Lecture 20: Self-supervised learning

CS 182/282A (“Deep Learning”)

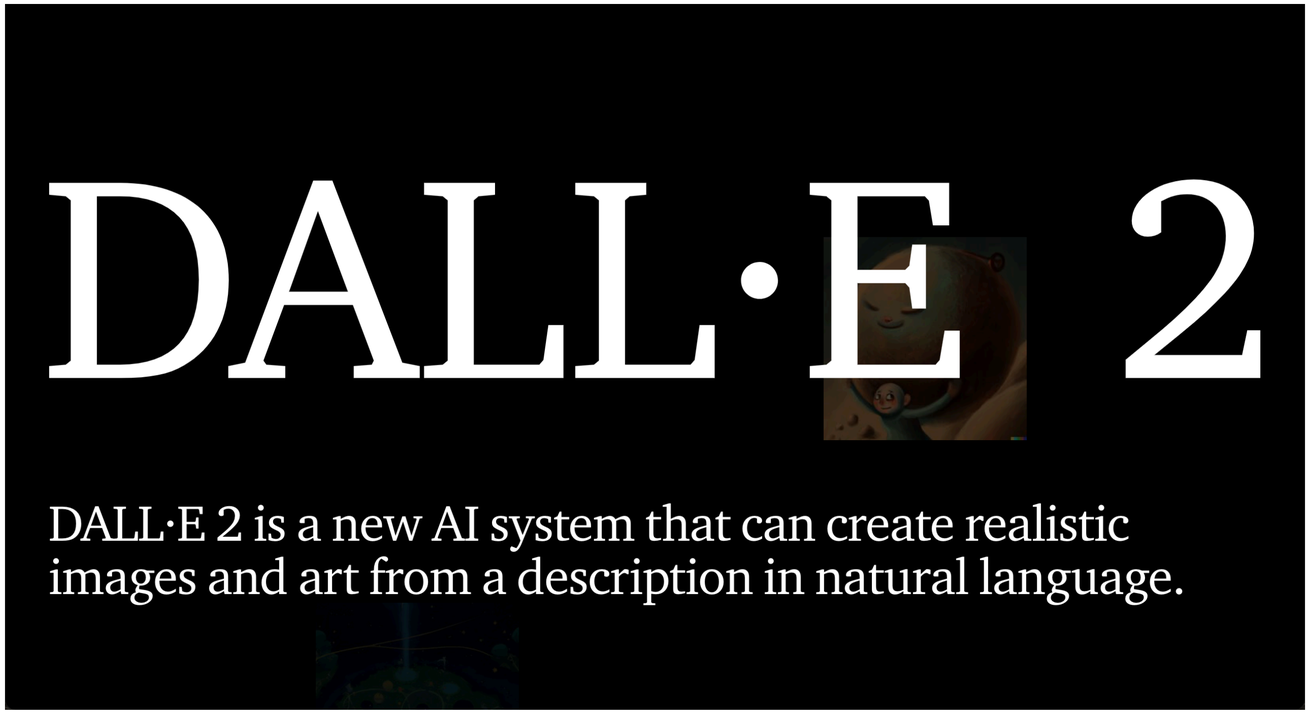
2022/04/11

Today's lecture

- Continuing on our journey of deep unsupervised learning, today we will focus on methods for **self-supervised learning** — "creating labels from unlabeled data"
- Sort of like: "hide" some information from the model and ask it to predict this
- Self-supervised learning and generative modeling are not mutually exclusive, but they are not the same in general
- Both self-supervised learning and generative modeling have been shown to be effective for representation learning in a number of domains
- Before that, though, let me show you something related to last time…

Generating images from language

- Remember when I said last lecture that we don't really yet have models that can generate realistic images given language descriptions? Well, that same day…



Some DALL·E 2 creations

“Gandalf the Grey in the style of Picasso”

Some DALL·E 2 creations

“A majestic horse standing next to an equally majestic zebra”

