Deep Learning 1

2025-2026 – Pascal Mettes

Lecture 8

From supervised to unsupervised deep learning

Previous lecture

Lecture	Title
1	Intro and history of deep learning
3	Deep learning optimization I
5	Convolutional deep learning
7	Graph deep learning
9	Multi-modal deep learning
11	What doesn't work in deep learning
13	Q&A

Lecture	Title
2	AutoDiff
4	Deep learning optimization II
6	Attention-based deep learning
8	From supervised to unsupervised deep learning
10	Generative deep learning
12	Non-Euclidean deep learning
14	Deep learning for videos

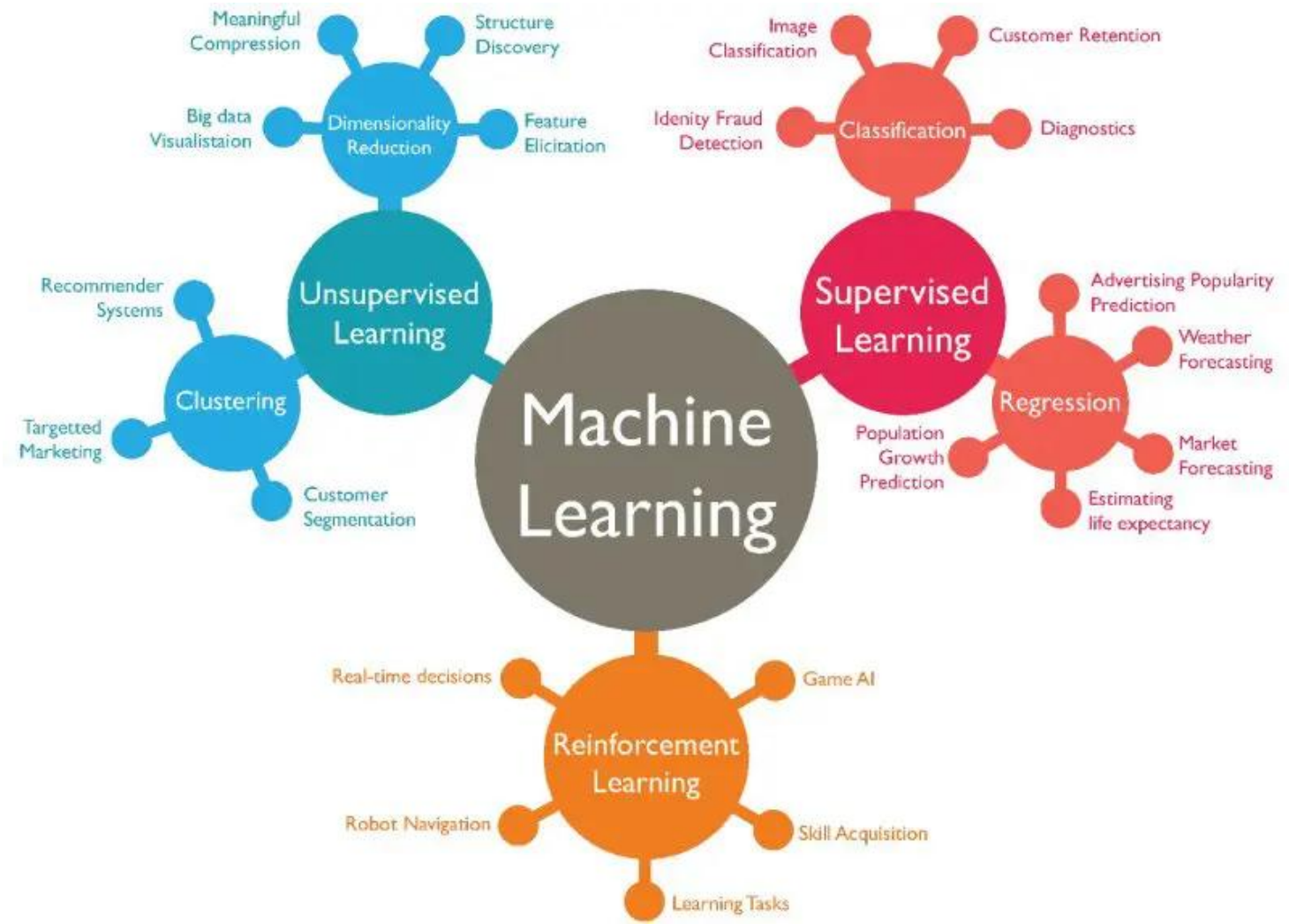
This lecture

Self-supervised learning for vision

Self-supervised learning for language

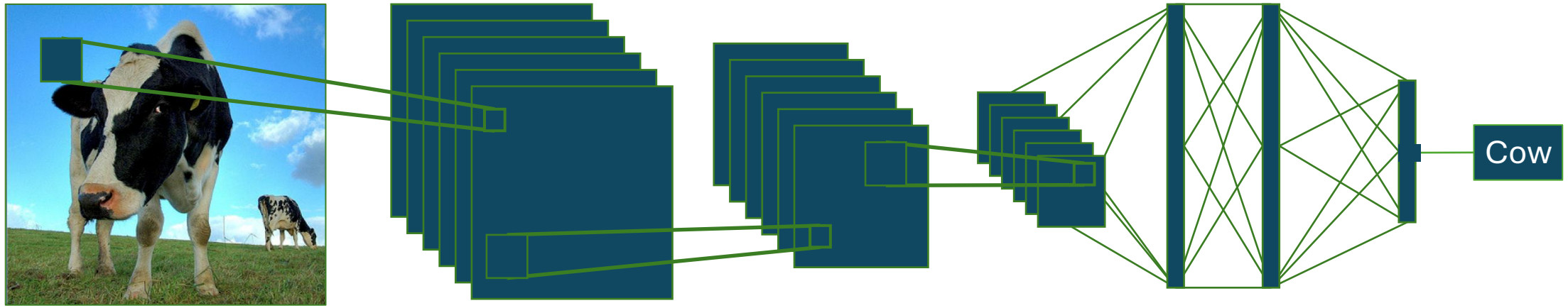
In between supervised and self-supervised learning

Traditional pillars of machine learning



Strength and weakness of supervision in DL

Supervision makes it possible to propagate signals back to train networks.



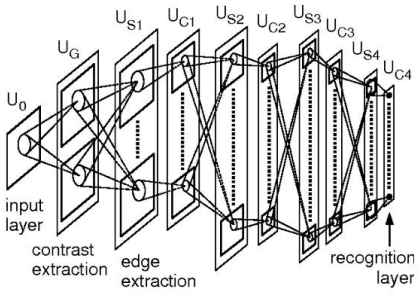
Classification labels no longer the backbone of latest models, why?

Self-supervised learning

Data as fuel for deep learning

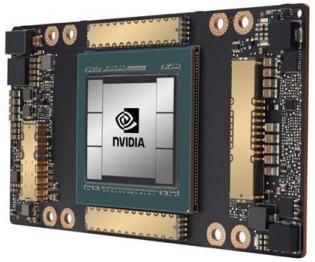
Algorithms

Deep neural networks



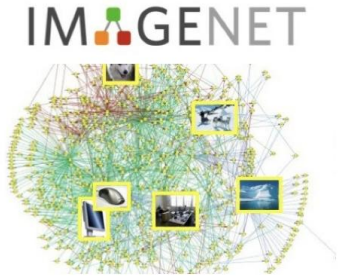
Hardware

GPUs



Data

Large scale datasets

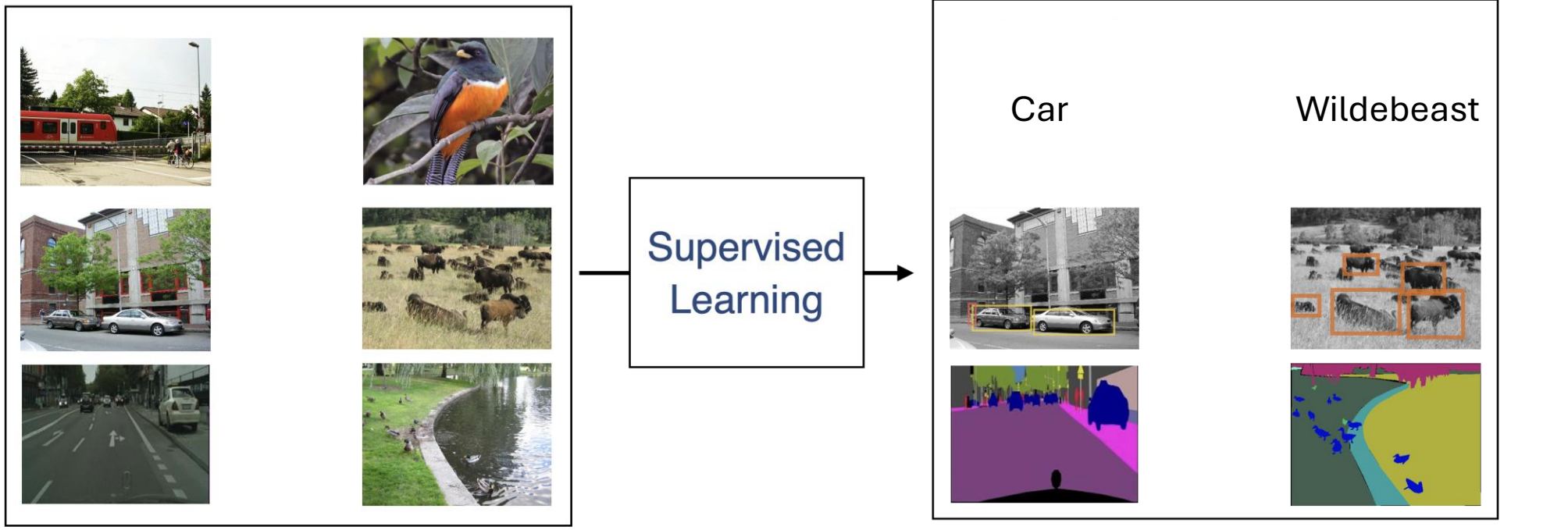


The bottleneck of data

often data are very cheap

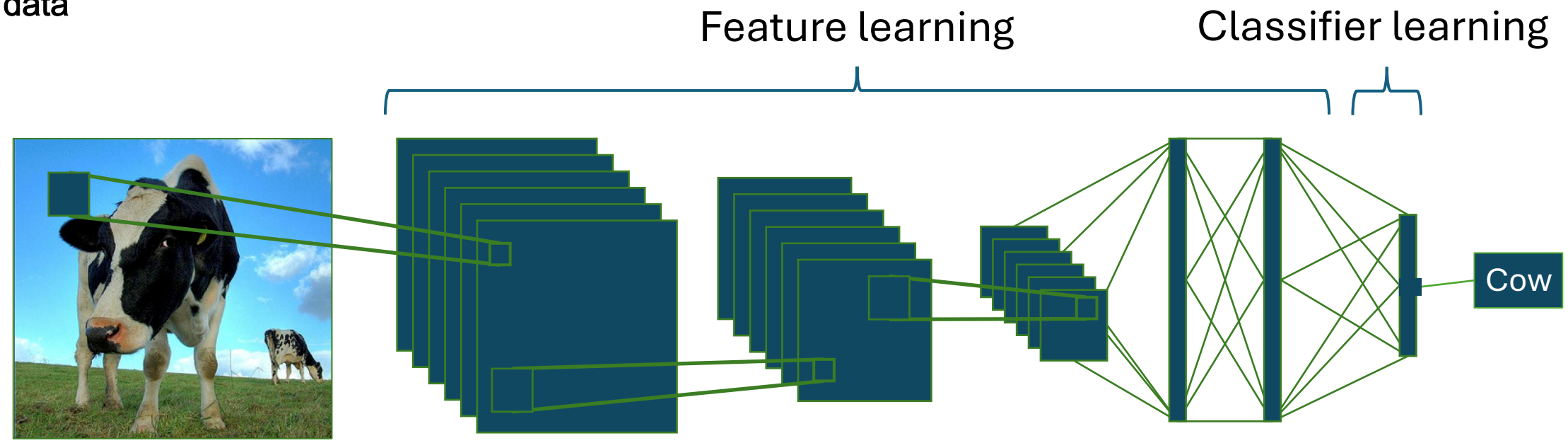
But manual annotations are expensive:

e.g. 30min per image / requiring experts



The two stages of deep learning

all the layers before may not need the labels, they would benefit more from more data



The final layer requires labels, but is that also true for all other layers?

