Learning Representations

CPSC 440/550: Advanced Machine Learning

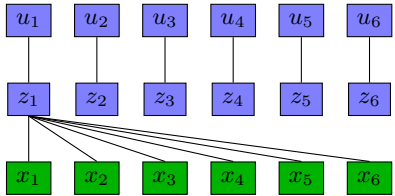
cs.ubc.ca/~dsuth/440/24w2

University of British Columbia, on unceded Musqueam land

2024-25 Winter Term 2 (Jan–Apr 2025)

Last time

- Transformers: process sequences with self-attention layers and independent MLPs



- Just stack a bunch of these together (+ tricks)

Training Transformers

- Scale the basic Transformer architecture (plus some tricks) to be huge
  - AlexNet (2012): 60 million parameters
  - ResNet-50 (2015): 25 million parameters
  - BERT: 110 million parameters
    - ModernBERT: 150 or 400 million parameters
  - GPT-2: 1.5 billion parameters
  - GPT-3: 175 billion parameters
  - GPT-4: 1.8 trillion parameters (mixture of experts)

- BERT-type (encoder-only) models: loss mostly masked language modeling
  - Replace 15% of tokens by `[MASK]`
  - Predict which token they should be

- GPT-type (decoder-only) models: loss mostly next-token prediction
  - What token should come next in the language sequence?

What to train it on?

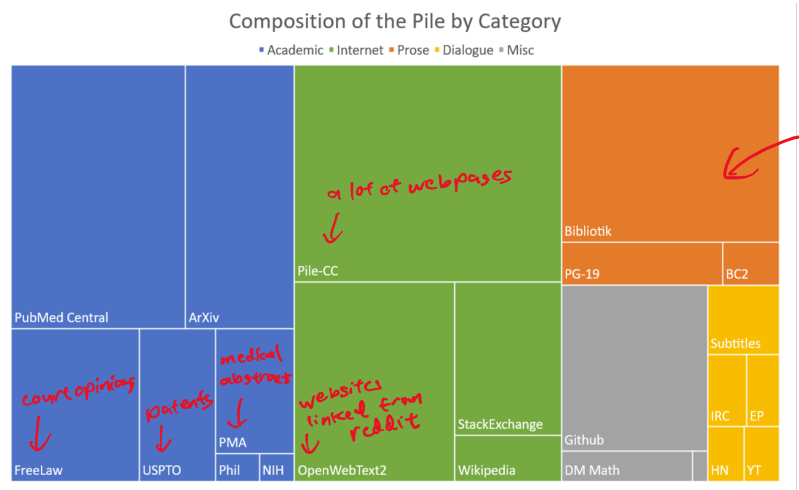
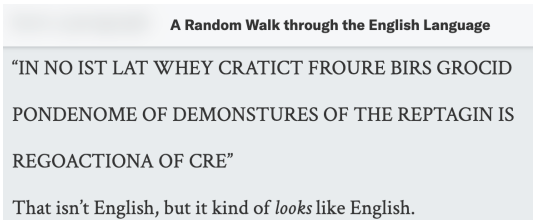


Figure 1: Treemap of Pile components by effective size.

https://arxiv.org/abs/2101.00027

Progress in language modeling

- Two-character Markov chain model of English (Shannon, 1948)



A Random Walk through the English Language

"IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID

PONDENOME OF DEMONSTURES OF THE REPTAGIN IS

REGOACTIONA OF CRE"

That isn't English, but it kind of *looks* like English.

https://www.scientificamerican.com/article/a-random-walk-through-the-english-language/

Progress in language modeling

- Two-character Markov chain model of English (Shannon, 1948)
- RNN model: became Sutskever et al., "Generating Text with RNNs," ICML 2011

George Dahl <george.dahl@gmail.com> Sep 24, 2010, 2:54 PM
to me

It is quite impressive. Look through the whole thing. Some regions are shockingly coherent. It generates the line breaks and everything. Often matching (and) and " and ". When it does a ^ it means it is a special character it doesn't know. It also doesn't know A and a are the same letter.
-G