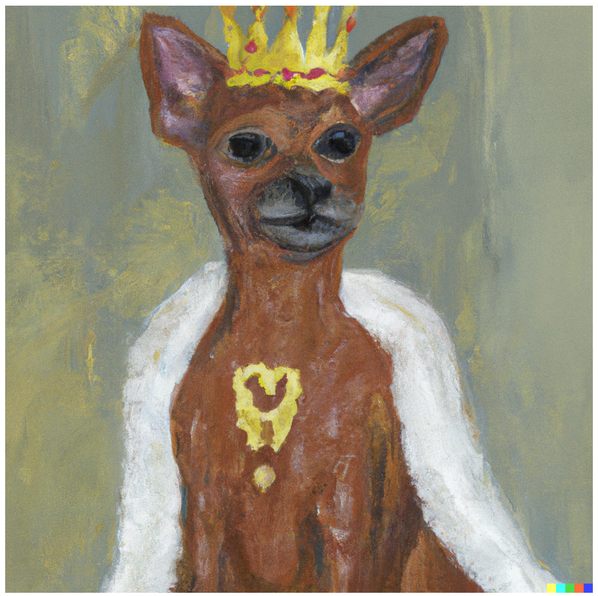
Some DALL·E 2 creations

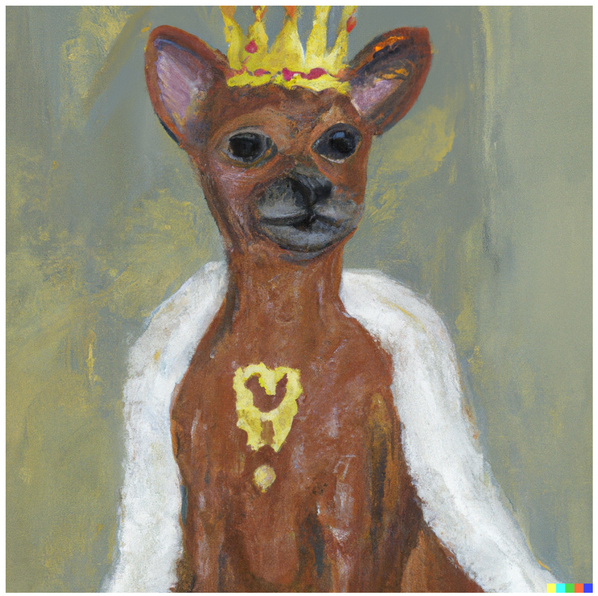
“A stage view of a lecture in session”

Some DALL·E 2 creations

“A painting of a small brown terrier with round ears wearing a regal robe and crown”

(my dog, Peanut)





Briefly: diffusion models

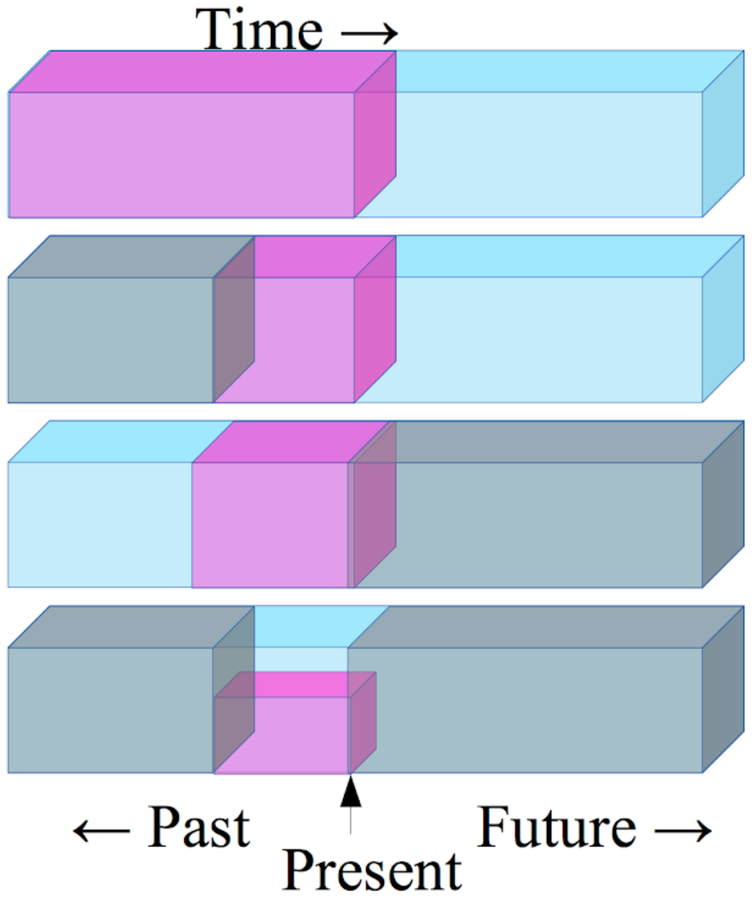
- Under the hood, the generative model being used is a **diffusion model**
- Diffusion models define a process by which $\mathbf{x}$ is transformed, little by little, via additive Gaussian noise, into $\mathbf{z}$, which is pure noise
- In particular, the model parameterizes the *reverse process* ($\mathbf{z}$ to $\mathbf{x}$), and the *forward process* ($\mathbf{x}$ to $\mathbf{z}$) is fixed as incrementally adding small amounts of noise
  - Similar to VAEs, we then train by maximizing the evidence lower bound
- Diffusion models generate impressive samples, however, sampling is expensive

