

CS 446/ECE 449: Machine Learning

Lecture 23: Contrastive Learning & Self-supervised Learning

Han Zhao
04/11/2024



Today

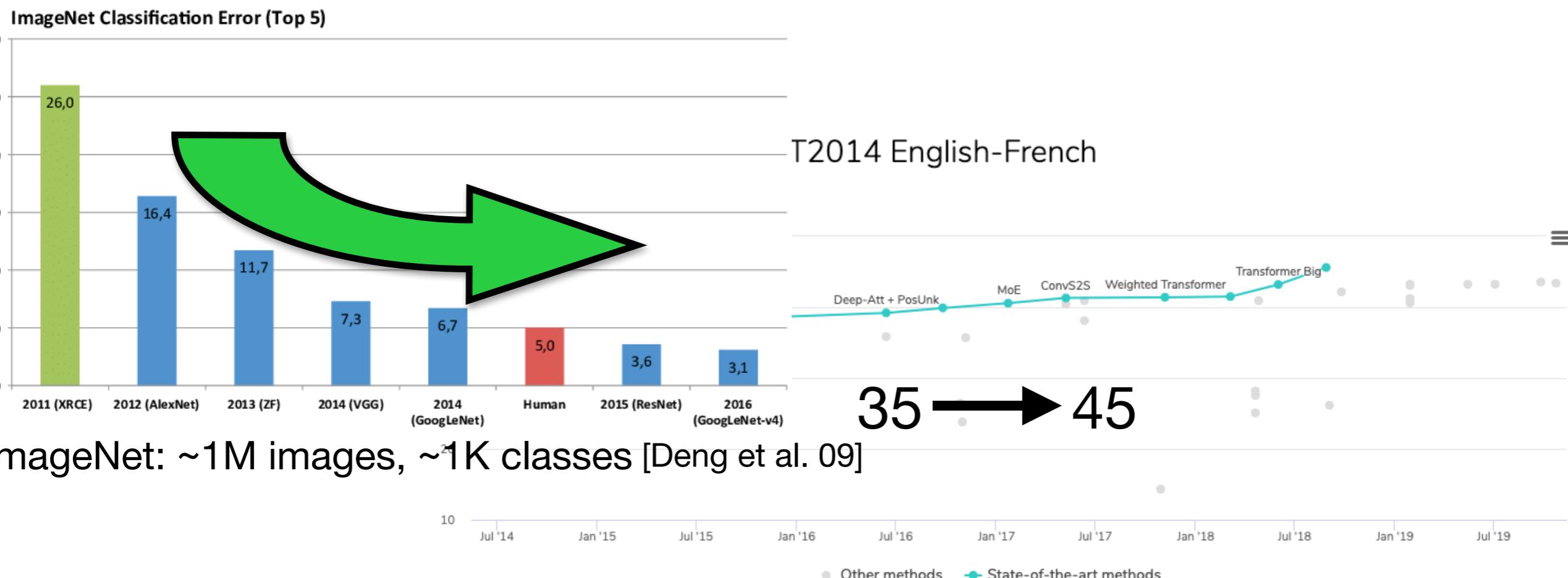
- Concept of self-supervised learning
- Pretext task training
- Contrastive representation learning

Recap: Supervised Learning

Classification / Regression:

With large-scale labeled data, we can achieve

- Image classification
- Speech recognition
- Machine translation
-



ImageNet: ~1M images, ~1K classes [Deng et al. 09]

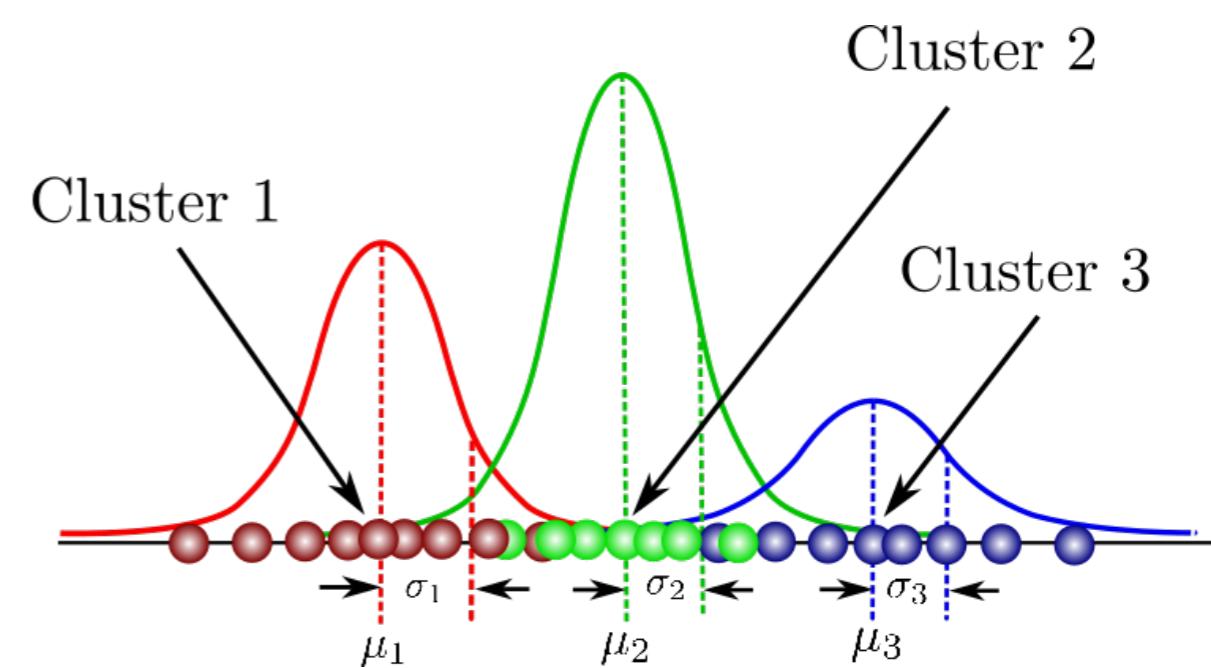
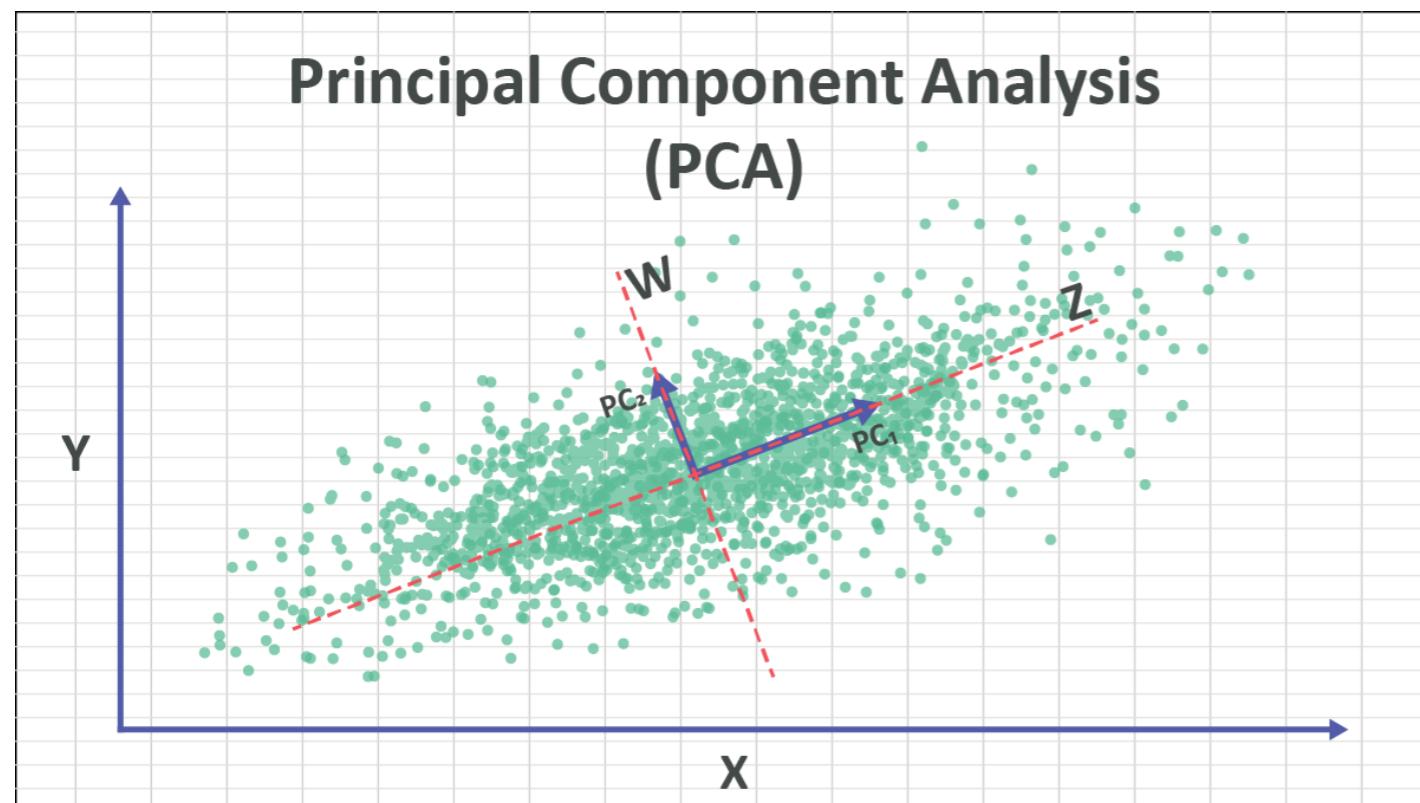
Machine Translation, ~3M parallel sentences [Cho et al.'14; Devlin et al.'14]³

