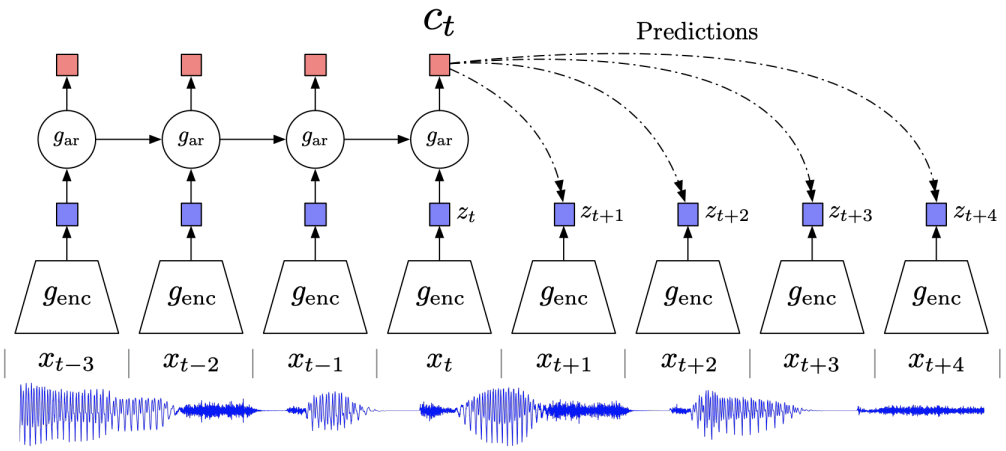
TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2022

Contrastive Coding

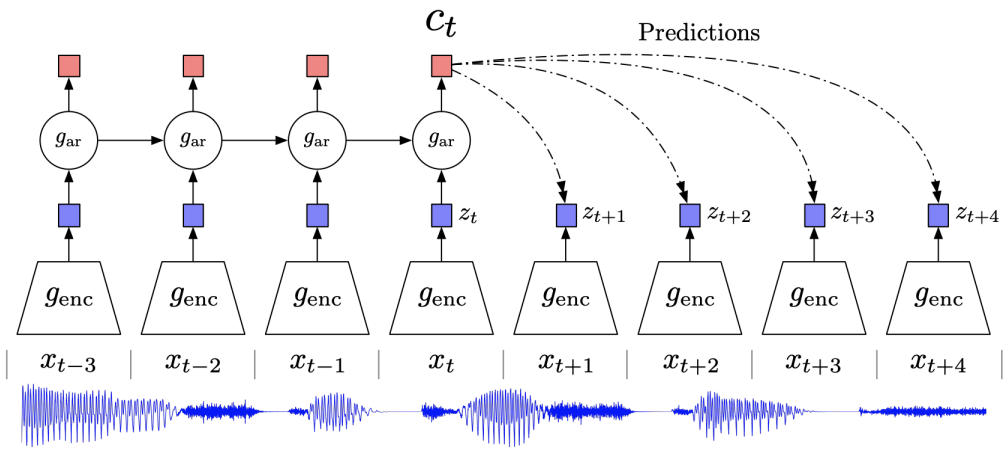
Contrastive Coding for Speech



van den ORD et al. 2018

Consider the problem of learning phonetic representations of speech sounds. In the figure each z_t is a symbol representing the sound at time t.

