TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2021

2021 Developments

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 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.

CLIP, January 2021, OpenAI

CLIP: Contrastive Language-Image Pre-training.

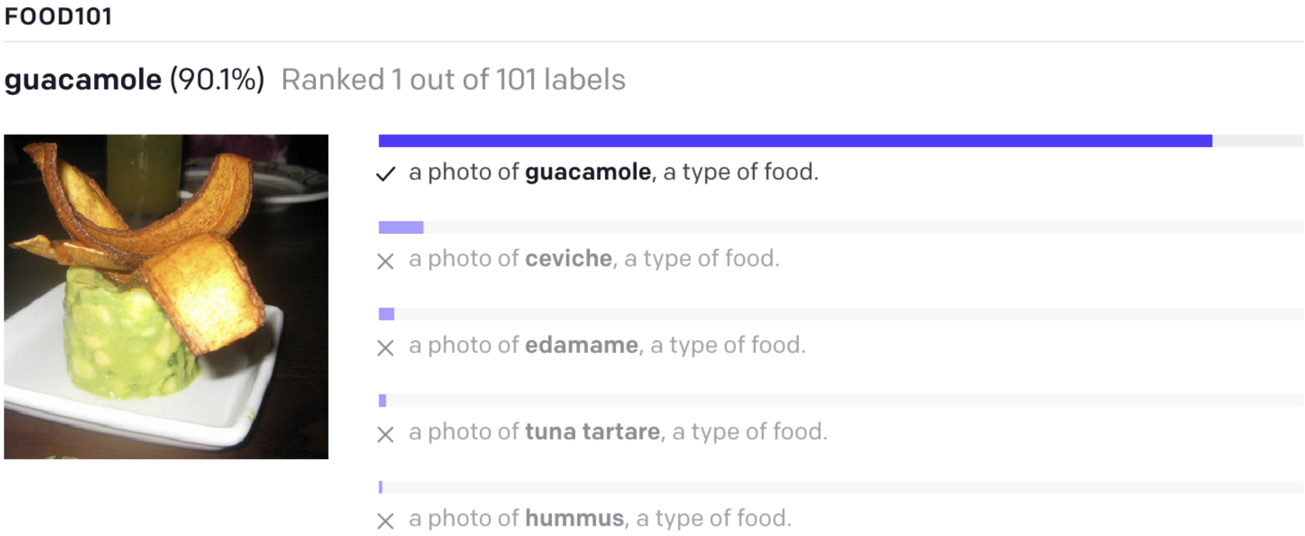
Trained on images and associated text (such as image captions or hypertext links to images).

The model computes a probability of text given 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



FOOD101

guacamole (90.1%) Ranked 1 out of 101 labels

✓ a photo of **guacamole**, a type of food.

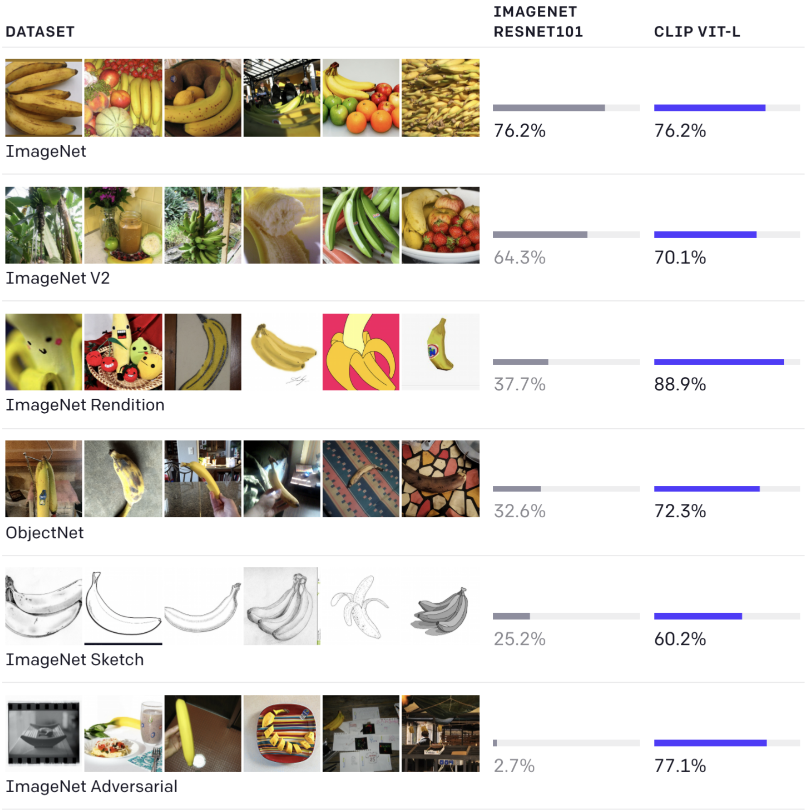
✗ a photo of **ceviche**, a type of food.

✗ a photo of **edamame**, a type of food.

✗ a photo of **tuna tartare**, a type of food.

