TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2022

History of Deep Learning Part Two

Wav2vec 2.0, June 2020, Facebook

Trained on 53k hours of unlabeled audio (no text) they convert speech to a sequence of discrete quantized vectors they call "pseudo-text units".

By training on only one hour of human-transcribed audio, and using the Wav2vec transcription into pseudo-text, the outperform the previous state of the art in word error rate for 100 hours of human-transcribed text.

GLSM, February 2021, Facebook

Generative Spoken Language Model (GLSM)

They then train a generative model of the sequences of pseudo-text units learned from unlabeled audio.

This model can continue speech from a speech prompt in much the same way that GPT-3 continues text from a text prompt.

Semantic and grammatical structure in a "unit language model" is recovered from speech alone.

Codex, July 2021, OpenAI

This is a language model trained on code, including comments, from public repositories.

Starting from an English prompt Codex continues with code — a form of automatic programming.

There is a published version (58 authors) and a production version that powers **GitHub Copilot**.

Copilot may supplant Stack Overflow for finding out how to do x in language y.

CLIP, January 2021, OpenAI

CLIP: Contrastive Language-Image Pre-training.

Trained on images and associated text (such as image captions or hypertext links to images) CLIP computes embeddings of text and embeddings of images ("co-embeddings") trained to capture the mutual information between the two.

This is done with contrastive learning.

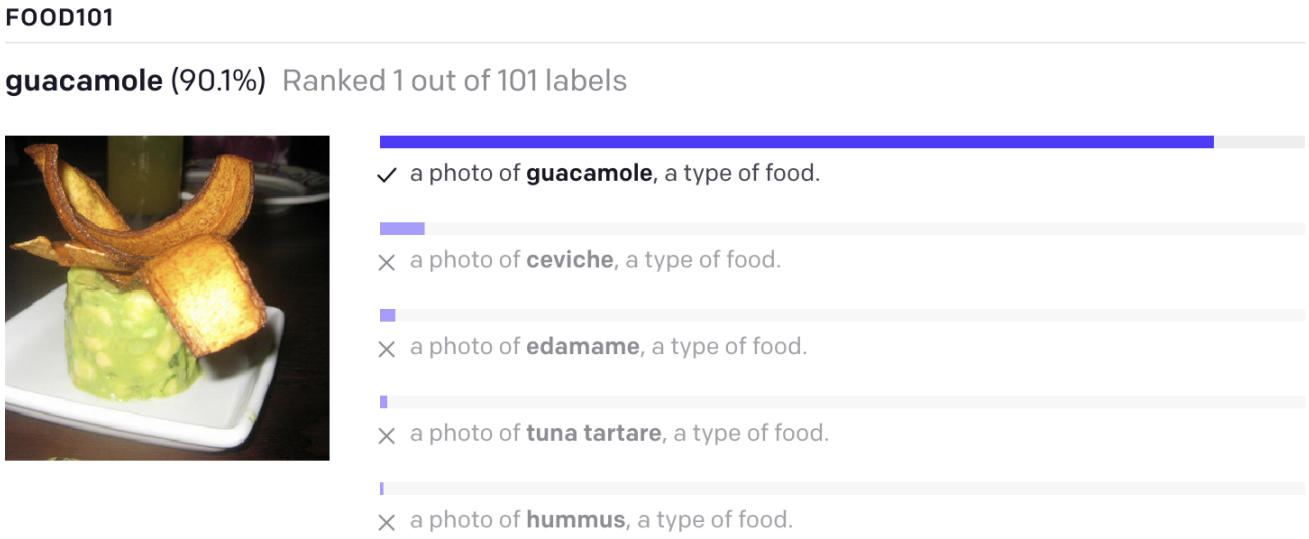
CLIP, January 2021, OpenAI

The model computes a probability of text given the co-embedding of the image.