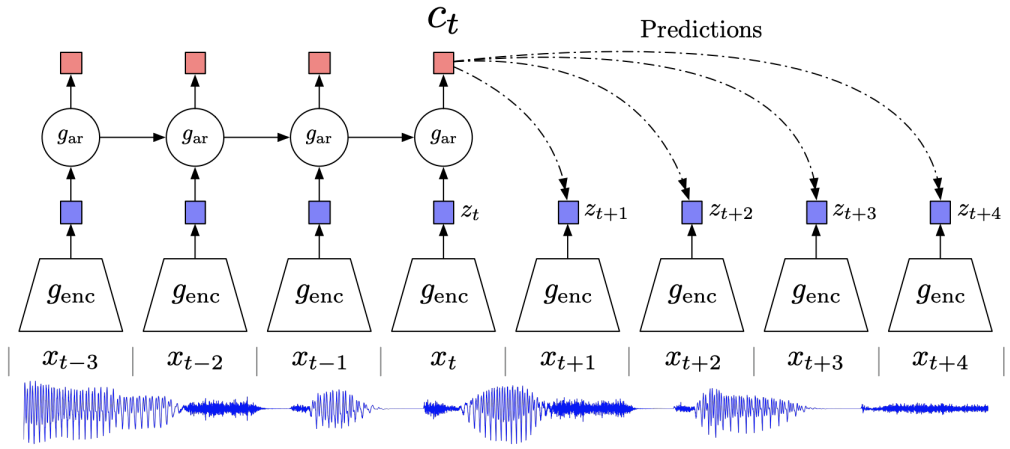
Contrastive Coding for Speech



van den ORD et al. 2018

Here we want to train the networks so as to capture the mutual information between $x_1, \ldots, x_t$ and x_{t+i}.

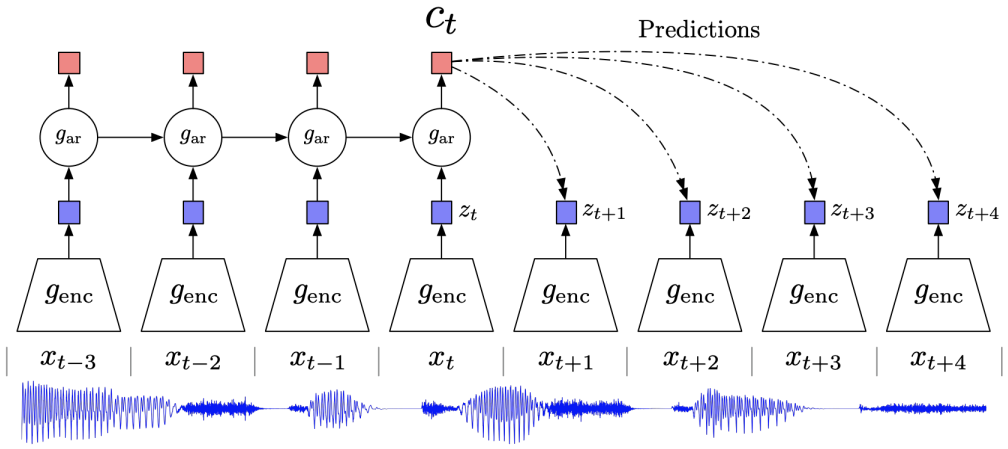
Contrastive Coding for Speech



van den ORD et al. 2018

Unlike VAEs, contrastive coding is about **capturing mutual information**. There is no attempt to model the input speech sound. Intuitively we want to **separate signal from noise** and avoid modeling noise.

Contrastive Coding for Speech



van den ORD et al. 2018

Here we want to train the networks so as to capture the mutual information between $x_1, \ldots, x_t$ and x_{t+i}.

We abstract this problem to that of capturing the mutual information between any two arbitrary random variables x and y.

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, they 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 (GSLM)

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.

Contrastive Coding

We draw a batch of pairs $(x_1, y_1), \ldots (x_B, y_B)$ from the population. For example x is the speech signal up to time t and y is the speech signal at $t + i$.

We then select b uniformly from 1 to B and construct the tuple $(x_b, y_1, \ldots, y_B, b)$.

Contrastive Coding

We draw pairs $(x_1, y_1), \ldots (x_B, y_B)$ from the population. We then select b uniformly from 1 to B and construct the tuple $(x_b, y_1, \ldots, y_B, b)$.

We then train a model to predict b.

$$\text{enc}_x^*, \text{enc}_y^* = \underset{\text{enc}_x, \text{enc}_y}{\text{argmin}} \; E_{(x, y_1, \ldots, y_B, b)} \left[-\ln P_{\text{enc}_x, \text{enc}_y}(b | x, y_1, \ldots, y_B) \right]$$

$$P_{\text{enc}_x, \text{enc}_y}(b | x, y_1, \ldots, y_B) = \underset{b}{\text{softmax}} \; \text{enc}_x(x)^\top \text{enc}_y(y_b)$$

The Contrastive Coding Theorem

For any distribution on pairs (x, y), with contrastive probabilities computed by

$$P(b|x, y_1, \ldots, y_B) = \underset{b}{\text{softmax}}\ s(x, y_b)$$

we have

$$I(x, y) \geq \ln B - \quad E_{(x,y_1,\ldots,y_B,b)}\left[-\ln P(b|(x, y_1, \ldots, y_B))\right]$$

Chen et al., On Variational Bounds of Mutual Information, May 2019.