Solving the problem of expensive annotations: self

The idea is that we are going to impose a certain structure on the data, which the network need to learn

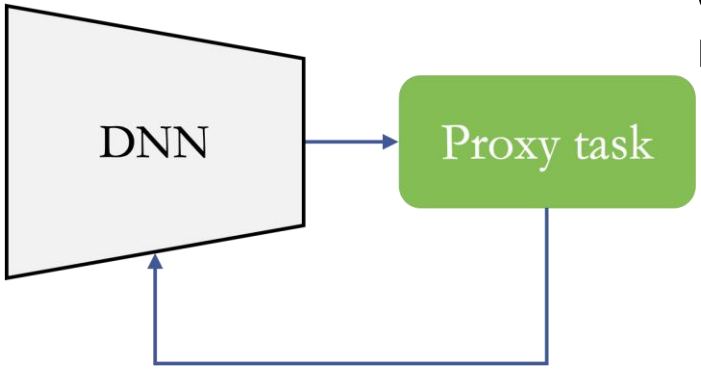
Self-supervision

Extract a supervisory signal from the raw data

Main idea of self-supervised learning

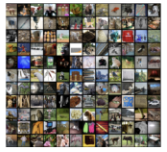
Phase 1: Pretraining

Unlabelled data
+ transformations

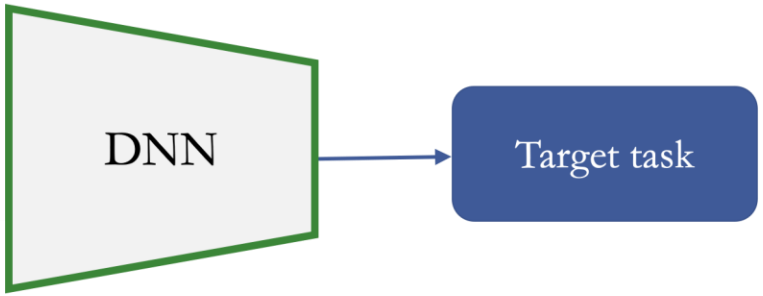


we're gonna define a custom task over the data to perform the training without really need all the labels

Phase 2: Downstream tasks



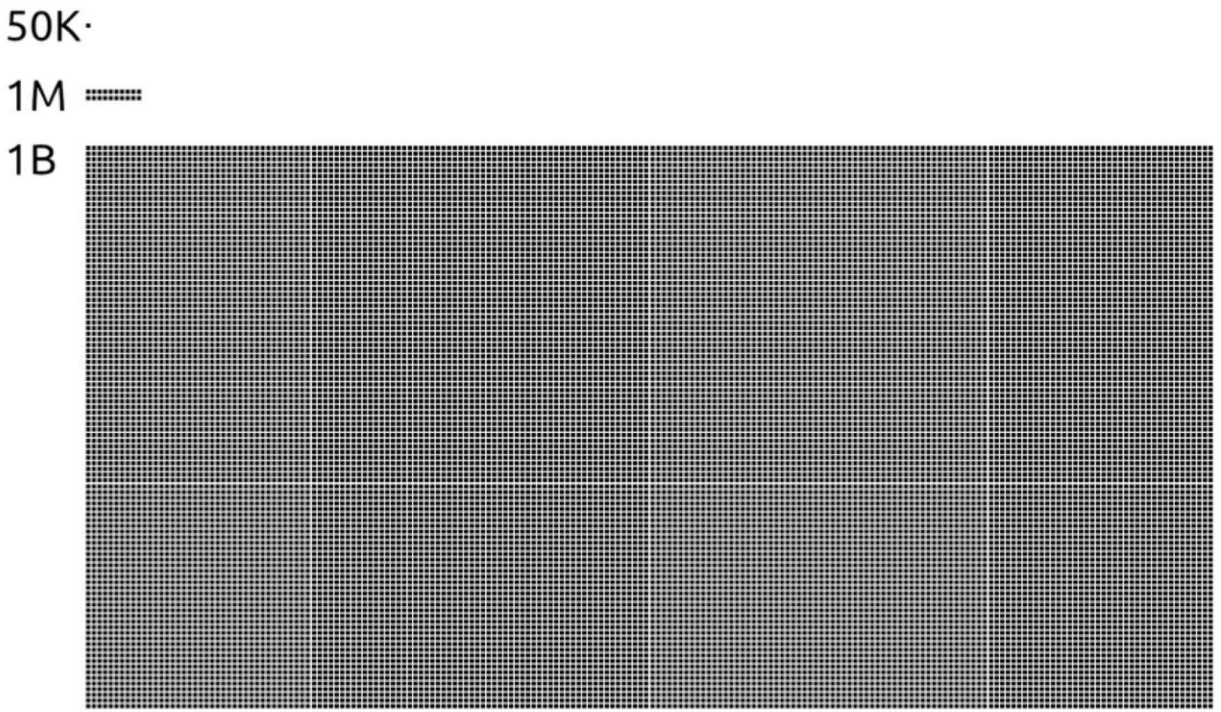
(Sparse) labeled data



Why do we want self-supervised learning?

Reason 1: Scalability

ImageNet
~1 million annotated images

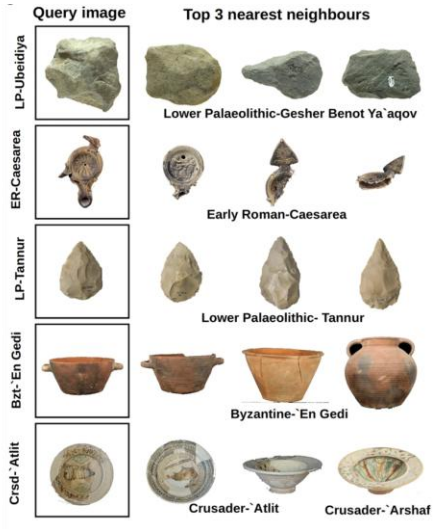


Instagram
~50 billion images floating about

The web is filled with unanontated data.

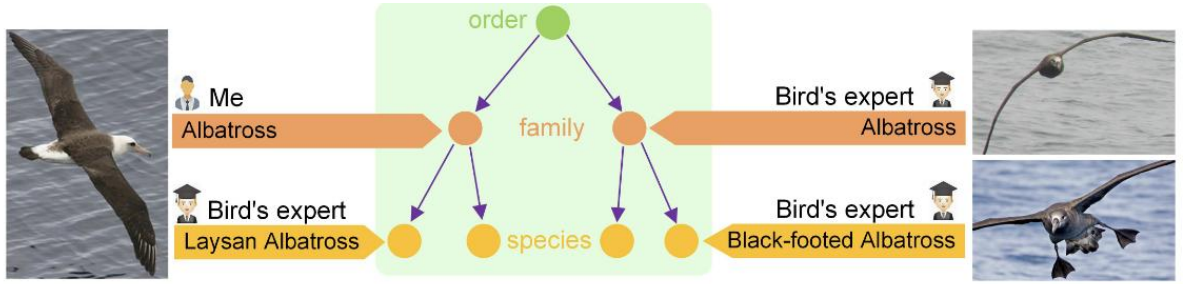
Reason 2: Generalizability

things change over time, but a label is static, so we eventually need to relabel

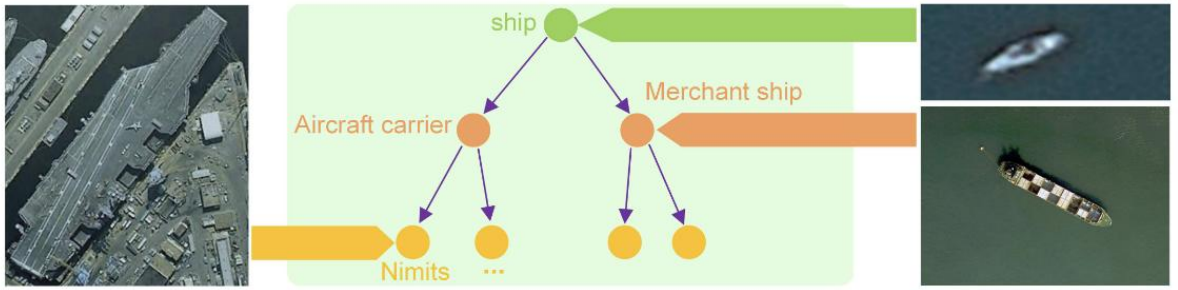


We want models that generalize to many domains and shifts.

Reason 3: Label are not perfect



(a) Differences in domain knowledge and interference from the image occlusion.



(b) Large variations of image resolutions.

Chen et al. 2022

Labels can be ambiguous, biased, or simply wrong.

Reason 4: Humans are self-supervised

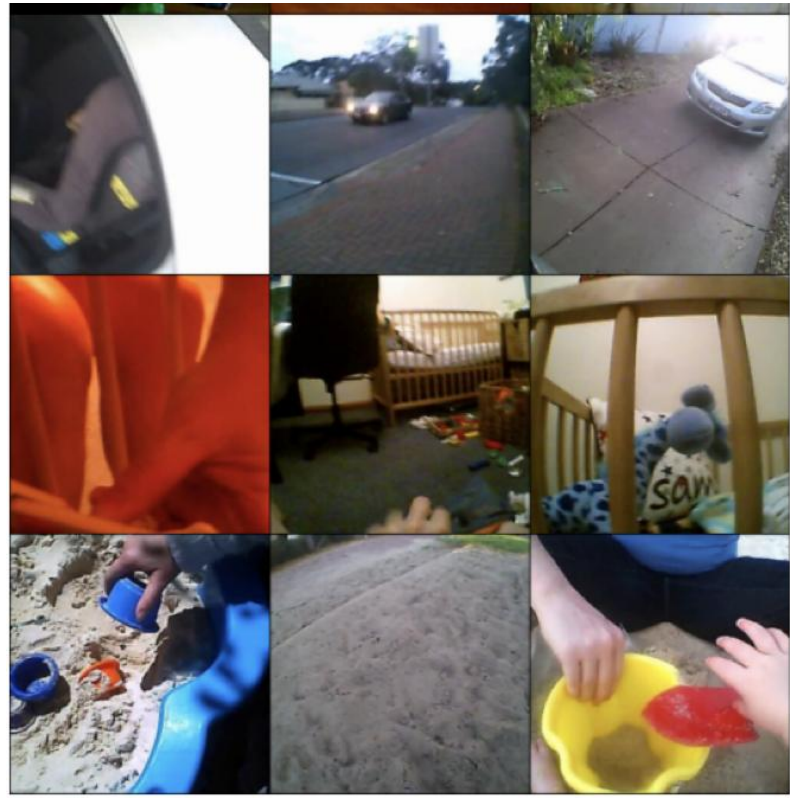
Meta AI Research Publications P

RESEARCH

Self-supervised learning: The dark matter of intelligence

March 4, 2021

As babies, we learn how the world works largely by observation. We form generalized predictive models about objects in the world by learning concepts such as object permanence and gravity. Later in life, we observe the world, act on it, observe again, and build hypotheses to explain how our actions change our environment by trial and error.



Still a lot of lessons from human learning that can be transferred.

How do we train deep networks without labels?

Self-supervied visual learning

The first popular domain for self-supervised learning.

Main idea: exploit the structure of images and videos to learn without labels.

