Multimodal Generative AI 2025
Self-Supervised Learning

VISTA
École Polytechnique

Today's lecture

Part I: Learning without labels

Part II: Self-Supervised Learning

Part III: Multimodal Self-Supervised Learning

Today's lecture

Part I: Learning without labels

Part II: Self-Supervised Learning

Part III: Multimodal Self-Supervised Learning

Pop Quiz

- Types of learning that do not require labeled data?

Pop Quiz - Hint

- Types of learning that do not require labeled data?
  - Supervised Learning
  - Unsupervised Learning

Pop Quiz Answers

- Types of learning that do not require labeled data?
  - Supervised Learning
  - Semi-supervised Learning
  - Weakly-supervised Learning
  - Self-supervised Learning
  - Unsupervised Learning

  - Active Learning

Part I: Outline
Learning without labels

- Learning from data:
  - Supervised Learning
  - Unsupervised Learning
  - Weakly-supervised Learning
  - Semi-supervised Learning
- Active Learning

Supervised Learning

Supervised Learning

- **Data**: (x,y), where x is data, y is label
- **Goal**: Learn a function to map x → y

Examples:
Classification,
regression,
object detection,
segmentation,
captioning…

Input: Image

Output: Assign Image to one of a fixed set of categories

Cat
Dog
Deer
Bird
Car

Unsupervised Learning

Supervised Learning

- **Data**: (x,y), where x is data, y is label
- **Goal**: Learn a function to map x → y

Examples:
Classification,
regression,
object detection,
segmentation,
captioning…

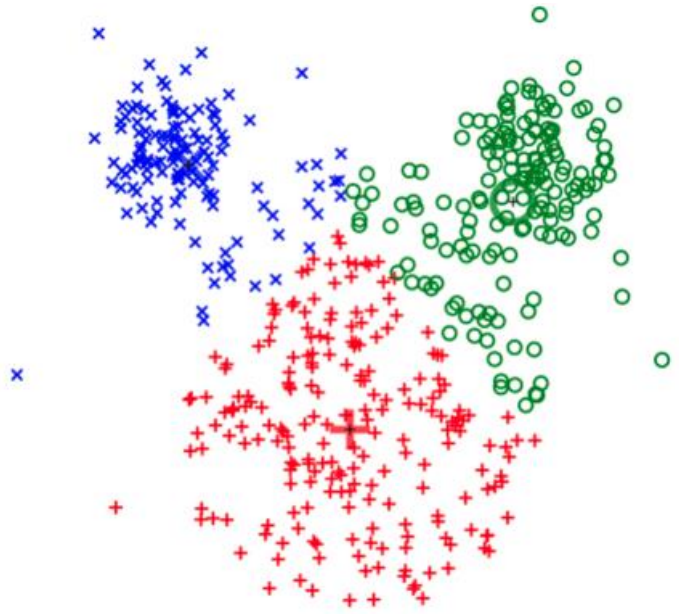
Unsupervised Learning

- **Data**: x, where x is data, NO label!
- **Goal**: Learn some underlying hidden structure of the data
- **Advantage**: No labelling

Examples:
Clustering
Dim reduction
Feature learning
Density estimation…

Unsupervised Learning

- **Data**: x, where x is data, NO label!
- **Goal**: Learn some underlying hidden structure of the data



Examples:

Clustering

Dim reduction

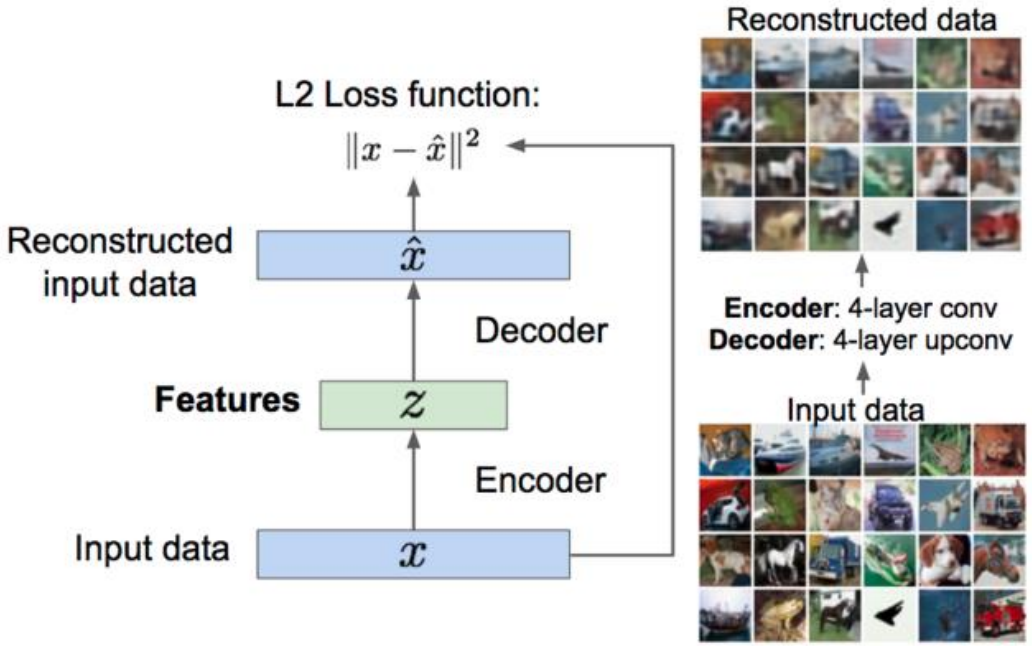
Feature learning

Density estimation…

K-Means Clustering

Unsupervised Learning

- **Data**: x, where x is data, NO label!
- **Goal**: Learn some underlying hidden structure of the data



Examples:

Clustering

Dim reduction

Feature learning

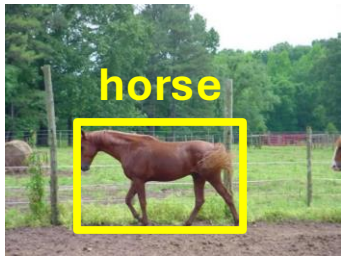
Density estimation…

Feature Learning
e.g. autoencoders

Weakly-Supervised Learning

Manual annotations are expensive

80-200s



40s

3s

(free)

Su et al., Crowdsourcing annotations for visual object detection, AAAI 2012

Deng et al., Scalable multi-label annotation, CHI 2014

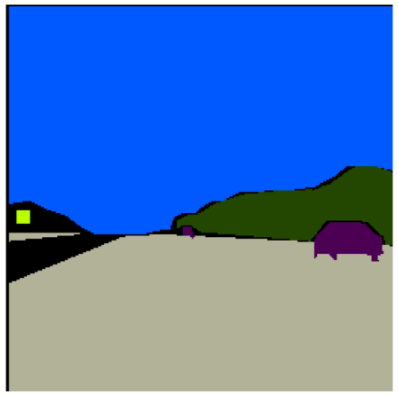
Lin et al., Microsoft COCO: common objects in context, ECCV 2014

Fully supervised setting

motorbike

...

Object detection model



Car
Road
Sign
Sky
Tree

...

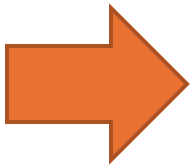
Semantic segmentation model

*Annotation to the **same** degree as outputs on test images*

Weakly-supervised learning

motorbike

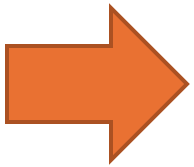
...



Object detection model

car, road, sign, sky, tree

...



Semantic segmentation model

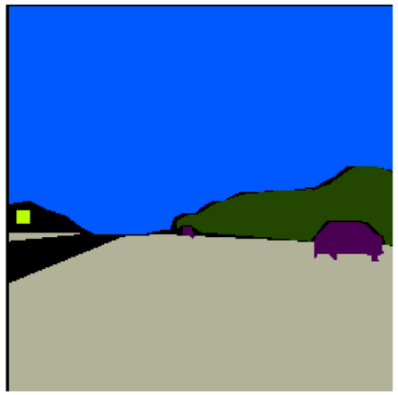
*Annotation to a **lower** degree than outputs on test images*

Weakly-supervised learning

...

motorbike

Object detection model



Car
Road
Sign
Sky
Tree

...

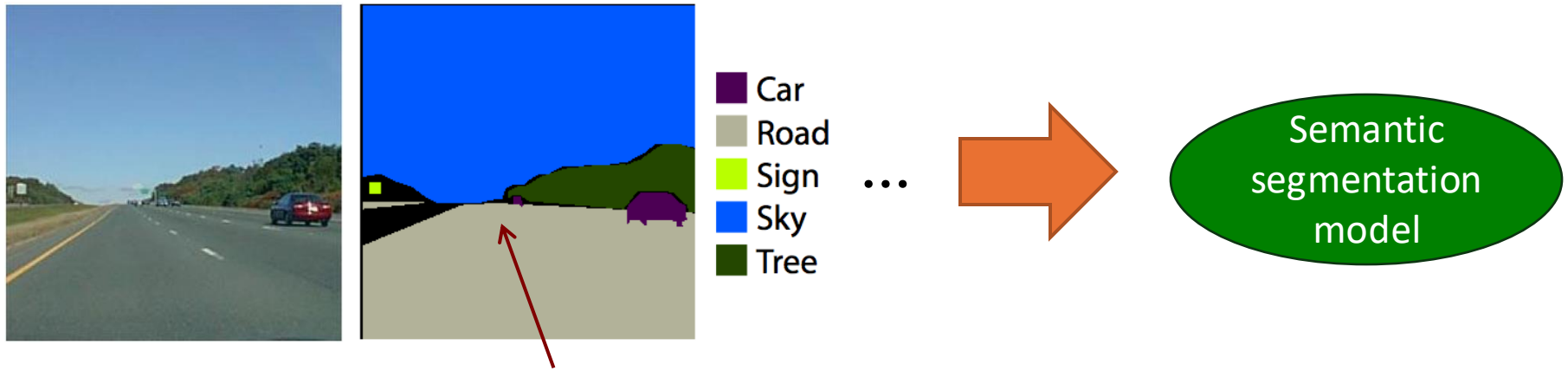
Semantic segmentation model

Core issue: need for auto-annotation

Weakly-supervised learning

- *Train with lower degree of supervision than necessary to fix latent (internal) variables*

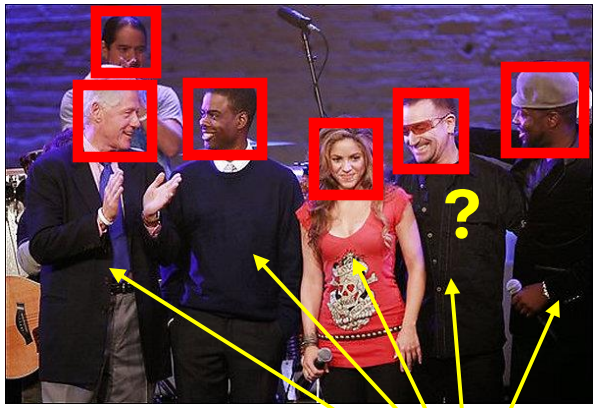
- *Need to inject additional assumptions (bias) in the learner, e.g. appearance classes are compact, objects form recurring appearance patterns*



Pixel labels = latent variables vs

appearance classifiers = model parameters

Learn face models from news captions



Former U.S. President Bill Clinton, comedian **Chris Rock**, musicians Shakira, Bono, and Wyclef jean participate …

Correspondence ambiguity

Source: Yahoo-news

McChrystal and Gates raise tension in Afghanistan debate

Weakly supervised learning here means:

- Recover latent name-to-face assignments
- Learn face appearance models specific to person sub-classes

Semi-Supervised Learning

Supervised Learning

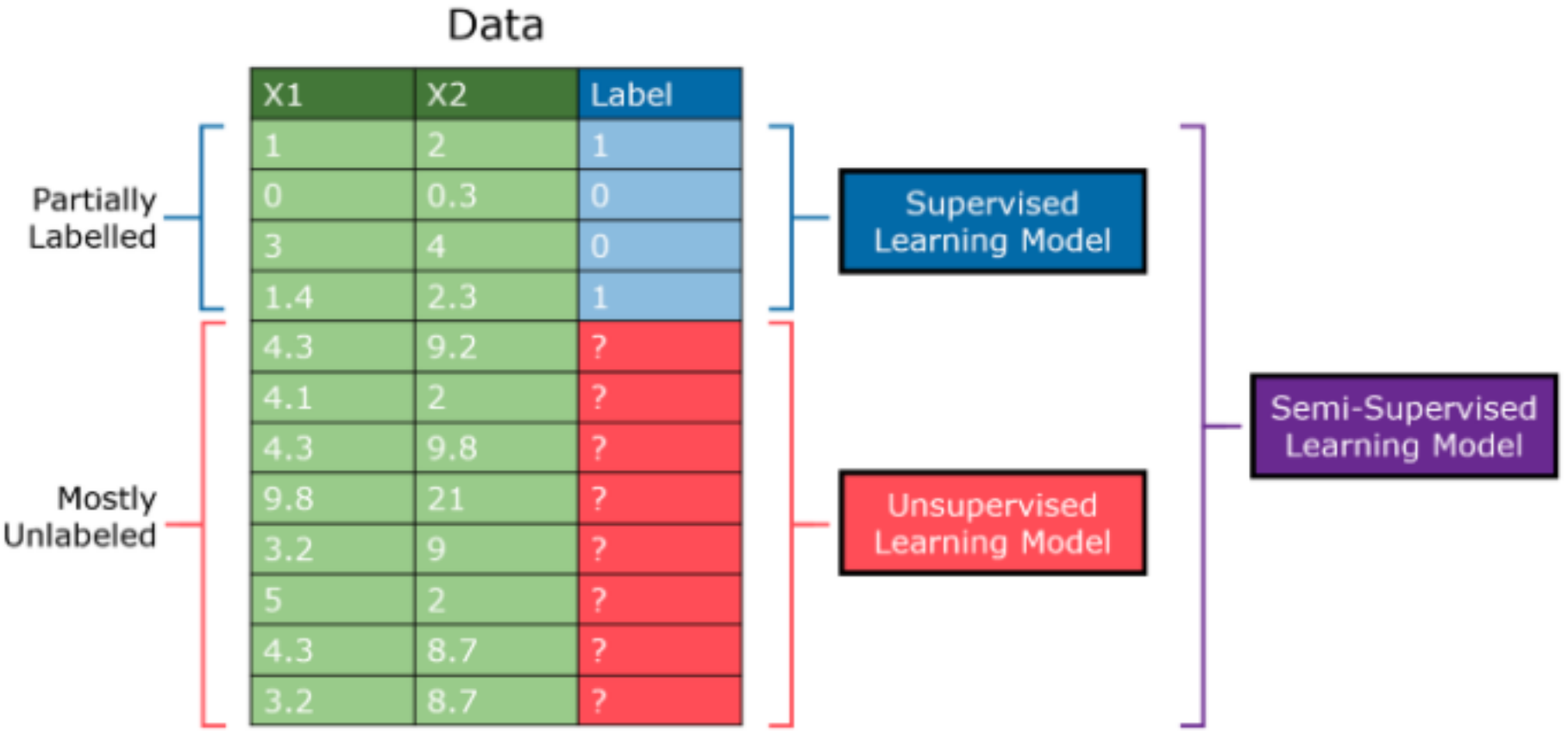
- **Data**: (x,y), where x is data, y is label
- **Goal**: Learn a function to map x → y

Examples:
Classification, regression, object detection, segmentation, captioning…

Semi-supervised Learning

- **Data**: labeled (x_1,y_1), where x is data, y is the label and (x_2) data with NO label!
  - $x_2 >> x_1$ → much more unlabeled data
- **Goal**: Learn a better prediction rule than based on labeled data alone
- **Advantage**: Partial labelling

Semi-supervised Learning



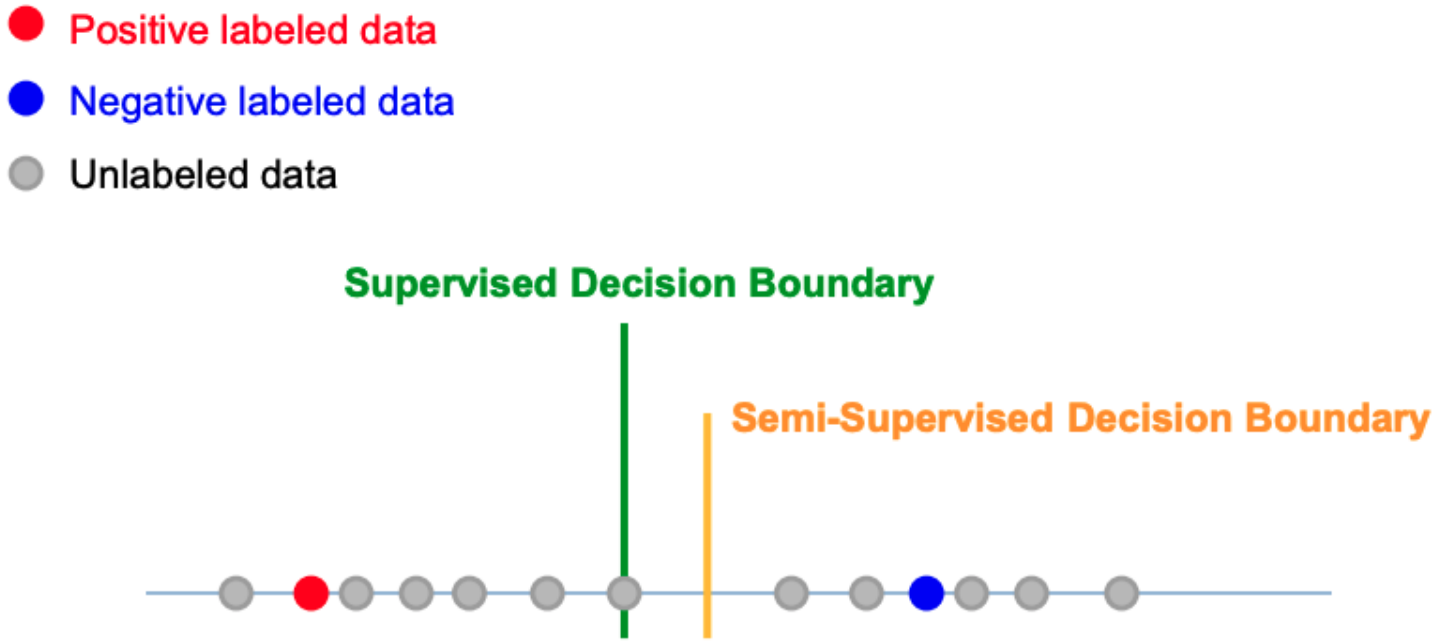
Semi-supervised Learning

- Limited set of labelled sample data to train itself
  → partially-trained model
- Use partially-trained model to label the unlabeled data
  → pseudo-labeled data
- Combine labeled and pseudo-labeled data

Semi-supervised Learning:
Supervised + Unsupervised:

1. Classification to identify data assets
2. Clustering to group it into distinct parts

Can unlabeled data help?

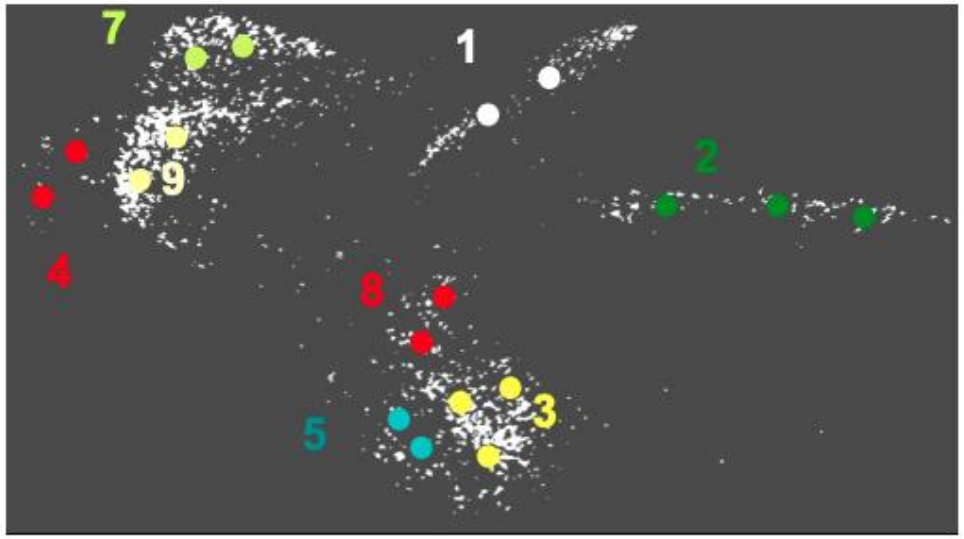


Assume each class is a coherent group (e.g. Gaussian)

Then unlabeled data can help identify the boundary more accurately.

Can unlabeled data help?

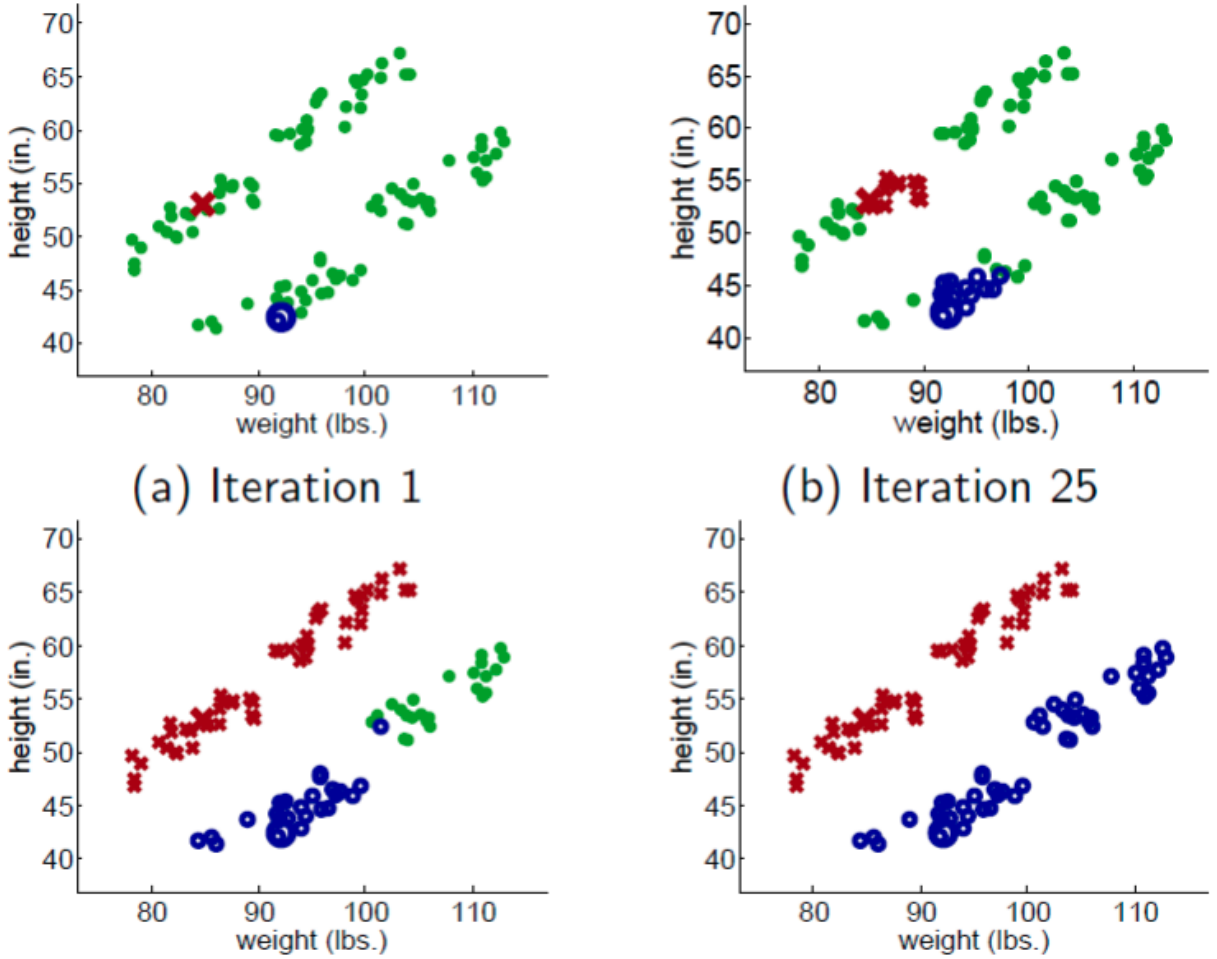
Unlabeled Images  Labels "0" "1" "2" …



This embedding can be done by manifold learning algorithms

"Similar" data points have "similar" labels

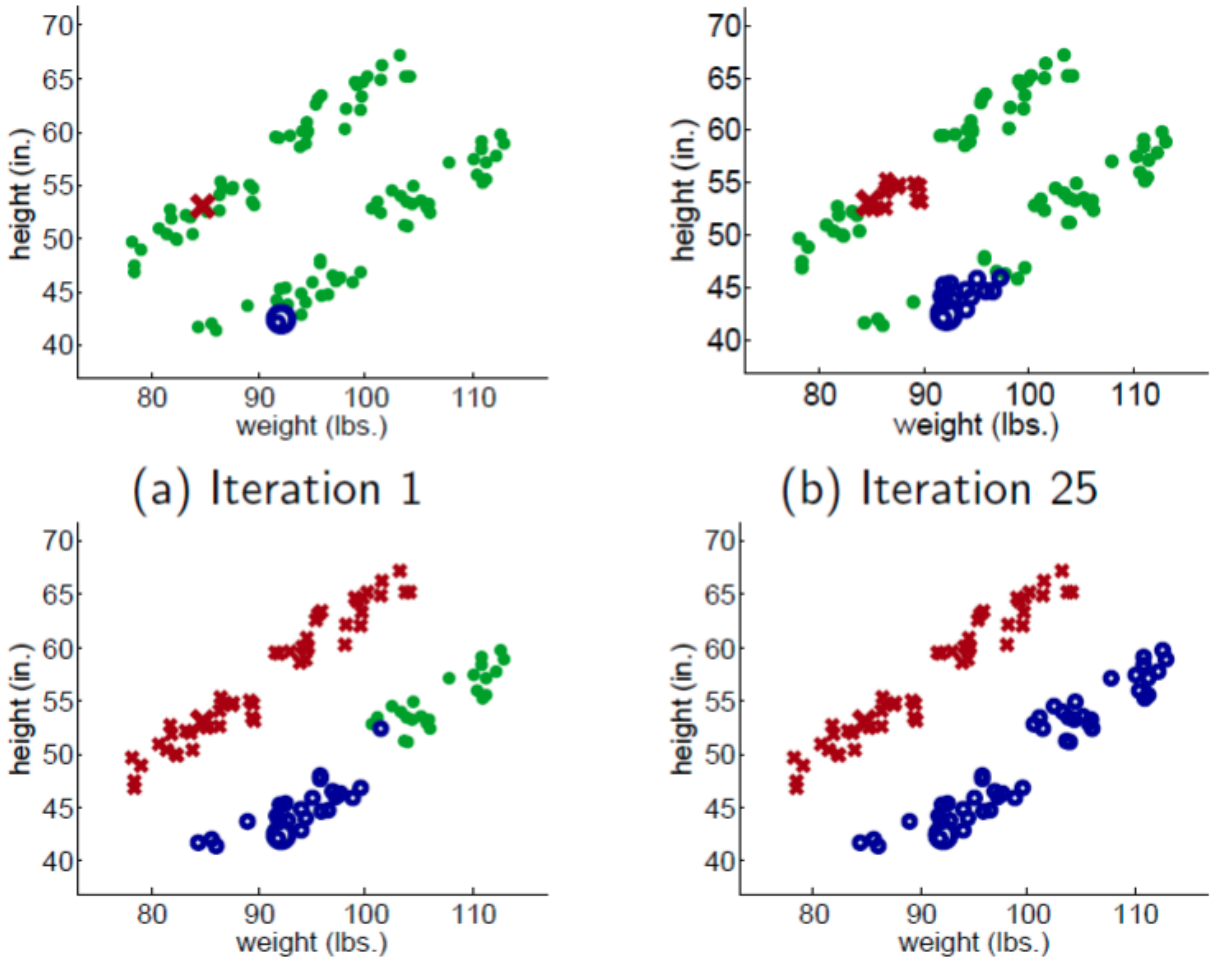
Propagating 1NN



(a) Iteration 1

(b) Iteration 25

Pop Quiz

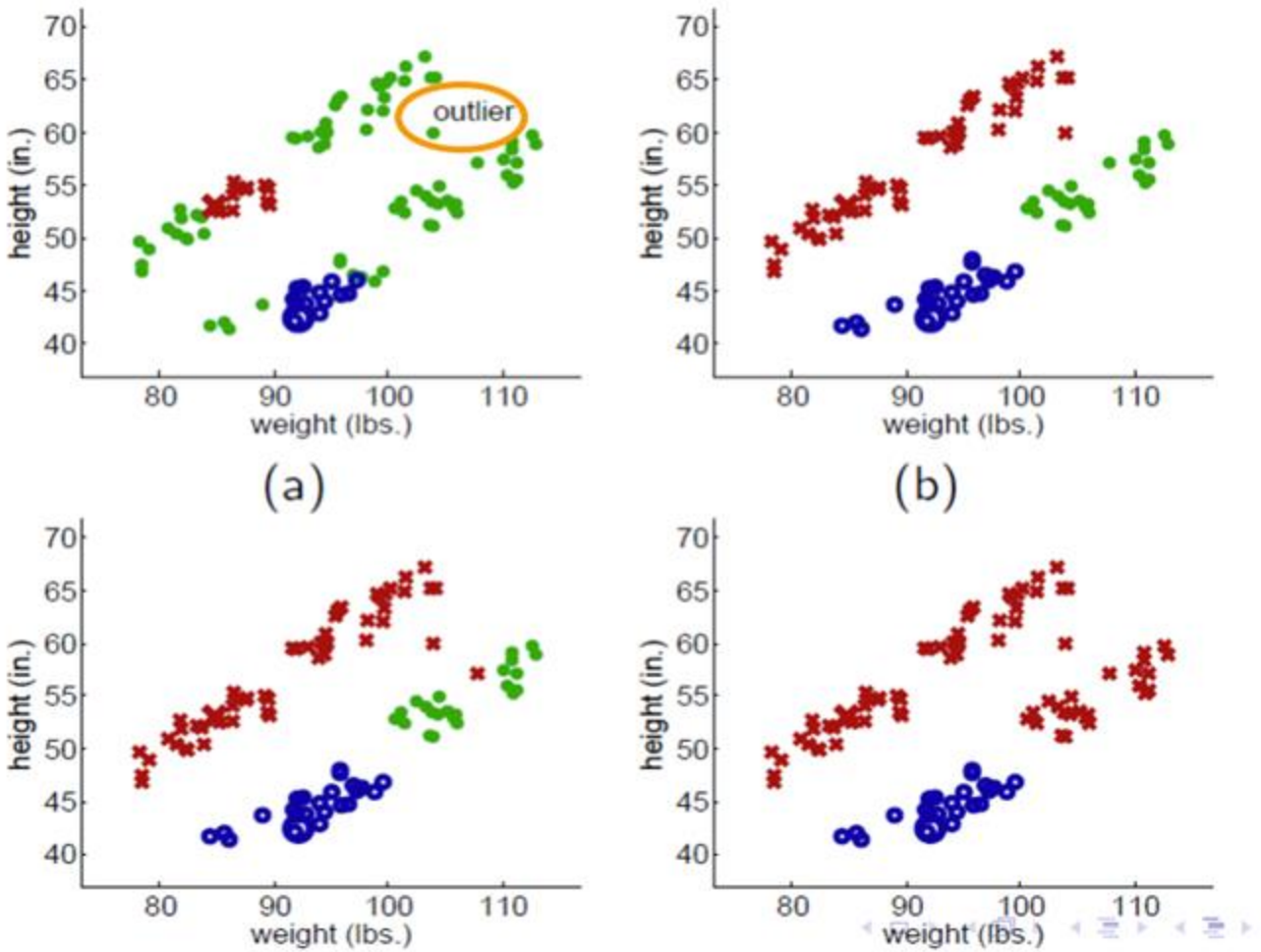


(a) Iteration 1

(b) Iteration 25

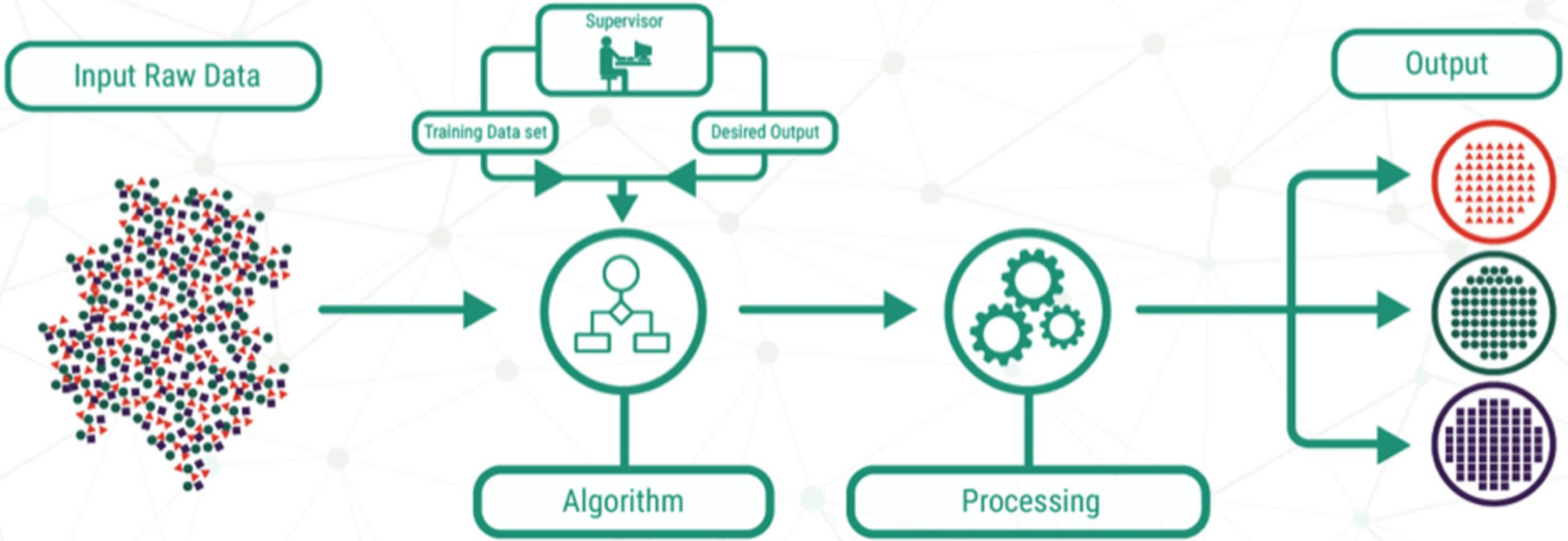
When the above won't work?

