CS167: Machine Learning

Large Language Model (LLM) application with PyTorch

Tuesday, May 7th, 2024

Announcements

- **Project#2**
  - Released and due on **05/12 (Sunday) by 11:59pm**
- **Quiz#3**
  - Released and due on **05/14 (Tuesday) by 11:59pm**

Course Evaluation (Section# 9:30am-10:45am)

- Finish the form below:

https://drake.qualtrics.com/jfe/form/SV_2meHaNblWCtnL9A

Course Evaluation (Section# 11am-12:15pm)

- Finish the form below:

https://drake.qualtrics.com/jfe/form/SV_6hdNkyRzLhDuyKq

Transformers

Recap: Transformers

- In 2017, a new mechanism is introduced for context learning called **attention mechanism**
  - more precisely, self-attention

- It takes less time to train advantage
- Transfer learning on a new task is possible advantage
- In subsequent years, it revolutionized the field of AI

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

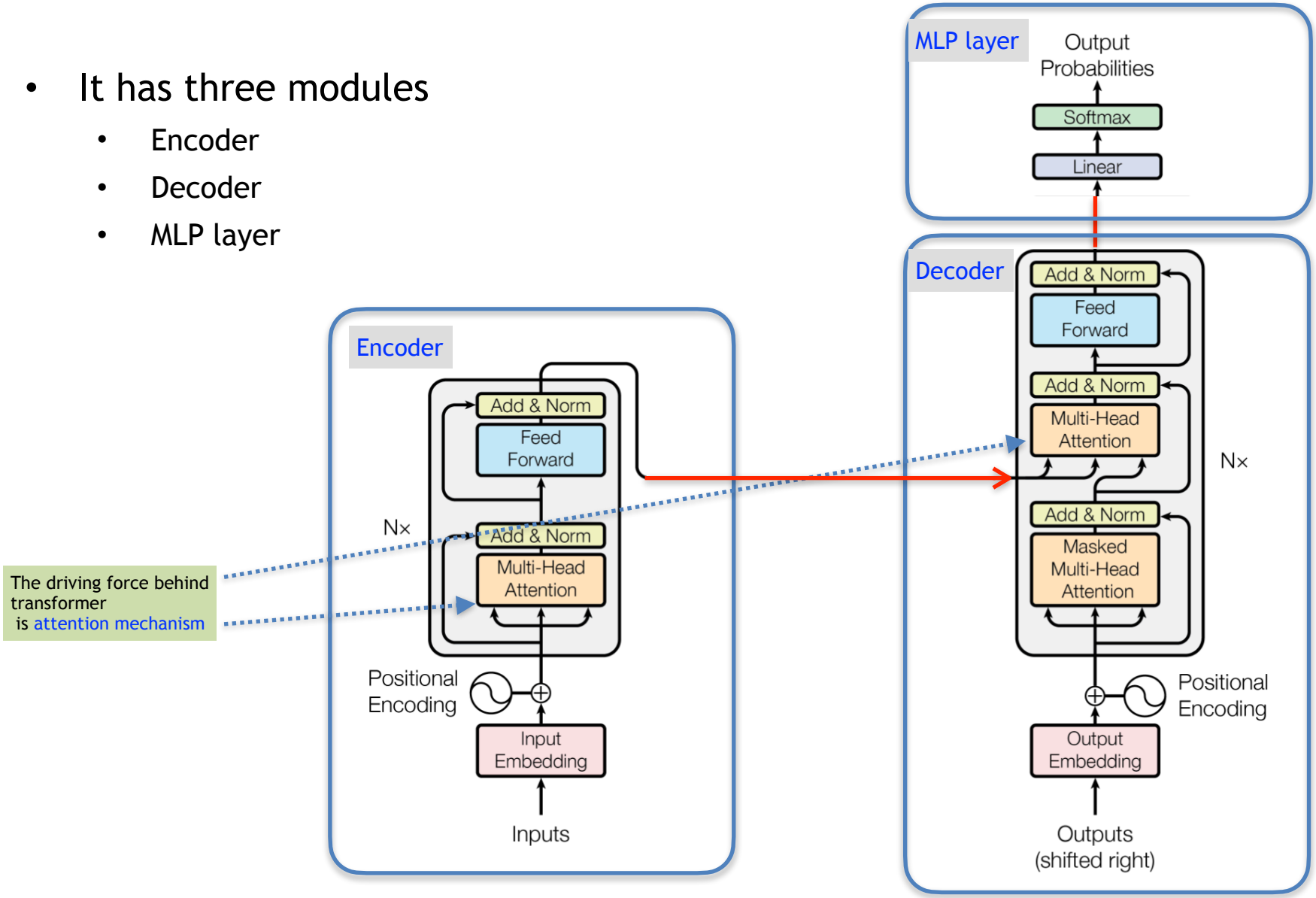
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Attention is all you need - NeurIPS'2017