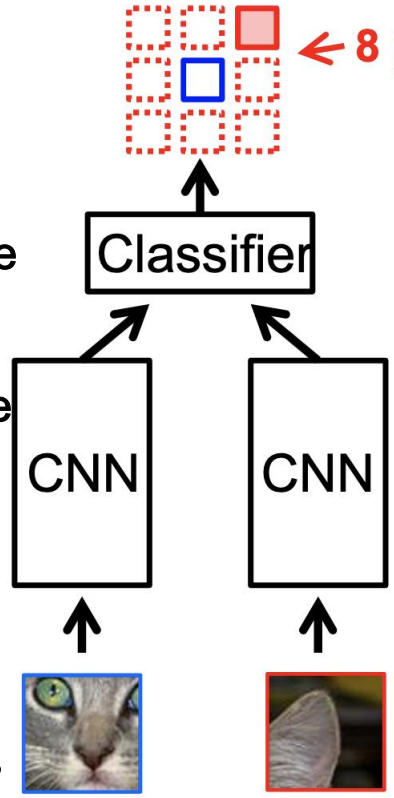
Goal is not to develop new algorithms, but borrow losses from supervised learning and think of your own loss functions.

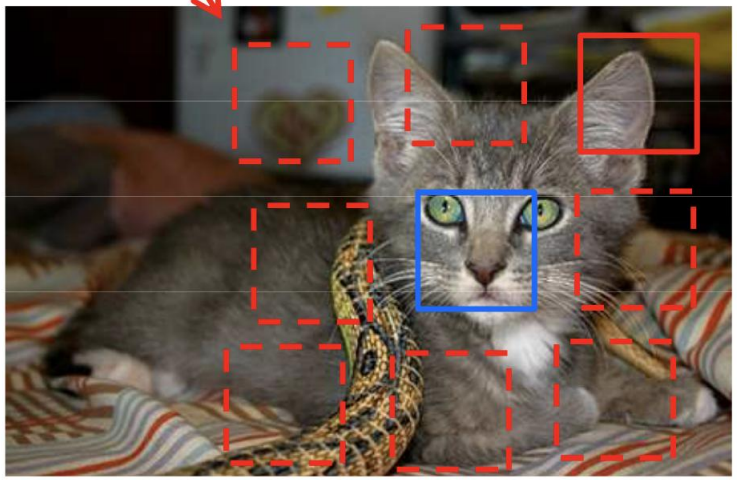
Early attempt: relative positioning

Randomly select one patch from an image, then sample a second patch from one of its 8 neighboring positions

Classification task: The network receives both patches as input and must predict which of the 8 possible spatial locations the second patch came from relative to the first

Automatic labels: Since you extracted the patches yourself, you know their relative positions - this becomes your free supervisory signal for backpropagation

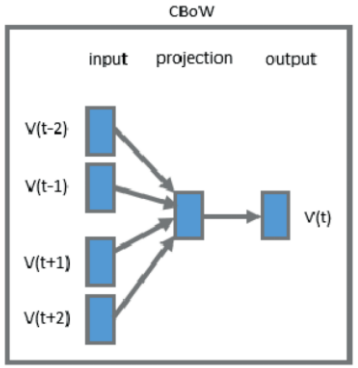




Randomly Sample Patch
Sample Second Patch

Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Word2Vec

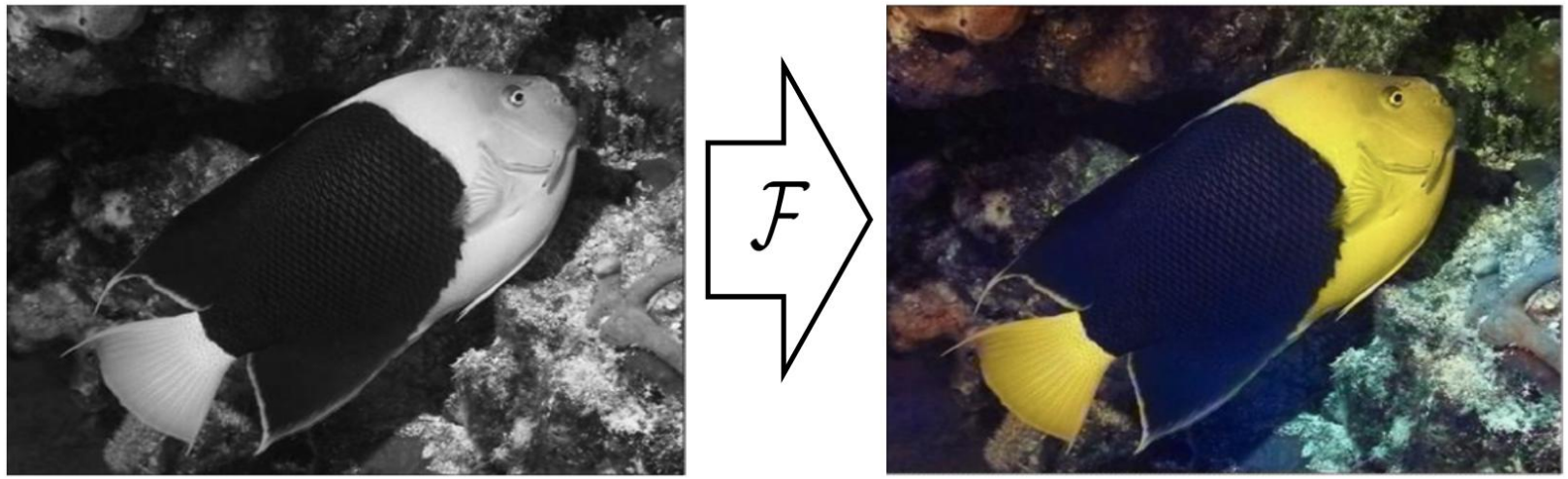


Motivated from NLP

[Zhang et al. ECCV 2016]

Learning by coloring

remove the color and learn to add them back



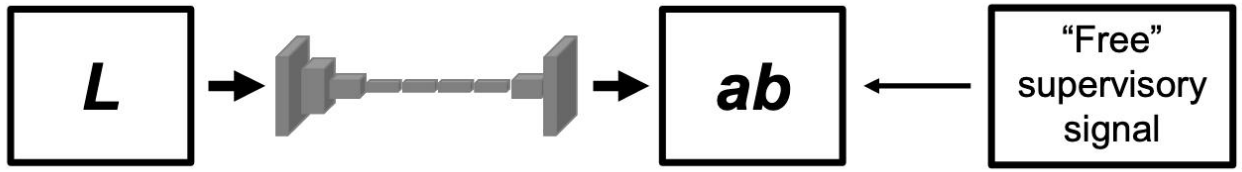
Grayscale image: *L* channel

$$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$$

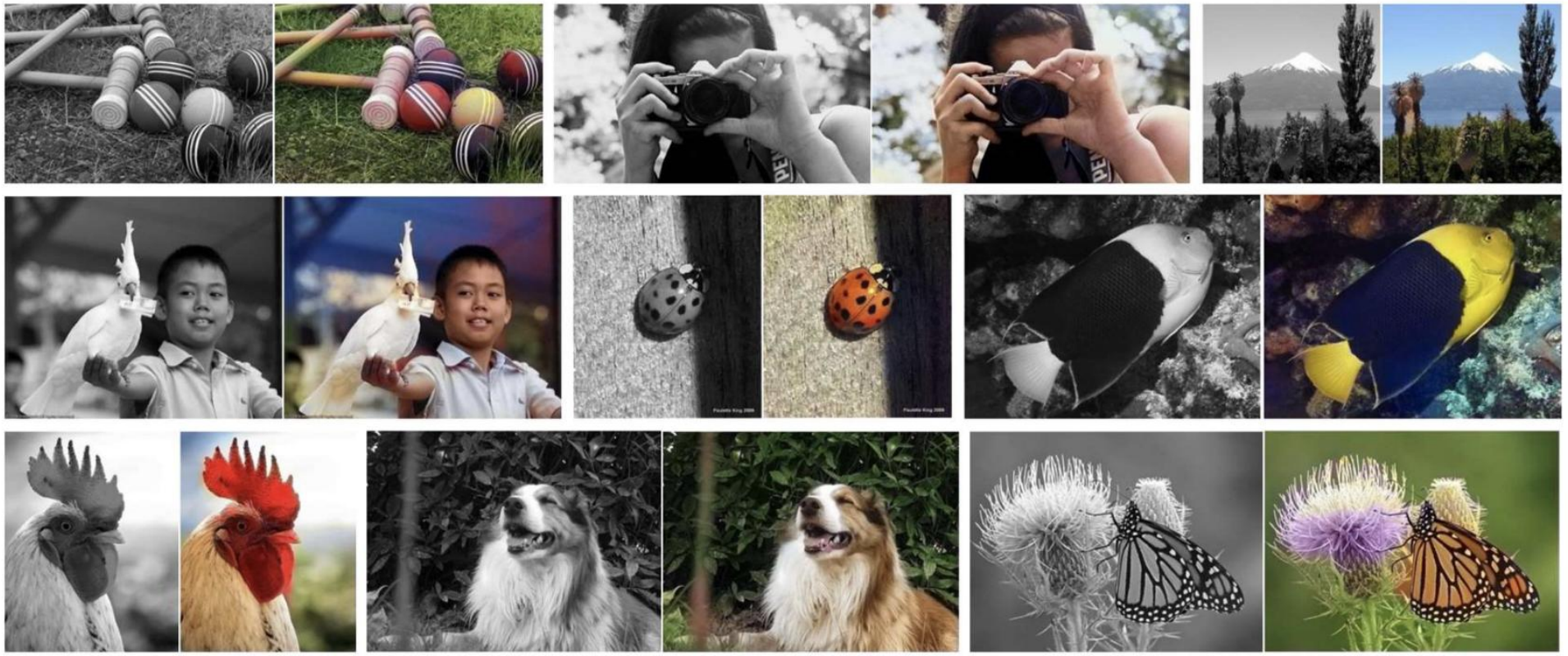
Concatenate (L,ab)

$$(\mathbf{X}, \hat{\mathbf{Y}})$$

so basically in self supervised learning from a modified input we want to learn to reconstruct the original input



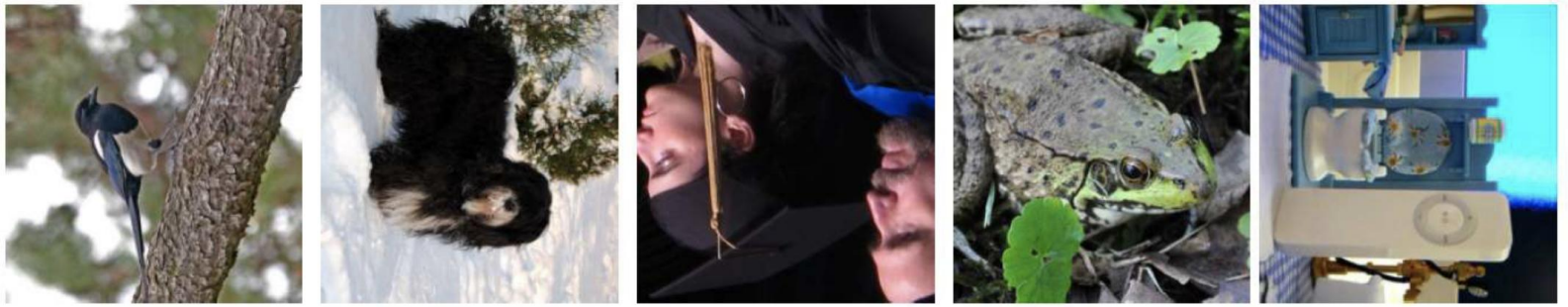
Visual results



The representations for learning color can be used to initialize a new network.

[Gidaris et al. ICLR 2018]

Learning by rotations



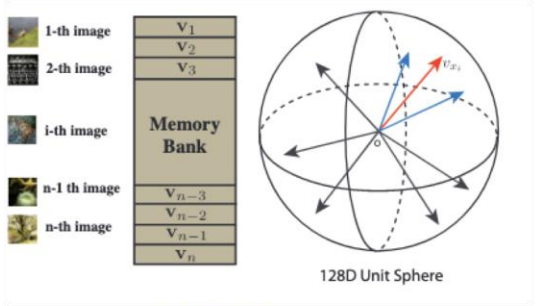
Assumption: if we know the object, we understand which rotation is most natural.

