  - Supervised learning (classification, prediction)
  - Unsupervised learning (auto encoders)
  - Self-supervised
  - Reinforcement learning (deep reinforcement learning)

- At the crux of Deep Learning

ANN Training



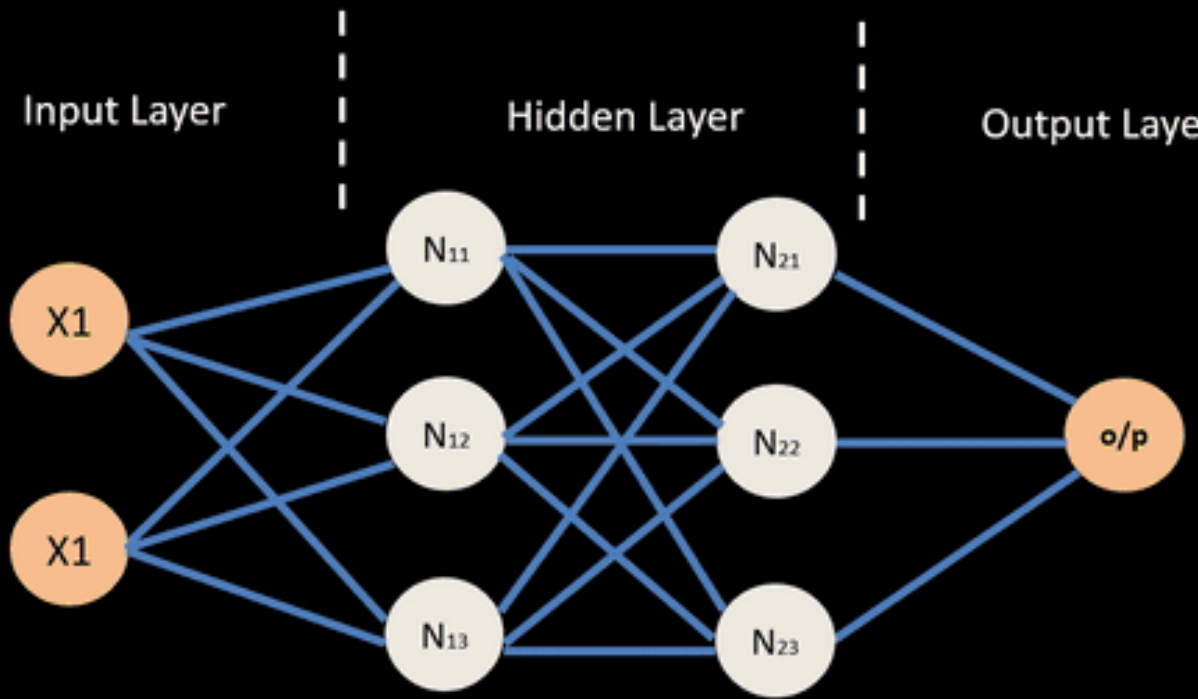
Independent variables
1 node per variable

Dependent variable
Single output node
(regression task)

Multiple output nodes for classification task (1 per class)

Hidden layers where values are transformed & fed forward to next layer
Deeper networks--> more hidden layers --> more complex computing