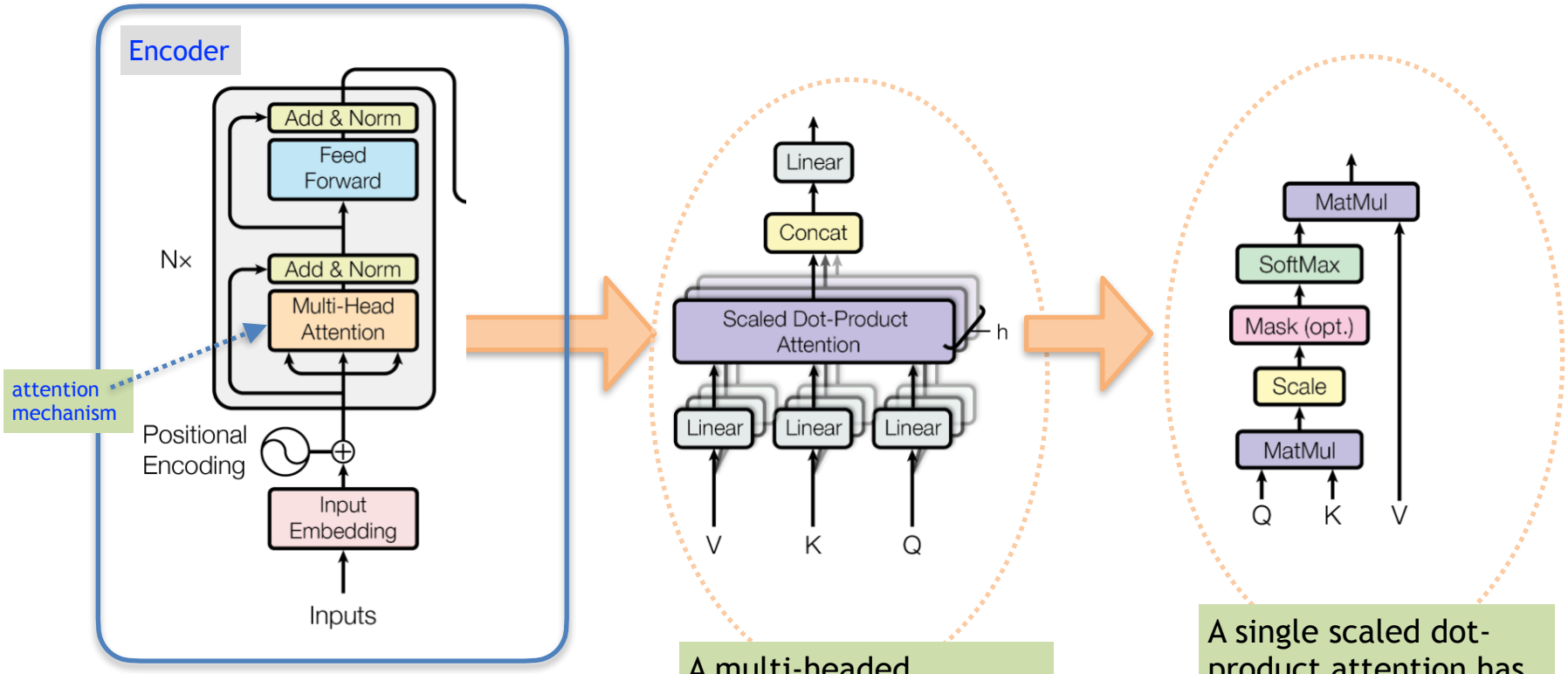
Recap: Transformers

- It has three modules
  - Encoder
  - Decoder
  - MLP layer



Recap: Transformer Encoder Module

- Lets focus on the encoder to understand what is this attention mechanism



Encoder has a **multi-headed attention**

A multi-headed attention has several **scaled dot-product attention**