Modern approach: contrastive self-supervision

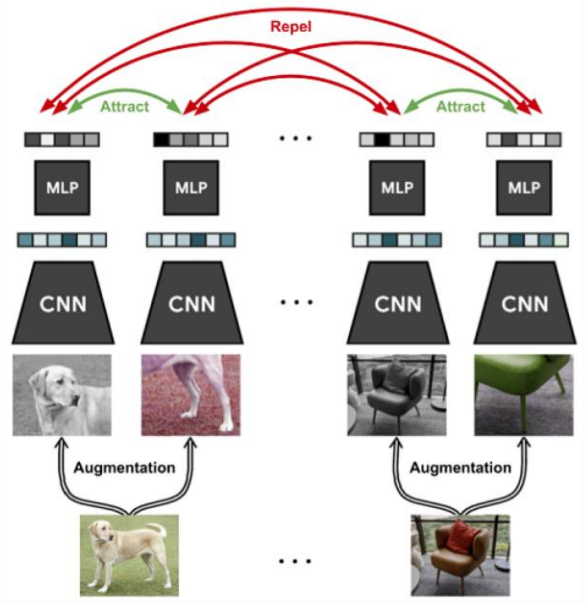
I'm gonna take an image from a batch and apply to it a set of different transformation, i want group my similar images together and respingere the others

contrastive learning, it want alignments and uniformity

the clustering wants groups as tight as possible (alignment) and as far as possible from each other (uniformity)



NPID



SimCLR

The contrastive loss for positive pairs i,j:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} {}_{[k\neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

with z_i, z_j embeddings for images i and j,

τ a temperature, sim() *is the dot-product*

"non-parametric" softmax

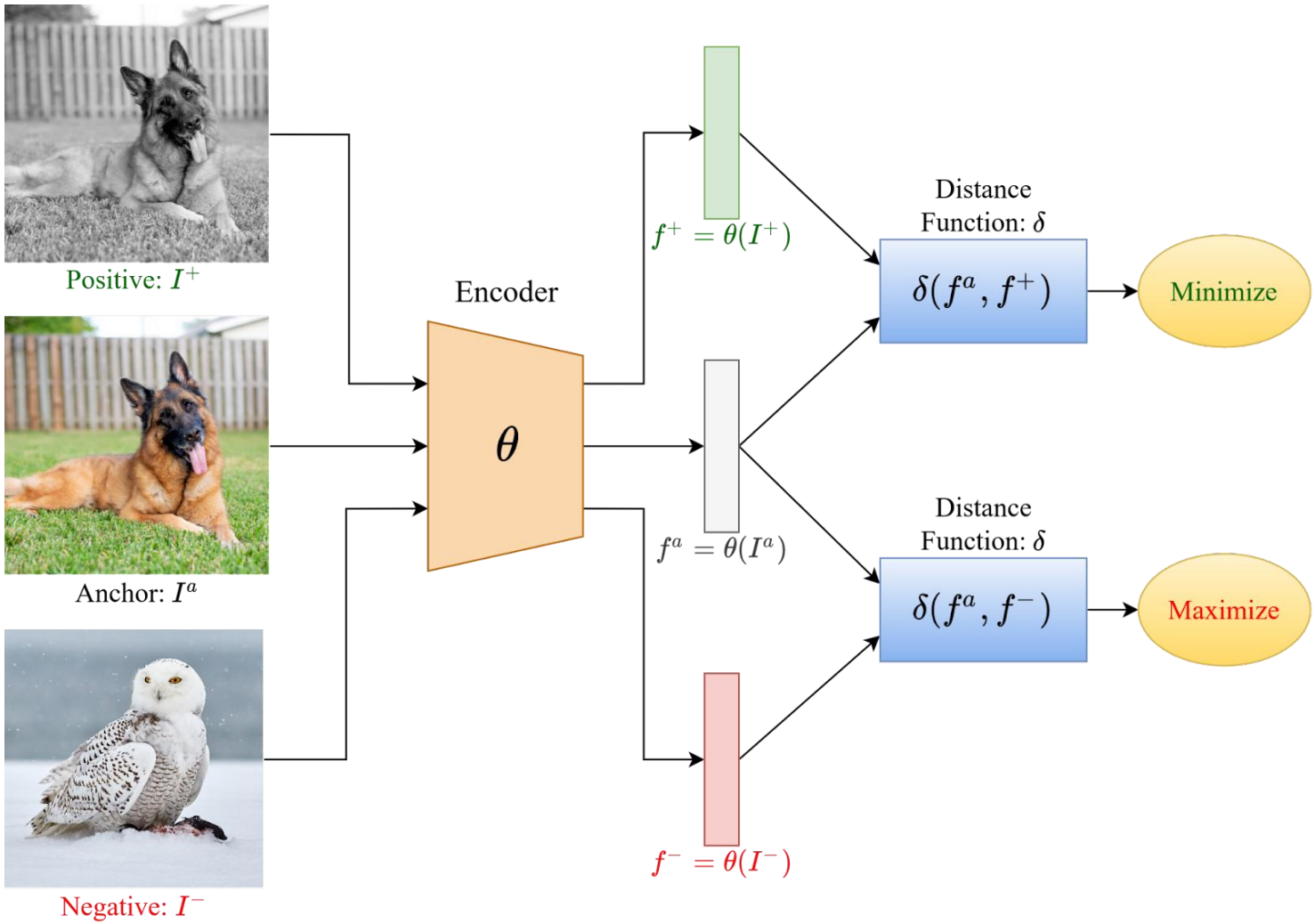
I want to cluster (attract) similar even though transformed images, and repel the different ones

contrastive learning is not tighted to self supervised is just a type of loss

note that the dimension of the denominatore is a function of the batch size and not the number of class like in the original softmax

Enforces image uniqueness and augmentation invariance.

How to train with contrastive losses



Self-supervised learning is conservative supervised learning

In supervised learning the learning is by labelled examples, because for each samples someone is gonna telling me the label.
In self-supervised instead, the learning is by rule, there are a set of transformations which doesn't change the semantic of my sample

Contrastive supervised learning:

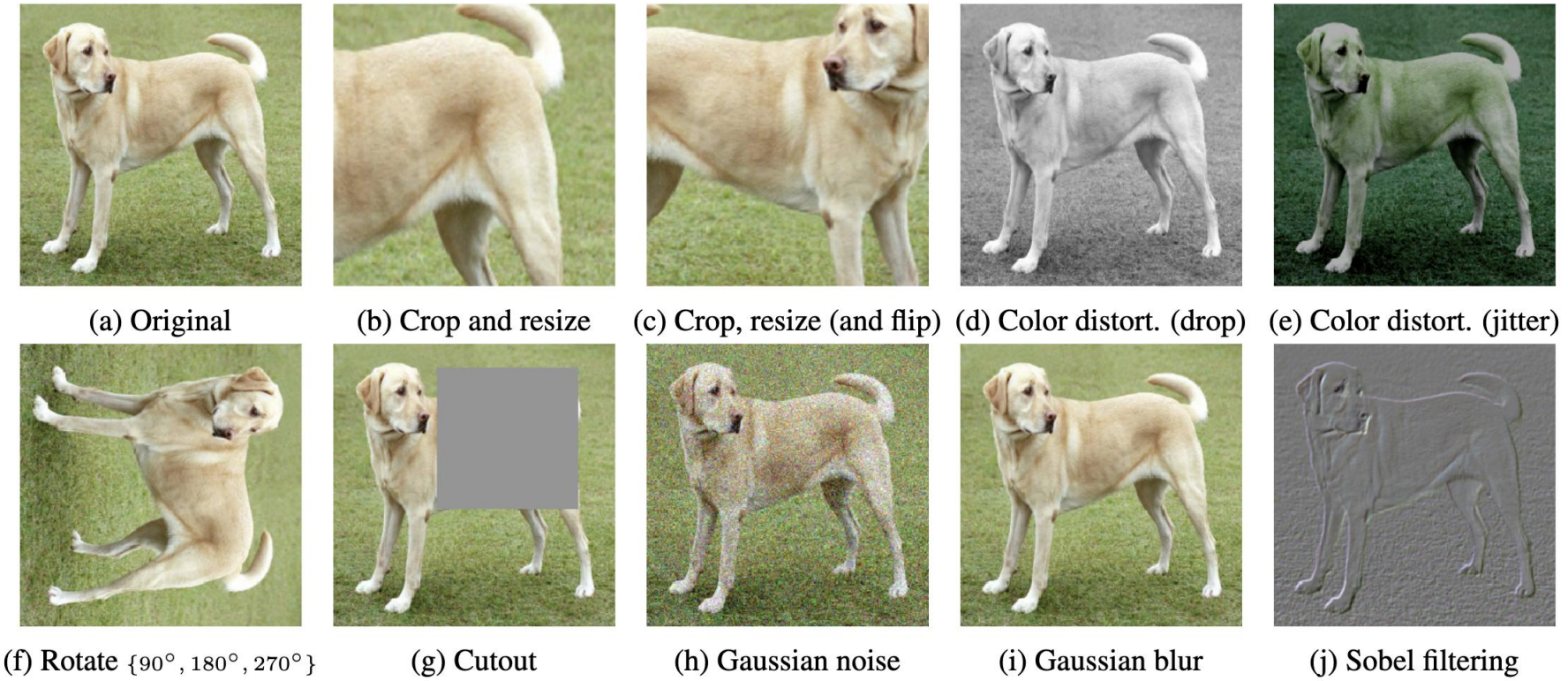
Pull samples of same class together, push others away.

Contrastive self-supervised learning:

Pull augmented versions of same sample together, push others away.

[Chen et al. ICLR 2020]

Self-supervised learning is learning augmentation invariance



(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

We want augmentations of a sample to lead to the same embedding representation.

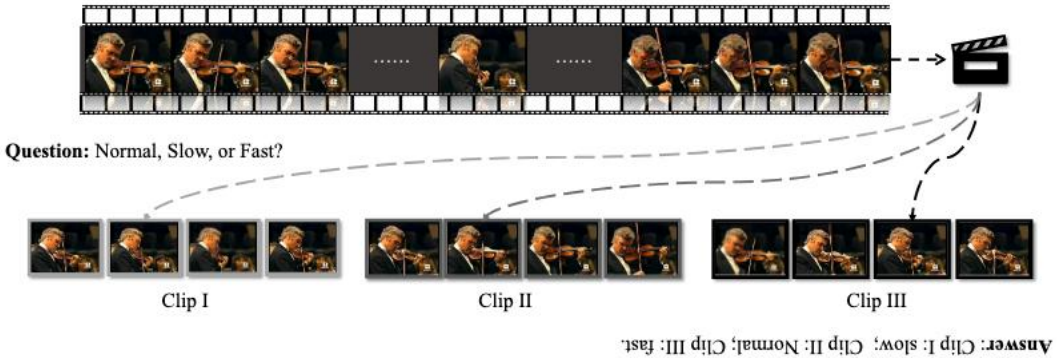
Self-supervised video learning

Videos have a super strong extra signal to learn from: time.

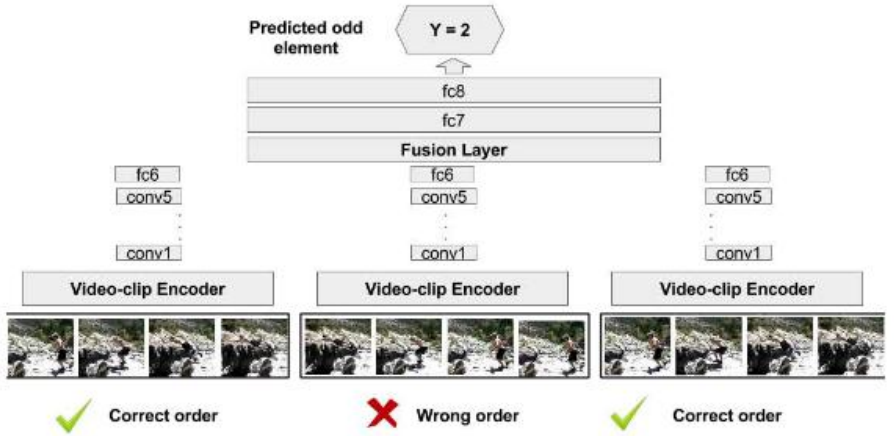
How can time be used for pretext tasks?