Propagating 1NN

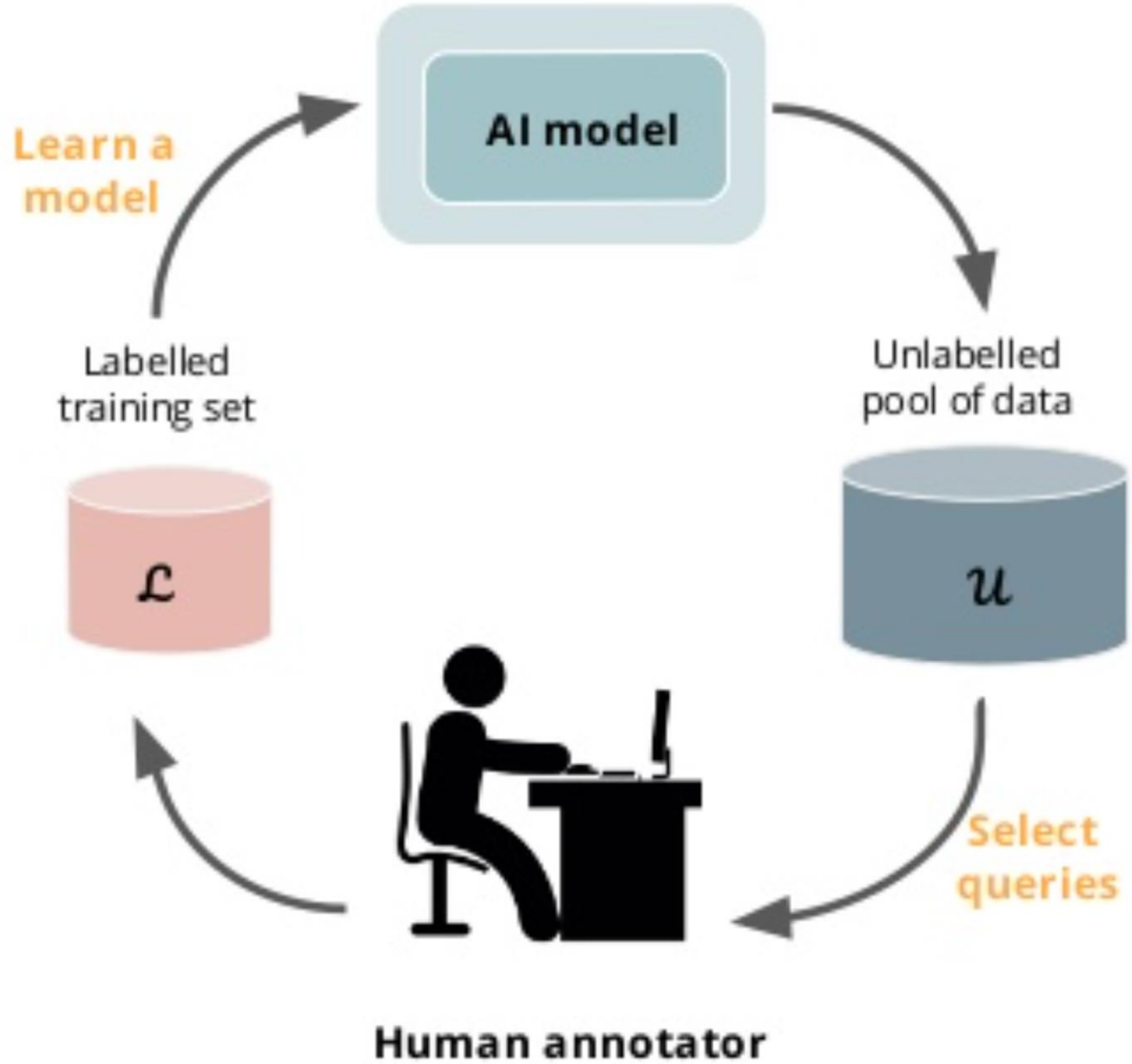


Active learning with human in the loop

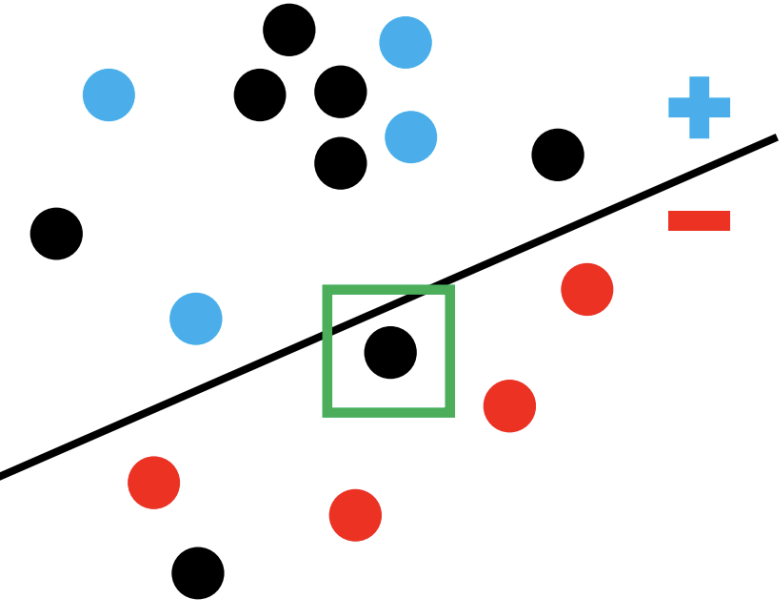
Fully supervised paradigm



Active learning



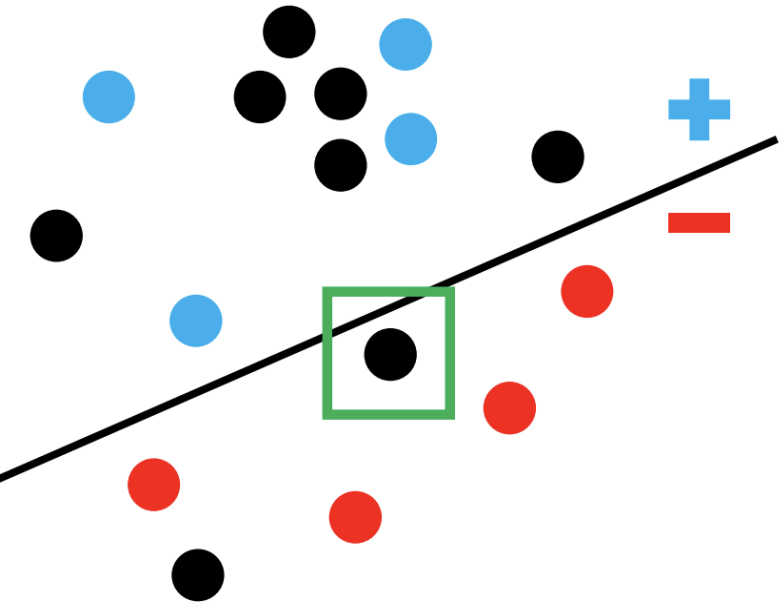
Active learning



(a) Uncertainty sampling
e.g. [Tong and Koller, 2002;
Kapoor et al., 2010]

[Tong 2002, Kappor 2010, Jain 2009, Kovashka 11, Vijayanarasimhan 2009, Branson 2010]

Active learning



(a) Uncertainty sampling
e.g. [Tong and Koller, 2002; Kapoor et al., 2010]

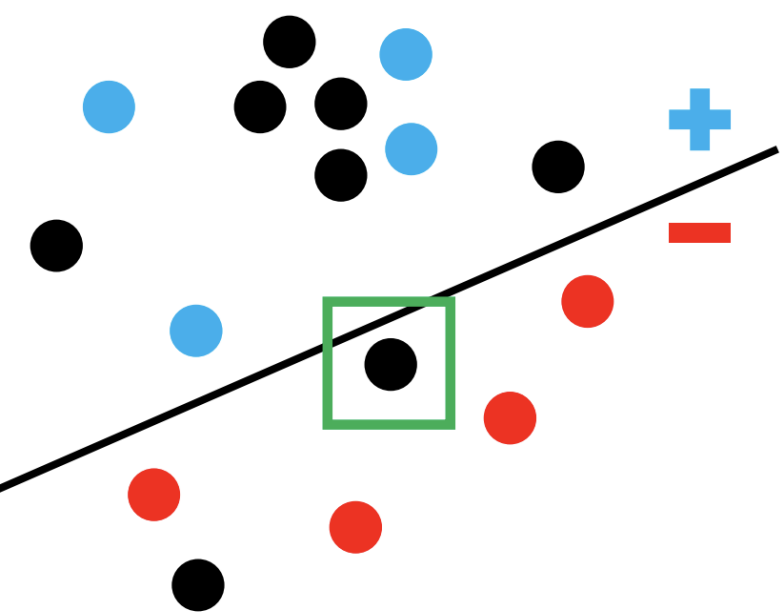
- Very efficient
- Labeling does not guarantee that will improve uncertainty on all images
- Annotated set is specific to one classifier

Alternatives

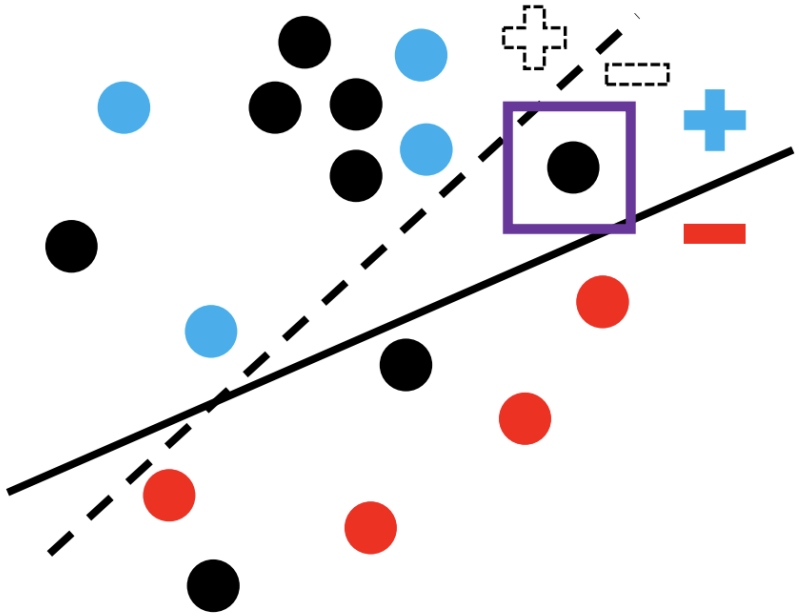
- Misclassification risk
- Measure expected entropy

[Tong 2002, Kappor 2010, Jain 2009, Kovashka 11, Vijayanarasimhan 2009, Branson 2010]

Active learning



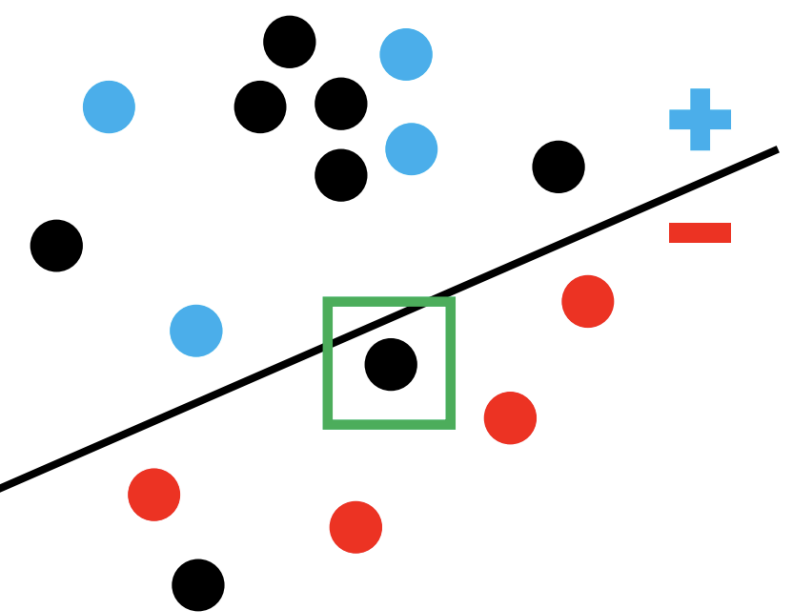
(a) Uncertainty sampling
e.g. [Tong and Koller, 2002; Kapoor et al., 2010]



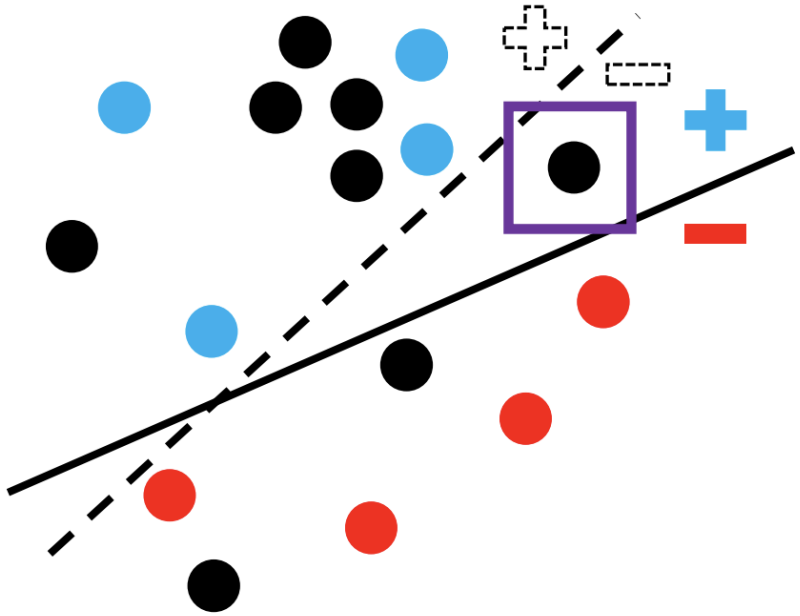
(b) Query by committee
e.g. [Seung et al., 1992; Loy et al., 2012]

[Tong 2002, Kappor 2010, Jain 2009, Kovashka 11, Vijayanarasimhan 2009, Branson 2010]

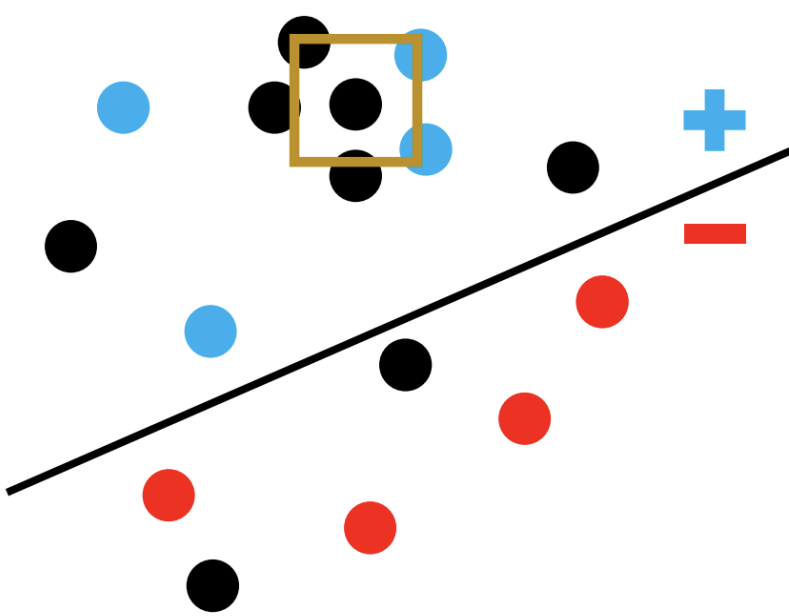
Active learning



(a) Uncertainty sampling
e.g. [Tong and Koller, 2002; Kapoor et al., 2010]



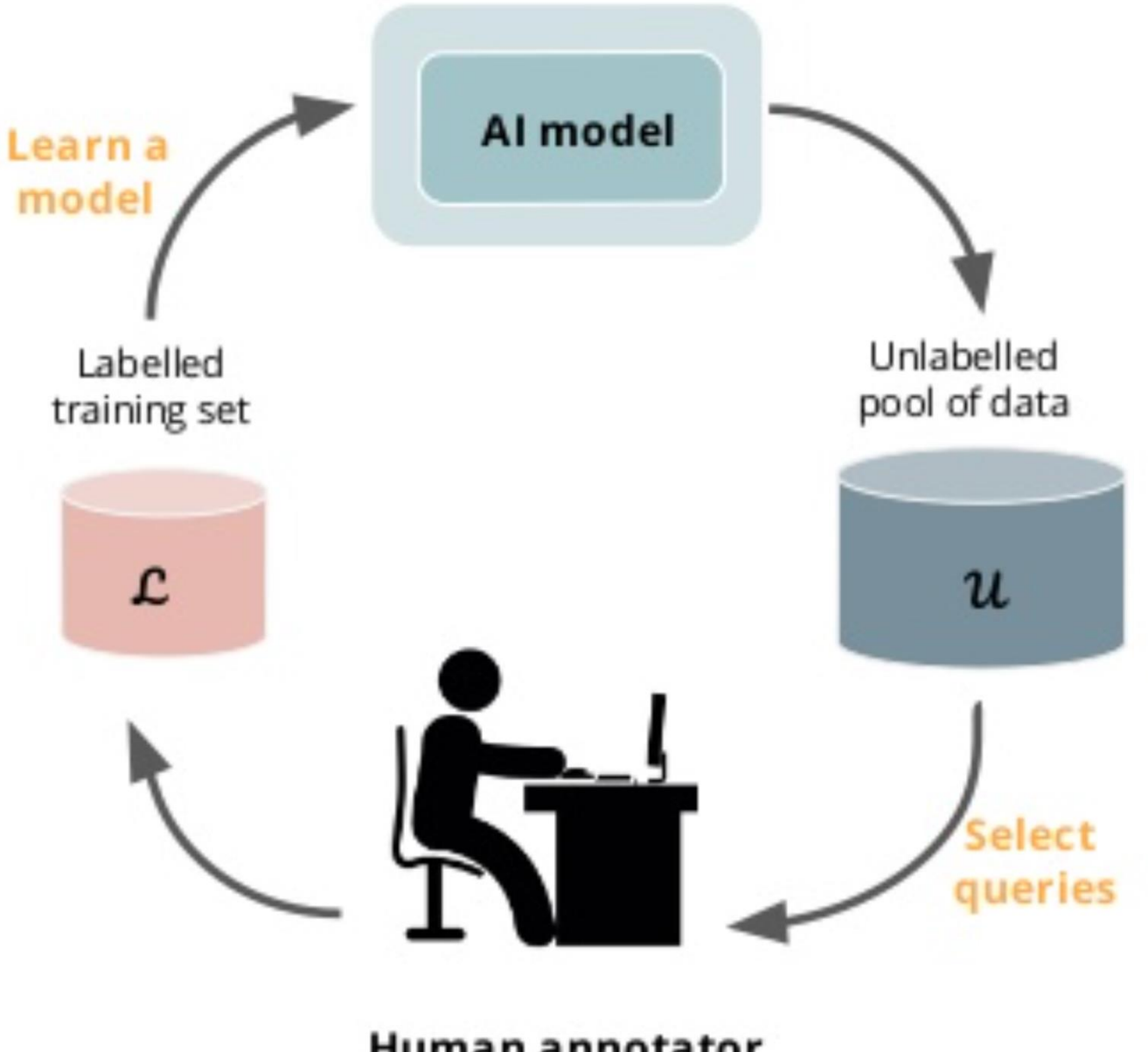
(b) Query by committee
e.g. [Seung et al., 1992; Loy et al., 2012]

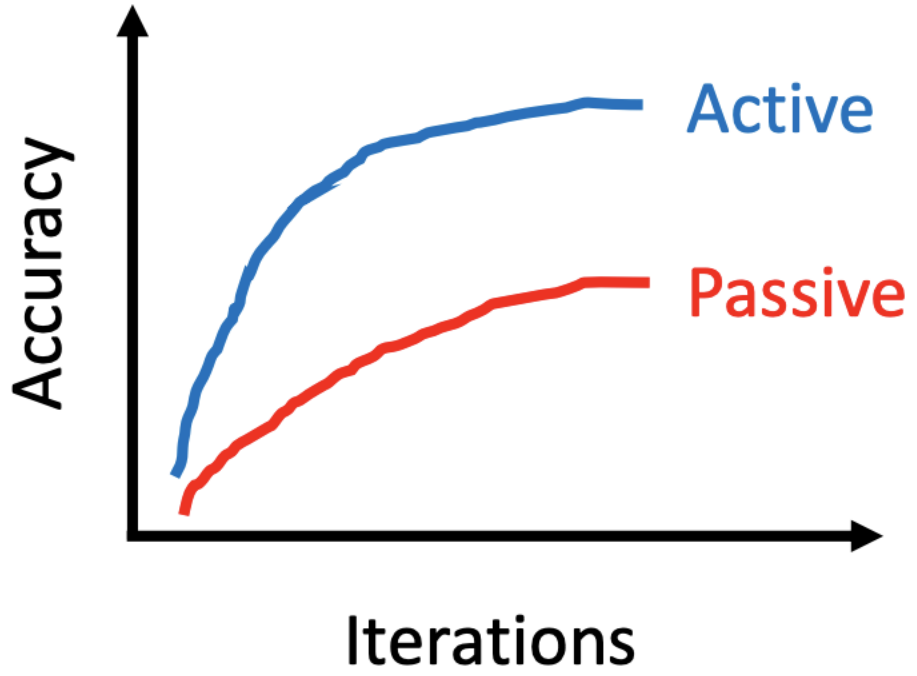


(c) Sampling from dense region
e.g. [Li and Guo, 2013]

[Tong 2002, Kappor 2010, Jain 2009, Kovashka 11, Vijayanarasimhan 2009, Branson 2010]

Active learning



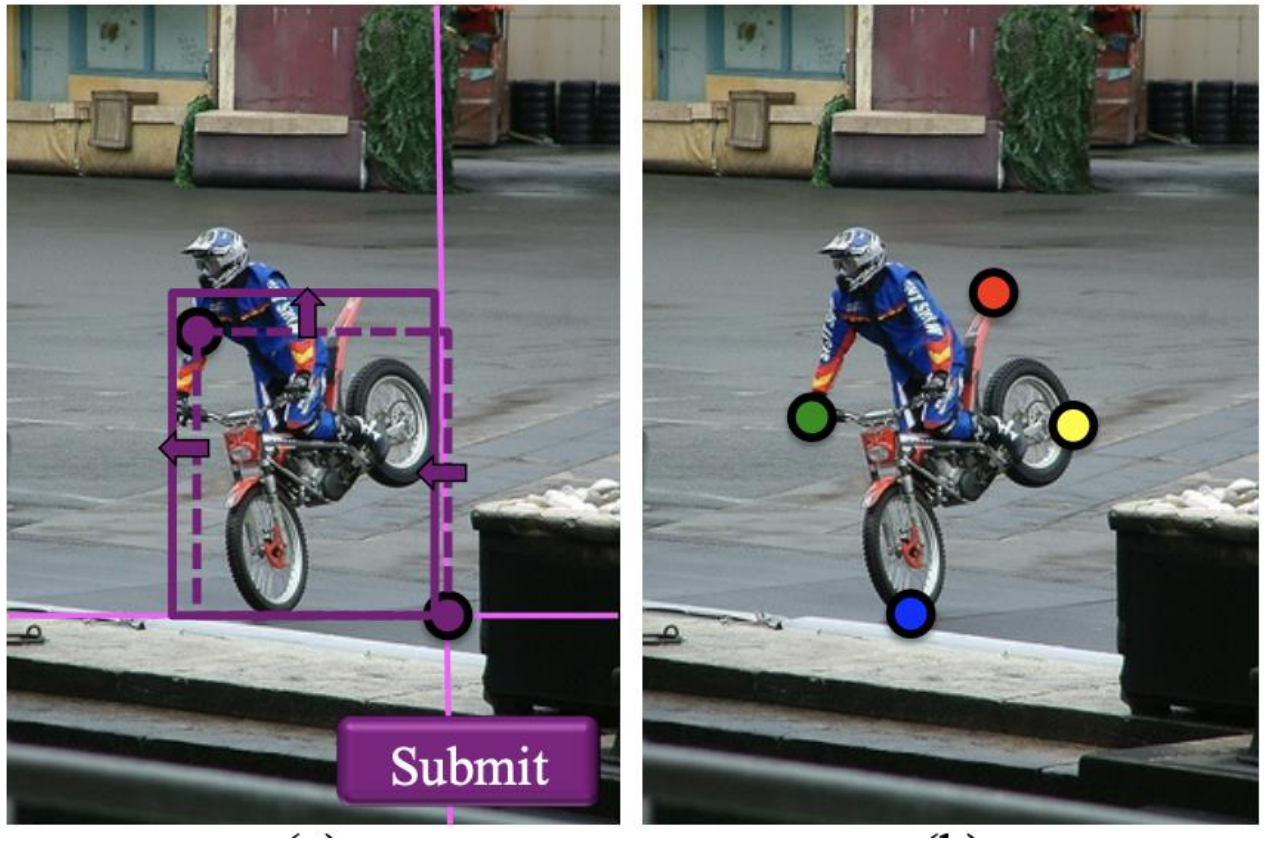


- Active vs. Passive:
In practice the gap is very small
Selection is specific to one classifier