Self-supervised learning

What is self-supervised learning?

According to Prof. Yann LeCun

- **Predict any part of the input from any other part.**
- **Predict the future from the past.**
- **Predict the future from the recent past.**
- **Predict the past from the present.**
- **Predict the top from the bottom.**
- **Predict the occluded from the visible**
- **Pretend there is a part of the input you don't know and predict that.**

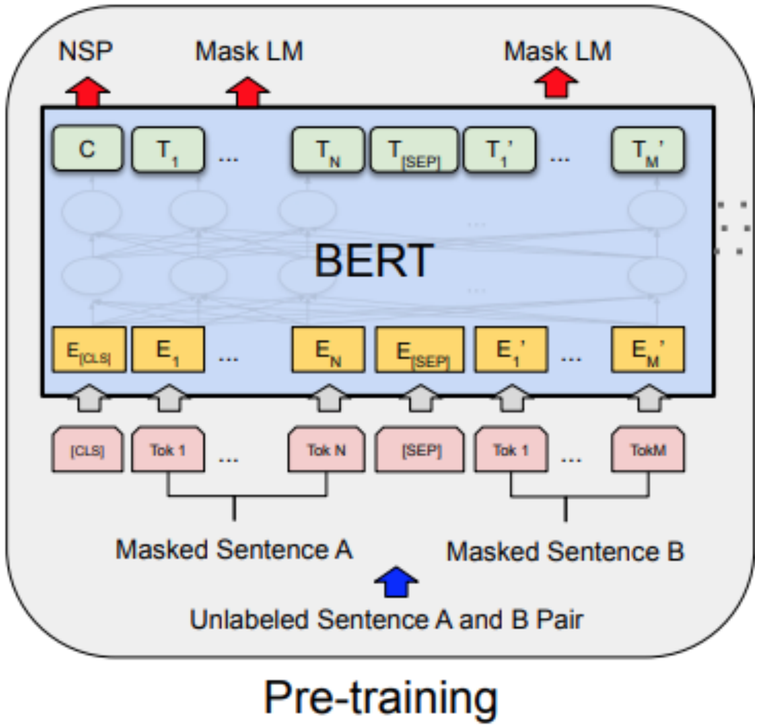


Why self-supervised learning?

