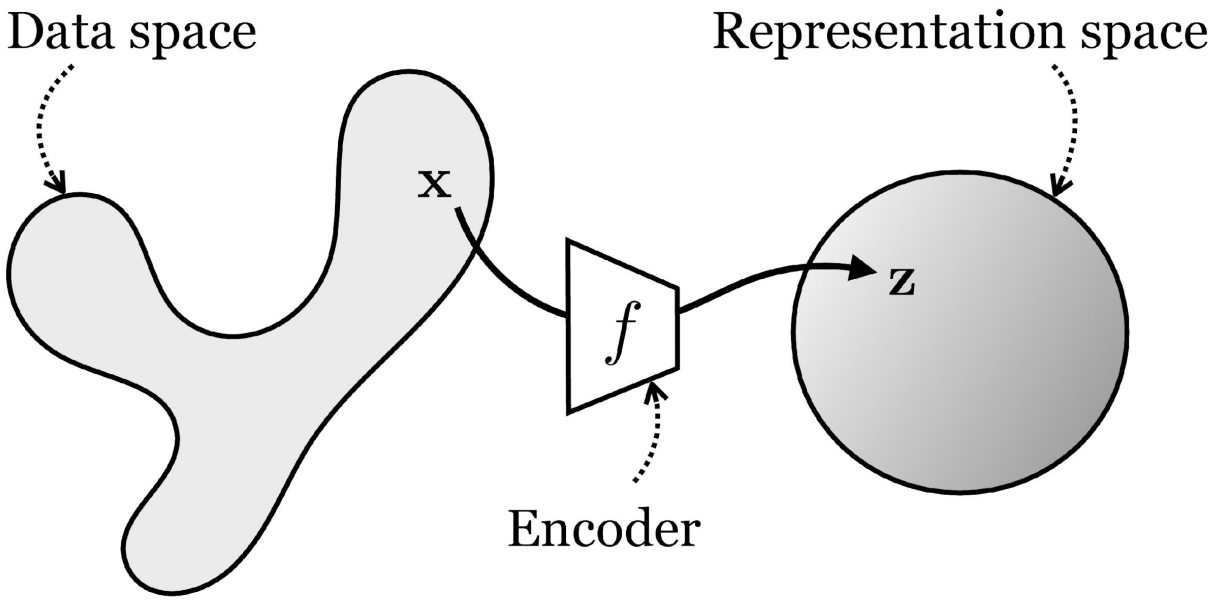
A Unifying Framework for Representation Learning

Shaden Alshammari

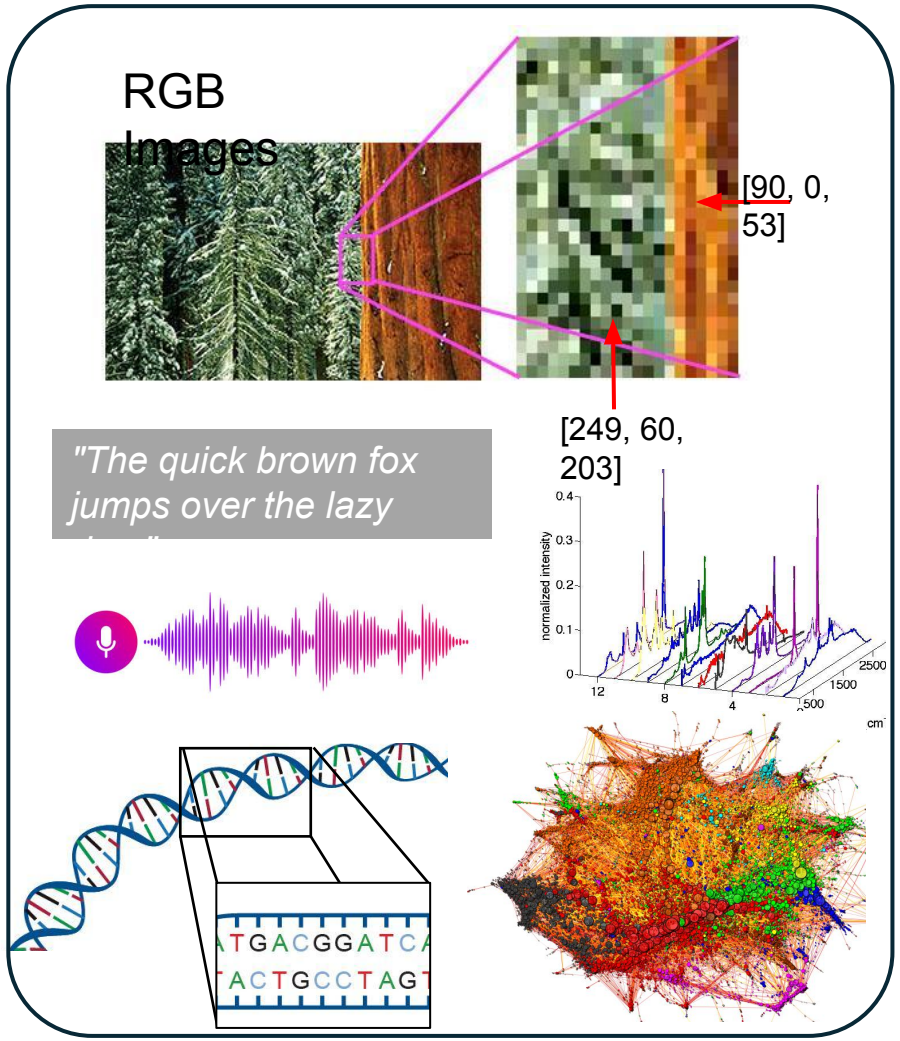
What's Representation Learning?

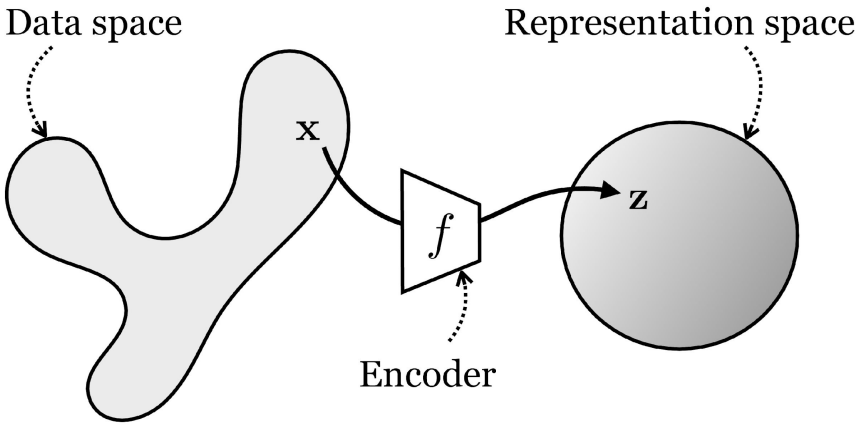


from raw, messy, high dimensional data to meaningful structure

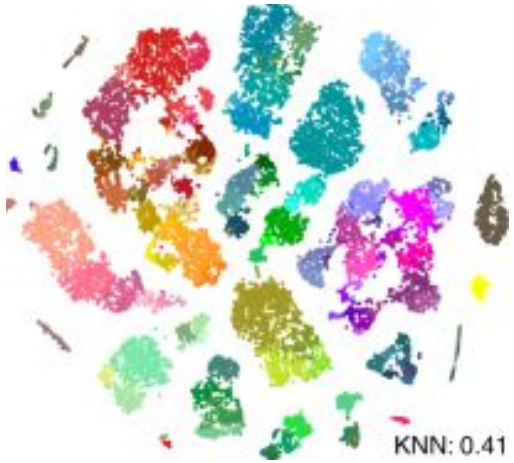
What's Representation Learning?

