Machine Learning

AI/ML/DL “buzzwords”.

Aleksandr Petiushko

ML Research

Content

1. AI/ML/DL “buzzwords”
   - Broad Concepts
   - Directions
   - State-of-the-Art

AI “buzzwords” requests: broad concepts

- AI in the real world and its evolution
- Bias in AI
- Explainable/Interpretable AI
- Machine learning related revolution
- AI Work Planning
- AI Ethics/Law

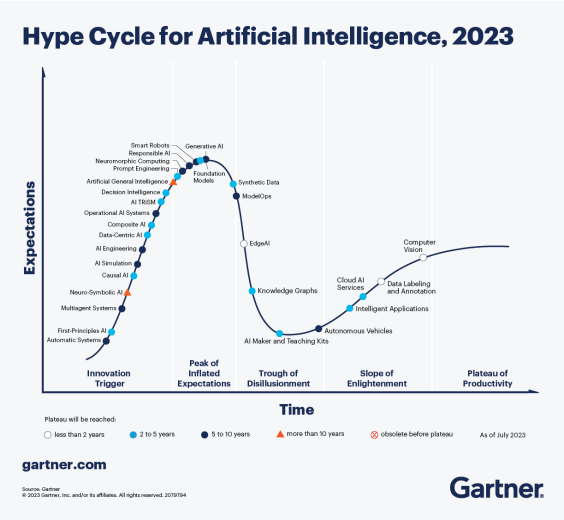
AI “buzzwords” requests: directions

- Deep Learning / Neural Nets
- Computer Vision
- NLP
- Speech Recognition
- Recommendation Systems
- Multi-modality
- Meta-learning
- RL and RLHF
- Edge Computing
- AutoML
- Self Driving
- AI in Healthcare

AI “buzzwords” requests: State-of-the-Art

- Generative AI
- AI Art Generator (Diffusion)
- LLM and ChatGPT
- Hallucinations
- Transformer Architecture
- Generative Adversarial Networks

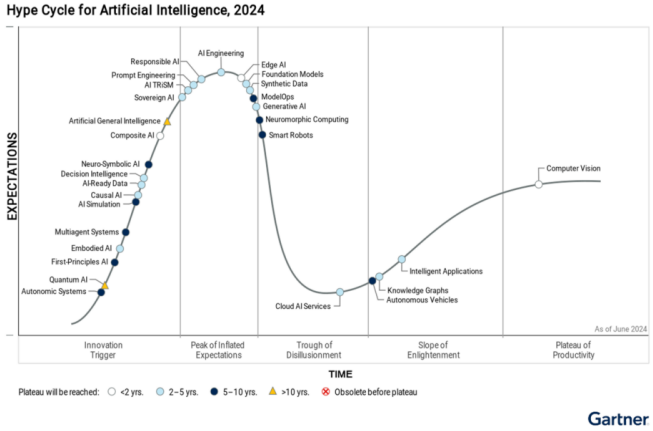
Broad concepts: AI Hype Cycle 2023[1]



[1] www.gartner.com

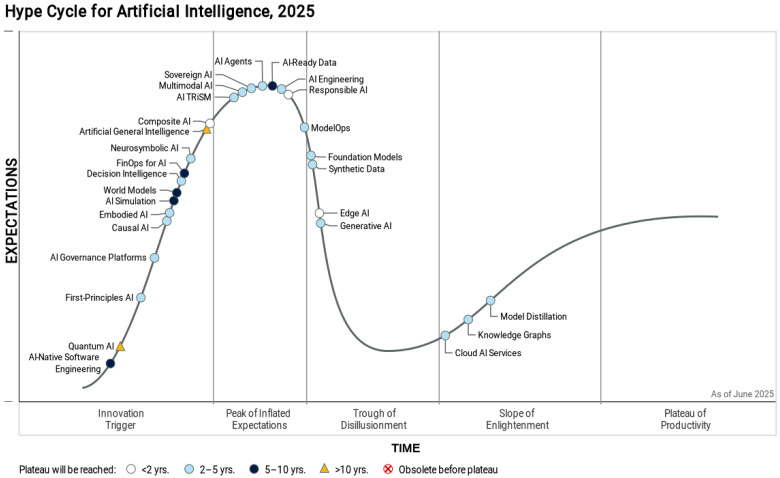
Broad concepts: AI Hype Cycle 2024[2]

