CS167: Machine Learning

Transformers
Large Language Model (LLM)

Thursday, May 2nd, 2024

Announcements

- **Project#2**
  - Released and due on **05/12 (Sunday) by 11:59pm**
- **Quiz#3**
  - Will be released early next week

Today's agenda

- Transformers
  - Transfer learning is possible
  - New type of network architecture
- Transformers Implementation in PyTorch

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

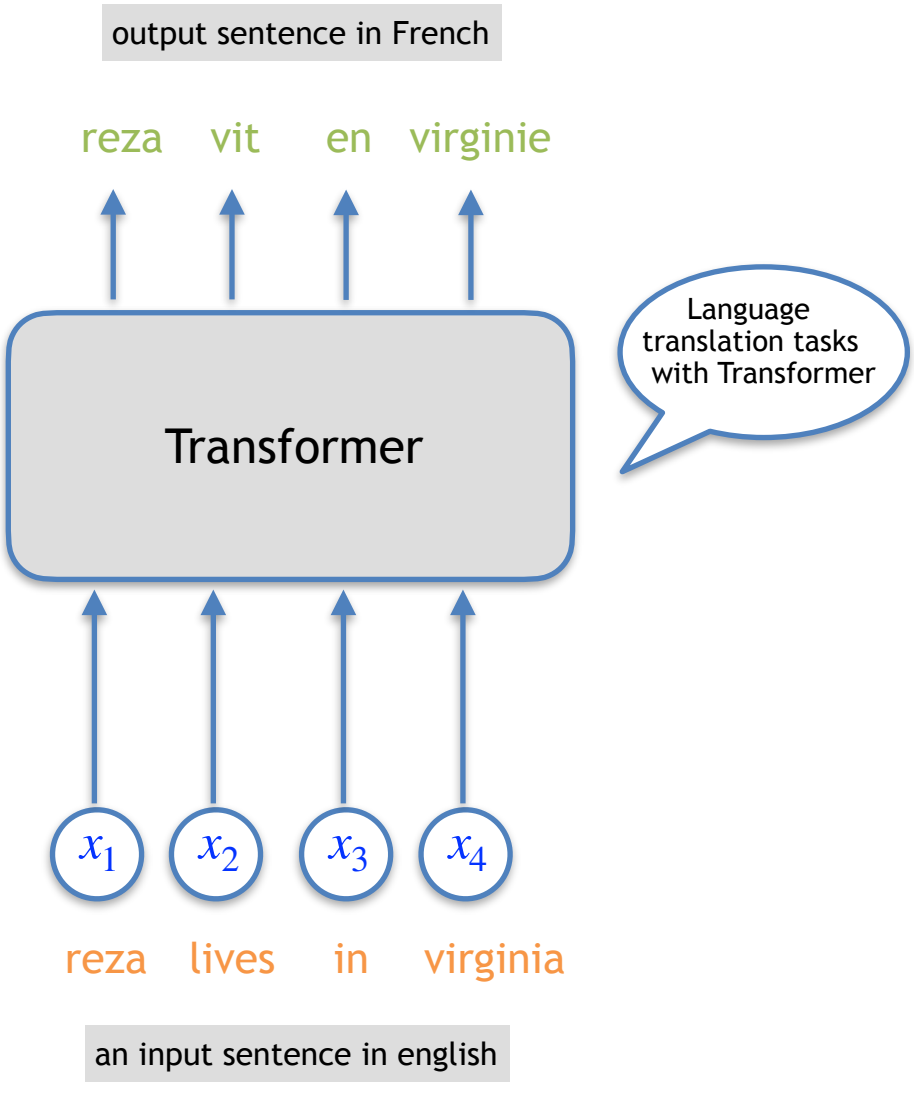
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

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.

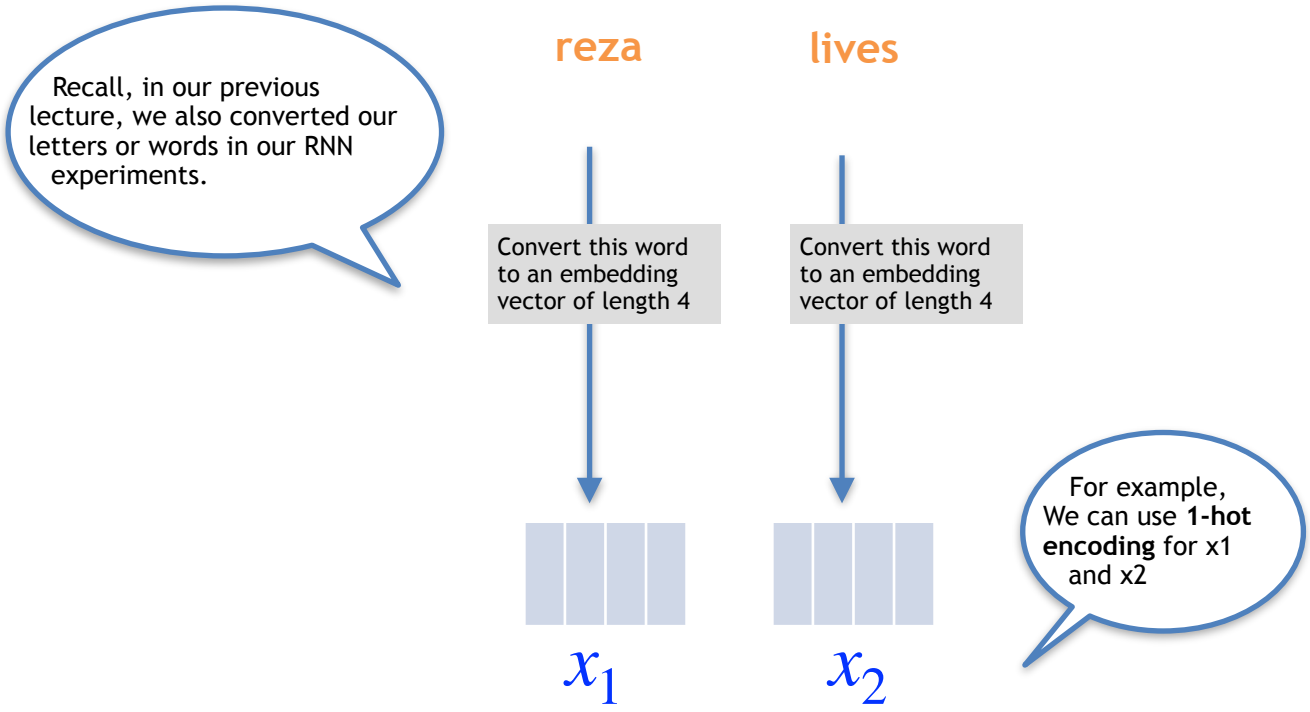
Attention is all you need - NeurIPS'2017