A single scaled dot-product attention has sequential matrix matrix multiplications in terms of Query (Q), Key (K), and Value (V)

Large Language Models (LLM)

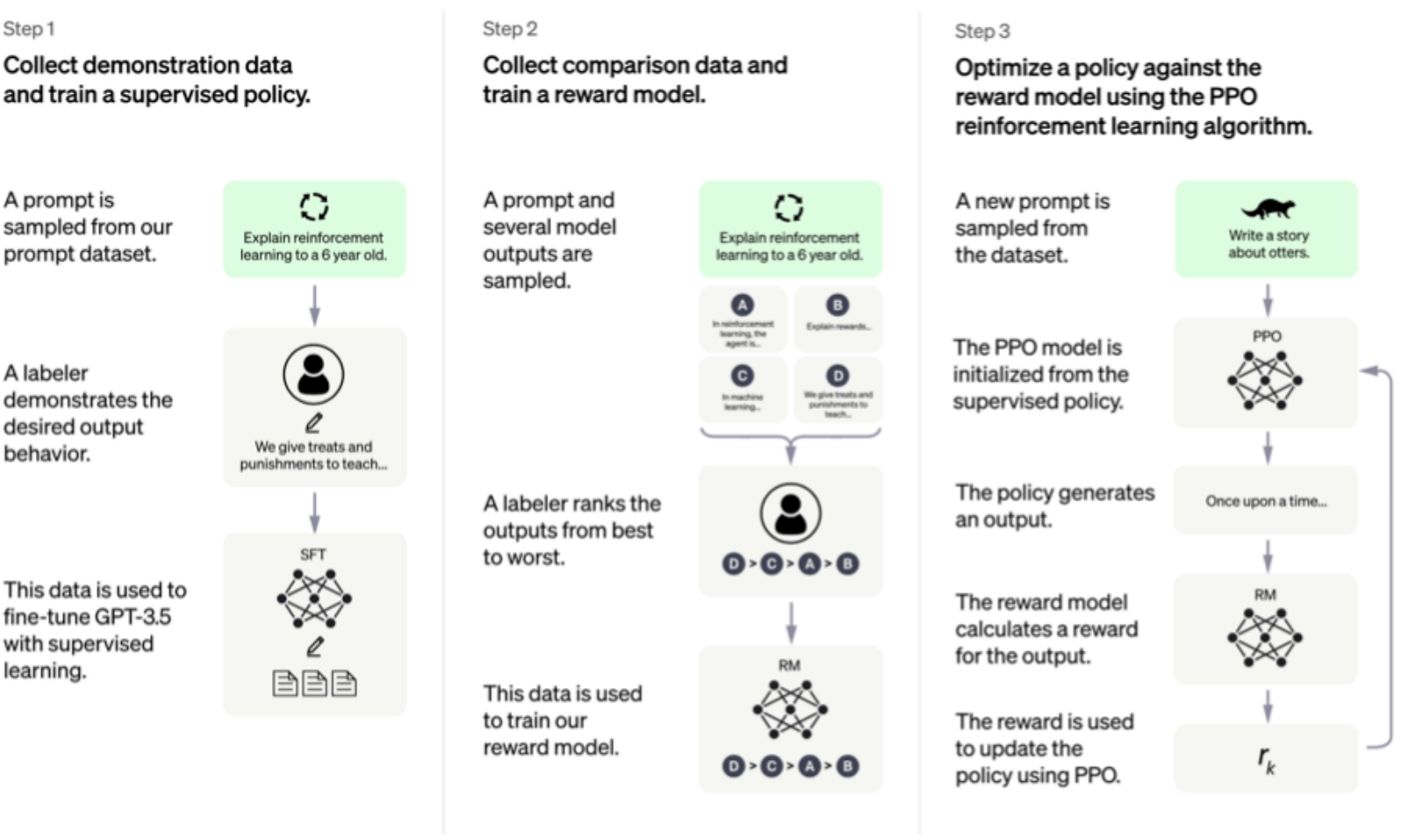
- LLMs are capable of exhibiting human or near-human-like performances in solving various natural language processing (NLP) tasks, such as
  - sentiment analysis
  - paraphrasing
  - question-answering
  - translation
  - text generation