Figure 1: Hype Cycle for Artificial Intelligence, 2024



[2] www.jaggaer.com

Broad concepts: AI Hype Cycle 2025[3]



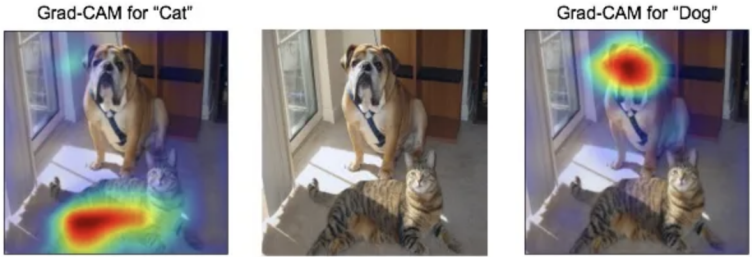
[3] www.mrak.at

Broad concepts: Bias in AI

- Provides unfair prediction based w.r.t. race, gender, age. etc
- The main reason: skewed/sparse training data
- Main technique to avoid: human-in-the-loop, alignment
- Read material: link old, link new

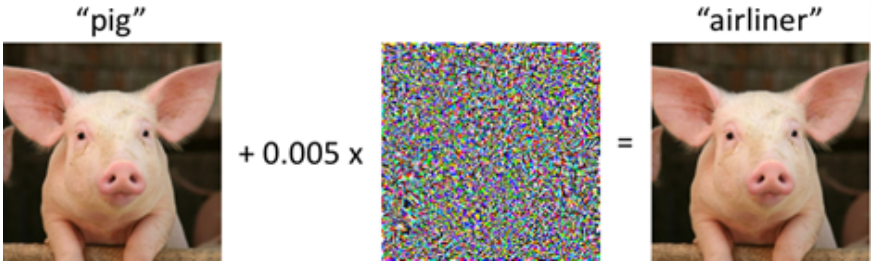
Broad concepts: Explainability and Interpretability

- AI/ML models *are not* black-boxes; they are huge (millions of weights) and structured (architecture) multi-dimensional and multi-variable functions
- Interpretability: understand the influence of any input sub-area/sub-feature on on the model output; ML Debug system
- Explainability: high-level interpretability (using human-like language)
- Interpretability can be done via input counterfactual analysis
- Explainability now can be done via LLM chain-of-thought (CoT) technique
- Read material: link old, link new



Broad concepts: ML-related revolution[4]

- Usually, the ML model performance is shown on a test/eval dataset
- Unfortunately, there is no guarantee of the similar performance on the different settings/datasets
- Providing these guarantees would put much more trust in ML and increase deployment speed
- Read material: link



[4]Personal opinion of a lector

Broad concepts: AI Work Planning[5]

- First of all, there is no a standard approach to AI Work Planning: it depends on many factors including the problem, the resources, the team capabilities, etc.
- One of the main practical approaches is what we've gone through "ML Pipeline" exercises during the course
- Practical advice: work in a "fail-fast" approach according to an Agile methodology
- Read material: link

[5] Personal opinion of a lector

Broad concepts: AI ethics[6]

Inequity and fairness

ML can contribute to and amplify social inequity

For foundation models, it is useful to separate:

- intrinsic biases (properties in the foundation model)
- extrinsic harms (harms in specific applications)

Source tracing to understand ethical/legal responsibility

Mitigations: proactive interventions/reactive recourse

Environment

Foundation models involve significant training/emissions

One perspective: amortised cost over re-use

Several factors would be beneficial to consider:

- compute-efficient models, hardware, energy grids
- environmental cost as a factor for evaluation
- greater documentation and measurement

Economics

Foundation models may have economic impact due to:

- novel capabilities
- potential applications in wide array of industries

Initial analyses have been conducted to understand implications for productivity, wage inequality, concentration of ownership