Recap: Unsupervised Learning

Clustering / Density estimation / dimensionality reduction

What we have covered so far:

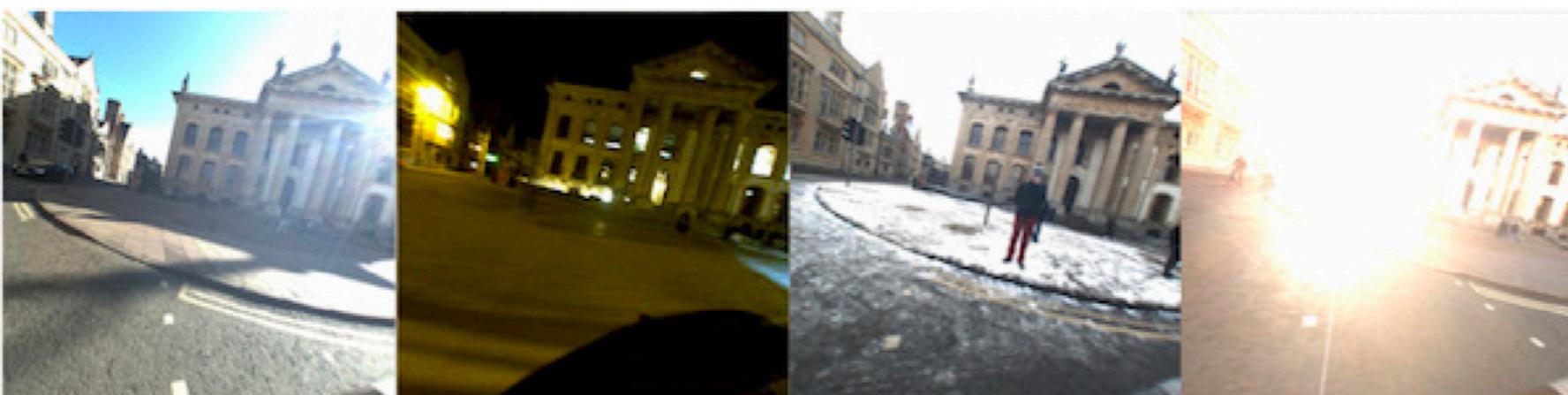
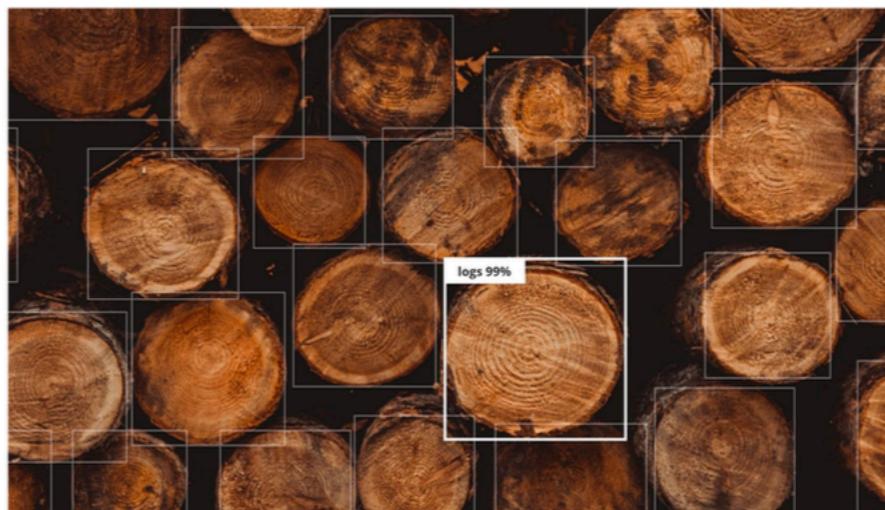
- K-means (clustering)
- Gaussian mixture model (density estimation/clustering)
- Principal component analysis (dimensionality reduction)
-



Representation Learning

Real-world ML deployment faces various challenges:

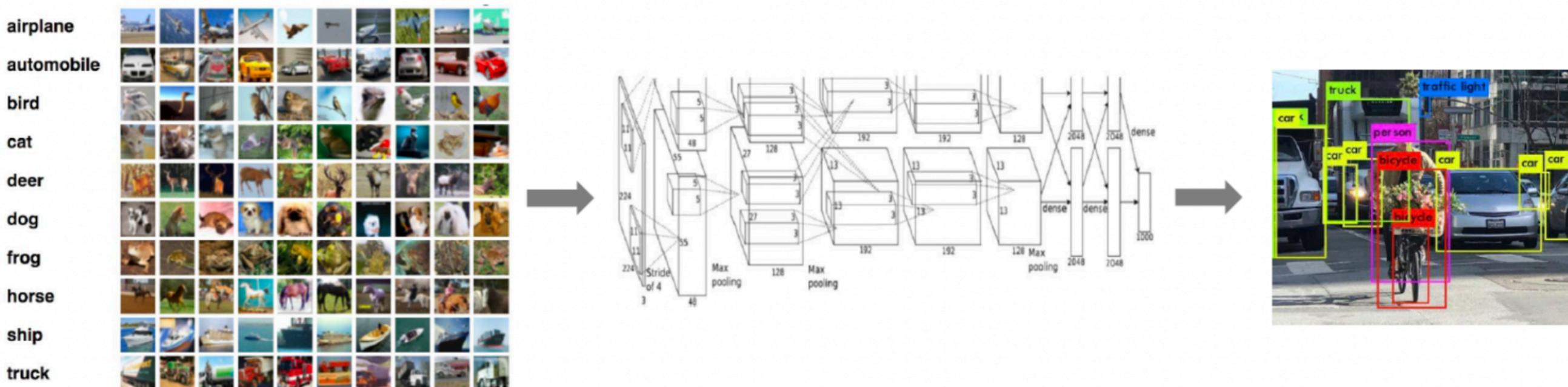
- Annotation is expensive
- Hard to collect large-scale high-quality data
- Potential distribution shifts
- Potential label noise in data collection
-



Representation Learning

Pretraining (representation learning) could be helpful:

- Pre-train the network on other labeled tasks (e.g., image classification on ImageNet)
- Fine-tune (part of) the weights on new down-stream tasks (e.g., object detection/recognition, etc)
- Typical paradigm of transfer learning:
 - 1st stage: Pre-training
 - 2nd stage: fine-tuning



What if we can only have access to large-scale unlabelled data? e.g., from ImageNet -> Internet-scale image data

Representation Learning

Self-supervised learning (SSL):

Learning good **data representation** from **unlabelled** data

Key idea: constructing supervised learning tasks from unlabelled datasets

Why?

- Data labelling could be very expensive
- Learning good and generic representations could benefit diverse downstream tasks

Self-supervised learning vs Unsupervised learning?

- SSL could be considered as a specific approach within the broader category of unsupervised learning
- In SSL the model generates its own supervisory signal from the input data itself, and then uses typical losses from supervised learning

Representation Learning

Self-supervised learning (SSL):

Learning good **data representation** from **unlabelled** data

Key idea: constructing supervised learning tasks from unlabelled datasets

Typical paradigms of SSL

- Pretext tasks
- Contrastive learning

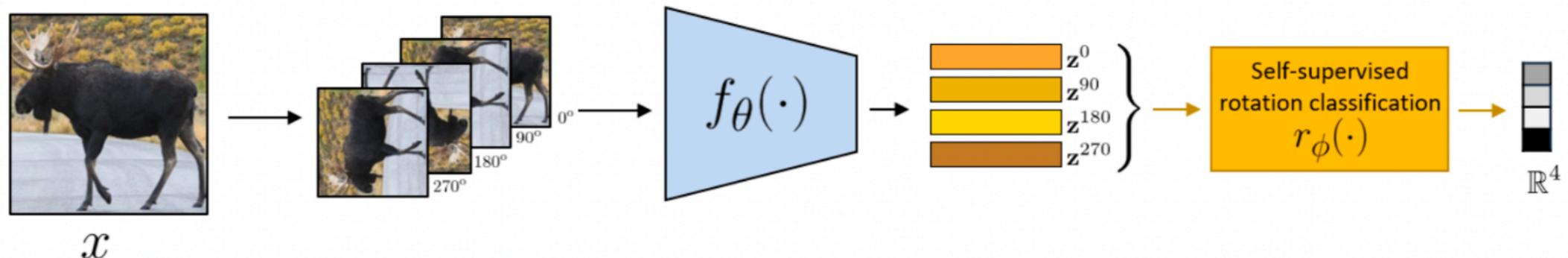
Representation Learning

Self-supervised learning (SSL):