Recap: Transformers for Language Translation



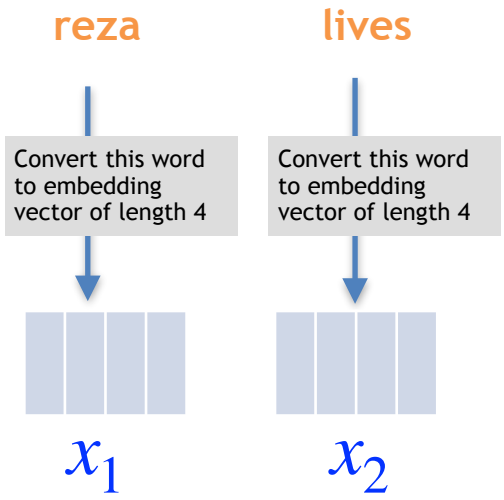
Recap: Attention Mechanism

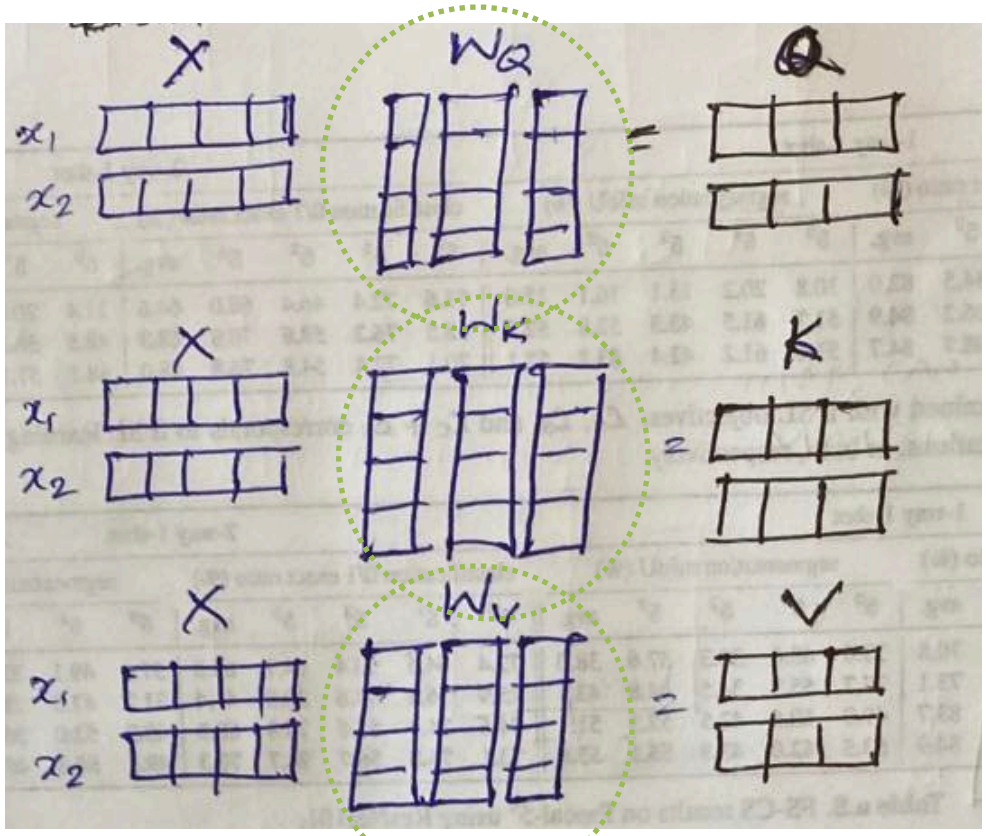
- Let's find out how to calculate the attention mechanism in a toy example
- Let's calculate attention with first two words of our sentence: "reza lives"



