Self-supervised Learning

Learning without labels

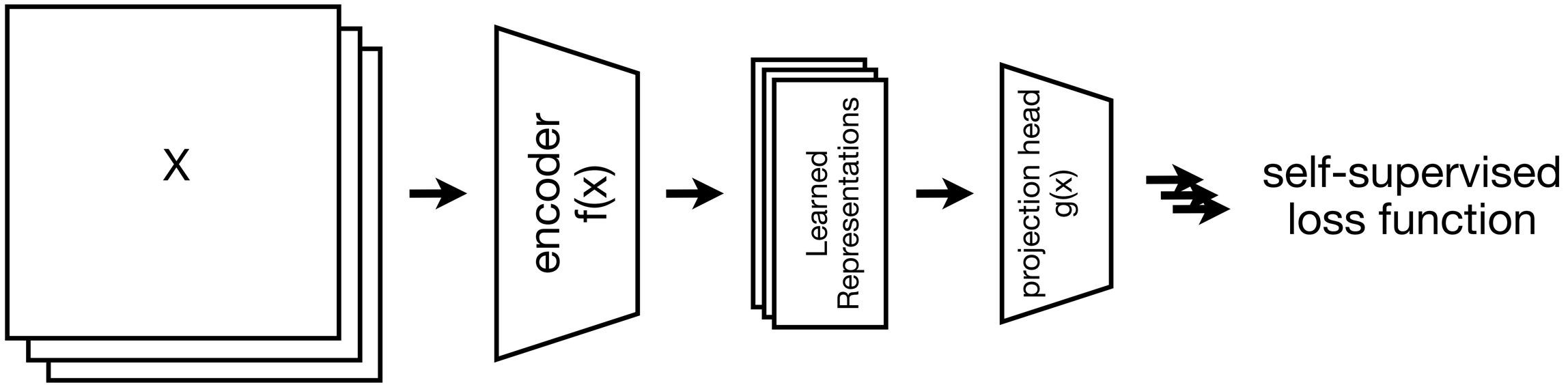
Antonia Marcu & Jonathon Hare

What and why?

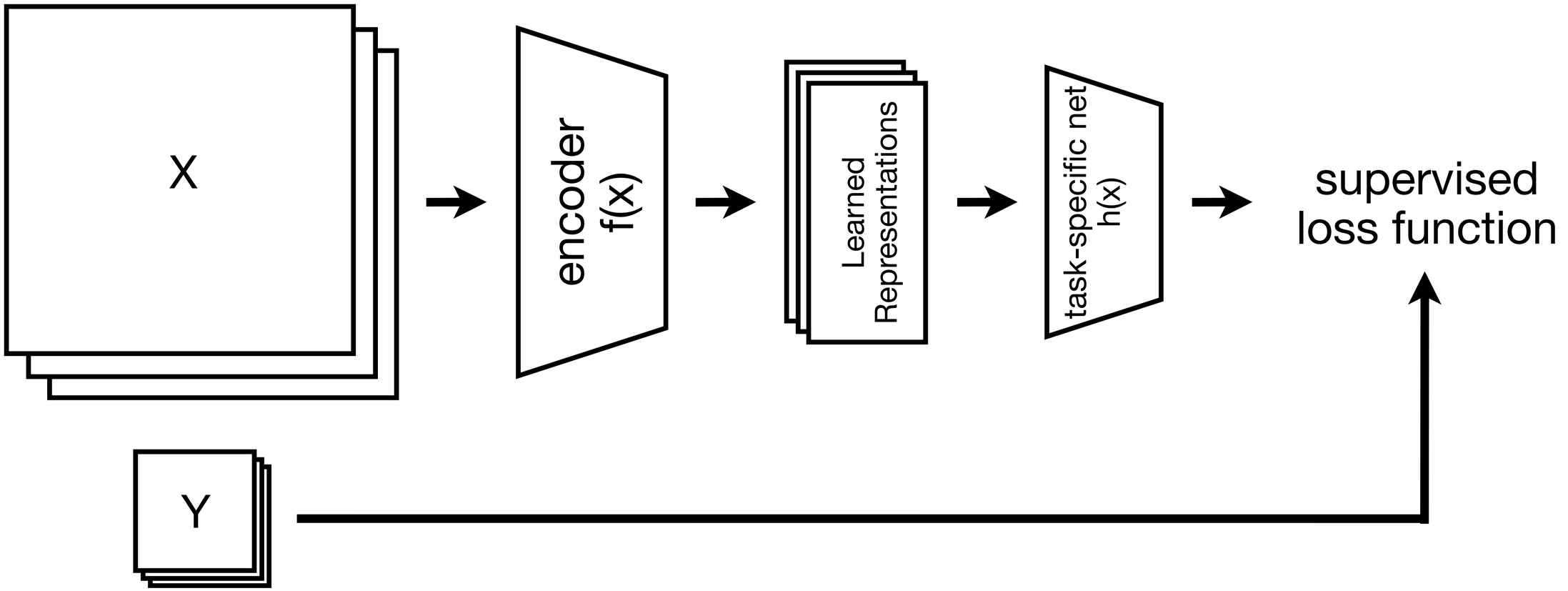
- How might we train large models to do something useful with very little labelled data (but lots of unlabelled data)?
- Can we somehow learn embeddings that are useable for downstream tasks?

- **Self-supervised Learning might be the answer!**

Basic idea: train model on the unlabelled data in some way (this is often referred to as learning the "pretext task")



Basic idea: then train a small task-specific network (and possibly fine-tune encoder) on a small labelled dataset



Types of SSL

- Auto-regressive SSL
- Auto-encoders
  - Noisy/masked auto-encoders
- Contrastive SSL
- Non-contrastive SSL

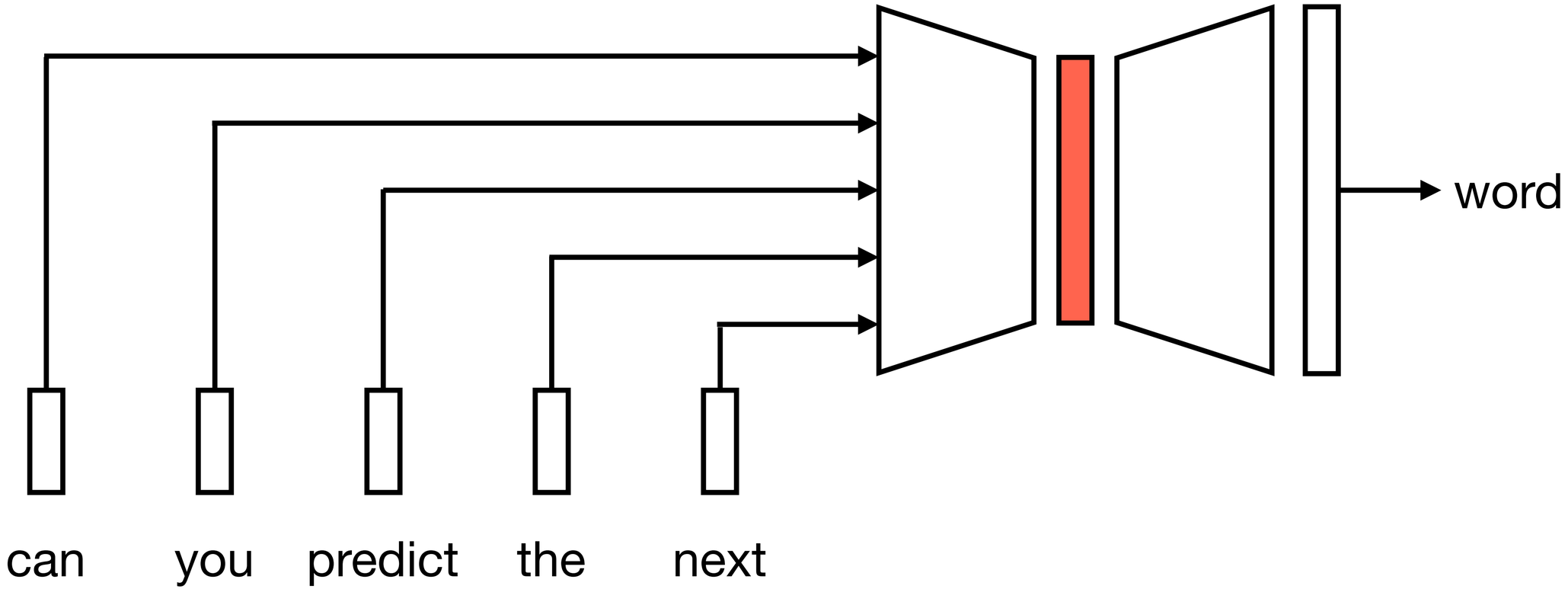
Types of SSL

- Auto-regressive SSL — *Encoder-decoder*
- Auto-encoders
  - Noisy/masked auto-encoders