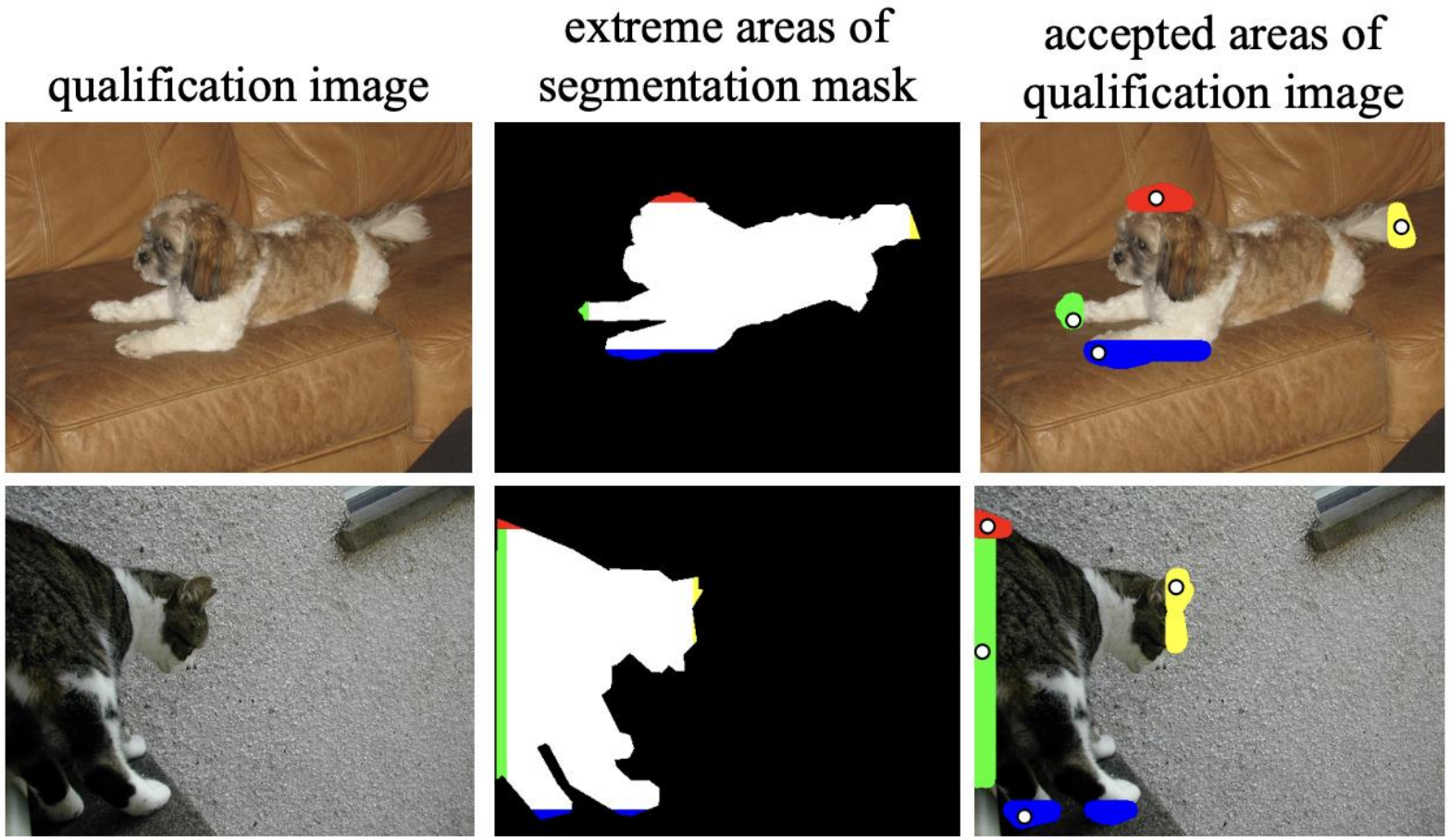
- Generalization (?):
Selection is based on one specific model

Active learning and annotation types

Clicking



Clicking


qualification image
extreme areas of segmentation mask
accepted areas of qualification image

Clicking

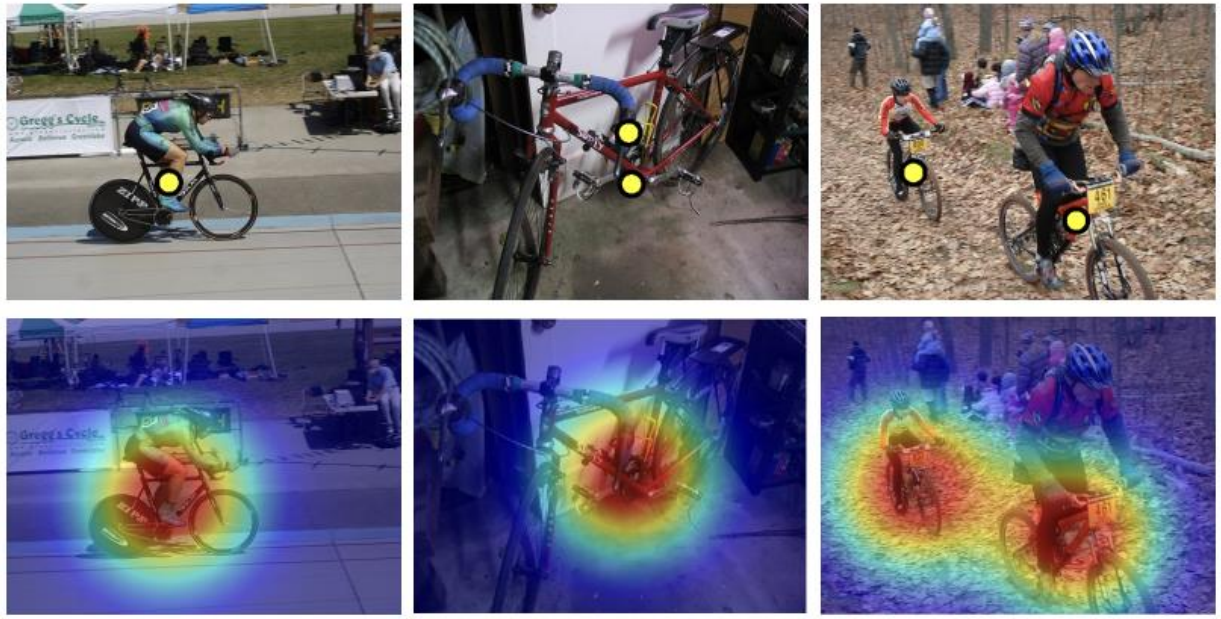
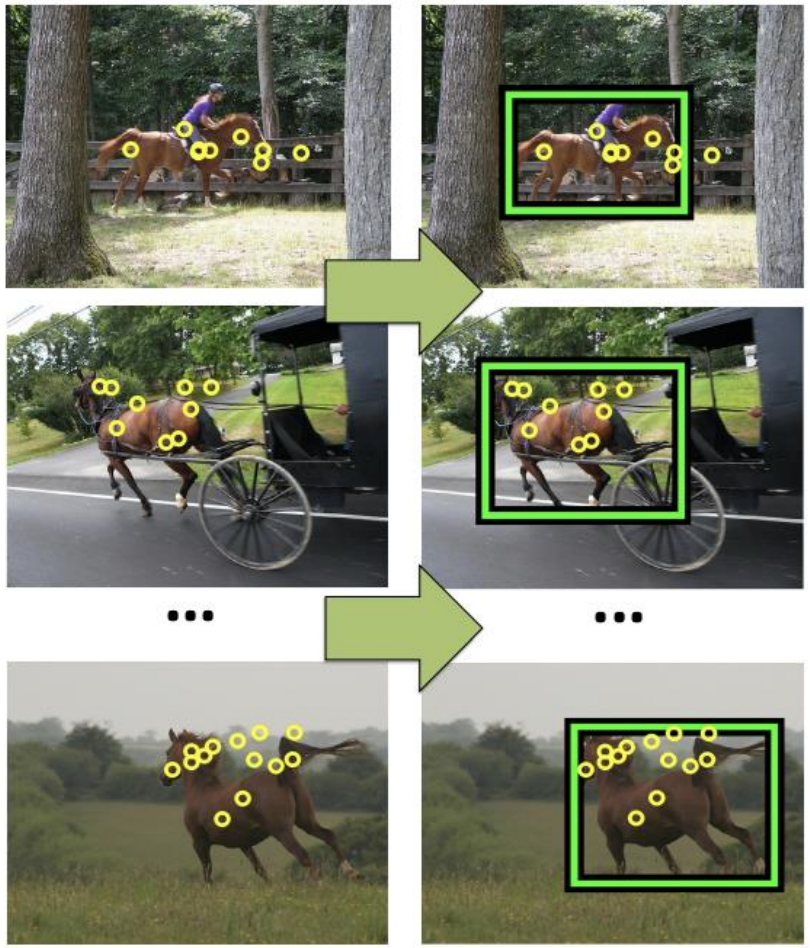


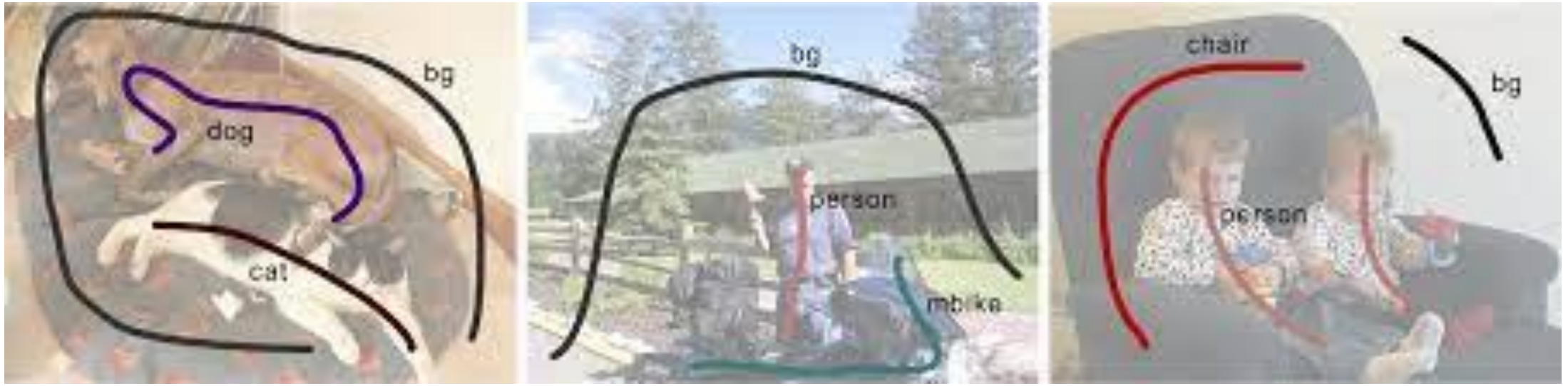
Figure 5. **Box center score** S_{bc} on bicycle examples. (left): One-click annotation. (middle): Two-click annotation on the same instance. (right): Two-click annotation on different instances. The values of each pixel in the heatmaps give the S_{bc} of an object proposal centered at that pixel.

Eye-tracking



- annotation time (1s per image)
- reduce annotation time by 6.8x
- simple annotation guidelines
- correct localizations in half images

Scribbles



More references

Actively selecting between different types of annotations (which images to label and at what level to label them)
e.g. [Vijayanarasimhan CVPR 2011]

Selecting batches of labels
e.g. [Liang 2014, Jain 2010, Vijayanarasimhan 2014]

Transfer learning
e.g. [Gavves 2015, Johns 2015]

Interactive annotation
e.g. [Branson ICCV 2011, Yao CVPR2012 ,Kovashka CVPR 2012, Wah CVPR 2013, Lad ECCV 2014,
Rubinstein ECCV 2012, Russakovsky CVPR 2015, Jain CVPR2016, Nagaraja ICCV2015, Castrejon CVPR 2017, Acuna CVPR 2018, Benenson CVPR 2019, Ling CVPR 2019]

Today's lecture

Part I: Learning without labels

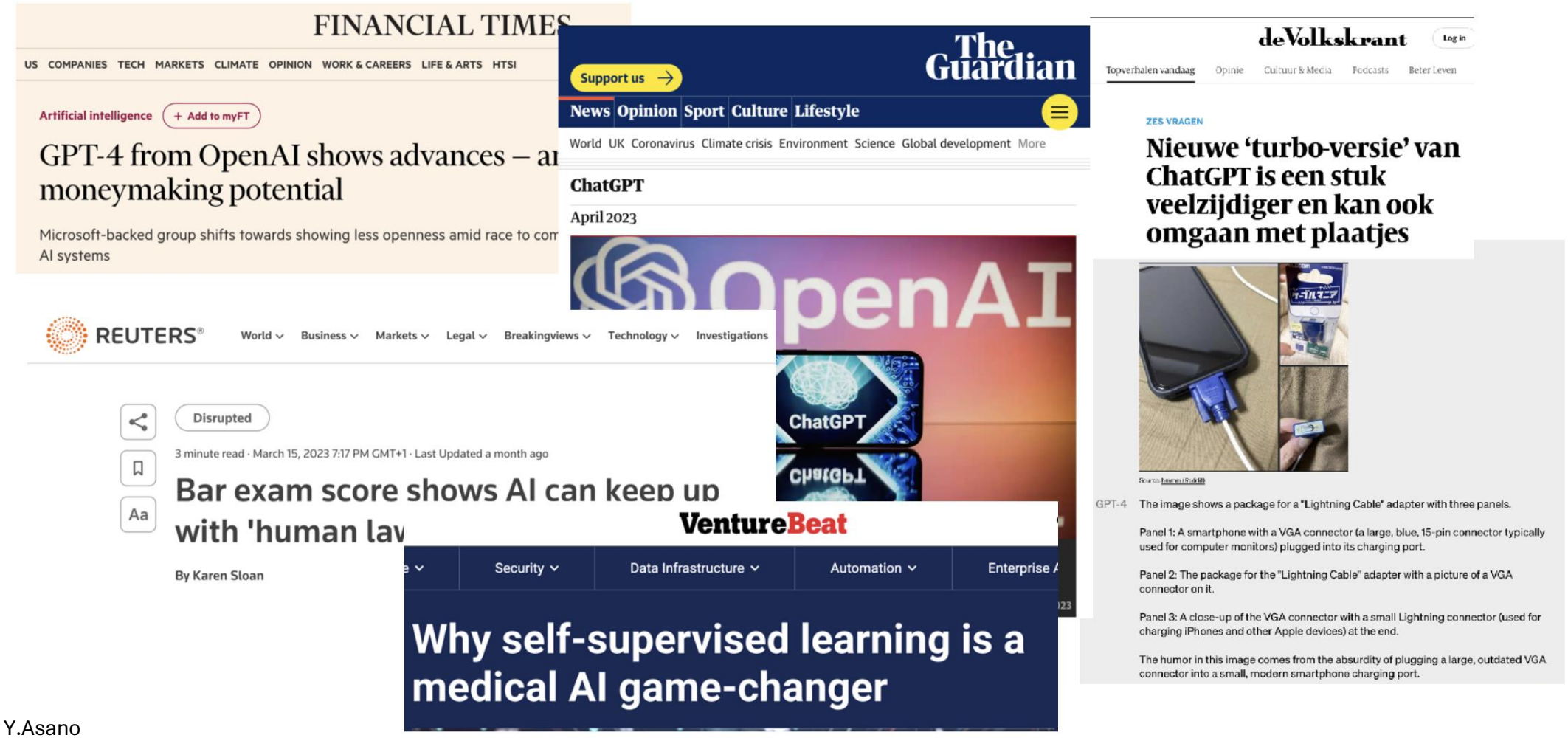
Part II: Self-Supervised Learning

Part III: Multimodal Self-Supervised Learning

Part II: Outline
Self-Supervised Learning (SSL)

- Definition
- Motivation
- Benefits
- Methods
- Contrastive learning
- Masked Modeling
- Examples

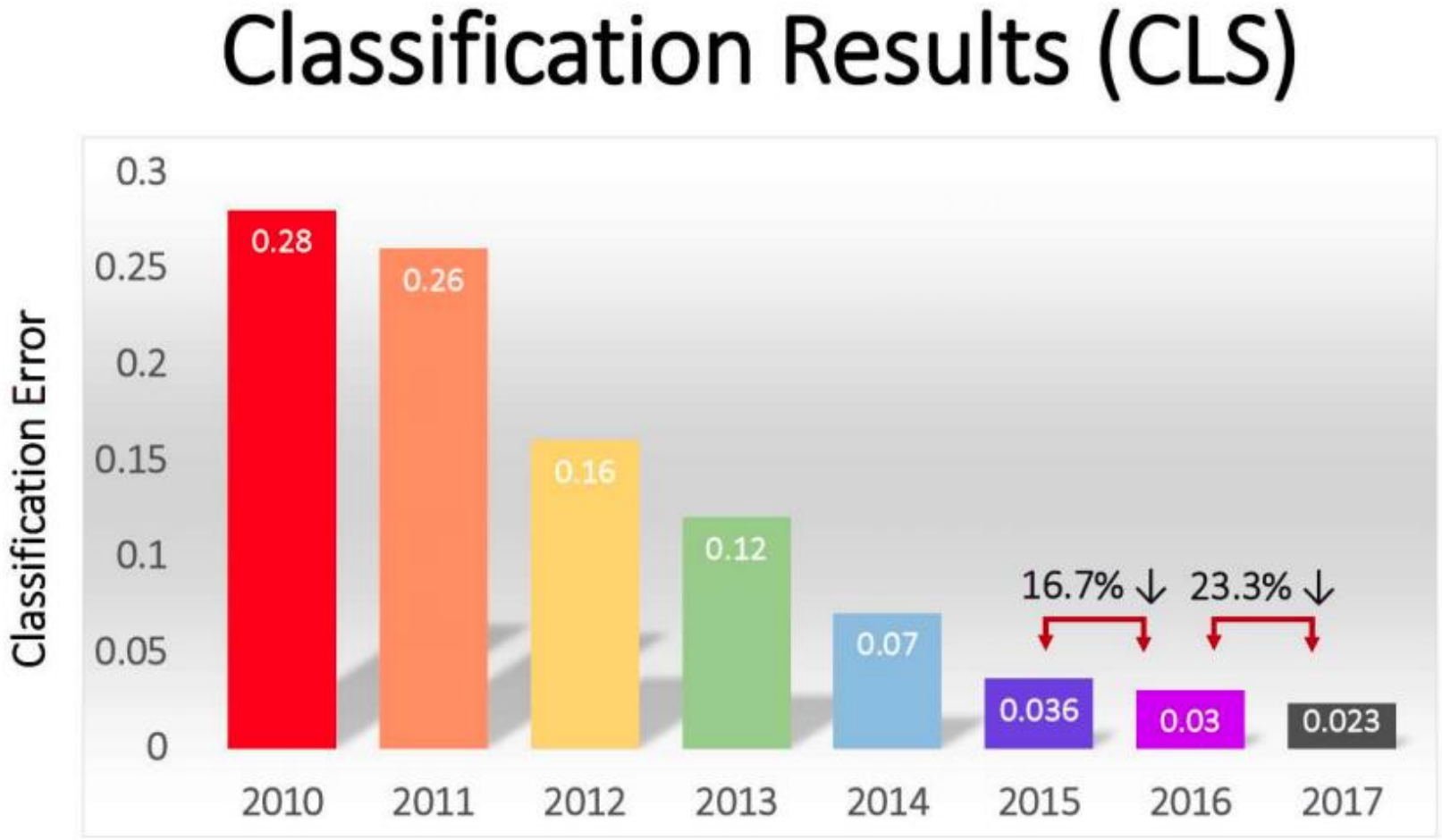
Multimodal Self-supervised learning is everywhere



Slide credit: Y.Asano

Self-Supervised Learning

The ImageNet Challenge Story



The ImageNet Challenge Story

- Strong supervision

- Features from networks trained on ImageNet can be used for other visual tasks, e.g. detection, segmentation, action recognition, fine grained visual classification

- To some extent, *any* visual task can be solved now by:
  - Construct a large-scale dataset labelled for that task
  - Specify a training loss and neural network architecture
  - Train the network and deploy

- What else can we do?

Self-supervised Learning Motivation

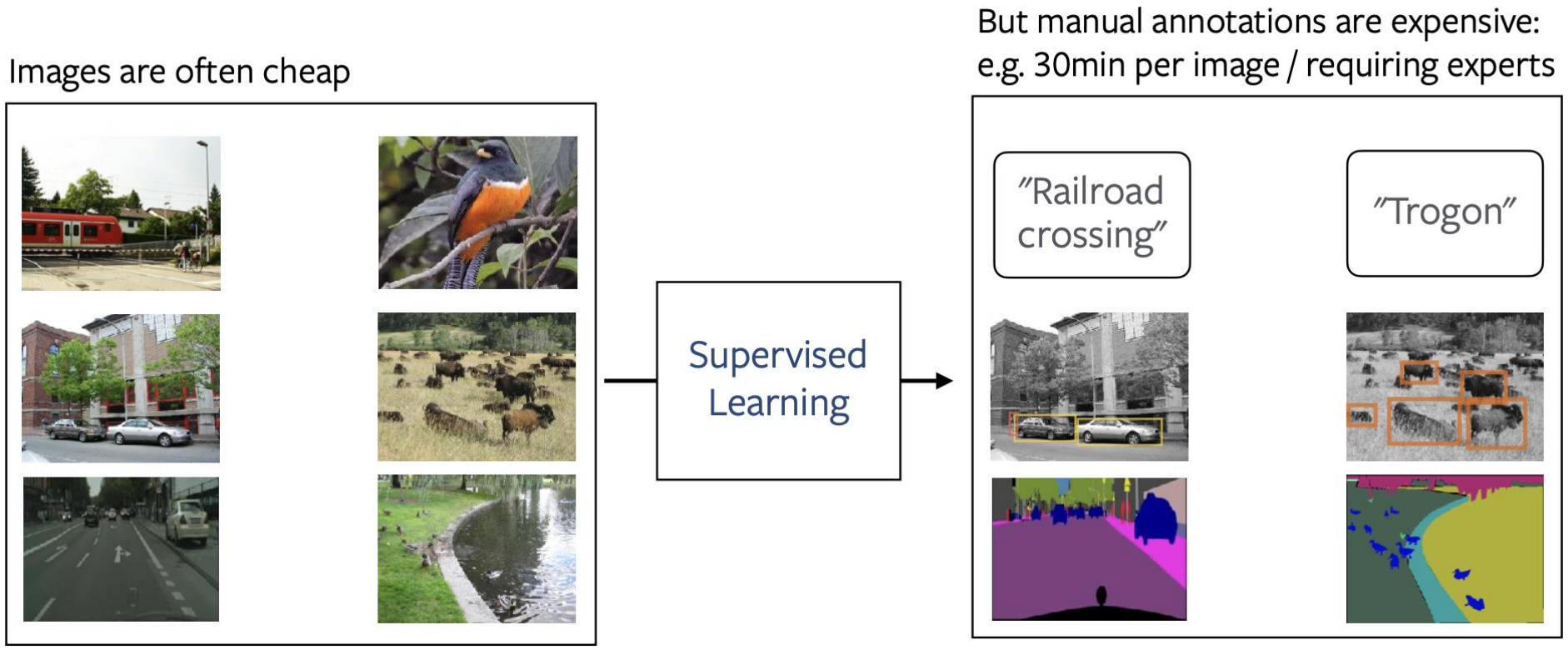
+ Expense of producing a new dataset for each new task

+ Some areas are supervision-starved, e.g. medical data, where it is hard to obtain annotations

+ Availability of vast numbers of unlabelled images/videos
  - Facebook: one billion images uploaded per day
  - 300 hours of video are uploaded to YouTube every minute

+ How infants may learn …

Self-supervised Learning Motivation



Slide credit: Y.Asano

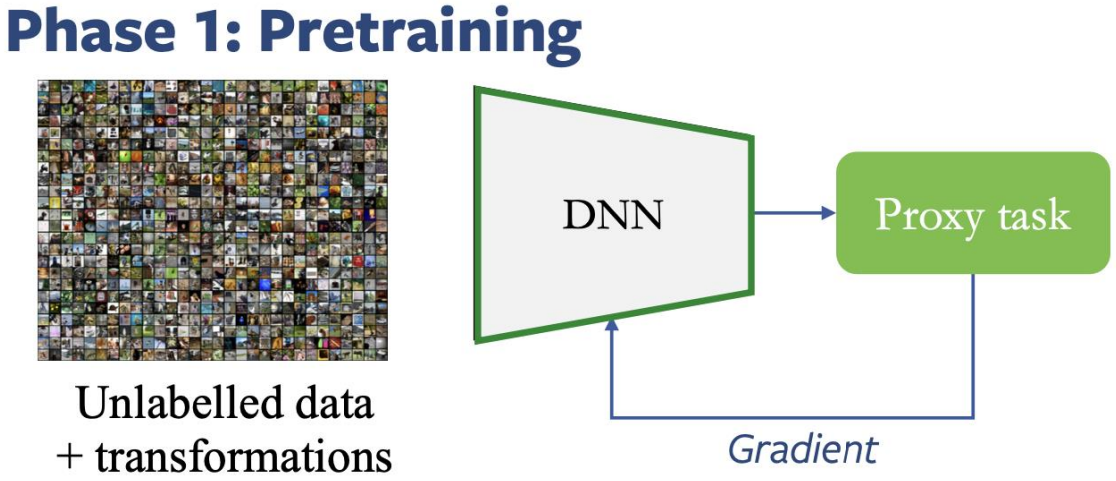
Self-supervised Learning Motivation

Slide credit: Y.Asano

Self-supervised Learning Definition

- A form of unsupervised learning where the data provides the supervision
- In general, withhold some part of the data, and task the network with predicting it
- The task defines a proxy loss, and the network is forced to learn what we really care about, e.g. a semantic representation, in order to solve it

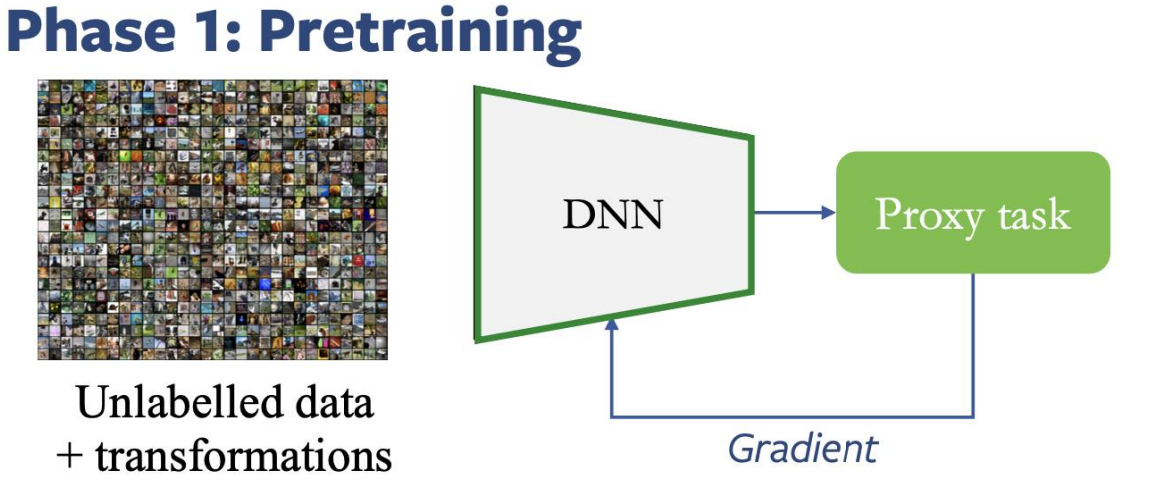
Self-supervised Learning Procedure



Types:
- Geometry based
- Clustering
- Contrastive
- Generative (partial/full)
- (more)

Slide credit: Y.Asano

Self-supervised Learning Procedure

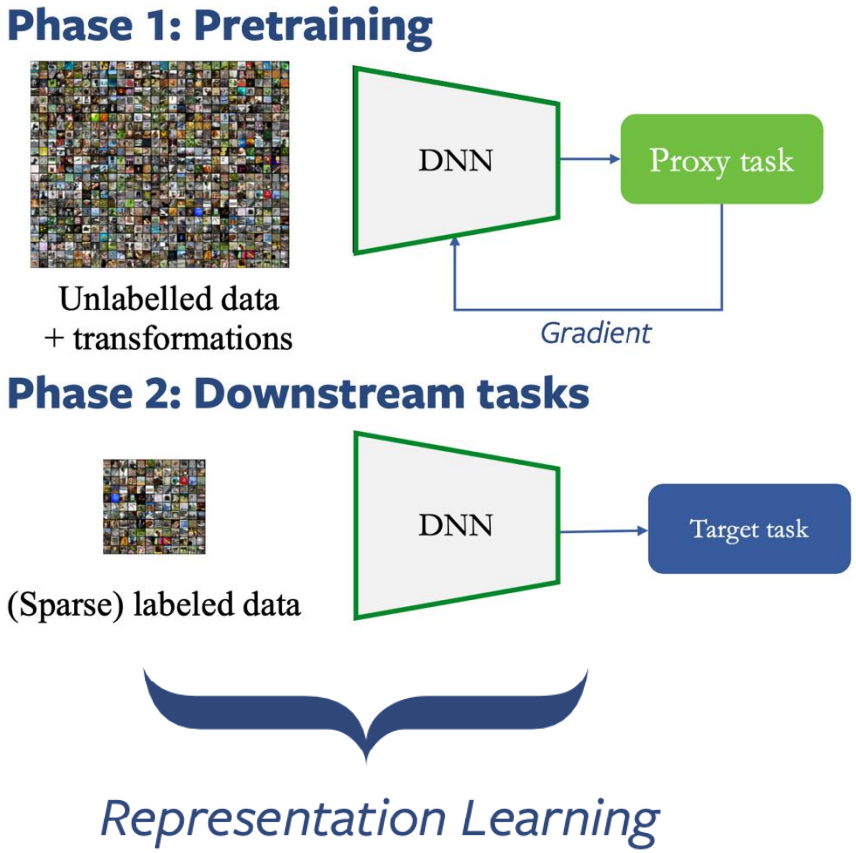


Types:
- Geometry based
- Clustering
- Contrastive
- Generative (partial/full)
- (more)

Types:
- Limited fine-tuning (e.g. linear layer)
- Finetuning (w/ full or fraction of datase

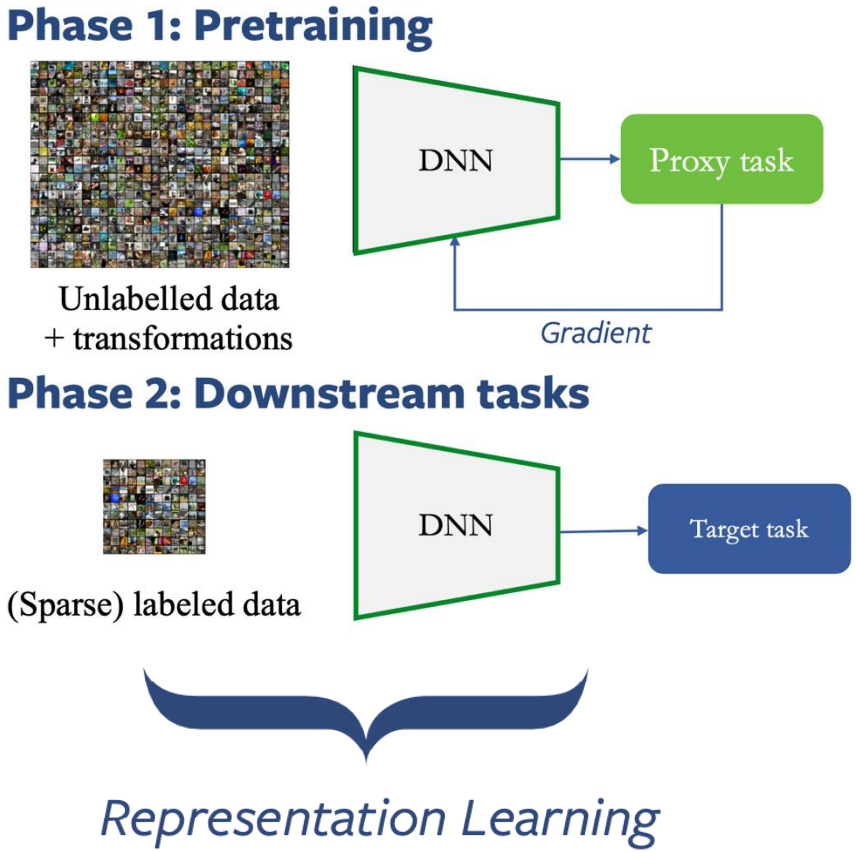
Slide credit: Y.Asano

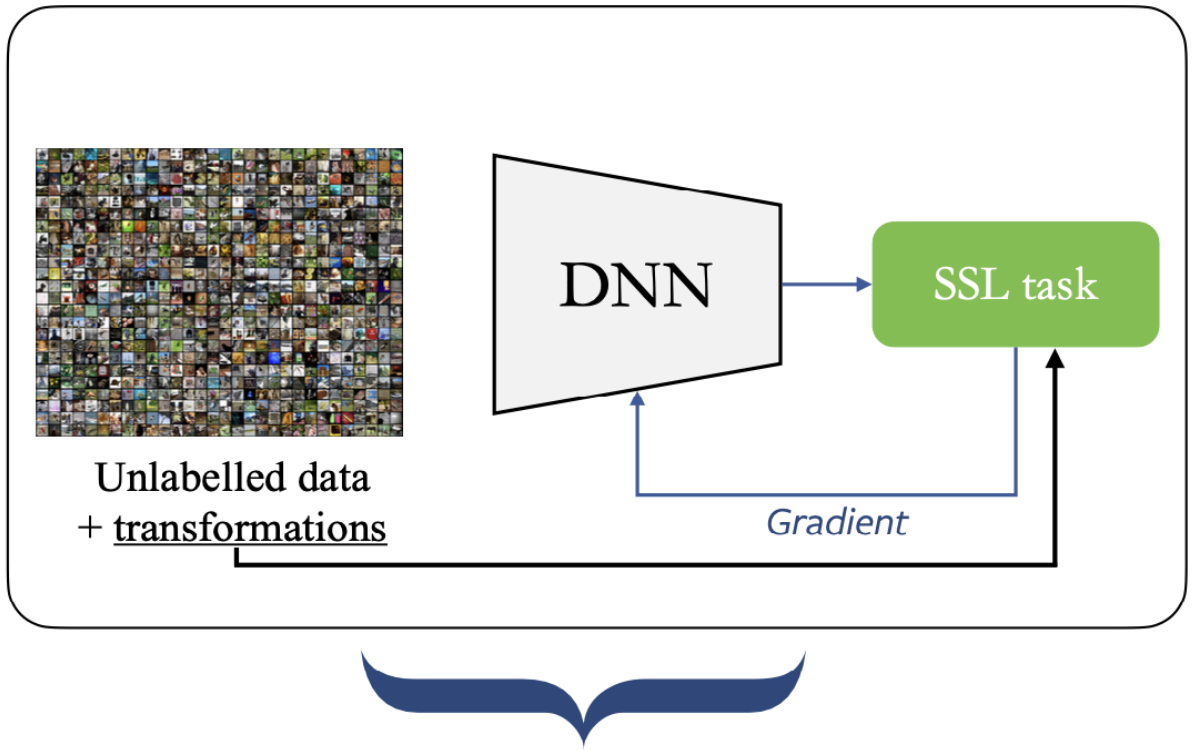
Self-supervised Learning Procedure



Slide credit: Y.Asano

Self-supervised Learning Procedure





Useful Self-supervised Learning, e.g.