Large Language Models (LLM)

- Popular large language models (LLMs) with their years of inception:
  - GPT (100M learnable parameters, introduced in 2018 by OpenAI)
  - BERT (300M learnable parameters, introduced in 2018)
  - GPT-2 (1.5B learnable parameters, introduced in 2019 by OpenAI)
  - Megatron-LM (8.0B learnable parameters, introduced in 2019 by NVidia)
  - T5 (11.0B learnable parameters, introduced in 2020 by Google)
  - T-NLG (17.0B learnable parameters, introduced in 2020 by Microsoft)
  - GPT-3 (175.0B learnable parameters, introduced in 2020 by OpenAI)
  - ChatGPT (introduced in 2022 by OpenAI)

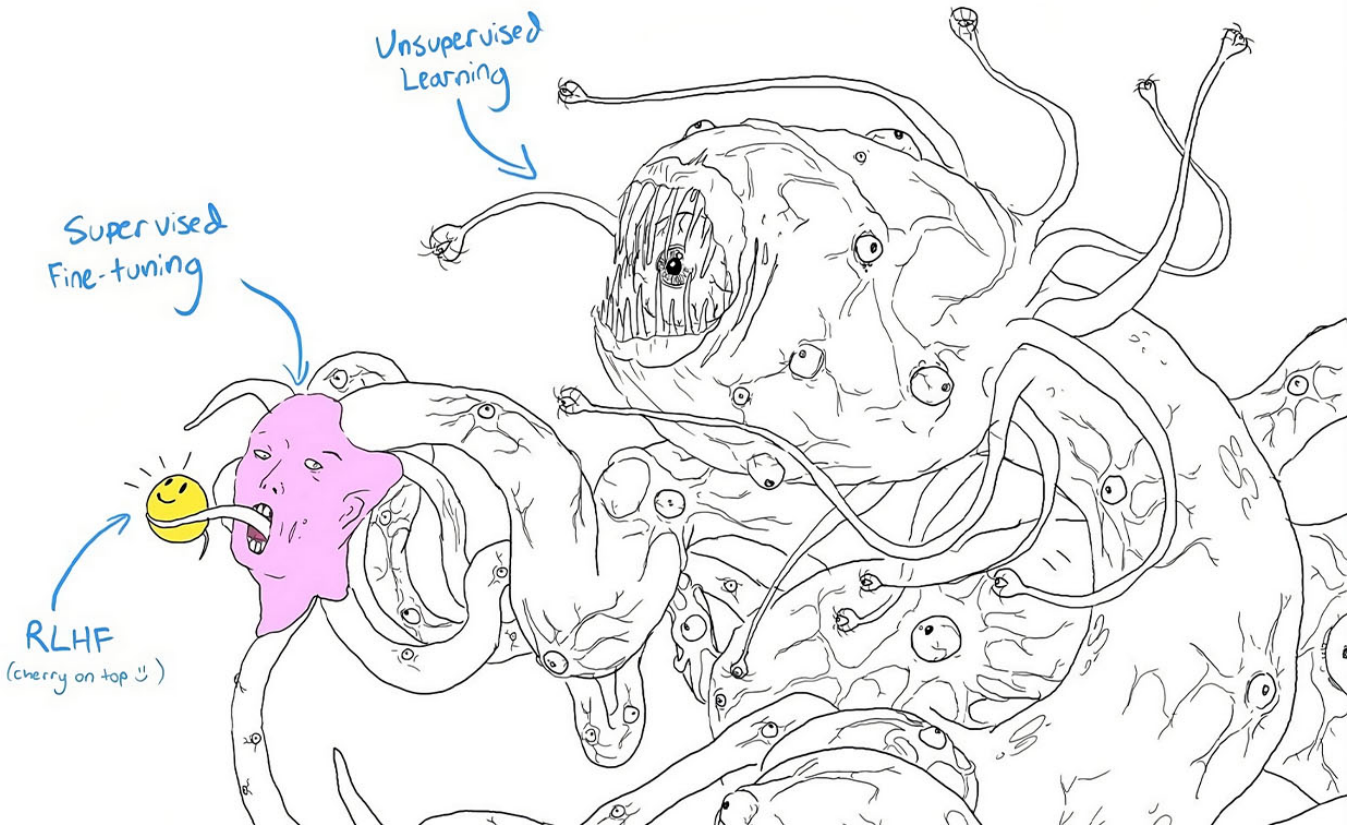
LLM: ChatGPT

- ChatGPT



LLM: ChatGPT