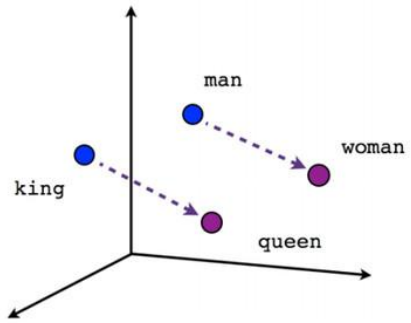
Raw Data → Meaningful Structure



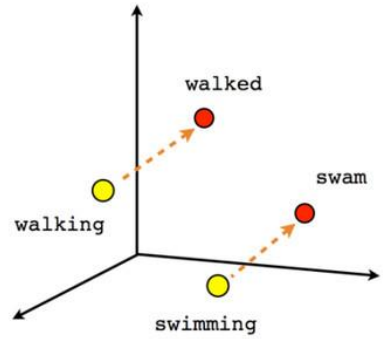


vectors





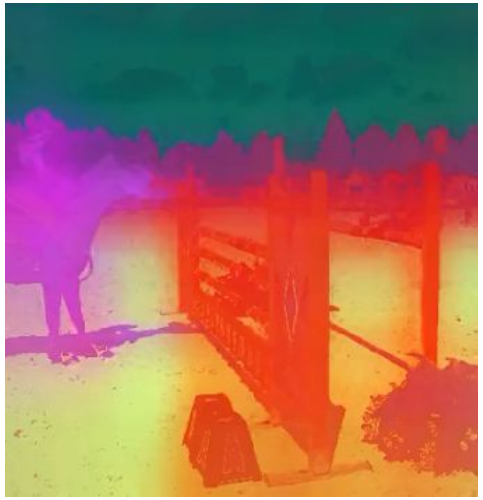
Male-Female



Verb tense

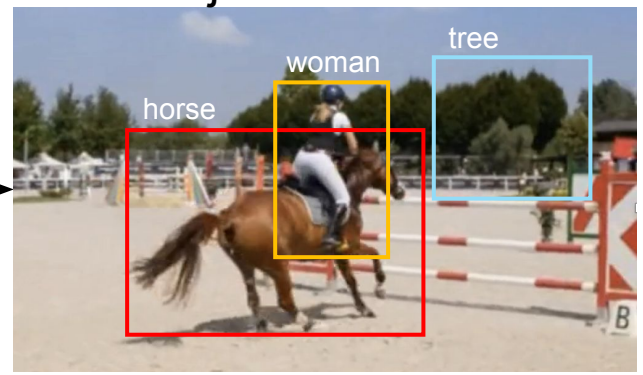
Why Learning Good Representations Matters?

Foundational Model



Adapter

Object Detection



Segmentatio

Localization,
...

~ Millions of Unlabeled Data

~ Thousands of Annotated Data

How Much Information is the Machine Given during Learning?

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

[Slide Credit: Yann LeCun]

Supervised Learning

Classification

Object Detection, Segmentation etc

Unsupervised Learning

Dimensionality Reduction

PCA, t-SNE

Clustering

K-Means

Self-Supervised Learning (SSL)

Compressive

Predictive

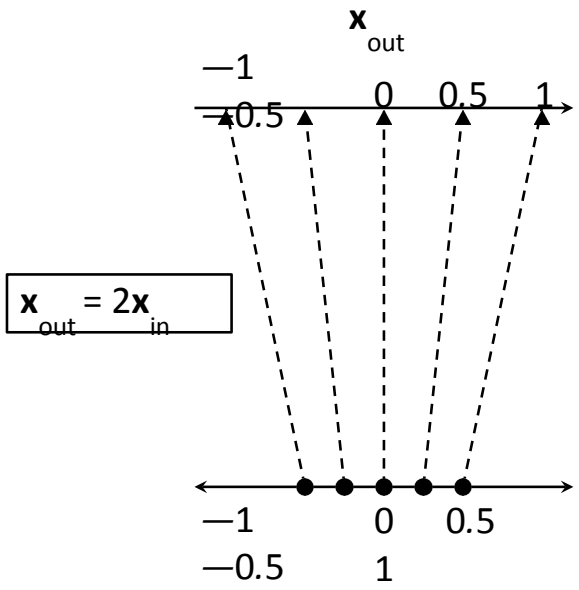
Contrastive

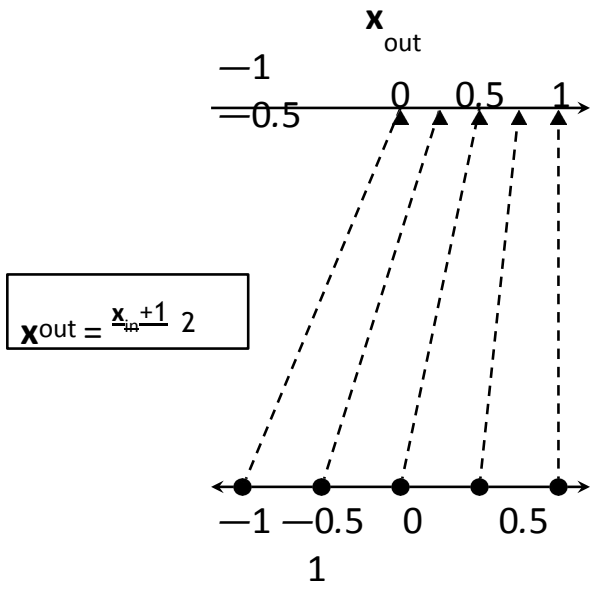
What Do Deep Networks Learn?