Augmentation Contrastive Coding for Images (SimCLR)

(SimCLR:) A Simple Framework for Contrastive Learning of Visual Representations, Chen et al., Feb. 2020 (self-supervised leader as of February, 2020).

They construct a distribution on pairs $\langle x, y \rangle$ defined by drawing an image from ImageNet and then drawing x and y as random "augmentations" of the image.

Augmentations include (among others) reflections, croppings, and changes in the color map.

Augmentation Contrastive Coding for Images (SimCLR)

They drawing an image from ImageNet and then draw x and y as random augmentations of the same image.

They then train a single coding function enc that applies to any augmentation and train the ecoding function by the the contrastive coding objective objective.

$$\text{enc}^* = \underset{\text{enc}}{\text{argmin}} \; E_{(x,y_1,\ldots,y_B,b)} \left[-\ln P_{\text{enc}}(b|(x, y_1, \ldots, y_B) \right]$$

$$P_{\text{enc}}(b|x, y_1, \ldots, y_B) = \underset{b}{\text{softmax}} \; \text{enc}(x)^\top \text{enc}(y_b)$$

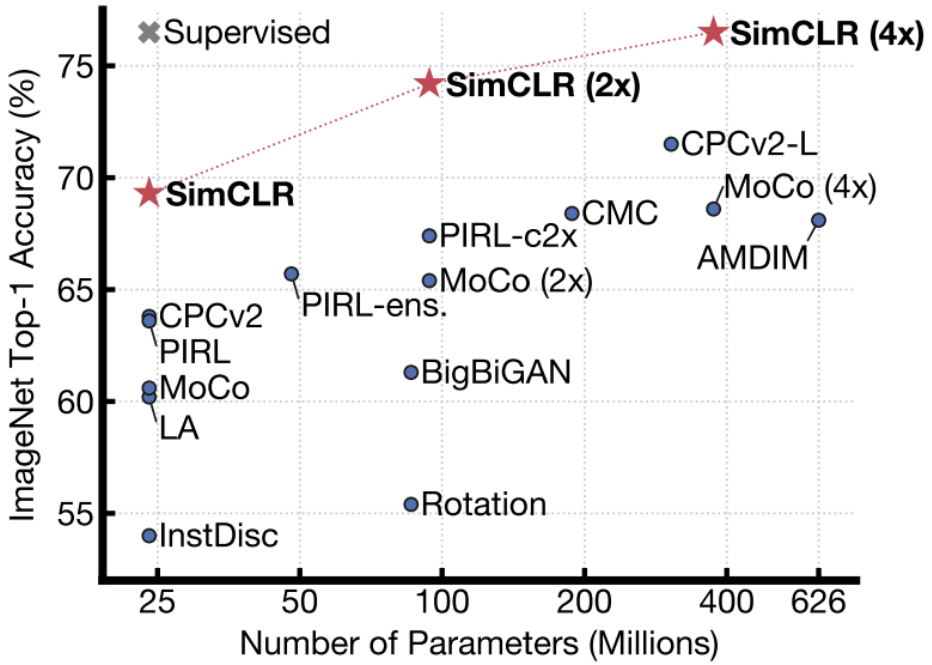
Augmentation Contrastive Coding for Images (SimCLR

The encoder is then used on images to define a feature vector for images.

They then train a **linear** imagenet classifier on the feature map defined by the encoder.

This is called linear probing.

Augmentation Contrastive Coding for Images (SimCLR)



Chen et al. 2020

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.