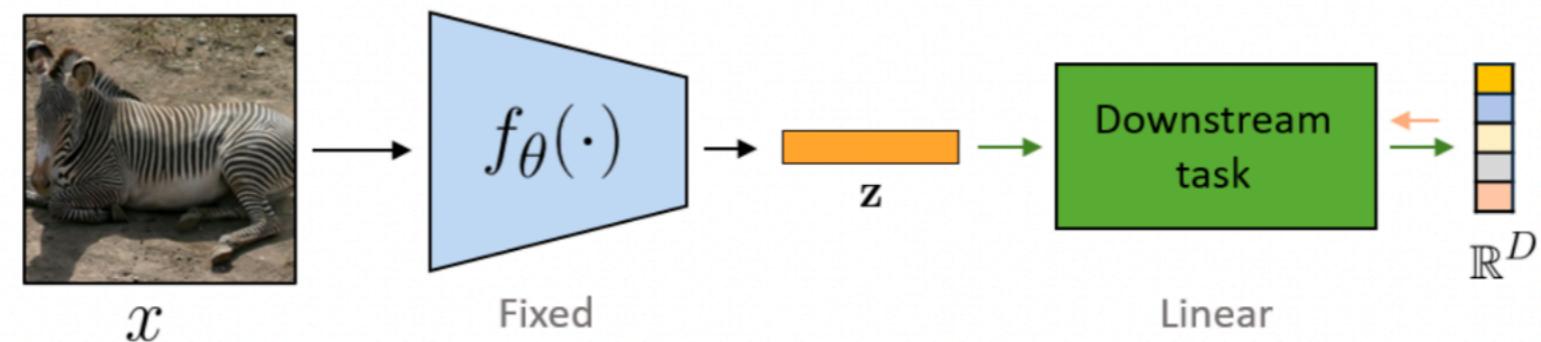
Learning good **data representation** from **unlabelled** data

Overview:

Stage 1: Train network on pretext task (without human labels)



Stage 2: Train classifier on learned features for new task with fewer labels



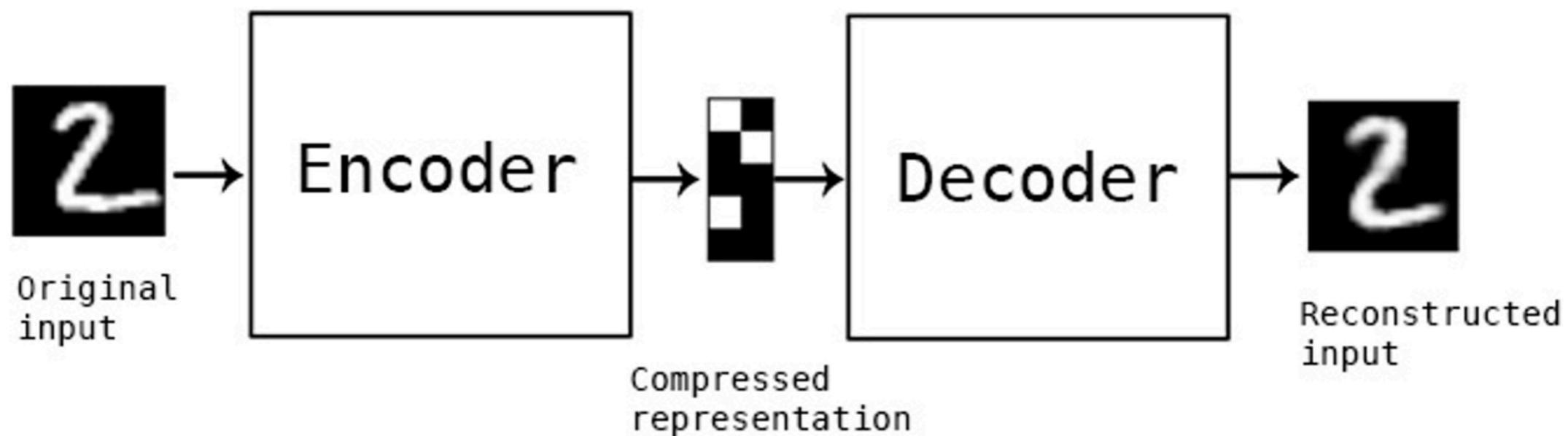
What is a good pretext task?

Representation Learning

Reconstruction loss: auto-encoders

Learns a low-dimensional compact representation that can encoder the input data for reconstruction

- The label is the input image itself
- Latent code learns to encode the data compactly



In fact, auto-encoders are generalization of principal component analysis
(from linear to nonlinear)

Representation Learning

Image-specific self-supervised losses:

- Colorize an image
- Infer the rotation angle of an image
- Solve the jigsaw puzzle
- Image completion
-

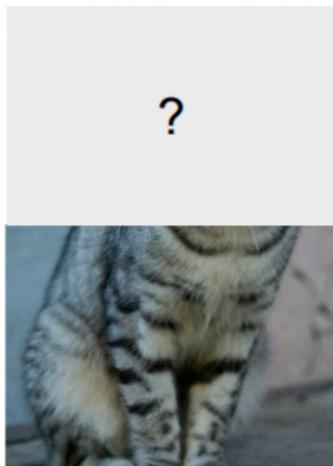
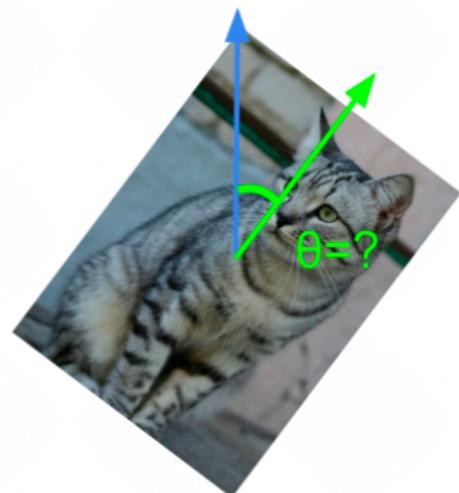


image completion



rotation prediction



“jigsaw puzzle”



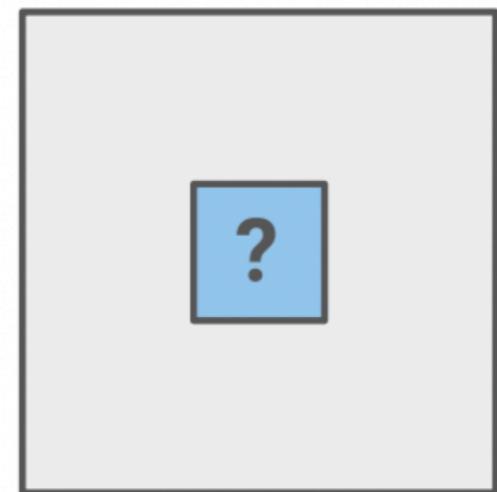
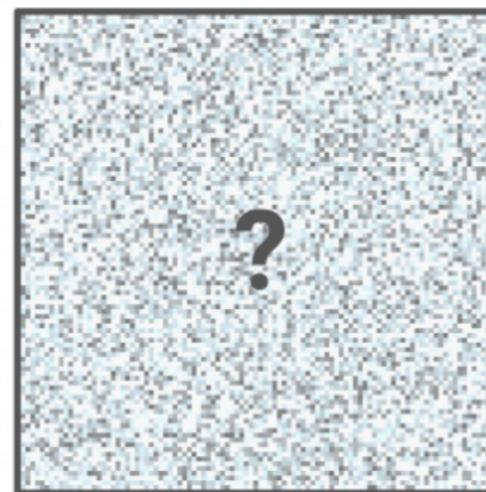
colorization

Representation Learning

Self-prediction:

Given an individual data sample, the task is to predict one part of the sample given the other part.

Pretending that the part to be predicted is missing and use it as the ground-truth:



Intuition: local smoothness of images

Representation Learning

Self-prediction:

Given an individual data sample, the task is to predict one part of the sample given the other part.

Pretending that the part to be predicted is missing and use it as the ground-truth:



Input (context)



reconstruction



adversarial



recon + adv

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{adv}}$$

Representation Learning

Masked language prediction:

Use context/prefix around a word to predict the word, e.g., word2vec, BERT, etc.

Pretending that the part to be predicted is missing and use it as the ground-truth:

Input: The man went to the [MASK]₁ . He bought a [MASK]₂ of milk .

Labels: [MASK]₁ = store; [MASK]₂ = gallon

Missing word prediction task.

Sentence A = The man went to the store.

Sentence B = He bought a gallon of milk.

Label = IsNextSentence

Sentence A = The man went to the store.

Sentence B = Penguins are flightless.