Recap: Attention Mechanism

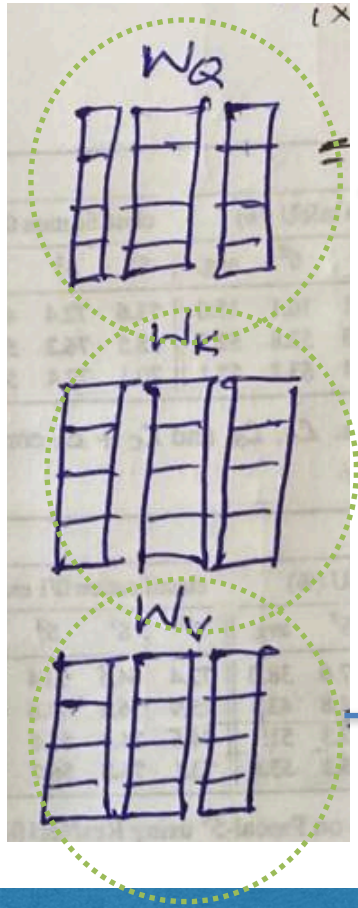
- It calculates three new matrices Q, K, and V with the help of three weight matrices W_Q,W_K, and W_V
- These three matrices (W_Q,W_K, and W_V) are learned during training



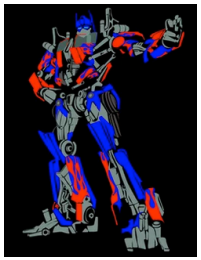


Recap: Attention Mechanism

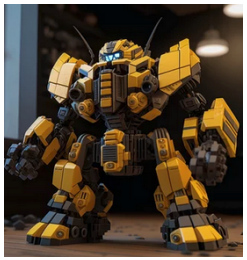
- It calculates three new matrices Q, K, and V with the help of three weight matrices W_Q,W_K, and W_V
- These three matrices (W_Q,W_K, and W_V) are learned during training