Misuse

Misuse: the use of foundation models as technically intended but for societal harm (e.g. disinformation)

Foundation models may make misuse easier by generating high-quality personalised content

Disinformation actors can target demographic groups

Foundation models may also help to detect misuse

Legality

How law bears on development/deployment is unclear

Legal/regulatory frameworks will be needed

In the US setting, important issues include:

- liability for model predictions
- protections from model behaviour

Legal standards must advance for intermediate models

Ethics of scale

Widespread adoption of foundation models poses ethical, political and social concerns

Ethical issues related to scale:

- homogenisation
- concentration of power

How can norms and release strategies address these?

[6] www.youtube.com

AI Regulations

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,

AI Regulations

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of “Harmful” (potentially dangerous) AI;

AI Regulations

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of “Harmful” (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)

AI Regulations

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of "Harmful" (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)
  - Bans (actually, requires a "kill switch") AI if it uses more than 10^{26} FLOPs / \$100M for training or more than 10^{25} FLOPs / \$10M for finetuning

AI Regulations

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of "Harmful" (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)
  - Bans (actually, requires a "kill switch") AI if it uses more than 10^{26} FLOPs / \$100M for training or more than 10^{25} FLOPs / \$10M for finetuning
  - Governor of CA (Gavin Newsom) vetoes it (Sep 2024)

AI Regulations

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of "Harmful" (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)
  - Bans (actually, requires a "kill switch") AI if it uses more than 10^{26} FLOPs / \$100M for training or more than 10^{25} FLOPs / \$10M for finetuning
  - Governor of CA (Gavin Newsom) vetoes it (Sep 2024)
- European Union's AI Act (Aug 2024) restricts AI in essential fields like education, employment, and law enforcement, and requires developers to provide details of their algorithms and data

AI Regulations

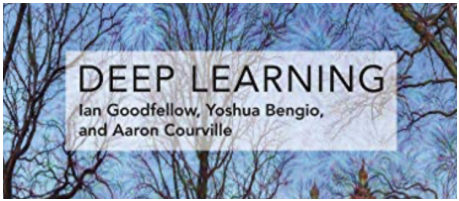
- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of "Harmful" (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)
  - Bans (actually, requires a "kill switch") AI if it uses more than 10^{26} FLOPs / \$100M for training or more than 10^{25} FLOPs / \$10M for finetuning
  - Governor of CA (Gavin Newsom) vetoes it (Sep 2024)
- European Union's AI Act (Aug 2024) restricts AI in essential fields like education, employment, and law enforcement, and requires developers to provide details of their algorithms and data
- Framework Convention on AI and Human Rights, Democracy, and the Rule of Law was signed by the US, UK, European and other countries (Sep 2024)

AI Regulations

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of "Harmful" (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)
  - Bans (actually, requires a "kill switch") AI if it uses more than 10^{26} FLOPs / \$100M for training or more than 10^{25} FLOPs / \$10M for finetuning
  - Governor of CA (Gavin Newsom) vetoes it (Sep 2024)
- European Union's AI Act (Aug 2024) restricts AI in essential fields like education, employment, and law enforcement, and requires developers to provide details of their algorithms and data
- Framework Convention on AI and Human Rights, Democracy, and the Rule of Law was signed by the US, UK, European and other countries (Sep 2024)
  - Requires AI models respect democracy and human rights

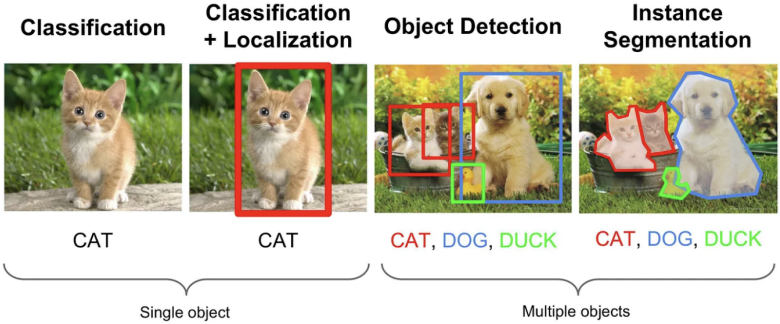
Directions: Deep Learning and Neural Nets