- Predict temporal order.
- Predict whether video is played in reverse or not.
- Predict alignment between video and audio.

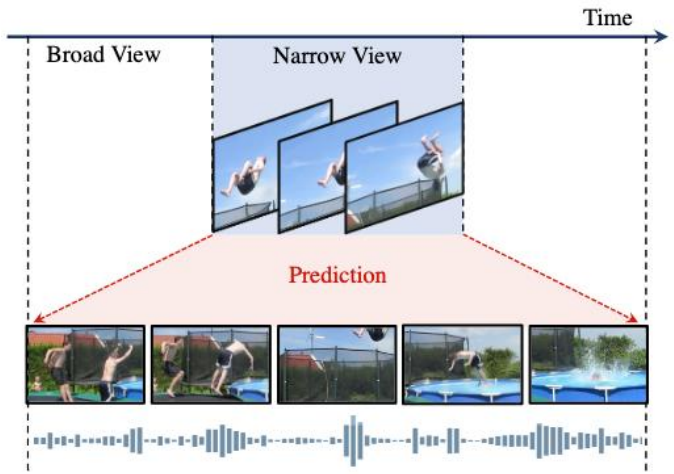
Examples of self-supervised video learning



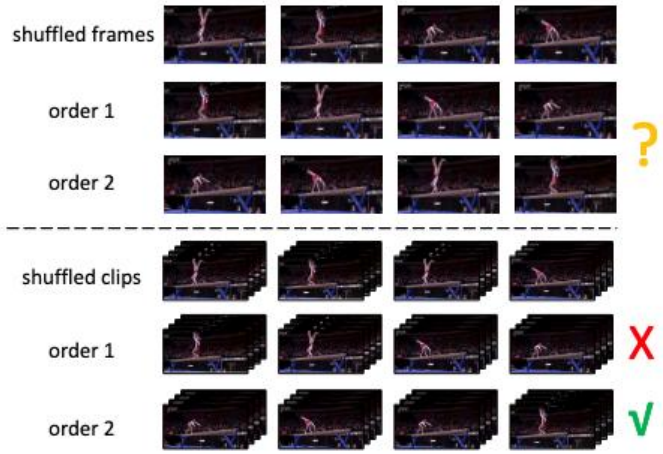
Wang et al. (2020): Predict pace.



Basura et al. (2017): Predict odd-one-out.



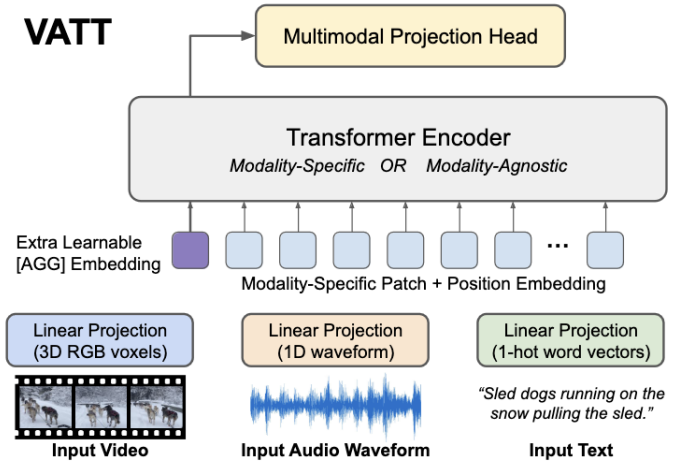
Recasens et al. (2021): Narrow to broad prediction.



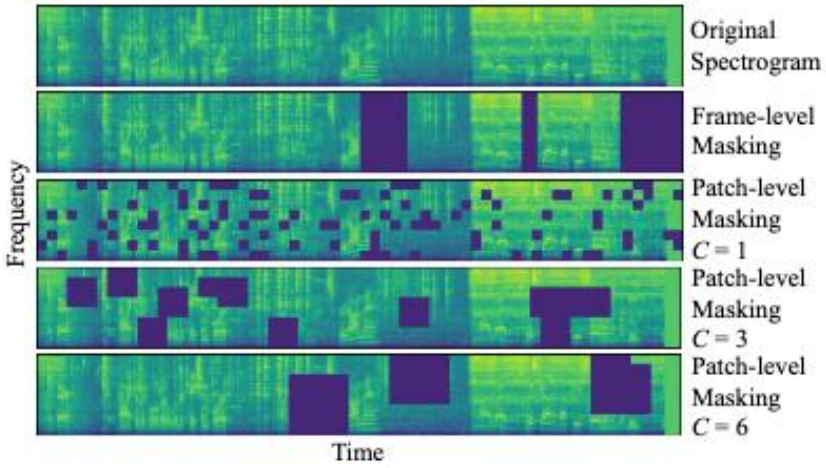
Xu et al. (2019): Predict clip order.

Break

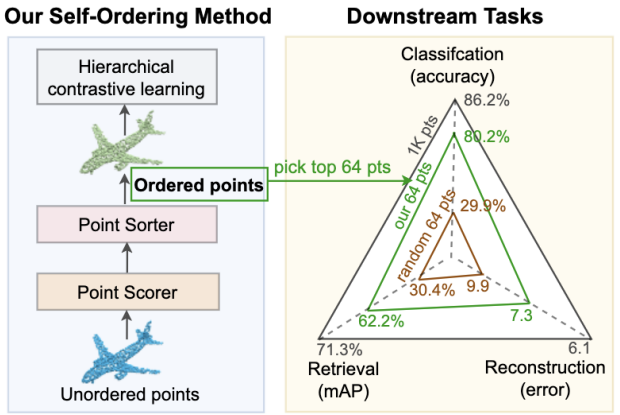
Self-supervised learning on other modalities



[Akbari et al. NeurIPS 2021]



[Gong et al. AAAI 2022]

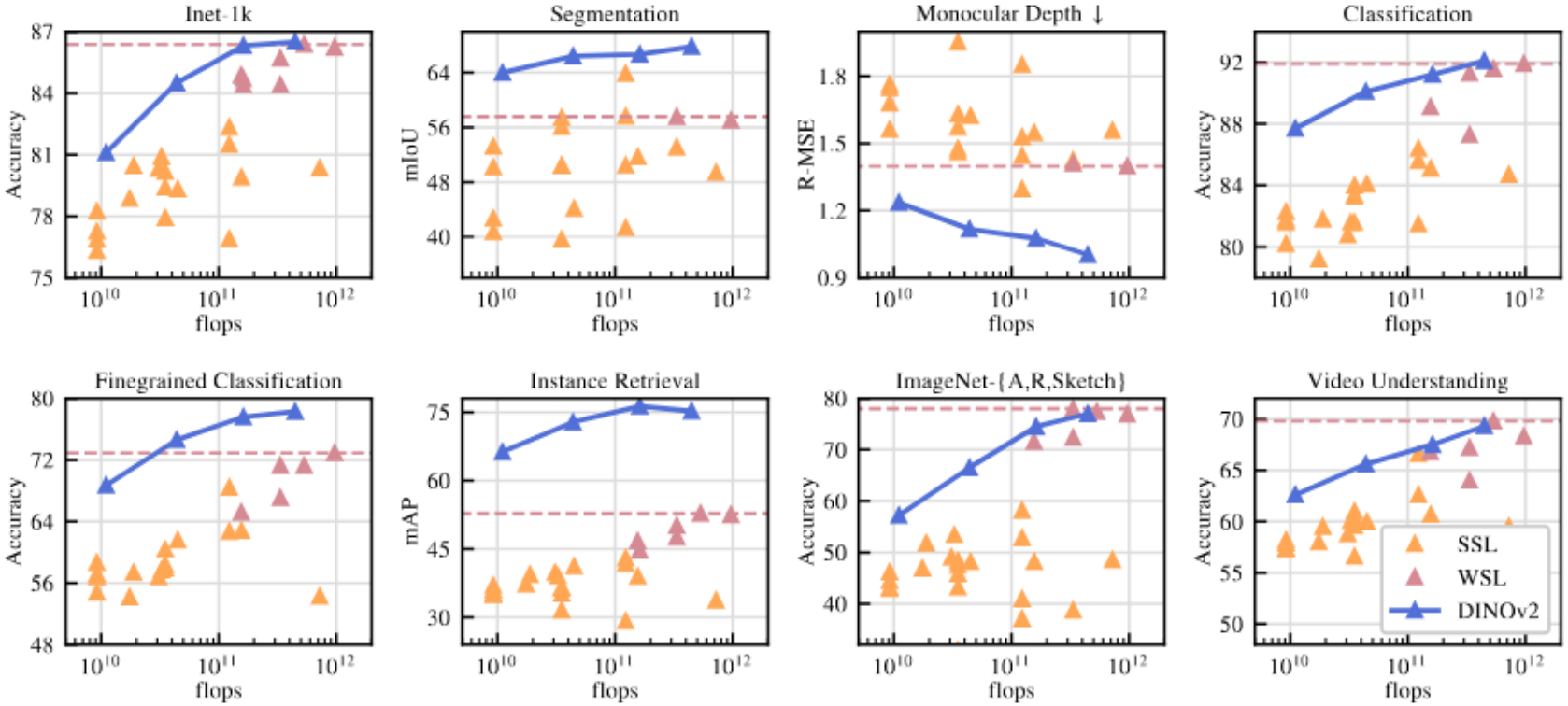


[Yang et al. ICCV 2023]

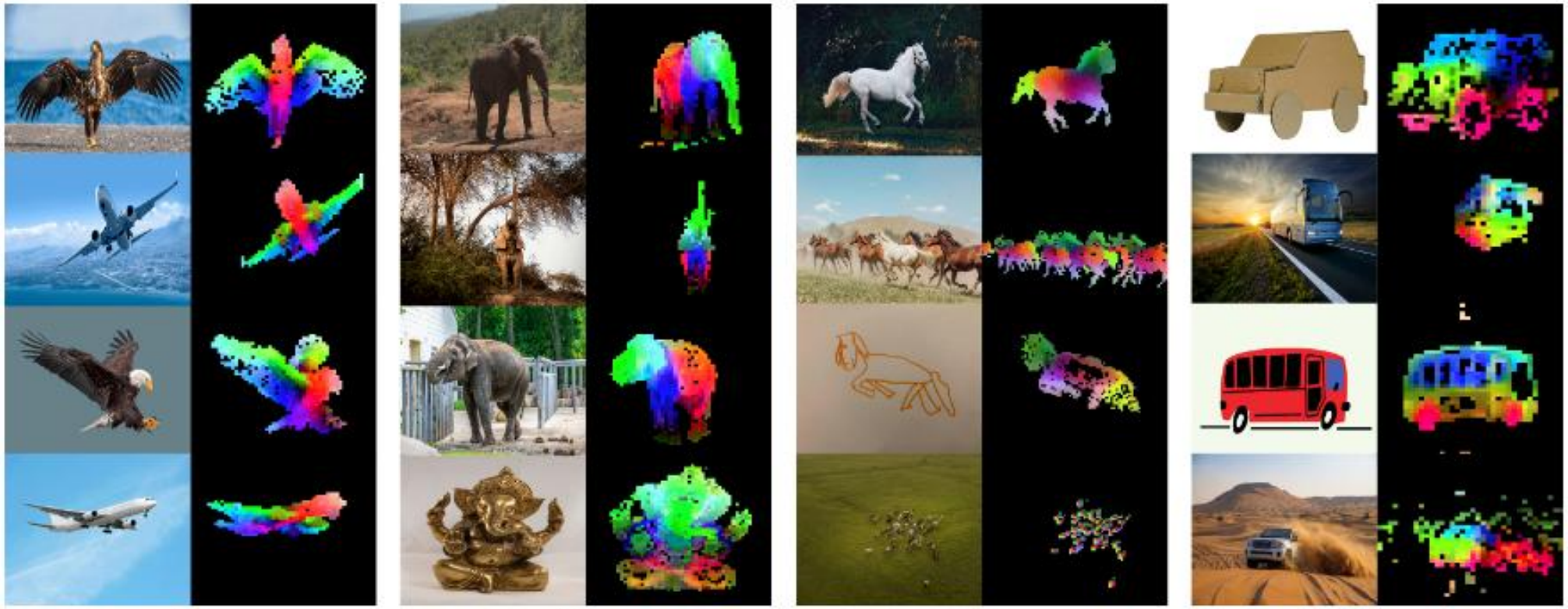
How good are self-supervised models? A DINOv2 study

DINOv2 (2024): Contrastive learning + patch-level masking + tricks + 142M dataset.

a model trained on many transformation without label, works better than a model trained on fewer samples but with label. Because the formers are able to learn structural invariances



How good are self-supervised models? A DINOv2 study



Supervised networks struggle when being deployed in new settings, self-supervised networks thrive in such settings.