Neural Network – Backpropagation



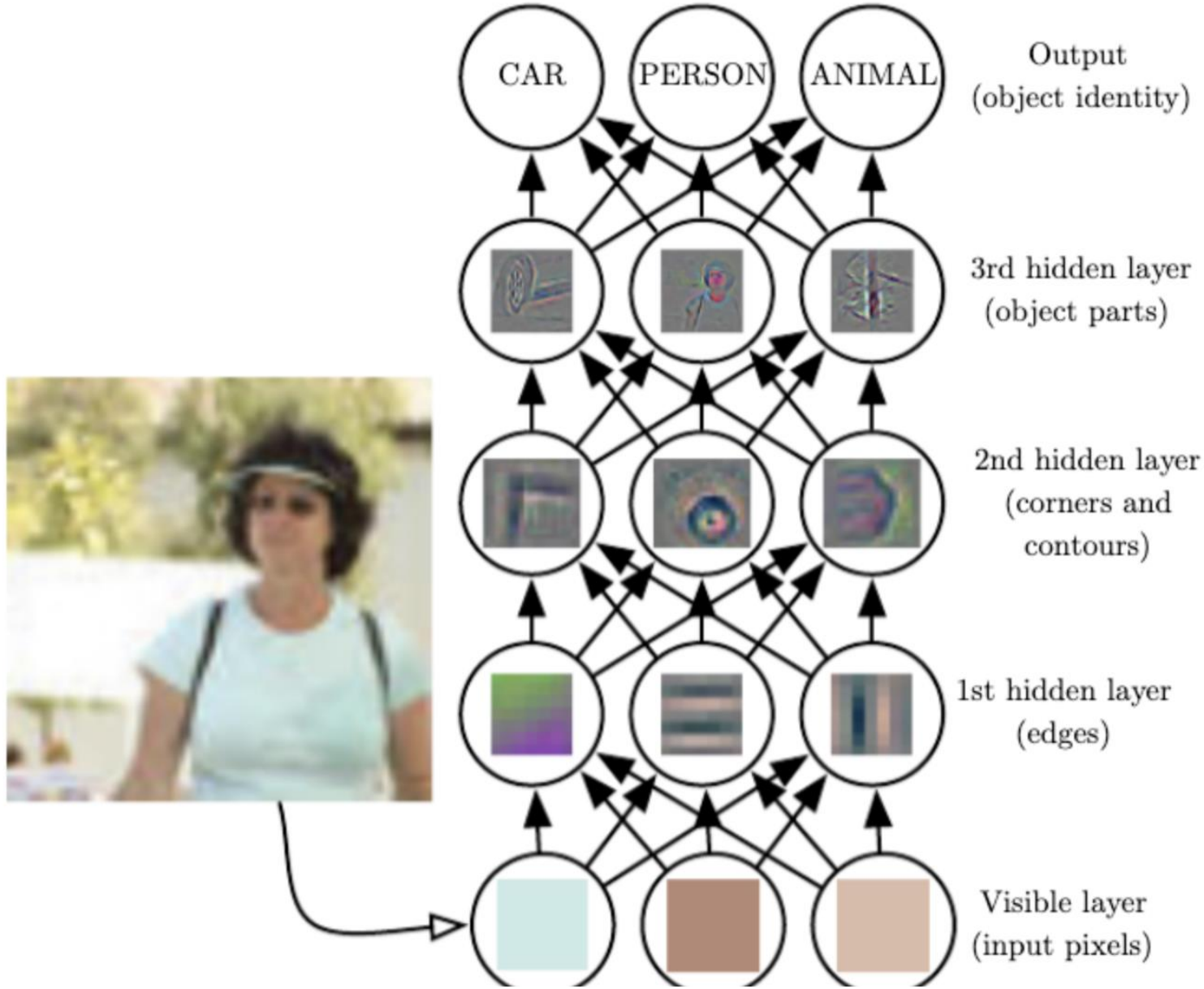
Deep Learning

- Deep Learning refers to Deep Artificial Neural Networks

- Many layers create ***deep*** architectures
  - Composition of transformations between layers
  - Enables learning of complex functions related to vision & speech
    - Image recognition
    - Machine translation
    - …

Deep Learning

- •Deep Learning (DL) are Representation Learning methods
  - Multiple representation levels obtained
  - Composing non-linear modules (interconnected neurons)
  - Each layer transform representation at one level into a representation at higher, more abstract level



