TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2023

Contrastive Coding

Review: Cross Entropy Loss

$$\operatorname*{argmin}_{\Phi} E_{(x,y)\sim \mathrm{Pop}} \left[-\ln P_{\Phi}(y|x) \right]$$

Under universality:

$$P_{\Phi^*}(y|x) = P_{\mathrm{Pop}}(y|x)$$

Review: GANs

Generative Adversarial Networks Ian J. Goodfellow et al, 2014

$$\underset{\text{gen}}{\text{argmax}} \min_{\text{disc}} E_{i\sim\{-1,1\},y\sim P_i} [-\ln P_{\text{disc}}(i|y)]$$

Under Universality:

$$P_{\text{gen}}(y) = P_{\text{Pop}}(y)$$

Review: VAEs

Auto-Encoding Variational Bayes, Kingma and Welling, 2013

$$\underset{\text{pri,dec,enc}}{\text{argmin}} \; E_{y\sim \text{Pop}, z\sim P_{\text{enc}}(z|y)} \left[-\ln \frac{P_{\text{pri}}(z) P_{\text{dec}}(y|z)}{P_{\text{enc}}(z|y)} \right]$$

Under Universality:

For for any encoder enc

$$P_{\text{pri}^*,\text{dec}^*}(z, y) = P_{\text{Pop,enc}}(z, y)$$

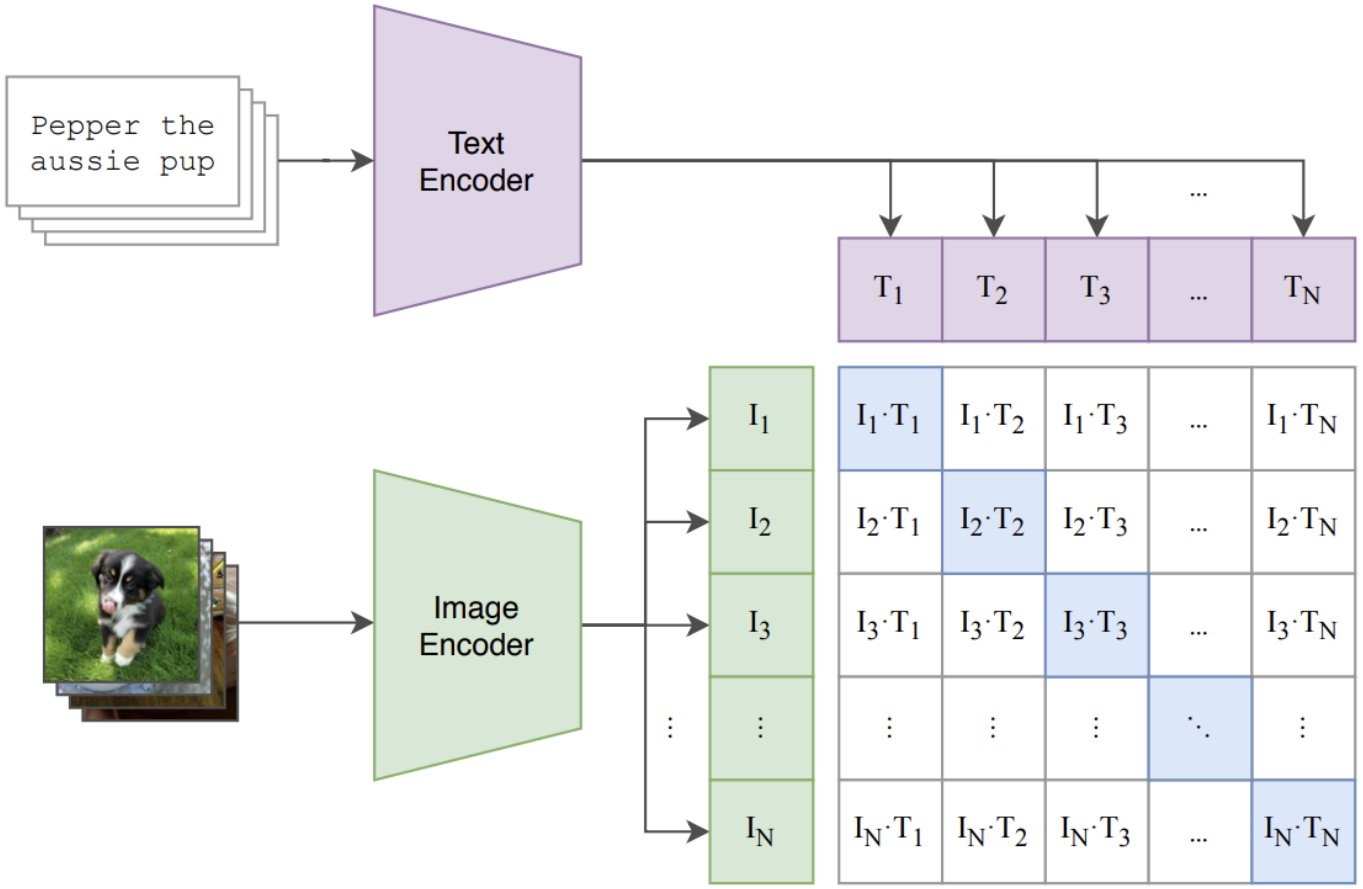
Under universality the Encoder can be anything (but in practice the encoder matters).