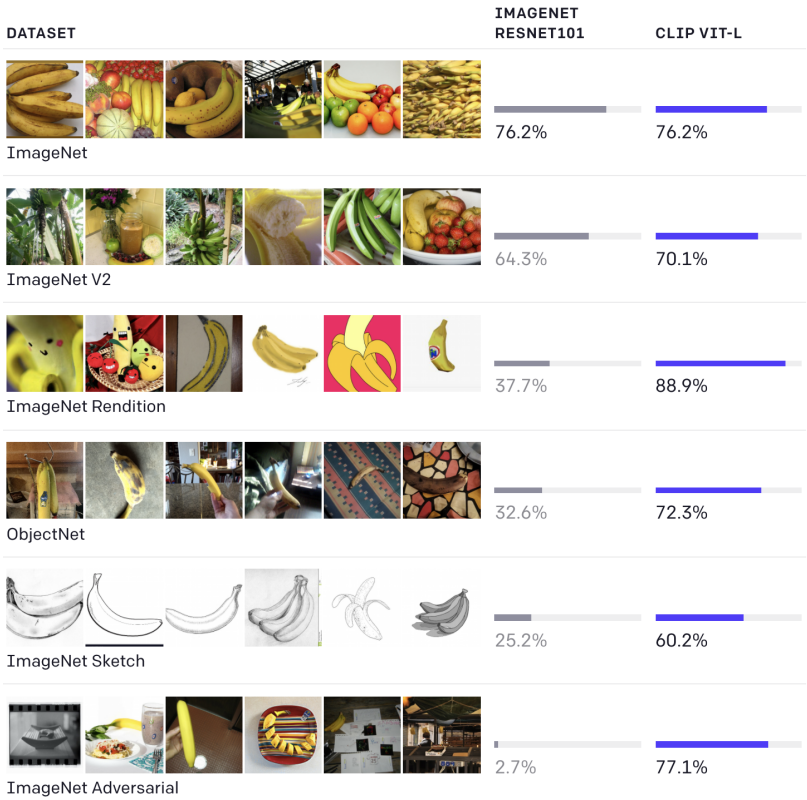
It is then used for zero-shot image classification on various datasets.

One can classify an image by comparing the probabilities that the model assigns to "prompts". There is a prompt for each class.

Zero-Shot Image Classification



Zero-Shot Image Classification



Although both models have the same accuracy on the ImageNet test set, CLIP's performance is much more representative of how it will fare on datasets that measure accuracy in different, non-ImageNet settings. For instance, ObjectNet checks a model's ability to recognize objects in many different poses and with many different backgrounds inside homes while ImageNet Rendition and ImageNet Sketch check a model's ability to recognize more abstract depictions of objects.

DALL·E, January 2021, OpenAI

DALL·E-2, April 2022

The name DALL·E is simply some kind of homage to the painter Dali and the Disney character WALL·E.

Both versions of DALL·E uses CLIP's co-embeddings of images and text.

Given text, DALL·E generates an image.