---------- Forwarded message ----------
From: Ilya Sutskever <ilya@cs.utoronto.ca>
Date: Fri, Sep 24, 2010 at 4:13 PM
Subject: Words of Wisdom
To: George Dahl <george.dahl@gmail.com>

between ETRO, ENOS Cervical Medical Series, and Security Services. The press , Senegal, Indianapolis, East Maryland, Davis, New York.^^cite web^url^http://www.publicamine.com). In 2001, RCA has reported he also played the first Eurolomadil influmbing style, undertaken in 1966 were Stratemeigh work and relies higher than that of the . In 1974 Northern Denver had been overshadowed by CNL, and several Irish intelligence agencies in the Mediterranean region. However, on the Victoria, Kings Hebrew stated that Charles decided to escape during an alliance. The mansion house was completed in 1882, the second in its bridge are omitted, while closing is the proton reticulum composed below it aims, such that it is the blurring of

Progress in language modeling

- Two-character Markov chain model of English (Shannon, 1948)
- RNN model: became Sutskever et al., "Generating Text with RNNs," ICML 2011
- GPT-2 (2019)

Context (WebText test)	
Pinky The pink ghost's AI is designed to "feel" opposite of the red ghost's behavior. Pinky actually attempts to get out in front of Pac-Man. This is accomplished by setting the target 4 tiles ahead of Pac-Man's current location in the direction that Pac-Man is travelling. One exception to this is when Pac-Man is traveling up. Due to an overflow bug in the code, the calculation includes a left offset equal to the expected up offset. The use of Pac-Man's expected position allows the player to exploit the navigation code for Pinky. This was most often seen in situations where Pac-Man and Pinky are in close proximity to each other. For example, if they were headed toward one another, once they are less than 4 tiles apart, the target tile is now behind Pinky. If he encounters an intersection while in this configuration, he will attempt to double-back (by turning) in order to reach the target tile. Player's would often perform what was referred to as a "head fake" by changing directions rapidly just as Pinky was approaching an intersection. By displacing the target tile just as Pinky	
Small model completion	**GPT-2 completion**
was coming to the left, they would get the "correct" direction. If the location is actually farther from what Pac-Man was at, the player's will not respond to the ghost's movement, and the ghost will instead return to the player's current location. This is because Pac-Man does not move by turning. The ghost will have to go to another location of some kind to return where he began, and this is achieved	was about to cross the intersection, Pac-Man would be able to dodge the ghosts' projectiles and return to the safe location of the safe house. Buster The yellow ghost's AI is specifically designed to try to avoid Pac-Man from leaving the safe house. The player must press the SPACE bar to reveal the safe house's doors, then go into them to find the yellow ghost hiding in the next room.

Worth noting: memorization

- LLMs absolutely memorize text that they've been trained on
- How representative of this is "normal" use case?
- There are technical routes to avoid memorization (mostly differential privacy), but we still haven't figured out how to get a useful private model at scale

EXAMPLE 18: IN SILICON VALLEY RARITY HUNDREDS OF GOOGLE EMPLOYEES UNIONIZE

https://www.nytimes.com/2021/01/04/technology/google-employees-union.html
Copyright number: TX 8-954-047 Effective Date: 2021-03-22

Prompt (taken from article):

OAKLAND, Calif. — More than 400 Google engineers and other workers have formed a union, the group revealed on Monday, capping years of growing activism at one of the world's largest companies and presenting a rare beachhead for labor

Output from GPT-4:

organizers in staunchly anti-union Silicon Valley.
The union's creation is highly unusual for the tech industry, which has long resisted efforts to organize its largely white-collar workforce. It follows increasing demands by employees at Google for policy overhauls on pay, harassment and ethics, and is likely to escalate tensions with top leadership.
The new union, called the Alphabet Workers Union after Google's parent company, Alphabet, was organized in secret for the better part of a year and elected its leadership last month. The group is affiliated with the Communications Workers of America, a union that represents workers in telecommunications and media in the United States and Canada.
But unlike a traditional union, which demands that an employer come to the bargaining table to agree on a contract, the Alphabet Workers Union is a so-called minority union that represents a fraction of the company's more than 260,000 full-time employees and contractors. Workers said it was primarily an effort to give structure and longevity to activism at Google, rather than to negotiate for a contract.
Chewy Shaw, an engineer at Google in the San Francisco Bay Area and the vice chair of the union's leadership council, said the union was a necessary tool to sustain pressure on management so that workers could force changes on workplace issues.
"Our goals go beyond the