SSL object detection & segmentation

SSL speaker detection, SSL dataset labelling etc..

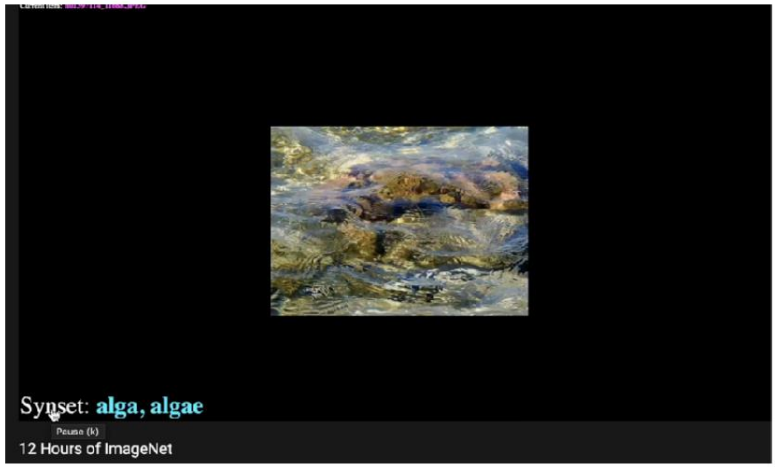
10

Slide credit: Y.Asano

Why SSL?

Why SSL?
1. Scalability

(above) x 50 =1.2M images



90ms * 1.2M = 30h

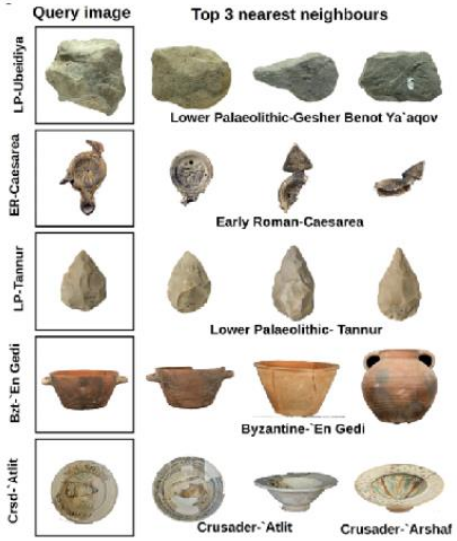
Instagram: >50B images

Annotation is expensive, yet datasets keep getting bigger

Why SSL?
2. Changing domains

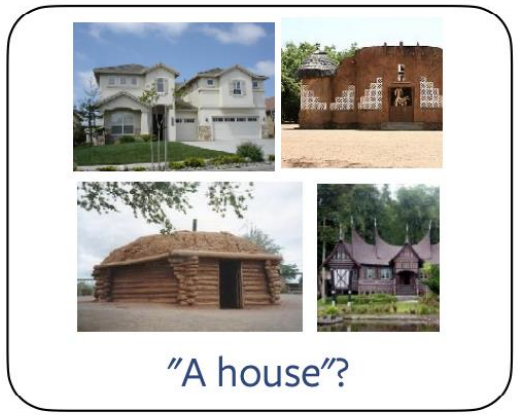
Unclear when & what to relabel. Again, large costs just to "keep up"

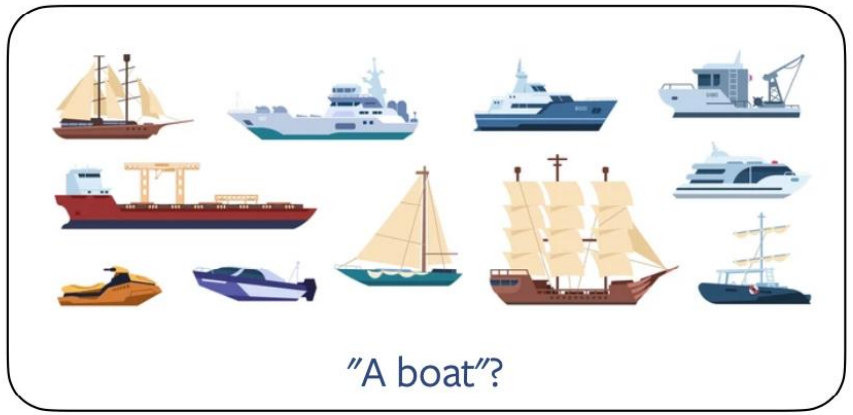
Why SSL?
3. Accessibility & generalisability

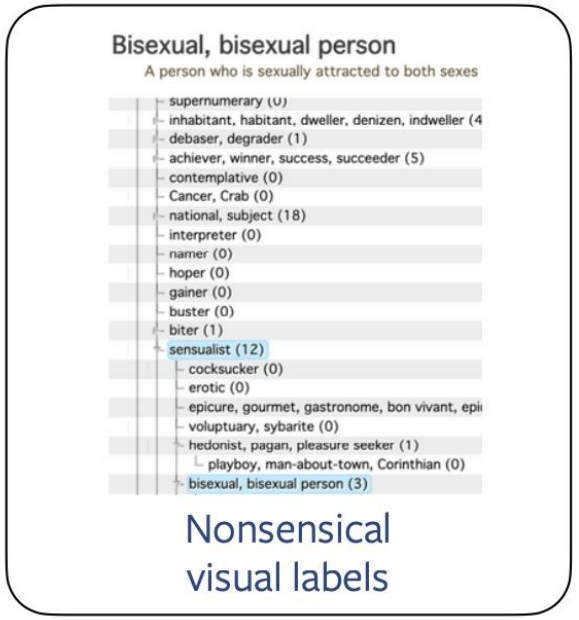


Pretrained models are very useful for a variety of tasks.

Why SSL?
4. Ambiguity of labels

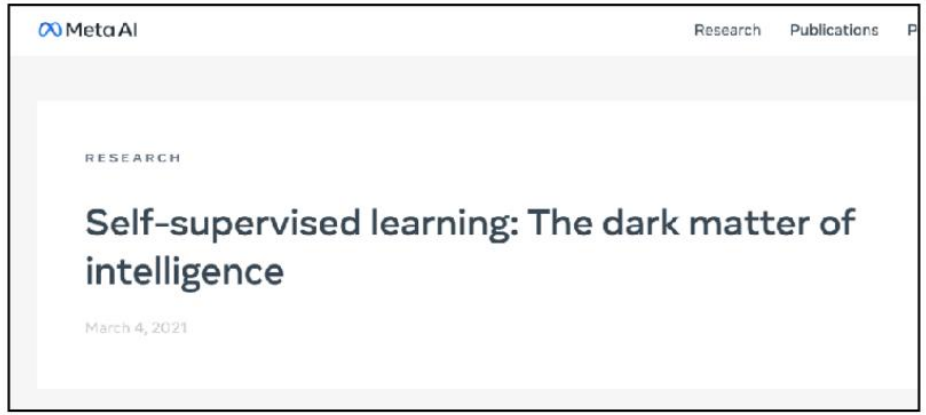


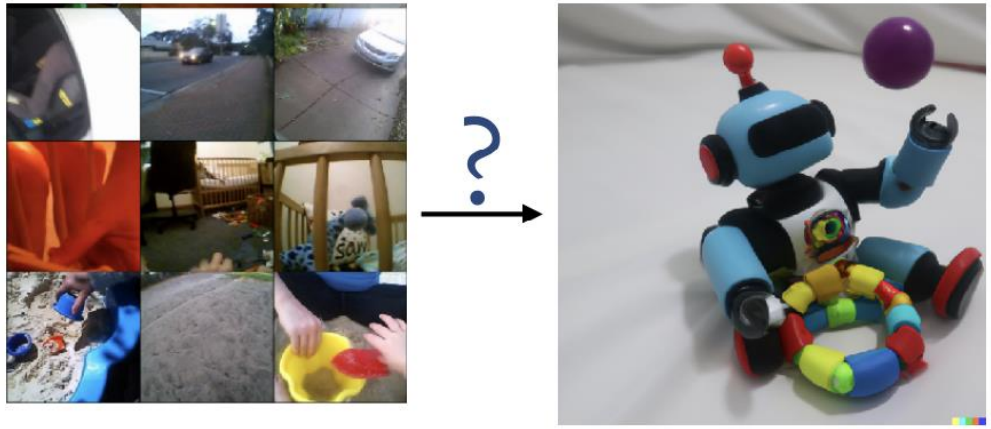




Labels are ambiguous at best, discriminating and bias-propagating at worst.
Do we really wish to provide our models with these priors?

Why SSL?
5. Investigating the fundamentals of visual understanding





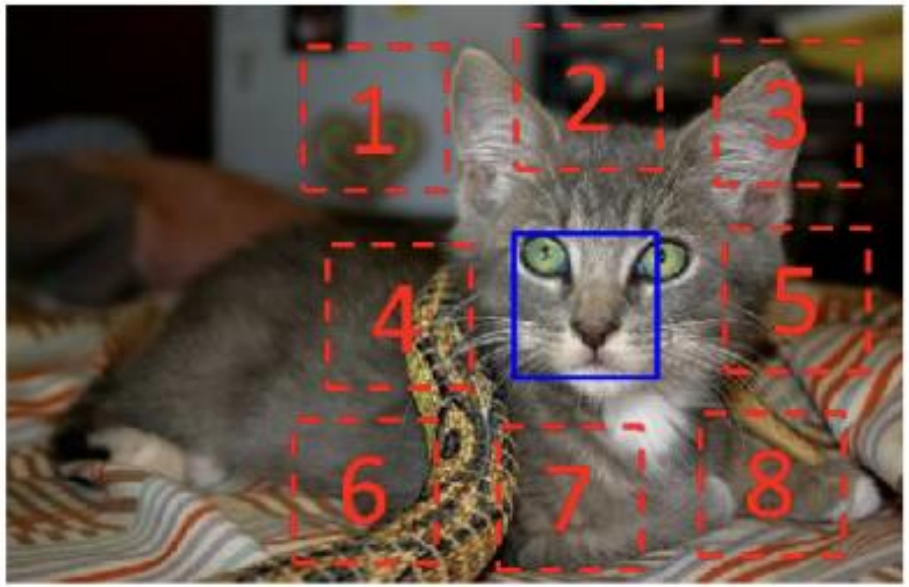
What, if there are, are the limits of learning without labels?

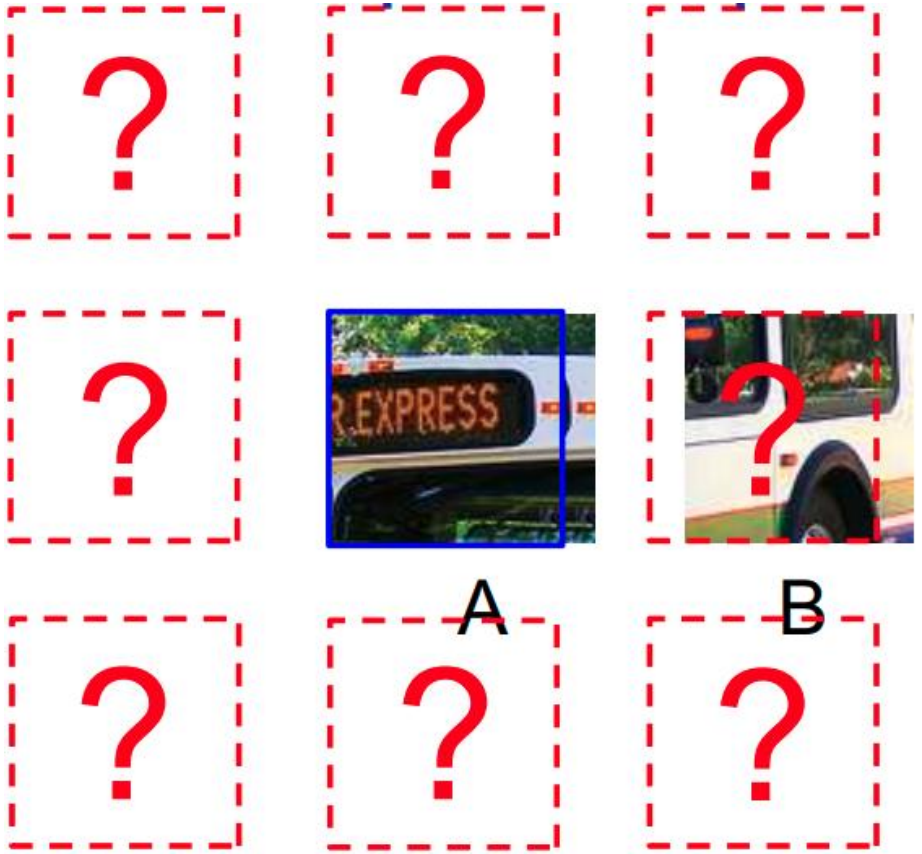
Pop Quiz

Can you think of examples?

Methods for SSL

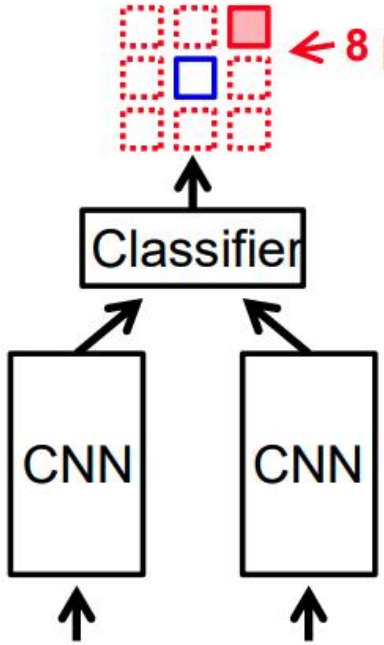
Self-supervised Learning
Relative positioning → Context prediction

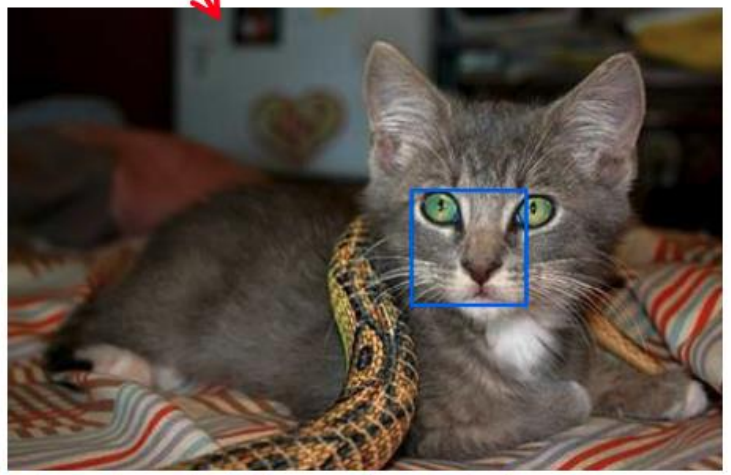




Self-supervised Learning
Relative positioning → Context prediction

Train network to predict relative position of two regions in the same image



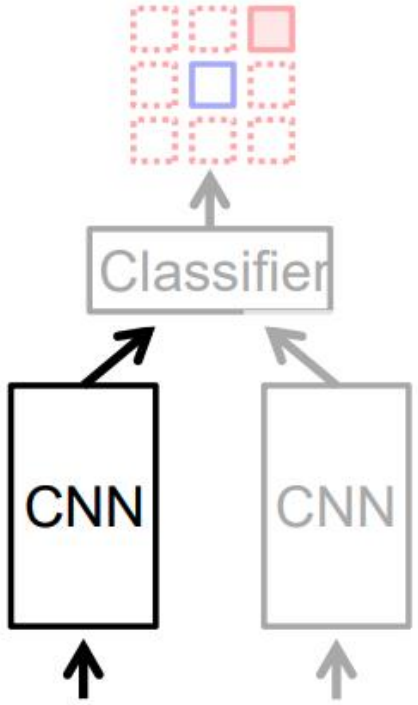


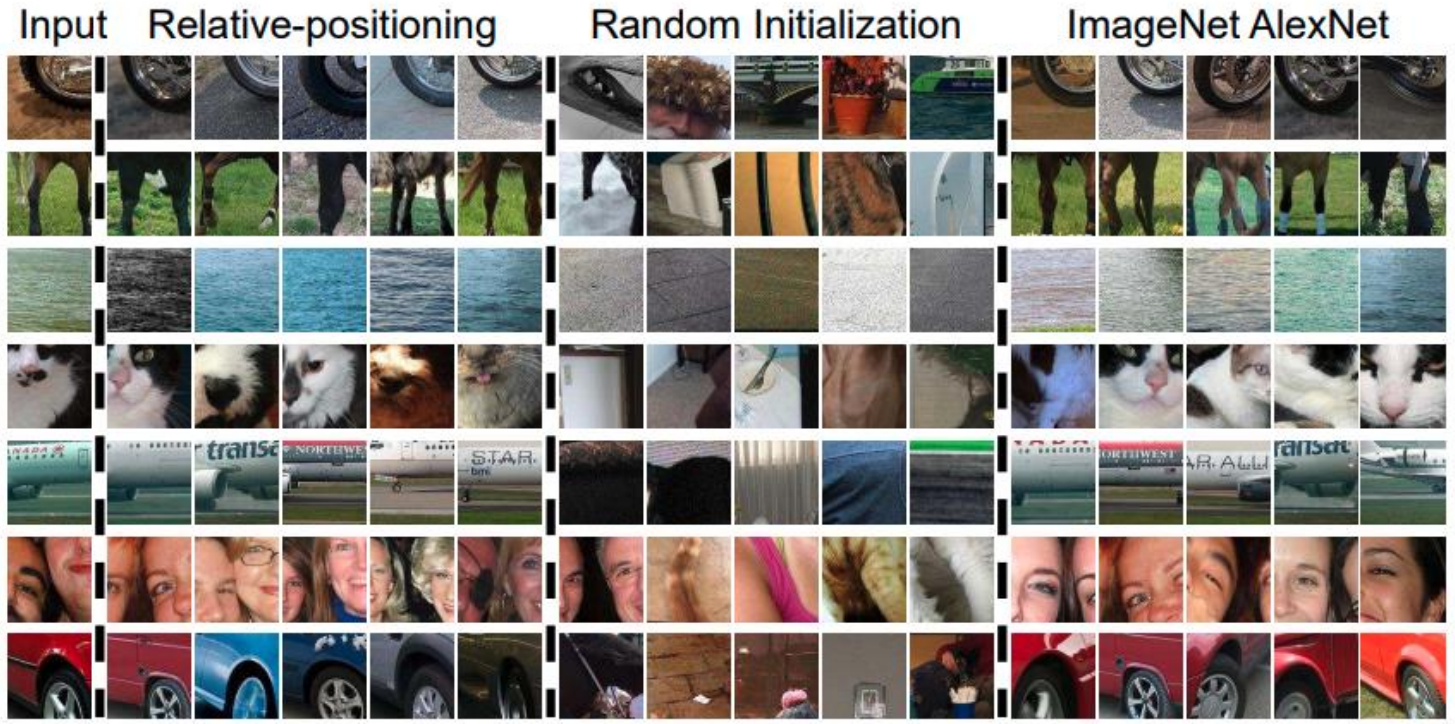
Randomly Sample Patch
Sample Second Patch

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

Self-supervised Learning
Relative positioning → Context prediction

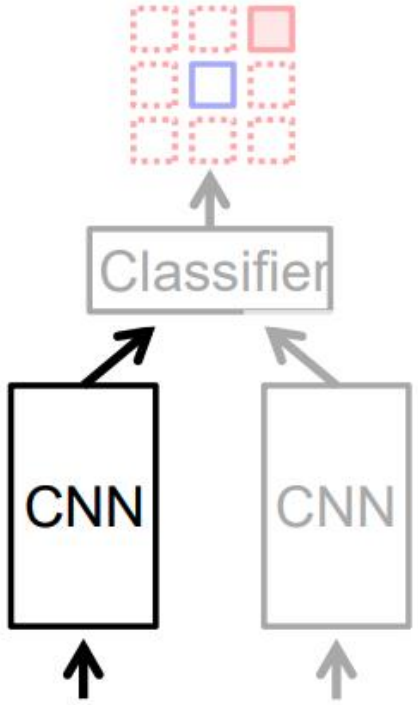
What is the network learning?

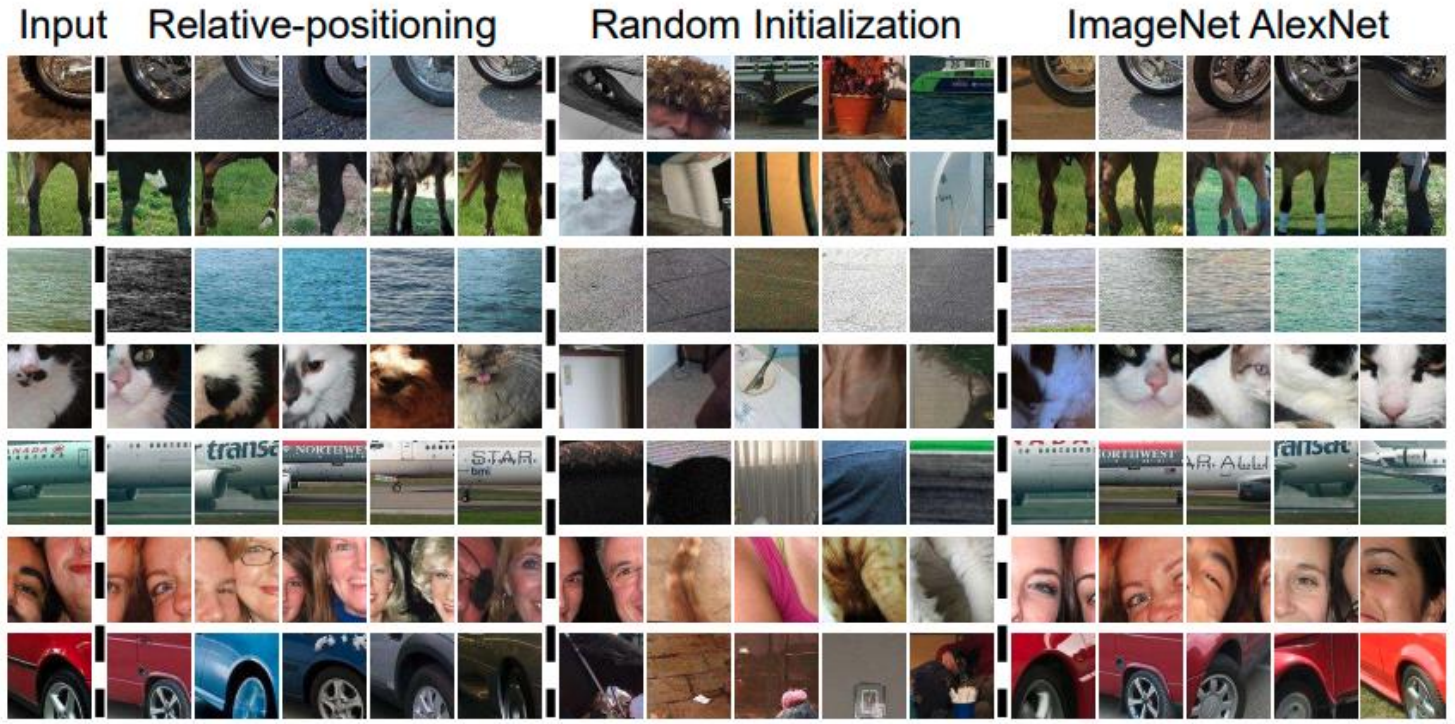




Self-supervised Learning
Relative positioning → Context prediction

What is the network learning?





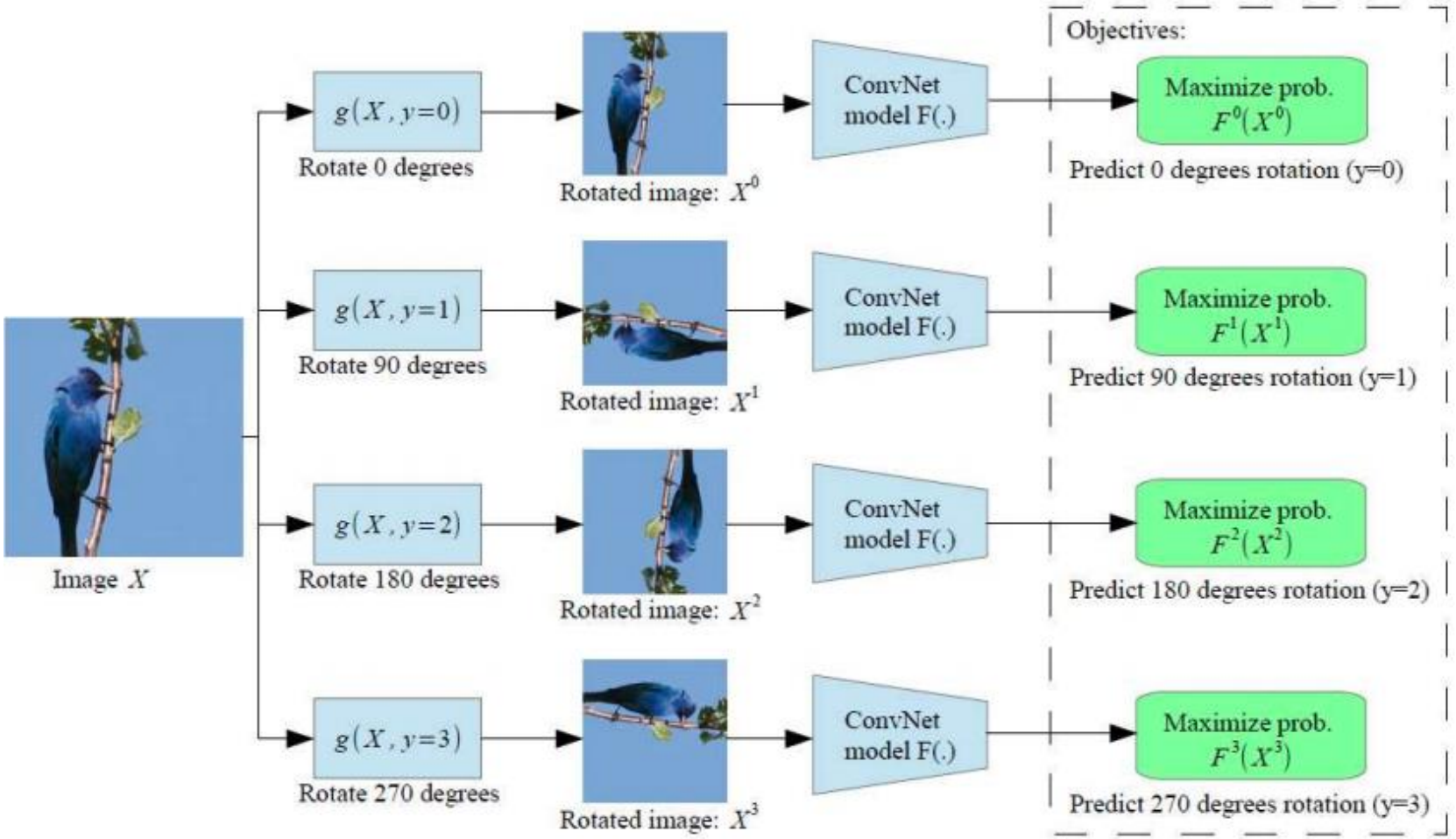
Self-supervised Learning
Context Encoder → Context prediction

(a) Input context

(c) Context Encoder
($L2$ loss)

https://www.researchgate.net/figure/Word2Vec-CBOW-and-Skip-gram-There-are-two-different-methods-in-the-Word2Vec-algorithm_fig2_320829283
Doersch et al. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015.
Pathak et al. Context Encoders: Feature Learning by Inpainting. CVPR 2016.