https://www.deeplearningbook.org

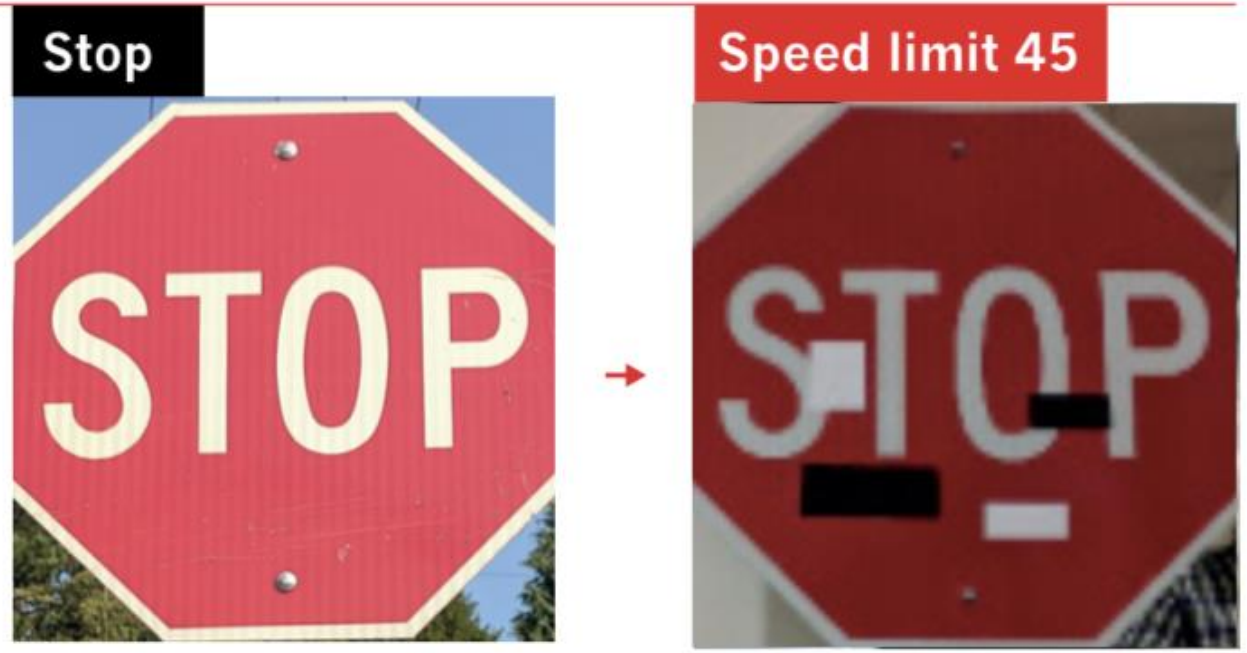
Ubiquitous Deep Learning Applications

- Search engines
  - Google Search
  - Microsoft Bing
- (Machine) Translation
  - DeepL
  - Google Translate
- Personal Assistant
  - Alexa
  - Siri

Powerful but Brittle

- Small alterations, noise in data
  - Drastic consequences

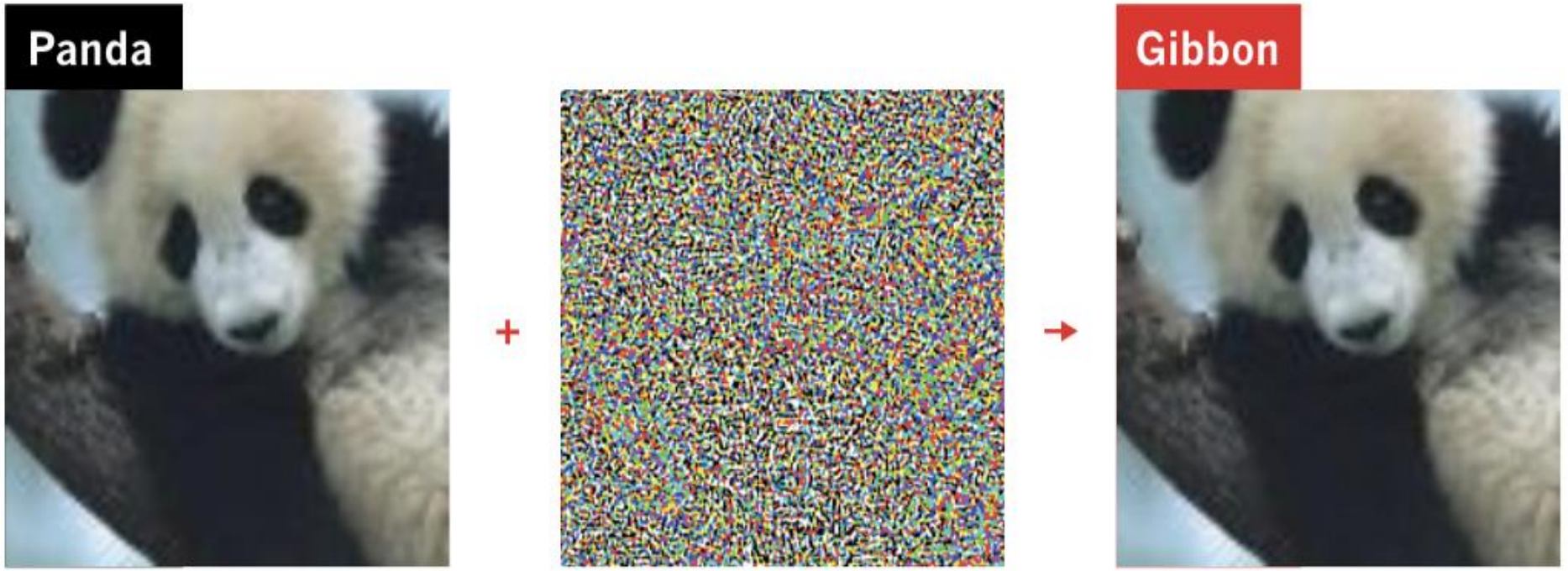
These stickers made an artificial-intelligence system read this stop sign as ‘speed limit 45’.