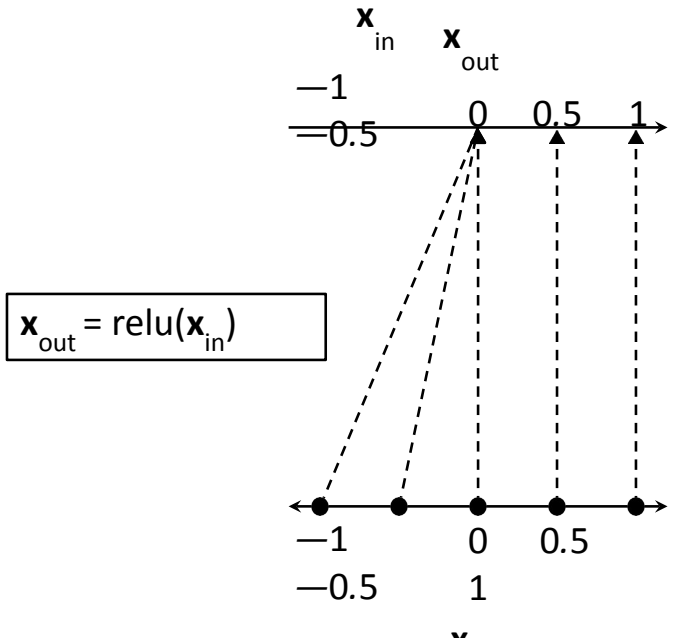
What Do ~~Deep~~ Networks Learn?

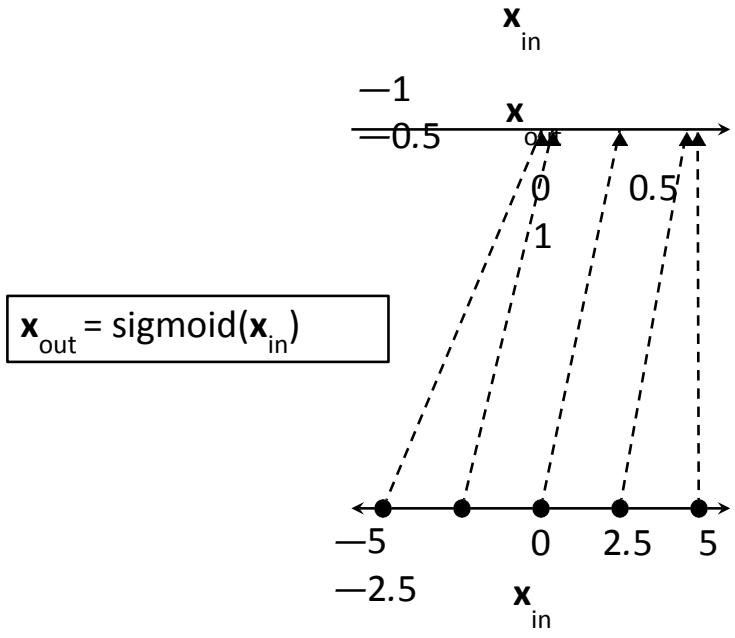
Shallow

Data transformations for a variety of neural net layers









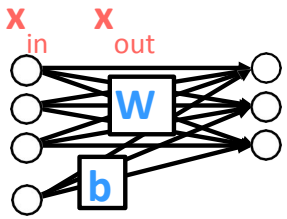
Activations

Parameters

Wiring graph

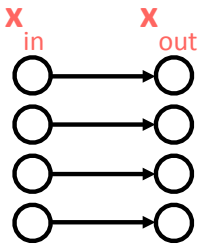
Equation

linear



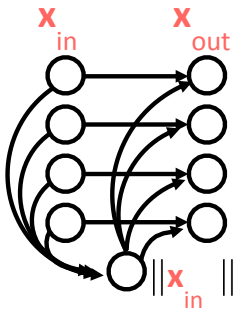
$$\mathbf{x}_{out} = \mathbf{W}\mathbf{x}_{in} + \mathbf{b}$$