Self-supervised Learning Image Transformations



Unsupervised Representation Learning by Predicting Image Rotations. Gidaris et al., ICLR 2018

Self-supervised Learning Image Transformations

But:

Unsupervised Representation Learning by Predicting Image Rotations. Gidaris et al., ICLR 2018

Self-supervised Learning Image Transformations

- AlexNet
- Closes gap between ImageNet and self-supervision

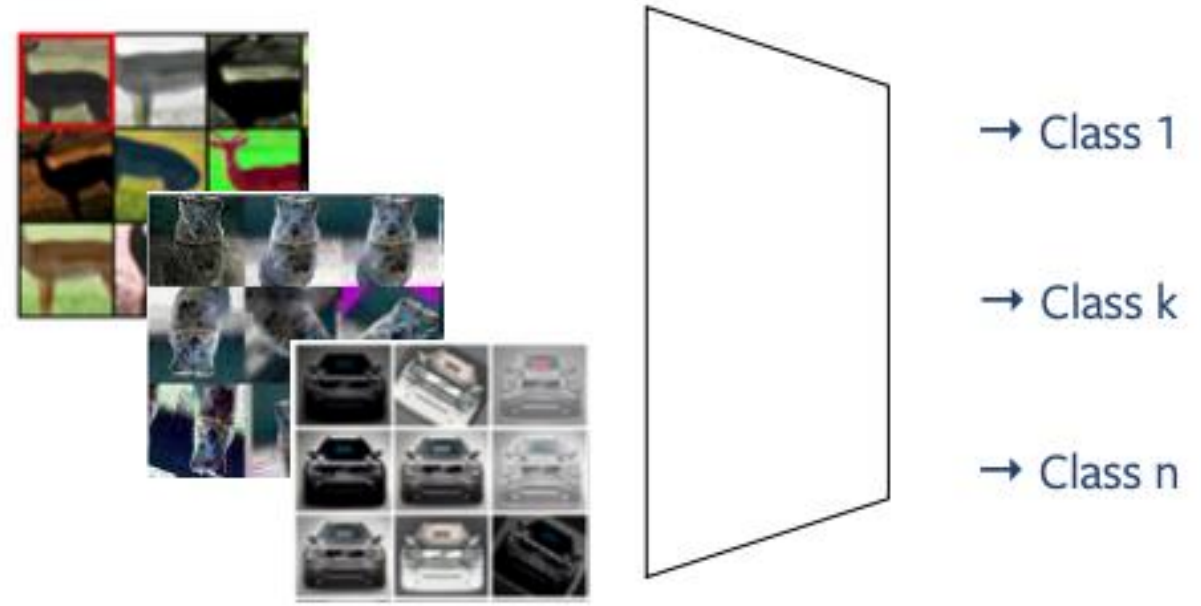
	PASCAL VOC Detection mAP
Random	43.4
Rel. Pos.	51.1
Colour	46.9
Rotation	54.4
ImageNet Labels	56.8

Unsupervised Representation Learning by Predicting Image Rotations. Gidaris et al., ICLR 2018

Contrastive learning

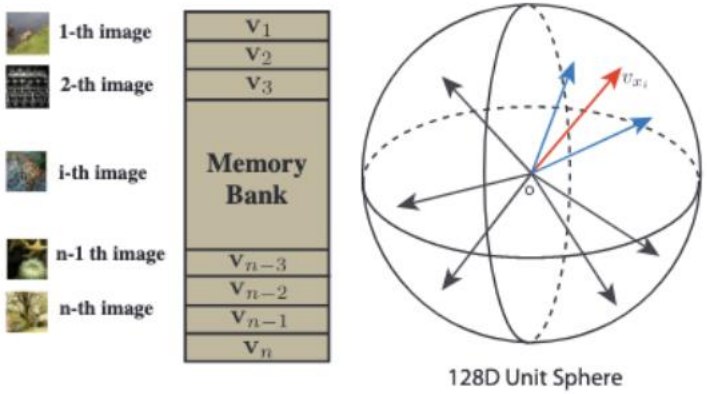
Self-supervised Learning
Image Uniqueness

- Exemplar CNN, precursor to contrastive learning

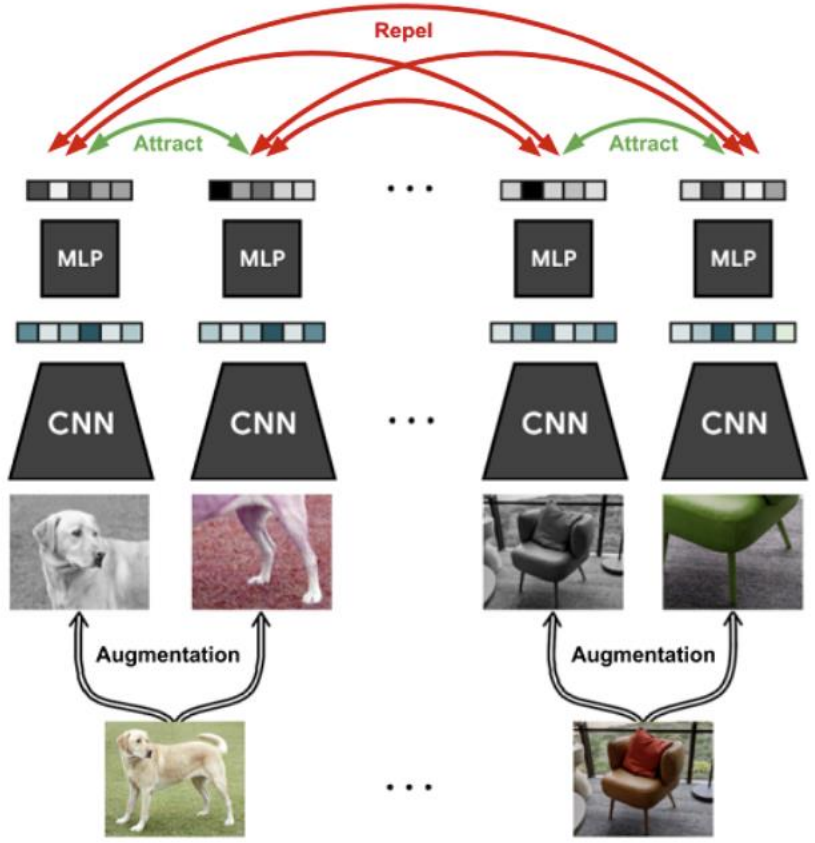


Uses image-uniqueness and enforces augmentation-invariance

Noise-contrastive self-supervised learning



NPID

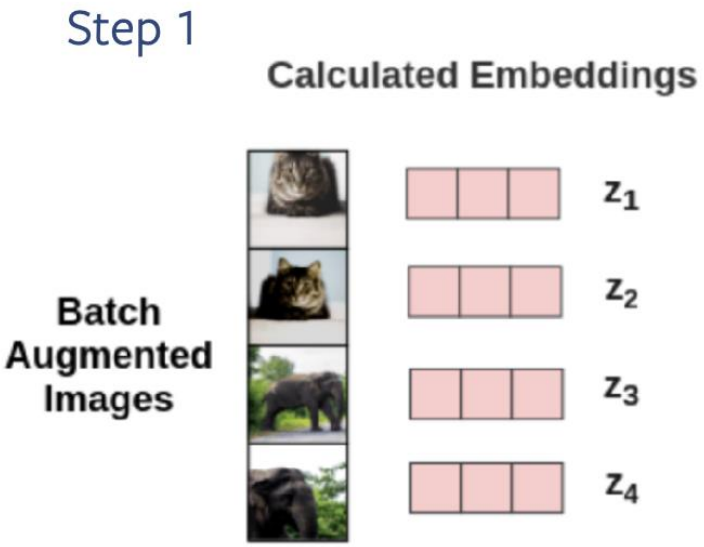


SimCLR

Wu et al. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. CVPR 2018

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR

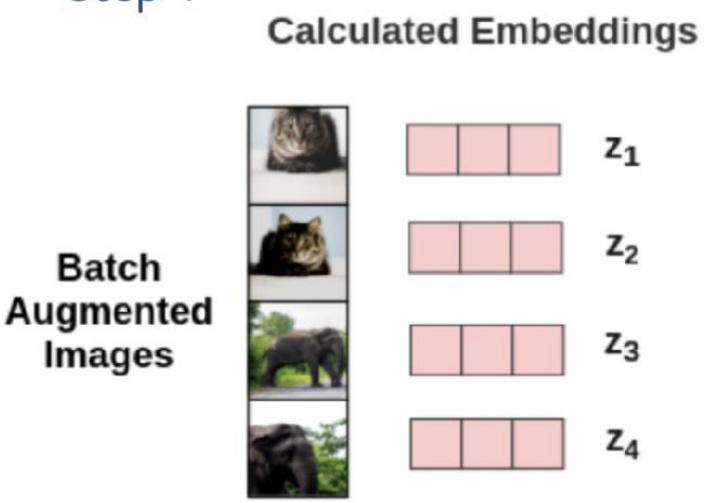


Slide credit: Y.Asano

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR

Step 1



Step 2

Similarity Calculation of Augmented Images

$$\text{similarity}(x_i, x_j) = \text{cosine similarity}(z_i, z_j)$$

$$s_{i,j} = \frac{z_i^T z_j}{(\tau ||z_i|| ||z_j||)}$$

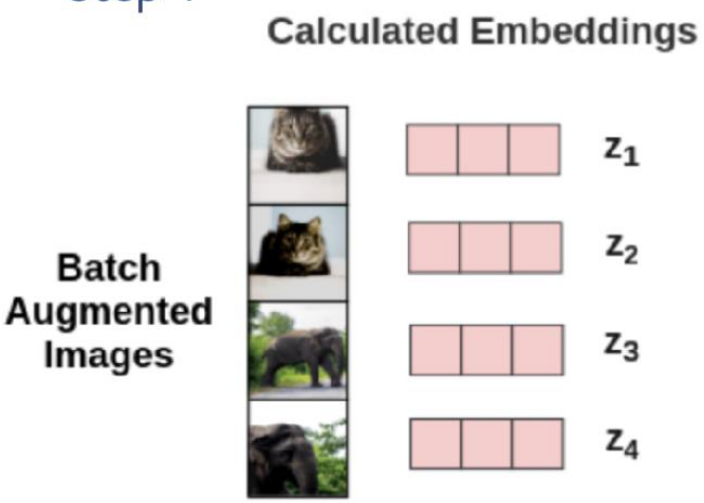
- τ is the adjustable temperature parameter. It can scale the inputs and widen the range `[-1, 1]` of cosine similarity
- ||z_i|| is the norm of the vector.

Slide credit: Y.Asano

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR

Step 1



Step 2

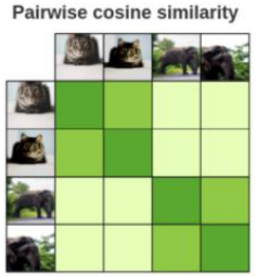
Similarity Calculation of Augmented Images

$$\text{similarity}(x_i, x_j) = \text{cosine similarity}(z_i, z_j)$$

$$s_{i,j} = \frac{z_i^T z_j}{(\tau ||z_i|| ||z_j||)}$$

- τ is the adjustable temperature parameter. It can scale the inputs and widen the range [-1, 1] of cosine similarity
- ||z_i|| is the norm of the vector.

Step 3



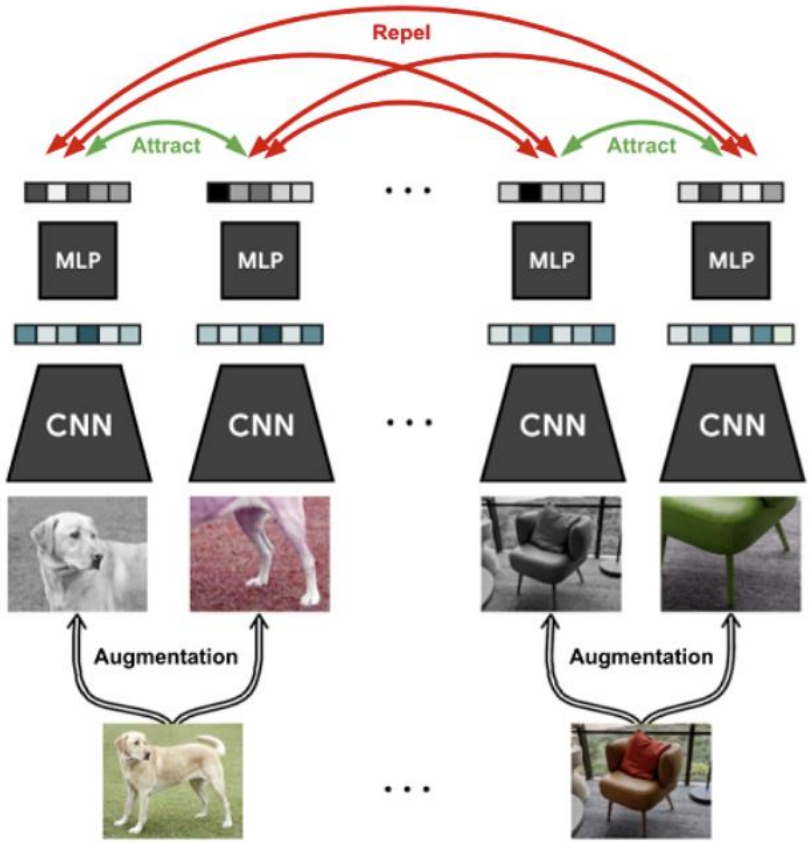
Loss: relatively increase similarity for pairs, decrease rest

What happens if you only try to increase the diagonal?

Slide credit: Y.Asano

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR



SimCLR

The contrastive loss for positive pairs i,j:

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

with z_i, z_j embeddings for images i and j,

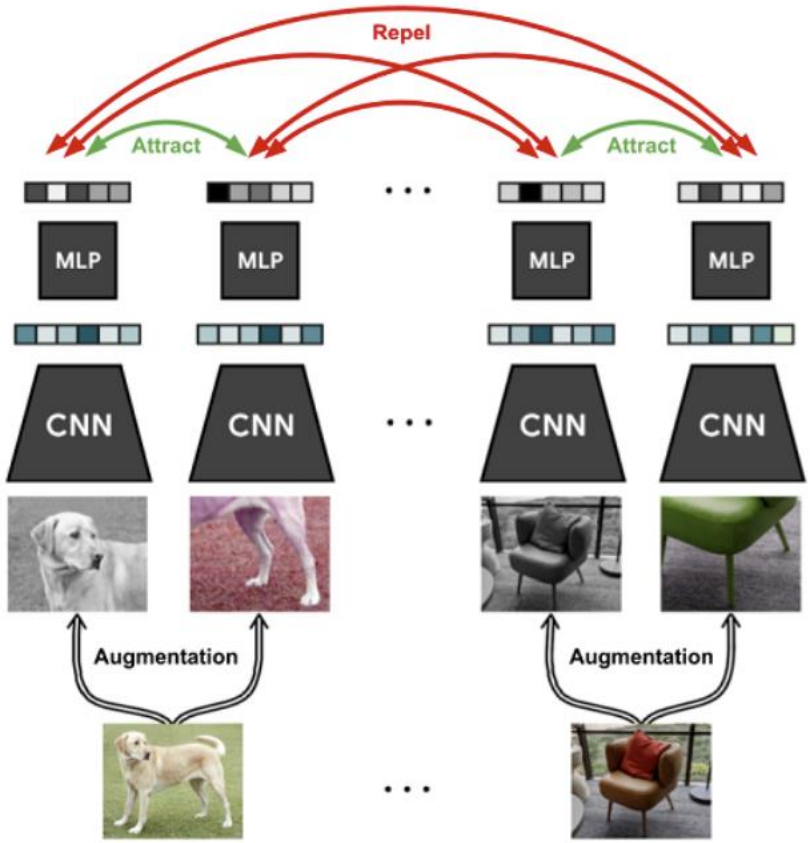
τ a temperature, sim() *is the dot-product*

"non-parametric" softmax

Slide credit: Y.Asano

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR



SimCLR

The contrastive loss for positive pairs i,j:

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

with z_i, z_j embeddings for images i and j,

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

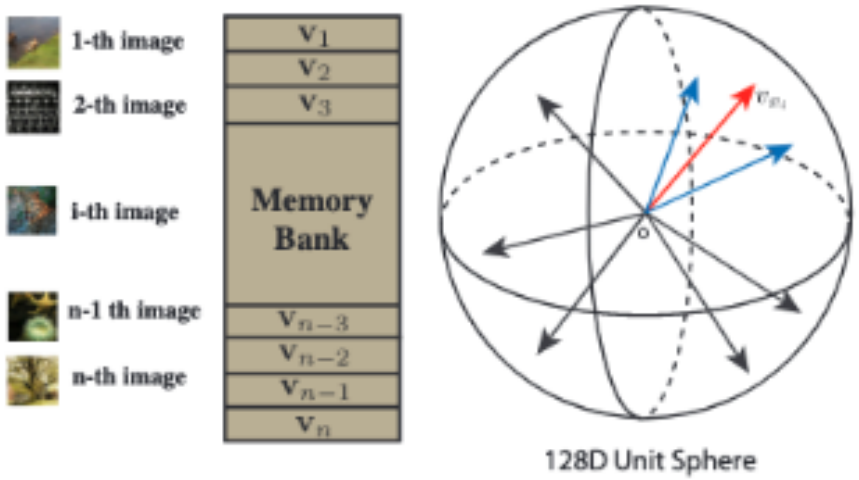
"non-parametric" softmax

Enforces image-uniqueness and
enforces augmentation-invariance

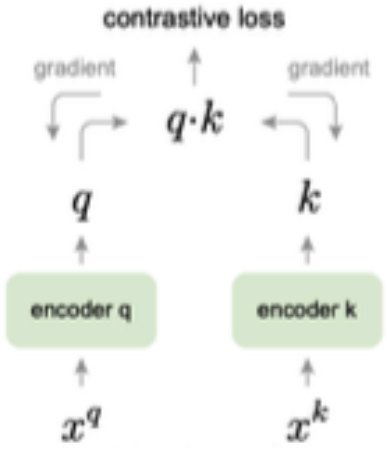
Slide credit: Y.Asano

Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

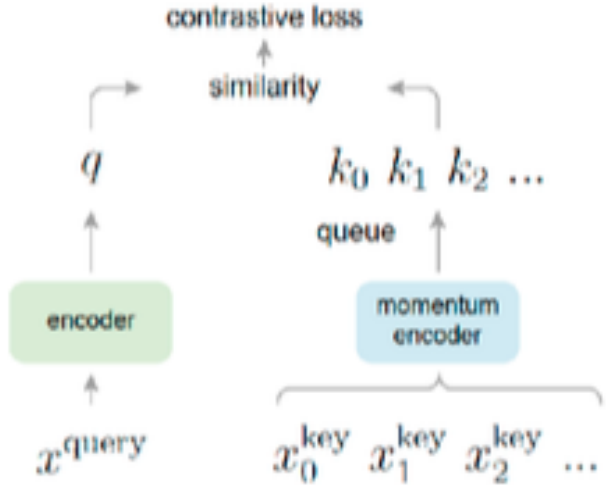
Modern Noise-contrastive self-supervised learning



NPID



SimCLR

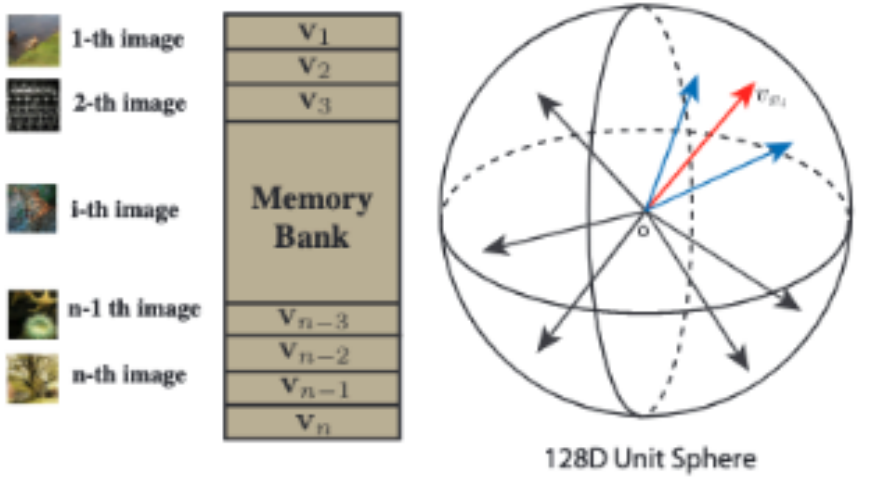


MoCo

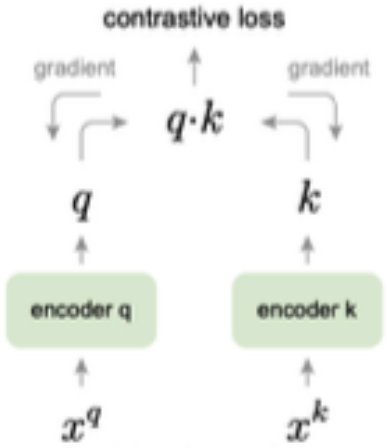
Wu et al. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. CVPR 2018
Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020
He et al. Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020

Modern Noise-contrastive self-supervised learning

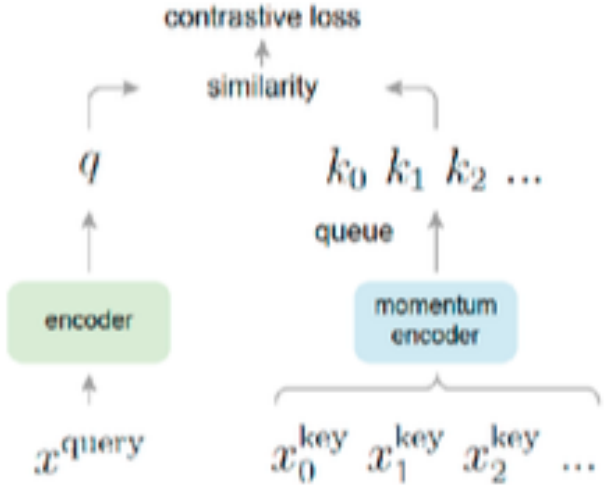
The start of large-scale & industrial self-supervised learning.
These works heavily rely on image augmentations.



NPID



SimCLR



MoCo

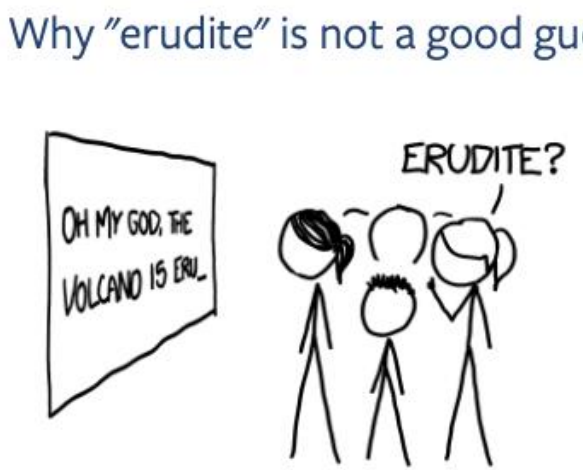
Momentum encoder:

```
# momentum update: key network
f_k.params = m*f_k.params+(1-m)*f_q.params
```


Wu et al. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. CVPR 2018
Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020
He et al. Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020

Masked Modeling

Language Modelling



Slide credit: Y.Asano

https://thegradient.pub/understanding-evaluation-metrics-for-language-models/
Radford et al. Improving Language Understanding by Generative Pre-Training . 2018

Language Modelling

Why "erudite" is not a good guess

Factor the probability of a datapoint (w_1,..., w_n):

$$P_{(w_1,w_2,\ldots,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)...p(w_n|w_1,w_2,..,w_{n-1})$$
$$= \prod_{i=1}^{n} p(w_i|w_1,...,w_{i-1})$$

Slide credit: Y.Asano

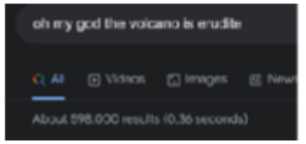
https://thegradient.pub/understanding-evaluation-metrics-for-language-models/
Radford et al. Improving Language Understanding by Generative Pre-Training . 2018

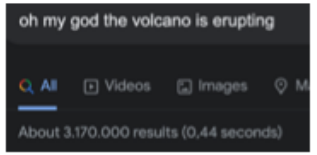
Language Modelling

Why "erudite" is not a good guess

Factor the probability of a datapoint (w_1,..., w_n):

$$P_{(w_1,w_2,\ldots,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)\ldots p(w_n|w_1,w_2,..,w_{n-1})$$
$$= \prod_{i=1}^{n} p(w_i|w_1,\ldots,w_{i-1})$$





S = Where are we going

Previous words (Context)

Word being predicted

P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

https://thegradient.pub/understanding-evaluation-metrics-for-language-models/
Radford et al. Improving Language Understanding by Generative Pre-Training . 2018

Slide credit: Y.Asano

Language Modelling: GPT

Generative Pretrained Transformer (aka GPT)
simply does language modelling with a Transformer (decoder)

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

https://thegradient.pub/understanding-evaluation-metrics-for-language-models/
Radford et al. Improving Language Understanding by Generative Pre-Training . 2018

Slide credit: Y.Asano

Language Modelling: GPT

Generative Pretrained Transformer (aka GPT)
simply does language modelling with a Transformer (decoder)

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

U = (u_k, . . . , u_1) is the context vector of tokens,
n is the number of layers,
W_e is the token embedding matrix,
W_p is the position embedding matrix

in practice: "causal" (left-to-right) context via masking

https://thegradient.pub/understanding-evaluation-metrics-for-language-models/
Radford et al. Improving Language Understanding by Generative Pre-Training . 2018

Slide credit: Y.Asano

Language Modelling: GPT

GPT-1,2,3: same loss. different training data and model sizes

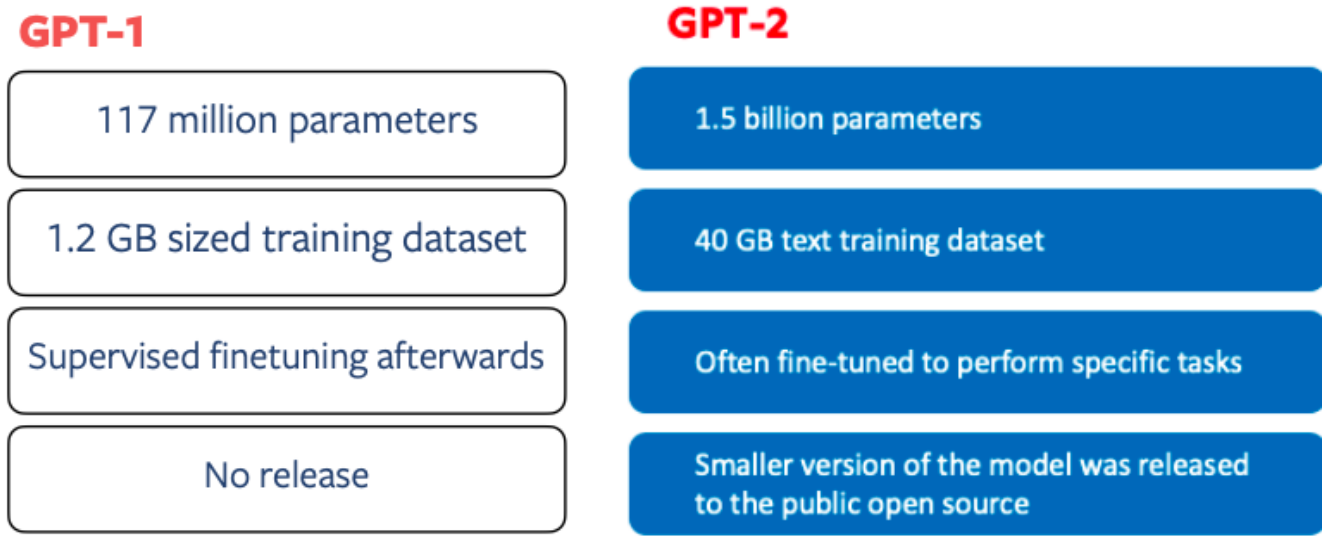
GPT-1

- 117 million parameters
- 1.2 GB sized training dataset
- Supervised finetuning afterwards
- No release

https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

Language Modelling: GPT

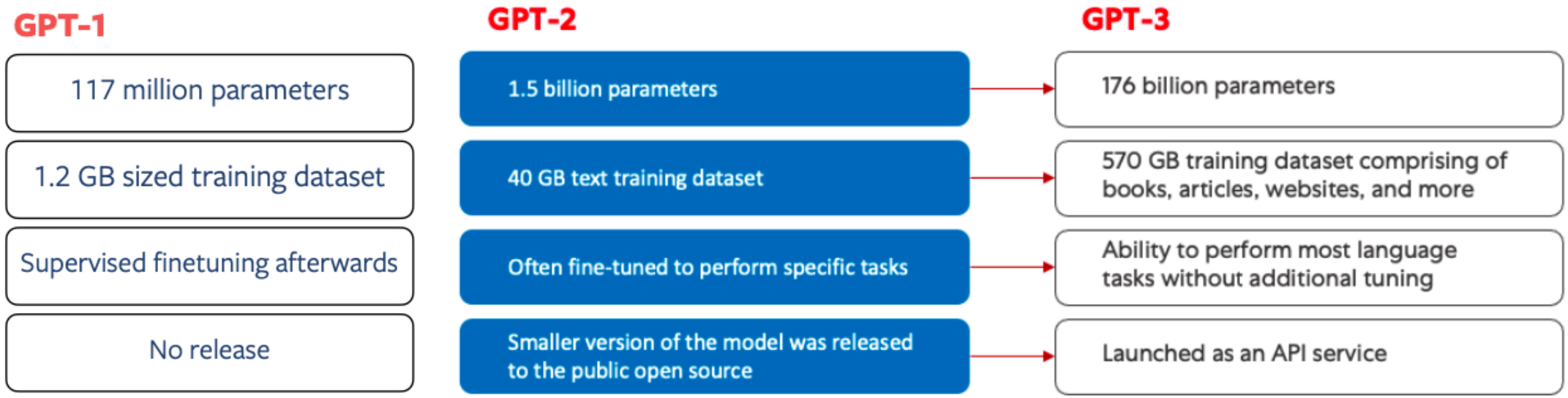
GPT-1,2,3: same loss. different training data and model sizes



