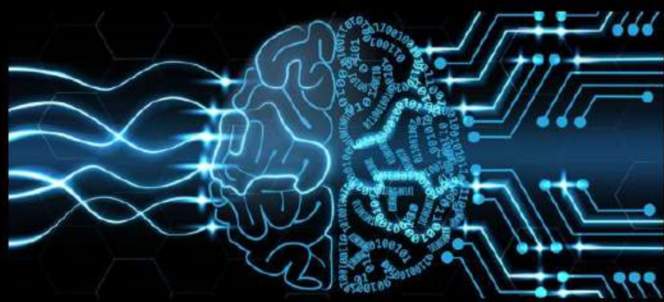
So far in Introduction to Deep Learning…

Data
- Signals
- Images
- Sensors
…

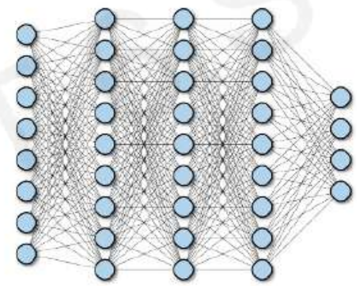


Decision
- Prediction
- Detection
- Action
…

Power of Neural Nets

Universal Approximation Theorem

A feedforward network with a single layer is sufficient to approximate, to an arbitrary precision, any continuous function.



Power of Neural Nets

Universal Approximation Theorem

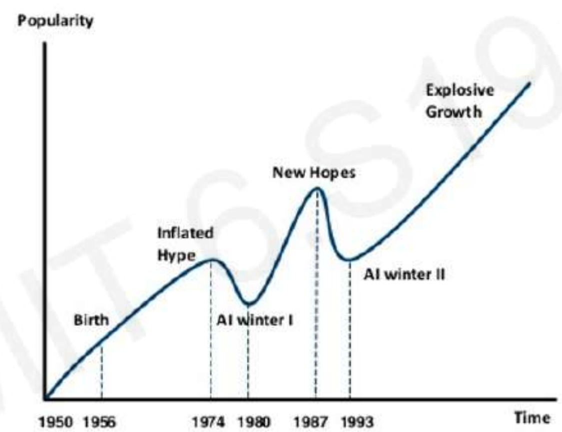
A feedforward network with a single layer is sufficient to approximate, to an arbitrary precision, any continuous function.

Caveats:

The number of hidden units may be infeasibly large

The resulting model may not generalize

Artificial Intelligence "Hype": Historical Perspective



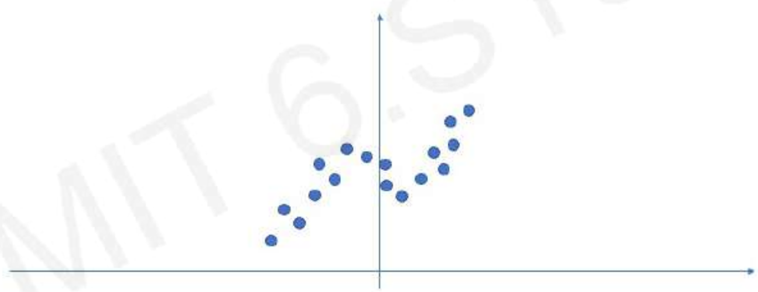
Capacity of Deep Neural Networks

Capacity of Deep Neural Networks

Capacity of Deep Neural Networks

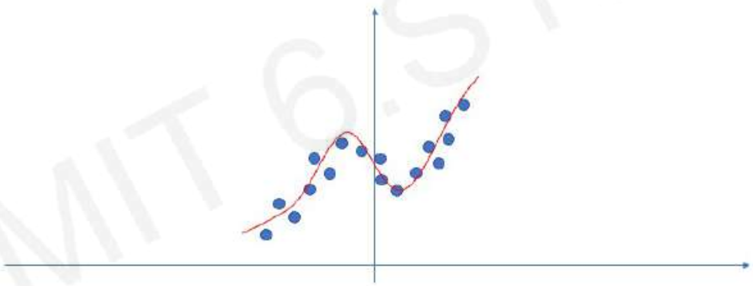
Neural Networks as Function Approximators