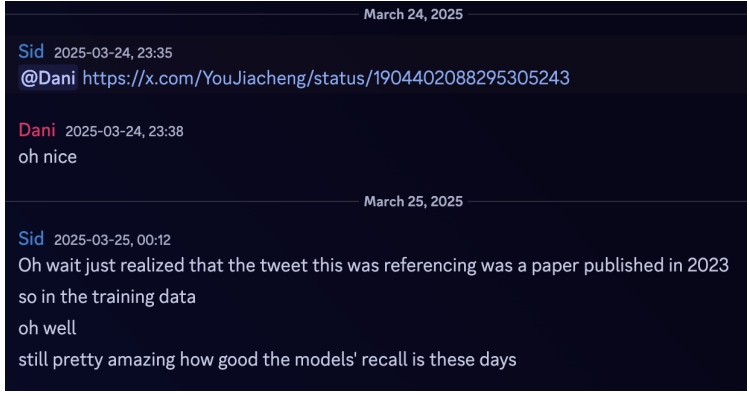
Actual text from NYTimes:

organizers in staunchly anti-union Silicon Valley.
The union's creation is highly unusual for the tech industry, which has long resisted efforts to organize its largely white-collar work force. It follows increasing demands by employees at Google for policy overhauls on pay, harassment and ethics, and is likely to escalate tensions with top leadership.
The new union, called the Alphabet Workers Union after Google's parent company, Alphabet, was organized in secret for the better part of a year and elected its leadership last month. The group is affiliated with the Communications Workers of America, a union that represents workers in telecommunications and media in the United States and Canada.
But unlike a traditional union, which demands that an employer come to the bargaining table to agree on a contract, the Alphabet Workers Union is a so-called minority union that represents a fraction of the company's more than 260,000 full-time employees and contractors. Workers said it was primarily an effort to give structure and longevity to activism at Google, rather than to negotiate for a contract.
Chewy Shaw, an engineer at Google in the San Francisco Bay Area and the vice chair of the union's leadership council, said the union was a necessary tool to sustain pressure on management so that workers could force changes on workplace issues.
"Our goals go beyond the

https://x.com/paul_cal/status/1740773165205479435

Less-direct memorization

- Training data is so huge that it's easy to trick yourself
- Often when "Our AI model aces X test!" it fails next year's exam
- Example of friend telling me about DeepSeek reproducing a neat proof



- We'll talk more about these kinds of issues soon
- First: what is this process of "language modeling" really doing?

Outline

BERT-type models

- Typical usage for encoder-only models:
  - Train to do masked language modeling on lots of text
  - Use learned representations as features for the problem you actually care about
- For example, `transformers.BertForSequenceClassification`:
  - Take per-token features that were input to language modeling task
  - Take their mean over the sentence/document/...
  - Train a linear classifier from these features to whatever label
- Or train a different per-word, per-sentence, ... task

- Compare to: learning a spam detector (or whatever) from scratch
  - Everything in the model can be absolutely geared to detecting spam!
  - ... but also you have to learn language from scratch

- Here, masked language modeling is a pretext task to help learn good features
- The downstream task is hopefully easier with good features!

Another pretext task: autoencoders

- PCA, also known as a linear autoencoder:

$$\min_{\text{linear } f,g} \sum_{i=1}^{n} \|x^{(i)} - g\left(f\left(x^{(i)}\right)\right)\|^2$$

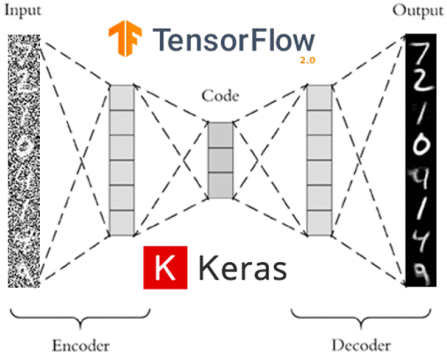
  - Train to reproduce the input data
  - We don't usually care about reproducing the input data
- Nonlinear "plain" autoencoders: exactly the same idea

$$\min_{\phi,\theta} \sum_{i=1}^{n} \|x^{(i)} - g_\theta\left(f_\phi\left(x^{(i)}\right)\right)\|^2$$