relu



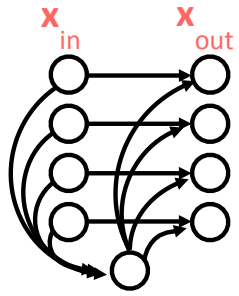
$$x_{out}[i] = \max(x_{in}[i], 0)$$

L2-norm



$$x_{out}[i] = \frac{x_{in}[i]}{\|\mathbf{x}_{in}\|_2}$$

softmax



$$x_{out}[i] = \frac{e^{-\tau x_{in}[i]}}{\sum_{k=1}^{K} e^{-\tau x_{in}[k]}}$$

Activations

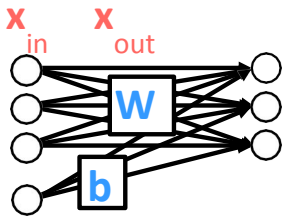
Parameters

Wiring graph

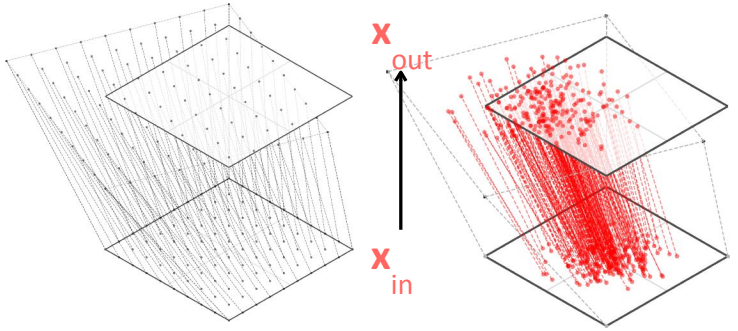
Equation

Mapping

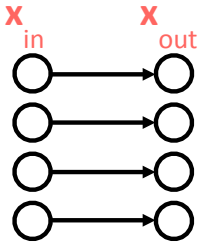
linear



$$\mathbf{x}_{out} = \mathbf{W}\mathbf{x}_{in} + \mathbf{b}$$

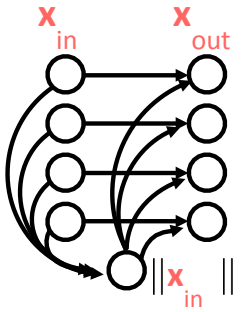


relu



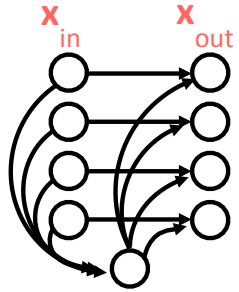
$$x_{out}[i] = \max(x_{in}[i], 0)$$

L2-norm



$$x_{out}[i] = \frac{x_{in}[i]}{\|\mathbf{x}_{in}\|_2}$$

softmax



$$x_{out}[i] = \frac{e^{-\tau x_{in}[i]}}{\sum_{k=1}^{K} e^{-\tau x_{in}[k]}}$$

Activations

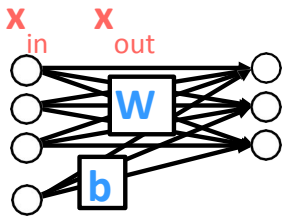
Parameters

Wiring graph

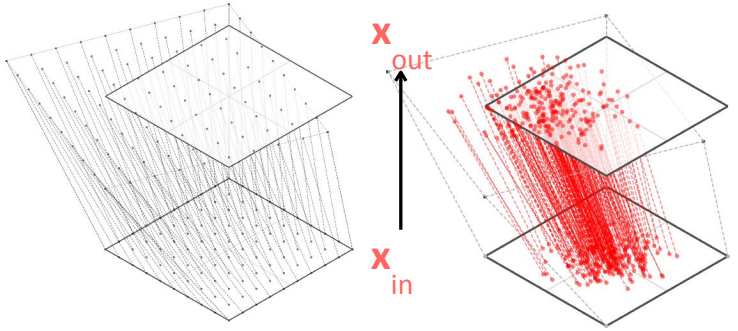
Equation

Mapping

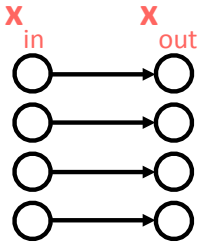
linear



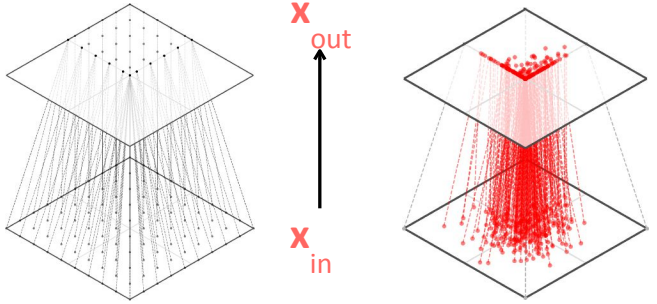
$$x_{out} = Wx_{in} + b$$



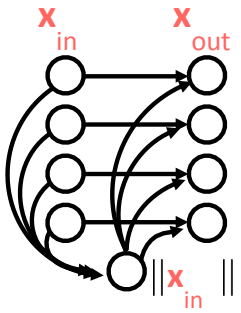
relu



$$x_{out}[i] = \max(x_{in}[i], 0)$$

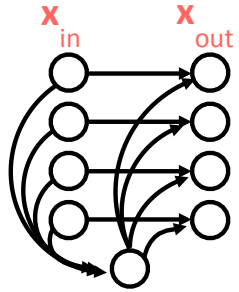


L2-norm



$$x_{out}[i] = \frac{x_{in}[i]}{\|x_{in}\|_2}$$

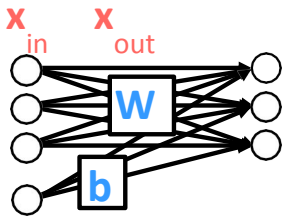
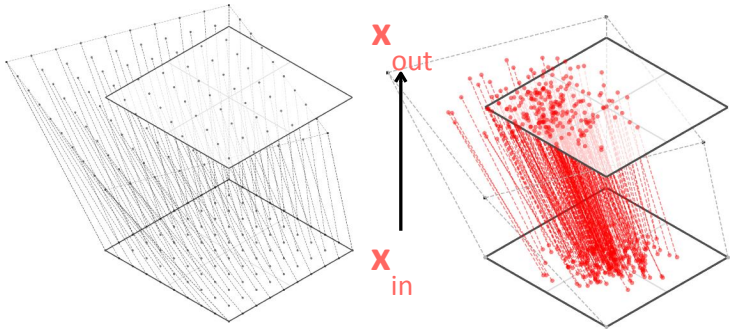
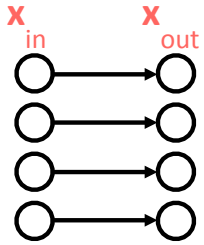
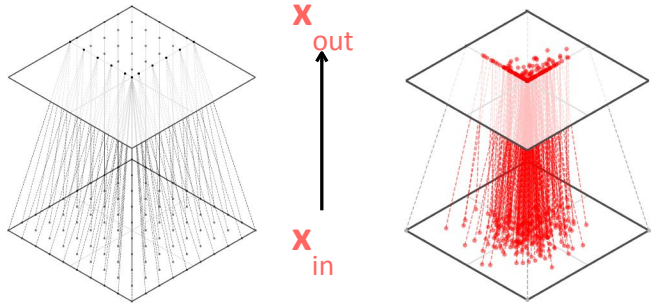
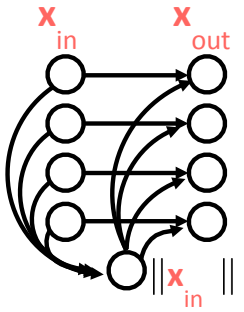
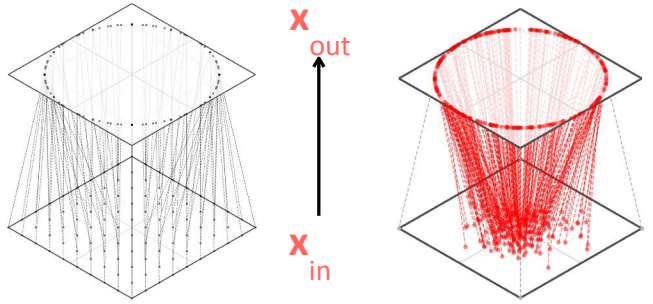
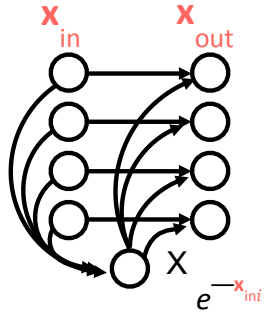
softmax