Label = NotNextSentence

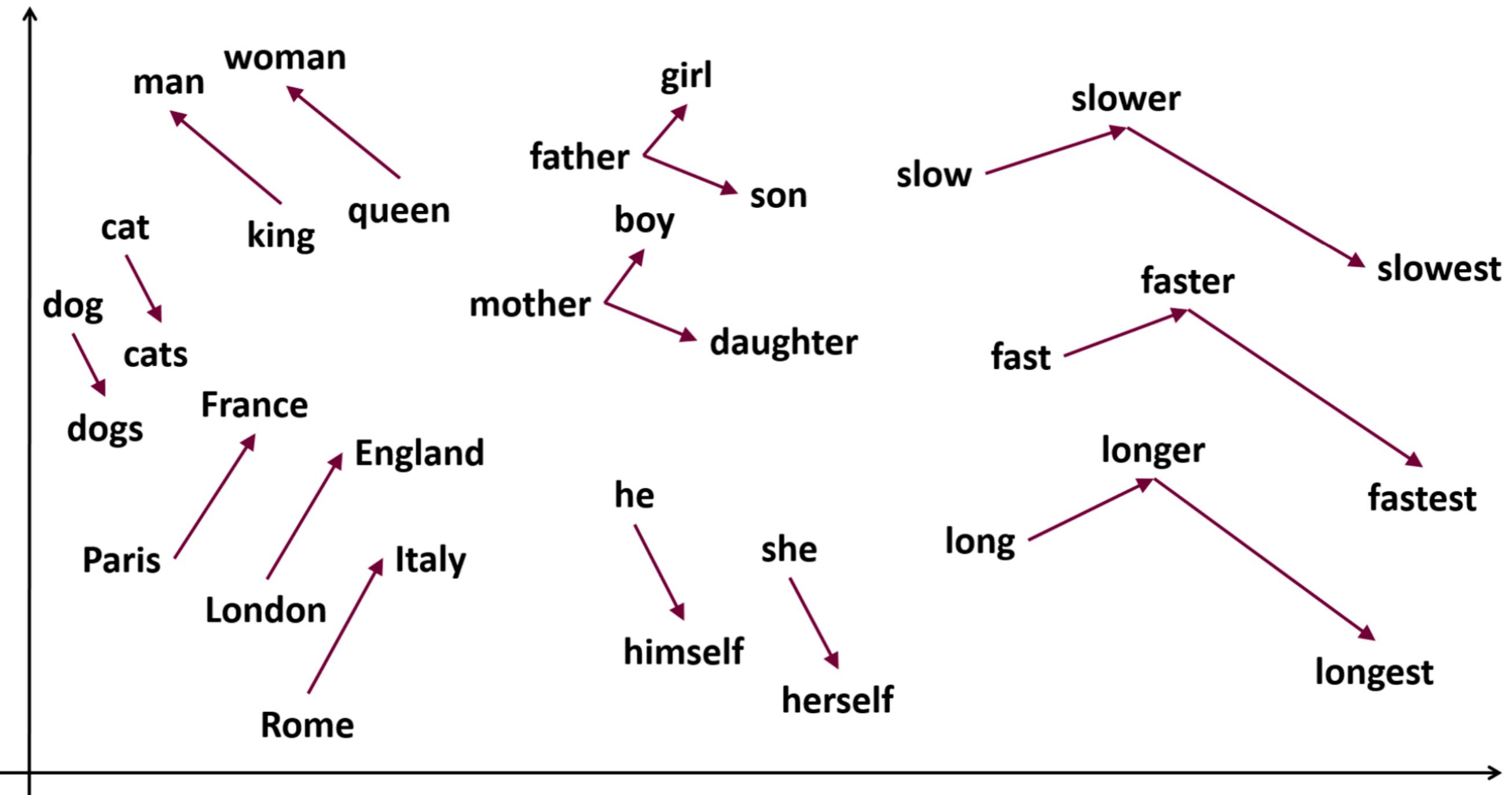
Next sentence prediction task.

Intuition: The meaning of a word can be defined by its context, or the other words and sentences that surround it.

Representation Learning

Masked language prediction:

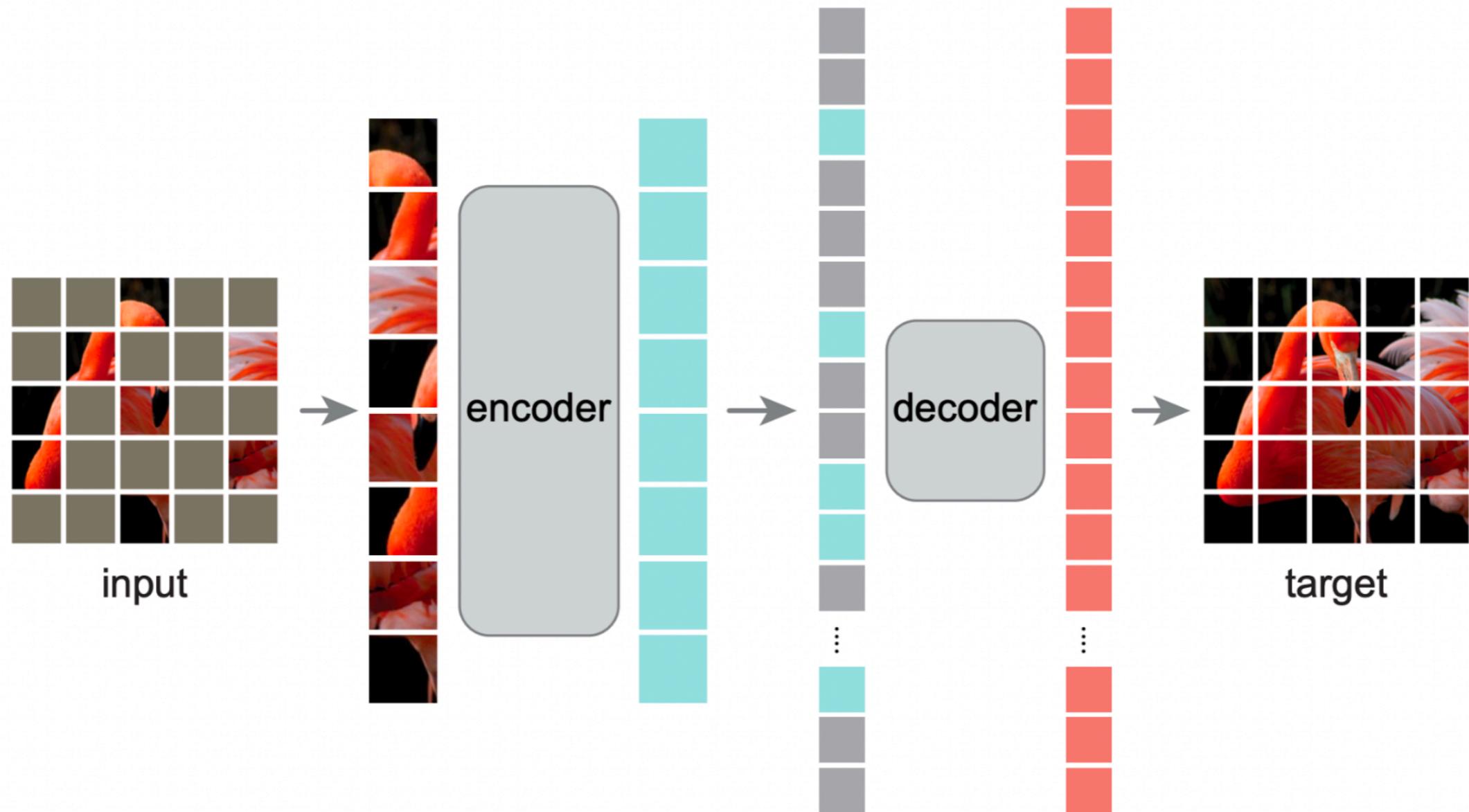
Use context/prefix around a word to predict the word, e.g., word2vec, BERT, etc.



Representation Learning

Masked auto-encoder:

Partition each image into patches and treat each patch as a “word”

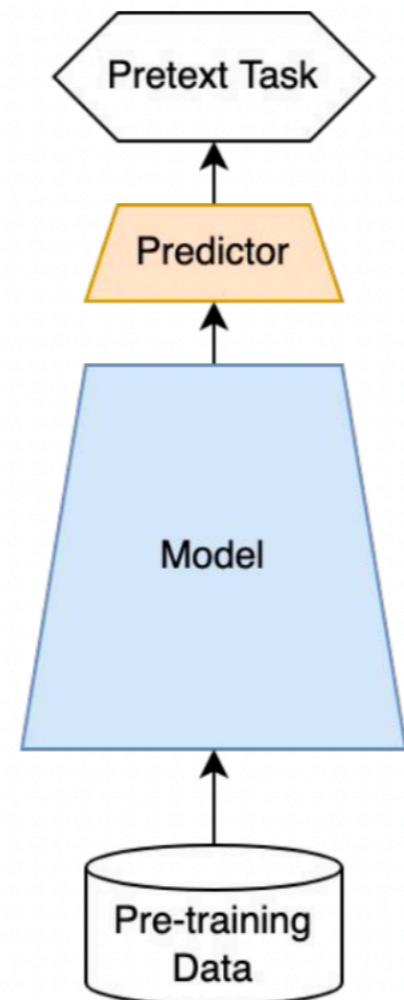


Representation Learning

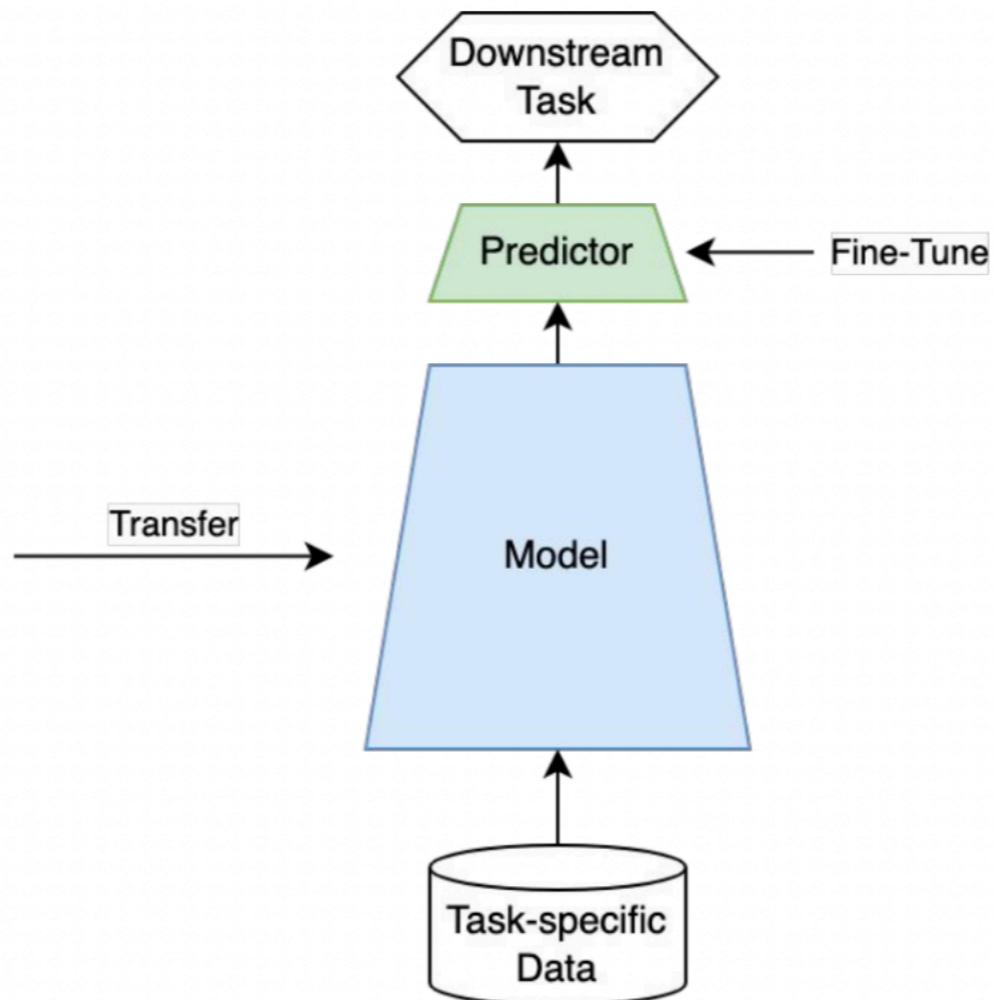
Pretext tasks for self-supervised learning:

- The model is forced to learn good representations to solve the pretext tasks
- The pre-trained features will be fine-tuned for downstream tasks

Step 1: Pre-train a model for a pretext task



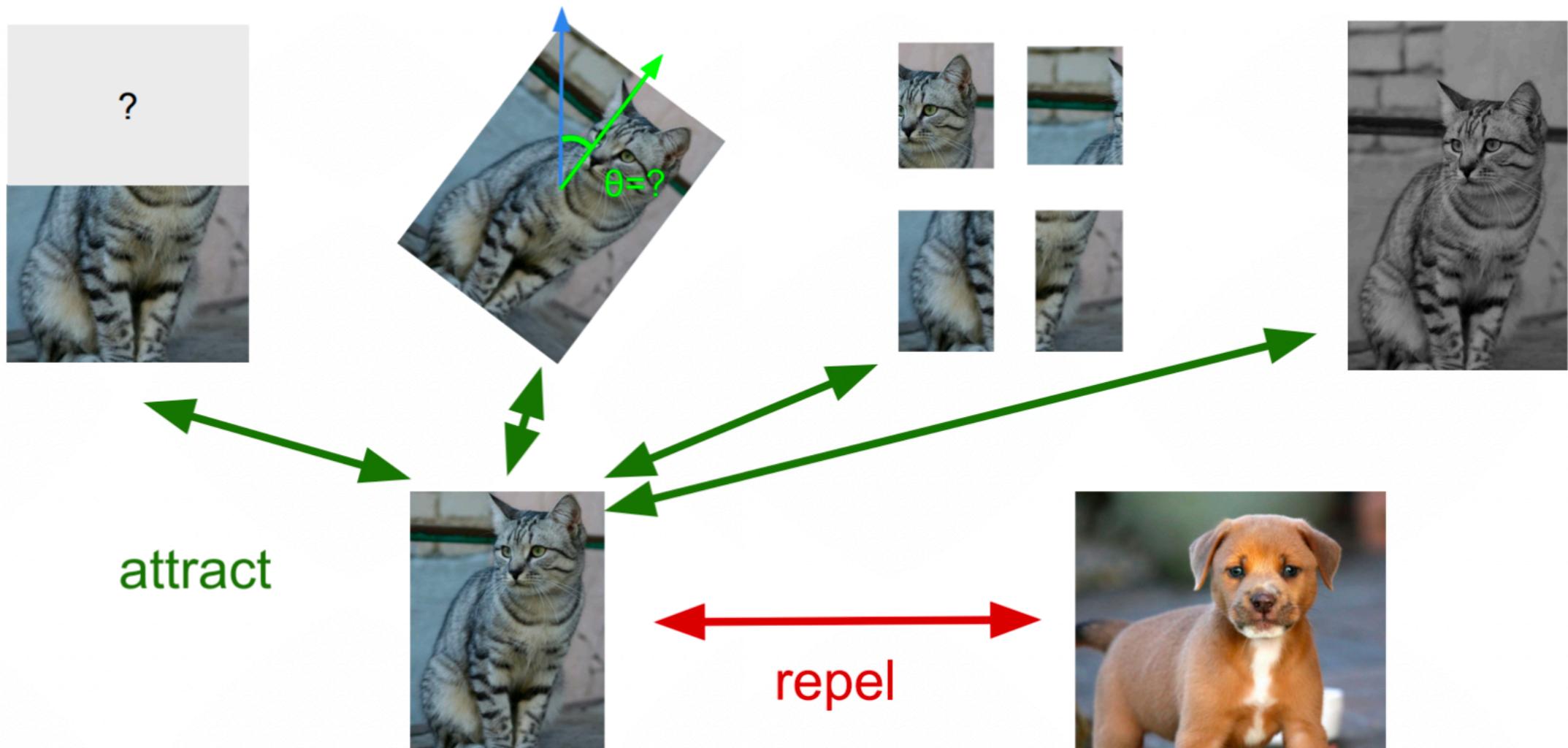
Step 2: Transfer to applications



Representation Learning

Contrastive learning: Invariance principle

Images are defined by its content, which should stay invariant under typical image transformations, including rotation, reflection, translation, etc.



Representation Learning

Contrastive learning: Invariance principle

- Given a reference sample x , we need to have a corresponding positive sample x^+ and a negative sample x^-
- We need to have a score function $s(\cdot, \cdot)$ that measures the similarity between a pair of inputs
- We aim to learn an encoder function $f(\cdot)$ that yields high scores for positive pairs and low scores for negative pairs:

$$s(f(x), f(x^+)) > s(f(x), f(x^-))$$

How could we obtain the positive pairs? What are typical choices of similarity functions?

Representation Learning

Contrastive learning: Invariance principle

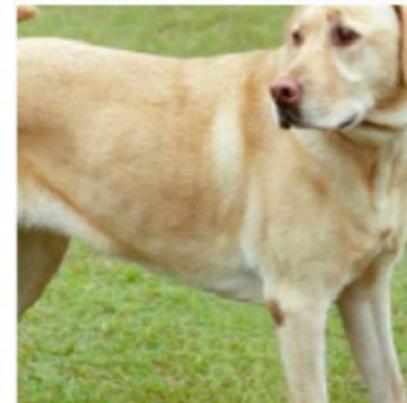
Positive samples are acquired through random data augmentation



(a) Original



(b) Crop and resize



(c) Crop, resize (and flip)



(d) Color distort. (drop)



(e) Color distort. (jitter)



(f) Rotate $\{90^\circ, 180^\circ, 270^\circ\}$



(g) Cutout



(h) Gaussian noise



(i) Gaussian blur



(j) Sobel filtering

Representation Learning

Contrastive learning: Invariance principle

Many loss functions can be used:

Contrastive Loss: minimize the ℓ_2 distance between positive samples and increase the ℓ_2 distance by a margin ϵ for negative samples

$$\mathcal{L}_c := \|f(x) - f(x^+)\|_2^2 + \max \left\{ 0, \epsilon - \|f(x) - f(x^-)\|_2^2 \right\}$$

InfoNCE (Noisy Contrastive Error): learn to find the positive samples

$$\mathcal{L}_{\text{NCE}} := -\mathbb{E} \left[\log \frac{s(f(x), f(x^+))}{s(f(x), f(x^+)) + s(f(x), f(x^-))} \right]$$