Neural networks are **excellent** function approximators



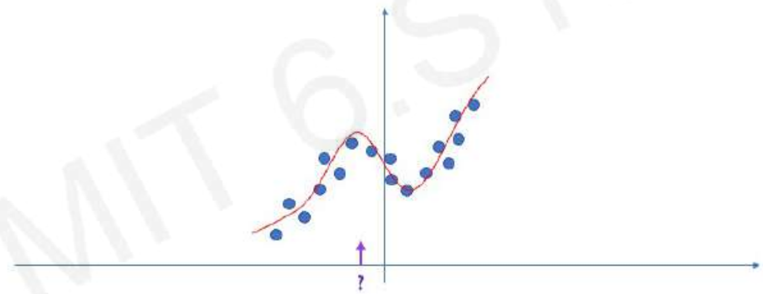
Neural Networks as Function Approximators

Neural networks are **excellent** function approximators



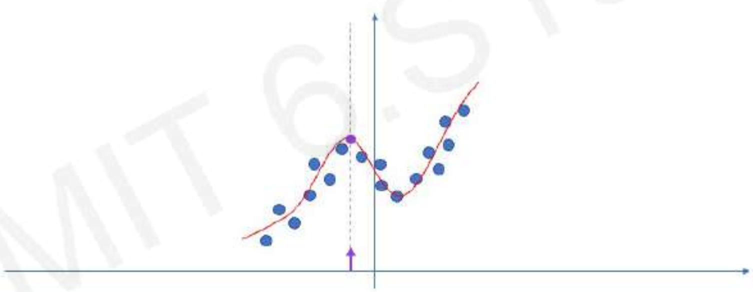
Neural Networks as Function Approximators

Neural networks are **excellent** function approximators



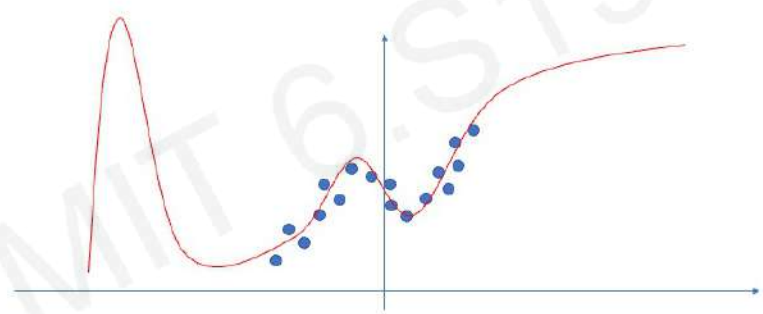
Neural Networks as Function Approximators

Neural networks are **excellent** function approximators



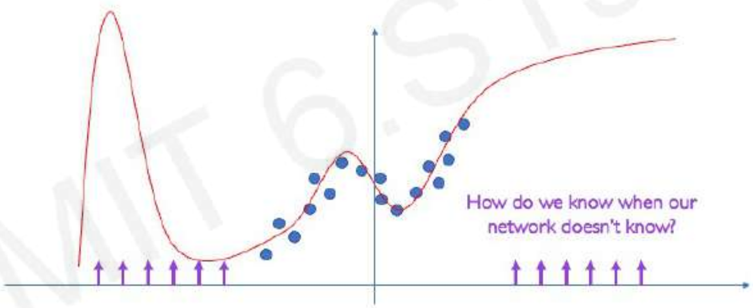
Neural Networks as Function Approximators

Neural networks are **excellent** function approximators

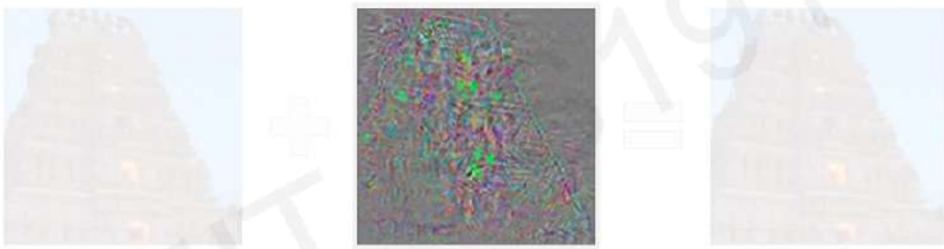


Neural Networks as Function Approximators

Neural networks are **excellent** function approximators
…**when they have training data**



Adversarial Attacks on Neural Networks



Original image | Perturbations | Adversarial example

Temple (97%) | | Ostrich (98%)

Adversarial Attacks on Neural Networks

Remember:

We train our networks with gradient descent

$$W \leftarrow W - \eta \frac{\partial J(W, x, y)}{\partial W}$$

"How does a small change in weights decrease our loss"

Adversarial Attacks on Neural Networks

Remember:

We train our networks with gradient descent

$$W \leftarrow W - \eta \frac{\partial J(W, x, y)}{\partial W}$$

"How does a small change in weights decrease our loss"

Adversarial Attacks on Neural Networks

Remember:

We train our networks with gradient descent

$$W \leftarrow W - \eta \frac{\partial J(W, x, y)}{\partial W}$$

Fix your image x, and true label y

"How does a small change in weights decrease our loss"

Adversarial Attacks on Neural Networks

Adversarial Image:

Modify image to increase error

$$x \leftarrow x + \eta \frac{\partial J(W, x, y)}{\partial x}$$

"How does a small change in the input increase our loss"

Adversarial Attacks on Neural Networks

Adversarial Image:

Modify image to increase error

$$\boxed{x} \leftarrow \boxed{x} + \eta \frac{\partial J(W, \boxed{x}, y)}{\partial \boxed{x}}$$

"How does a small change in the input increase our loss"

Adversarial Attacks on Neural Networks

Adversarial Image:

Modify image to increase error

$$\boxed{x} \leftarrow \boxed{x} + \eta \frac{\partial J(\boxed{W}, \boxed{x}, \boxed{y})}{\partial \boxed{x}}$$

Fix your weights θ, and true label y

"How does a small change in the input increase our loss"

Synthesizing Robust Adversarial Examples

classified as turtle classified as rifle classified as other

Algorithmic Bias

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk – A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

6.S191 Lab

Neural Network Limitations…

- Very **data hungry** (eg. often millions of examples)
- **Computationally intensive** to train and deploy (tractably requires GPUs)
- Easily fooled by **adversarial examples**
- Can be subject to **algorithmic bias**
- Poor at **representing uncertainty** (how do you know what the model knows?)
- Uninterpretable **black boxes**, difficult to trust
- Often require **expert knowledge** to design, fine tune architectures
- Difficult to **encode structure** and prior knowledge during learning
- **Extrapolation**: struggle to go beyond the data

Neural Network Limitations…

- Very **data hungry** (eg. often millions of examples)
- **Computationally intensive** to train and deploy (tractably requires GPUs)
- Easily fooled by **adversarial examples**
- Can be subject to **algorithmic bias**
- Poor at **representing uncertainty** (how do you know what the model knows?)
- Uninterpretable **black boxes**, difficult to trust
- Often require **expert knowledge** to design, fine tune architectures
- Difficult to **encode structure** and prior knowledge during learning
- **Extrapolation**: struggle to go beyond the data

6.S191 Lab

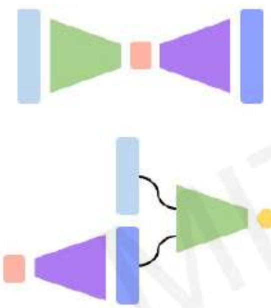
Neural Network Limitations…

- Very **data hungry** (eg. often millions of examples)
- **Computationally intensive** to train and deploy (tractably requires GPUs)
- Easily fooled by **adversarial examples**
- Can be subject to **algorithmic bias**
- Poor at **representing uncertainty** (how do you know what the model knows?)
- Uninterpretable **black boxes**, difficult to trust
- Often require **expert knowledge** to design, fine tune architectures
- Difficult to **encode structure** and prior knowledge during learning
- **Extrapolation**: struggle to go beyond the data

New Frontiers I:
Generative AI & Diffusion Models

The Landscape of Generative Modeling

Lecture 4:
VAEs and GANs

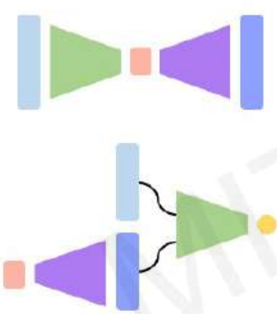


Limitations
- Mode collapse
- Generating OOD
- Hard to train

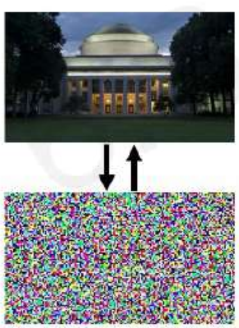
Challenges
- Stability
- Efficiency
- Quality
- Novelty