$$x_{out}[i] = \frac{e^{-\tau x_{in}[i]}}{\sum_{k=1}^{K} e^{-\tau x_{in}[k]}}$$

Activations

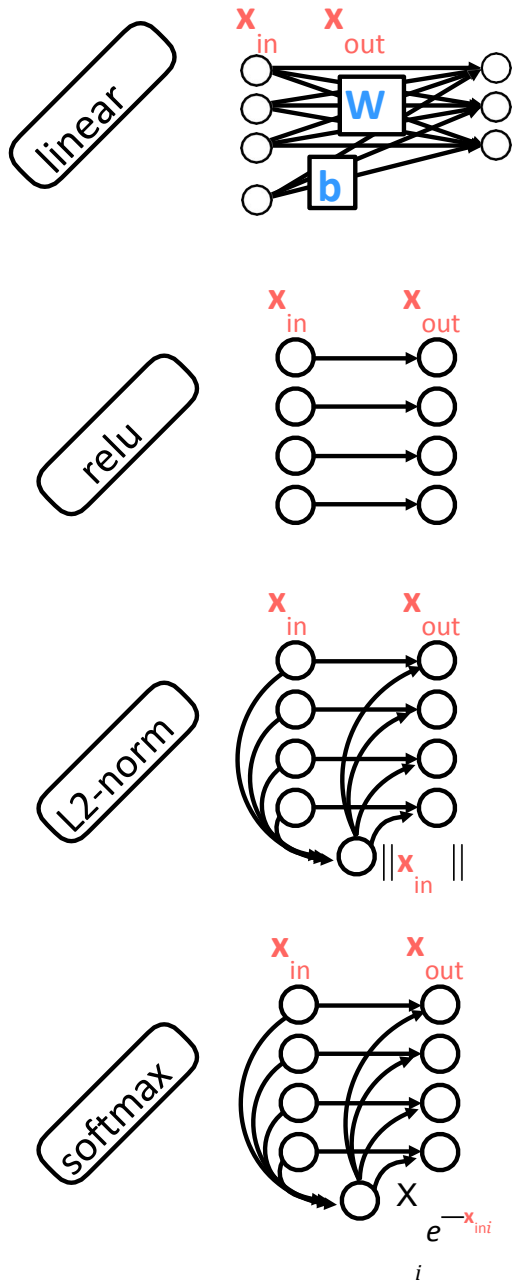
Parameters

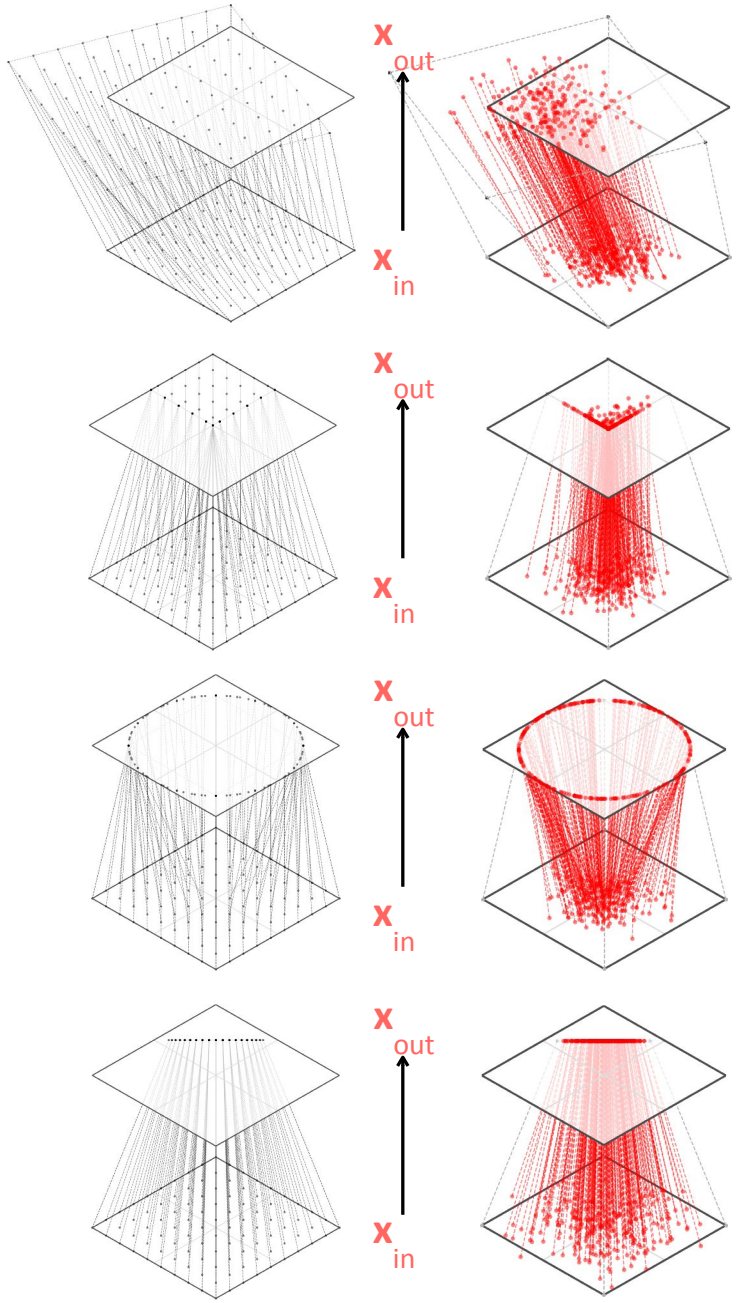
Wiring graph	Equation	Mapping
linear 	$\mathbf{x}_{out} = \mathbf{W}\mathbf{x}_{in} + \mathbf{b}$	
relu 	$x_{out}[i] = \max(x_{in}[i], 0)$	
L2-norm 	$x_{out}[i] = \frac{x_{in}[i]}{\|\mathbf{x}_{in}\|_2}$	
softmax 	$x_{out}[i] = \frac{e^{-\tau x_{in}[i]}}{\sum_{k=1}^{K} e^{-\tau x_{in}[k]}}$	

Activations

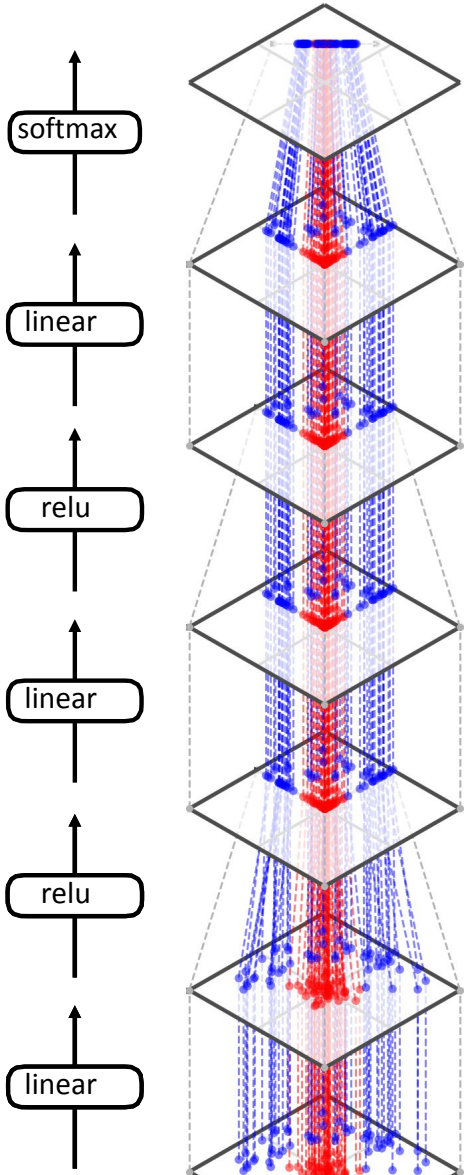
Parameters

Wiring graph	Equation	Mapping
linear	$x_{out} = Wx_{in} + b$	
relu	$x_{out}[i] = \max(x_{in}[i], 0)$	
L2-norm	$x_{out}[i] = \frac{x_{in}[i]}{\|x_{in}\|_2}$	
softmax	$x_{out}[i] = \frac{e^{-\tau x_{in}[i]}}{\sum_{k=1}^{K} e^{-\tau x_{in}[k]}}$	