Self-supervised learning for language

What if our data is a collection of sentences?

why is it not unsupervised?
because the human wrote the phrases in an ordered way

“The quick brown fox jumps over the lazy dog.”

I.e., how can we train Large Language Models on the internet?

Masked Language Modelling

Main idea is simple: remove some tokens and predict them.

Set of classification labels: All tokens.

Targets: Tokens that were removed.

Just like self-supervised visual learning, the problem falls back to a standard classification setup, but now with “free labels”.

Masked Language Modelling

Standard setting: Sample 15% of tokens and replace with [MASK].

“The quick brown [MASK] jumps over the [MASK] dog.”

Modified MLM: Sample 15% of tokens. Replace 80% with [MASK], 10% with random token, and 10% left unchanged.

“The quick brown [oven] jumps over the [maybe] dog.”

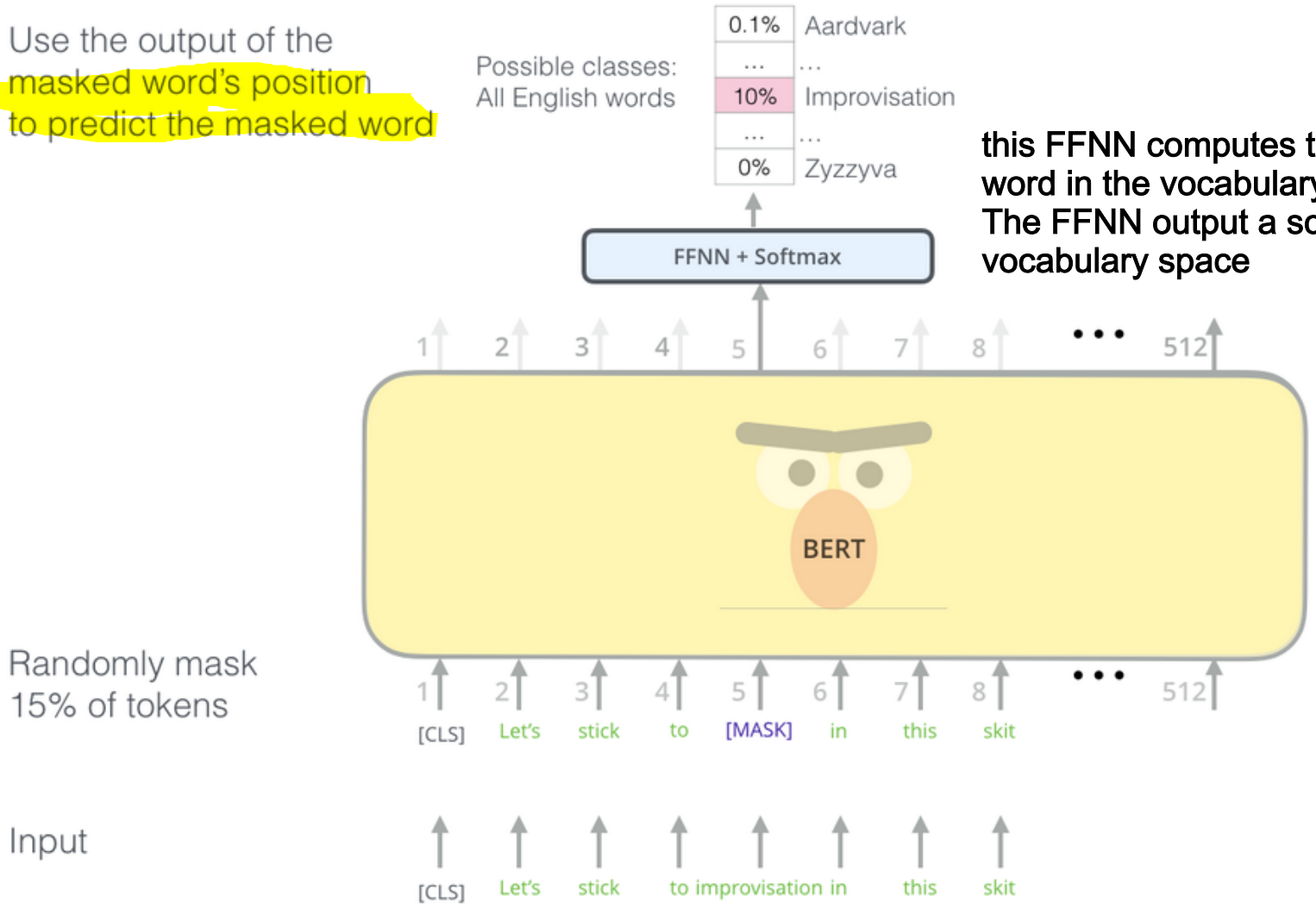
“The quick brown [fox] jumps over the [lazy] dog.”

Modified Masked Language Modelling

For 80% of the sampled tokens, we simply need to predict the masked input.

For 10% of the tokens, the model needs to figure out that the word needs to be replaced.

For the remaining 10%, the model needs to figure out to do nothing.

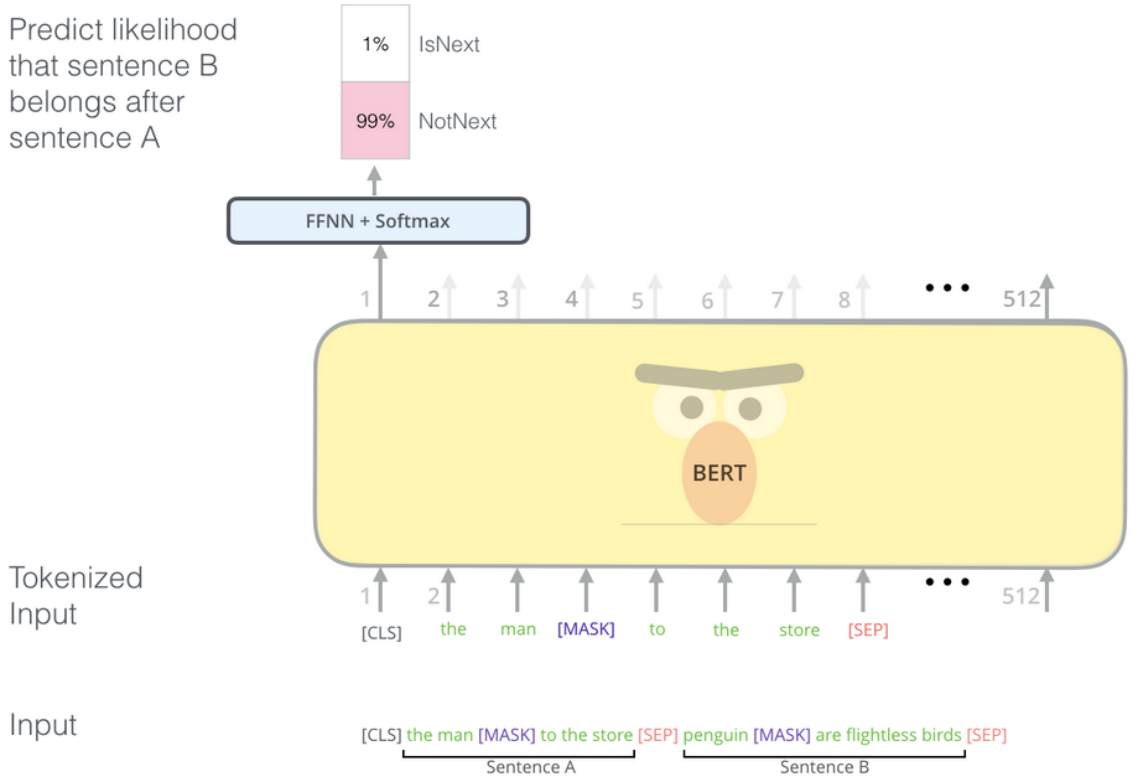


this FFNN computes the scores (logits) for each word in the vocabulary being the masked word. The FFNN output a score vector in the vocabulary space

https://jalammar.github.io/illustrated-bert/

Next Sentence Prediction

Main idea: Given two sentences, predict whether the first follows the second.



Take two sentences that actually appear consecutively in the original text corpus. Since they were originally adjacent, you label this pair as "IsNext" or positive

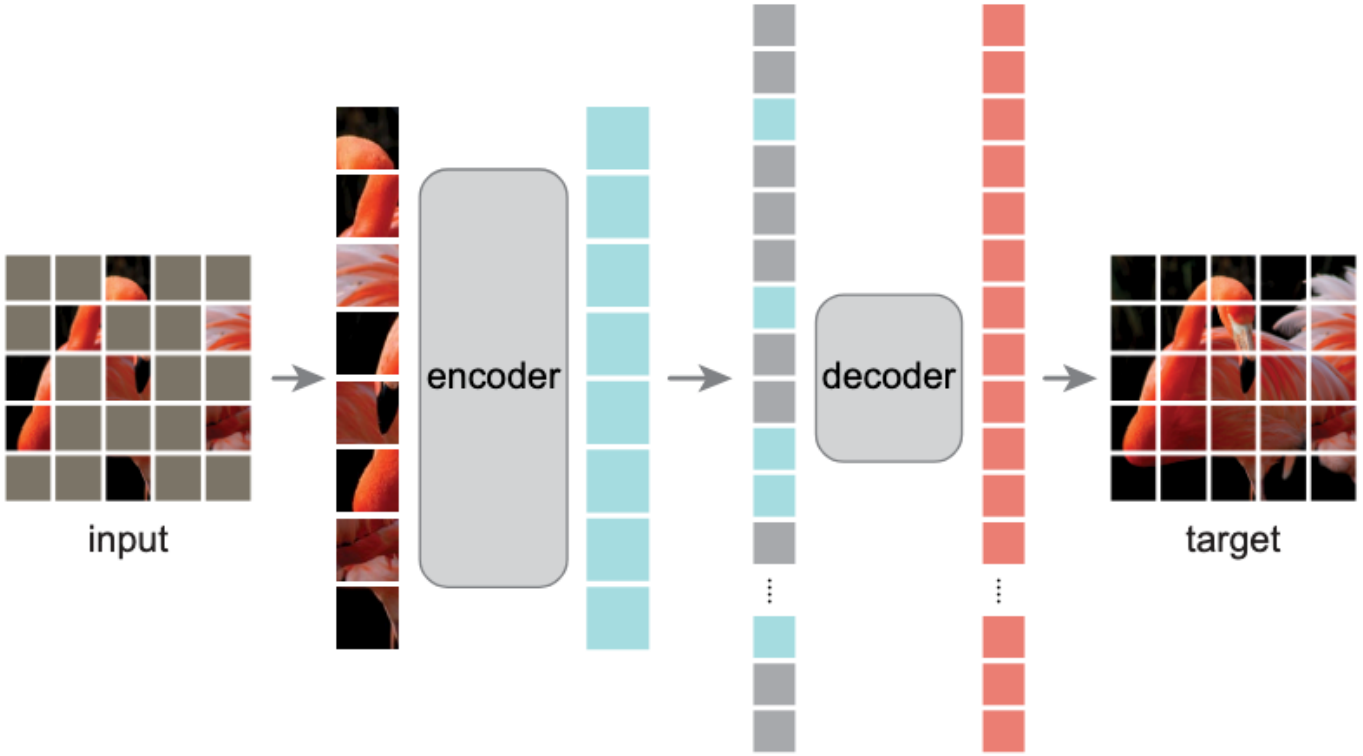
Take the first sentence, then randomly select a second sentence from somewhere else in the corpus (from a different document or a different part of the same document). Since these were not originally adjacent, you label this pair as "NotNext" or negative.