https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

Language Modelling: GPT

GPT-1,2,3: same loss. different training data and model sizes

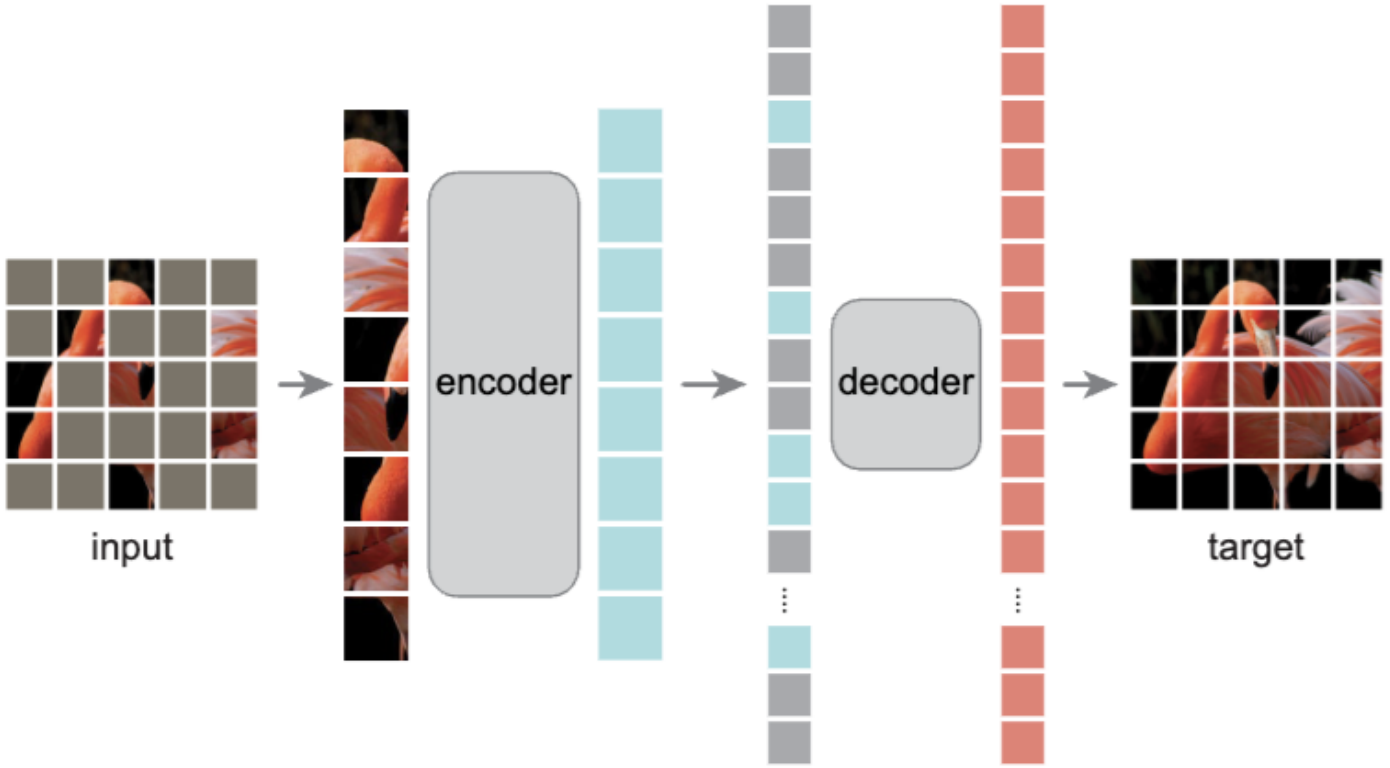


Slide credit: Y.Asano

https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

Masked Image Modelling

Masked AutoEncoder (MAE) ~ same, but instead of words use image patches

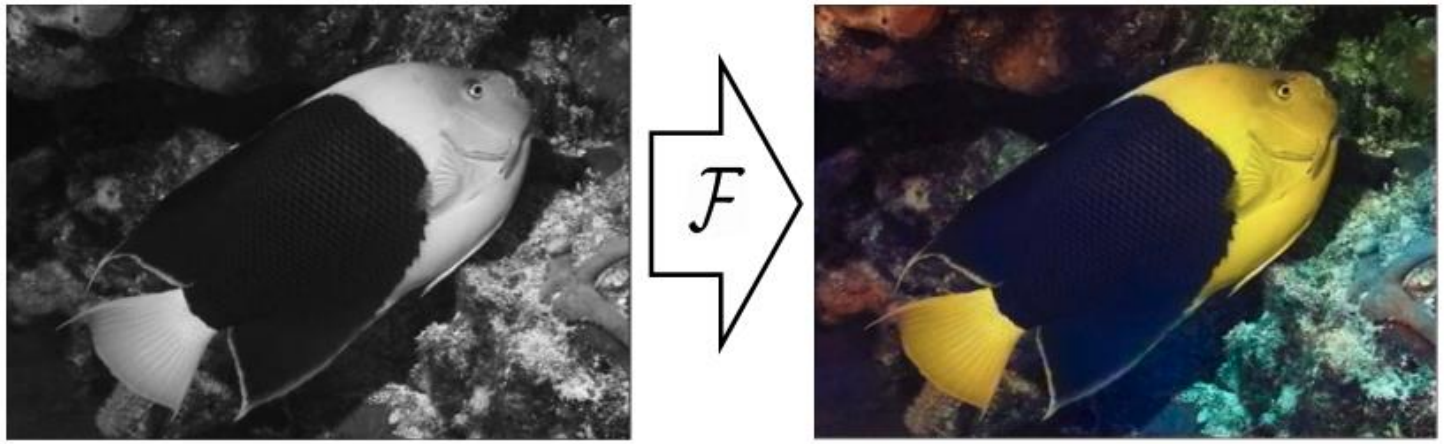


He et al. Masked Autoencoders Are Scalable Vision Learners. CVPR'21
Xie et al. SimMIM: A Simple Framework for Masked Image Modeling. ArXiv
Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Ima

Other examples

Self-supervised Learning Colorization

Train network to predict pixel colour from a monochrome input

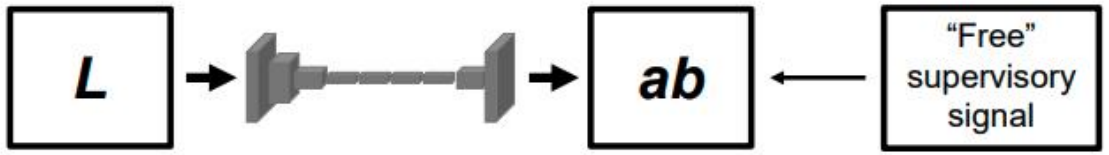


Grayscale image: *L* channel

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

Concatenate (L,ab)

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



Self-supervised Learning Colorization

Self-supervised Learning
Exemplar Networks

- Exemplar Networks (Dosovitskiy et al., 2014)
- Perturb/distort image patches, e.g. by cropping and affine transformations
- Train to classify these exemplars as same class

Self-supervised Learning
Image Transformations

- Which image has the correct rotation?

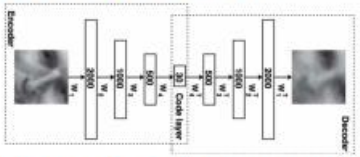
Self-supervised Learning
Image Transformations

- Which image has the correct rotation?

Figure 1: Images rotated by random multiples of 90 degrees (e.g., 0, 90, 180, or 270 degrees). The core intuition of our self-supervised feature learning approach is that if someone is not aware of the concepts of the objects depicted in the images, he cannot recognize the rotation that was applied to them.

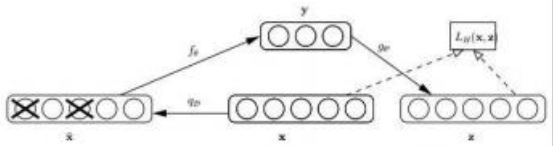
Self-supervised Learning Images

Autoencoders



Hinton & Salakhutdinov. Science 2006.

Denoising Autoencoders

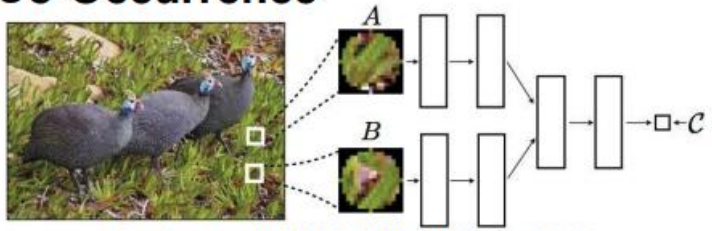


Vincent *et al.* ICML 2008.

Exemplar networks

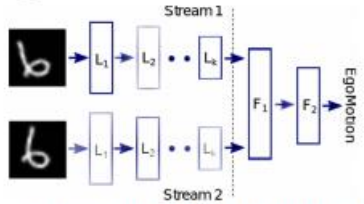
Dosovitskiy *et al.*, NIPS 2014

Co-Occurrence



Isola *et al.* ICLR Workshop 2016.

Egomotion



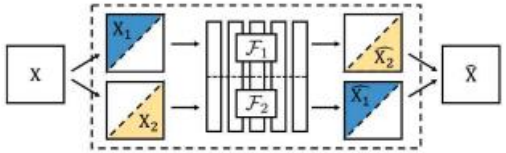
Agrawal *et al.* ICCV 2015 Jayaraman *et al.* ICCV 2015

Context

Noroozi et al 2016

Pathak *et al.* CVPR 2016

Split-brain auto-encoders



Zhang *et al.* CVPR 2017

Today's lecture

Part I: Learning without labels

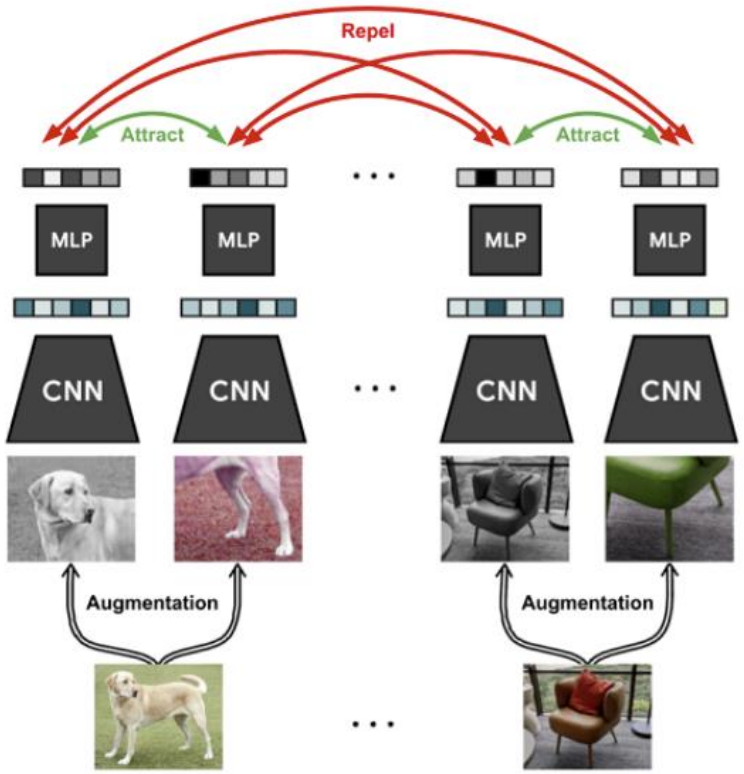
Part II: Self-Supervised Learning

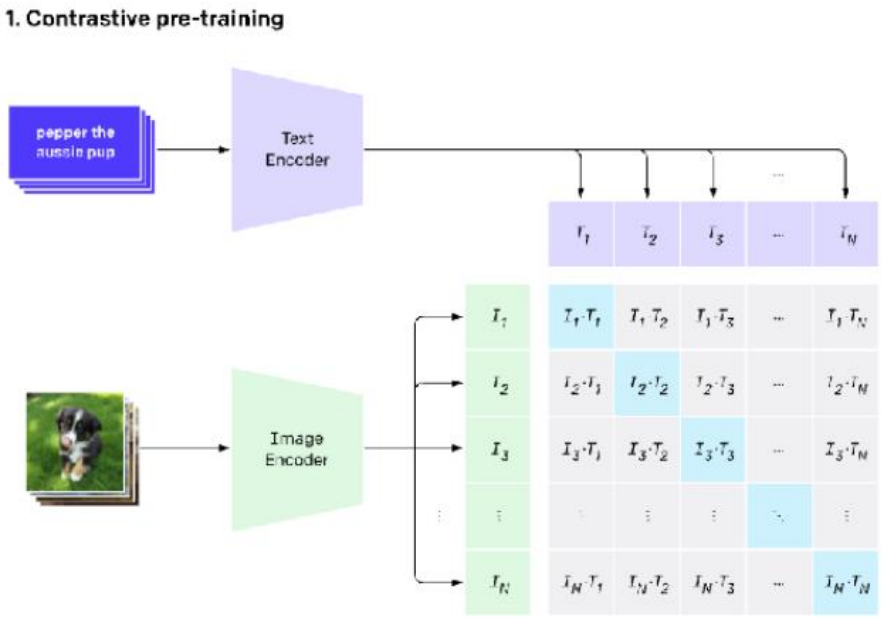
Part III: Multimodal Self-Supervised Learning

Part II: Outline
Multimodal Self-Supervised Learning (SSL)

- CLIP
- Videos
- Sound

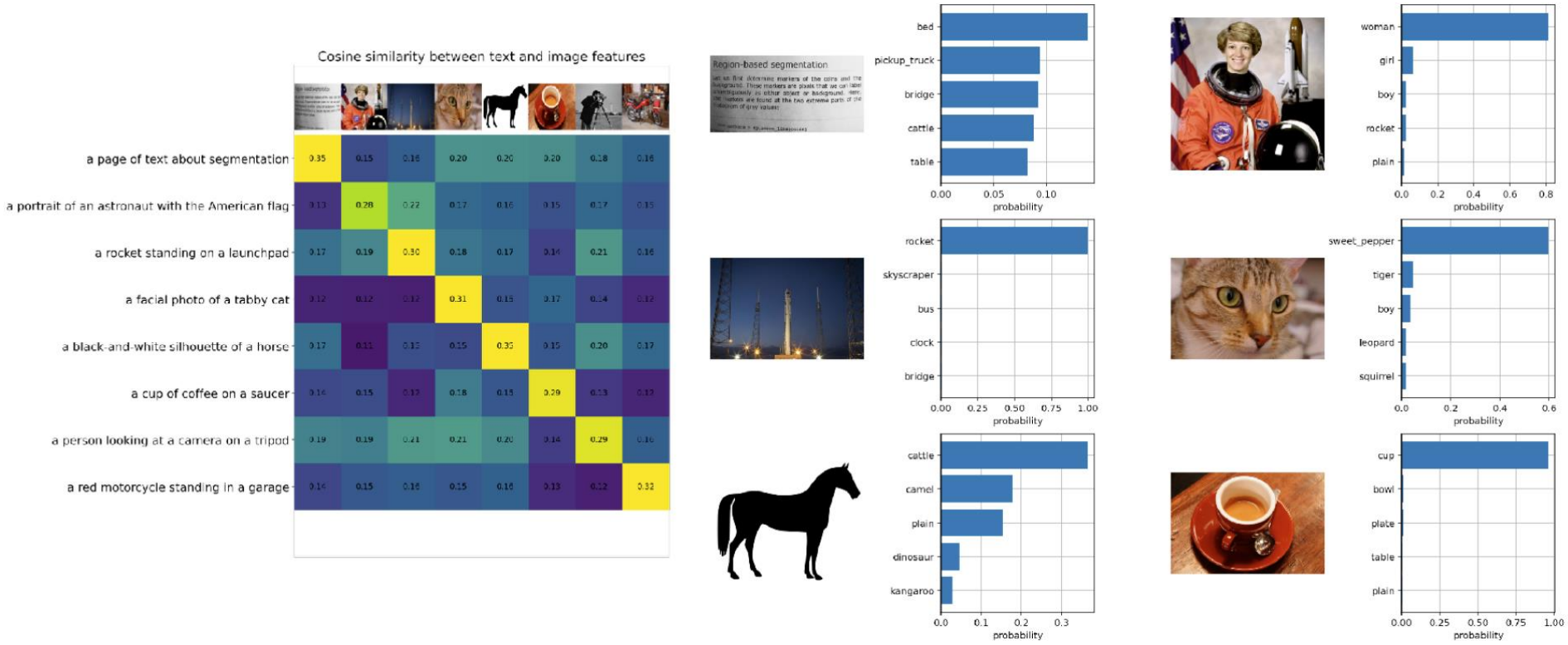
CLIP





Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2020
but also: Desai & Johnson VirTex: Learning Visual Representations from Textual Annotations. CVPR 2021 etc.

CLIP
Zero-shot classification



https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2020
but also: Desai & Johnson VirTex: Learning Visual Representations from Textual Annotations. CVPR 2021 etc.

When comparing pretrained image and language models, which one to adapt (more?)

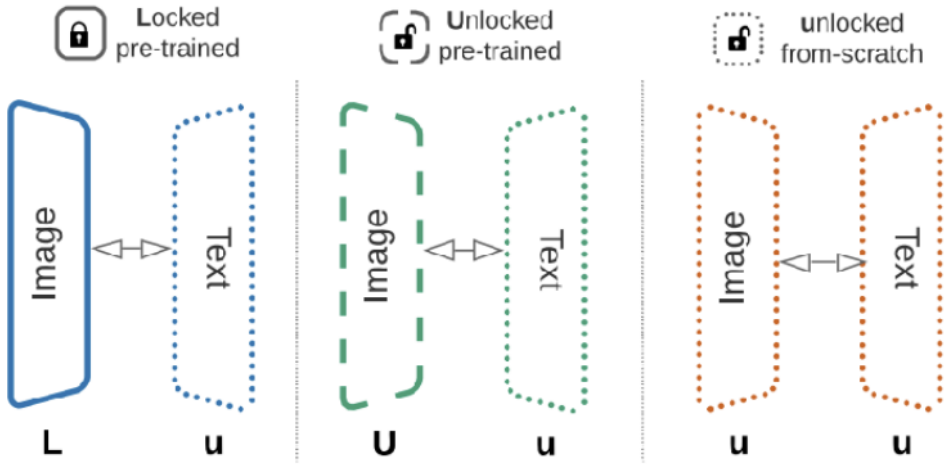


Figure 2. Design choices for contrastive-tuning on image-text data. Two letters are introduced to represent the image tower and text tower setups. `L` stands for locked variables and initialized from a pre-trained model, `U` stands for unlocked and initialized from a pre-trained model, `u` stands for unlocked and randomly initialized. `Lu` is named as "Locked-image Tuning" (LiT).

Method	ImgNet	ImgNet-v2	Cifar100	Pets
Lu	70.1	61.7	70.9	88.1
Uu	57.2	50.2	62.1	74.8
uu	50.6	43.3	47.9	70.3

Locking the image model is better.

Table 3: Zero-shot transfer results on ImageNet (variants).

Model	IN	IN-v2	IN-R	IN-A	ObjNet	ReaL
CLIP	76.2	70.1	88.9	77.2	72.3	-
ALIGN	76.4	70.1	92.2	75.8	72.2	-
BASIC	85.7	80.6	95.7	85.6	78.9	-
CoCa	86.3	80.7	96.5	90.2	82.7	-
LiT-g/14	85.2	79.8	94.9	81.8	82.5	88.6
LiT-e/14	85.4	80.6	96.1	88.0	84.9	88.4
LiT-22B	85.9	80.9	96.0	90.1	87.6	88.6

With only requiring one forward pass for getting image embeddings, can combine this with using a 22B parameter ViT

Zhai et al. LiT: Zero-Shot Transfer with Locked-image text Tuning. CVPR 2022
Dheghani et al. Scaling Vision Transformers to 22 Billion Parameters. 2023

Slide credit: Y.Asano

ALIGN

```
<figure class="wp-block-image size-large"><img
src="https://yourdomain.com/wp-content/uploads/2020/06/space-gray-
iphonex" alt="This is where you enter the text description."
class="wp-image-1204"/></figure>
```

"motorcycle front wheel"

"*thumbnail for version as of 21 57 29 june 2010*"

"file frankfurt airport skyline 2017 05 jpg"

"file london barge race 2 jpg"

"moustache seamless wallpaper design"

"st oswalds way and shops"

Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021

ALIGN

```
<figure class="wp-block-image size-large"><img
src="https://yourdomain.com/wp-content/uploads/2020/06/space-gray-
iphonex" alt="This is where you enter the text description."
class="wp-image-1204"/></figure>
```

"motorcycle front wheel"

"*thumbnail for version as of 21 57 29 june 2010*"

"file frankfurt airport skyline 2017 05 jpg"

"file london barge race 2 jpg"

"moustache seamless wallpaper design"

"st oswalds way and shops"

Novelty:

Start with very noisy dataset and:

- Filter based on images:
  - remove small ones, remove ones with >1k captions/alt texts
- Filter based on text:
  - alt-text with >10 occurences are removesd(e.g. "1920x10280")
  - too short or too long, or too rare
- Result:
  - dataset size ~2B (CLIP: 400M)

Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021

Slide credit: Y.Asano

ALIGN

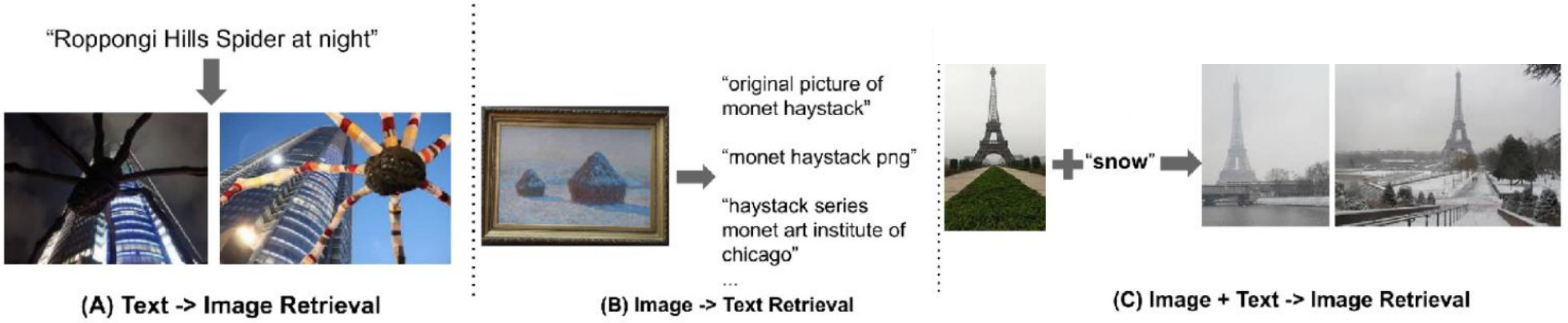


Image-text retrieval

Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021

Slide credit: Y.Asano

Self-Supervised Learning from Videos

Self-supervised Learning Videos

- A temporal sequence of frames

Self-supervised Learning Videos

- A temporal sequence of frames

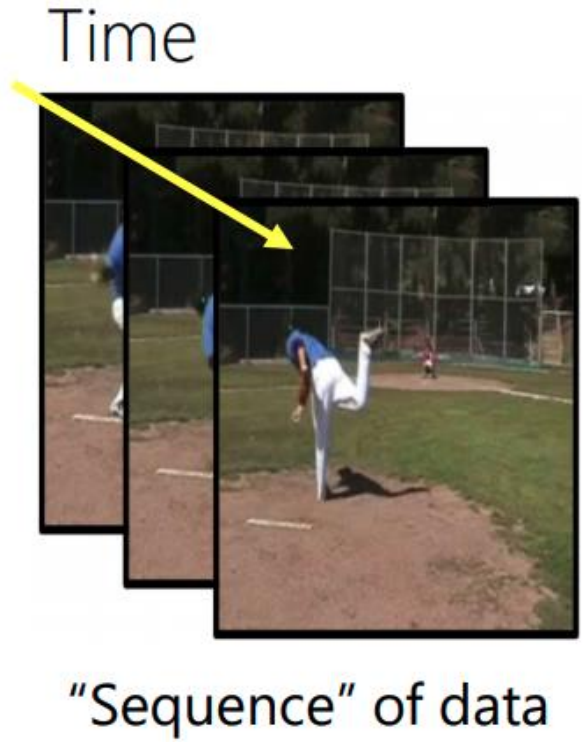
What can we use to define a proxy loss?

- Nearby (in time) frames are strongly correlated, further away may not be
- Temporal order of the frames
- Motion of objects (via optical flow)
- …

Temporal structure

Self-supervised Learning
Temporal structure in videos

- Shuffle and Learn: Unsupervised Learning using Temporal Order Verification



Self-supervised Learning
Temporal structure in videos

- Is this a valid sequence?

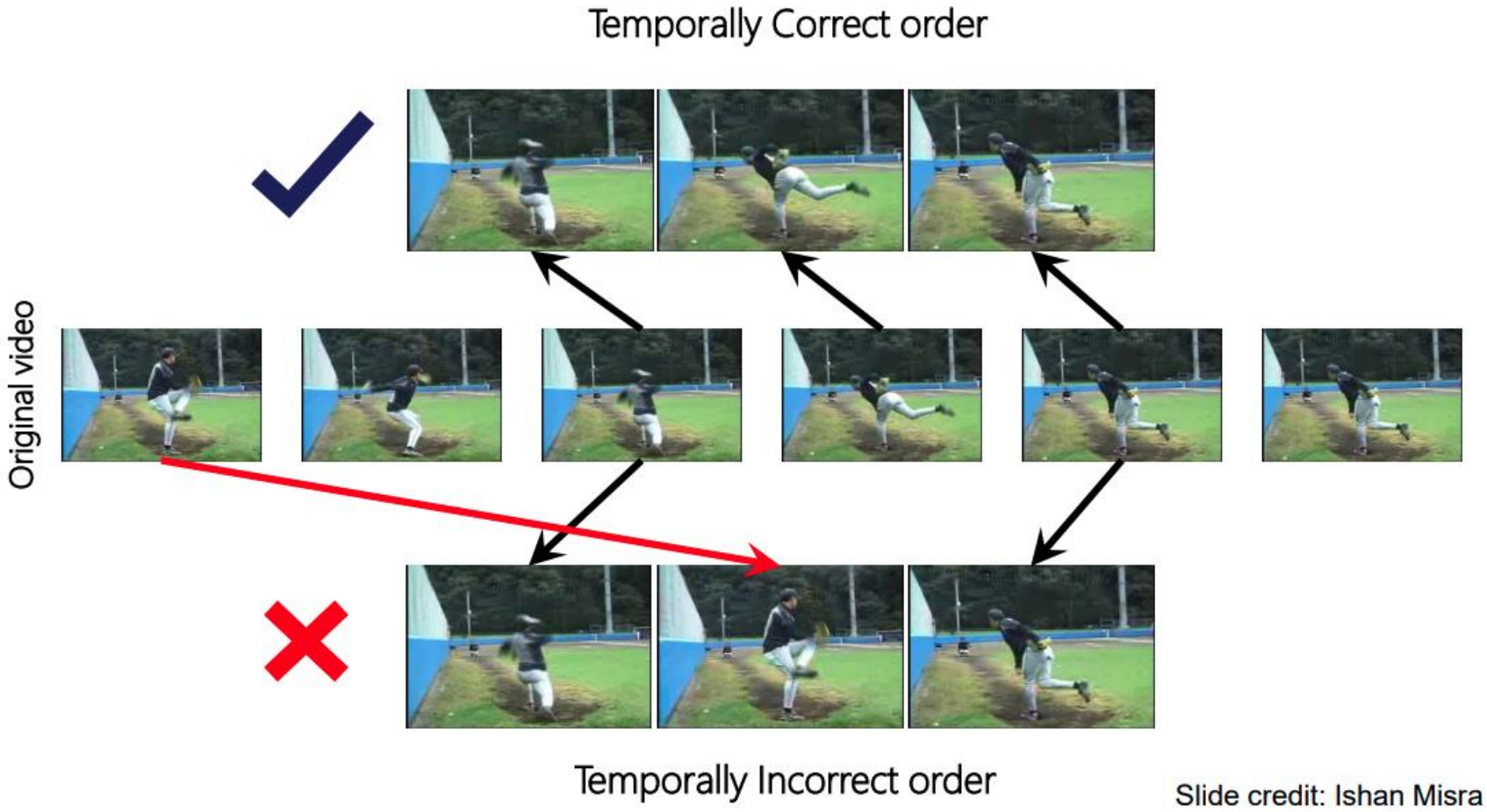


Self-supervised Learning
Temporal structure in videos

- Is this a valid sequence?