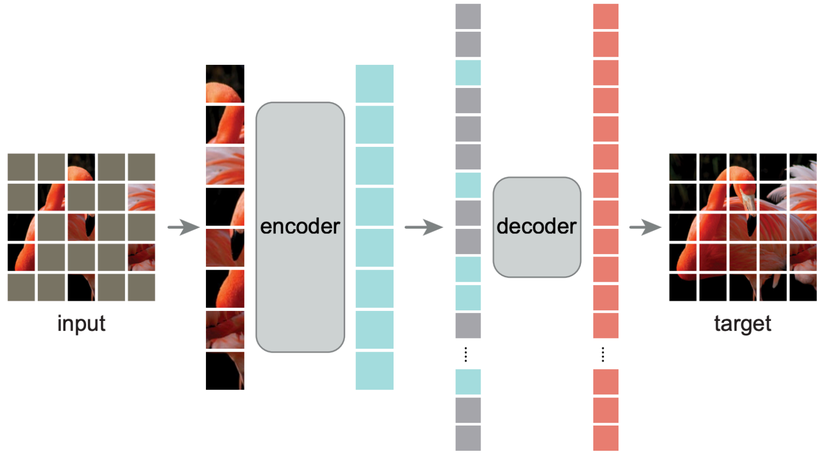
- “Give a robot a label and you feed it for a second, teach a robot to label and you feed it for a lifetime.” — Pierre Sermanet
- Good supervision is often not cheap, and we already have vast amounts of unlabeled data on the internet for a number of different modalities
- These data may be able to teach models about the *structure* of the domain
  - E.g., the ability to predict parts of the data given other parts
- Or, we may, as humans, have additional *domain knowledge* we can pass on to the model via training, e.g., through data augmentation

Have we seen self-supervised learning before?

- Yes we have!



Pre-training



Prompt: Recycling is good for the world, no, you could not be more wrong.

GPT-2: Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources.

Masked autoencoding

- Randomly masking out parts of the input (e.g., 15% of tokens or 75% of image patches) and predicting these parts is a very effective self-supervised approach
- MAE vision transformers and BERT are both primarily transformer *encoders* that turn (masked) inputs into representations that are useful for downstream tasks
- During training, they are equipped with simple *decoders*, e.g., token classifiers for BERT and a small transformer (but not a transformer decoder) for MAE
    - These decoders attempt to recover the original input that was masked out, and the encoder is thus trained to produce useful contextual representations

Self-supervised learning in vision

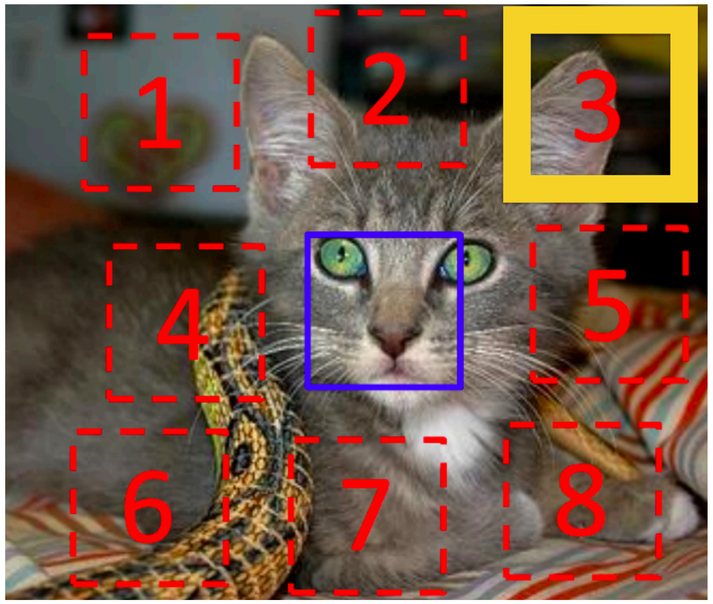
- For a long time, masked autoencoding was not a performant or scalable approach to self-supervised learning in computer vision
- As a result, many other techniques have been developed for this domain
- Some ideas have persisted through many of these techniques: leveraging the known structure of images, **contrastive learning**, and data augmentation
- We will go through these techniques in (roughly) chronological order

