TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2022

Contrastive Coding

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.

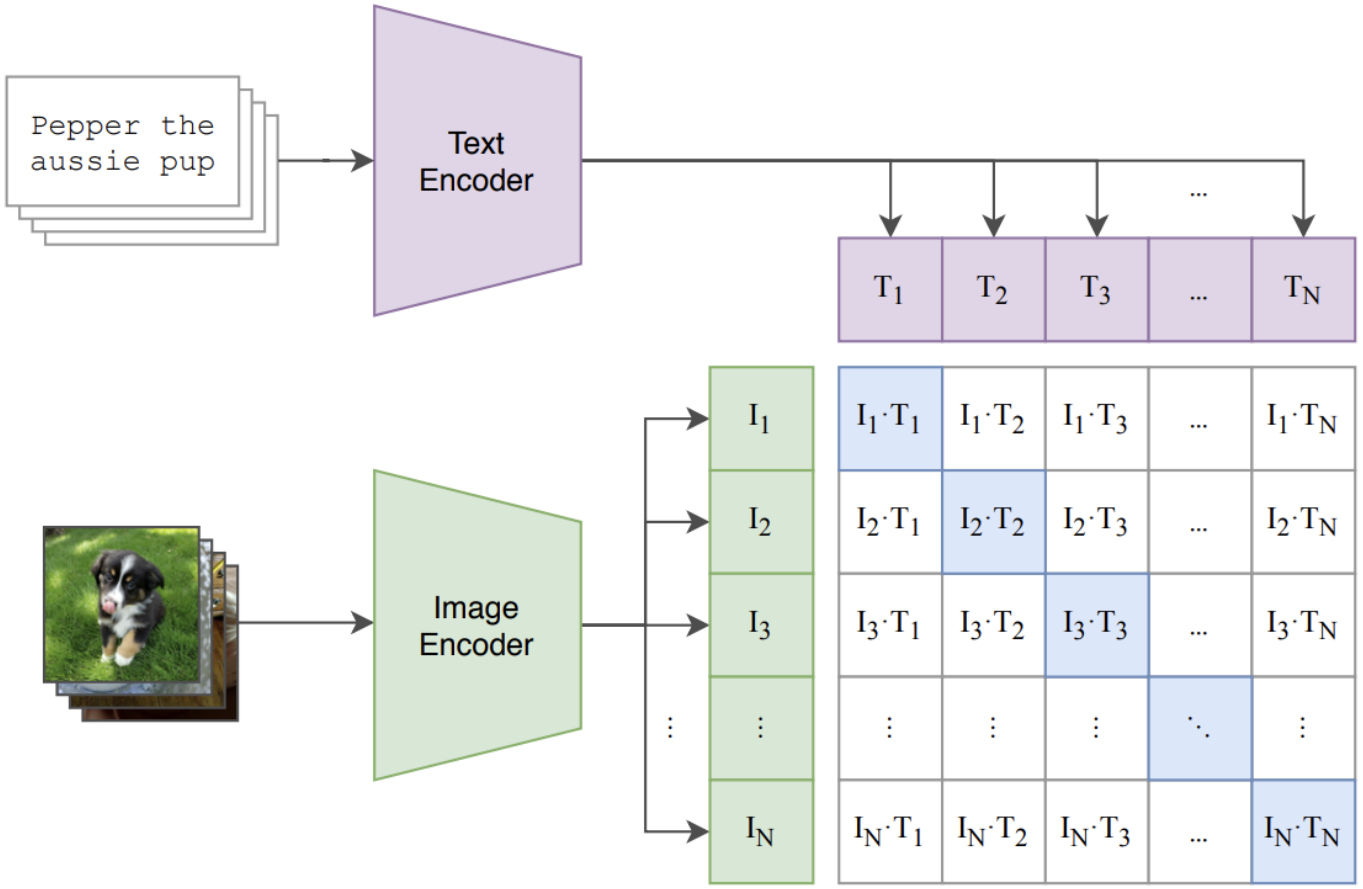
This is done with contrastive coding.