- Contrastive SSL — *Siamese*
- Non-contrastive SSL

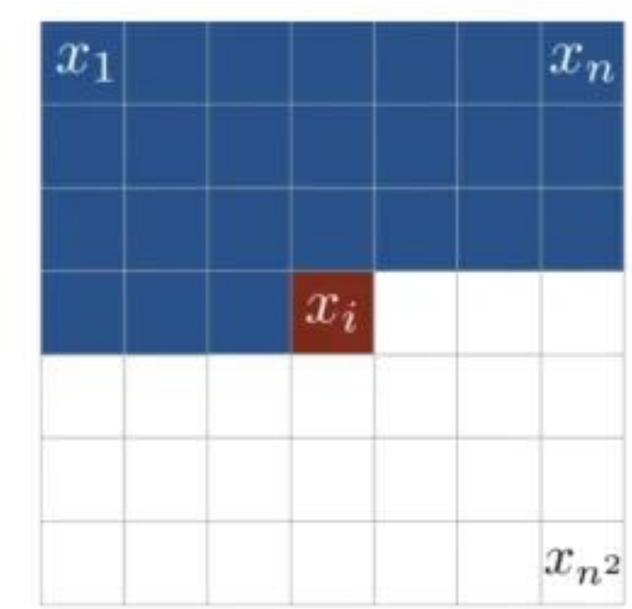
Auto-regressive SSL

Learning what comes next



Auto-regressive Image Modelling

PixelRNN

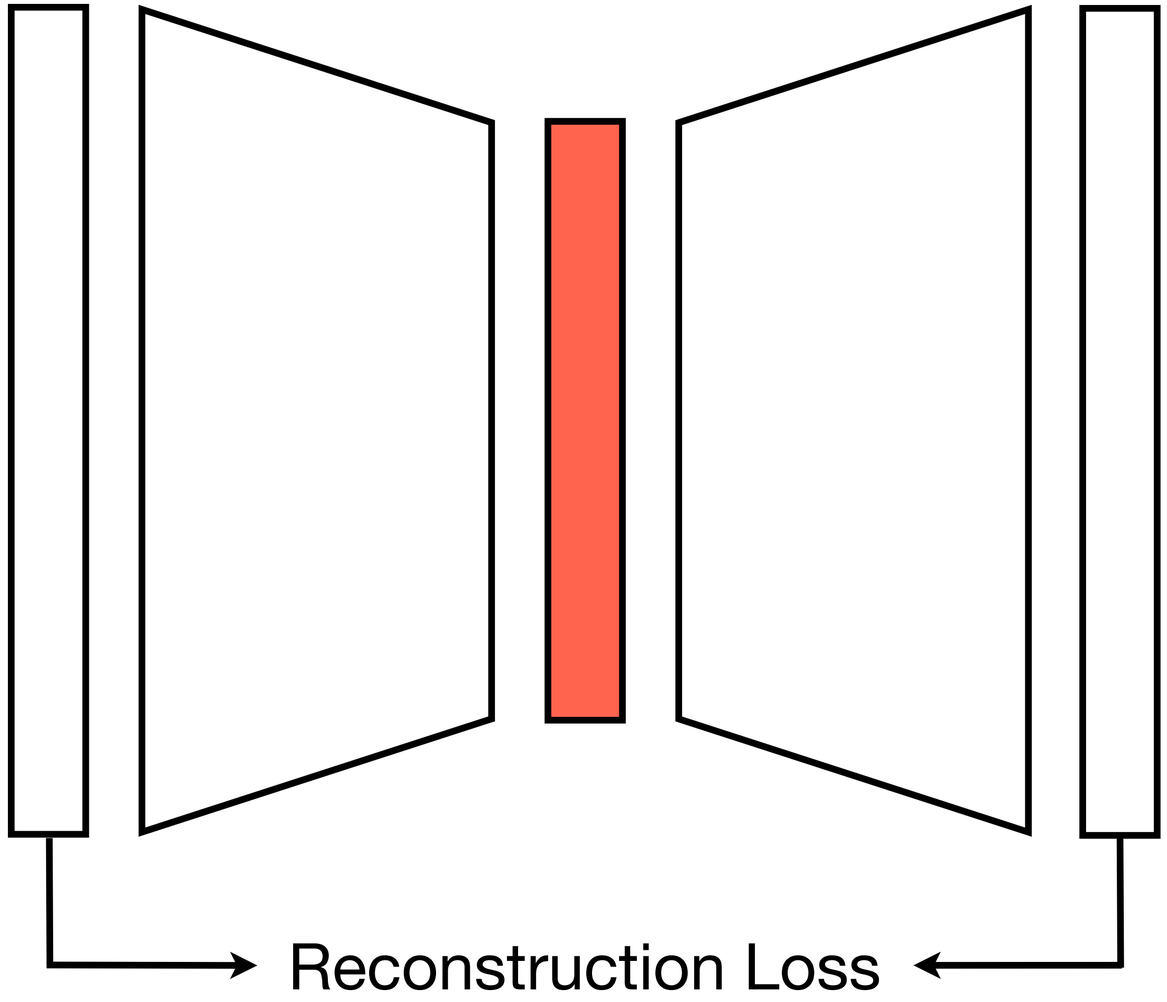


$$p(\mathbf{x}) = \prod_{i=1}^{n^2} p(x_i | x_1, ..., x_{i-1})$$

Pixel Recurrent Neural Networks

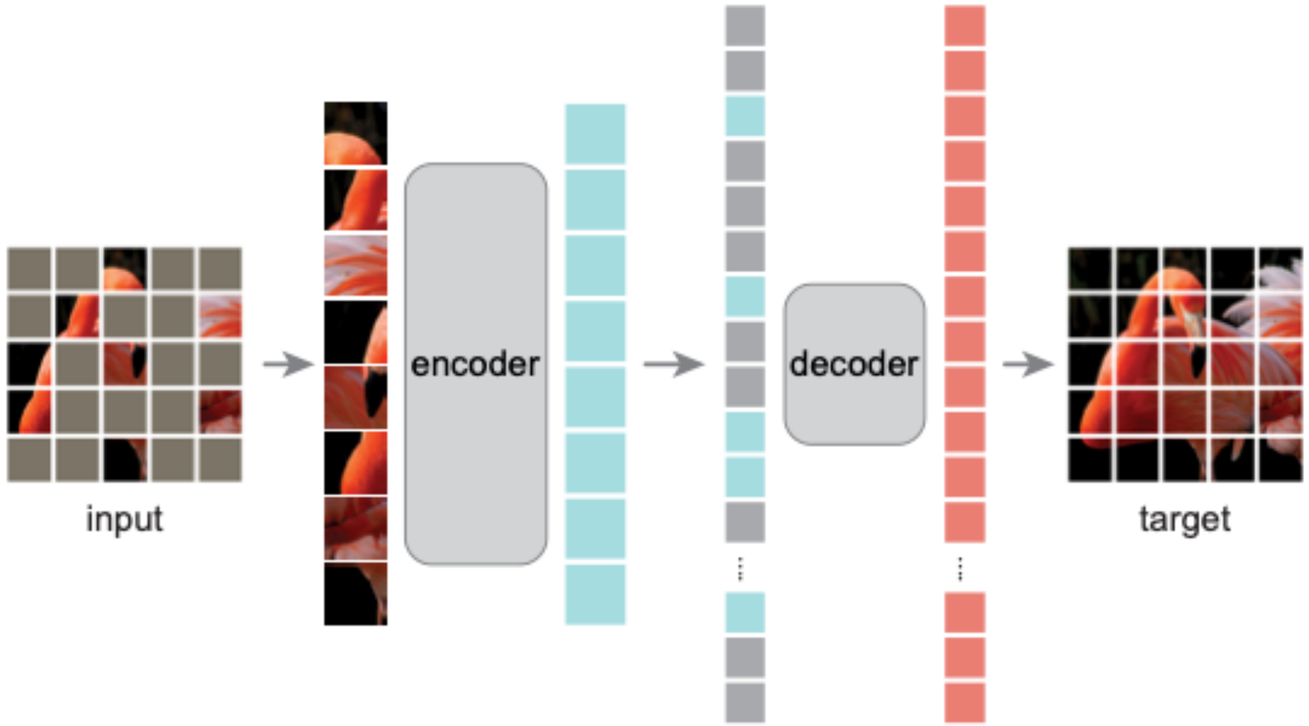
Autoencoders

Learning to compress / Learning to reconstruct



Autoencoders: Image Modelling

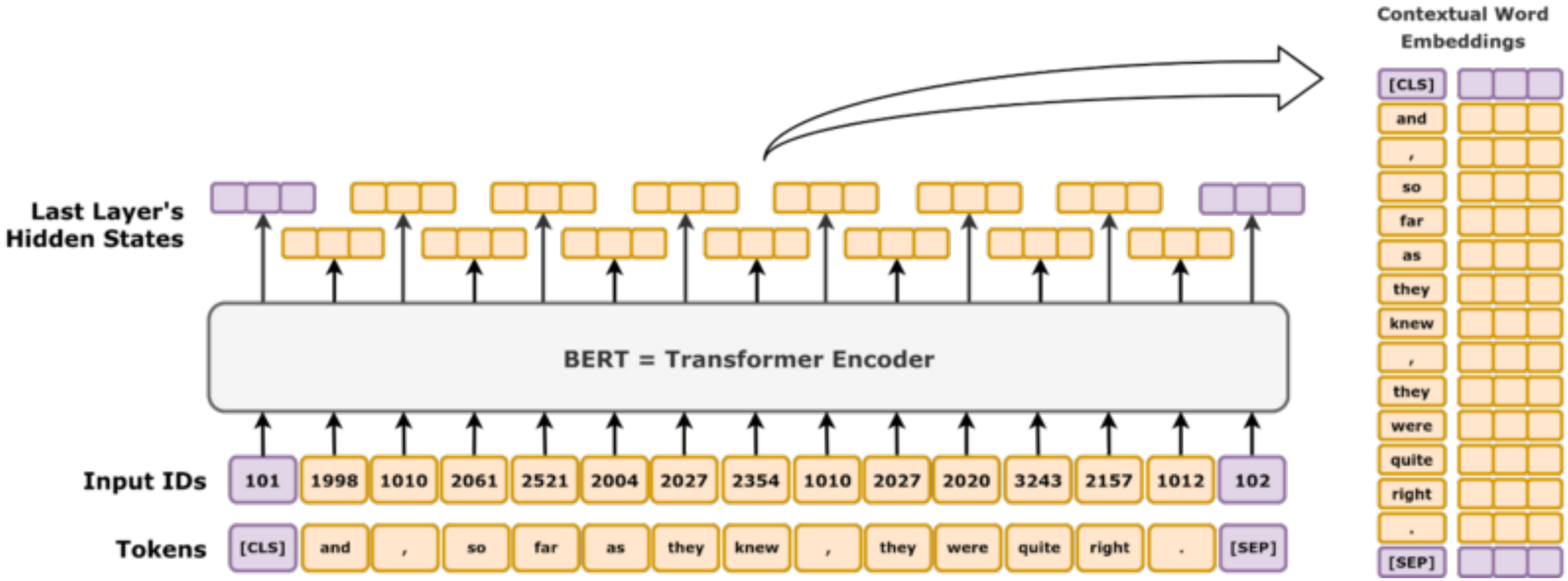
Masked Autoencoders



Masked Autoencoders Are Scalable Vision Learners

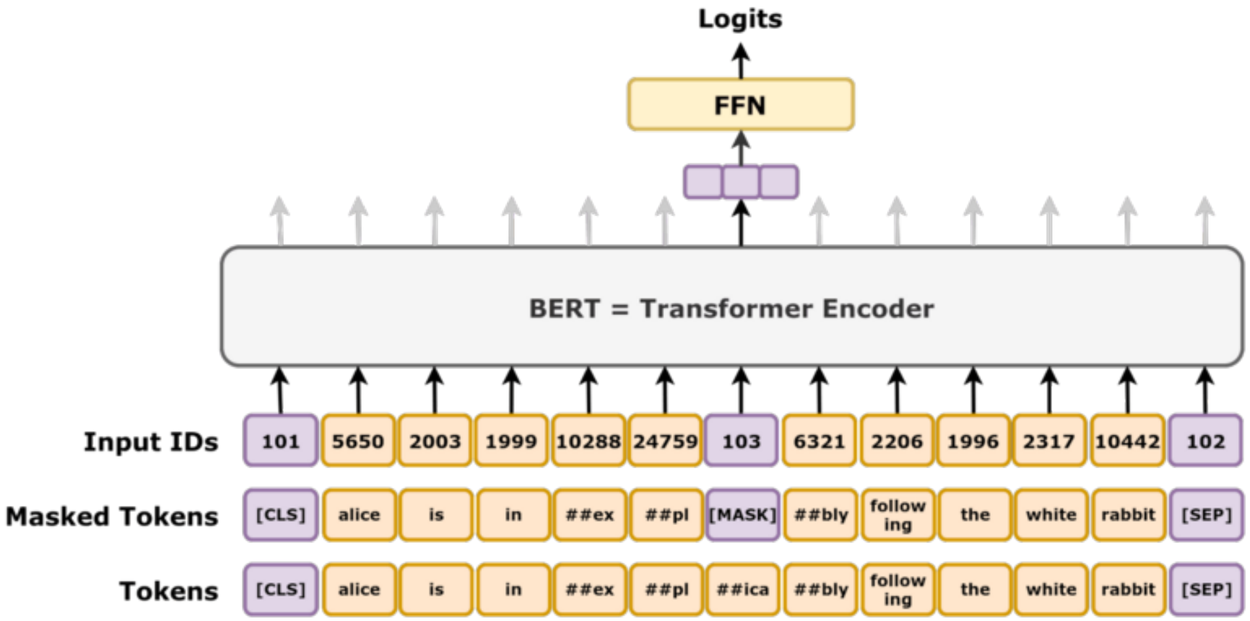
Autoencoders: Text Modelling

BERT

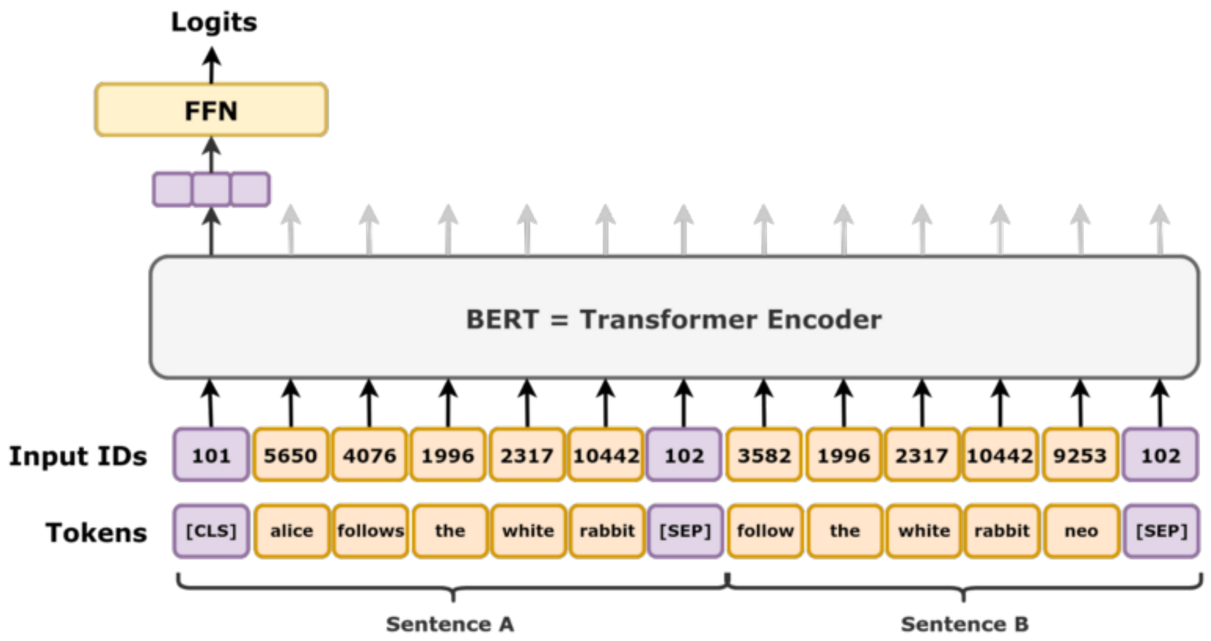


Autoencoders: Text Modelling

BERT



Masked prediction

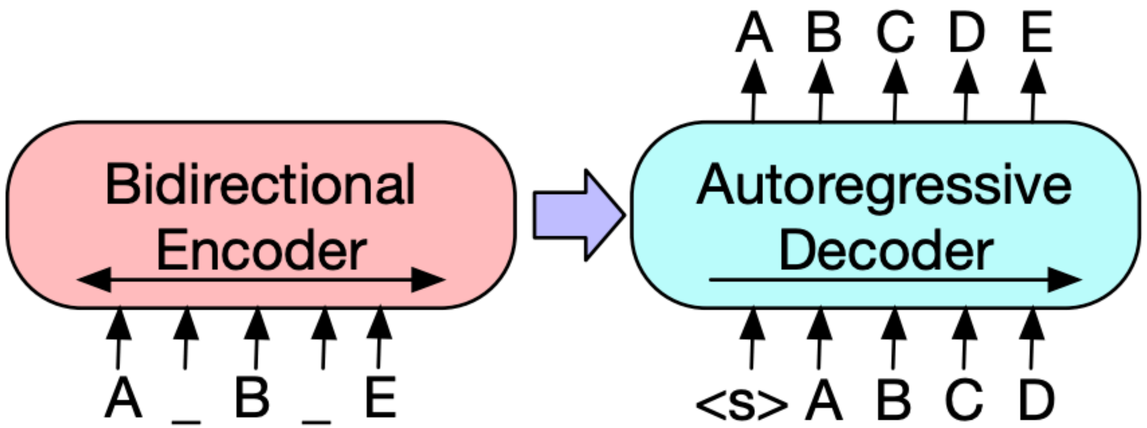


Next sentence prediction

Autoencoders: Text Modelling

BART

predict clean text



train on noisy text

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.