DALL·E-1 Zero-Shot Image Rendering from Language

TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

Edit prompt or view more images↓

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

DALL·E-2

vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

a dolphin in an astronaut suit on saturn, artstation

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

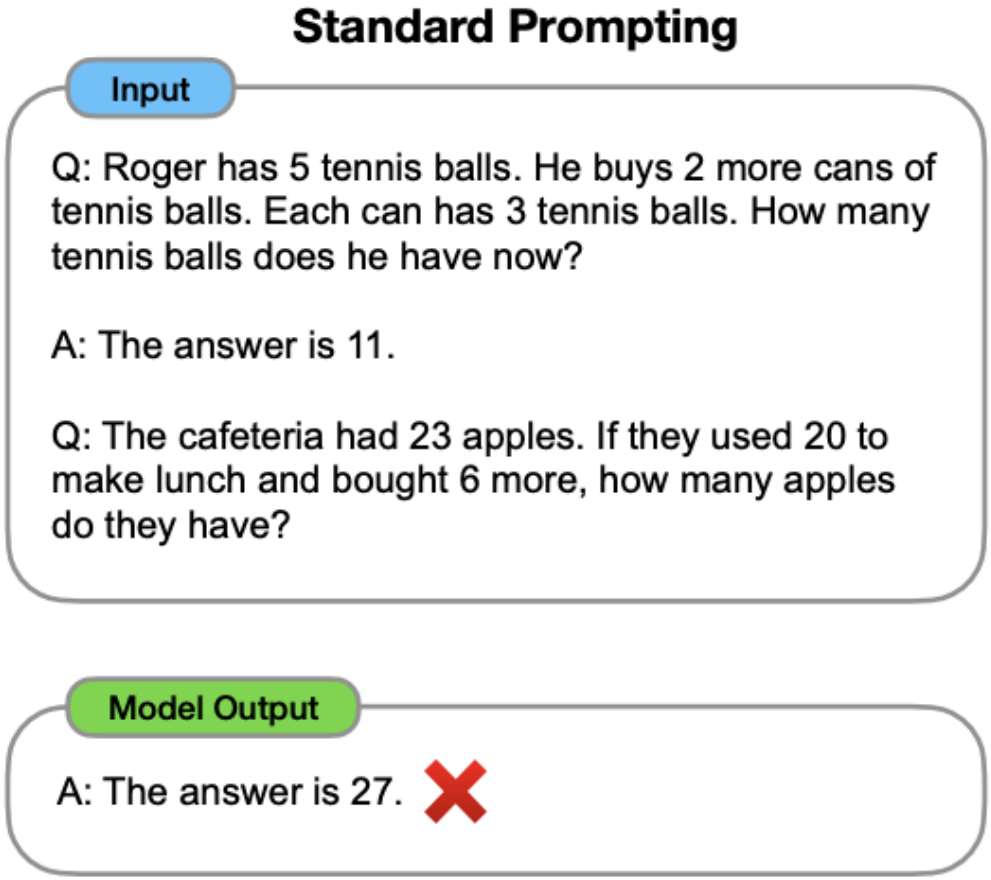
Diffusion Models

DALL·E-2 uses diffusion models. Although originally defined in 2015, they have become very prominent in the last year.

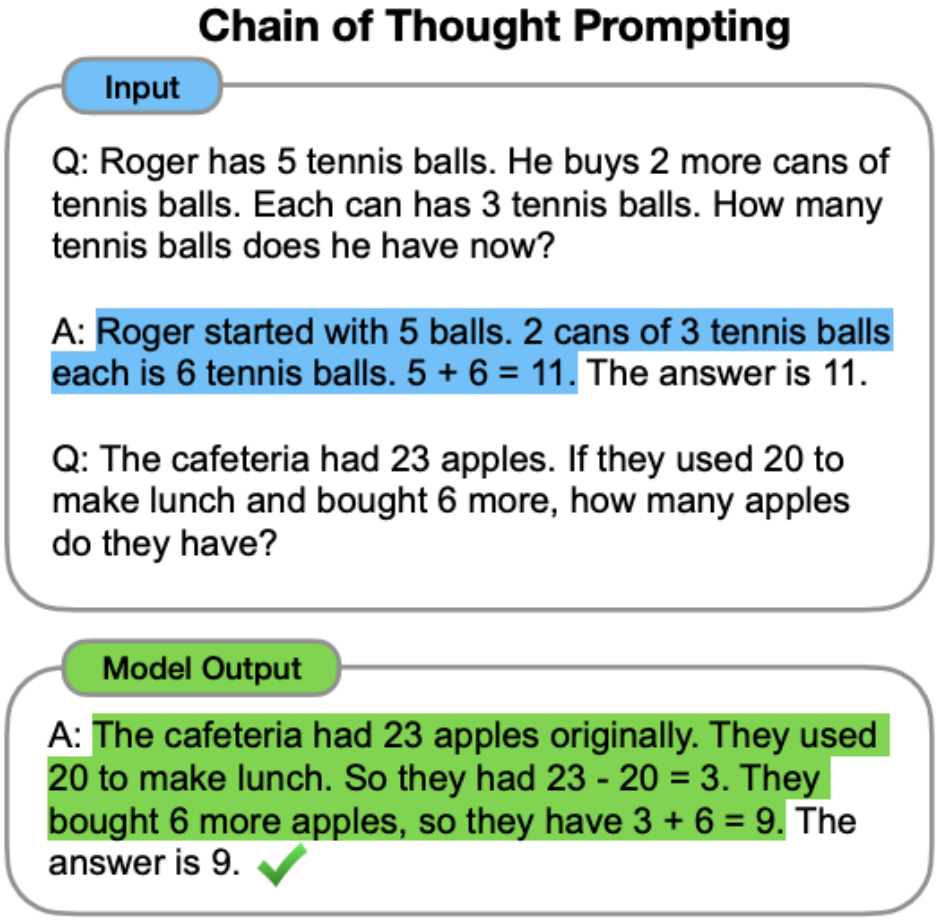
Chain of Thought Prompting, January 2022

Give examples of "chains of thought" for few shot learning of reasoning steps.

Naive Prompting



Chain of Thought Prompting, January 2022



Step by Step Prompting, June 2022

It turns out that adding the simple instruction "take it step by step" elicits powerful chain of thought reasoning in GPT-3.

Humanoid Soccer, September 2022, Deep Mind

Deep mind demonstrated two-on-two humanoid soccer in simulation (MuJoCo).

This a startling advance in the state of the art in humanoid humanoid control.

It also continues Deep Mind's effort in reinforcement learning.

Application Advancements vs. Architecture Advancements

Advancements in the general principles of learning are having applications over very diverse applications.

When considering Moore's law of AI it seems worth distinguishing architectural advancements (new general learning methods) from new applications of established architectures.

This course will focus on general, architectural, ideas.

END