Optimus prime



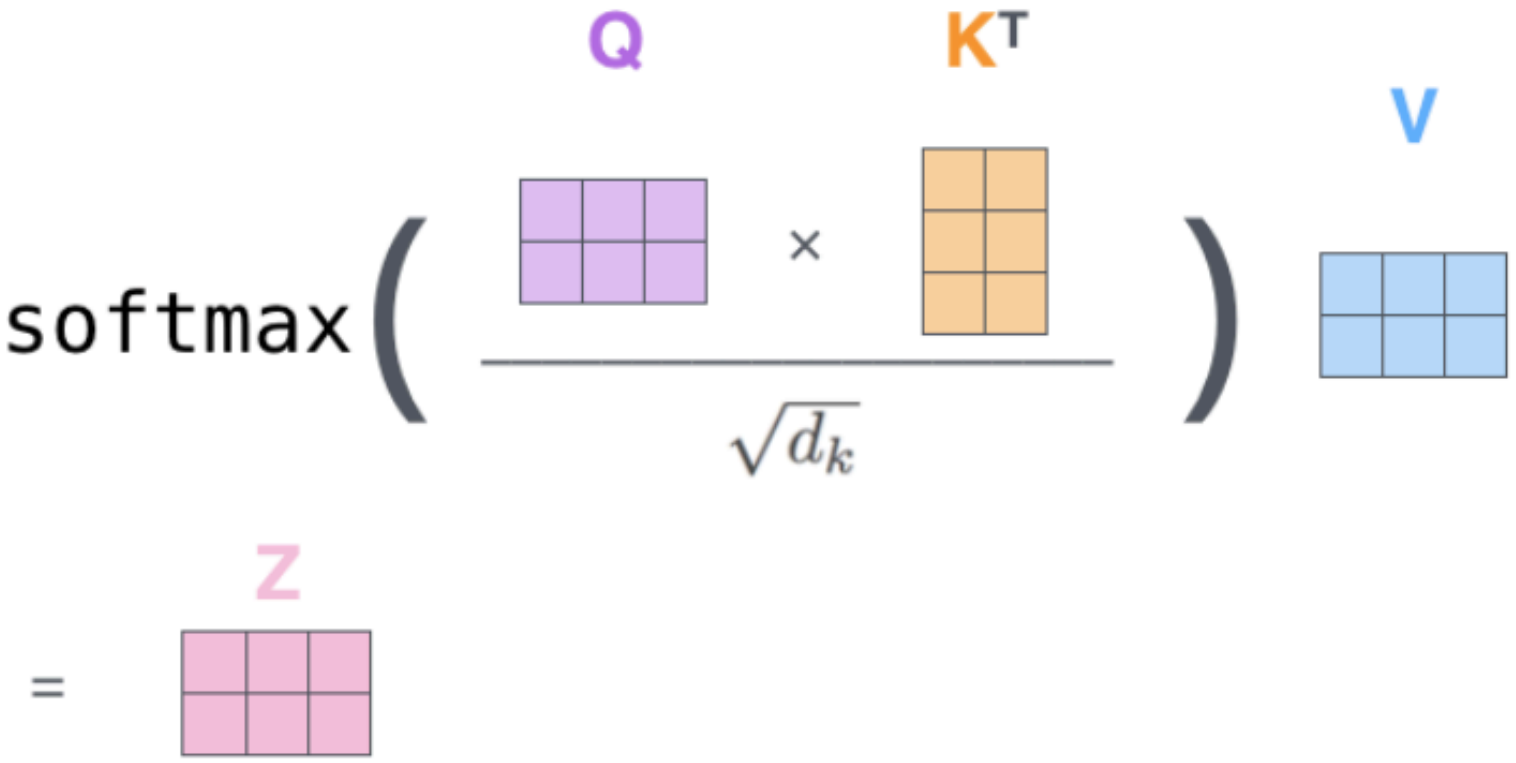
Bumble Bee



Ironhide

Recap: Attention Mechanism

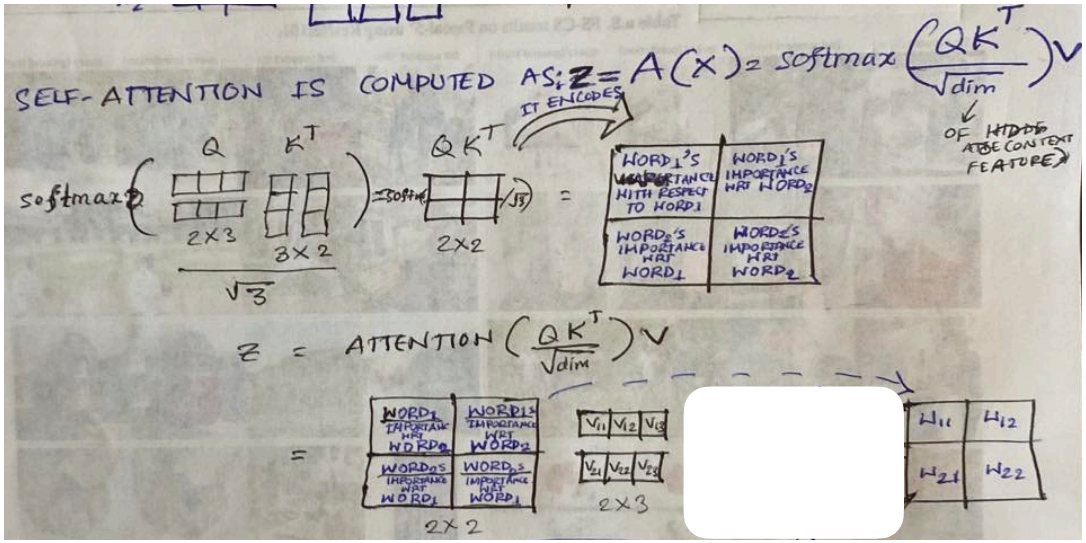
- Finally, attention is calculated using Q, K, and V matrices using the following equation:



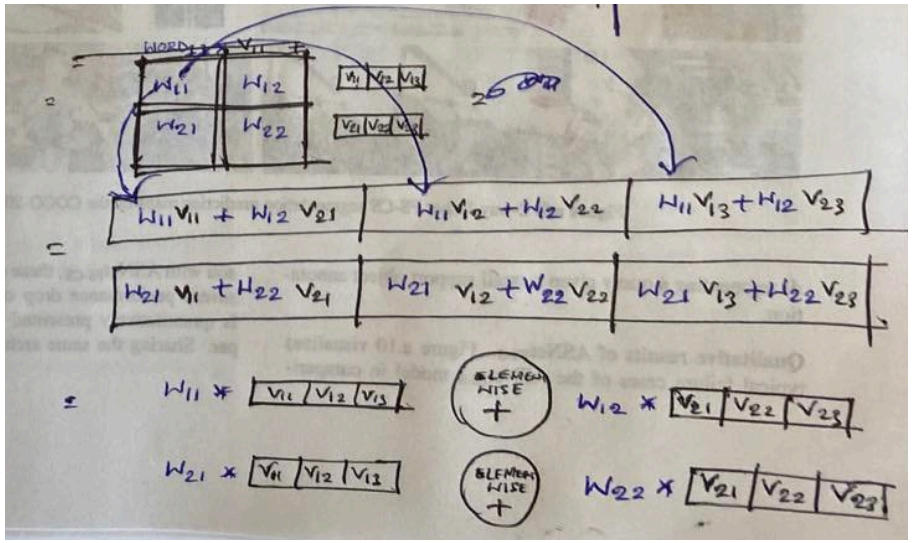
Reference: Illustrated Transformer

Attention

My hand-notes



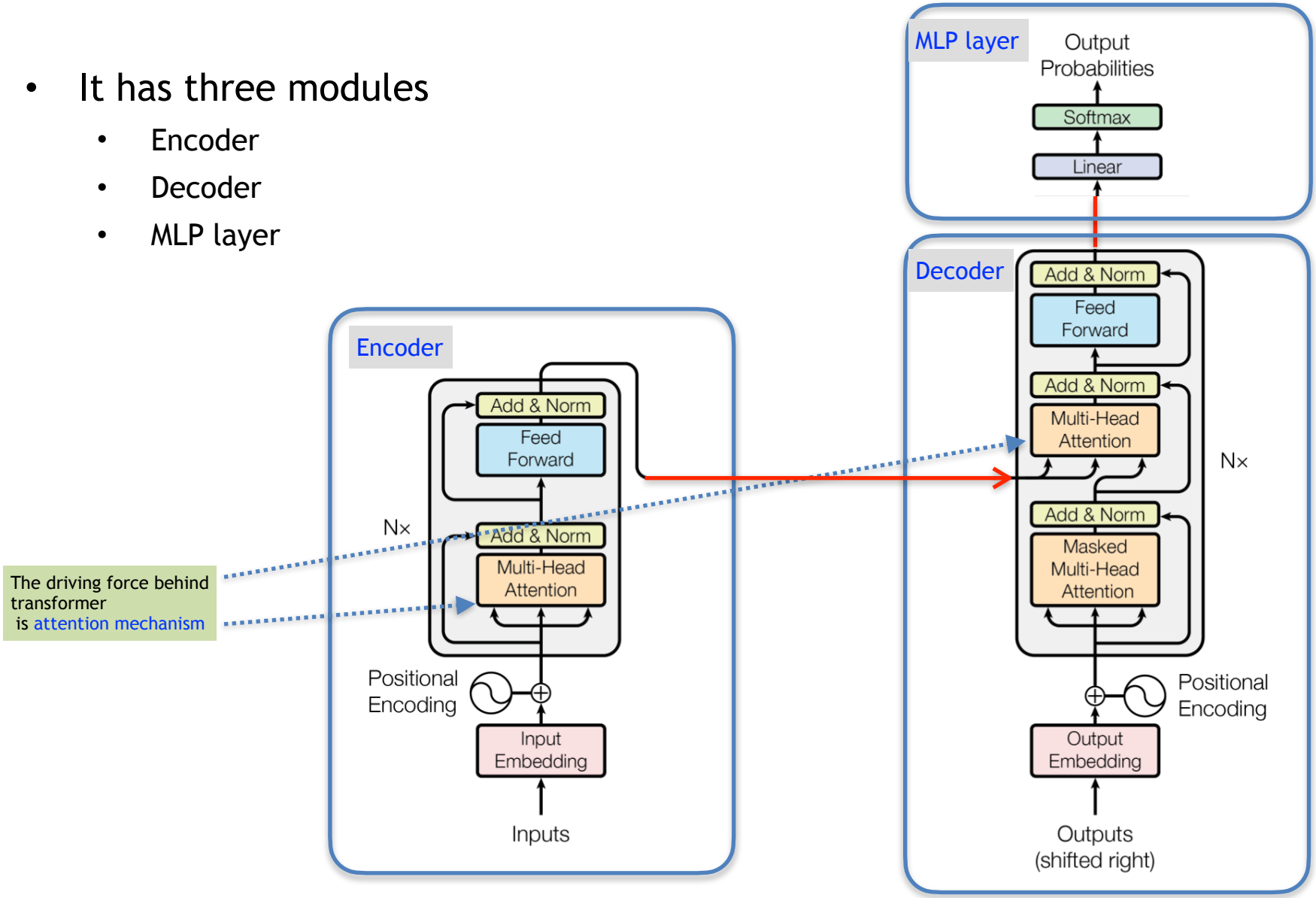
My hand-notes



Reference: Illustrated Transformer

Transformers

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

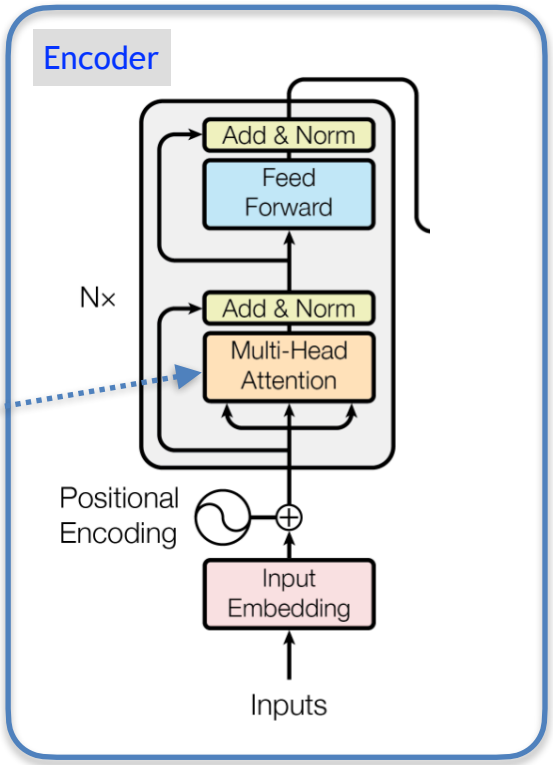


Participate in the following Poll

https://forms.gle/TiYmPhxgqf7yMJZS9

Transformers: Encoder

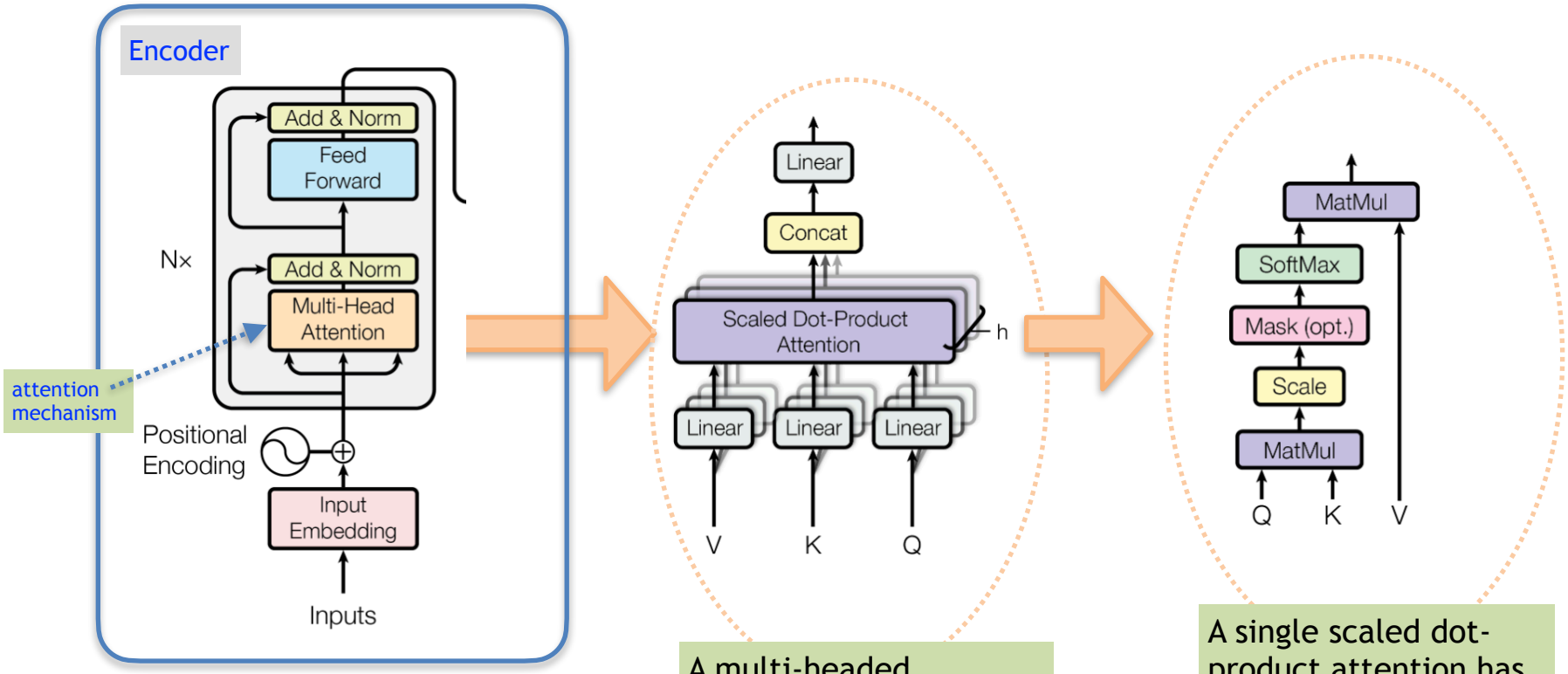
- Lets focus on the encoder to understand what is this attention mechanism