  - Original video

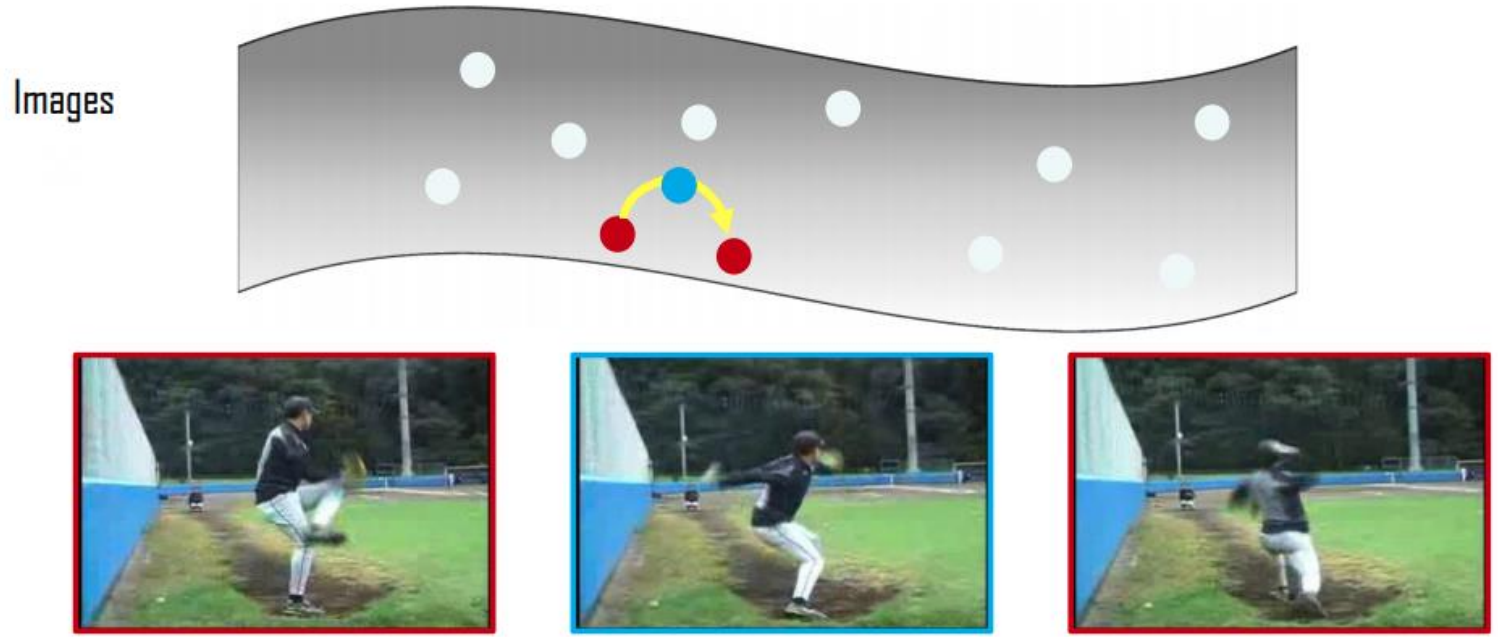
Self-supervised Learning
Temporal structure in videos



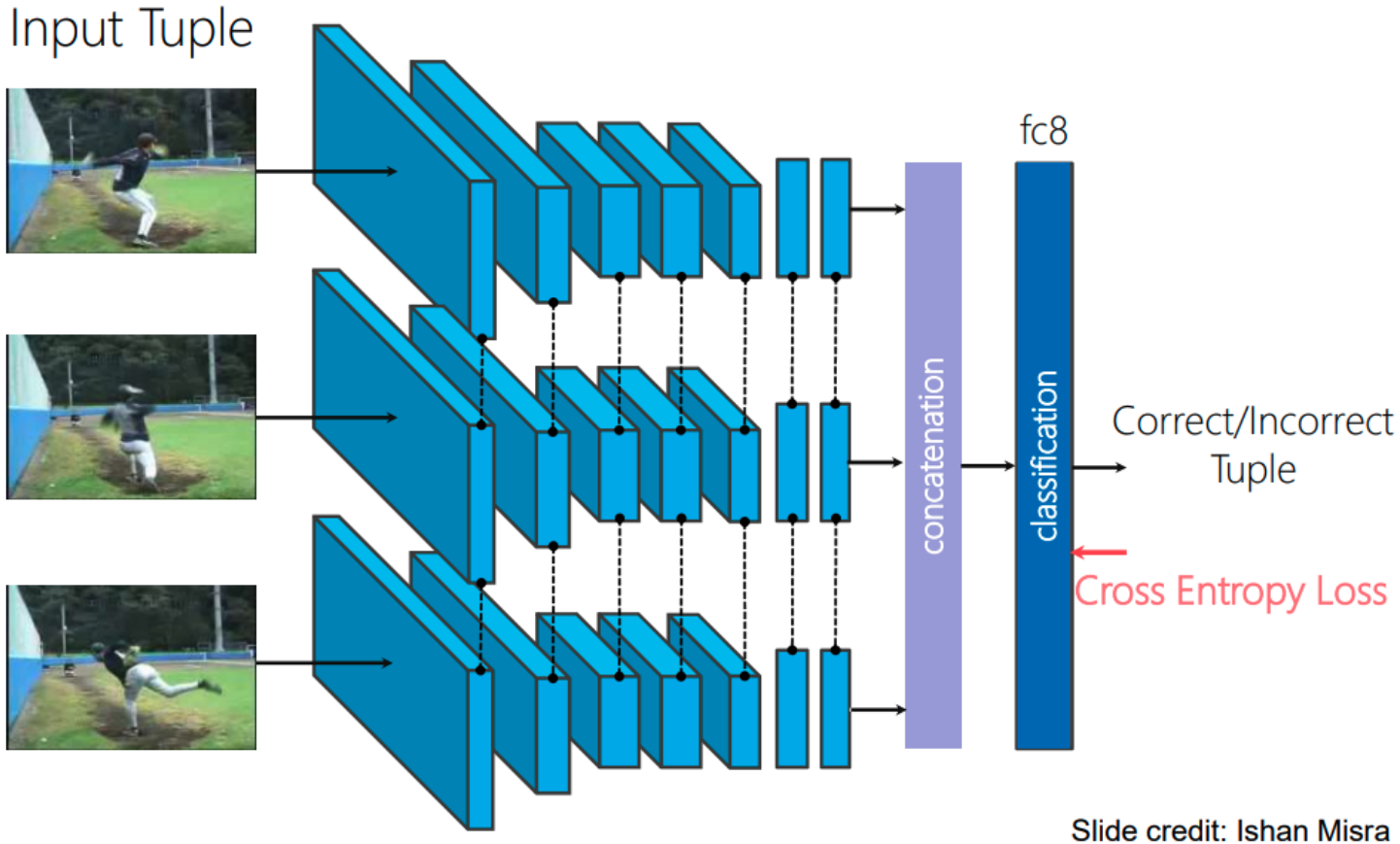
Slide credit: Ishan Misra

Self-supervised Learning
Temporal structure in videos

- Geometric View

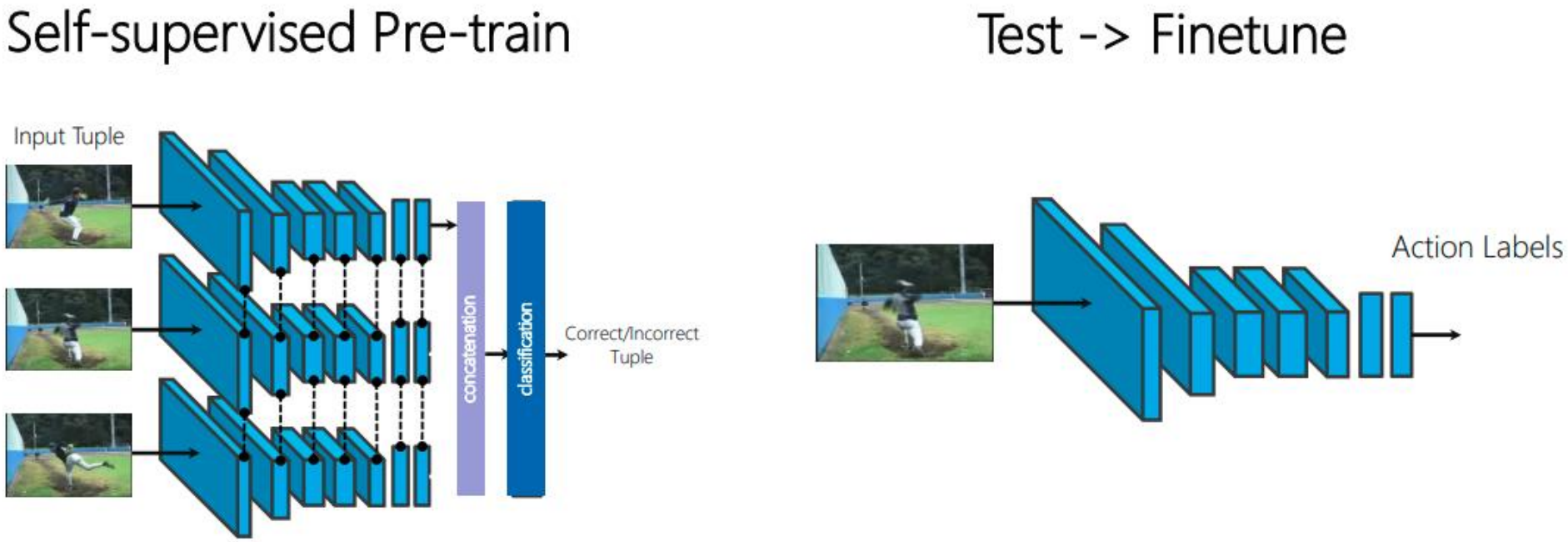


Self-supervised Learning
Temporal structure in videos



Self-supervised Learning
Temporal structure in videos

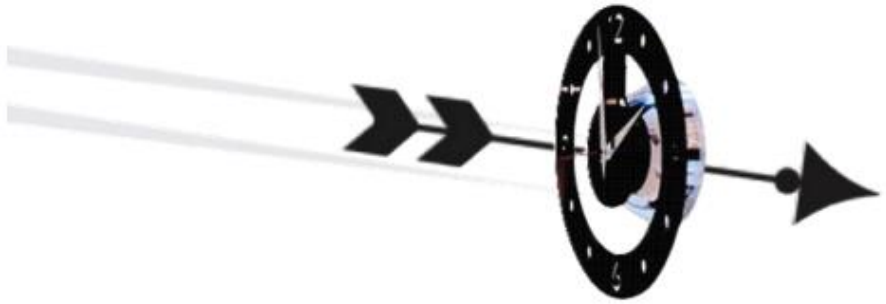
- Fine-tuning setup



Learning the arrow of time

Self-supervised Learning
Learning the arrow of time

Task: predict if video playing forwards or backwards



Supervision:

Positive training samples: video clips playing forwards

Negative training samples: video clips playing backwards

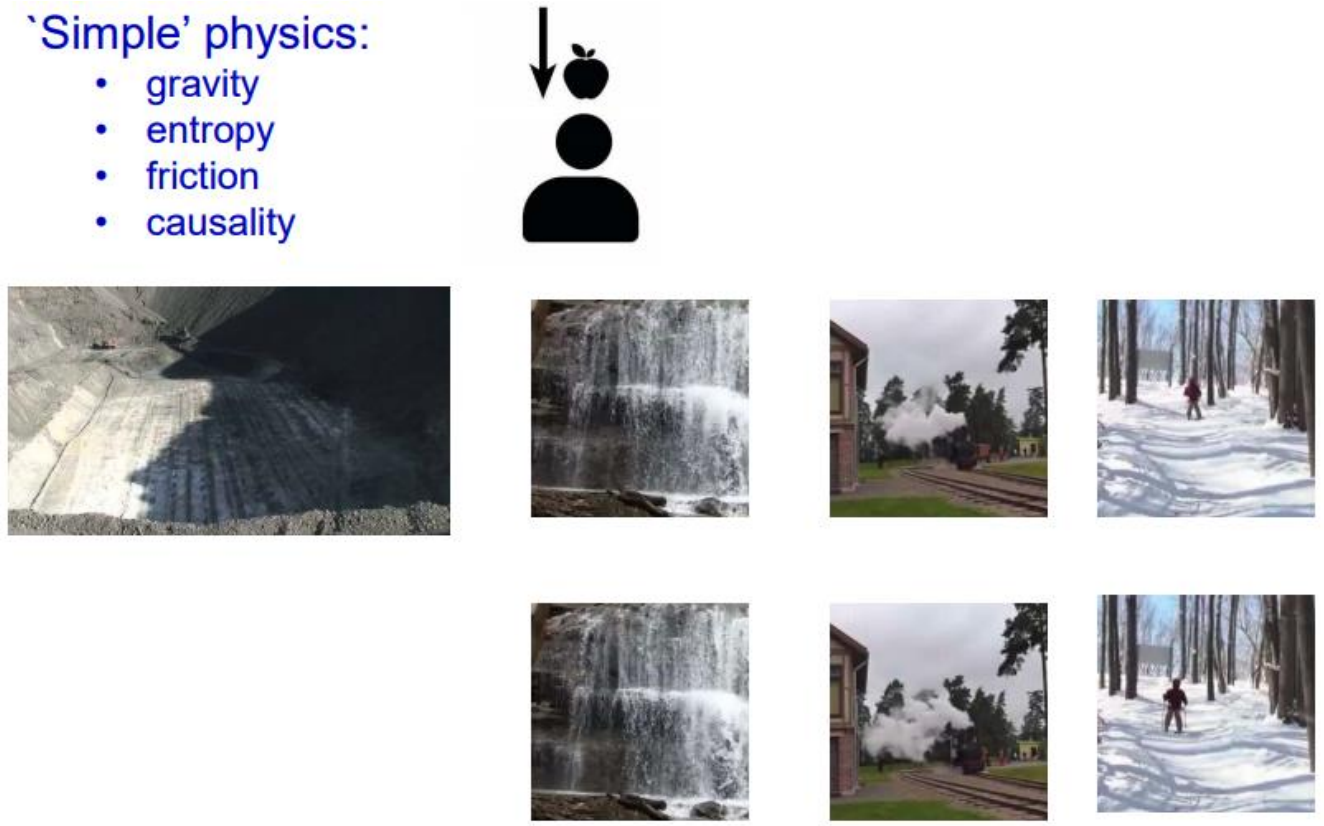
Self-supervised Learning
Learning the arrow of time

- Strong cues

Semantic, face motion direction, ordering

Self-supervised Learning
Learning the arrow of time

- Strong cues



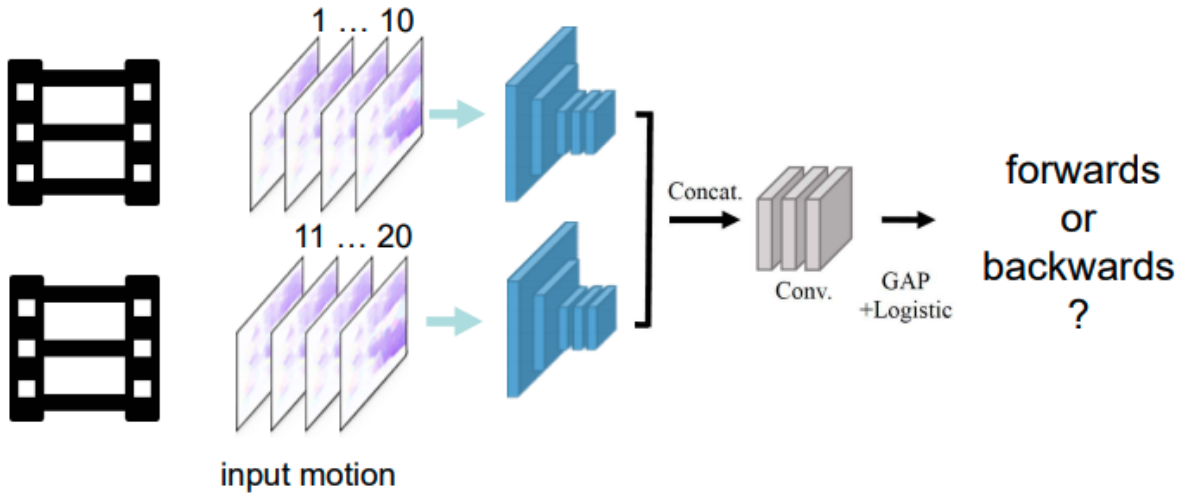
Self-supervised Learning
Learning the arrow of time

- Weak or no cues

Symmetric in time, constant motion, repetitions

Self-supervised Learning
Learning the arrow of time

- Temporal class-activation Map Network



T-CAM Model:

Input: optical flow in two chunks

Final layer: global average pooling to allow class activation map (CAM)

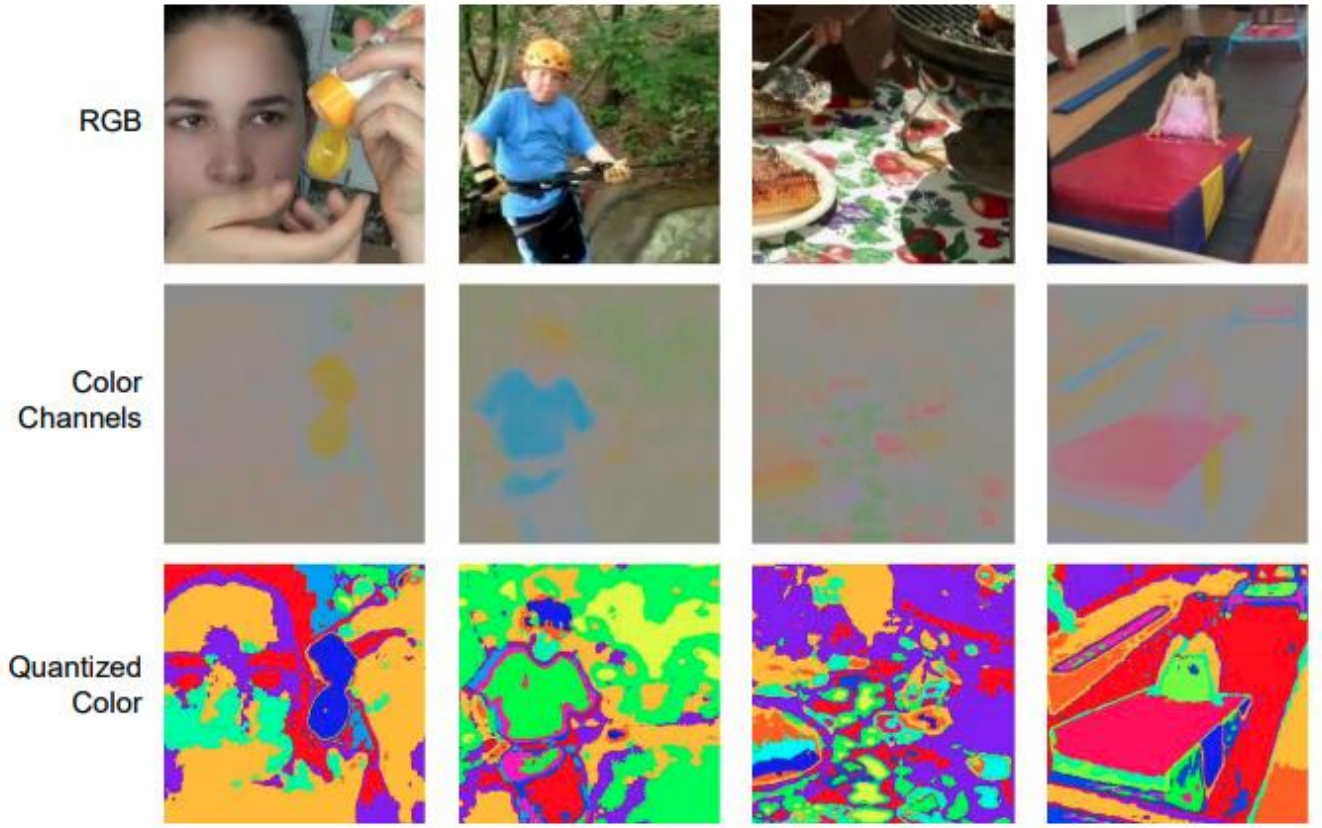
Self-supervised Learning
Learning the arrow of time

- Semantic motions

Tracking emerges by colorization

Self-supervised Learning
Tracking emerges by colorization

- Temporal coherence of color



Self-supervised Learning
Tracking emerges by colorization

- Self-supervised tracking

Task: given a color video …
Colorize all frames of a gray scale version using a reference frame

Reference Frame

Gray-scale Video

Self-supervised Learning
Tracking emerges by colorization

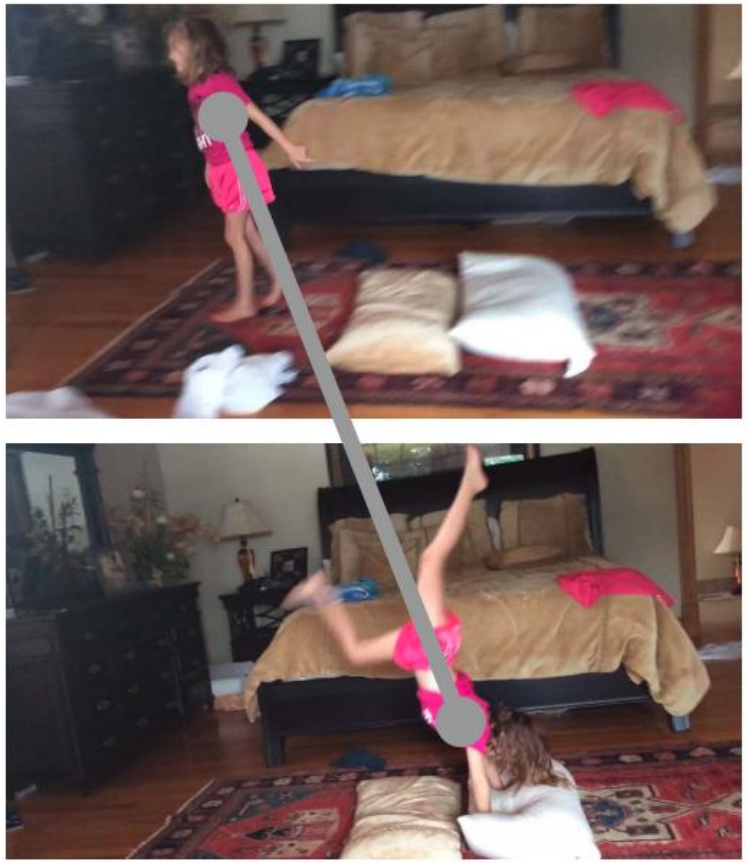
- What is this color?

Self-supervised Learning
Tracking emerges by colorization

- Where to copy color from?

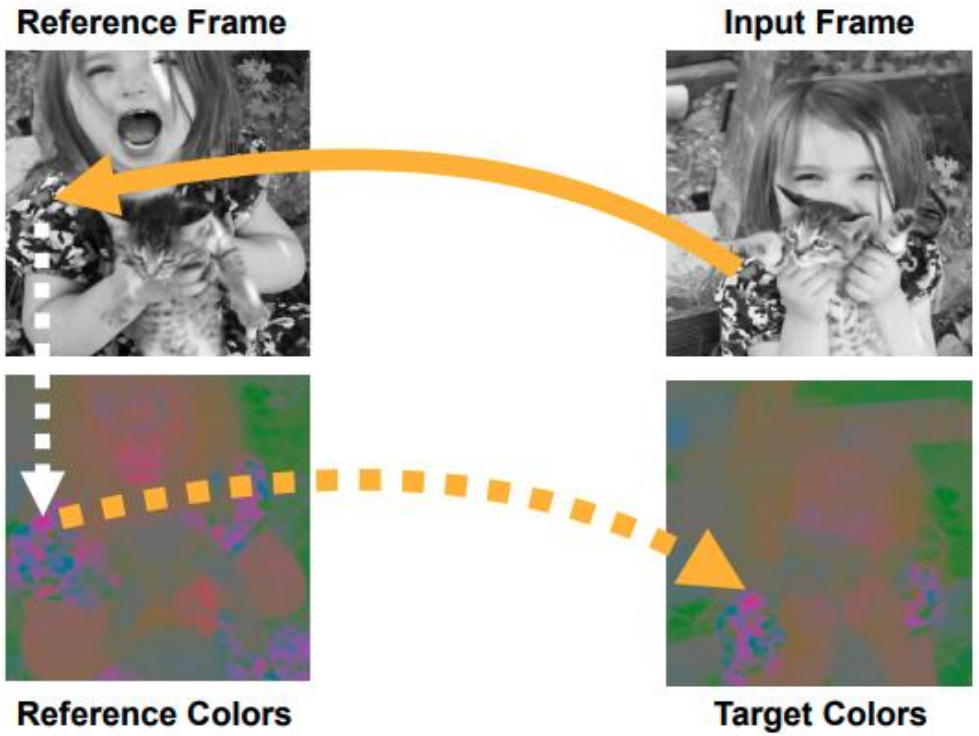
Self-supervised Learning
Tracking emerges by colorization

- Semantic correspondence



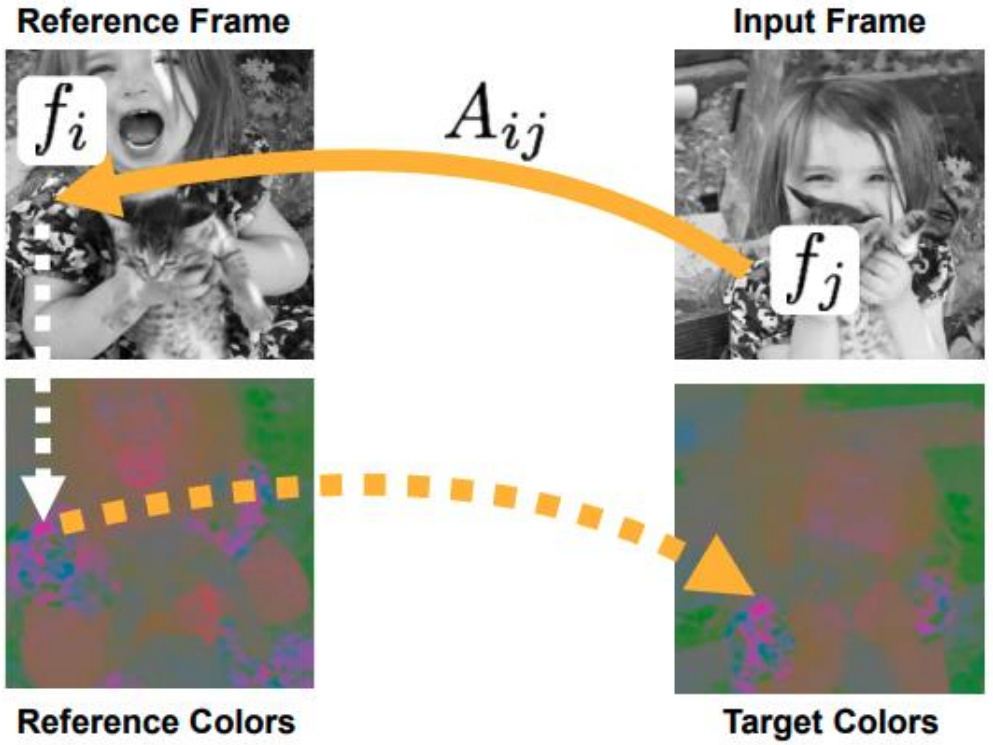
Self-supervised Learning
Tracking emerges by colorization

- Colorize by pointing



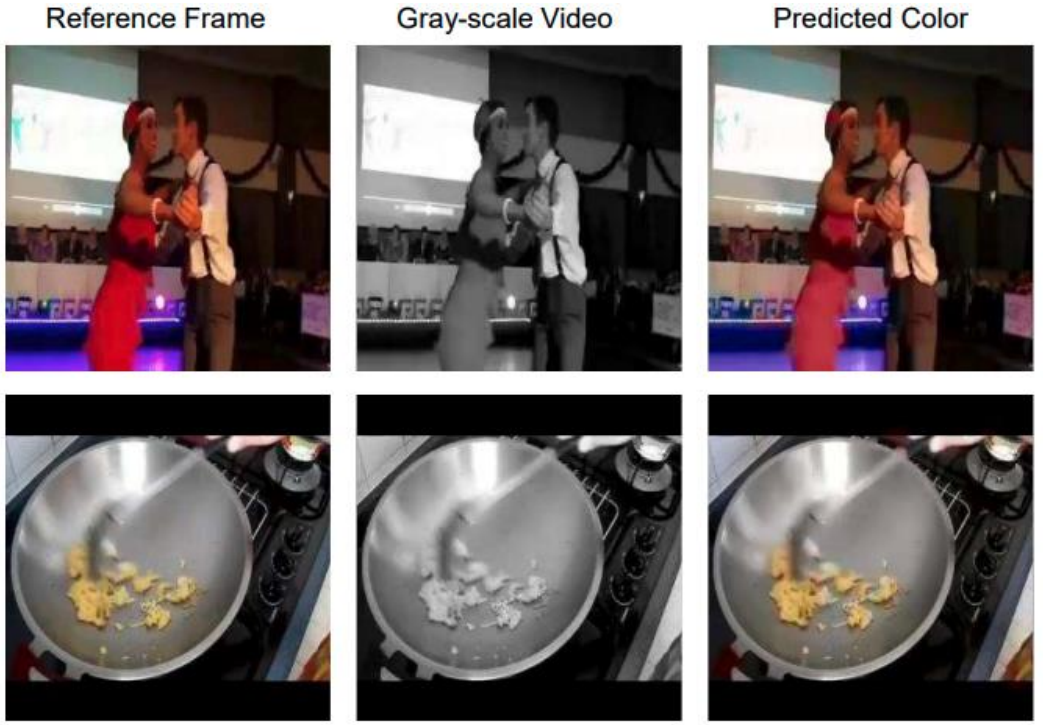
Self-supervised Learning
Tracking emerges by colorization

- Colorize by pointing



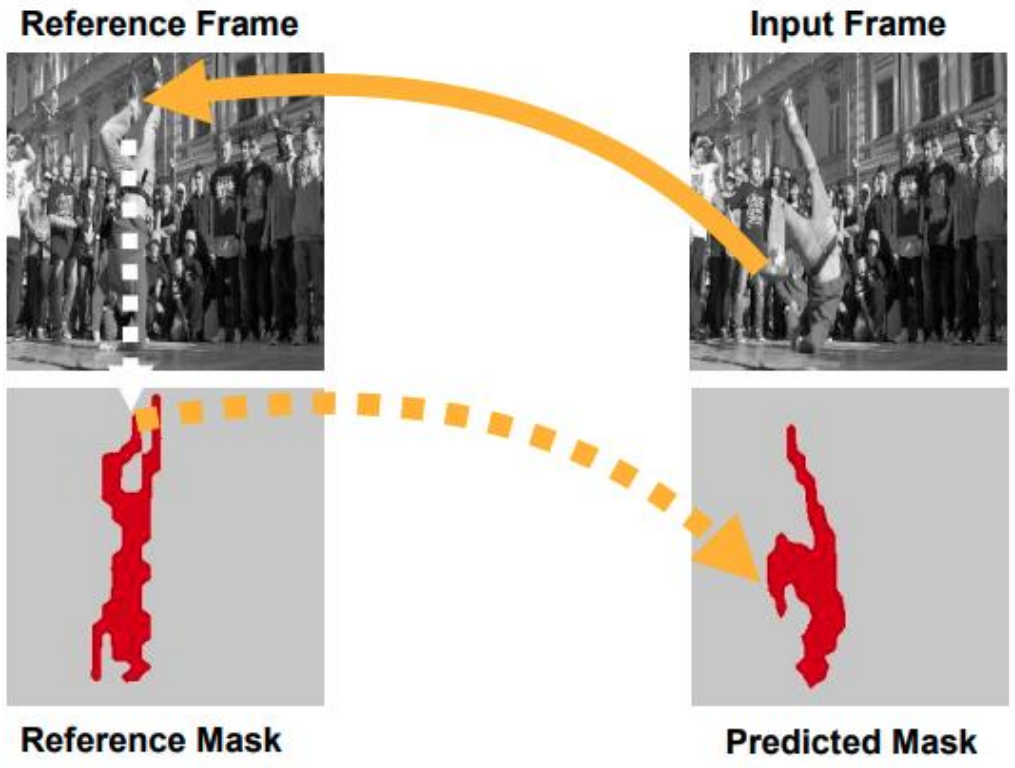
Self-supervised Learning
Tracking emerges by colorization

- Video colorization



Self-supervised Learning
Tracking emerges by colorization

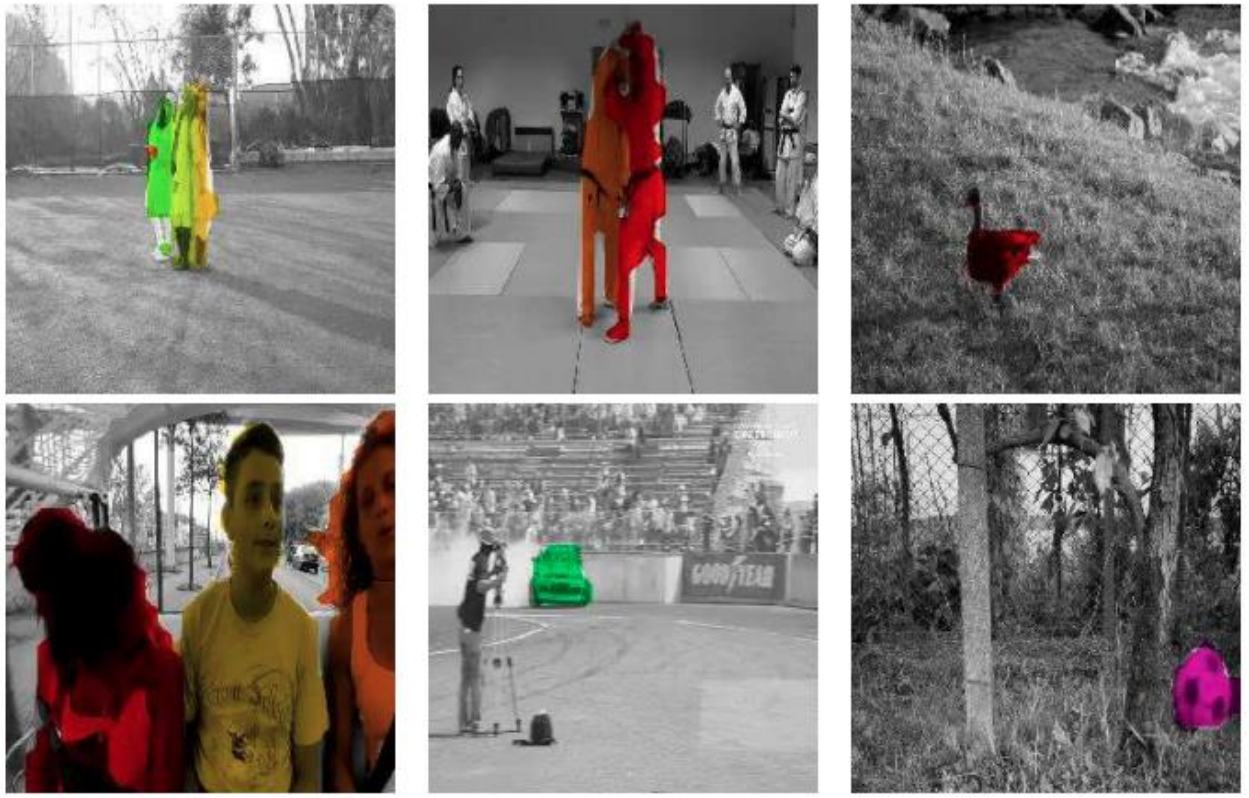
- Tracking emerges!