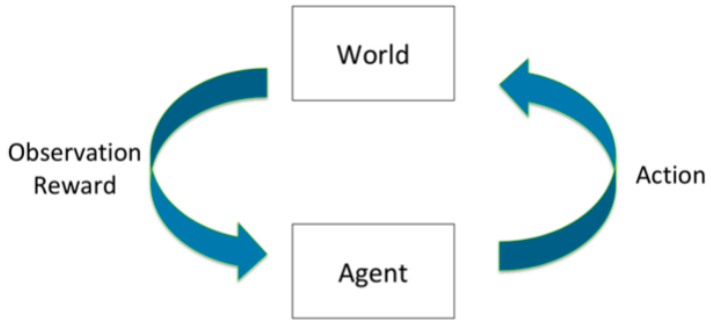
- An important component in ChatGPT is its use of Reinforcement learning with Human Feedback (RLHF)



https://knowyourmeme.com/memes/shoggoth-with-smiley-face-artificial-intelligence

LLM: ChatGPT

- Types of Machine Learning:
  - Supervised Learning: Inferring a function from labeled training data. The training data consist of a set of training examples.
  - Unsupervised Learning: Draw inferences from datasets consisting of input data without labeled responses to unearth hidden structure in the data.
  - Reinforcement learning is the third paradigm of teaching machines with different form of labels (a.k.a rewards)

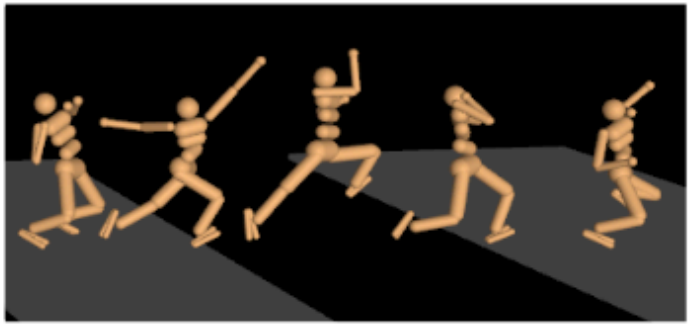


Reinforcement Learning (RL)