Contrastive Coding: A Fourth Training Objective

Representation Learning with Contrastive Predictive Coding, Aaron van den Oord, Yazhe Li, Oriol Vinyals, July 2018.

CLIP: Contrastive Language-Image Pre-training, January 2021, OpenAI

CLIP Contrastive Coding



Contrastive Coding: the Fourth Training Objective

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 $(x_b, y_1, \ldots, y_B, b)$.

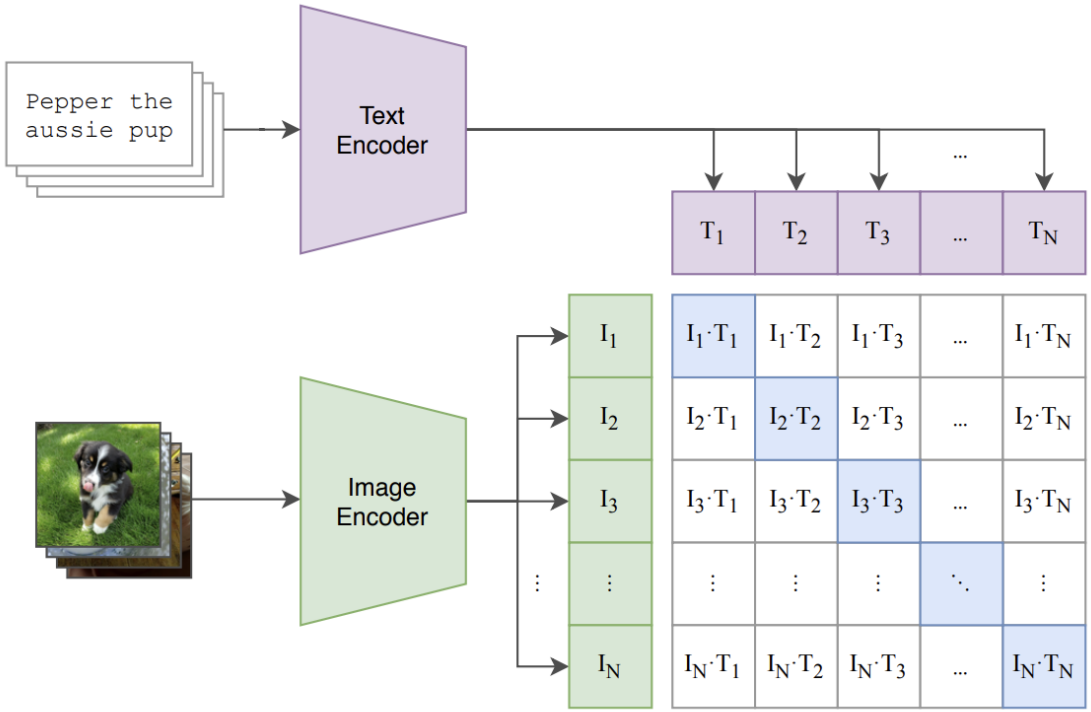
We then train a model to predict b.