- Neural Net (NN): a (usually!) non-linear function mapping a (usually) multi-dimensional input to some output (which can be of the same dimension, or a bigger/smaller one)
- Most common NN atomic operations: addition, multiplication, scalar non-linearity, aggregation/normalization
- Deep Learning: a NN consisting of more 2 layers of atomic operations (that's why deep) and the corresponding procedure of the training ("learning") it weights using back propagation process
- Read material: link



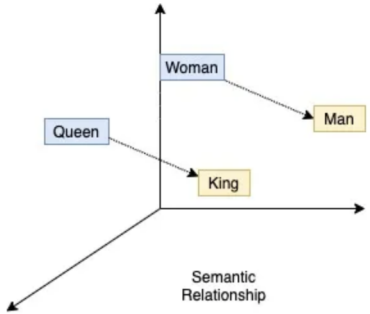
Directions: Computer Vision

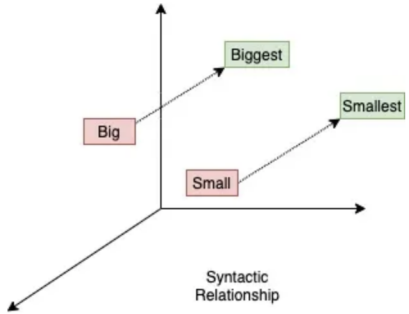
- Computer Vision (CV): Direction targeted to analyze vision information: mostly images and videos
- Most common CV directions: classification, detection, segmentation
- Main research is concentrated around architectures of CV models
- Read material: link



Directions: NLP

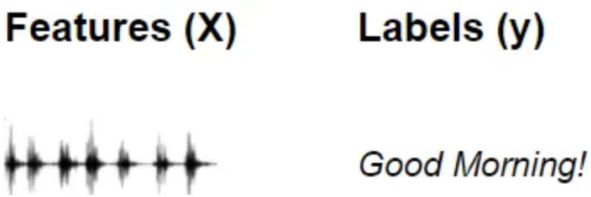
- Natural Language Processing (NLP): Direction targeted to analyze human languages (e.g., English)
- Most common NLP directions: understanding (classification), generation, QA
- The difficulties start with language representation + the engine to process this representation
- Read material: link





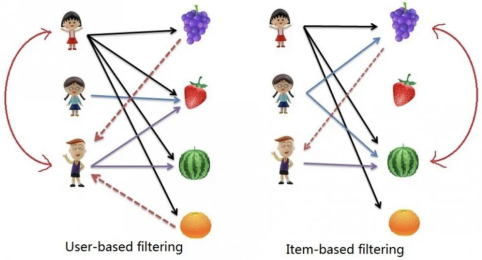
Directions: Speech Recognition

- Automatic Speech Recognition (ASR): Direction targeted to map a sequence of audio inputs to text outputs
- ASR mains differences with CV: 1) temporal sequence; 2) can benefit from signal pre-processing (like Fourier Transform, Mel-Frequency Cepstral Coefficients, etc.)
- Main research is concentrated around architectures of ASR models and how to omit pre-processing stage
- Read material: link