✗ a photo of **hummus**, a type of food.

Zero-Shot Image Classification



DATASET	IMAGENET RESNET101	CLIP VIT-L
ImageNet	76.2%	76.2%
ImageNet V2	64.3%	70.1%
ImageNet Rendition	37.7%	88.9%
ObjectNet	32.6%	72.3%
ImageNet Sketch	25.2%	60.2%
ImageNet Adversarial	2.7%	77.1%

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.

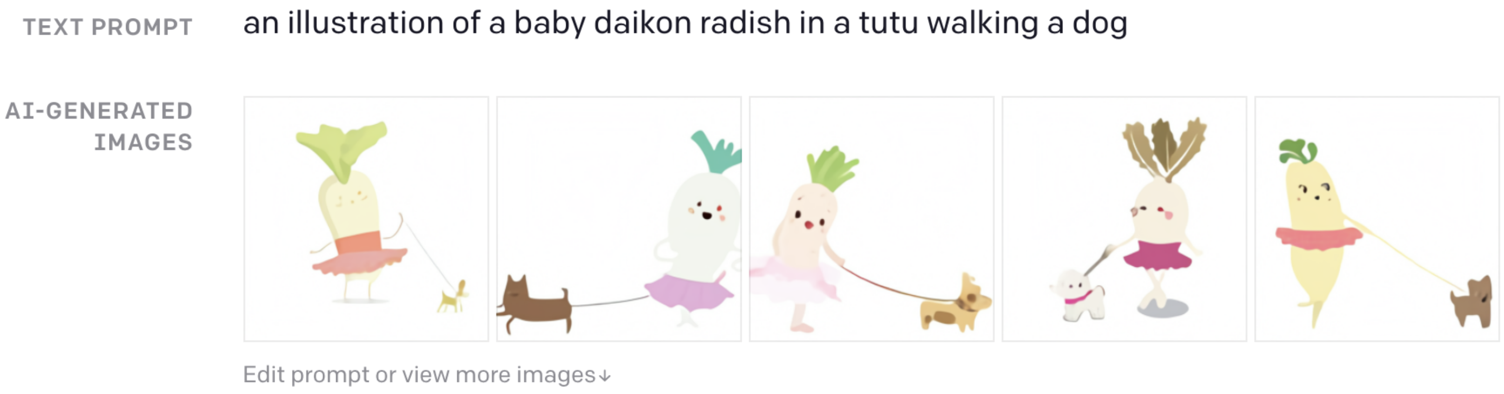
DALL·E, January 2021, OpenAI

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

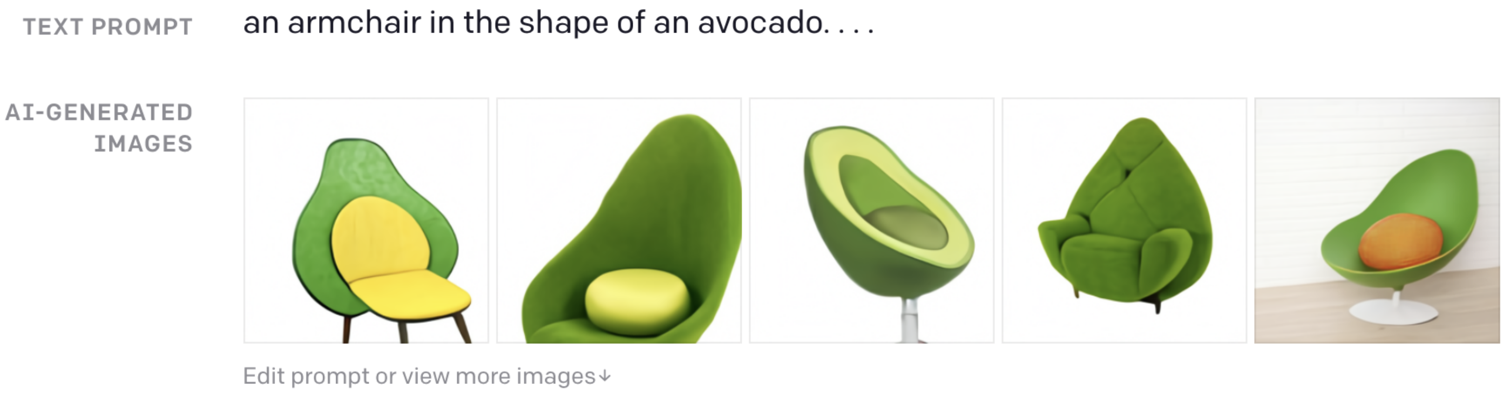
Like CLIP, DALL·E is trained on images paired with text (pressumably the same data as CLIP). DALL·E and CLIP were announced by OpenAI on the same day, although they are different systems.

Given text, DALL·E generates an image.

Zero-Shot Image Rendering from Language



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



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

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 stackoverflow for finding out how to do x in language y.

It also seems possible that some version Codex will replace, or at least augment, speech assistants such as Siri.

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 and how they have advanced in recent years.

END