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

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

CLIP Contrastive Coding

In CLIP we make B^2 predictions for a batch of size $N = B$



The Mutual Information Contrastive Coding Theorem

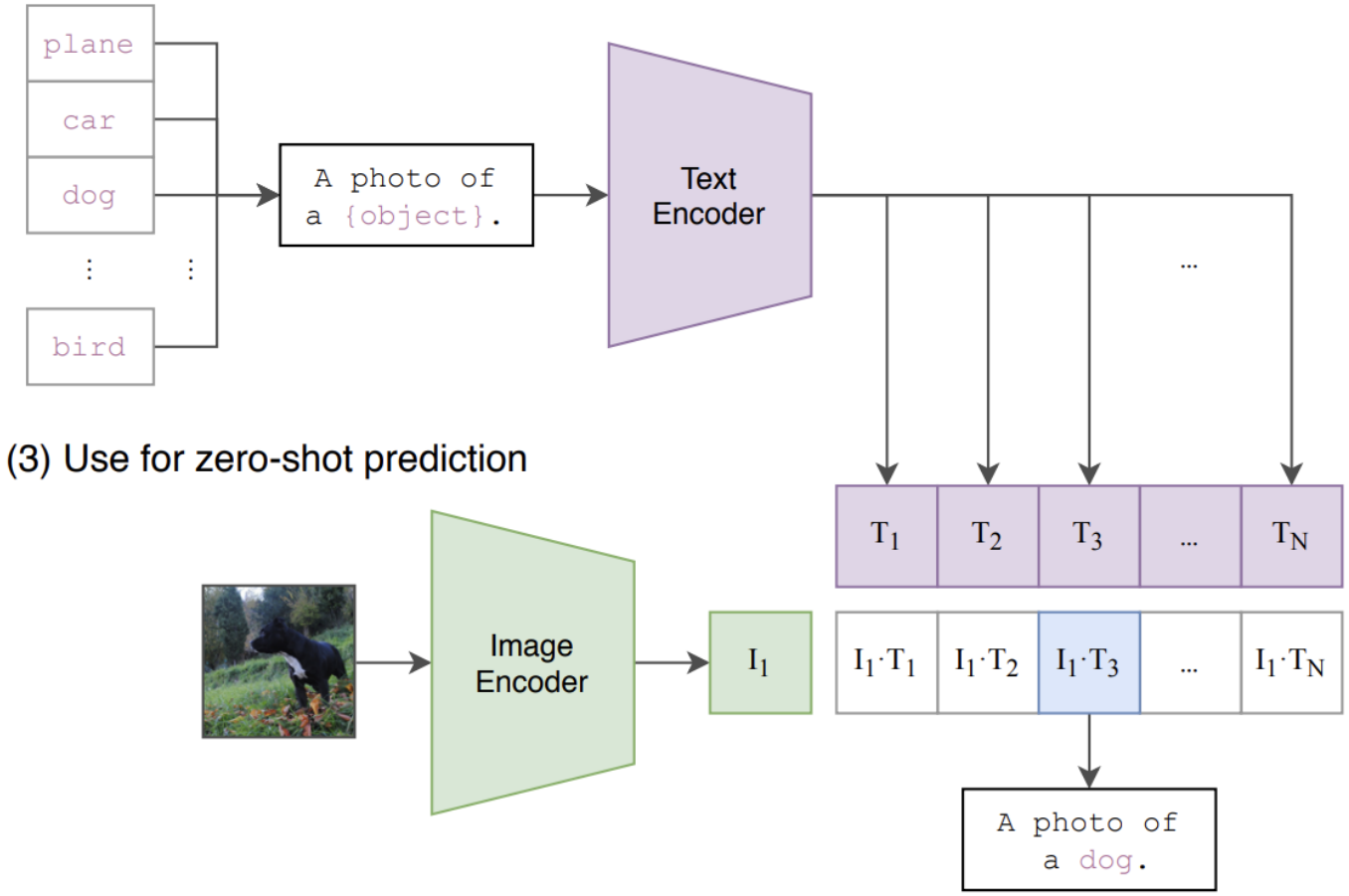
For any distribution on pairs (x, y), with contrastive probabilities computed by

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

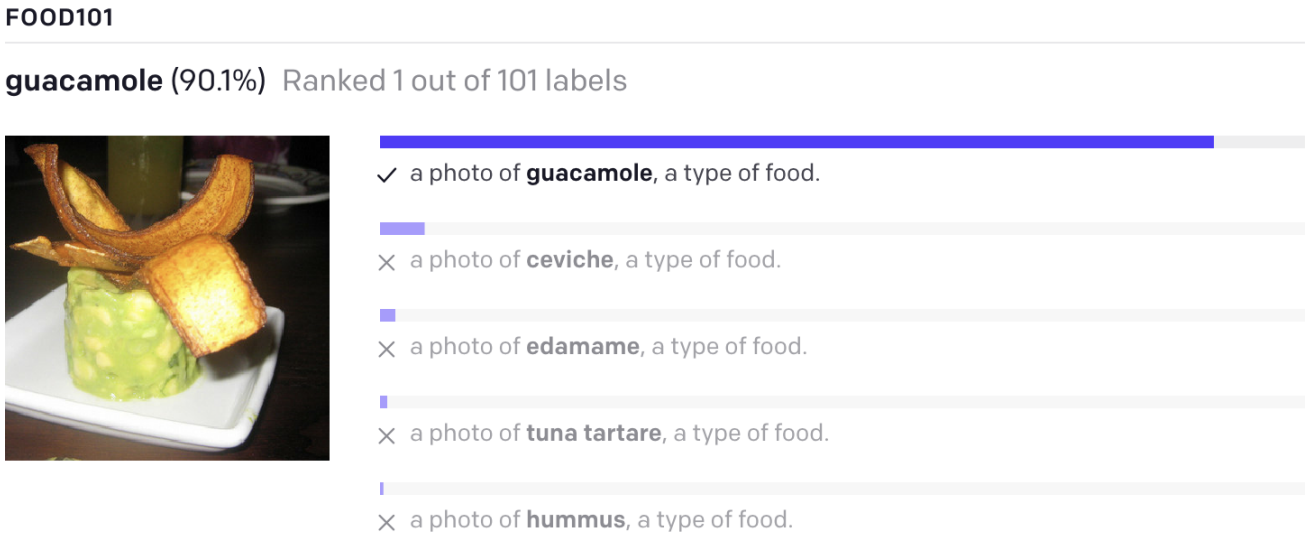
$$I(x, y) \geq \ln B - \quad E_{(x_b, y_1, \ldots, y_B, b)} [-\ln P(b|(x_b, y_1, \ldots, y_B))]$$