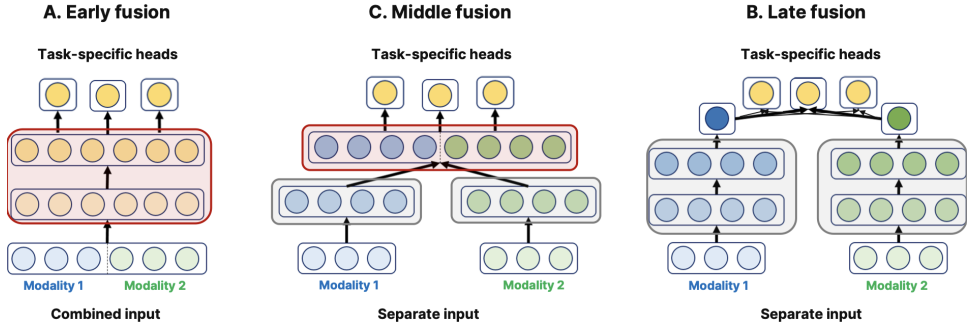
Directions: Recommendation Systems

- Recommendation Systems (RecSys): Direction targeted to predict user preference of the unseen product
- Most common RecSys approaches: Collaborative Filtering (based on User Similarity), Content-based Filtering (based on Item Similarity), and Hybrid
- Main research is concentrated around the Representations (matrix factorization, neural-based representations)
- Read material: link



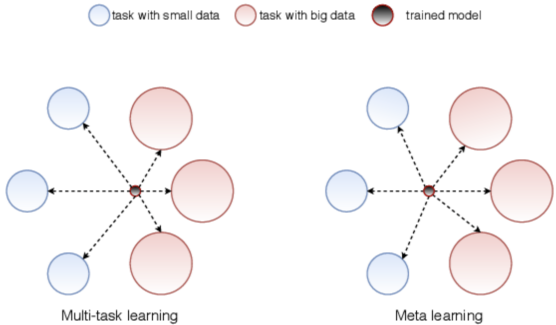
Directions: Multi-modality

- It seems that incorporating different modalities (like human does: vision, hearing, flair, etc.) can improve the performance of an ML modal
- Two main modalities to combine now: text and images
- Techniques differ in the way where we fuse the modality: early or late fusion
- Read material: link



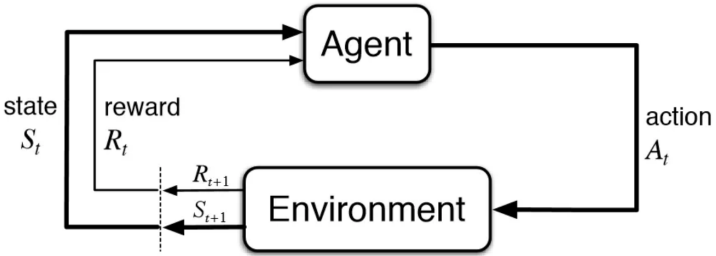
Directions: Meta-learning and Multi-tasking

- How to solve different tasks by a single model?
- If it is done sequentially, then we come to meta-learning: trying to provide the model's weights initialization as close as possible to all the tasks
- If it is done simultaneously, then w come to multi-tasking: training weights to satisfy all of the objectives at once
- Read material: link



Directions: RL and RLHF

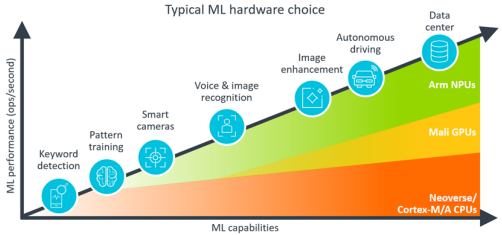
- Reinforcement Learning (RL): trying to train the ML model of action selection (e.g., of a robotic arm) based on the environment dynamics and external reward
- RL is the closest to real-life setting and at the same time the most complex one (hard to train, hard to get data, etc.)
- RL from Human Feedback (RLHF): training reward model based on human-in-the-loop preferences (e.g., pairwise comparison)
- Read material: link



Directions: Edge Computing

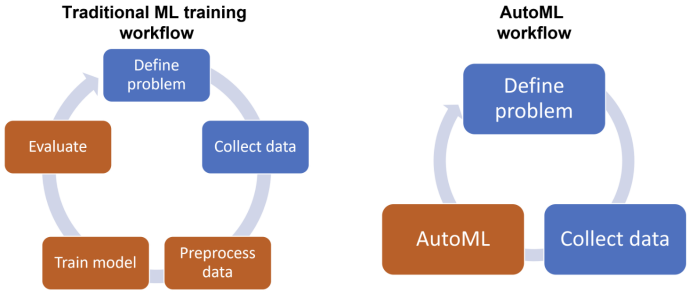
- State-of-the-Art models are huge (GB scale in both ROM and RAM), need to adapt them in order to use on edge devices
- Main techniques are: 1) distillation to faster/smaller architectures, 2) quantization, 3) prunning, 4) parallelization
- Need to be careful: small model doesn't mean low latency!
- Read material: link