How to automatically generate “labels” for this setting?

MLM in vision: masked autoencoding

Many ideas from one domain are inspiration for the next domain.

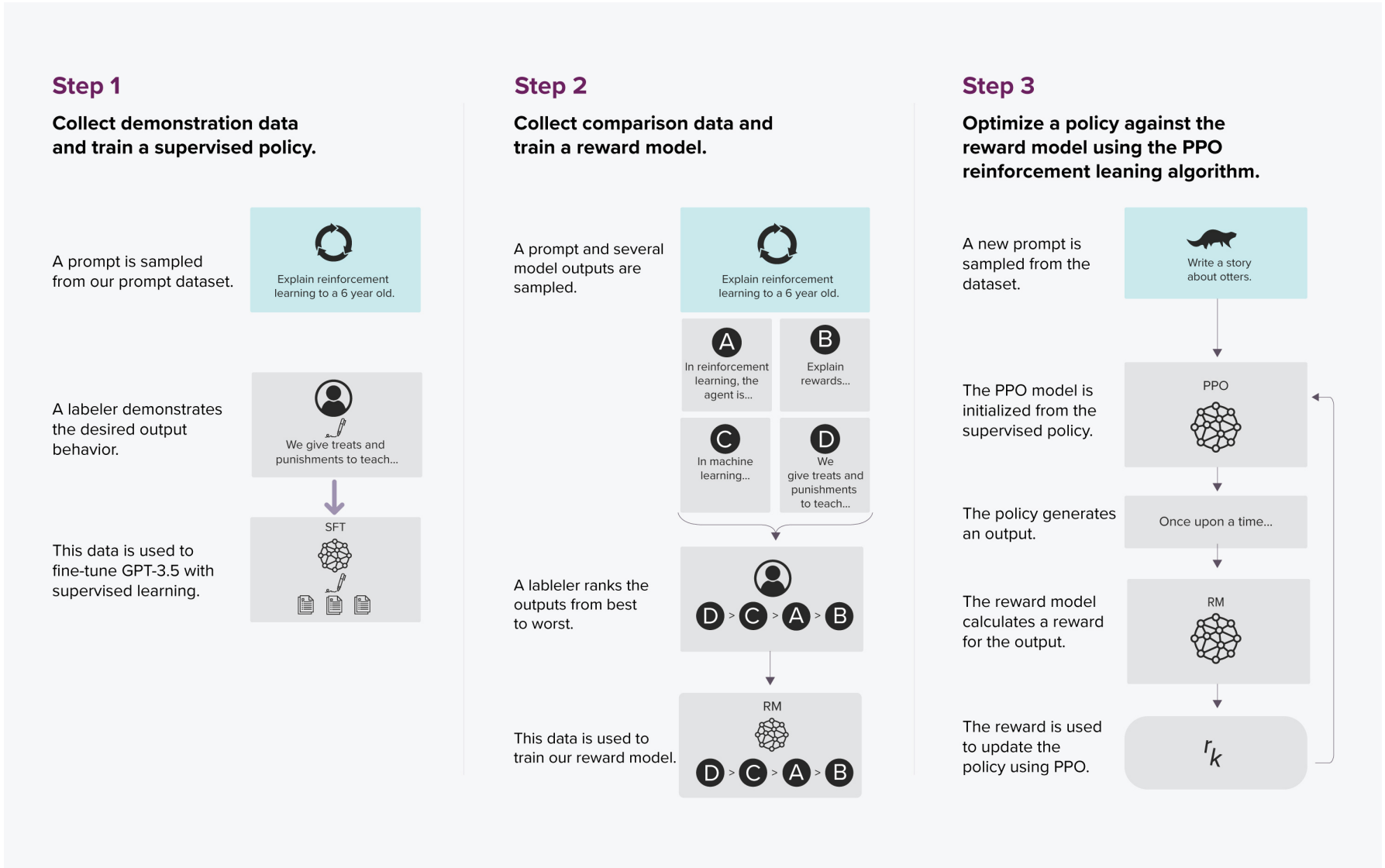
I crossed out some patches, I feed them to the net and I learn to predict them back



“Masked Autoencoders Are Scalable Vision Learners” He et al. CVPR 2022

Revisiting RLHF

reinforcement learning human feedback



RLHF from a supervision perspective

Self-supervision (GPT) is not enough for Large Language Models (ChatGPT).

Prompt: Explain why we need water to survive as humans.

(Imaginary) GPT: Explain why humans die when not given any water.

What is going on here?

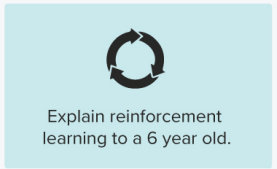
We lack alignment with the intent of the user input.

This needs human supervision.

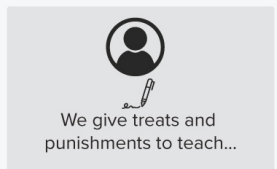
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.

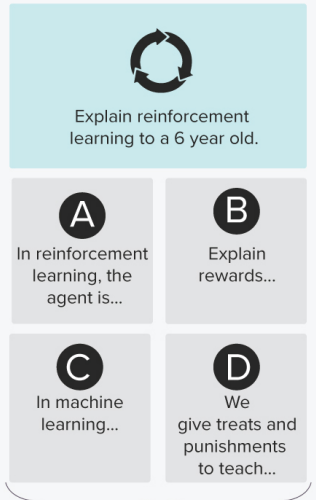


This data is used to fine-tune GPT-3.5 with supervised learning.

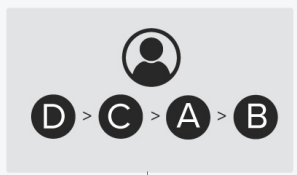
Step 2

Collect comparison data and train a reward model.

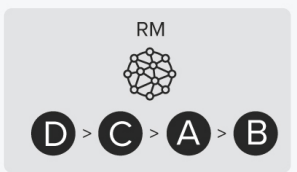
A prompt and several model outputs are sampled.



A lableler ranks the outputs from best to worst.



This data is used to train our reward model.



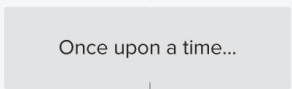
Step 3

Optimize a policy against the reward model using the PPO reinforcement leaning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.



The reward model calculates a reward for the output.

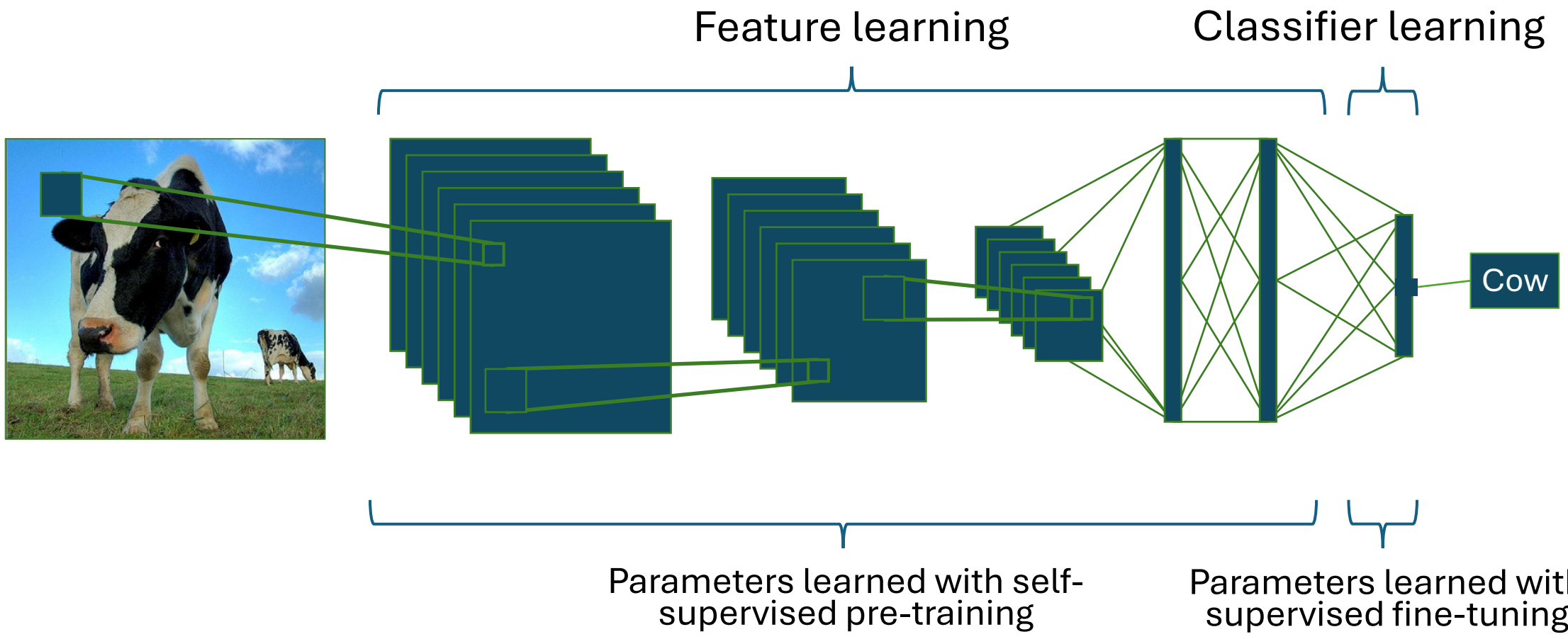
The reward is used to update the policy using PPO.

LLM = Transformer + self-supervised pre-training + human aligned tuning

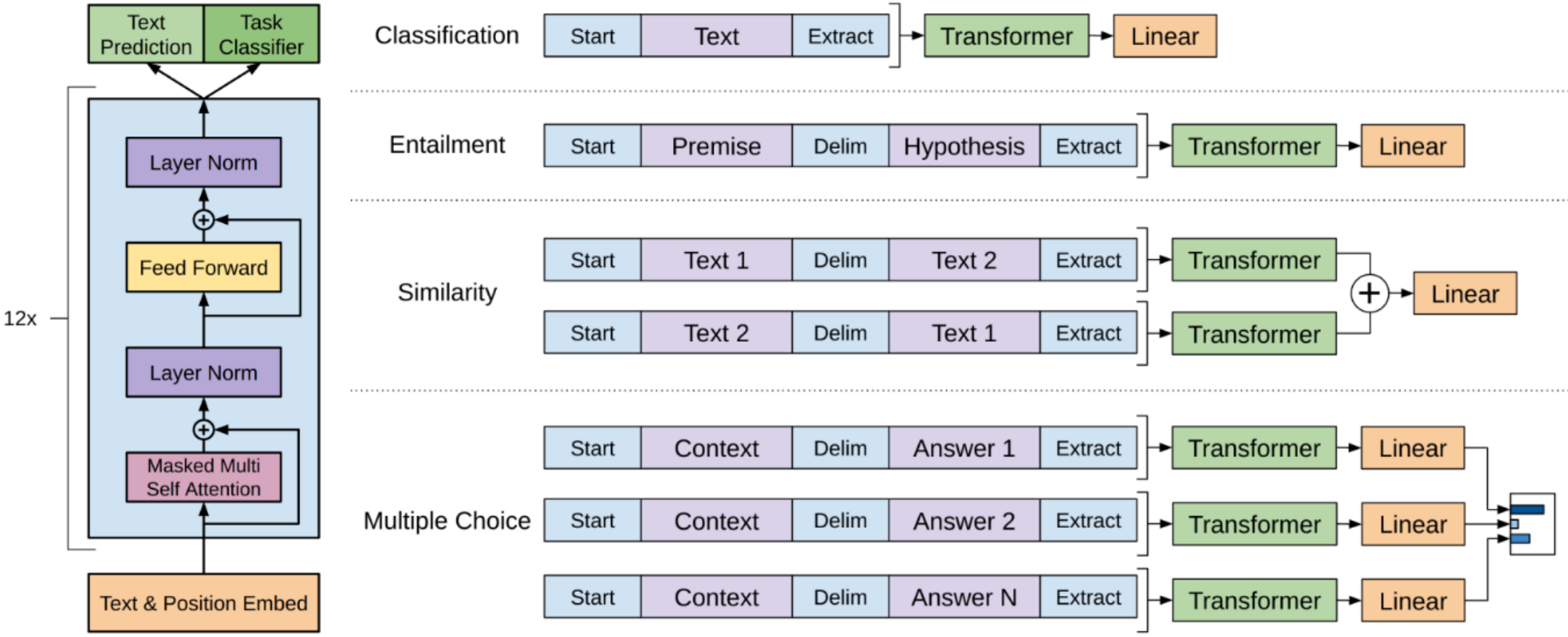
Discussion: why are they so good? And is this all you need for deep learning?