Leveraging spatial context

Doersch et al, ICCV 2015

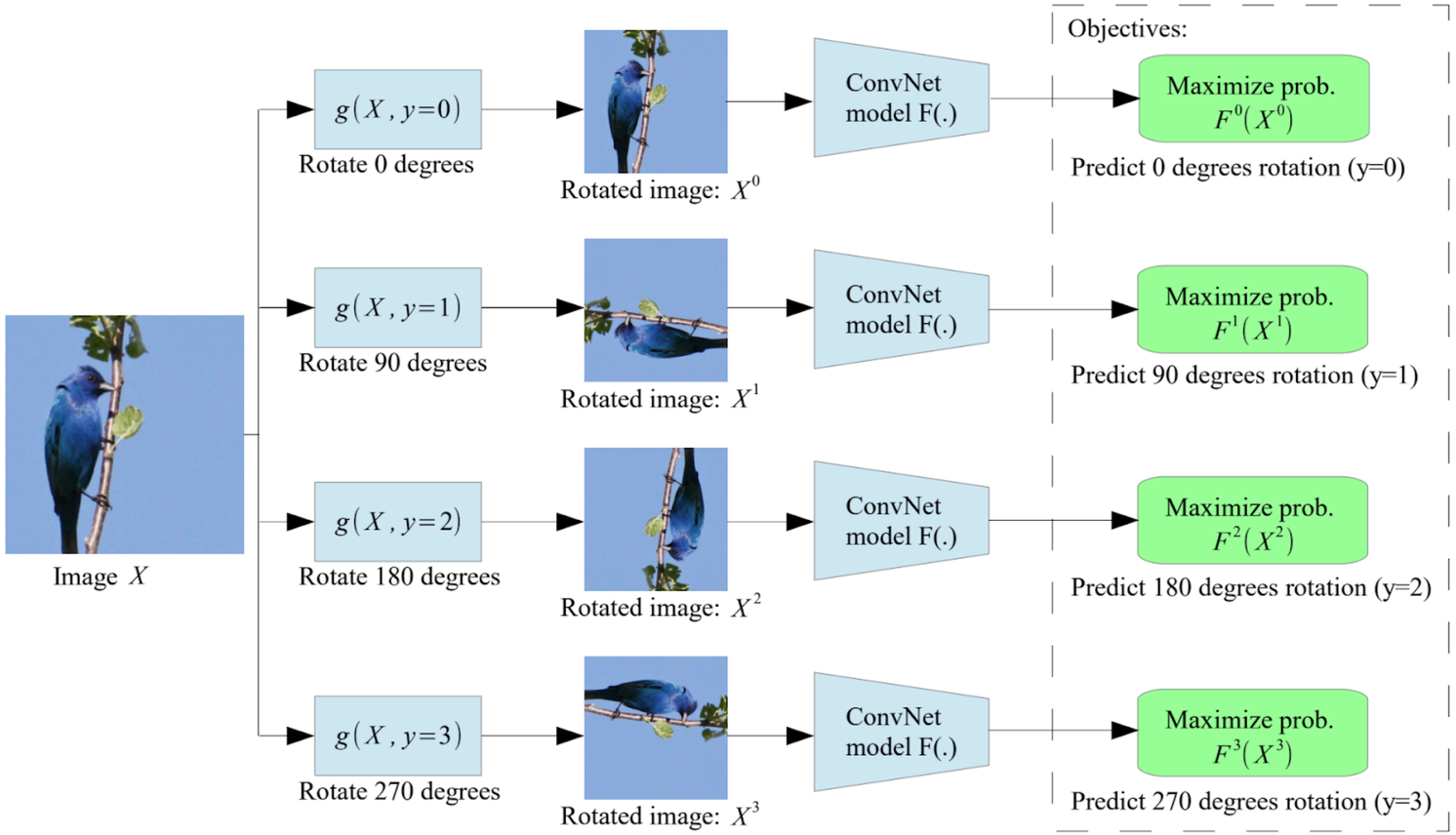
- Self-supervised task: extract two patches from the image, and predict the relative position of the second patch with respect to the first

- This task proved to be useful in pretraining conv nets for downstream object detection tasks, likely because the network learns about objects and their parts