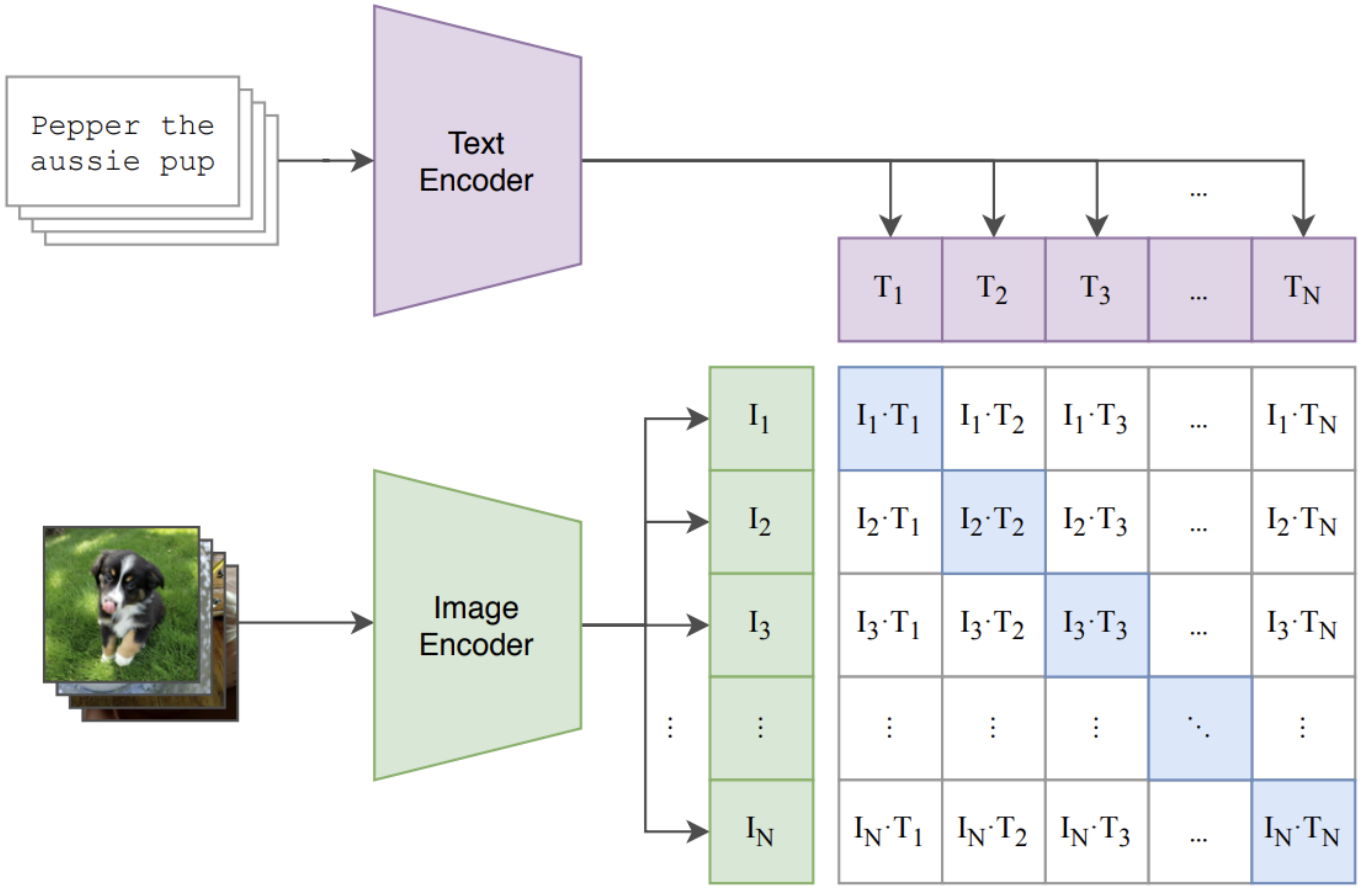
This is done with contrastive coding.

Contrastive Coding

Consider a population distribution on pairs $\langle x, y \rangle$ (such as images and associated text).

We are interested in finding embedding functions enc_x and enc_y such that $\text{enc}_x(x)$ and $\text{enc}_y(y)$ are in the same embedding space and capture the mutual information between x and y.