The driving force behind transformer is attention mechanism

Transformers: Encoder

- 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)

Today's agenda

- Transformers
  - Transfer learning is possible
  - New type of network architecture
- Transformers Implementation in PyTorch

Transformers Implementation in PyTorch

- PyTorch official website has an excellent tutorial demonstrating how to train a Transformer for language modeling

Open this notebook: transformer.ipynb

Today's agenda

- Transformers
  - Transfer learning is possible
  - New type of network architecture
- Transformers Implementation in PyTorch
- Large Language Model (LLM)
  - GPT (commercial)
  - BERT (research; useful for fine-tuning on a new task)
  - ChatGPT (commercial)

Large Language Models (LLM)

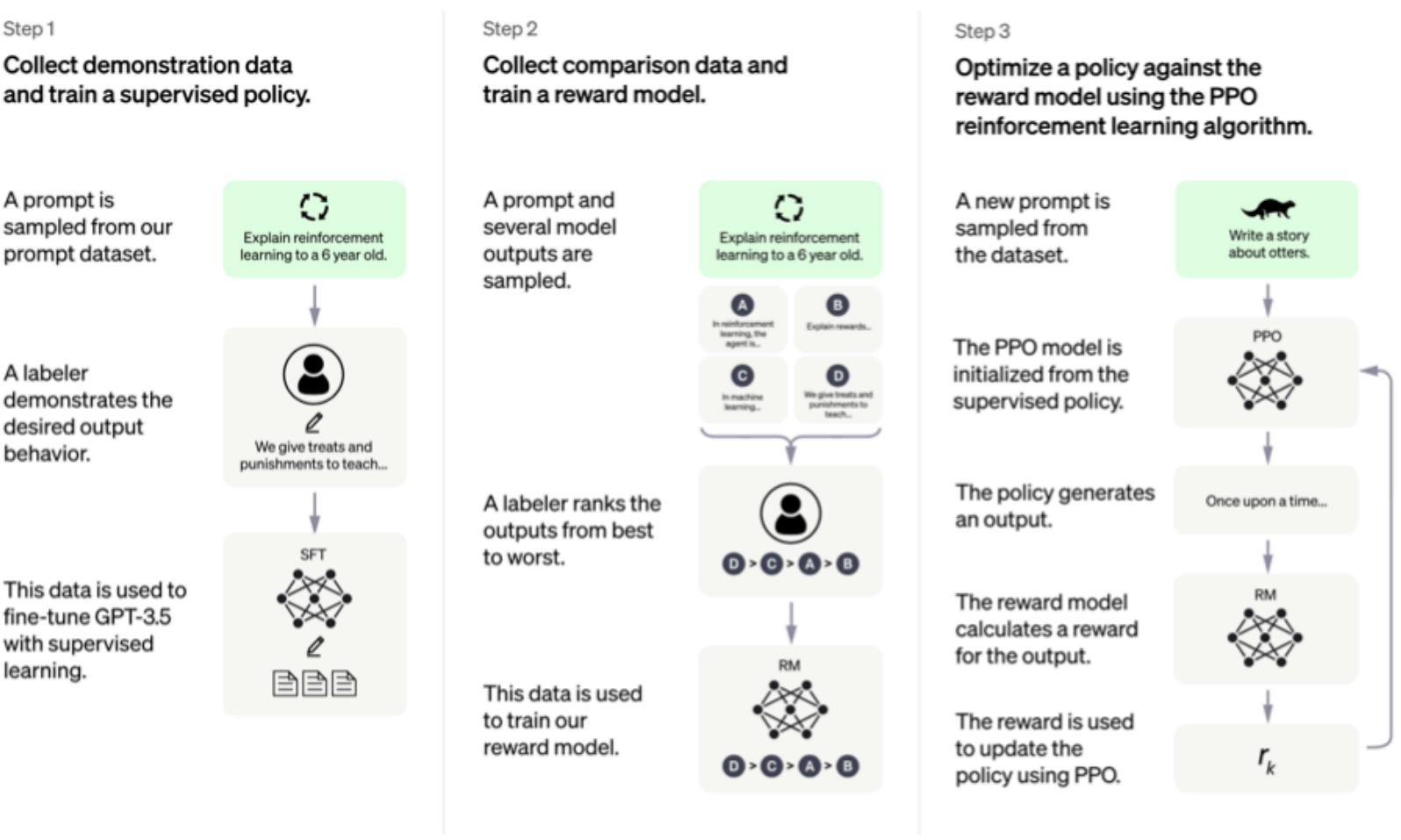
- LLMs are capable of exhibiting human or near-human-like performances in solving various natural language processing (NLP) tasks, such as
  - paraphrasing
  - question-answering,
  - translation (e.g., translate a sentence from English to French),
  - text generation
  - sentiment analysis
  - and various others.