Predicting rotations

Gidaris et al, ICLR 2018



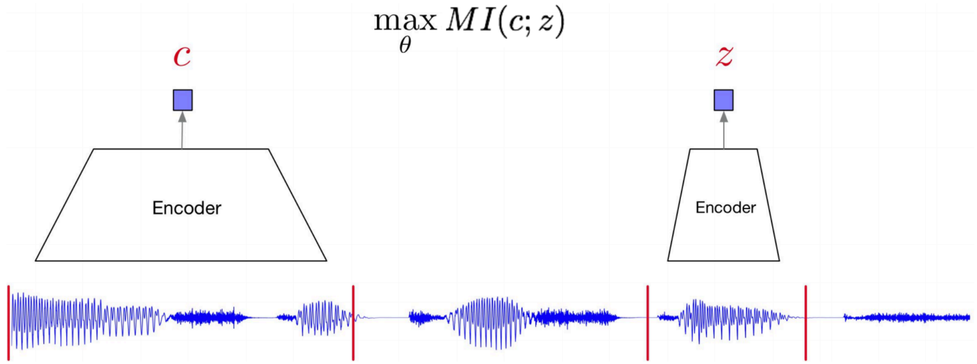
Contrastive learning

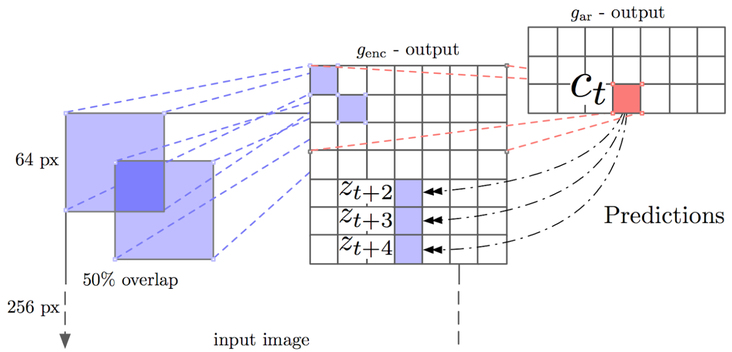
- Several self-supervised learning methods are based on **contrastive learning**: learned representations should be close together for "similar" inputs and far apart for "dissimilar" inputs (we will define "similar" and "dissimilar" shortly)
- Suppose we have a representation $\mathbf{z}$, generated from some input, along with representations $\mathbf{z}_1, \ldots, \mathbf{z}_K$ generated from other inputs
  - Suppose one of $\mathbf{z}_1, \ldots, \mathbf{z}_K$ is generated from a similar input, call that $\mathbf{z}_+$
- The most common contrastive learning loss is $-\log \frac{\exp\{\mathbf{z}^\top \mathbf{z}_+ / \tau\}}{\sum_{i=1}^K \exp\{\mathbf{z}^\top \mathbf{z}_i / \tau\}}$

Contrastive predictive coding (CPC)

van den Oord et al, 2018

- CPC defines “similar” as coming from the same input, which requires splitting up the input into multiple parts, and “dissimilar” as coming from different inputs
- This general approach has been successfully applied to audio, images, and text
- However, extracting a single representation from the whole input then takes a bit more work





The usefulness of data augmentation

- Many other types of data augmentation, besides rotations, can be leveraged to produce useful self-supervised learning signals in computer vision
- However, rather than trying to predict the augmentation, as we did for rotations, we will be using data augmentations in conjunction with contrastive learning
- Specifically, we will consider different augmentations of the same image as similar and different images as dissimilar
- This allows us to learn representations of entire images directly, since we don’t need to split the image to obtain similar inputs