Details: Eykholt,K. et al., Robust Physical Attacks on Deep Learning Visual Classification, IEEE/CVF Conf. Comp.Vision Pattern Recog., (2018)

Adversarial Attacks

Adding carefully crafted noise to a picture can create a new image that people would see as identical, but which a DNN sees as utterly different.



https://tinyurl.com/3aj63u6p

From Deep Learning to Transformers

Transformers

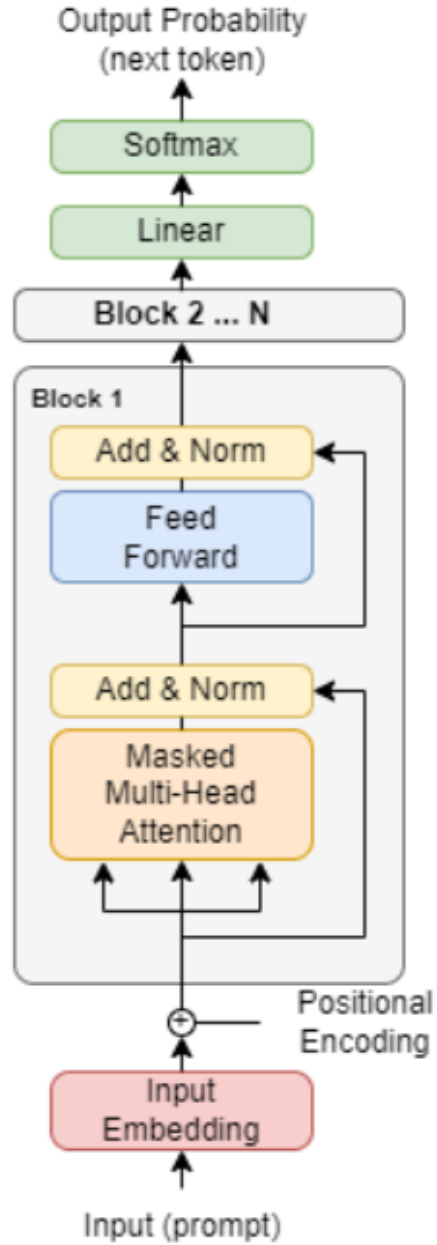
- Transformers as most LLMs backbone

- Transformer
  - Developed by Google in 2017 ("Attention is all you need")
  - Based on shallow neural networks

  - Ability consider word context & useful for disambiguation
    - "the *bank* is going bankrupt" vs. "the *bank* is flooding"

  - Implements (self-)attention mechanism
    - Memorizes long contexts (long sentences)

  - Lends itself to parallelization
    - Words in a sentence processed together vs. sequentially (token by token)
    - Massive gains in run time