- VAEs: can view as randomized, regularized version of plain autoencoders
- Or as a fancy "continuous" version of clustering: same idea

Denoising autoencoders

- If f, g are too powerful, $\min_{\phi,\theta} \sum_{i=1}^{n} \|x^{(i)} - g_\theta\left(f_\phi\left(x^{(i)}\right)\right)\|^2$ can become trivial
- Alternative: learn to remove noise from inputs



https://www.pyimagesearch.com/2020/02/24/denoising-autoencoders-with-keras-tensorflow-and-deep-learning/

- Easy to get a lot of data!
- Interesting pretext task

Pretext tasks in vision: relative patch prediction

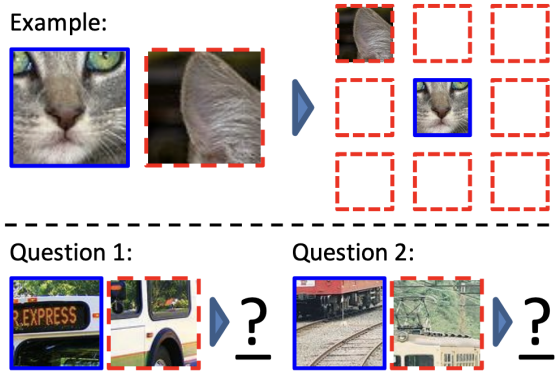


Figure 1. Our task for learning patch representations involves randomly sampling a patch (blue) and then one of eight possible neighbors (red). Can you guess the spatial configuration for the two pairs of patches? Note that the task is much easier once you have recognized the object!

Pretext tasks in vision: jigsaw puzzles

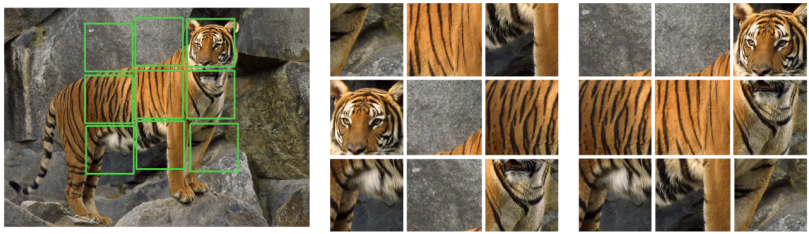


Fig. 1: Learning image representations by solving Jigsaw puzzles. (a) The image from which the tiles (marked with green lines) are extracted. (b) A puzzle obtained by shuffling the tiles. Some tiles might be directly identifiable as object parts, but others are ambiguous (*e.g.*, have similar patterns) and their identification is much more reliable when all tiles are jointly evaluated. In contrast, with reference to (c), determining the relative position between the central tile and the top two tiles from the left can be very challenging [10].

https://arxiv.org/abs/1603.09246

Pretext tasks in vision: colourization

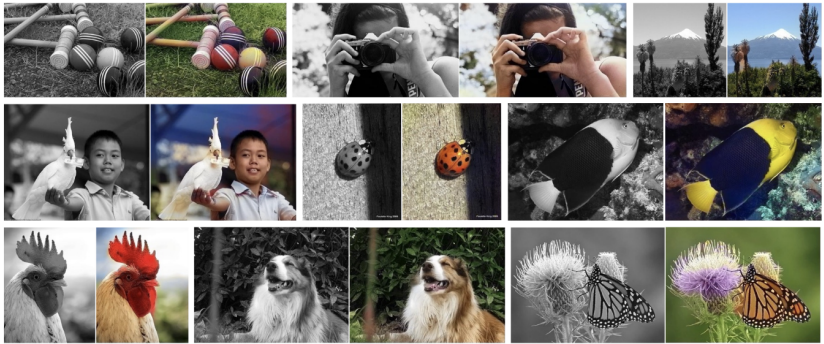


Fig. 1. Example input grayscale photos and output colorizations from our algorithm. These examples are cases where our model works especially well. Please visit `http://richzhang.github.io/colorization/` to see the full range of results and to try our model and code. Best viewed in color (obviously).