In this case the score function is defined as the cosine similarity, i.e.,

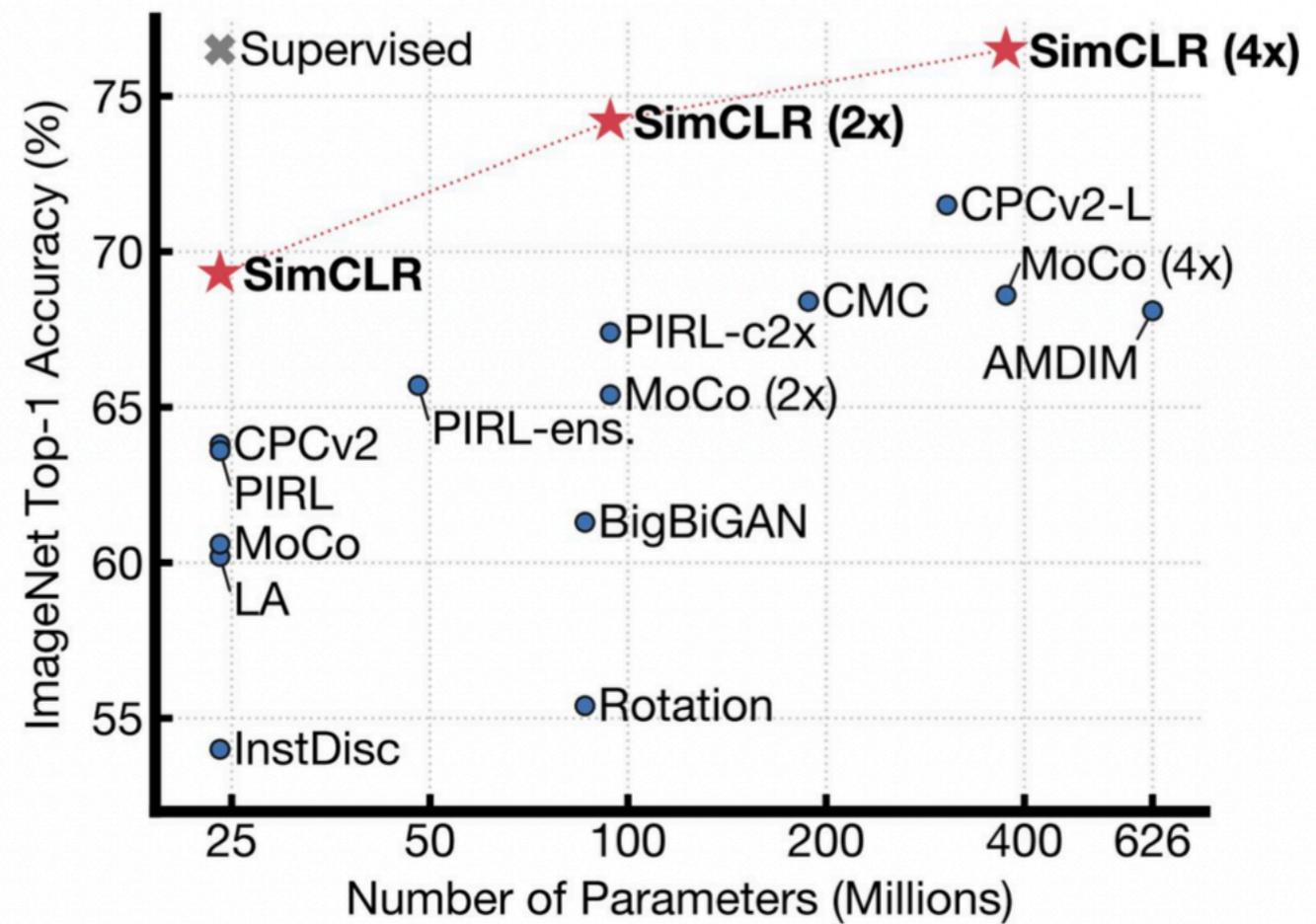
$$s(f(x), f(x')) = \frac{f(x)^\top f(x')}{\|f(x)\|_2 \cdot \|f(x')\|_2}$$

Representation Learning

Contrastive learning: Invariance principle

SimCLR:

- Generate a positive pair by randomly data-augmenting each sample
- Use all the other image samples in the same batch as negative samples
- Mapping images into embedding space with $f(\cdot)$
- Measure the cosine similarity
- Minimize the InfoNCE loss

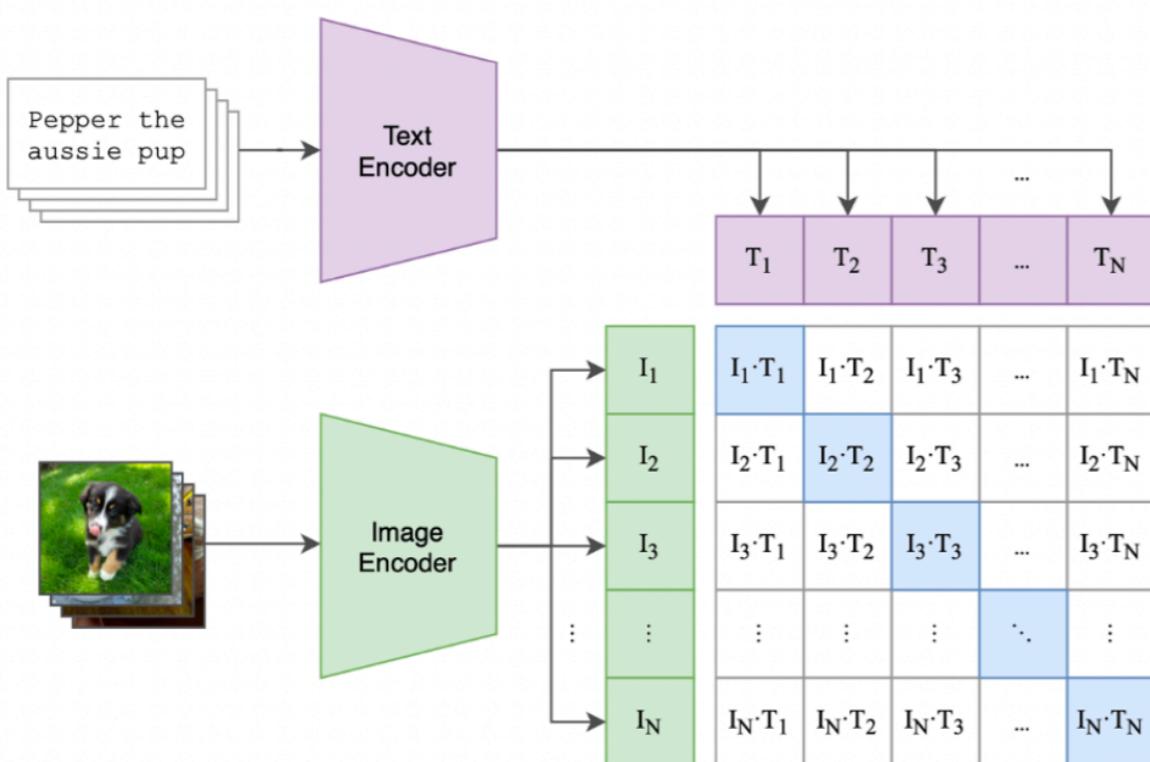


Representation Learning

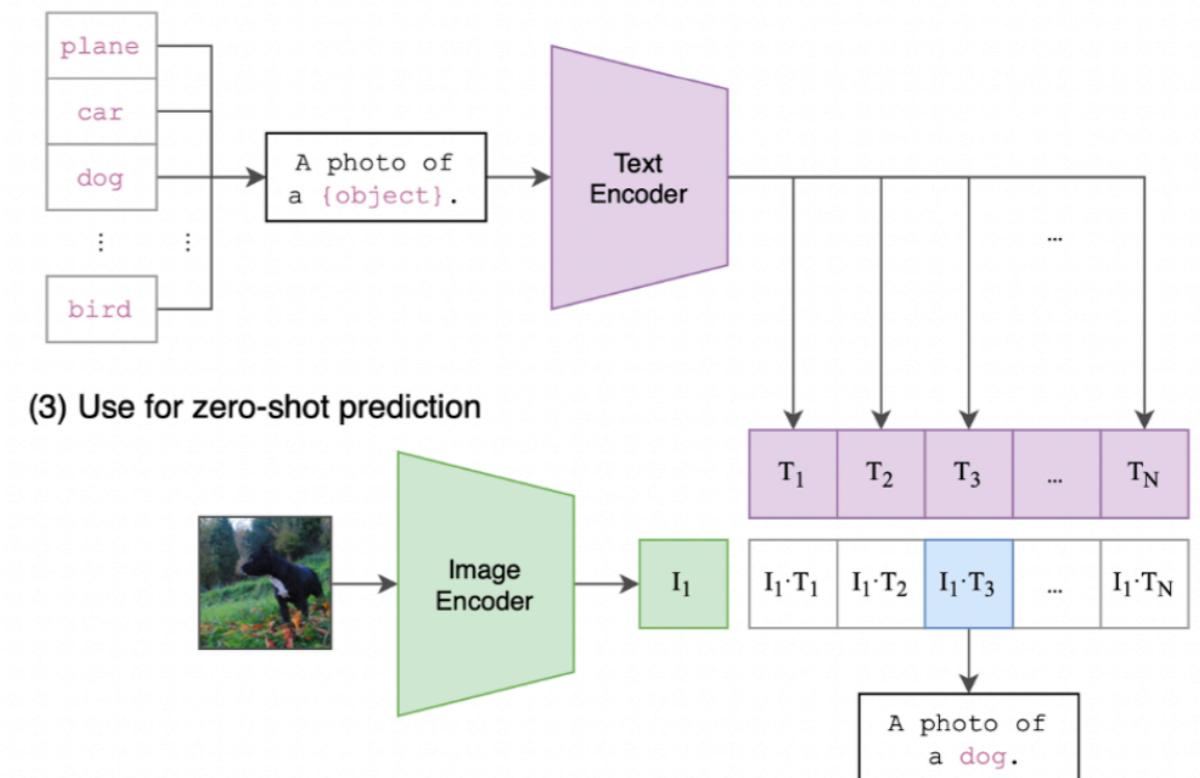
Contrastive Language Image Pre-training (CLIP):

- Pre-training with paired datasets (captions vs images)