The Landscape of Generative Modeling

Lecture 4:
VAEs and GANs



Diffusion Models

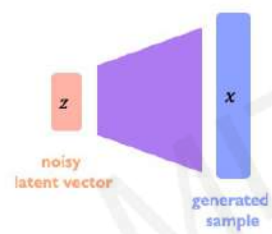


Text-to-Image

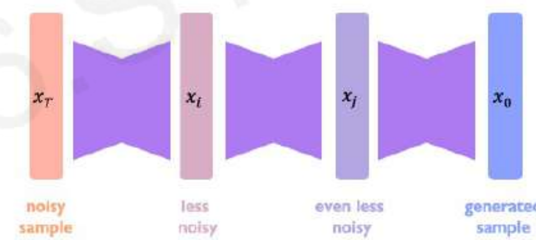
"Two cats doing research"

Diffusion Models

VAEs/GANs: Generating samples in one-shot directly from low-dimensional latent variables

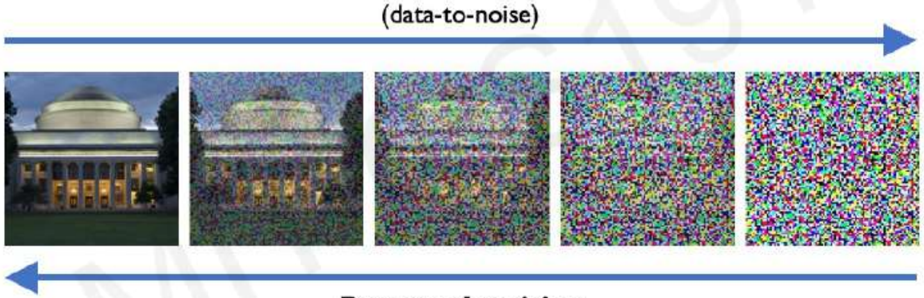


Diffusion: Generating samples iteratively by repeatedly refining and removing noise



The Diffusion Process

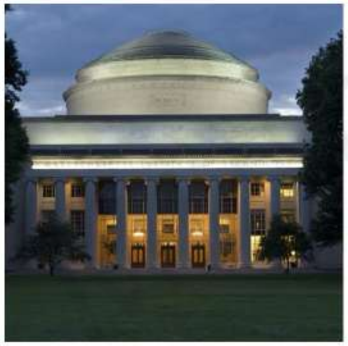
Forward noising
(data-to-noise)

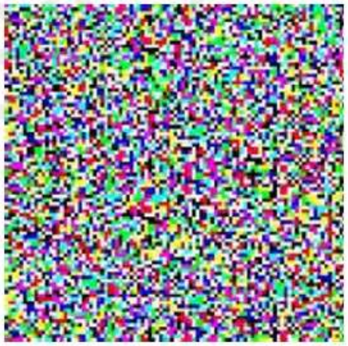


Reverse denoising
(noise-to-data)

Forward Noising

Step 1: Given an image (left), sample a random noise pattern (right)





Forward Noising

Step 2: Progressively add more and more of the noise to your image

T = 0 | T = 1 | T = 2 | T = 3 | T = 4

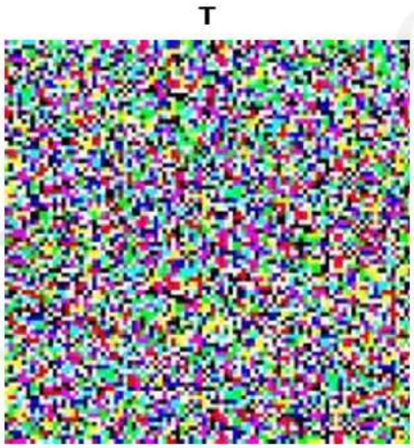
T = 0	T = 1	T = 2	T = 3	T = 4
100% image 0% noise	75% image 25% noise	50% image 50% noise	25% image 75% noise	0% image 100% noise