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



Contrastive Coding

We draw pairs $(x_1, y_1), \ldots (x_n, y_n)$ from the population, and i uniformly from 1 to N.

We then construct $(z_x, z_y^1, \ldots, z_y^N, i)$ by and setting $z_x = \text{enc}_x(x_i)$ and $z_y^i = \text{enc}_y(y_i)$.

We then train a model to predict i.

$$\text{enc}_x^*, \text{enc}_y^* = \underset{\text{enc}_x, \text{enc}_y}{\text{argmin}} \; E_{(z_x, z_y^1, \ldots, z_y^N, i)} \left[-\ln P(i | (z_x, z_y^1, \ldots, z_y^N) \right]$$

$$P(i | z_x, z_y^1, \ldots z_y^N) = \underset{i}{\text{softmax}} \; z_x^\top z_y^i$$

The Contrastive Coding Theorem

For any distribution on pairs (z_x, z_y), if contrastive probabilities are computed by

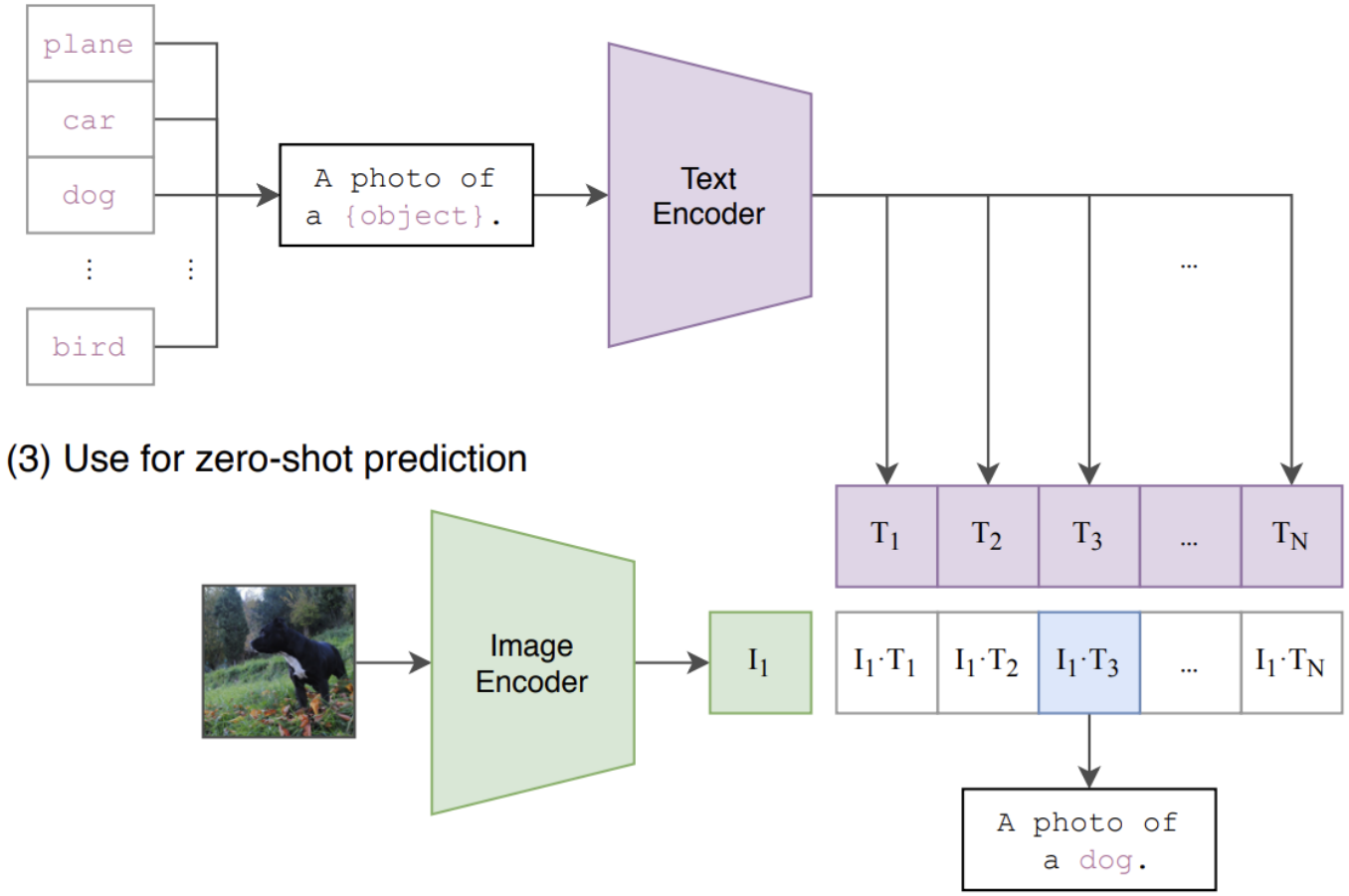
$$P(i|z_x, z_y^1, \ldots z_y^N) \;=\; \underset{i}{\text{softmax}}\; s(z_x, z_y^i)$$