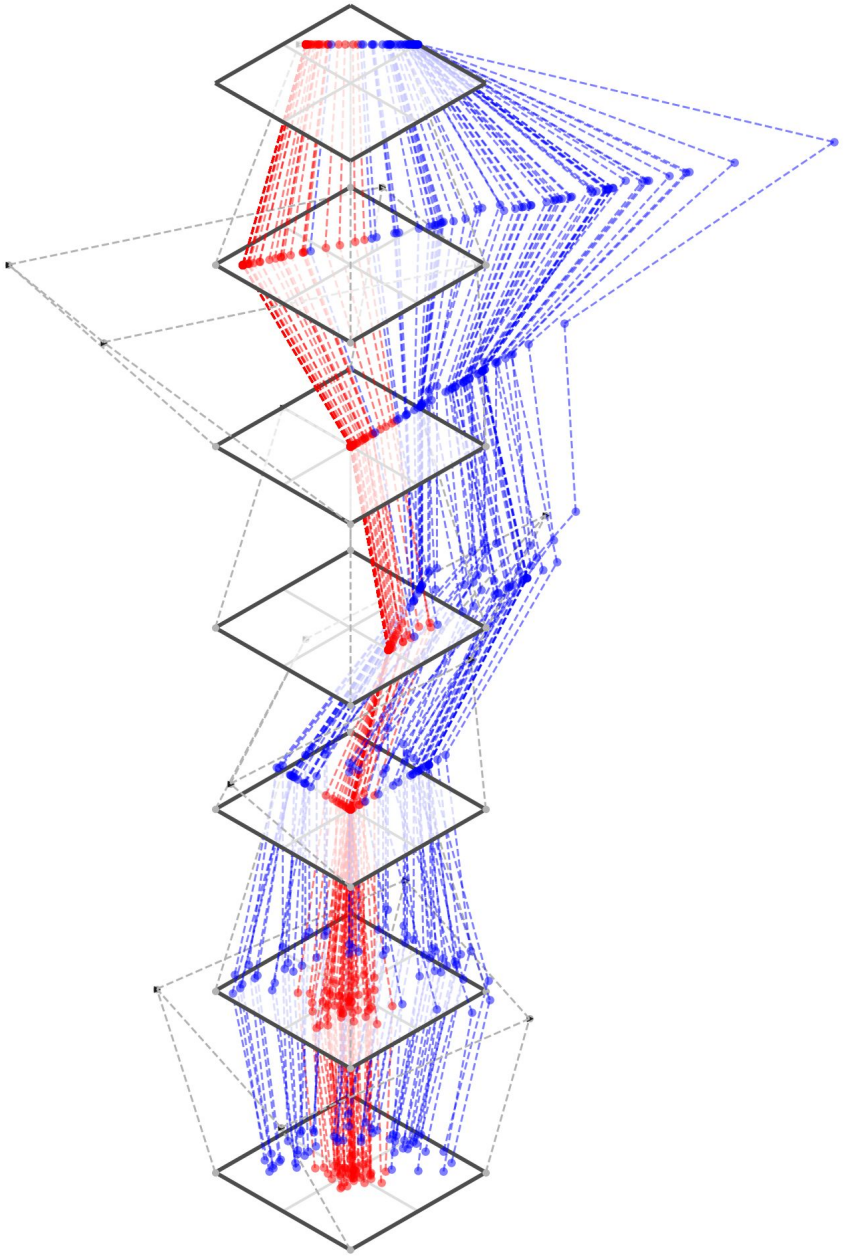




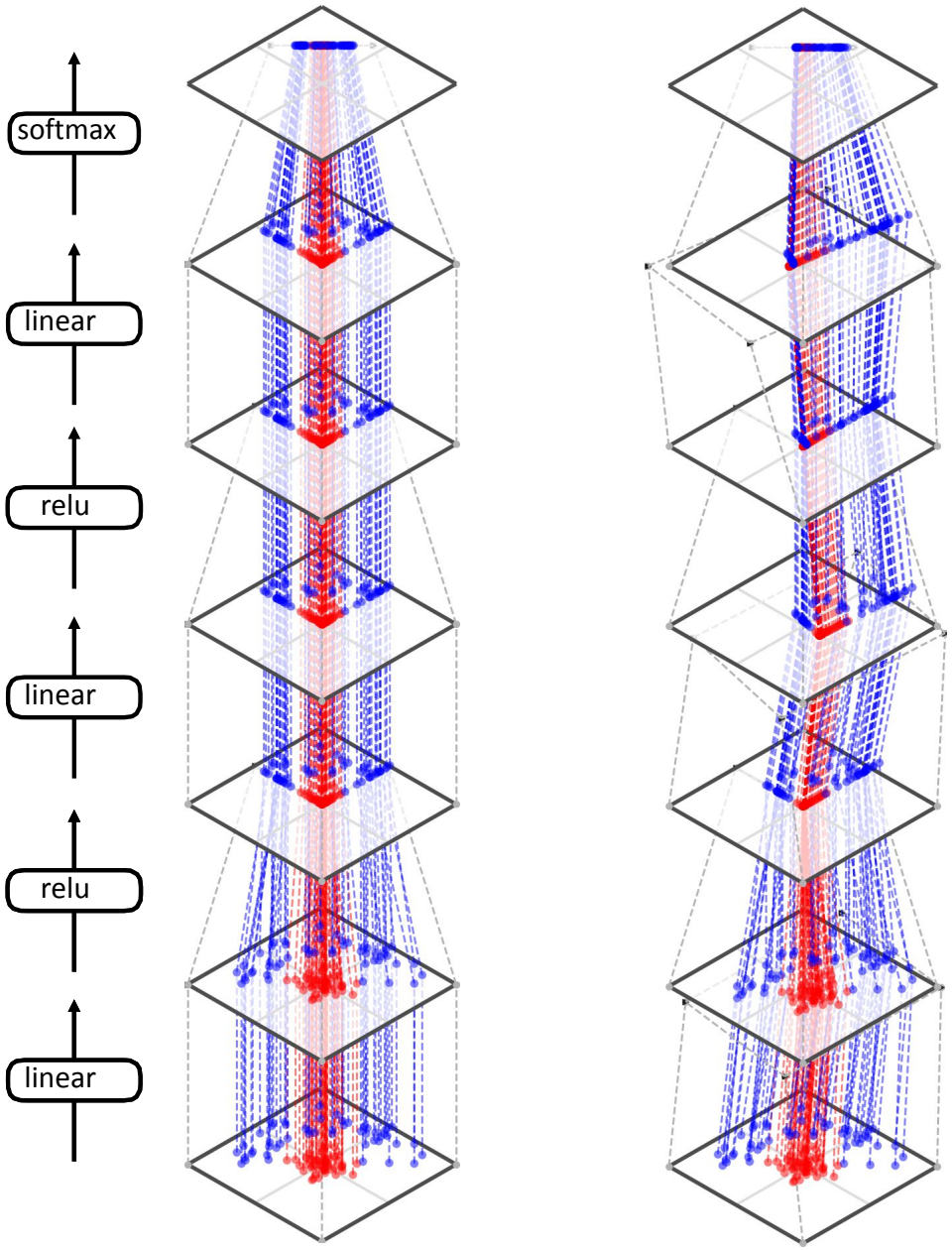
MLP



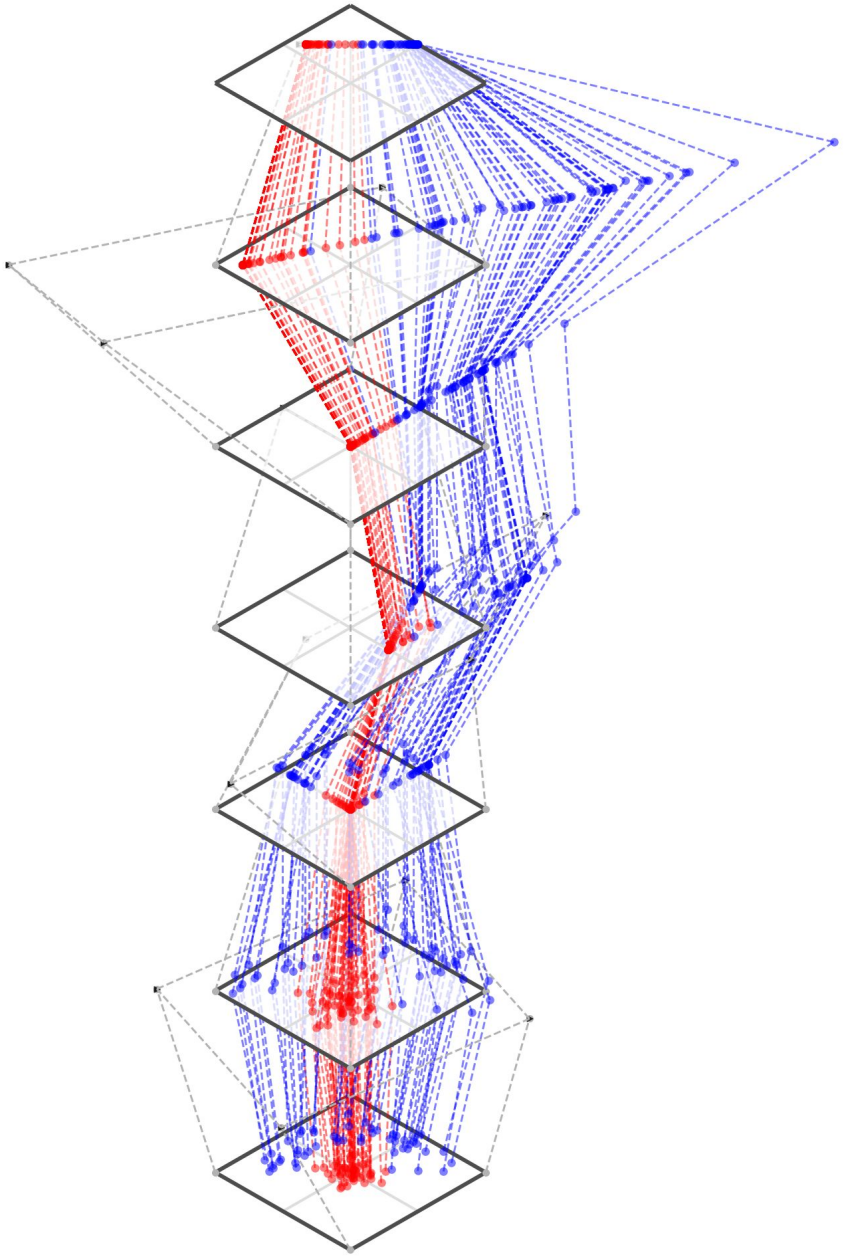