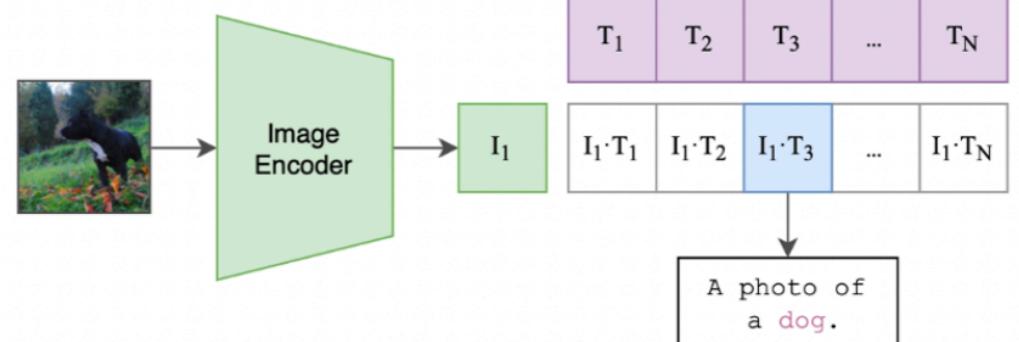
(1) Contrastive pre-training



(2) Create dataset classifier from label text

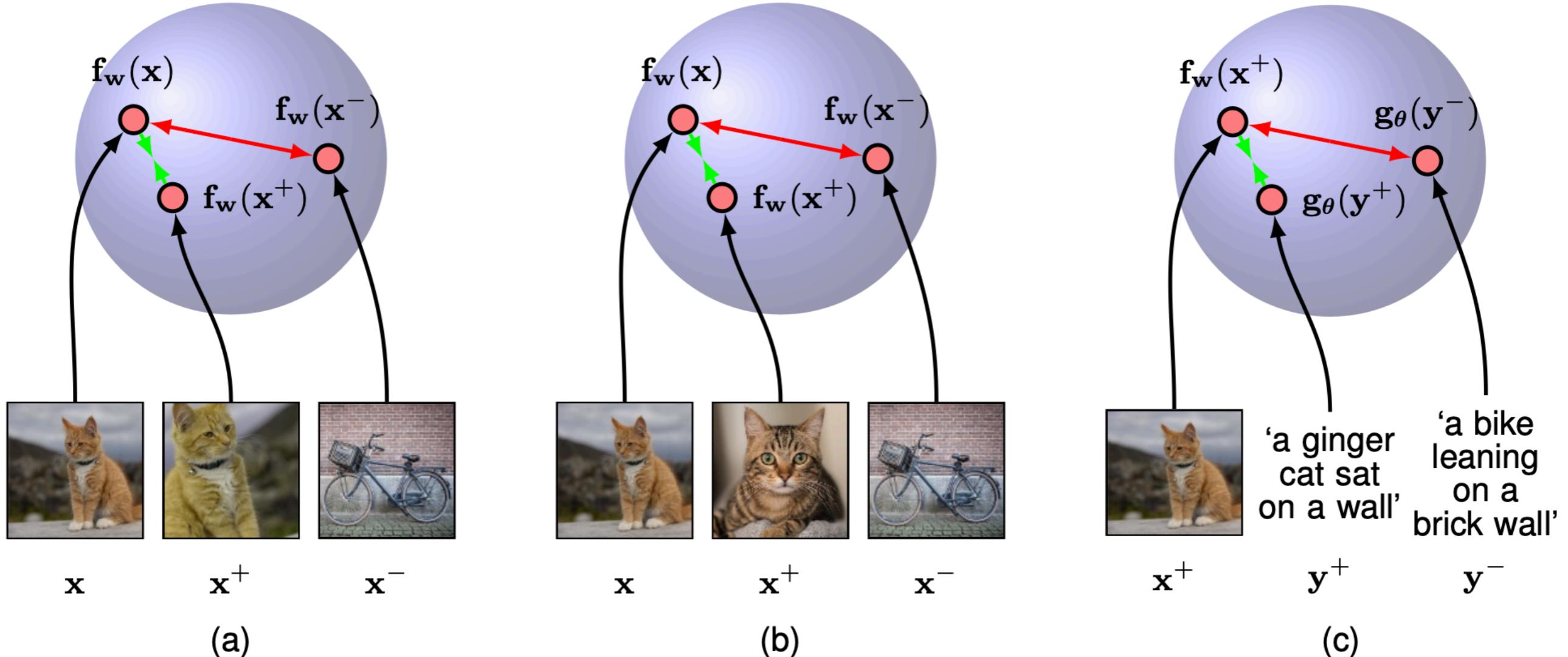


(3) Use for zero-shot prediction



Representation Learning

Other variants of contrastive learning:



(b) Positive pairs through data-augmentation

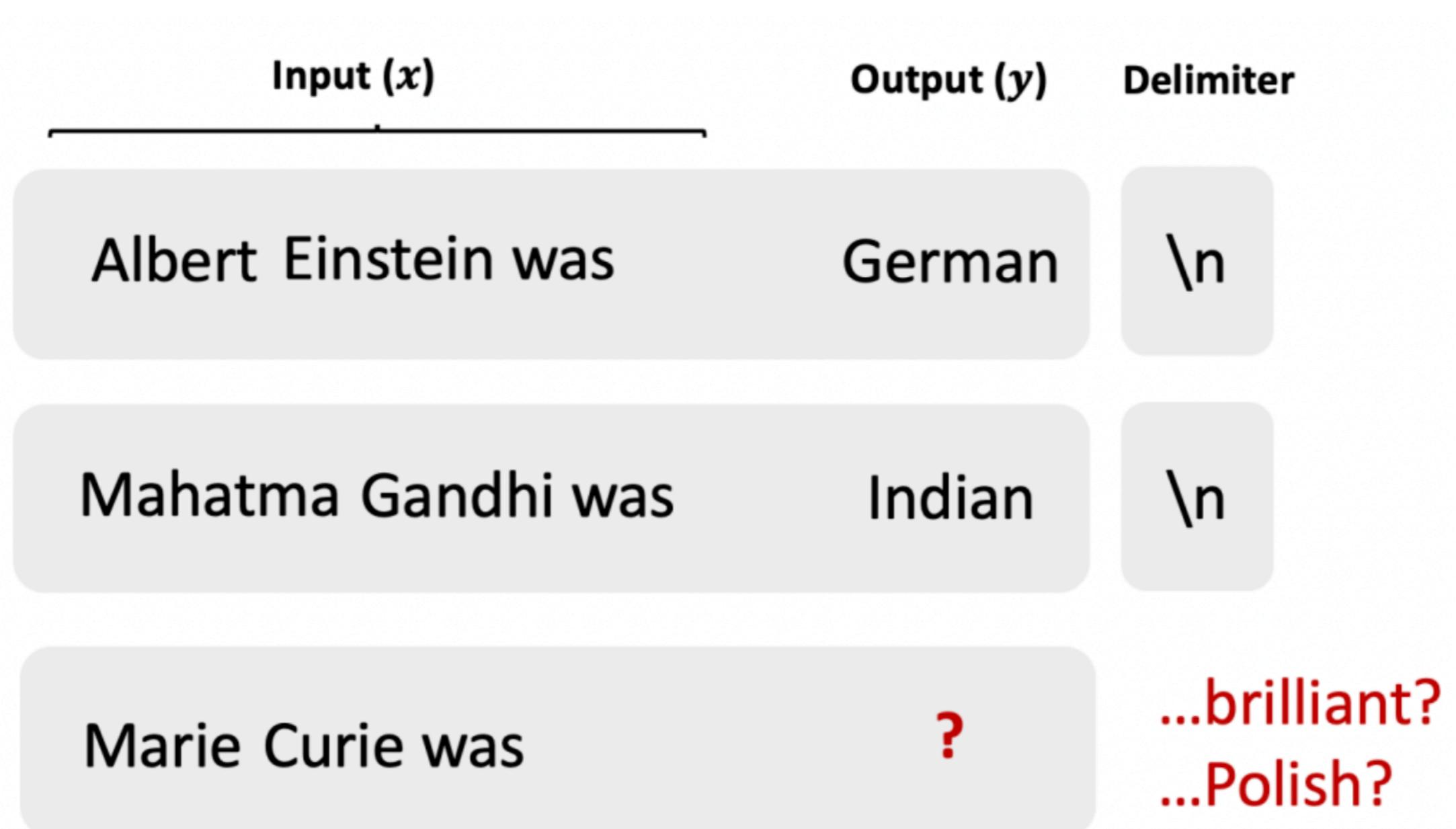
(c) Supervised contrastive learning: positive pairs through class label

(d) Multi-modal contrastive learning: positive pairs through paired image/text

Representation Learning

In-context learning: a paradigm shift of transfer learning:

- Reframing/Reformulating a classification problem into language prediction



Representation Learning

In-context learning: a paradigm shift of transfer learning:

- Reframing/Reformulating a classification problem into language prediction

Translate each sentence into a string of emojis.

English: That cat ate the fish.

Emojis: 🐱😊🐟

English: What is this, a house for ants?

Emojis: 🏠🐜❓

English: The quick brown fox jumps over the lazy dog.

Emojis: 🌬🦊🏃😴🐶

English: One small step for man, one giant leap for mankind.

Emojis: 🚀🌌earth🏃

Representation Learning

Large-language Models



find the bug in the following code piece



```
a = 1
while a==1:
    b = input("what's your name?")
    print("Hi", b, ", welcome to an amazing loop")
```



If the intention was for the code to ask the user for their name and then print a greeting only once, then there is a bug in the code. The `while` loop condition is checking if `a` is equal to 1, but `a` is initialized to 1 and never changed inside the loop, so the loop will continue indefinitely and the code will not stop running.



To fix this bug, the initial value of `a` should be changed to 0, and a condition should be added inside the `while` loop to break out of the loop after the first iteration. For example:

```
a = 0
while a==1:
    b = input("what's your name?")
    print("Hi", b, ", welcome to an amazing loop")
    a = 1
```

Copy code

This change will fix the bug

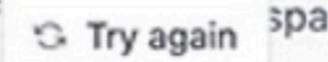
Representation Learning

Large-language Models

 what are wormholes. Explain like I am 5 

 A wormhole is like a tunnel that connects two different places in space. Imagine that you are standing on one end of the tunnel and you want to get to the other end. Normally, you would have to walk through the tunnel, but a wormhole is like a shortcut that allows you to get to the other end without having to walk all the way.  