Exploiting Similarity in the Embedding Space

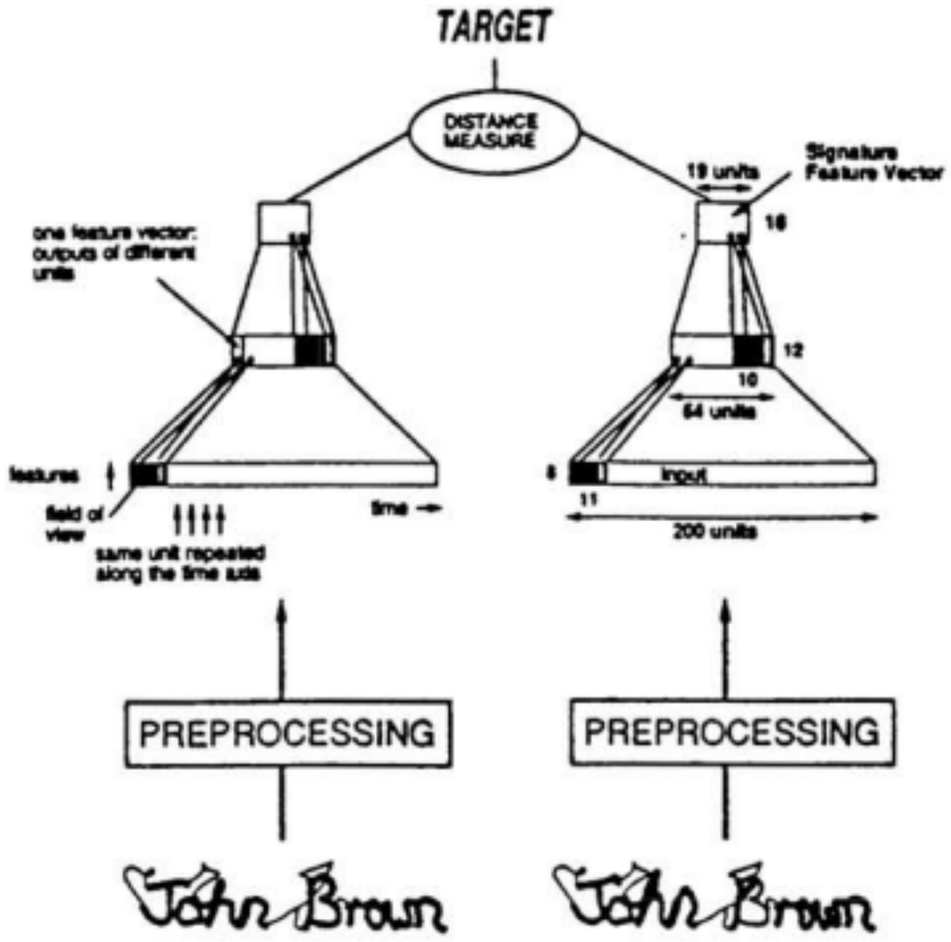
Quick Nomenclature

Embedding space = Latent space = Representation Space*

*ish. Some of the papers we will discuss today use **Representation Space** to mean something very specific. Other use **Embedding space** to mean something specific

Similarity in the Embedding Space

Siamese Nets



Signature Verification using a "Siamese" Time Delay Neural Network

Can we turn this on its head?

Joint Embedding Learning

- Intuition: Learn embeddings by training two identical networks to represent similar inputs in similar ways

Joint Embedding Learning

- Intuition: Learn embeddings by training two identical networks to represent similar inputs in similar ways

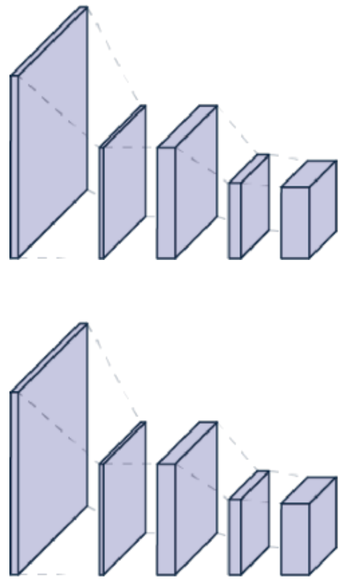
Joint Embedding Learning

- Intuition: Learn embeddings by training two identical networks to represent similar inputs in similar ways

How can we get similar inputs in an unsupervised way?