Training iteration



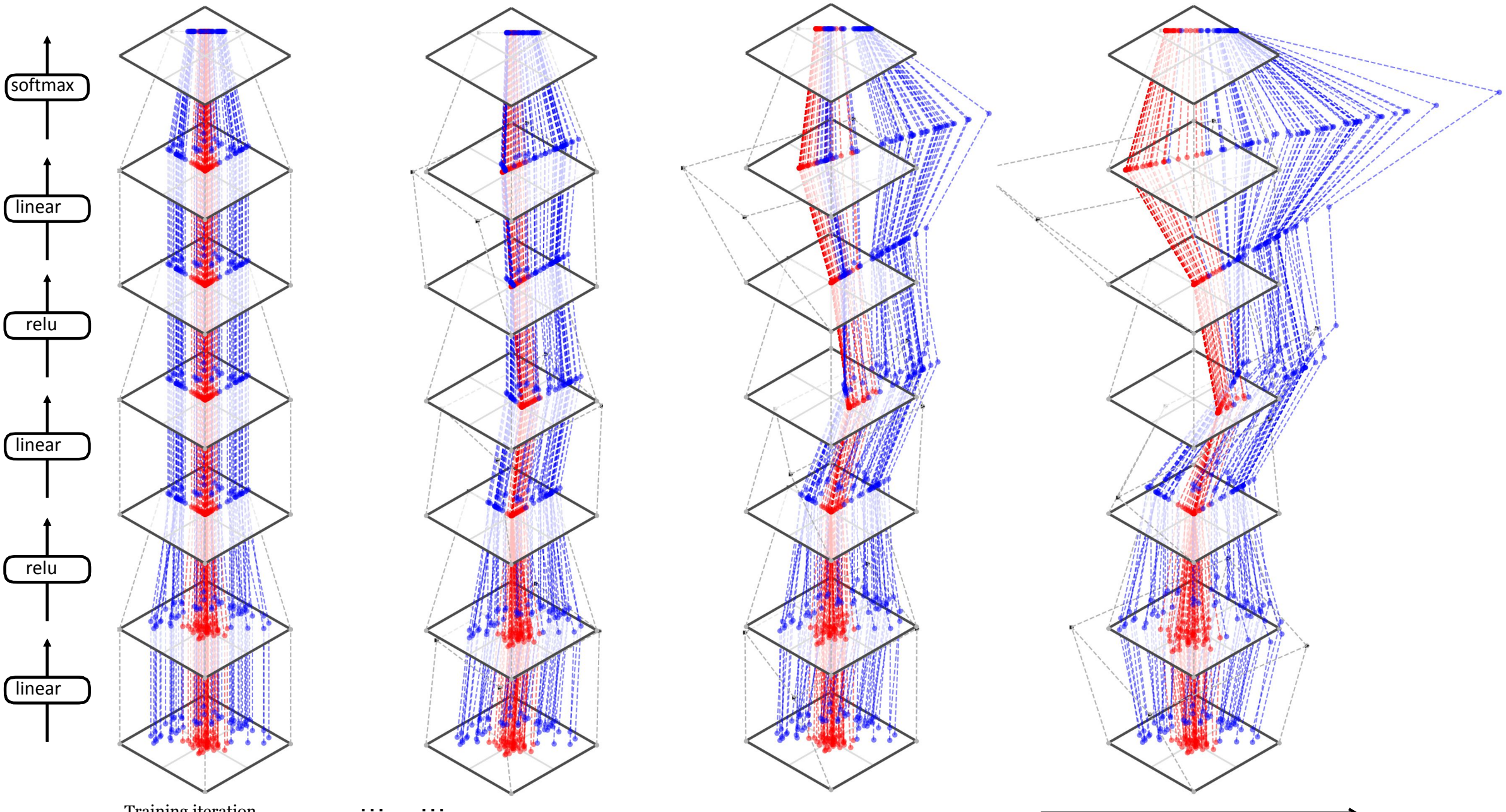
MLP



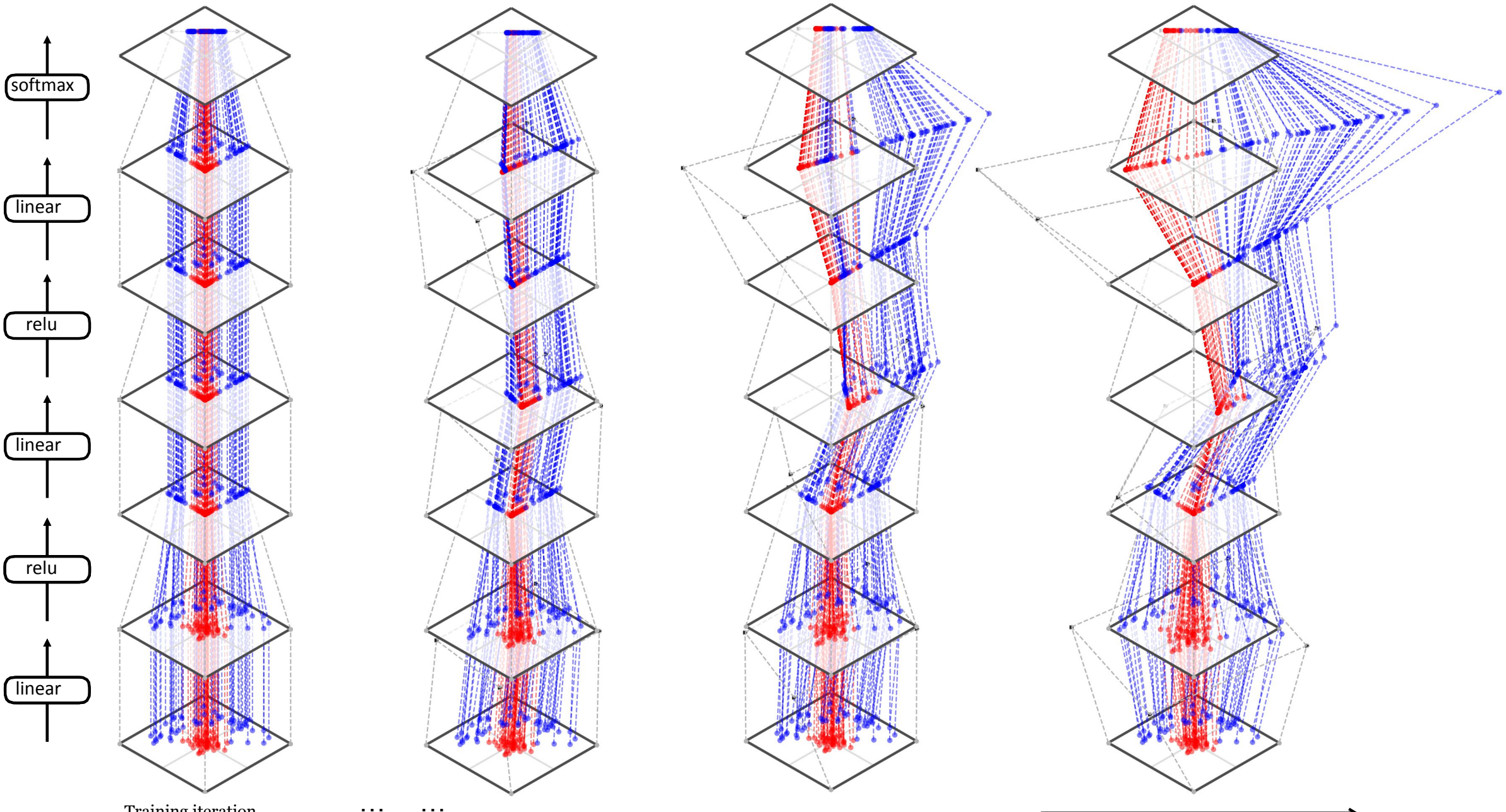