Reverse Denoising

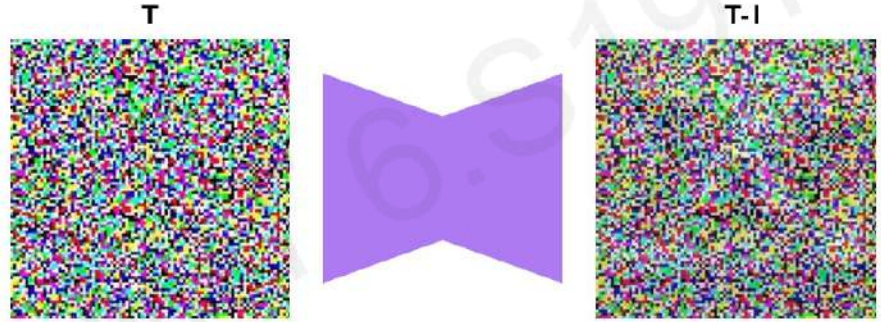
Goal: Given image at **T**, can we **learn** to estimate image at **T-1**?

How can we train this network?

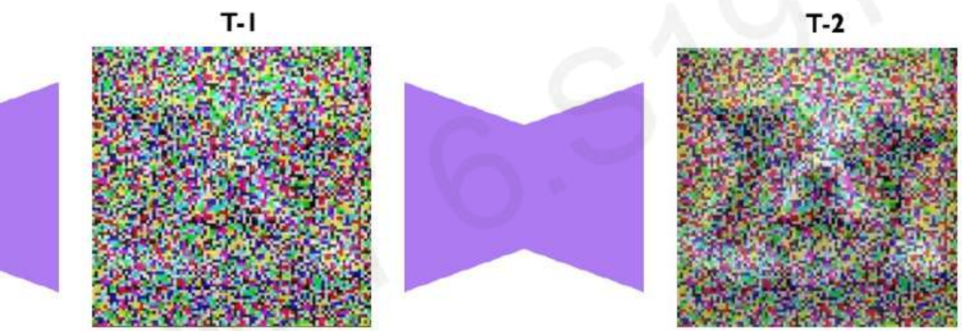
Sampling Brand New Generations



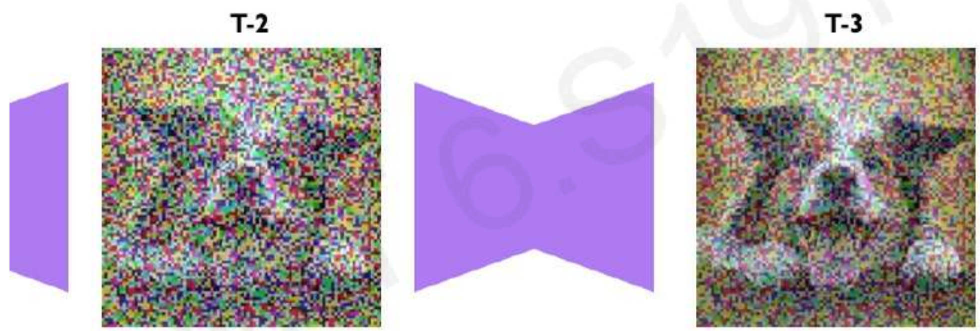
Sampling Brand New Generations



Sampling Brand New Generations

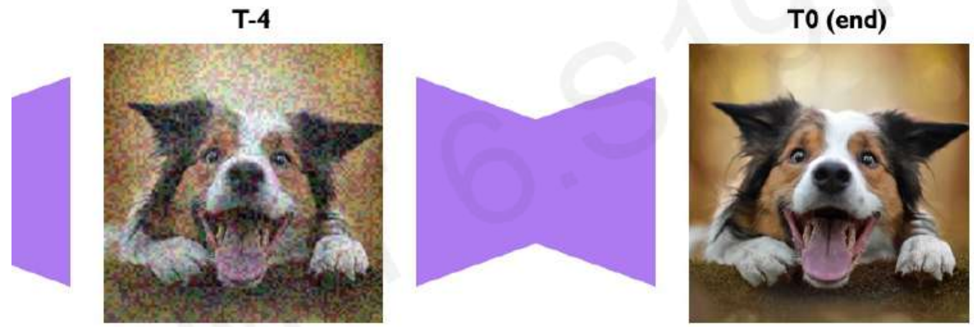


Sampling Brand New Generations