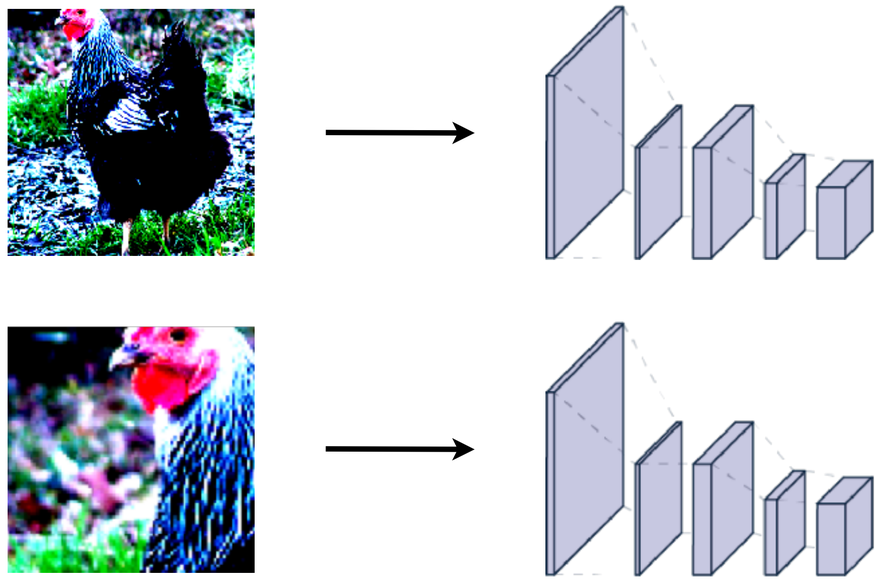
Joint Embedding Learning

- Intuition: Learn embeddings by training two identical networks to represent similar inputs in similar ways



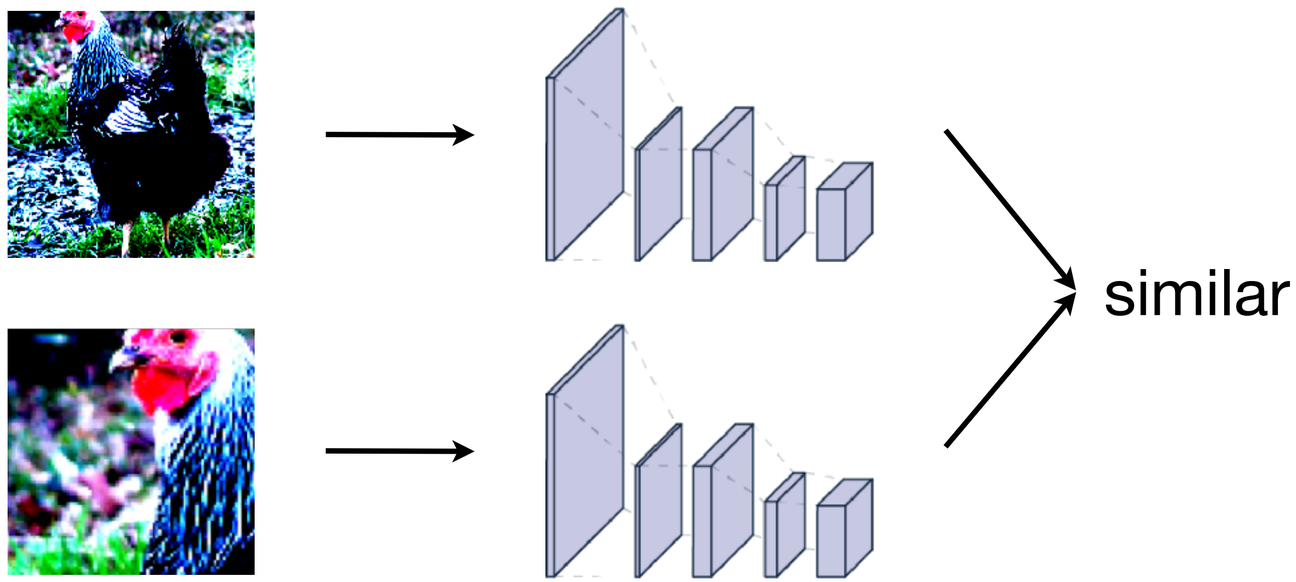
Joint Embedding Learning

- Intuition: Learn embeddings by training two identical networks to represent similar inputs in similar ways



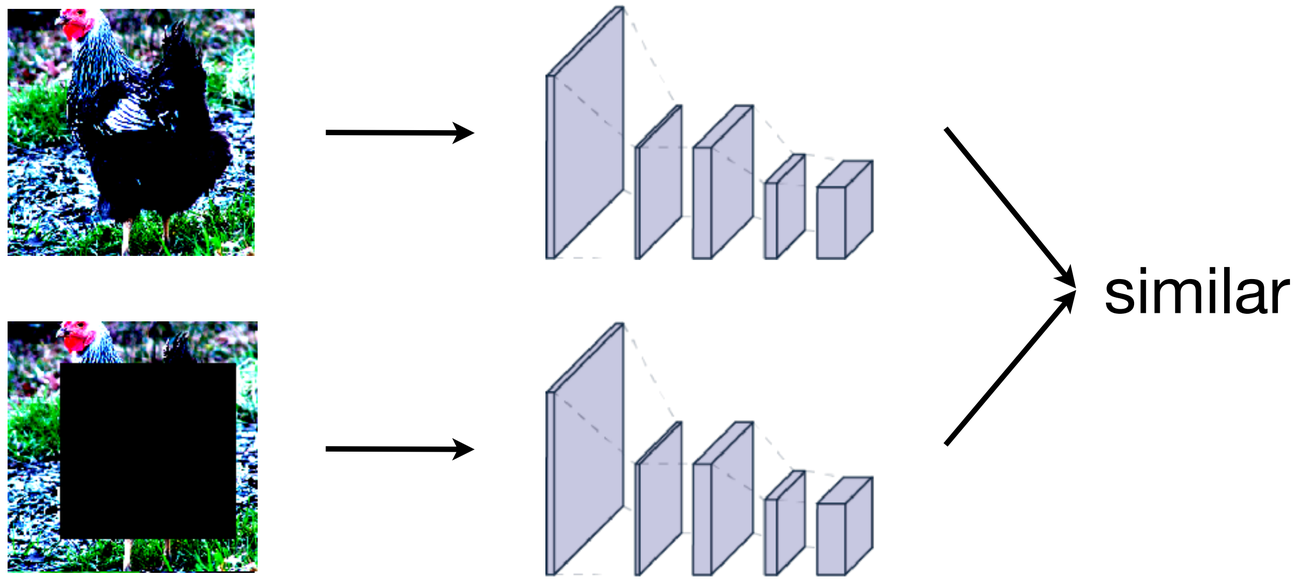
Joint Embedding Learning

- Intuition: Learn embeddings by training two identical networks to represent similar inputs in similar ways



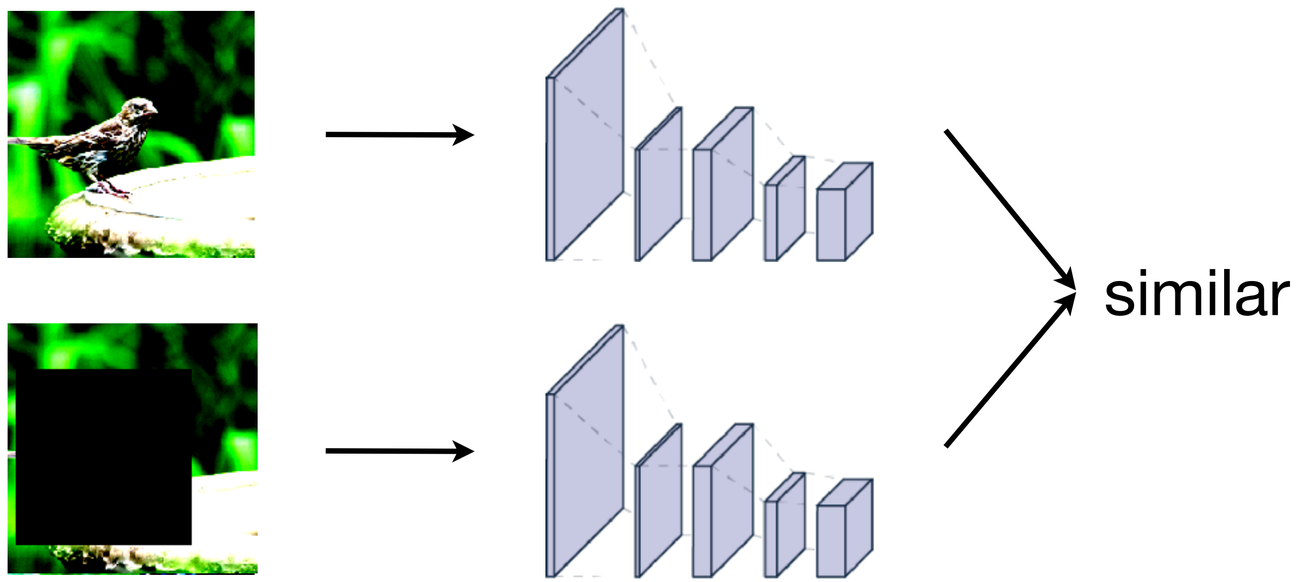
Joint Embedding Learning

- Intuition: Learn embeddings by training two identical networks to represent similar inputs in similar ways



Joint Embedding Learning

- Intuition: Learn embeddings by training two identical networks to represent similar inputs in similar ways



What can go wrong?

Preventing Collapse

Pushing dissimilar embeddings away from each other

- Contrastive Learning
- Information Maximisation

Contrastive SSL

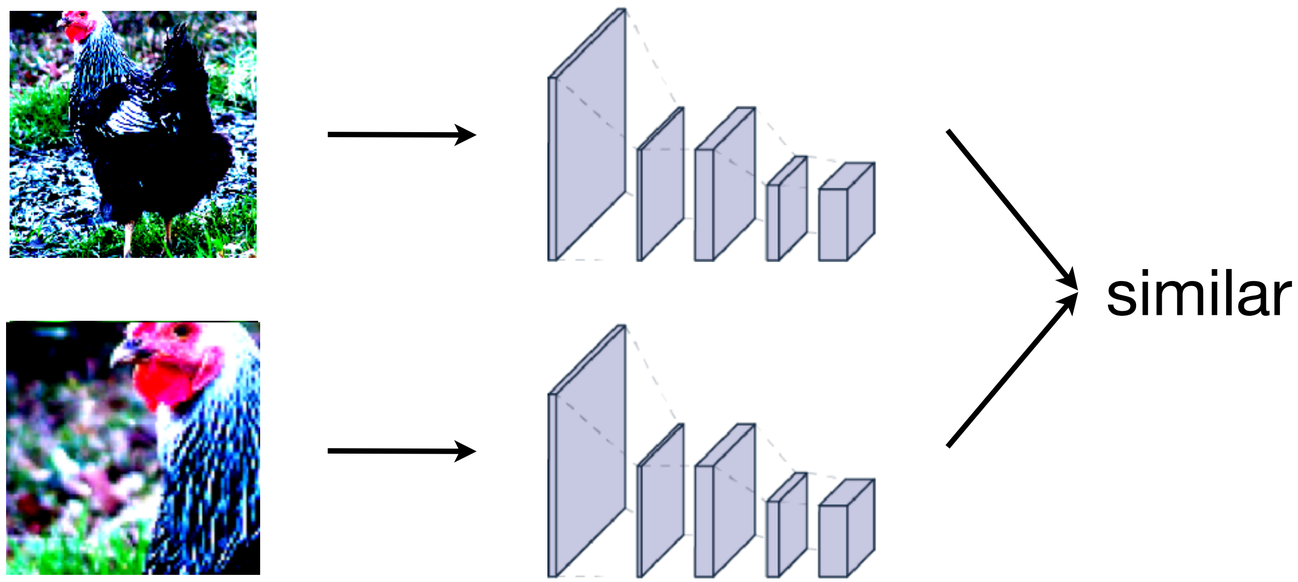
Learning by maximising similarity and differences