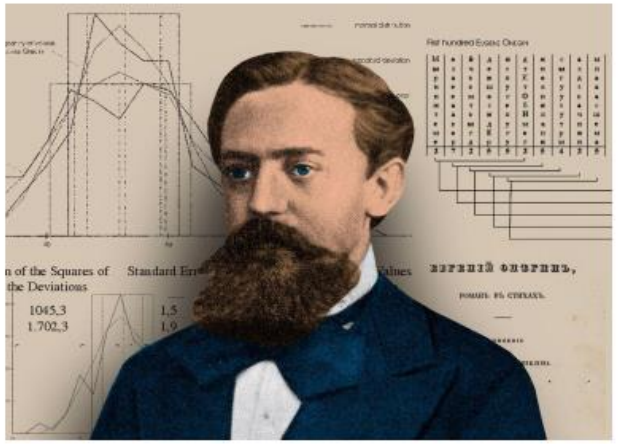

Output Probability
(next token)
Softmax
Linear
Block 2 ... N
Block 1
Add & Norm
Feed
Forward
Add & Norm
Masked
Multi-Head
Attention
Positional
Encoding
Input
Embedding
Input (prompt)

Generative Pretrained Transformers (GPT)

- GPT: Type of Transformer model from OpenAI

- Designed for & applied to *Language Modelling*
  - Predict next word given previous words (contexts)

What is a Language Model?



In 1913, Russian mathematician Andrey Markov counted letters from "Eugene Onegin" and showed that the chance of a letter appearing depends on the letter before it.

Science in Context **19**(4), 591–600 (2006). Copyright © Cambridge University Press
doi:10.1017/S0269889706001074 Printed in the United Kingdom

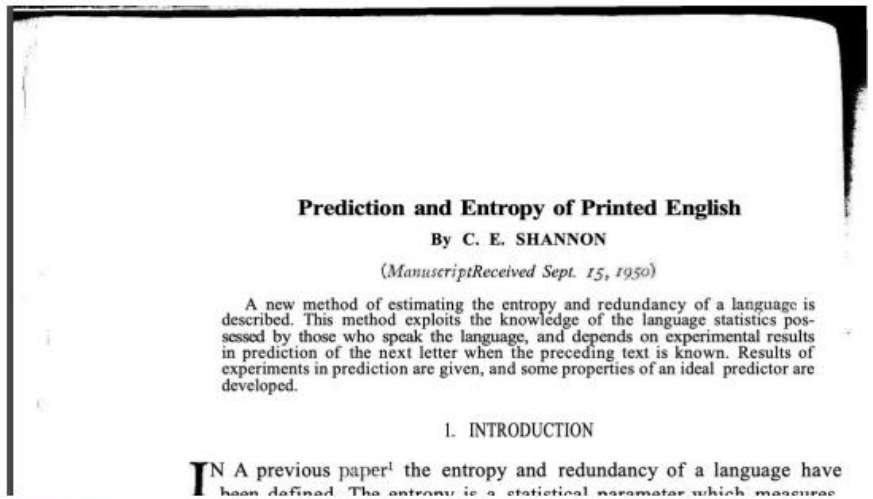
Classical Text in Translation

An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains

A. A. Markov

(Lecture at the physical-mathematical faculty, Royal Academy of Sciences, St. Petersburg, 23 January 1913)[1]

In 1951, Claude Shannon published the first paper on the prediction of English letters



Prediction and Entropy of Printed English

By C. E. SHANNON

(*ManuscriptReceived Sept. 15, 1950*)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