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)

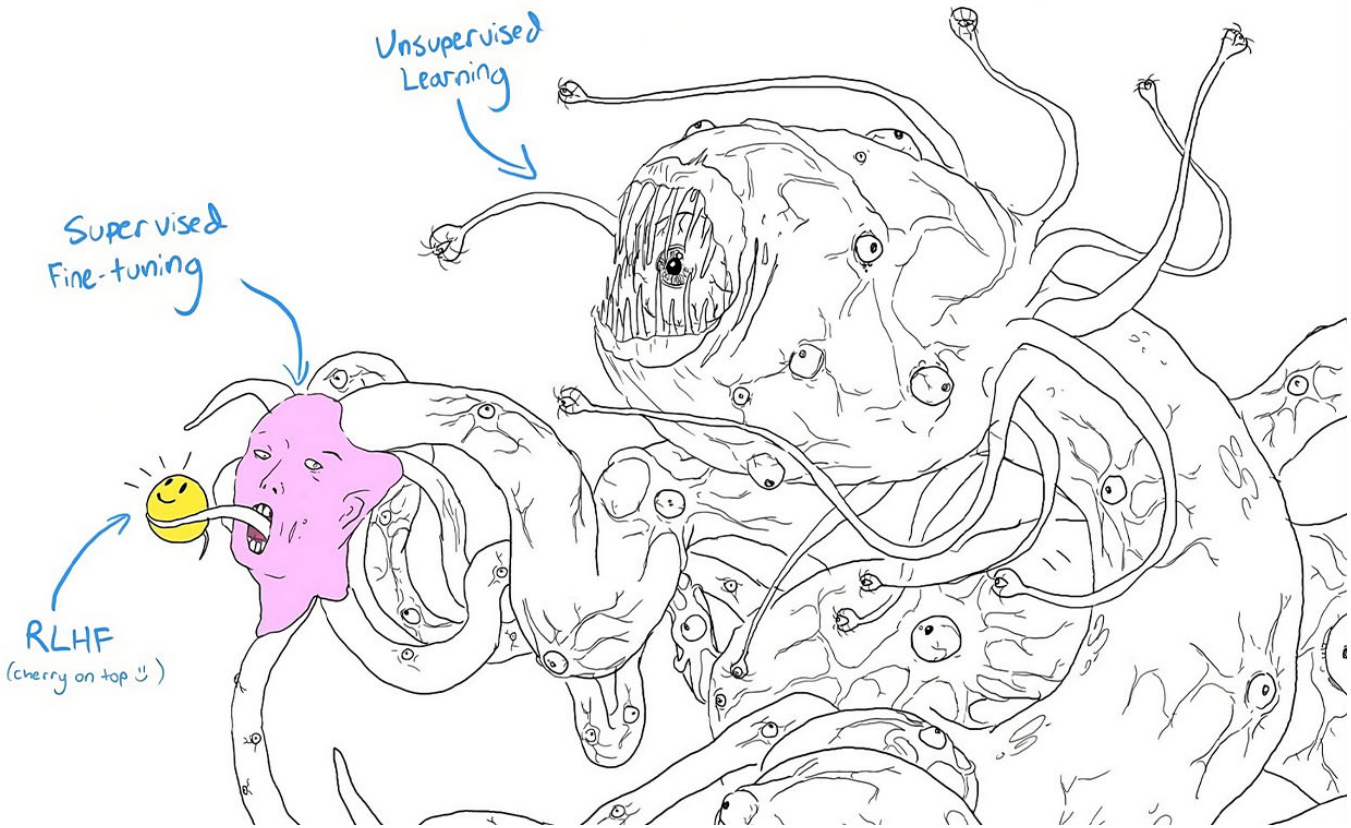
LLM: ChatGPT

- ChatGPT



LLM: ChatGPT

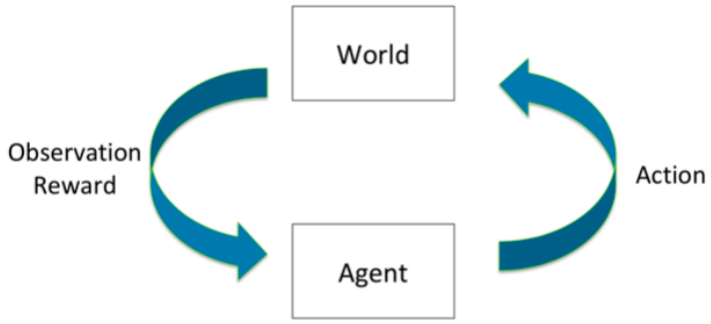
- An important component in ChatGPT is its use of Reinforcement learning with Human Feedback (RLHF)



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)