Multiple ML Use Cases on Edge Devices



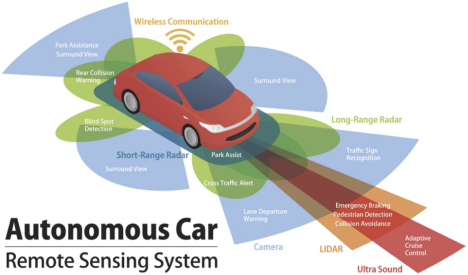
Directions: AutoML and NAS

- NN Architecture design is hard and non-formalized
- AutoML and Neural Architecture Search (NAS): an approach to automatically construct the architecture out of some atomic building blocks and the rules to combine them (NAS) and build an automatic ML pipeline around it (AutoML)
- Usually, NAS is very resource-greedy and time-consuming; but can lead to a smaller/better models (unfortunately, some times only marginally)
- Read material: link



Directions: Self Driving

- Self-driving is a super challenging area comprising literally all the ML Research directions. The main concern is safety that comes through the robustness and generalization
- Highly dependent on the training/eval data, hardware used, settings etc. — that's why there is no open self-driving model (at least, now)
- Main modules are Mapping, Perception, Prediction, Planning, Control
- Read material: link

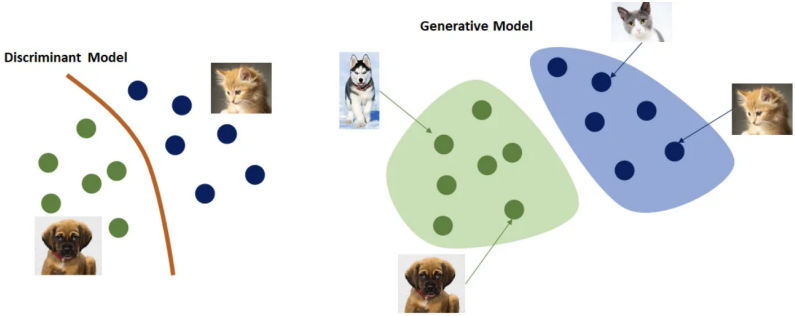


Directions: AI in Healthcare

- Healthcare is a very conservative field
- Primary challenges are 1) Data amount (small!); 2) Data privacy (hard to share); 3) Metrics are different; 4) Extremely need in explainability
- That's why most of the ML models used in Healthcare are still simple
- Successful applications include diagnostics and medical image descriptions (link)
- Read material: link

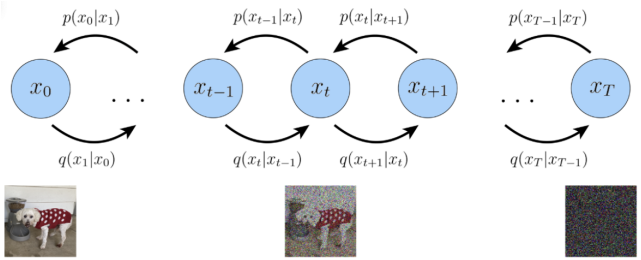
State-of-the-Art: Generative AI

- Generative AI main goal: to work on top of distributions, not single data examples
- Generative models can be conditioned: e.g., generate the image given some textual caption for it
- Output is **probabilistic**: so we can produce multiple in-distribution outputs based on the same condition
- Read material: link