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

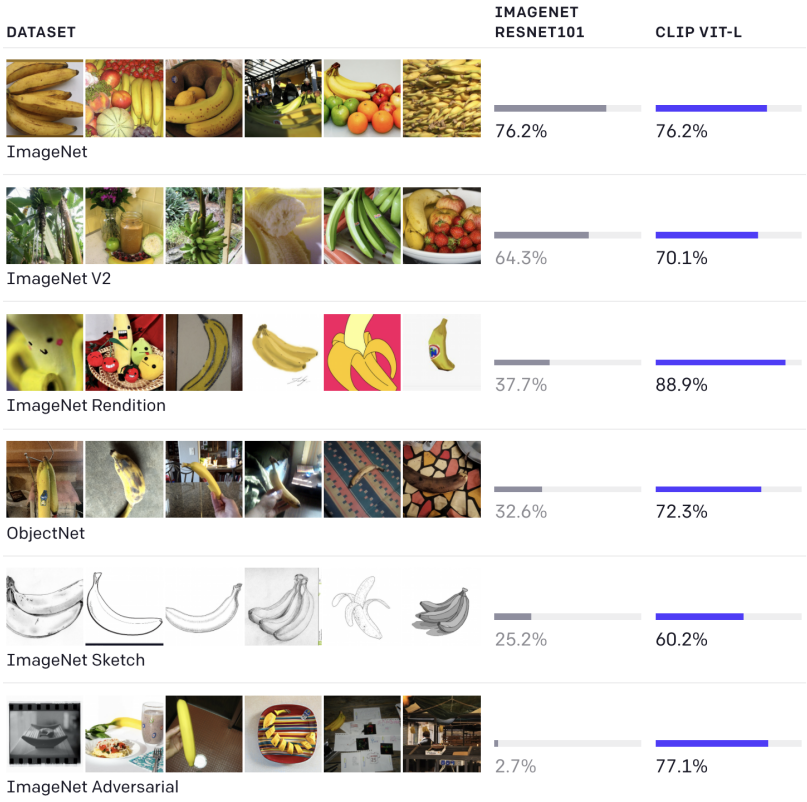
CLIP Image Classification



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.

An abstract Formulation

We consider a population distribution on pairs (x, y).

For example:

- x might be an image and y might be the text of a caption for image x (CLIP).
- x might be an video frame and y video frame a second later.
- x might be a window of a sound wave and y a later window (Wav2Vec).
- $x = f(z)$ and $y = f(z)$ where f and g are transformation functions on an image z such as translation, rotation, color shift, or cropping. (augmentation) of x. (SimCLR)

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.

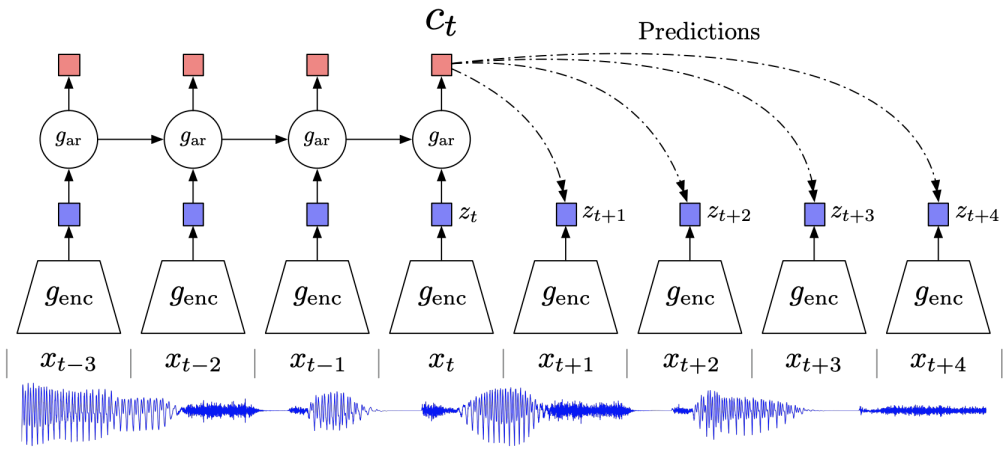
Suppose we have 100 bits of mutual information as seem plausible for translation pairs.