then

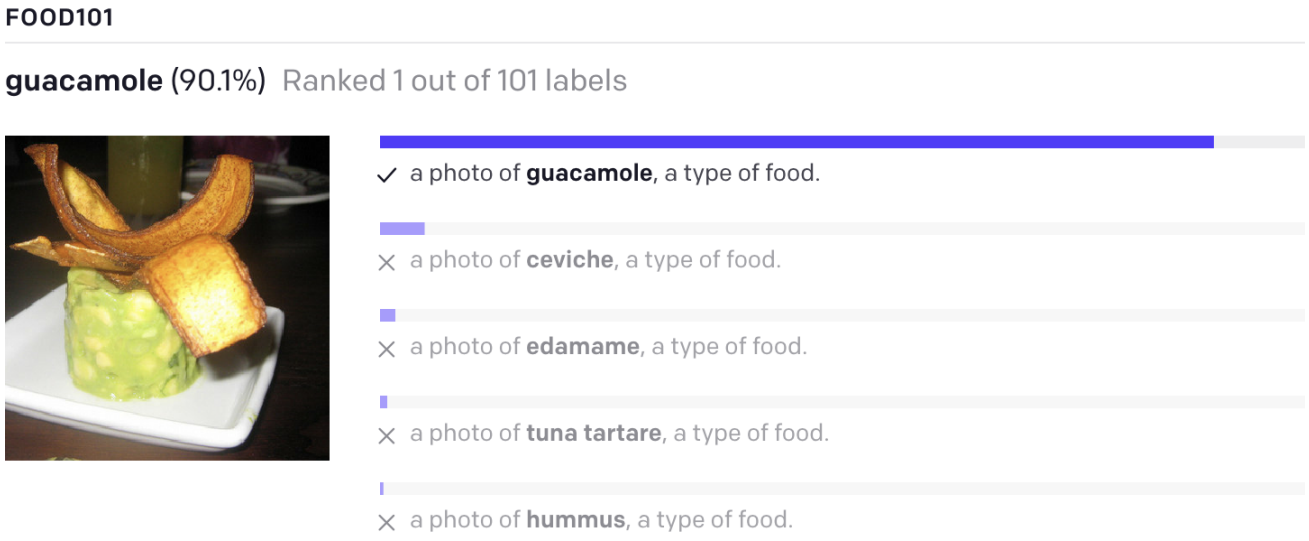
$$I(z_x, z_y) \;\geq\; \ln N - \;\; E_{(z_x, z_y^1, \ldots, z_y^N, i)} \left[-\ln P(i|(z_x, z_y^1, \ldots, z_y^N) \right]$$