1. INTRODUCTION

IN A previous paper[1] the entropy and redundancy of a language have

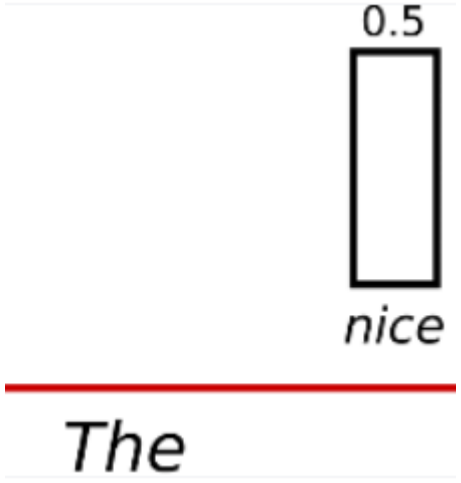
Illustration of Andrey Markov: https://spectrum.ieee.org/andrey-markov-and-claude-shannon-built-the-first-language-generation-models
Photo of "Eugene Onegin" https://www.themoscowtimes.com/2019/11/29/pushkins-classic-eugene-onegin-sells-for-150k-in-london-a68410
Photo of Claude Shannon https://spectrum.ieee.org/claude-shannon-information-theory

Source: Pascale Fung, Special Topics in Deep NLP

Language Model Example

- Learn how to predict next word
  - Prob (next word | current word)
  - From huge text corpus

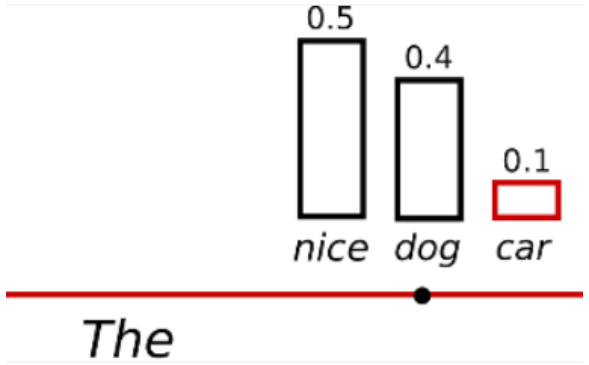
- Prompt model with any word, e.g. "the"
  - Model generates most frequent word given "the" (from what it has learnt)
  - Suppose next word "nice"



Language Model Example

- Then given "the", "nice", predict the next word
  - Conditional prob. P(w|the, nice)

- Next word according to training data



Probabilistic Approach

- Doubtful as to whether they reason
- Stochastic parrots
- More on this later

Generative Pretrained Transformers (GPT)

- GPT-2 gained public attention
  - Model with 1.5 billion parameters
  - Trained on ~ 8 million webpages

- Input sentence:
  - *In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

- Asked to predict next words...

- Output produced by Language Model learnt in GPT-2

Coherent Text

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

Inserted line breaks

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

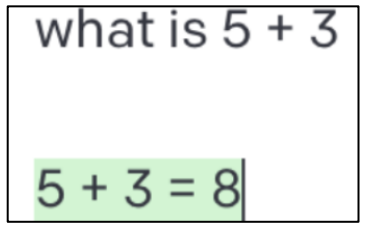
While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

Inserted quotes

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

GPT

- GPT-2 spurred further research in language models (LLMs)
- Led to GPT-3
  - Displayed higher order cognitive abilities
- Basic arithmetic