Self-supervised Learning
Tracking emerges by colorization

- Segment tracking results

Only the first frame is given. Colors indicate different instances.



Videos with sound

Self-supervised Learning
Audio-Visual Co-supervision

- Sound and frames are:
  - Semantically consistent
  - Synchronized

Self-supervised Learning
Audio-Visual Co-supervision

- Goal: use vision and sound to learn from each other
- Two types of proxy task:
  - Predict audio-visual correspondence
  - Predict audio-visual synchronization

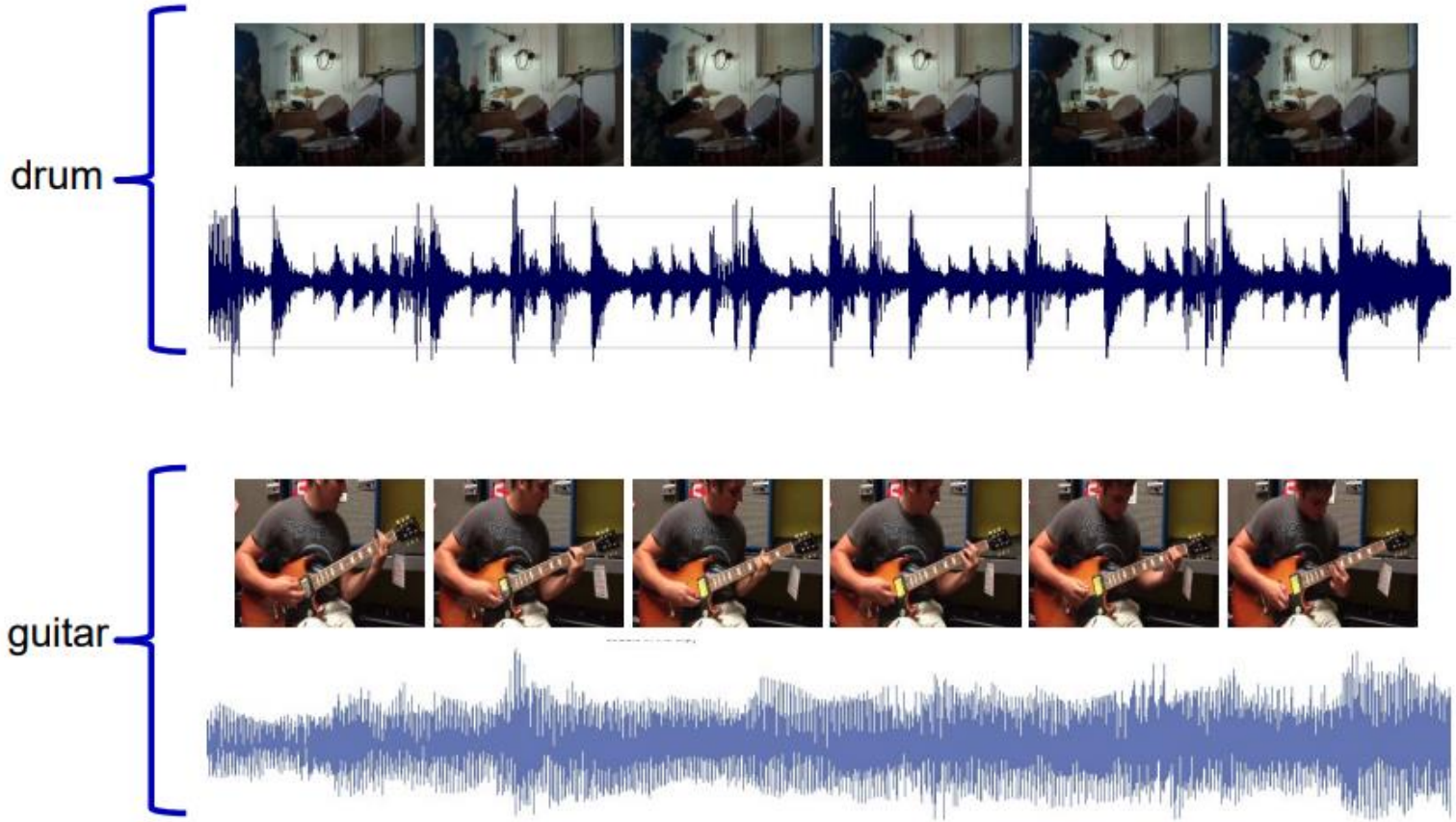
Self-supervised Learning
Audio-Visual Co-supervision

- Train a network to predict if image and audio clip correspond

Correspond?

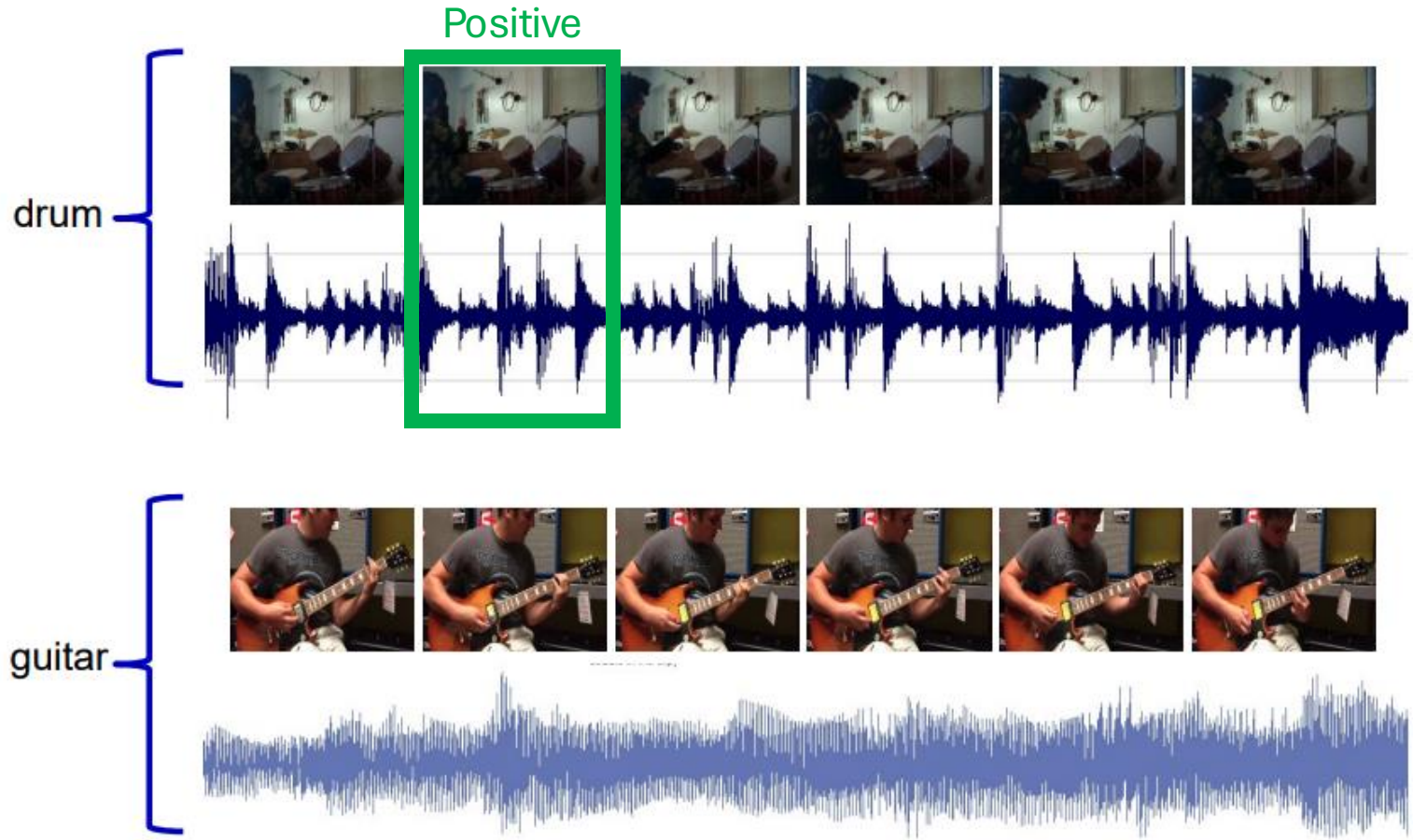
Self-supervised Learning
Audio-Visual Co-supervision

- Audio-visual Correspondence



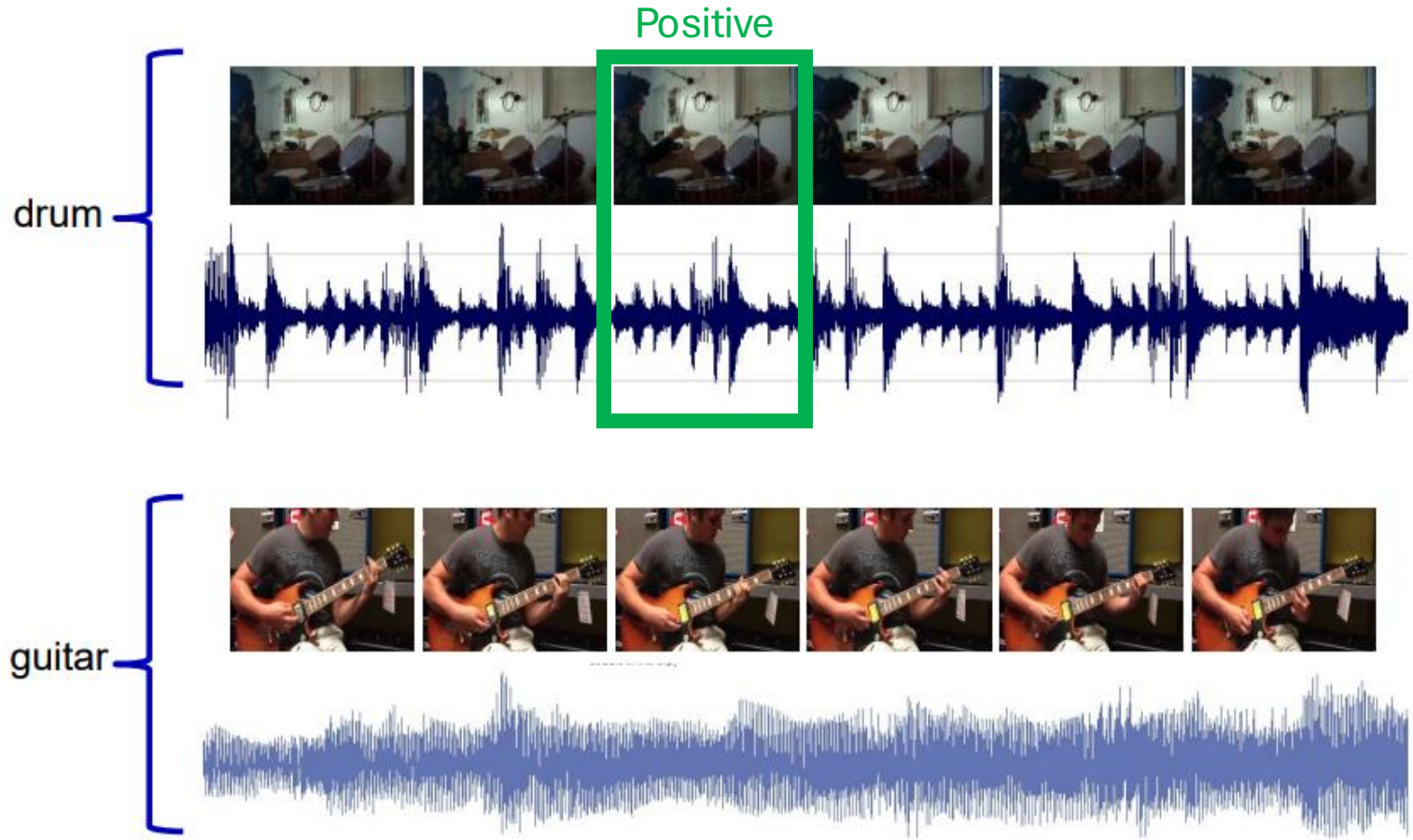
Self-supervised Learning
Audio-Visual Co-supervision

- Audio-visual Correspondence



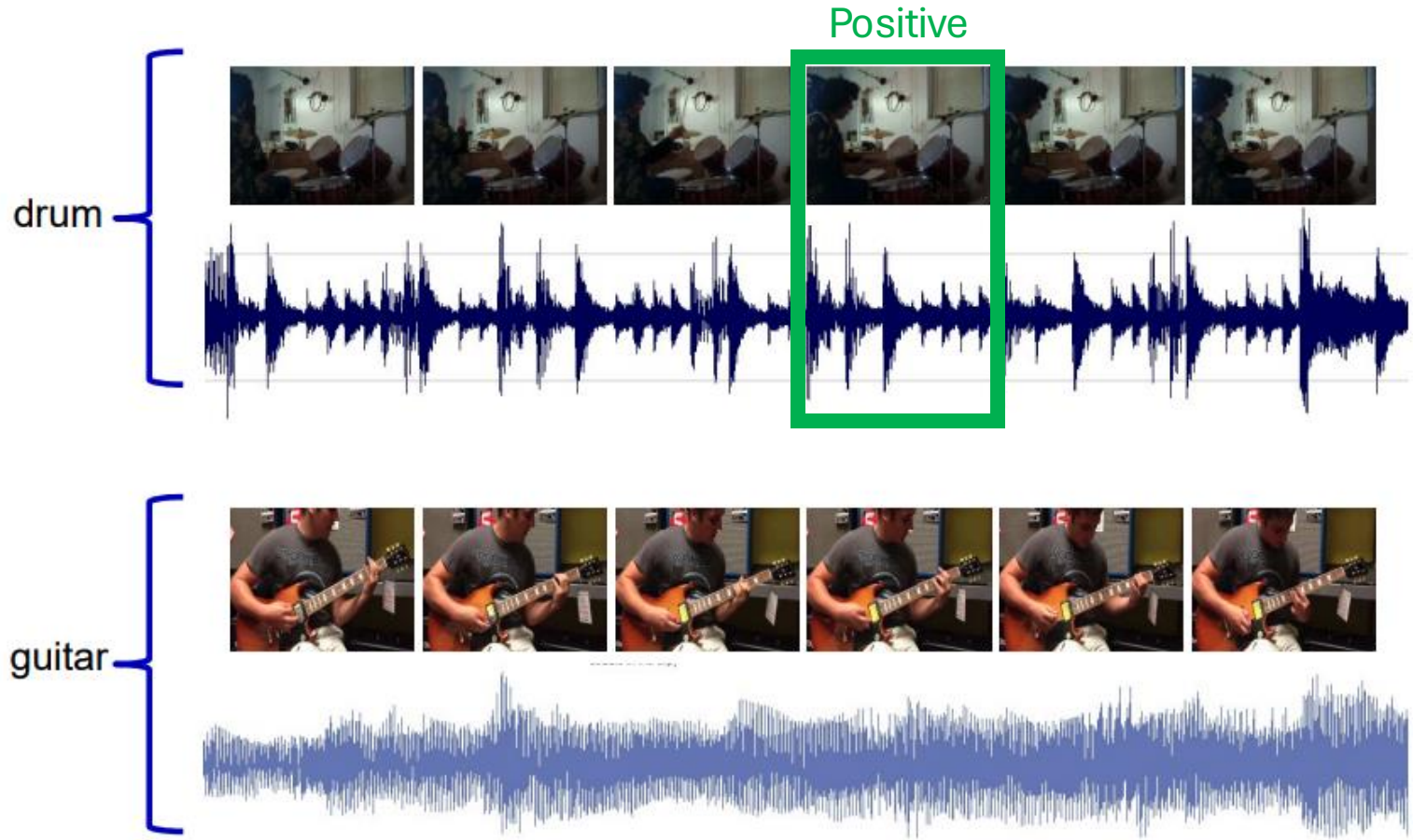
Self-supervised Learning
Audio-Visual Co-supervision

- Audio-visual Correspondence



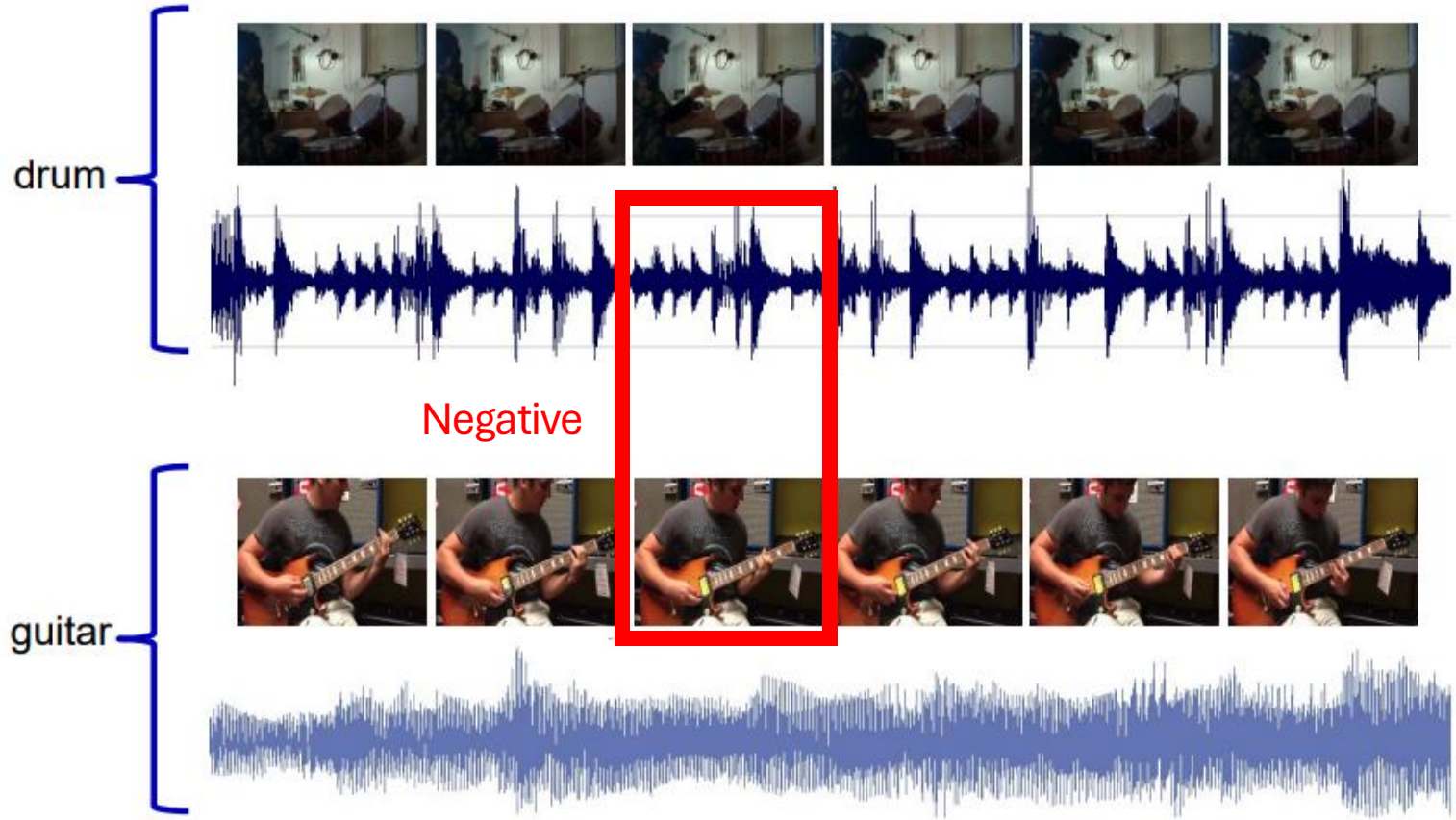
Self-supervised Learning
Audio-Visual Co-supervision

- Audio-visual Correspondence



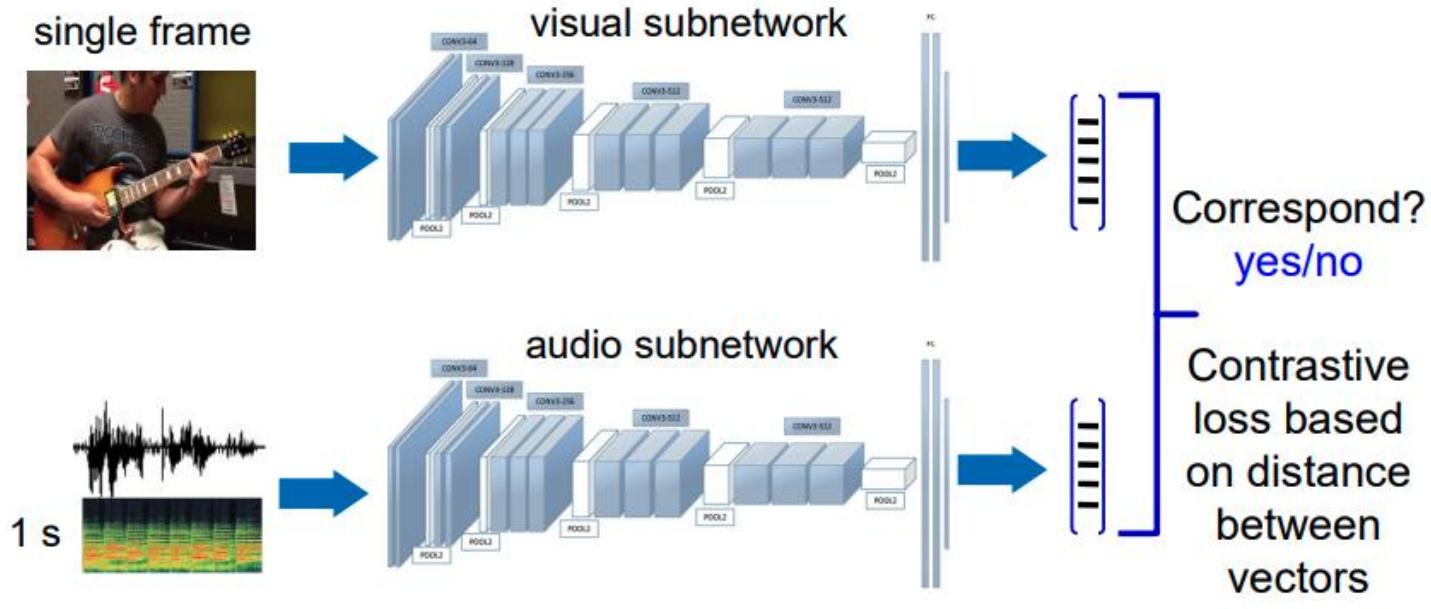
Self-supervised Learning
Audio-Visual Co-supervision

- Audio-visual Correspondence



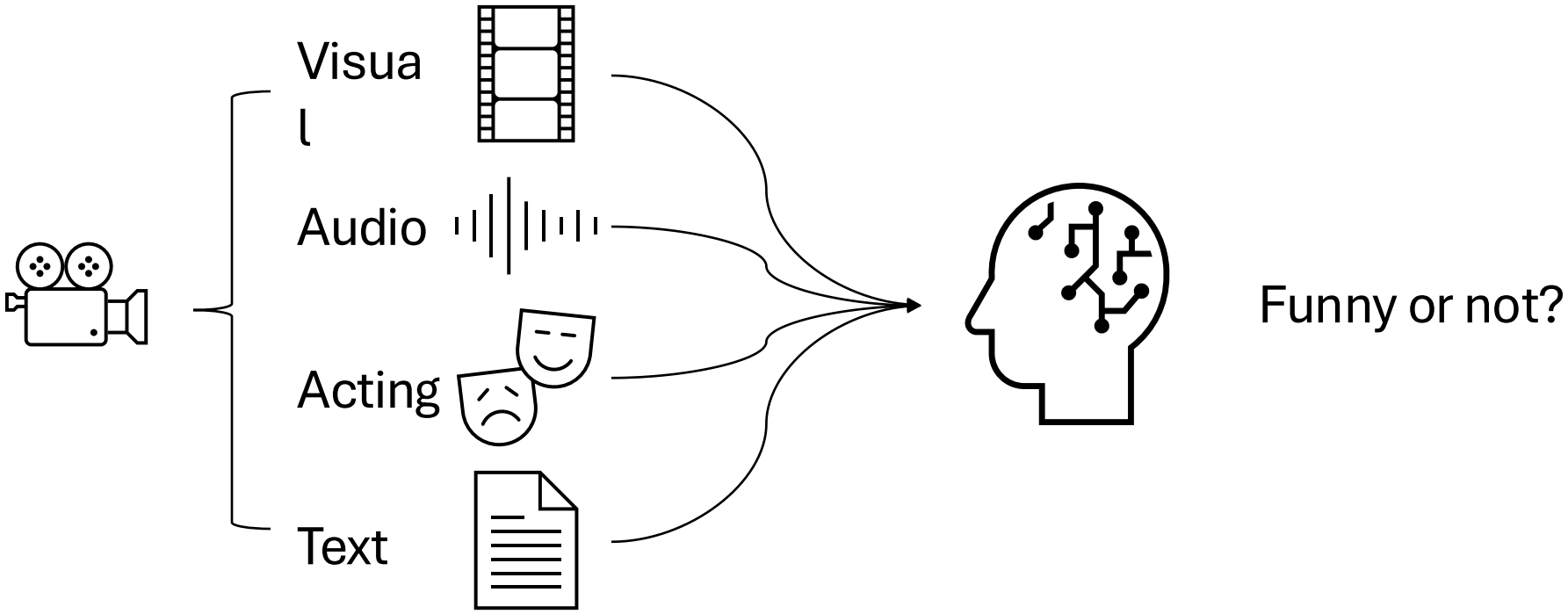
Self-supervised Learning
Audio-Visual Co-supervision

- Audio-visual Embedding (AVE-Net)



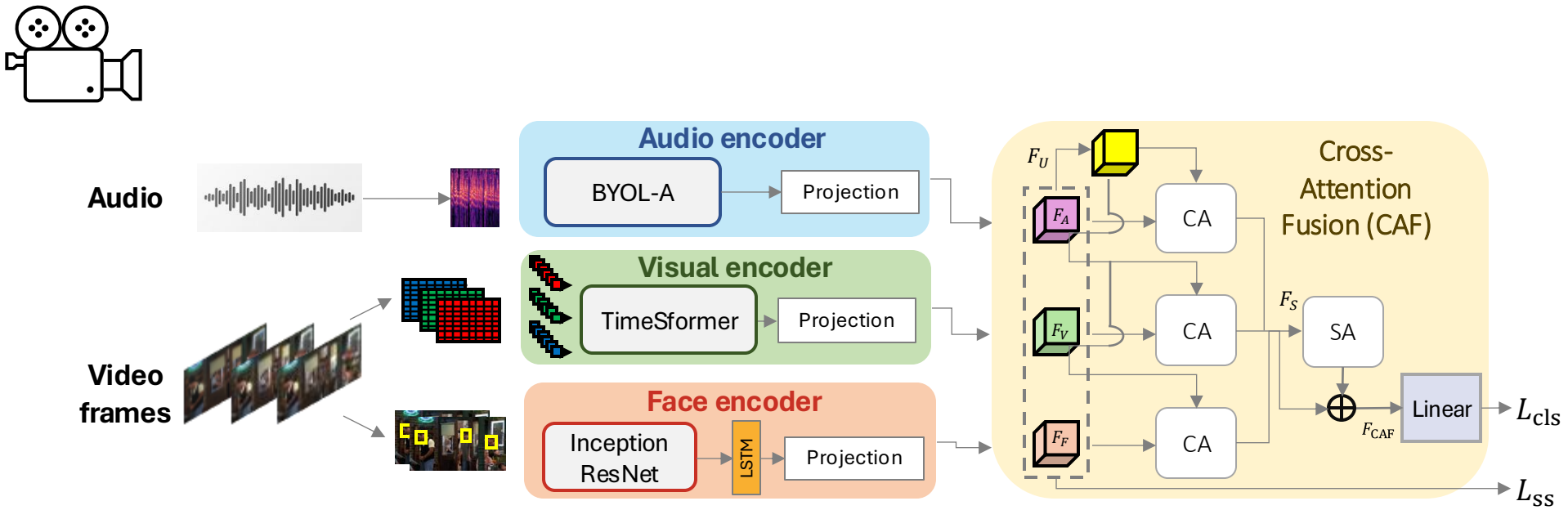
Distance between audio and visual vectors:
- Small: AV from the same place in a video (Positives)
- Large: AV from different videos (Negatives)

FunnyNet: Audiovisual learning of funny moments in videos



[Liu et al, ACCV 2022, IJCV 2024]

FunnyNet


Audio
Video frames
Audio encoder
BYOL-A
Projection
Visual encoder
TimeSformer
Projection
Face encoder
Inception ResNet
LSTM
Projection
Cross-Attention Fusion (CAF)
F_U
F_A
F_V
F_F
CA
CA
CA
F_S
SA
F_{CAF}
Linear
L_{cls}
L_{ss}

FunnyNet-W
Sitcom with Canned Laughter

FunnyNet-W
Sitcom without Canned Laughter

Self-supervised learning categories

Thank you