- Minimise representational distance between similar samples while maximising distance between dissimilar samples

Contrastive SSL

Learning by maximising similarity and differences

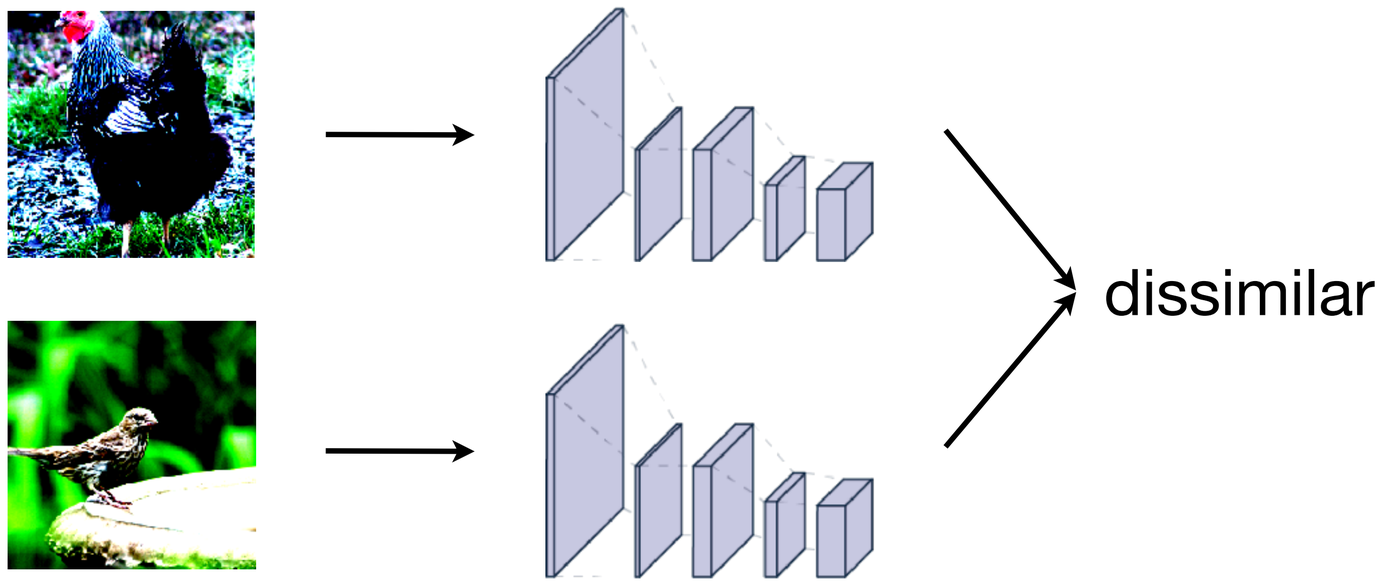
- Minimise representational distance between similar samples while maximising distance between dissimilar samples



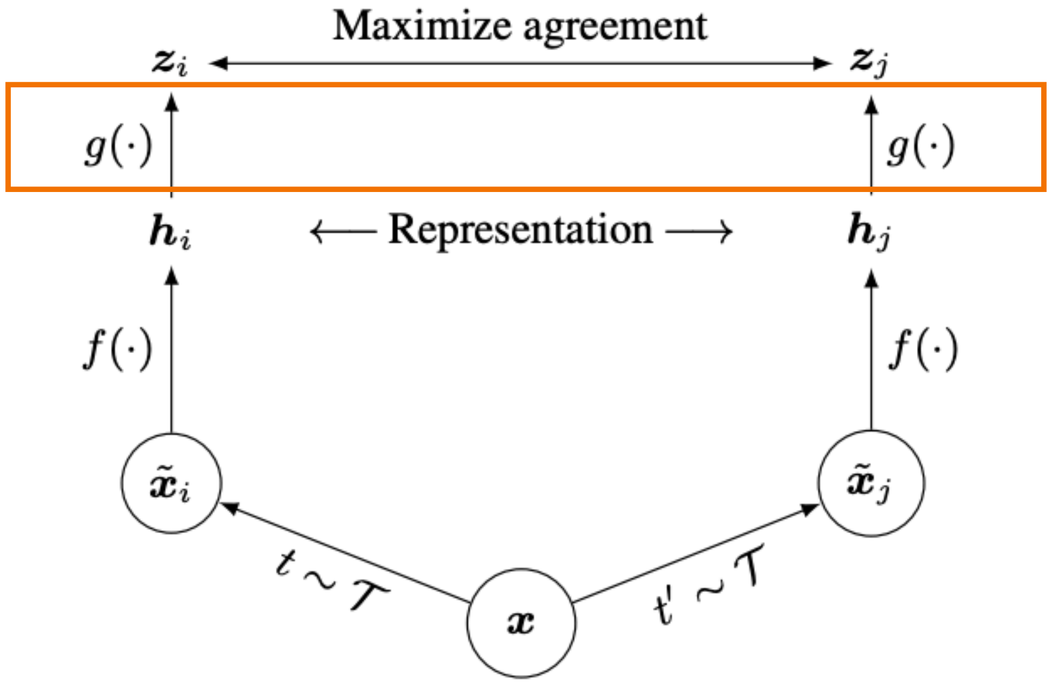
Contrastive SSL

Learning by maximising similarity and differences

- Minimise representational distance between similar samples while maximising distance between dissimilar samples



SimCLR



MLP with
one hidden layer

"A Simple Framework for Contrastive Learning of Visual Representations"

SimCLR

- In an N-dimensional batch:
  - Positive samples (similar): augmented versions of each image (2N total)
  - Negative samples (different): all other samples in the batch

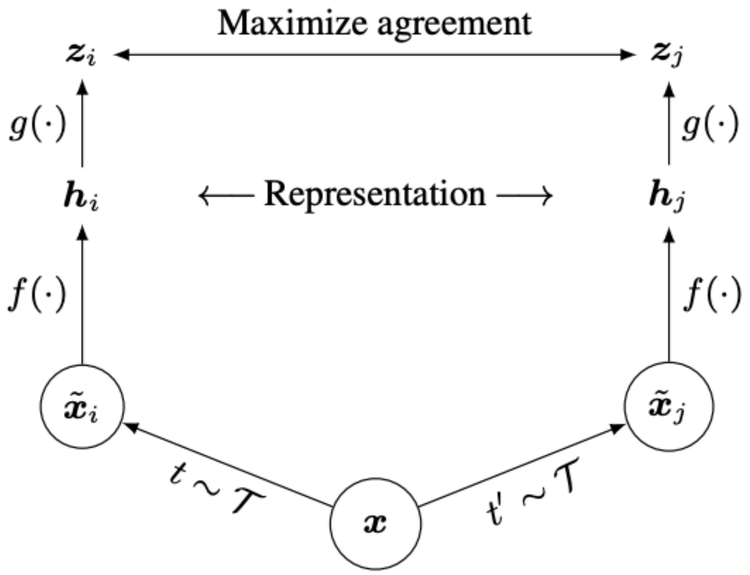
"A Simple Framework for Contrastive Learning of Visual Representations"

SimCLR

- In an N-dimensional batch:
  - Positive samples (similar): augmented versions of each image (2N total)
  - Negative samples (different): all other samples in the batch

"A Simple Framework for Contrastive Learning of Visual Representations"

SimCLR



cosine

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

temperature scaling

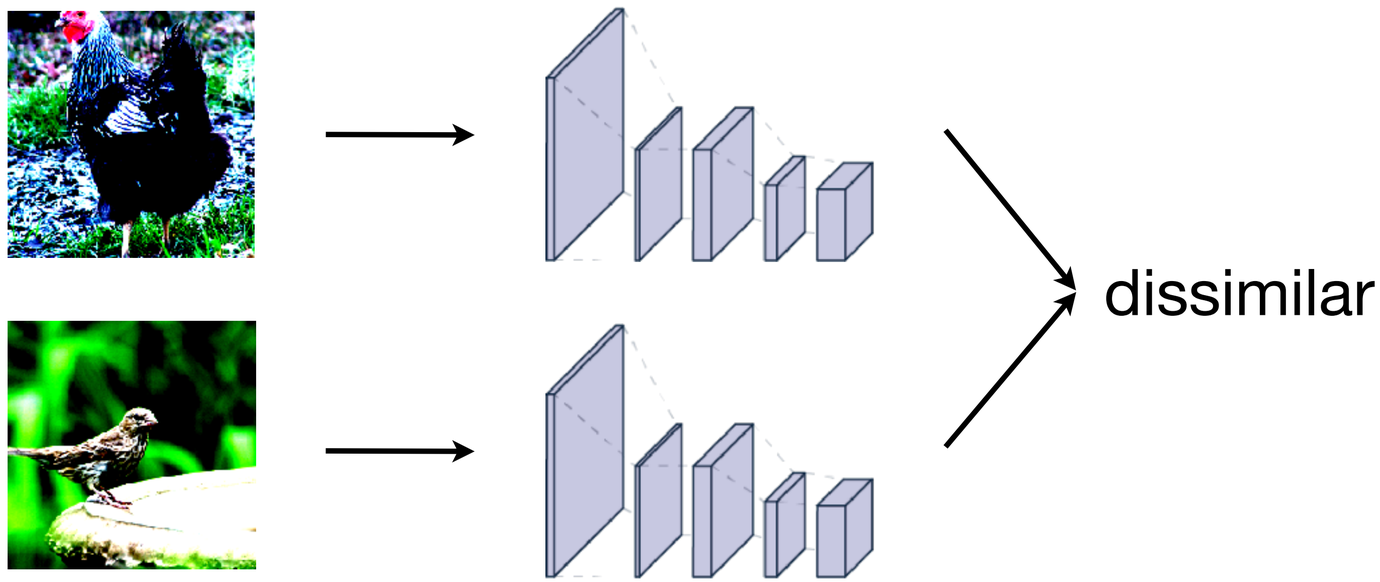
"A Simple Framework for Contrastive Learning of Visual Representations"

What can go wrong?