CLIP Constrastive Coding

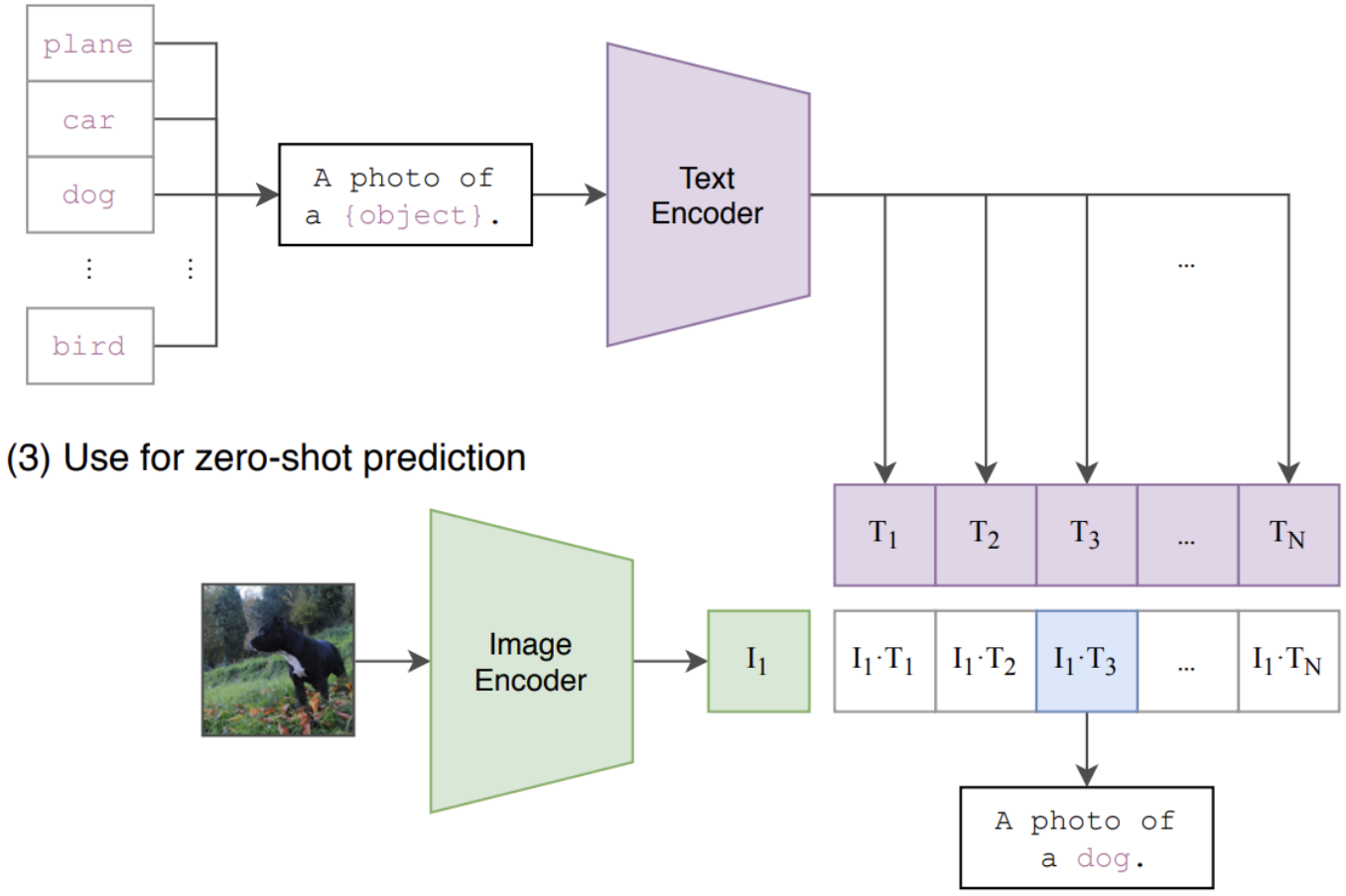


The Contrastive Coding Theorem

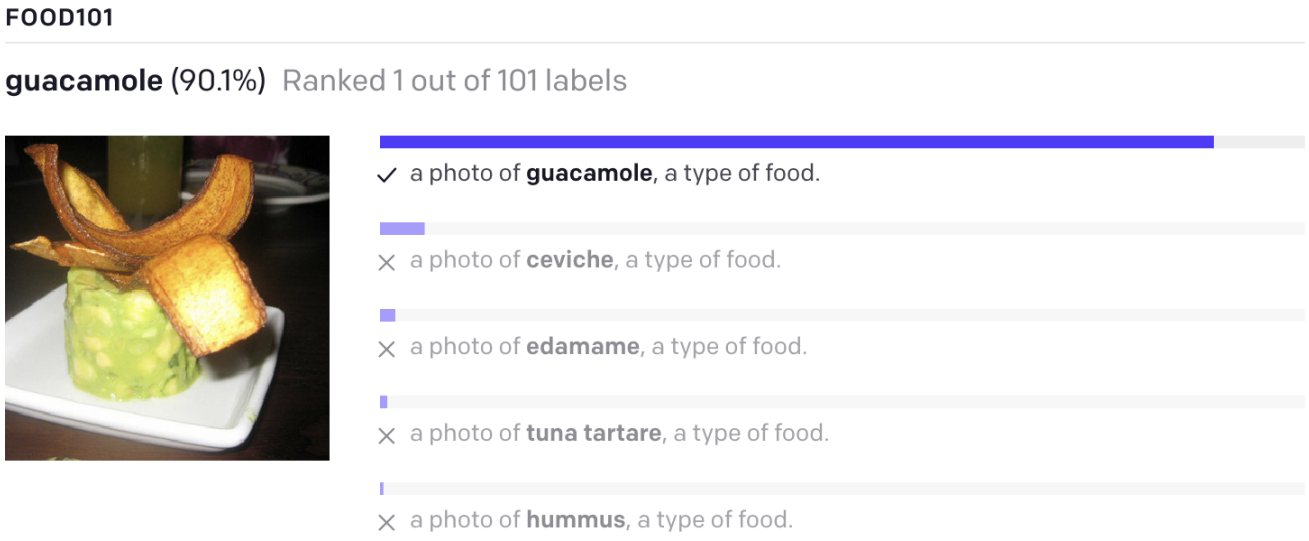
$$I(x, y) \geq \ln B - \quad E_{(x,y_1,\ldots,y_B,b)}\left[-\ln P(b|(x, y_1, \ldots, y_B)\right]$$