https://arxiv.org/abs/1603.08511

Don't get too excited about colourization

bonus!

https://www.youtube.com/watch?v=Xn36bUkSHVY

Don't get too excited about colourization

bonus!

https://www.youtube.com/watch?v=Xn36bUkSHVY

Pretext tasks in vision: contrastive learning

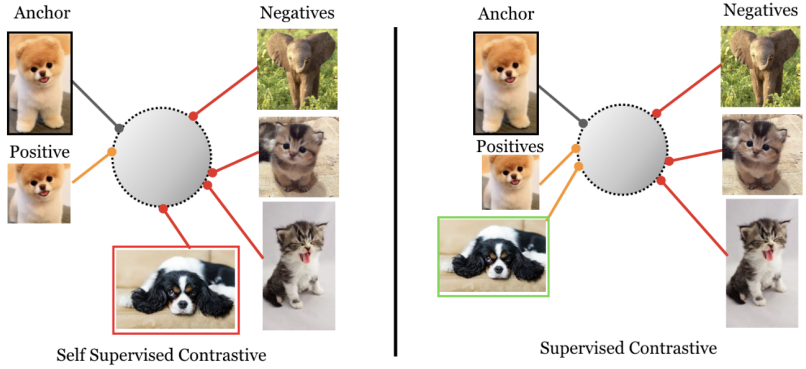
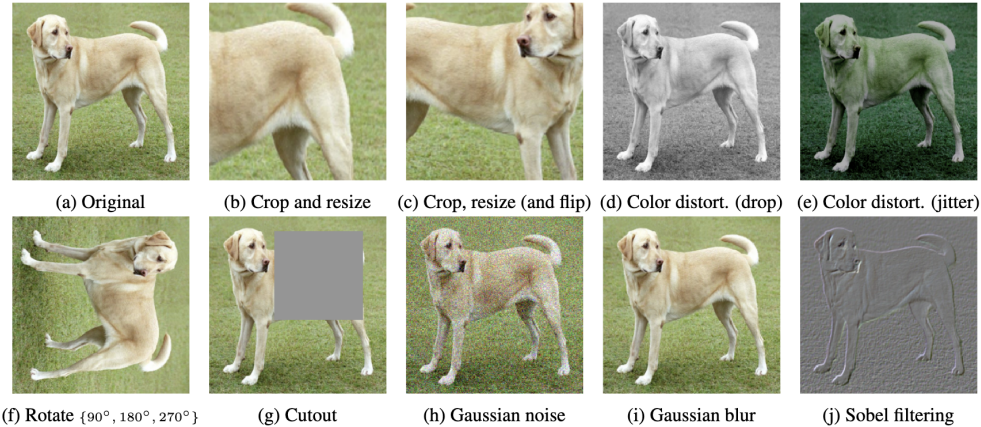


Figure 2: Supervised vs. self-supervised contrastive losses: The self-supervised contrastive loss (left, Eq. 1) contrasts a *single* positive for each anchor (i.e., an augmented version of the same image) against a set of negatives consisting of the entire remainder of the batch. The supervised contrastive loss (right) considered in this paper (Eq. 2), however, contrasts the set of *all* samples from the same class as positives against the negatives from the remainder of the batch. As demonstrated by the photo of the black and white puppy, taking class label information into account results in an embedding space where elements of the same class are more closely aligned than in the self-supervised case.