- Reinforcement learning is the third paradigm of teaching machines with different form of labels (a.k.a rewards)
  - we observing a rise in general-purpose solutions

AI beating Go champion

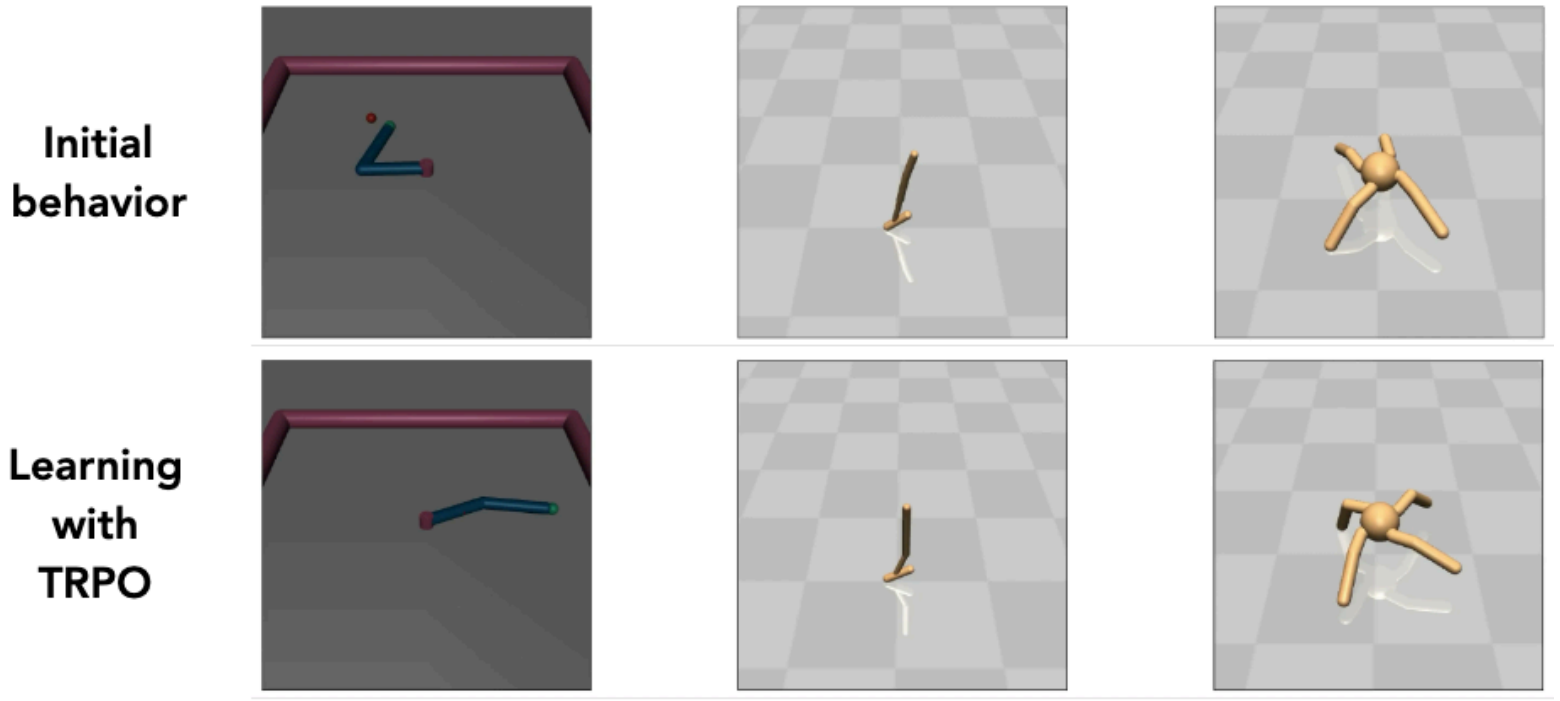
Human-level control of Atari games



Emergence of locomotion behaviors

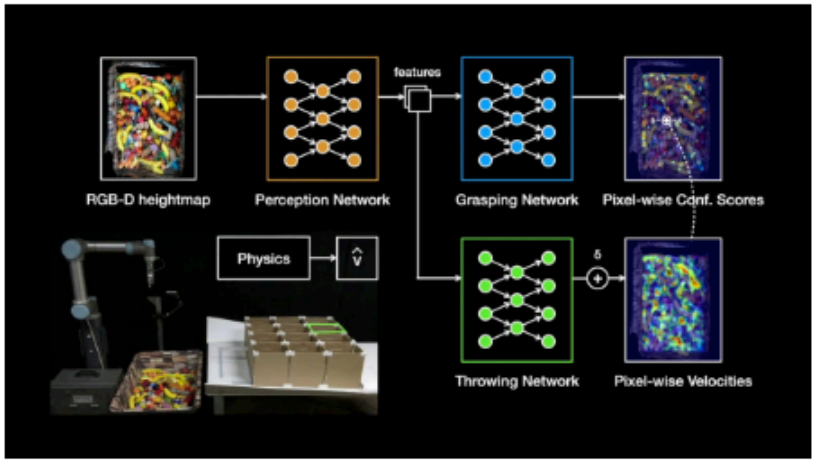
Reinforcement Learning (RL)

- Reinforcement learning is the third paradigm of teaching machines with different form of labels (a.k.a rewards)
  - Deep reinforcement learning agents show excellent general learning capabilities in virtual worlds



Reinforcement Learning (RL)

- Reinforcement learning is the third paradigm of teaching machines with different form of labels (a.k.a rewards)
  - Hybrid approaches are becoming popular in robotics



Zeng A, Song S, Lee J, Rodriguez A, Funkhouser T (2019).
TossingBot: Learning to Throw Arbitrary Objects with Residual Physics.

LLM applications: PyTorch Code

- LLMs are capable of exhibiting human or near-human-like performances in solving various natural language processing (NLP) tasks, such as
  - sentiment analysis
    - eg, given a text, classifying whether its a positive or negative sentiment
  - paraphrasing
    - eg, given a two phrase, determining if one is rephrasing of another
  - question-answering
    - eg, given large paragraph of textual context, ask specific question within the paragraph
  - translation
    - eg, translate a sentence from English to French