For CLIP the batch size $B = 2^{15}$ so we can potentially guarantee 15 bits of mutual information.

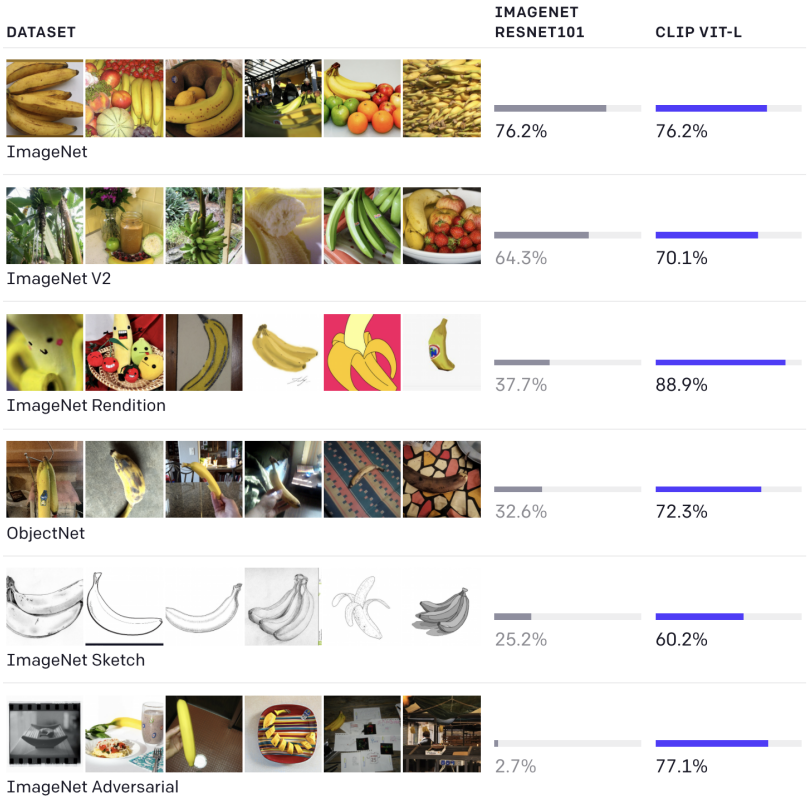
CLIP Image Classification



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.

A Weakness of Contrastive Coding

$$I(x, y) \geq \ln B - \quad E_{(x, y_1, \ldots, y_B, b)} [-\ln P(b|(x, y_1, \ldots, y_B)]$$

The discrimination problem may be too easy.

The guarantee can never be stronger than $\ln B$ where B is the batch size.

Suppose we have 100 bits of mutual information as seem plausible for translation pairs.

A possibly Better Estimate of Mutual Information

We might be able to get a better estimate of the mutual information using

$$I(x, y) \geq I(\text{enc}(x), y) = H(\text{enc}(x)) - H(\text{enc}(x)|y)$$

and estimating $H(\text{enc}(x))$ and $H(\text{enc}(x)|y)$ separately.

For this to be meaningful it seems best to use a discrete code $\text{enc}(x)$ such as might be achieved with K-means clustering.

We upper bound $H(\text{enc}(x)|y)$ by a cross entropy model and estimate (but not lower bound) $H(\text{enc}(z))$ by a cross entropy model.

END