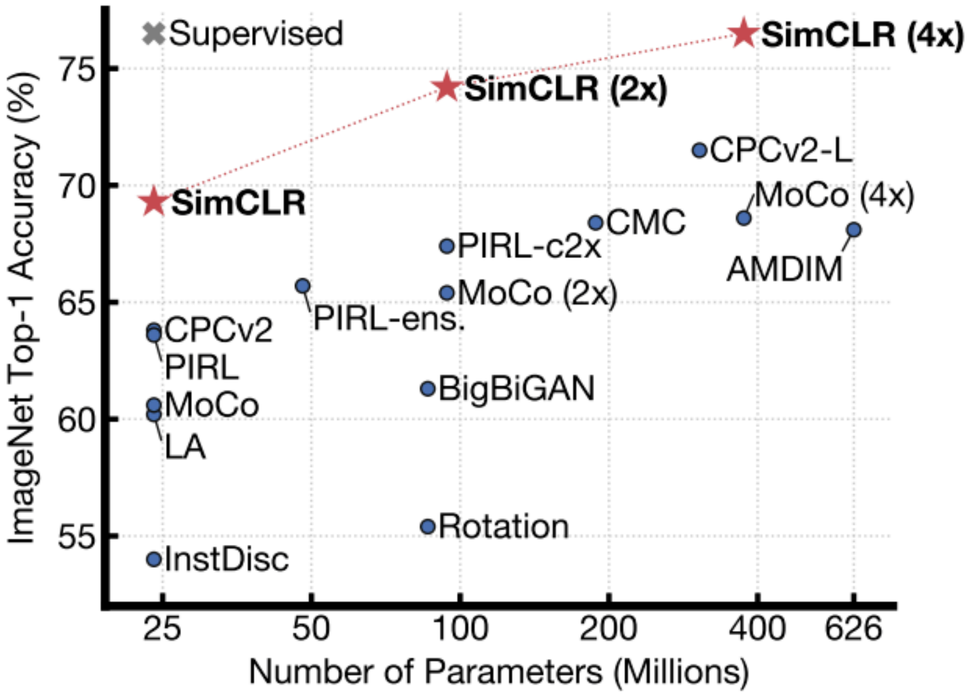
Momentum contrast (MoCo)

He et al, 2019

- A key challenge in contrastive learning is sampling a sufficiently large and diverse set of dissimilar (**negative**) examples for training
- If it is too large, then generating representations for all negatives will be expensive
- MoCo maintains a *queue* of negatives and adds to it every training iteration with the representations from the latest mini batch, pushing out the oldest mini batch
- Since the representation itself is changing during training, MoCo uses an *exponential moving average (EMA)* of the model parameters to generate representations for the queue, thus encouraging similarity across mini batches
- MoCo and subsequent improvements (MoCo v2 and v3) have worked quite well

SimCLR

- We can remove the need for a queue for contrastive learning if we incorporate: larger batch sizes, longer training, larger models, stronger data augmentations, and a few other technical improvements
- These findings are summarized by the simple framework for contrastive learning of visual representations, or **SimCLR**



Bootstrap your own latent (BYOL)

Grill et al, NeurIPS 2020

- Contrastive learning is a popular paradigm for self-supervised learning in vision, but there are other methods that take different approaches
- Another common motif is the use of two networks during training (sometimes called the *online* and *target* networks, or the *student* and *teacher* networks)
- In BYOL, the online and target networks are given two augmentations of the same input, and the online network is optimized to try and predict the target network's representation
- The target network is updated as an EMA of the online network
- This general scheme ends up working surprisingly well

Self-distillation with no labels (DINO)

Caron et al, ICCV 2021

- Another way to use two networks is to simply have one (the student) try to match the probability values of the other (the teacher)
- This idea is called **distillation** in the case where the teacher “knows something”, e.g., it has been pretrained on some dataset
- **Self-distillation** is the self-supervised learning equivalent of this idea, and it is interesting that this approach can work even when the teacher “knows nothing”

Multimodal contrastive learning: CLIP

Radford et al, 2021

- We are increasingly seeing powerful **multimodal** models which combine information across modalities — most commonly, images and text
- Contrastive language-image pretraining (CLIP) is one such model trained on a large dataset of (image, text) pairs collected from the internet
- The model contains both a text encoder and an image encoder, which are trained with a contrastive objective where the (image, text) pairs are similar and all other non pairs are dissimilar
- CLIP is an important piece of the DALL·E 2 system

Summary

- We have now seen two broad branches of deep unsupervised learning: **deep generative modeling** and **self-supervised learning**
- Deep generative modeling aims to do distribution/density modeling, whereas self-supervised learning more directly targets representation learning
  - These characterizations are not mutually exclusive, and sometimes it is the combination of ideas that leads to the most impressive systems
- Deep unsupervised learning is the collection of techniques that have the best chance of scaling to massive models and massive datasets, and there is still plenty of work to be done