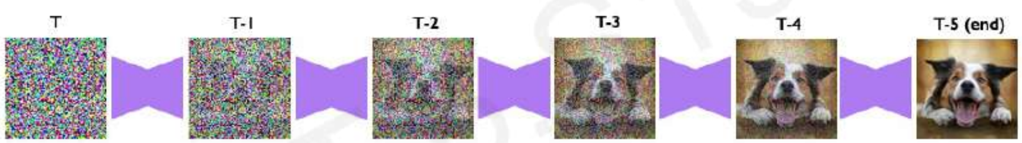


Sampling Brand New Generations

Sampling Brand New Generations

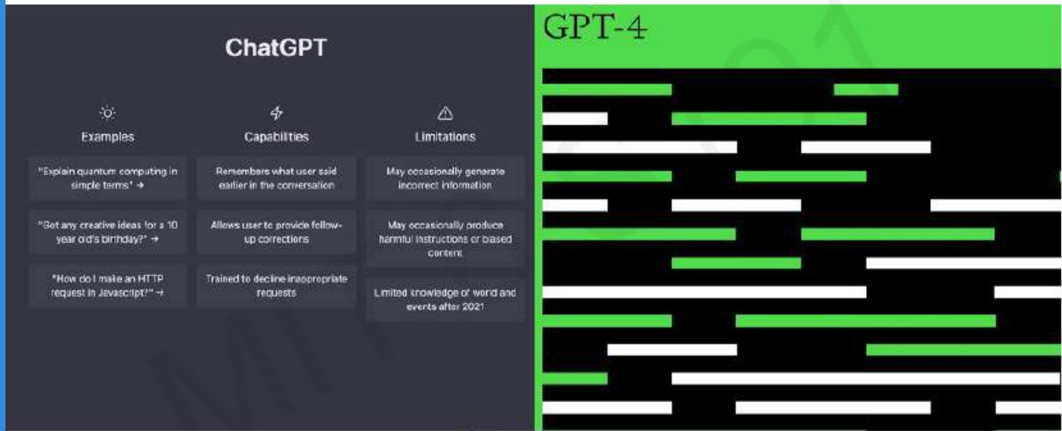


Sampling Brand New Generations



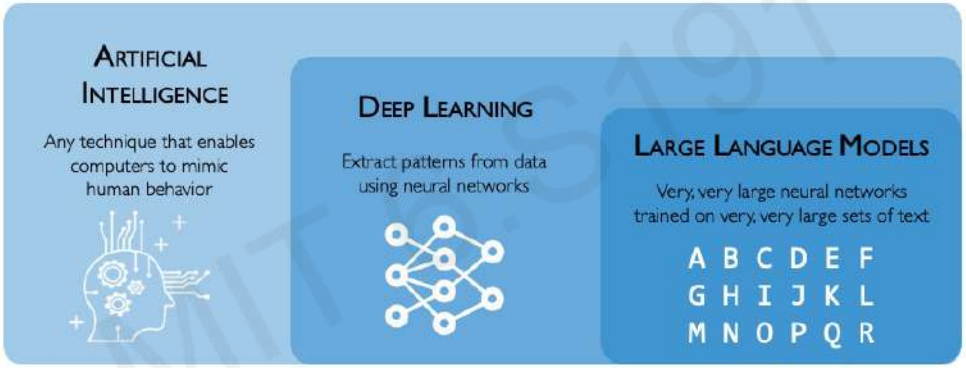
New Frontiers II:
Large Language Models

Large Language Models (LLMs) and the World



OpenAI

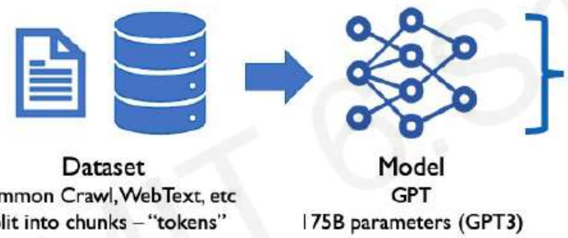
What are LLMs?



6.S191 Guest Lectures!

How do LLMs like GPT work?