  - text generation
    - eg, although, we trained RNN for text generation. You can do better with a Transform

LLM applications: Sentiment Classification using PyTorch

- LLMs are capable of exhibiting human or near-human-like performances in solving various natural language processing (NLP) tasks, such as

  - sentiment analysis
    - eg, given a text, classifying whether its a positive or negative sentiment

- Use the following link for PyTorch code:

  https://github.com/alimoorreza/CS167-sp24-notes/blob/main/Day26a_LLM_sentiment_classfication.ipynb

LLM applications: Paraphrase Detection using PyTorch

- LLMs are capable of exhibiting human or near-human-like performances in solving various natural language processing (NLP) tasks, such as
  - paraphrasing
    - eg, given a two phrase, determining if one is rephrasing of another

- Use the following link for PyTorch code:

https://github.com/alimoorreza/CS167-sp24-notes/blob/main/Day26b_LLM_paraphrase_classification.ipynb

LLM applications: Question-Answering using PyTorch

- LLMs are capable of exhibiting human or near-human-like performances in solving various natural language processing (NLP) tasks, such as

  - question-answering
    - eg, given large paragraph of textual context, ask specific question within the paragraph

- Use the following link for PyTorch code:

  https://github.com/alimoorreza/CS167-sp24-notes/blob/main/Day26c_LLM_question_answering.ipynb