Wormholes are like shortcuts because they can bend space and time. This means that they can make the distance between two places in space shorter, and they can also make the time it takes to travel between those places faster. Imagine that you are in a spaceship and you want to travel to a planet that is very far away. Normally, it would take a long time to get there, but if you find a wormhole, you can use it to take a shortcut and get there faster.

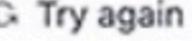
Wormholes are still just a theory, and we don't know for sure if they actually exist. Some scientists think that they might, but we don't have any proof yet. Even if they do exist, we don't know how to find them or how to use them. But if we ever do figure out how to use wormholes, they could be very helpful  space.

Representation Learning

Large-language Models

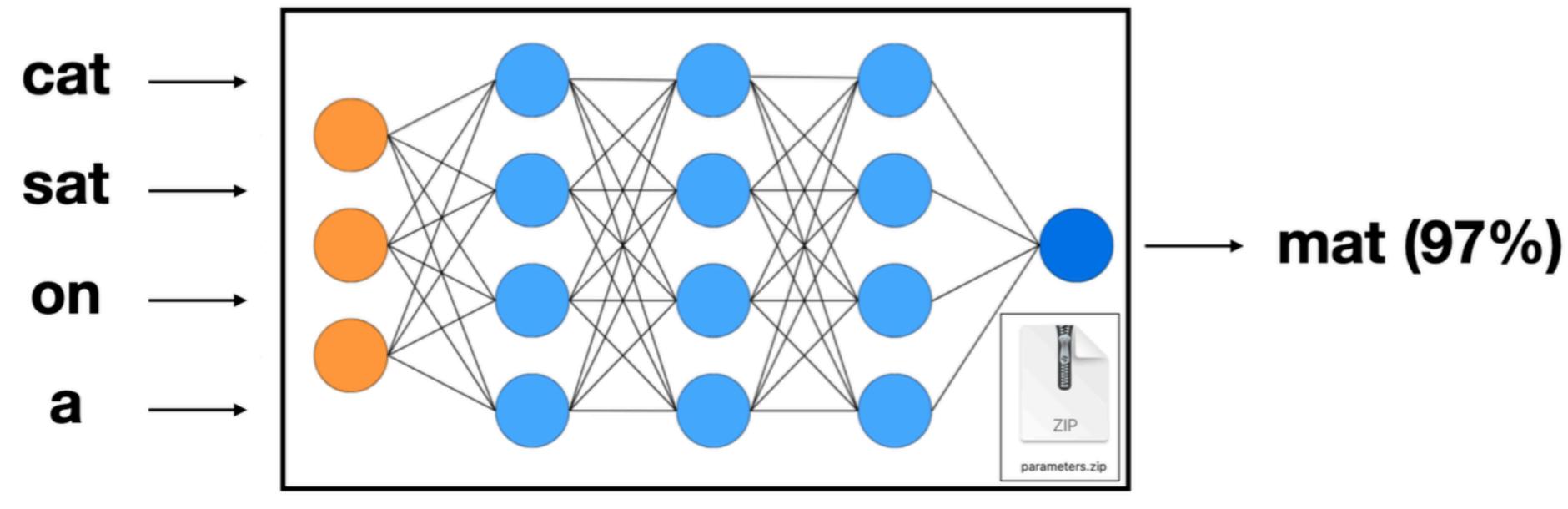
 give me party ideas for the birthday party of a 29 year old 

 Here are a few party ideas for the birthday party of a 29 year old:  

1. A cocktail party: This can be a fun and sophisticated way to celebrate the 29th birthday. You can invite friends to come over and enjoy some delicious cocktails and appetizers. You can even have a cocktail-making contest and see who can come up with the most creative drink.
2. A game night: If the birthday person and their friends are competitive, a game night can be a great way to celebrate. You can set up a variety of games, such as board games, card games, and video games, and let everyone compete for prizes. This can be a fun and energetic way to spend the evening.
3. A movie night: If the birthday person is a movie buff, a movie night can be a great way to celebrate. You can set up a big screen and a projection system, and invite friends to come over and watch some of the birthday person's favorite movies. You can even have a themed movie night, where everyone has to dress up as a character from one of the movies.
4. A spa day: If the birthday person is in need of some relaxation, a spa day can be a great way to celebrate. You can invite friend  and enjoy some pampering, such as

Representation Learning

Large-language Models: next token prediction



e.g. context of 4 words

predict next word

$$\log p(x_T | x_1, \dots, x_{T-1}; \theta)$$

Representation Learning

Large-language Models: next token prediction

Stage 1: random masking

The **University of Illinois Urbana-Champaign** (U of I, Illinois, University of Illinois, or UIUC)^{[12][13]} is a public land-grant research university in [REDACTED], and Urbana, Illinois. It is the flagship institution of the University of Illinois system and was founded in [REDACTED]. With over 56,000 students, the University of Illinois is one of the largest [REDACTED] in the United States.

The university contains [REDACTED]^[14] and offers more than 150 undergraduate and over 100 graduate programs of study. The university holds 651 buildings on 6,370 acres (2,578 ha)^[9] and its annual operating budget in 2016 was over \$2 billion.^[15] The University of Illinois Urbana-Champaign also operates a Research Park home to innovation centers for over 90 start-up companies and multinational corporations.^[16]

Representation Learning

Large-language Models: next token prediction

Stage 1: random masking

GPT-3.5 & GPT-4 GPT-3 (Legacy)

The University of Illinois Urbana-Champaign (U of I, Illinois, University of Illinois, or UIUC)[12][13] is a public land-grant research university in Champaign, Illinois, and Urbana, Illinois. It is the flagship institution of the University of Illinois system and was founded in 1867. With over 56,000 students, the University of Illinois is one of the largest public universities by enrollment in the United States.



Clear

Show example

Tokens	Characters
92	419

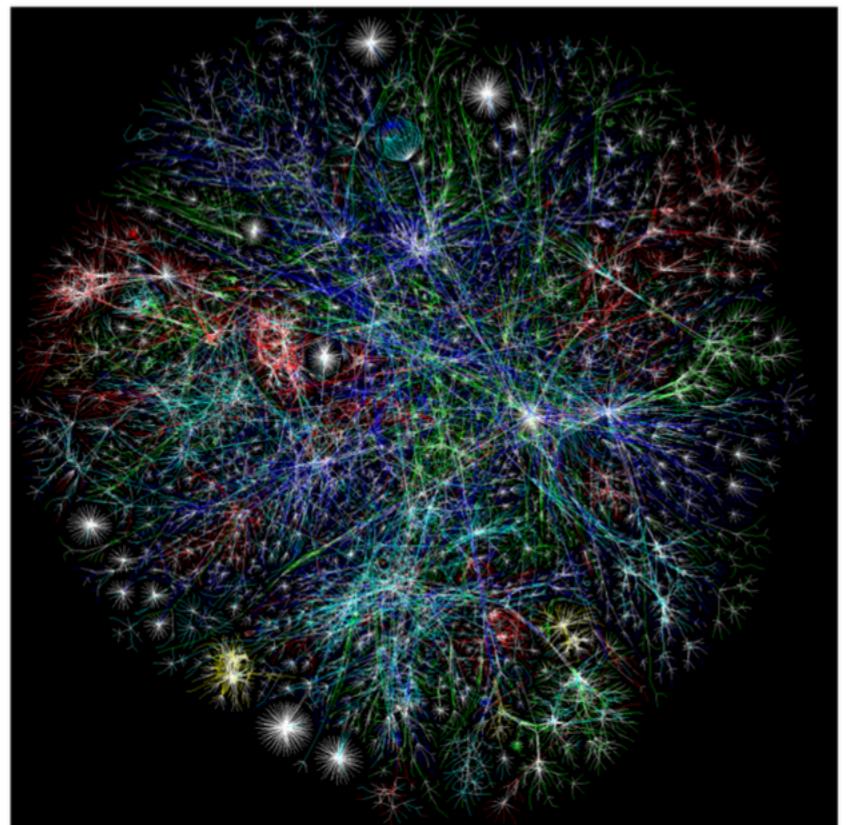
The University of Illinois Urbana-Champaign (U of I, Illinois, University of Illinois, or UIUC)[12][13] is a public land-grant research university in Champaign, Illinois, and Urbana, Illinois. It is the flagship institution of the University of Illinois system and was founded in 1867. With over 56,000 students, the University of Illinois is one of the largest public universities by enrollment in the United States.

TEXT TOKEN IDS

Representation Learning

Large-language Models: next token prediction

Stage 2: fine-tuning



<USER>

Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.

<ASSISTANT>

"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions...

~100K conversations

Representation Learning

Large-language Models: next token prediction

Stage 3: alignment — reinforcement learning from human-feedback

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

Explain the moon landing to a 6 year old



Some people went to the moon...



This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

Explain the moon landing to a 6 year old



D > C > A = B

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



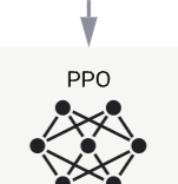
D > C > A = B

Step 3

Optimize a policy against the reward model using reinforcement learning.

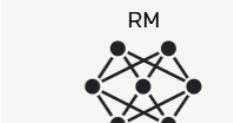
A new prompt is sampled from the dataset.

Write a story about frogs



The policy generates an output.

Once upon a time...



The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

r_k

Next

Language modeling / foundation models

- GPT