Contrastive SSL

Learning by maximising similarity and differences

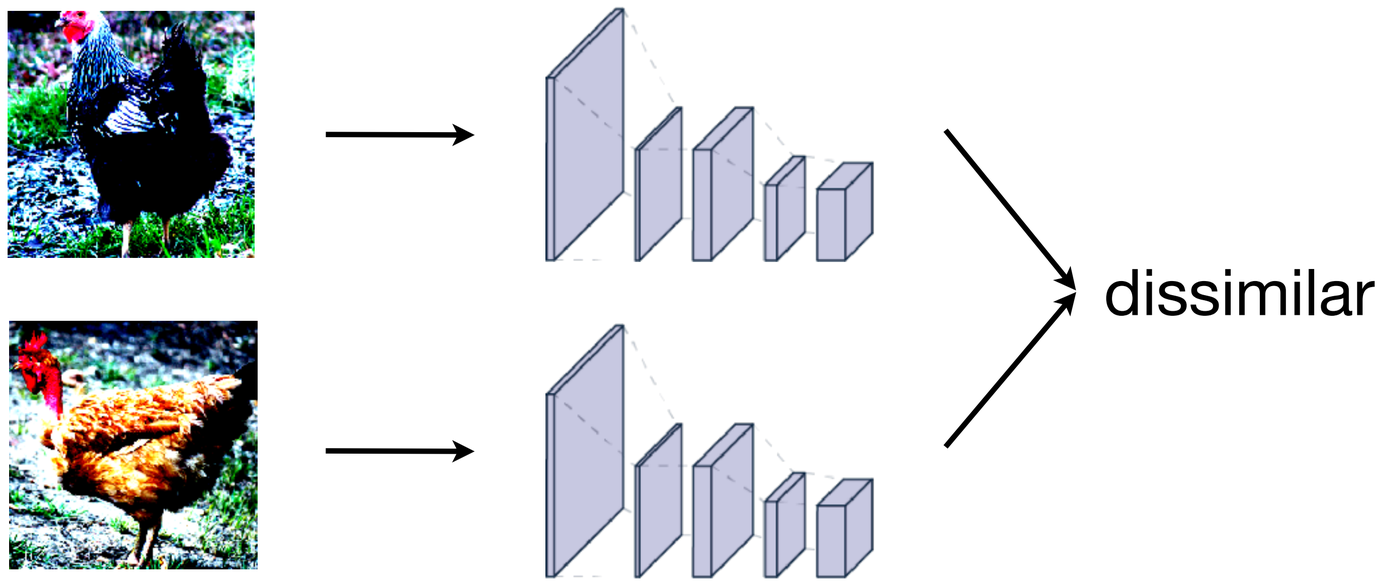
- Minimise representational distance between similar samples while maximising distance between dissimilar samples



Contrastive SSL

Learning by maximising similarity and differences

- Minimise representational distance between similar samples while maximising distance between dissimilar samples



BYOL

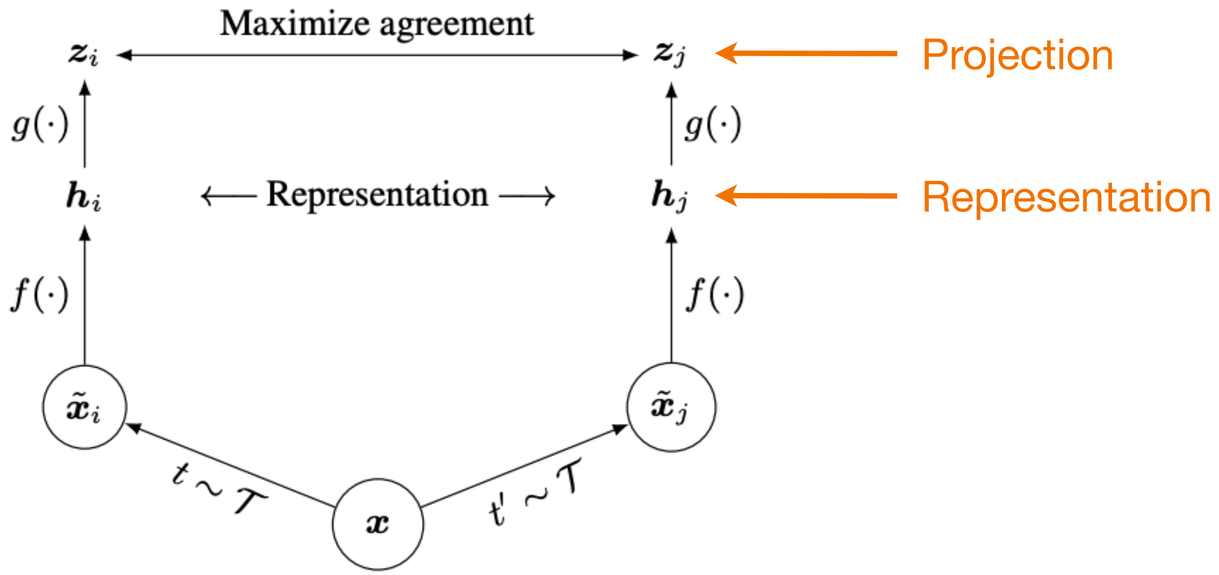
No negative samples needed

BYOL

- Intuition: Given the representation* of a sample, can we learn to predict what the representation* of the augmented sample will look like?

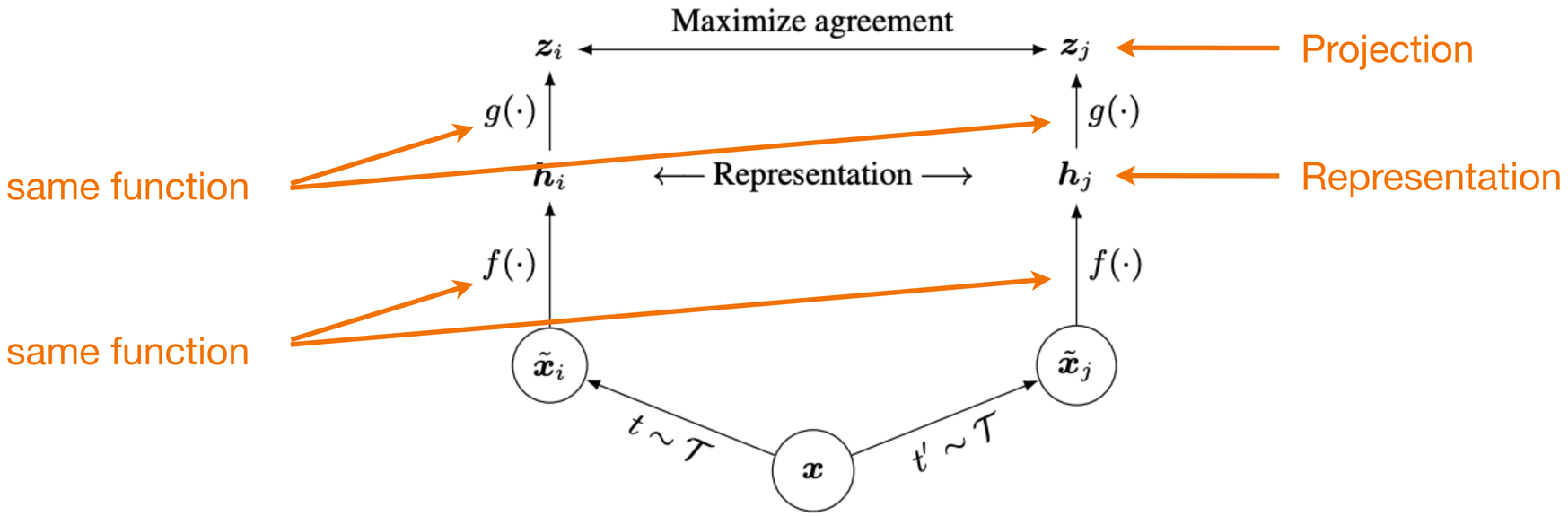
* used in the generic sense here

Reminder



"A Simple Framework for Contrastive Learning of Visual Representations"

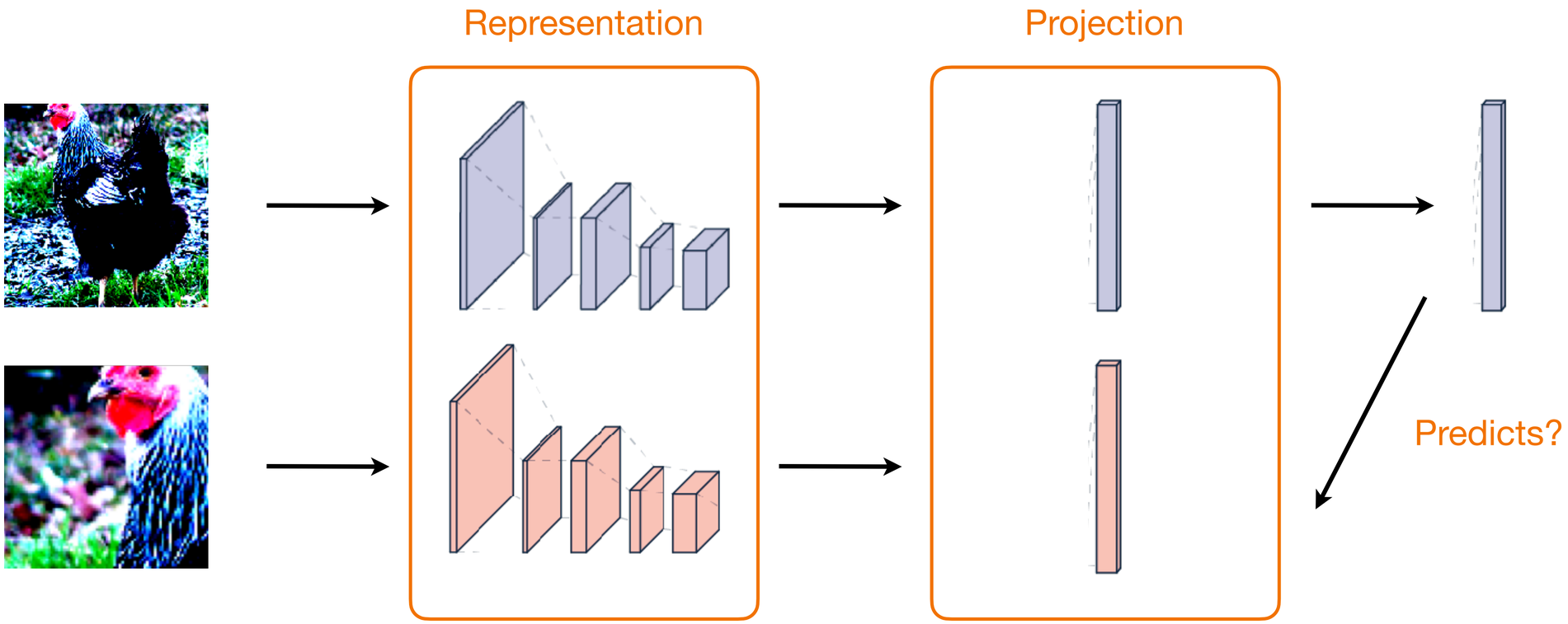
Reminder



"A Simple Framework for Contrastive Learning of Visual Representations"