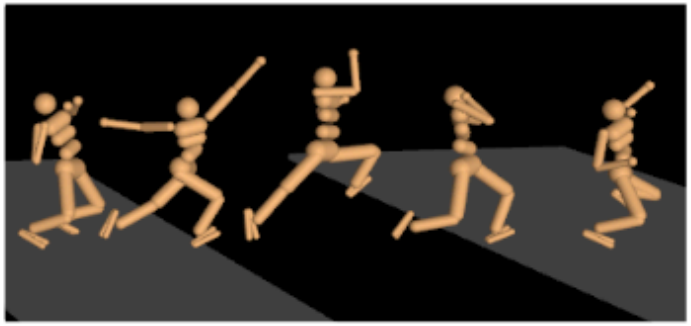


Reinforcement Learning (RL)

- 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

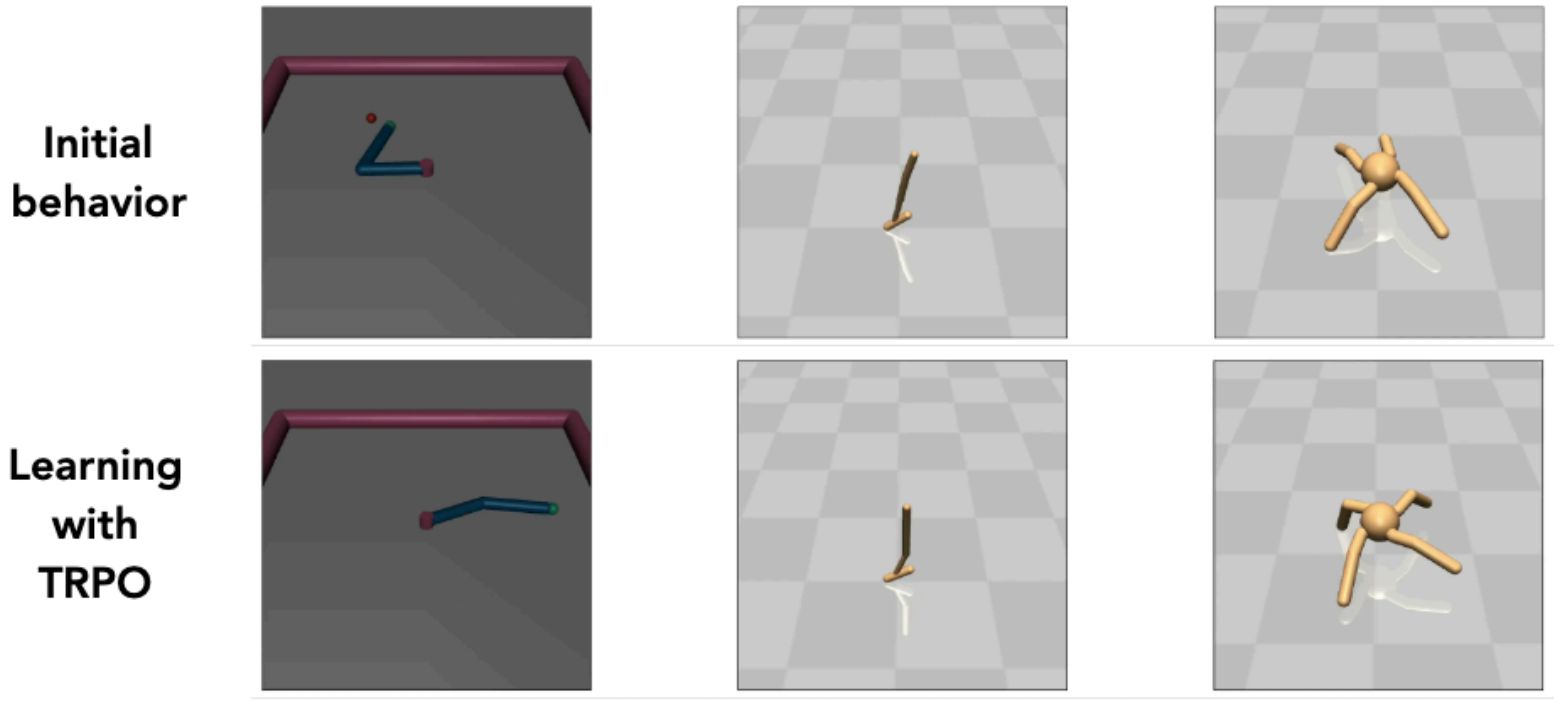
Human-level control of Atari games



Emergence of locomotion behaviors

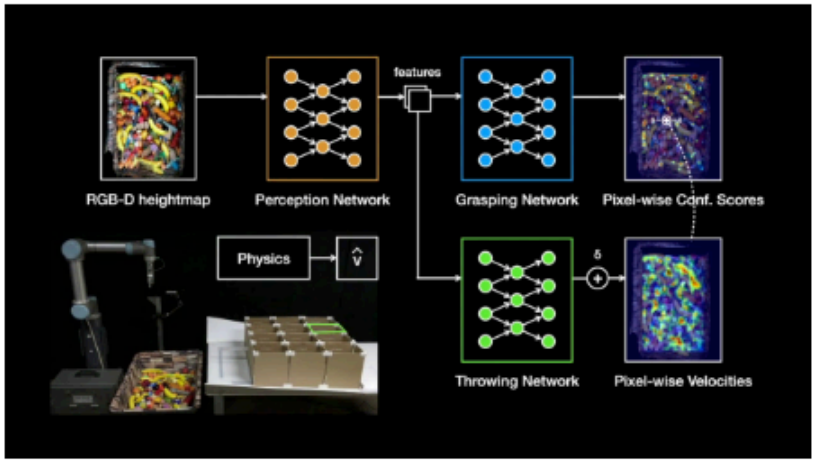
Reinforcement Learning (RL)

- 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.