

TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2021

2021 Developments

Wav2vec 2.0, June 2020, Facebook

Wav2vec pre-trains on audio-only data to get a (discrete) representation of speech sounds.

Trained on 53k hours of unlabeled audio, and one hour of transcribed audio, Wav2vec outperforms the previous state of the art for 100 hours of transcribed text.

GLSM, February 2021, Facebook

This is a speech analogue of a VQ-VAE.

Training on audio-only (no text) they convert speech to a sequence of discrete quantized vectors they call pseudo-text units.

They then train a generative model of the sequences of pseudo-text units.

Semantic and grammatical structure in a “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.

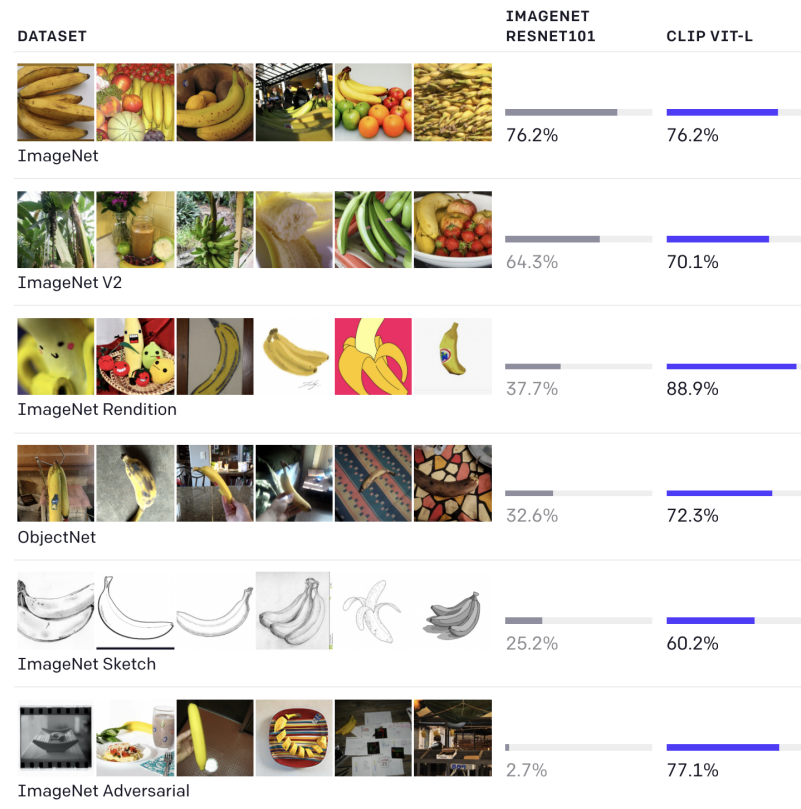
✗ a photo of **ceviche**, a type of food.

✗ 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



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.

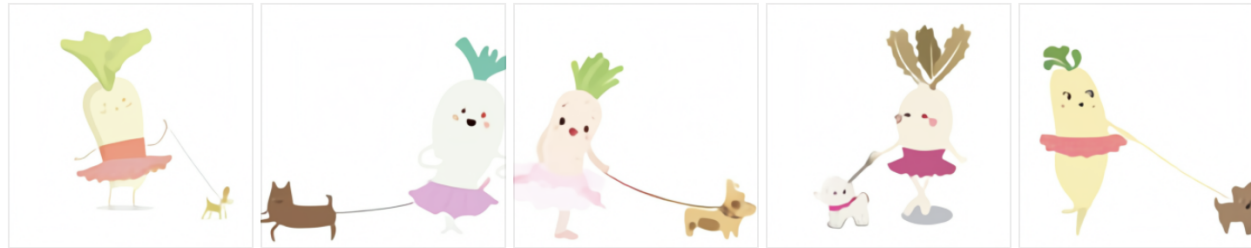
Like CLIP, DALL·E is trained on images paired with text (presumably the same data as CLIP). DALL·E and CLIP were announced by OpenAI on the same day, although they are different systems.

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

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 ↓](#)

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.

Demos indicate a major advance in programmer productivity.

This is perhaps a first example of AI being used to bootstrap AI programming.

END