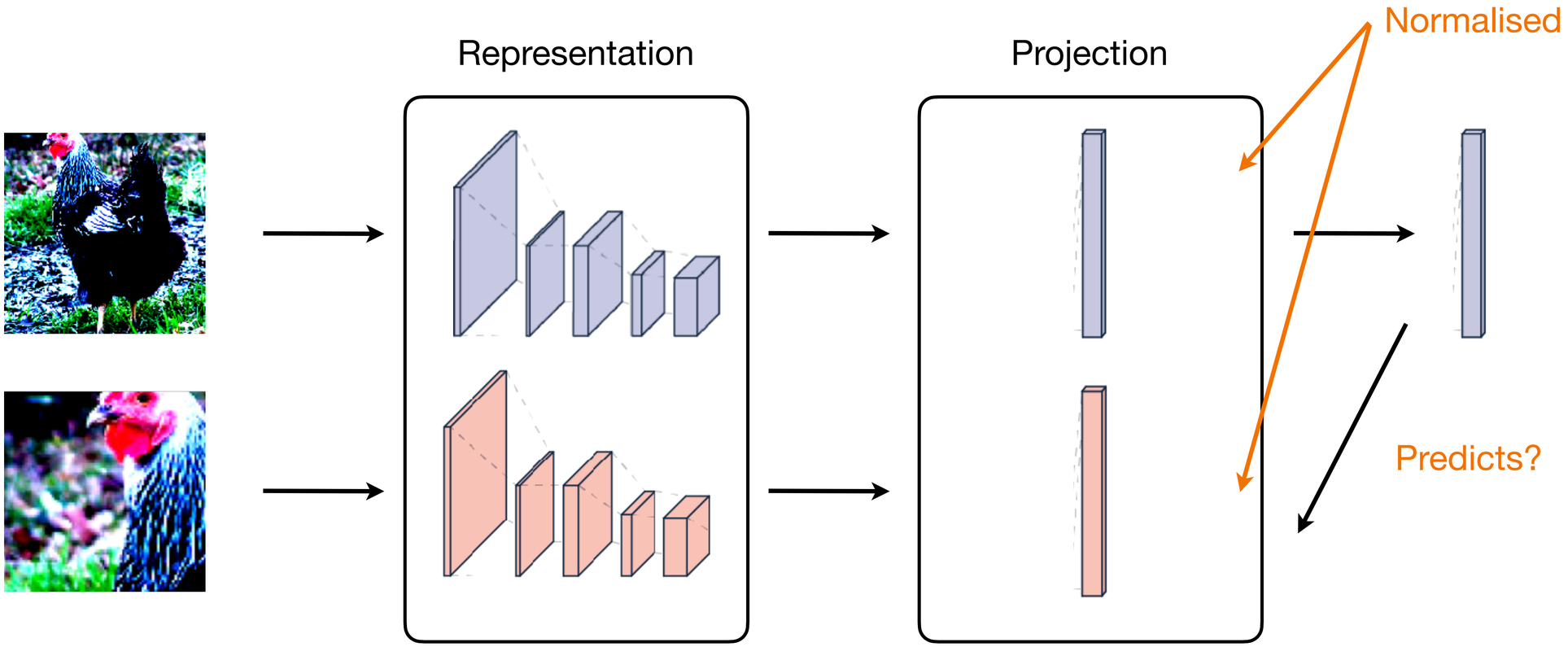
BYOL

No negative samples needed



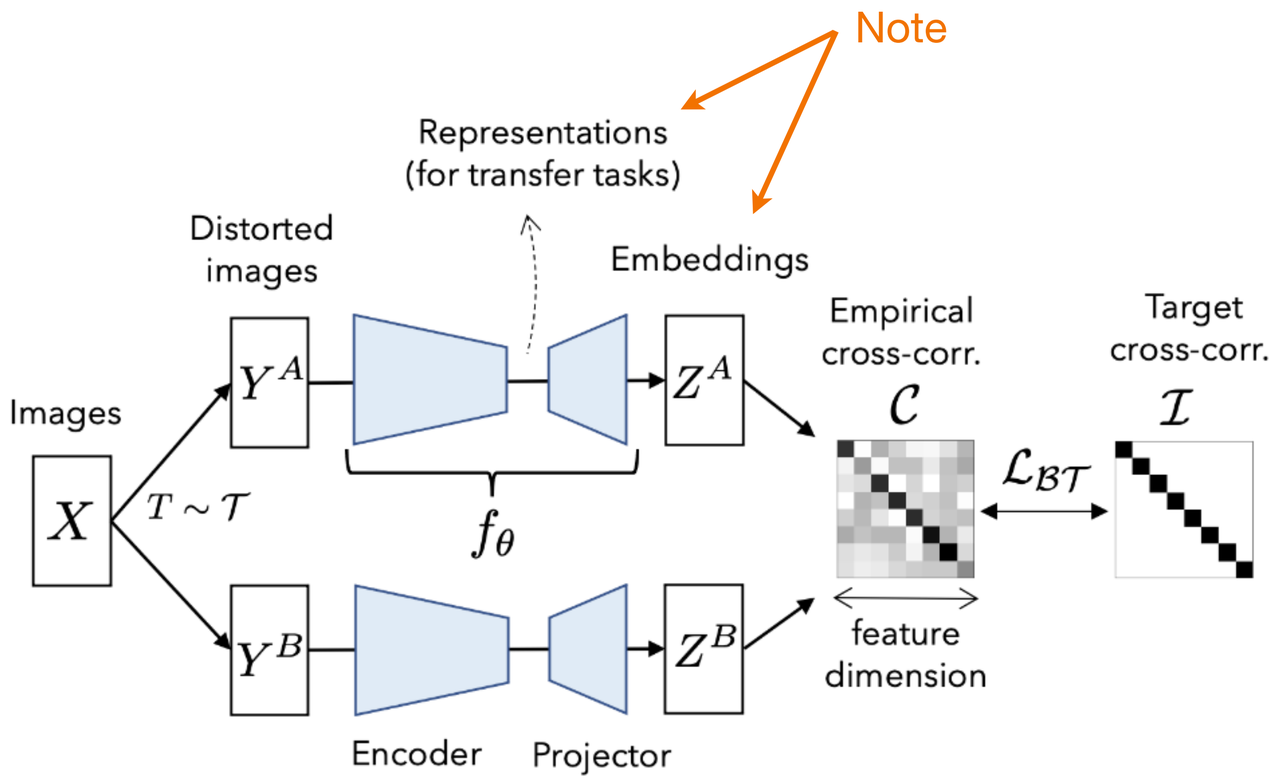
BYOL

No negative samples needed



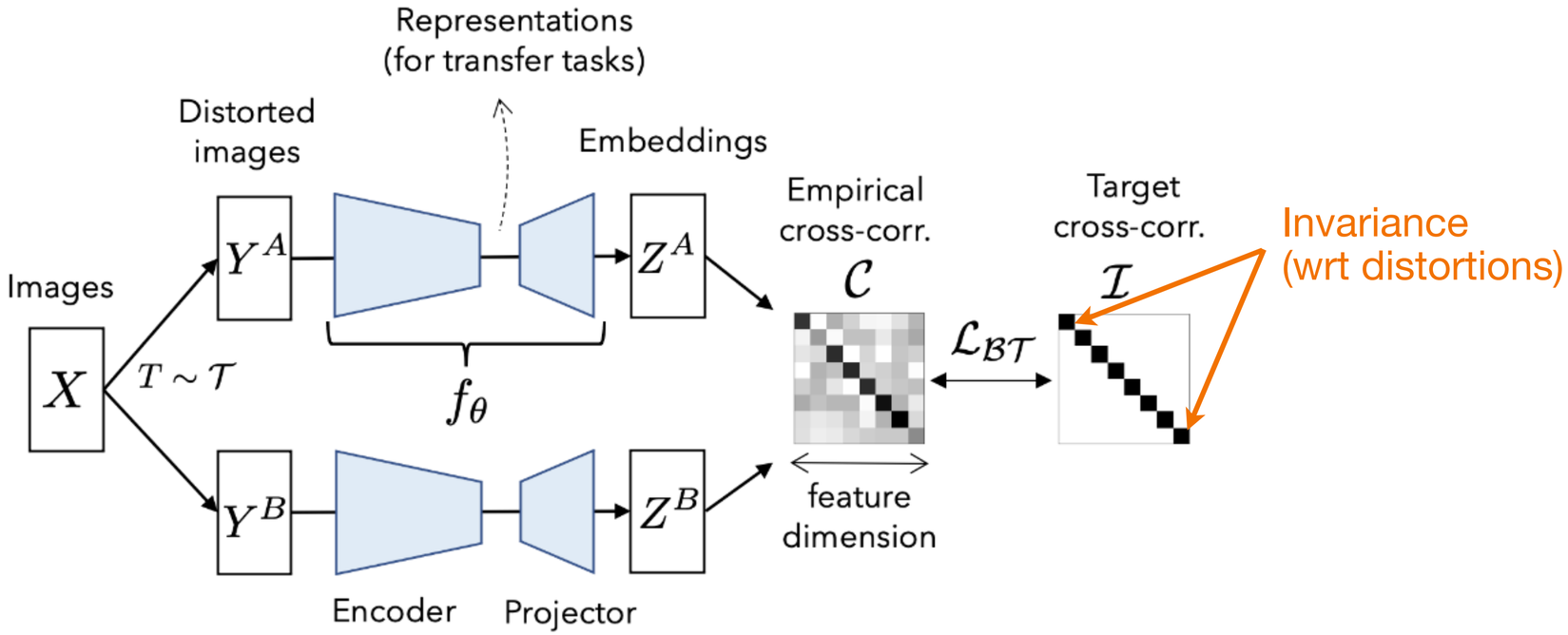
Information Maximisation

BarlowTwins



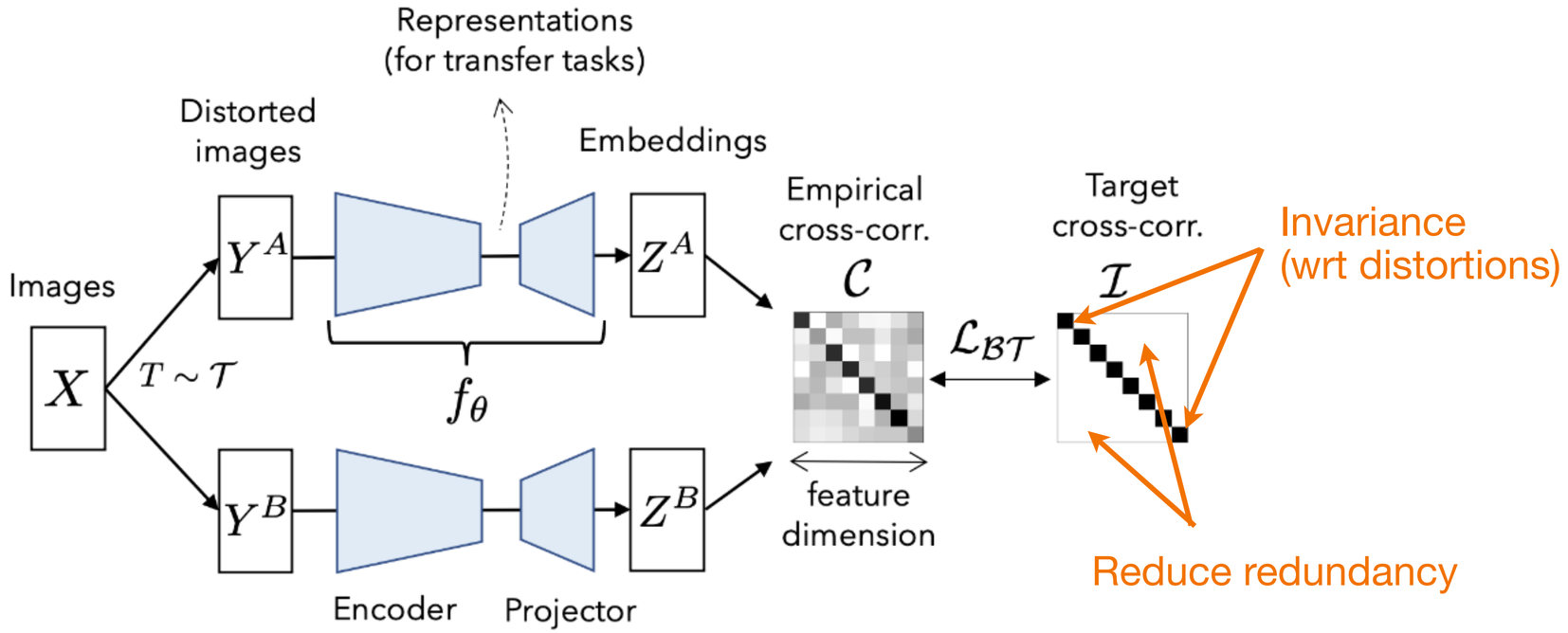
"Barlow Twins: Self-Supervised Learning via Redundancy Reduction"

BarlowTwins



"Barlow Twins: Self-Supervised Learning via Redundancy Reduction"

BarlowTwins

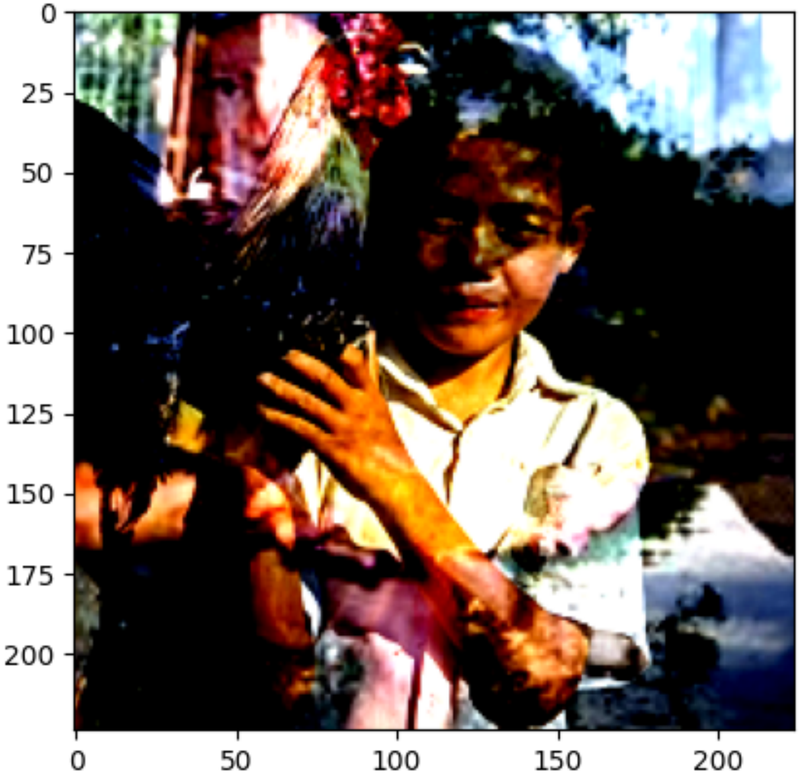