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

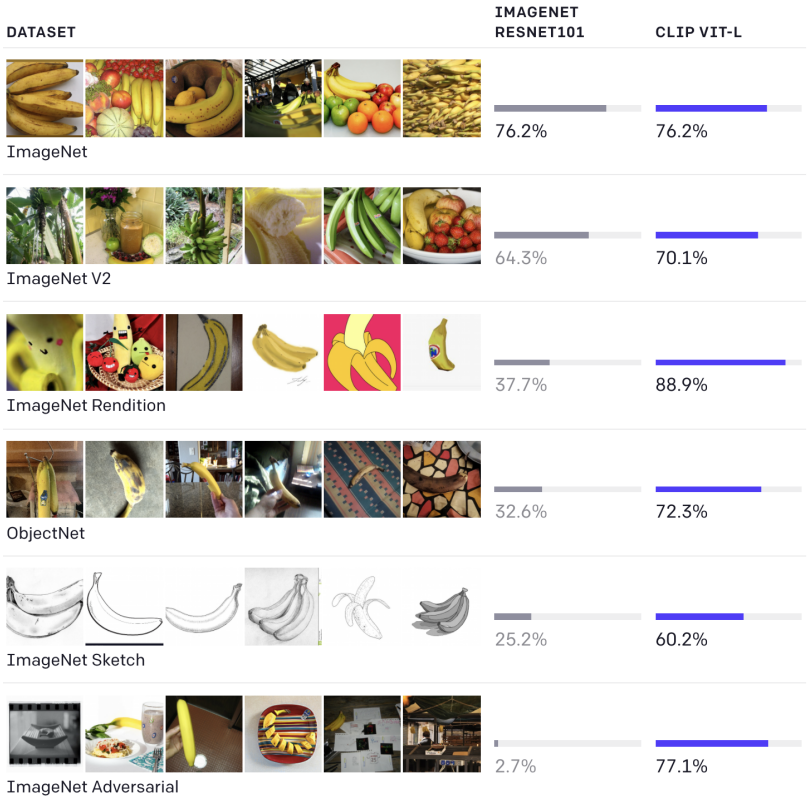
CLIP Image Classification



Zero-Shot Image Classification

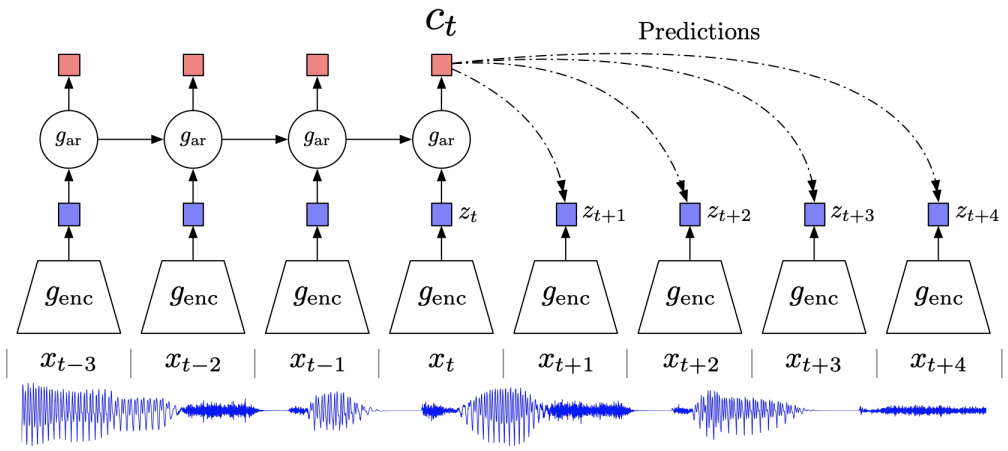


Zero-Shot Image Classification



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.