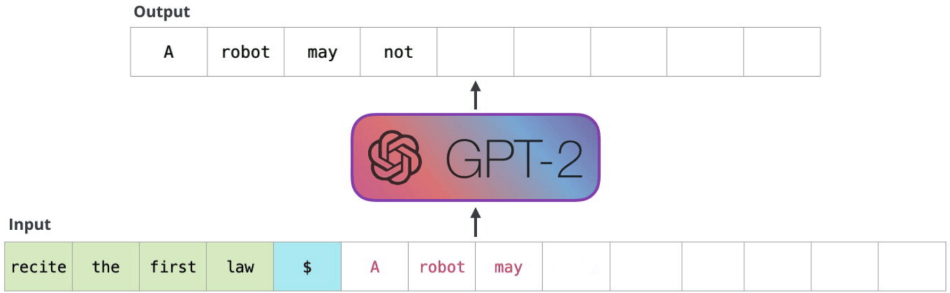
State-of-the-Art: AI Art Generator / Diffusion

- Diffusion Model is the generative model based on hierarchical markovian variational auto encoder
- Main idea is to reverse the noise (diffusion) by the learned model
- Stable diffusion is the process to work on top of the (latent) representations that have much smaller dimensions making it quite stable
- Famous representatives: Midjourney, Imagen, DALL-E (Images), SORA (Video)
- Read material: link



State-of-the-Art: LLMs and ChatGPT

- Large Language Model: A model trained with a super simple principle: either to predict a masked word (encoder; BERT) or to predict the next word (decoder; GPT)
- ChatGPT is an auto regressive decoder model trained on a huge amount of textual data (like a half of the Internet) using the RLHF
- Main problems with them right now: how to restrict the outputs (e.g., obscene language), how to ensure fairness, and how to avoid hallucinations
- Read material: link



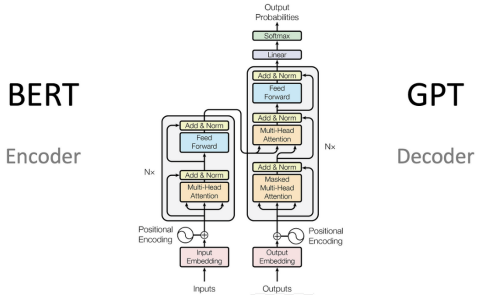
State-of-the-Art: Hallucinations

- Hallucination: one of the problems with LLMs
- It's the output that is not a part of the LLM training data
- Can be overcome with chain-of-thoughts, iterative output adjustment, and using external memory (retrieval augmented generation)
- Read material: link

Passage	Scenario #1 - Hallucination
Original Answer	Philip Hayworth was an English barrister and politician who served as Member of Parliament for Thetford from 1859 to 1868.
Sample 1	*Philip Hayworth was a **British politician** who served as the Member of Parliament for **Bolton West** from **1931 to 1945**. He was also a member of the Free Trade Union and served on several government committees.*

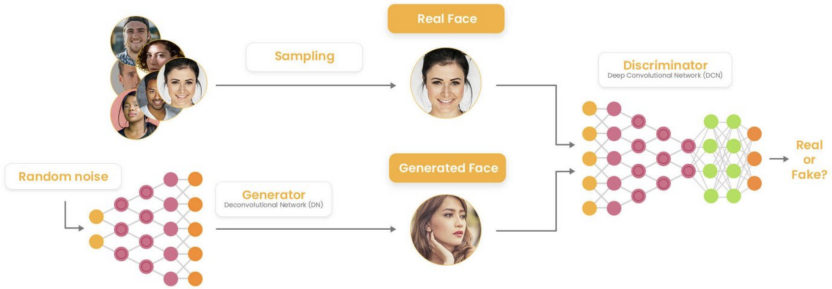
State-of-the-Art: Transformer Architecture

- Transformer is the architecture with the minimal inductive bias (e.g., in comparison to CNN or RNN): we are looking at the correlation ("attention") between every pair of input elements
- Super efficient architecture for literally every ML direction: CV, ASR, LLM, etc
- Can be encoder-decoder (machine translation), encoder (language modeling), and decoder (text generation - `http://jalammar.github.io/illustrated-gpt2/`)
- Read material: link



State-of-the-Art: GAN

- Generative Adversarial Network (GAN): one more generative model
- Main idea is to train two model in parallel: 1) Generator to produce the in-distribution data, and 2) Discriminator to discriminate between the real (GT) and synthetic (generated) data
- Quite unstable because of minmax problem and now mostly replaced by Diffusion models
- Read material: link



Thank you!