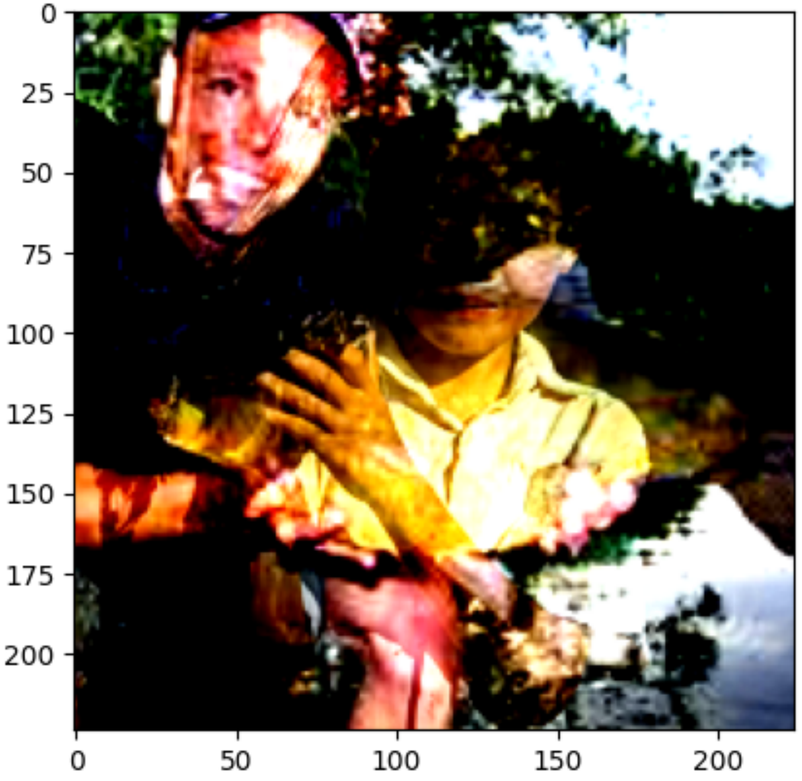
"Barlow Twins: Self-Supervised Learning via Redundancy Reduction"

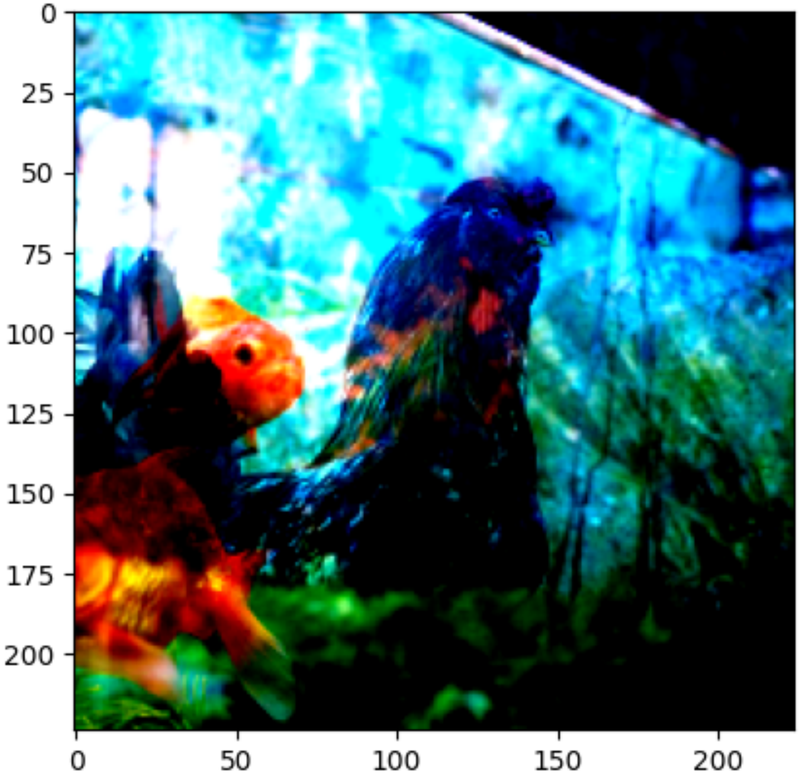
Distortions

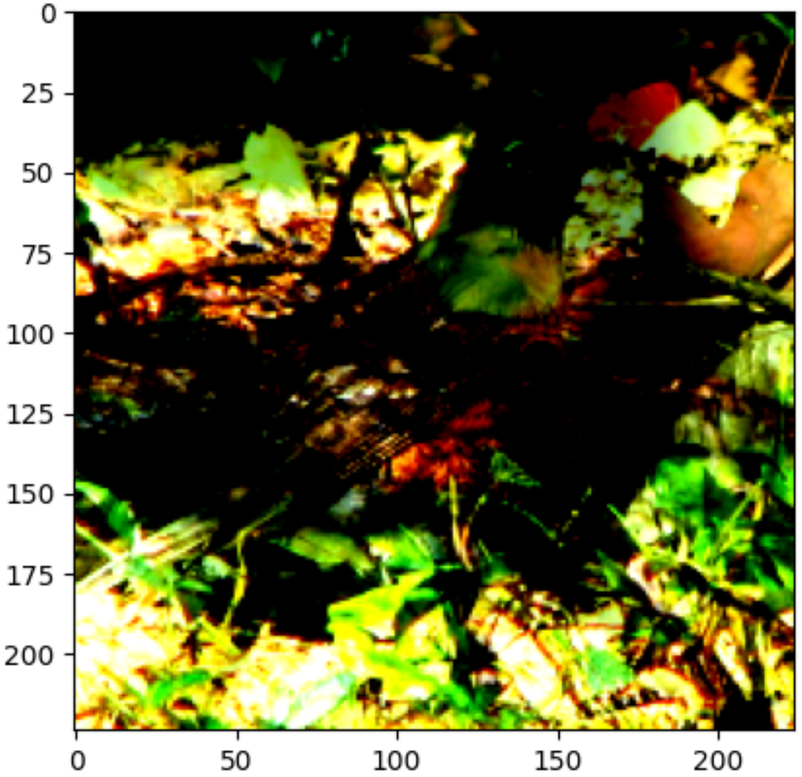
- Defining rich enough augmentations is crucial for the success of such methods

- What makes an augmentation "good"?


0
25
50
75
100
125
150
175
200
0
50
100
150
200


0
25
50
75
100
125
150
175
200
0
50
100
150
200


0
25
50
75
100
125
150
175
200
0
50
100
150
200


0
25
50
75
100
125
150
175
200
0
50
100
150
200