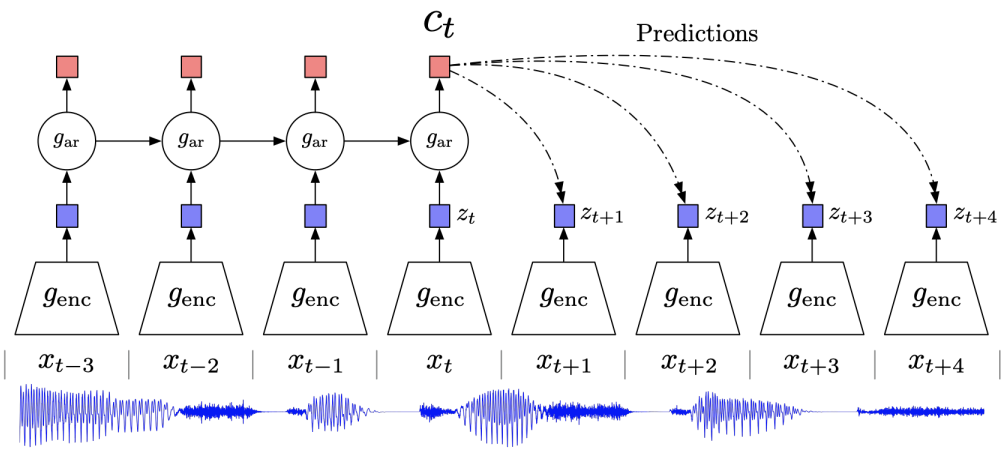
Contrastive Coding for Speech



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

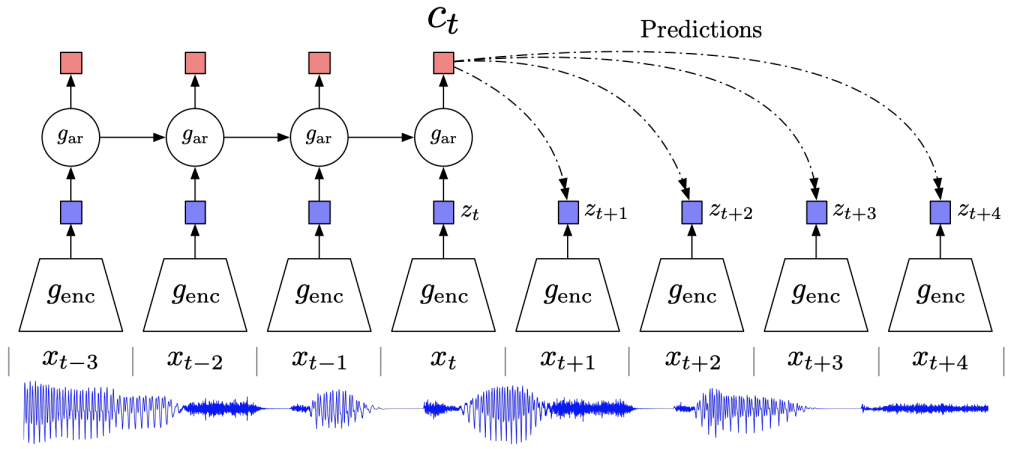


van den ORD et al. 2018

Here we want to train the networks so as to capture the mutual information between c_t and z_{t+i}.

This is done with the same contrastive loss function as in CLIP (and predates CLIP).

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.

wav2vec 2.0

Trained on 53k hours of **unlabeled** audio they convert speech to a sequence of symbols they call "pseudo-text units".

Using this pre-trained transcription of speech into pseudo-text they reduce the amount of labeled data needed for a given accuracy in speech recognition by a factor of 100.