Pretext tasks in vision: contrastive learning



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

https://arxiv.org/abs/2002.05709

Pretext tasks in vision: contrastive learning

- Contrastive loss as in SimCLR:

$$\ell_{i,j} = -\log \frac{\exp(z^{(i)} \cdot z^{(j)}/\tau)}{\sum_{k \neq i} \exp(z^{(i)} \cdot z^{(k)}/\tau)}$$

  for $\|z^{(i)}\| = 1$, τ a constant temperature
- "Does z_j look most like z_i out of the batch?"
- Construct a batch of two transforms of each of a bunch of images

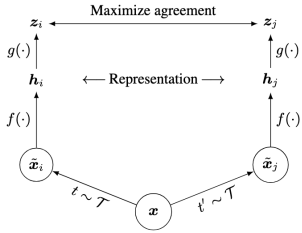
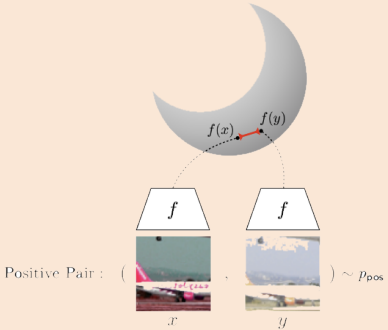


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

https://arxiv.org/abs/2002.05709

What contrastive loss does

bonus!



Alignment: Similar samples have similar features.
(Figure inspired by Tian et al. (2019).)

Uniformity: Preserve maximal information.

https://arxiv.org/abs/2005.10242

Contrastive loss: easy to adapt to other settings

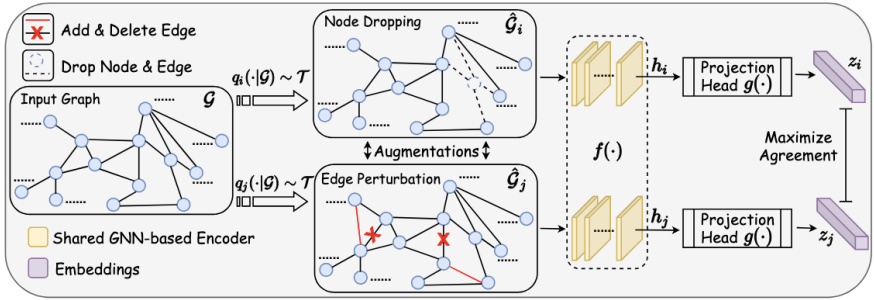


Figure 1: A framework of graph contrastive learning. Two graph augmentations $q_i(\cdot|\mathcal{G})$ and $q_j(\cdot|\mathcal{G})$ are sampled from an augmentation pool $\mathcal{T}$ and applied to input graph $\mathcal{G}$. A shared GNN-based encoder $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize the agreement between representations z_i and z_j via a contrastive loss.

https://arxiv.org/abs/2010.13902

Challenges in contrastive learning

bonus!

Less can be more in contrastive learning

Jovana Mitrovic **Brian McWilliams** **Melanie Rey**
DeepMind, UK
{mitrovic, bmcw, melanierey}@google.com

Abstract

Unsupervised representation learning provides an attractive alternative to its supervised counterpart because of the abundance of unlabelled data. Contrastive learning has recently emerged as one of the most successful approaches to unsupervised representation learning. Given a datapoint, contrastive learning involves discriminating between a matching, or positive, datapoint and a number of non-matching, or negative, ones. Usually the other datapoints in the batch serve as the negatives for the given datapoint. It has been shown empirically that large batch sizes are needed to achieve good performance, which led the the belief that a large number of negatives is preferable. In order to understand this phenomenon better, in this work investigate the role of negatives in contrastive learning by decoupling the number of negatives from the batch size. Surprisingly, we discover that for a fixed batch size performance actually degrades as the number of negatives is increased. We also show that using fewer negatives can lead to a better signal-to-noise ratio for the model gradients, which could explain the improved performance.

https://proceedings.mlr.press/v137/mitrovic20a/mitrovic20a.pdf

Information-theoretic view of contrastive learning

bonus!

- Can motivate SimCLR-type loss through mutual information
- Minimizing the SimCLR loss maximizes (a bound on) $\mathrm{MI}\left(f(X^{(1)}), f(X^{(2)})\right)$ where $X^{(1)}$ and $X^{(2)}$ are two transformations of the same source image

- Mutual information measures dependence between two random variables

$$\mathrm{MI}(X, Y) = \mathrm{KL}(p(x, y) \parallel p(x)p(y)) = \mathop{\mathbb{E}}_{X,Y}\left[\log \frac{p(X, Y)}{p(X)p(Y)}\right]$$

- Zero if and only if $X \perp\!\!\!\perp Y$
- Invariant to invertible transformations of the variables

- Problem: any invertible f has same mutual information, but can have vastly different downstream "usefulness"!