Training:



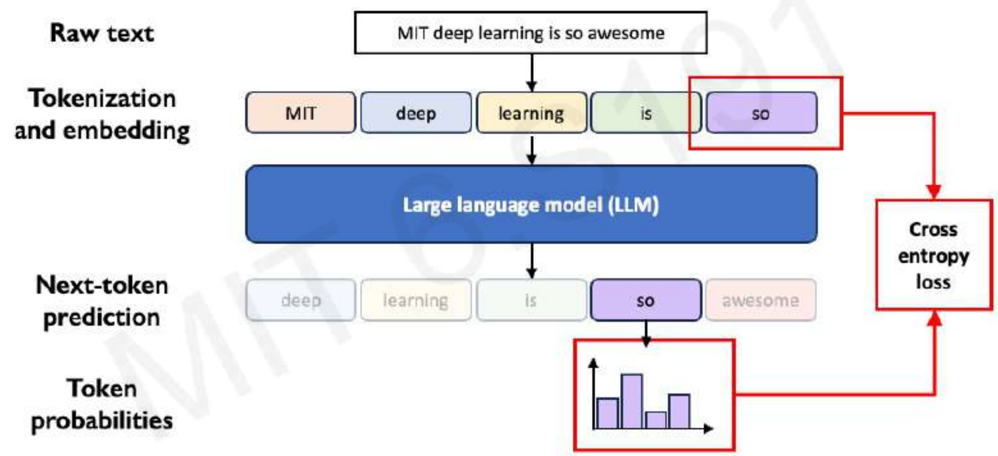
Task and Objective:

Given a sequence of tokens, predict the next token.

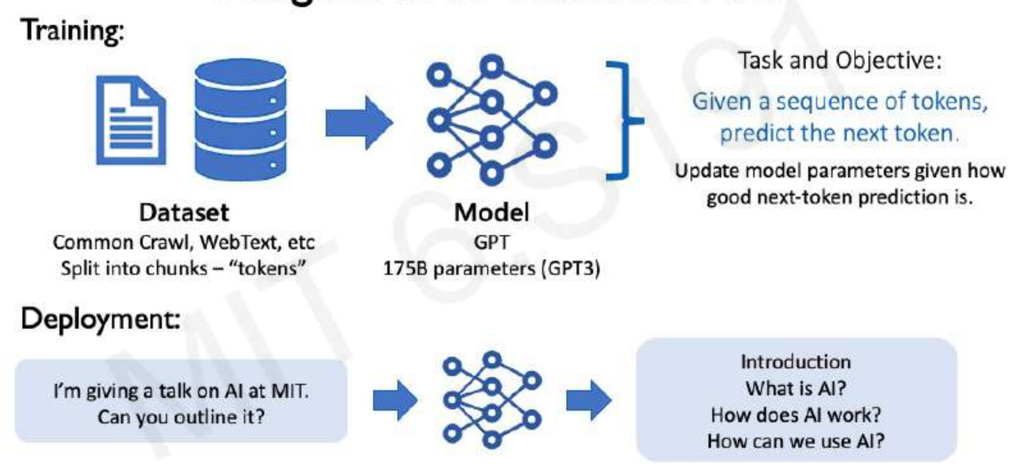
Update model parameters given how good next-token prediction is.

How does next token prediction work? 🤔

Next Token Prediction



Using LLMs to Generate Text

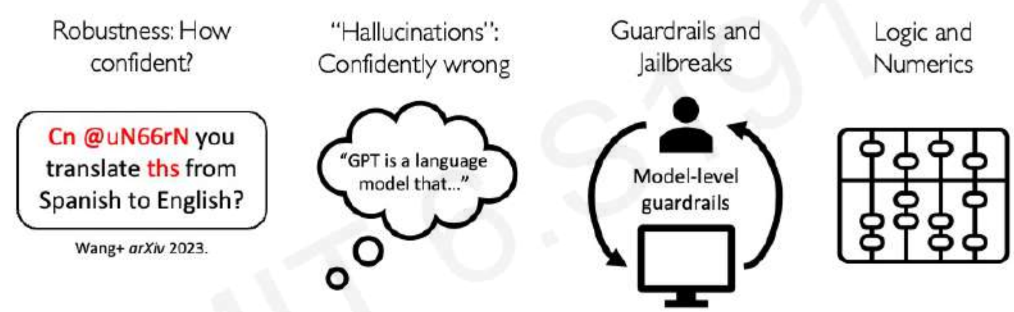


What capabilities do LLMs have?

Capabilities that are feasible and reliable now:

LLMs like GPT have shown mastery over natural language.

Limitations of LLMs



Key challenges motivated by the high-level thinking process:
robustness + confidence; long-term planning; logic and discovery

What can LLMs do?
Emergent Abilities with Scale.

An ability is **emergent** if it is not present in smaller models but is present in larger models.