Baevski et al., 2020

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.

$$\text{enc}^* = \underset{\text{enc}}{\text{argmin}} \; E_{(z_x, z_y^1, \ldots, z_y^N, i)} \left[-\ln P(i|(z_x, z_y^1, \ldots, z_y^N) \right]$$

$$P(i|z_x, z_y^1, \ldots z_y^N) = \underset{i}{\text{softmax}} \; z_x^\top z_y^i$$

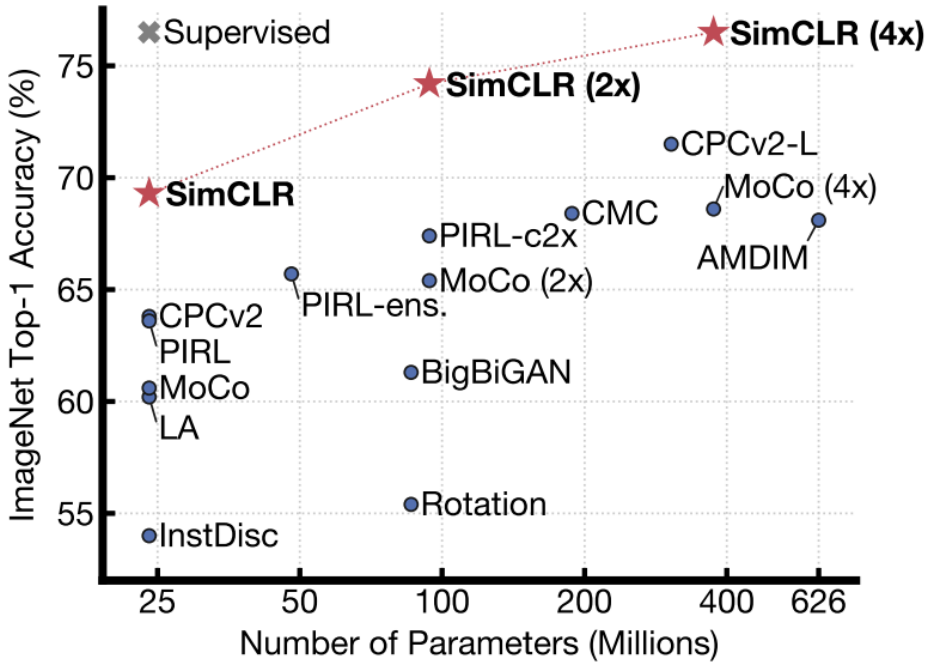
Augmentation Contrastive Coding for Images (SimCLR

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

A Weakness of Contrastive Coding

$$I(z_x, z_y) \geq \ln N - \quad E_{(z_x, z_y^1, \ldots, z_y^N, i)} \left[-\ln P(i | (z_x, z_y^1, \ldots, z_y^N) \right]$$

The discrimination problem is too easy.

The guarantee can never be stronger than $\ln N$ where N is the number of choices in the discrimination task.

Contrastive Cluster Assignments

For a population on pairs $\langle x, y \rangle$ we might consider the following training objective.

$$\text{enc}_x^*, \text{enc}_y^* = \underset{\text{enc}_x, \text{enc}_y}{\text{argmax}} \; I(z_x, z_y) \quad z_x = \text{enc}_x(x), \; z_y = \text{enc}_y(y)$$

Unfortunately this has a degenerate solution of $\text{enc}_x(x) = x$ and $\text{enc}_y(y) = y$.

Contrastive Cluster Assignments

To block the degenerate solution we first rewrite the mutual information objective function.

$$
\begin{aligned}
\text{enc}_x^*, \text{enc}_y^* &= \underset{\text{enc}_x, \text{enc}_y}{\text{argmax}} \; I(z_x, z_y) \\
&= \underset{\text{enc}_x, \text{enc}_y}{\text{argmax}} \; H(z_y) - H(z_y|z_x) \\
&= \underset{\text{enc}_x, \text{enc}_y}{\text{argmin}} \; H(z_y|z_x) - H(z_y)
\end{aligned}
$$

Contrastive Cluster Assignments

$$\text{enc}_x^*, \text{enc}_y^* = \underset{\text{enc}_x, \text{enc}_y}{\text{argmin}} \; H(z_y|z_x) - H(z_y)$$

We can block the solution of $\text{enc}_y(y) = y$ by requiring that $\text{enc}_y(y)$ is a symbol from a finite vocabulary of size K. We then have

$$H(z_y) \leq \ln K$$

We can then allow $\text{enc}_x(x) = x$.