Addresses the Weakness with Large Batch Size

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

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

Contrastive Coding for Speech

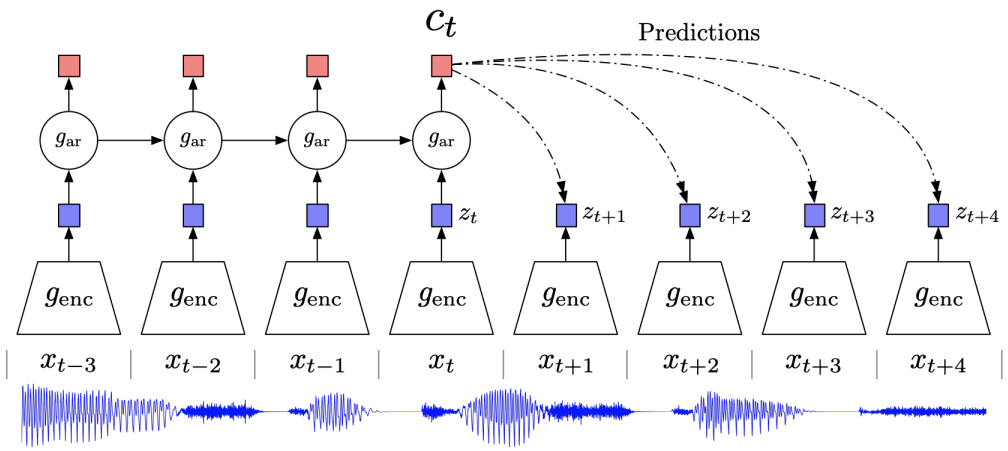


van den Oord, Li and Vinyals,

Representation Learning with Contrastive Predictive Coding, 2018

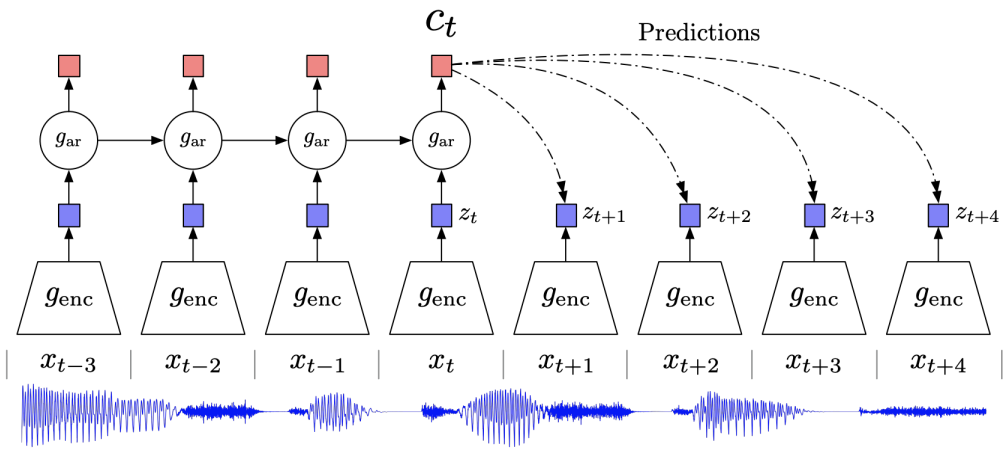
What should we abstract from the past that is relevant to the future?

Contrastive Coding for Speech



Unlike VAEs, contrastive coding is about **capturing mutual information**. Intuitively we want to **separate signal from noise** and avoid modeling noise.

Contrastive Coding for Speech



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

Tishby's Information Bottleneck

The Information Bottleneck Method
Tishby, Pereira and Bialeck, 1999

Design $P_{\text{enc}}(z|x)$ with the following objective.

$$\text{enc}^* = \underset{\text{enc}}{\text{argmin}} \; I(z,x) - \beta I(z,y)$$

END