- Code generation from natural language
  - JXS
  - REACT

ChatGPT

- Focuses on dialogues

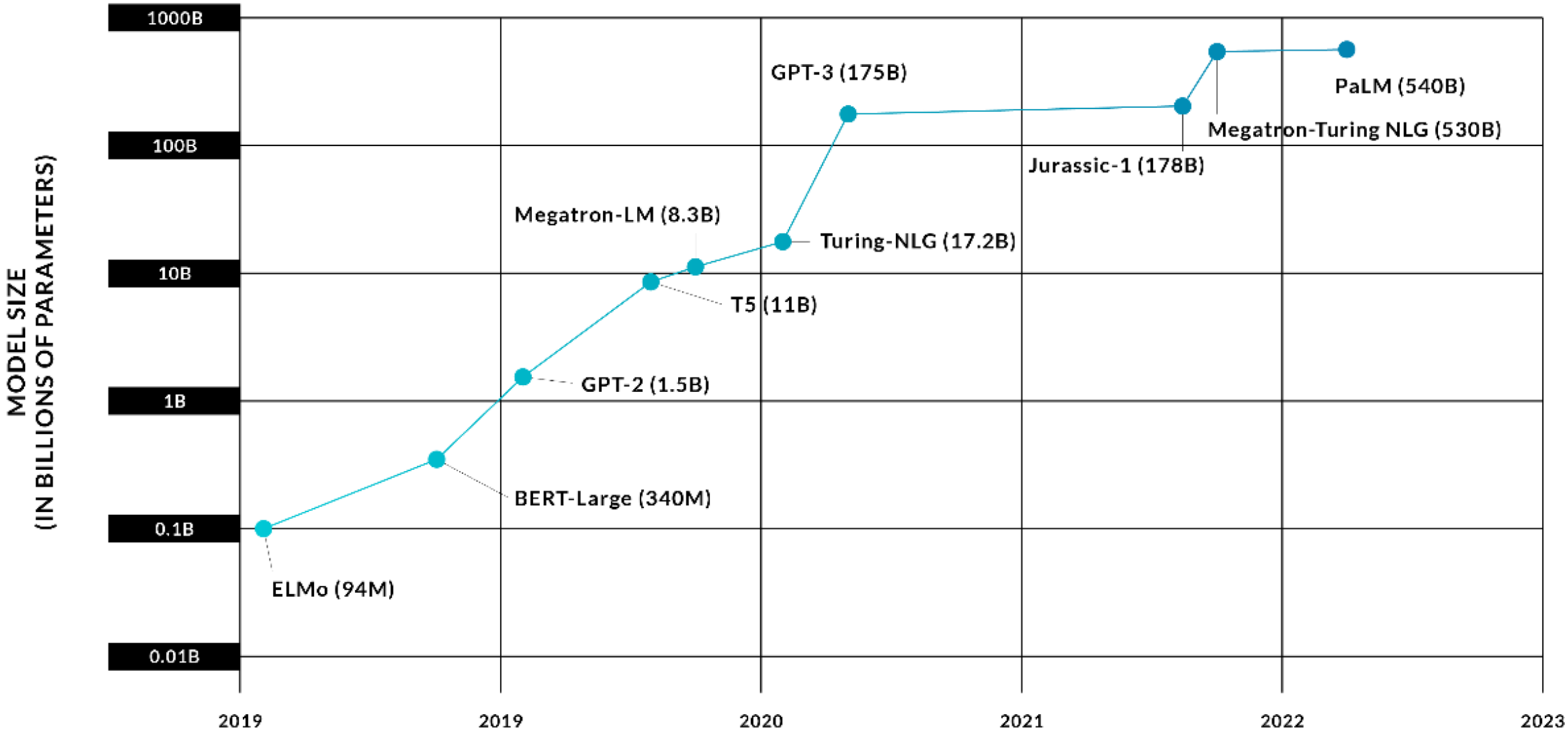
- But LLM at its core
  - Trained on large text corpora
  - Language modelling task

- Safeguards against deviating into malicious behaviour
  - Hate speech, bullying, antisemitism,…

- Implemented by another neural network
  - Trained on human-labelled data to learn human preferences
  - Used to fine-tune original GPT3 model to prefer certain answers over others
  - Paradigm known as Reinforcement Learning with Human Feedback

LLMs

- Language models evolution to Large Language Models (LLMs)
- Trained on massive datasets
  - Wikipedia (~20GB)
  - BookCorpus (~5GB)
  - Common Crawl (>20TB)
- Complex & huge number of parameters



Source: https://twosigmaventures.com/blog/article/the-promise-and-perils-of-large-language-models/

Useful Reading

- Deep Learning, Goodfellow et al. , MIT Press(2016)
- Deep Learning, LeCun et al., Nature (2015)
- Attention is all you need,Vaswani et al., Neurips (2017)
- OpenAI Codex
  - Accessed 6th March 2023
- OpenAI Learning from Human Preferences
  - Accessed 6th March 2023
- Microsoft KOSMOS-1

Useful Reading

- Kurakin et al., Adversarial Examples in the Physical World, Tech. Report, Google Inc. (2016)
- Heaven, Deep Trouble for Deep Learning, Nature (2019)

- Heaven, Why deep-learning AIs are so easy to fool, Nature (2019)

- Ghaffari Laleh et al., Adversarial attacks and adversarial robustness in computational pathology, Nature (2022)