Is there something in between supervised and self-supervised learning?

Classical setup: pre-training and fine-tuning



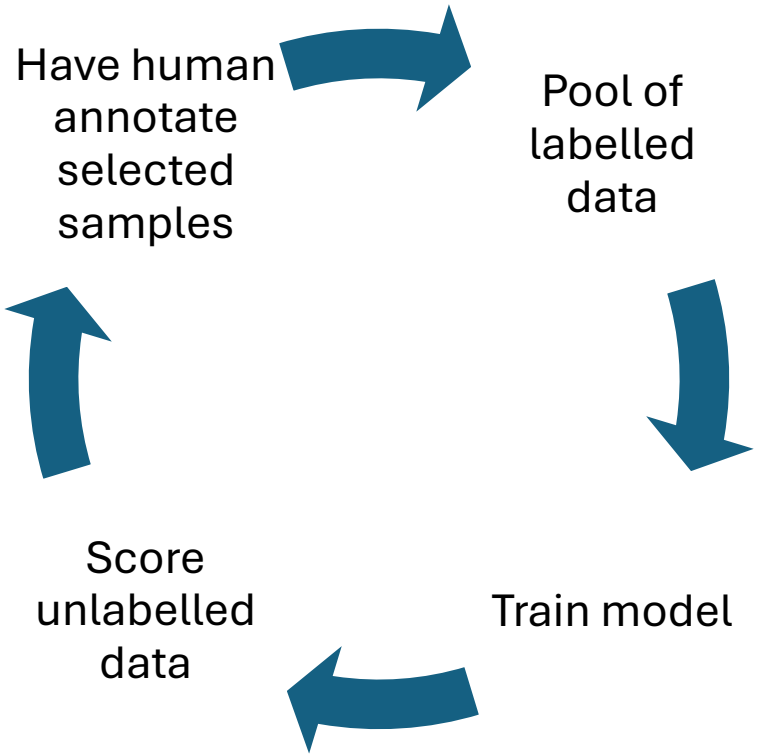
Examples of fine-tuning in language



Active learning

i have 2 sets, one unlabelled and one labelled, I want to use the feature learned from the labeled set on the unlabelled set

Assume we only have an unlabelled training set. Is it possible to simultaneously label samples and train a model?



Which samples to select?

Random: randomly select (ignore scoring).

Most uncertain: closest to the decision bounary, or lowest norm in embedding space (second to last layer), or highest likelihood entropy.

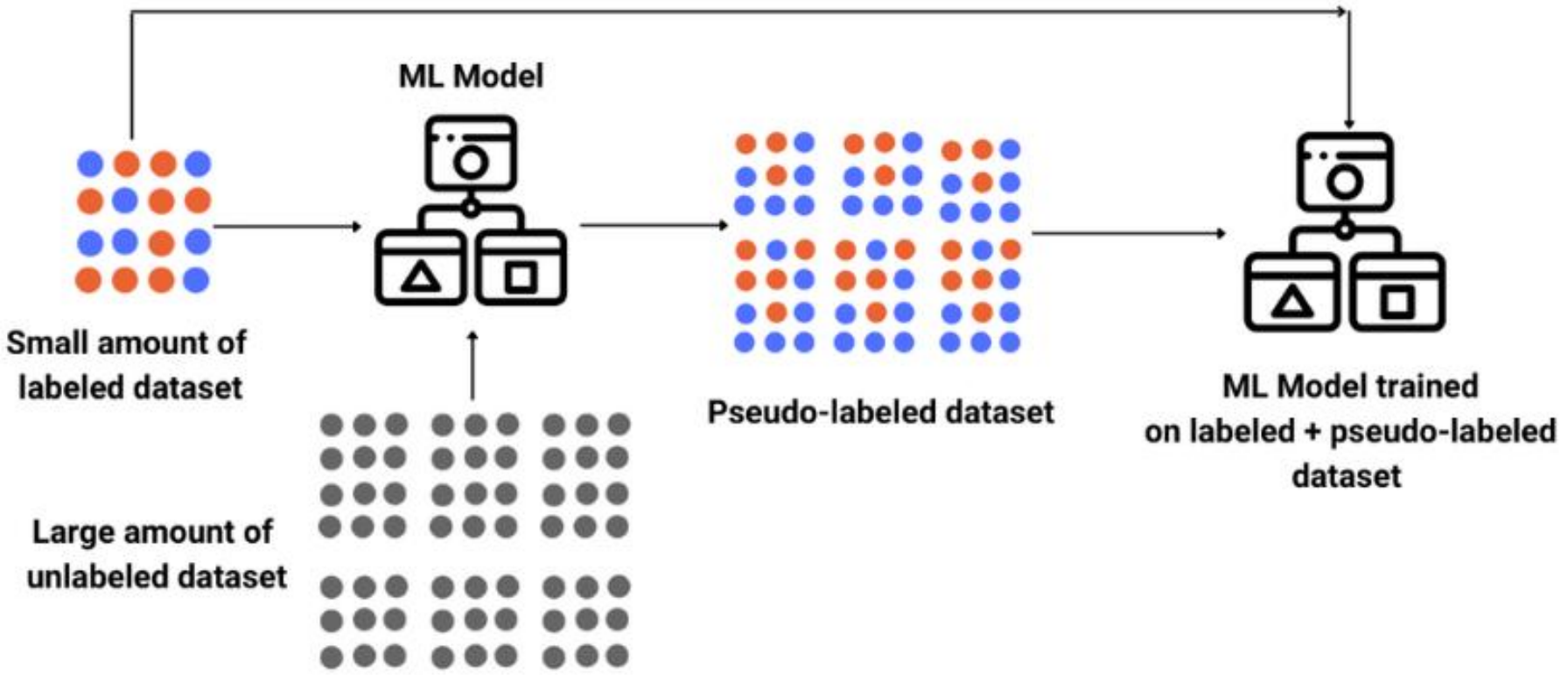
Group-based metrics: Uniformity over classes to avoid biases.

Mix: Combine a mix of X% random and (100-X)% uncertain+group.

Semi-supervised learning

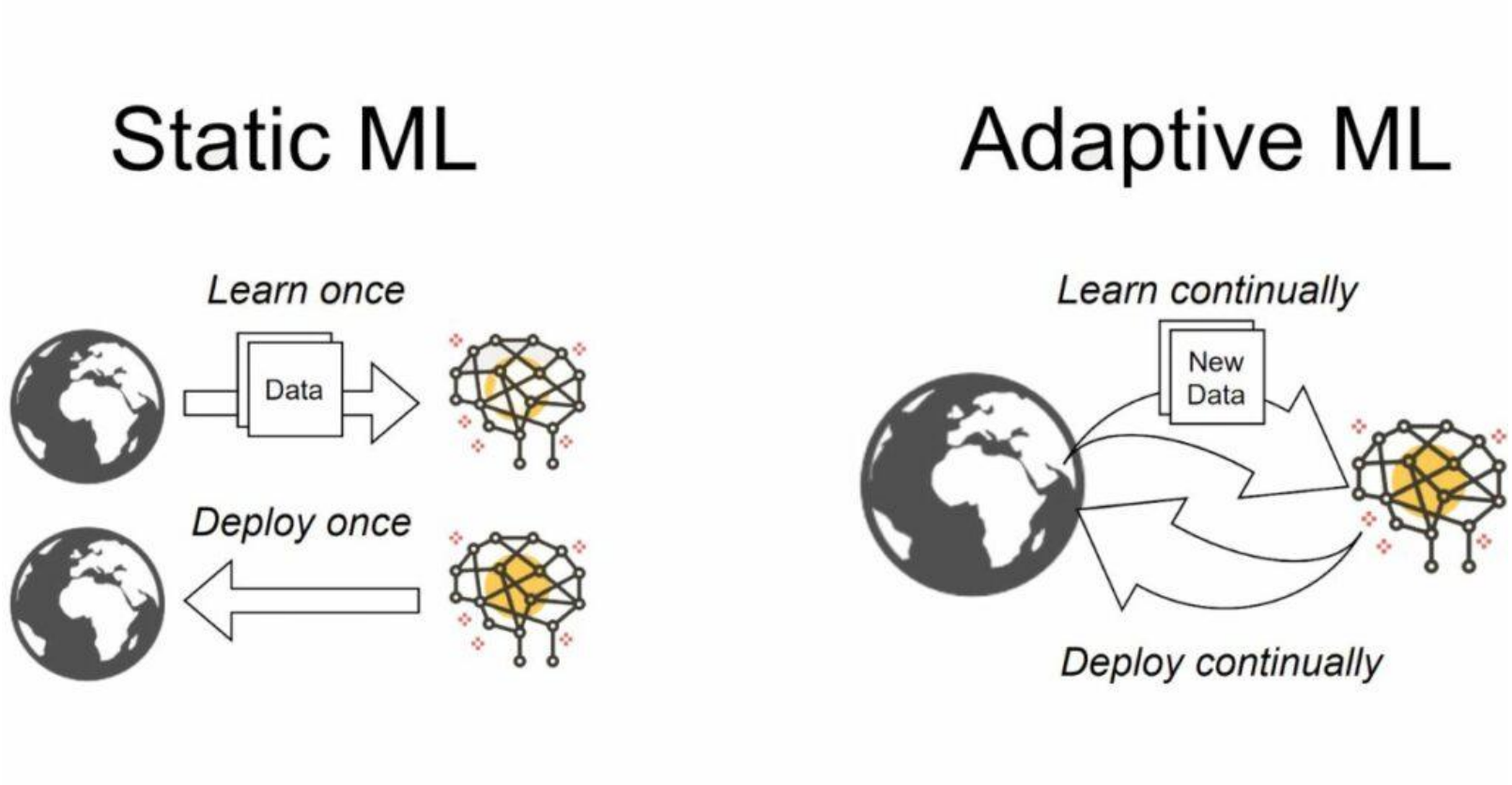
pseudo-labeling (or self-training)

can be seen as a static moment in time of active learning



Continual learning

In the real world, there is no such thing as a static dataset.



catastrophic forgetting

Shockingly, we can’t just train on new data! Much more in lecture 11.

Is self-supervised learning truly without supervision?

My view on supervised vs self-supervised learning

Supervised learning	**Self-supervised learning**
Label by sample.	Label by rule.
Invariance defined at global semantic level.	Invariance defined at geometric or local semantic level.

Self-supervised learning is conservative supervised learning from pre-defined invariances.

Next lecture

Lecture	Title
1	Intro and history of deep learning
3	Deep learning optimization I
5	Convolutional deep learning
7	Graph deep learning
9	Multi-modal deep learning
11	What doesn't work in deep learning
13	Q&A

Lecture	Title
2	AutoDiff
4	Deep learning optimization II
6	Attention-based deep learning
8	From supervised to unsupervised deep learning
10	Generative deep learning
12	Non-Euclidean deep learning
14	Deep learning for videos

Learning and reflection

TODO

Thank you