So why does contrastive learning work?

bonus!

- Possibility: in a SimCLR-type architecture, we don't actually maximize the MI
- The InfoNCE lower bound on MI that we maximize depends on geometry
- . . . and particularly the same linear geometry we use for fine-tuning
- SimCLR has close connections to a kernel method :)

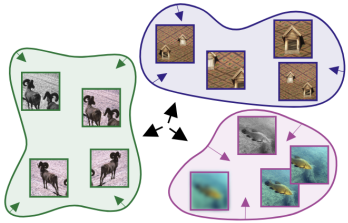
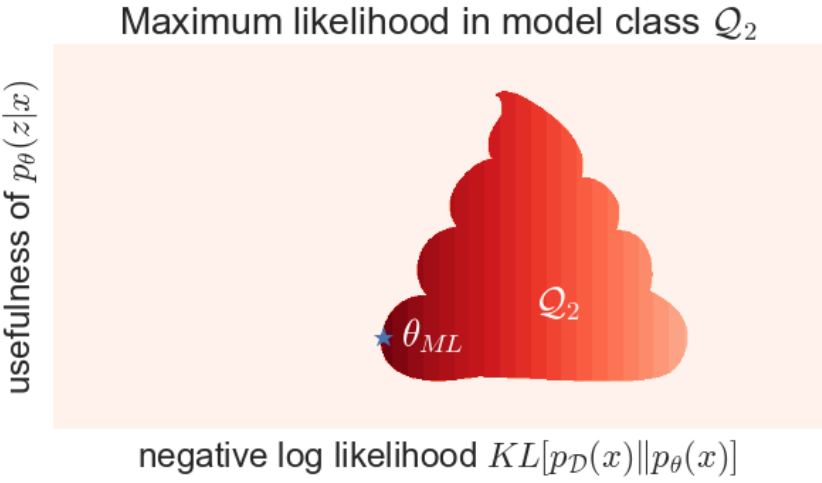


Figure 2: Statistical dependence view of contrastive learning: representations of transformed images should highly depend on image identity. Measuring dependence with HSIC, this pushes different images' representation distributions apart (black arrows) and pulls representations of the same image together (colored shapes).

Reminder



https://www.inference.vc/maximum-likelihood-for-representation-learning-2/

Pretext learning is hard in general!

- True for getting a "useful representation" out of ELBO maximization
- Also true for getting a "useful representation" out of contrastive learning/...
- Same thing is true for turning large language models into chatbots! (More next time)

Multimodal contrastive learning: CLIP

bonus!

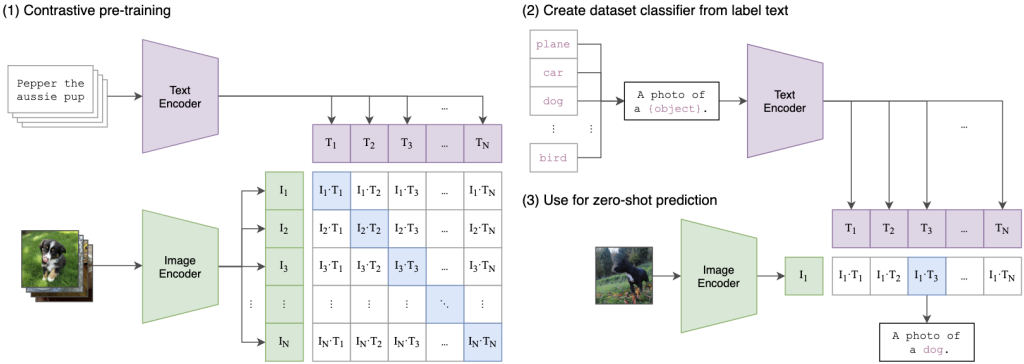


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

https://arxiv.org/abs/2103.00020