Direct MI Coding

$$\text{enc}_x^*, \text{enc}_y^* = \underset{\text{enc}_x, \text{enc}_y}{\text{argmin}} \; H(z_y|z_x) - H(z_y)$$

We can train a model pred to predict z_y from z_x and train on cross-entropy loss.

We can estimate $H(z_y)$ from an empirical histogram over the symbols and int

Direct MI Coding

$$\text{enc}_x^*, \text{enc}_y^*, \Psi^* = \underset{\text{enc}_x, \text{enc}_y, \Psi}{\text{argmin}} \; E_{(x,y)\sim\text{Pop}} \left[-\ln P_\Psi(z_y|z_x) + \ln \hat{P}(z_y) \right]$$

where $\hat{P}(z_y)$ is an estimate (perhaps an exponential moving average) of the probability of z_y over the draw of $(x, y) \sim$ Pop.

Direct MI Coding Theorem

$$\mathrm{enc}_x^*, \mathrm{enc}_y^*, \Psi^* = \underset{\mathrm{enc}_x, \mathrm{enc}_y, \Psi}{\operatorname{argmin}} \; E_{(x,y)\sim \mathrm{Pop}} \left[-\ln P_\Psi(z_y | z_x) + \ln P(z_y) \right]$$

$$I(z_x, z_y) \geq H(z_y) - \hat{H}(z_y | z_x)$$

$$\hat{H}(z_y | z_x) = E_{(x,y)\sim \mathrm{Pop}} \left[-\ln P_\Psi(z_y(y) | z_x(x) \right]$$

Direct MI Coding Theorem

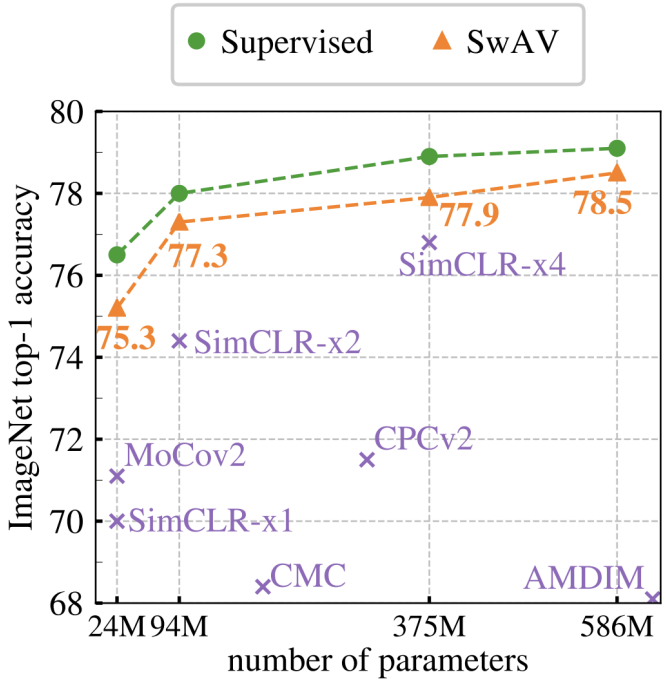
For $H(z_y) = \ln K$ where K is the number of symbols, and $\hat{H}(z_y|z_x) = 0$ we get

$$I(z_x, z_y) \geq \ln K$$

Which typically improves significantly on the best possible bound $I(z_x, z_y) \geq \ln N$ from CPC.

SwAV

SwAV uses direct MI coding rather than SimCLR's CPC.



Caron et al. 2021

END