Training iteration



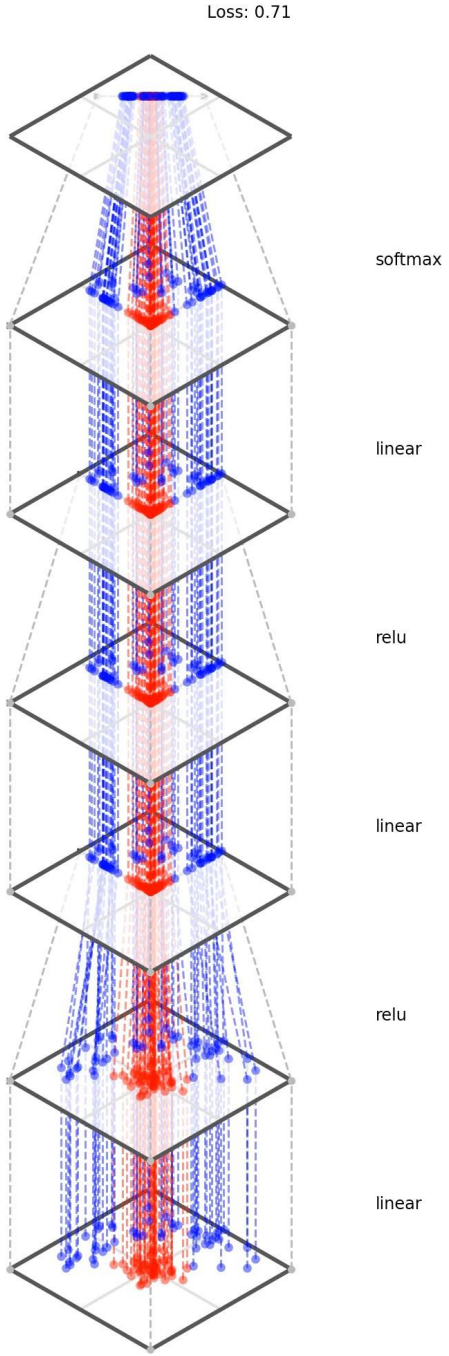
MLP


softmax
linear
relu
linear
relu
linear
Training iteration

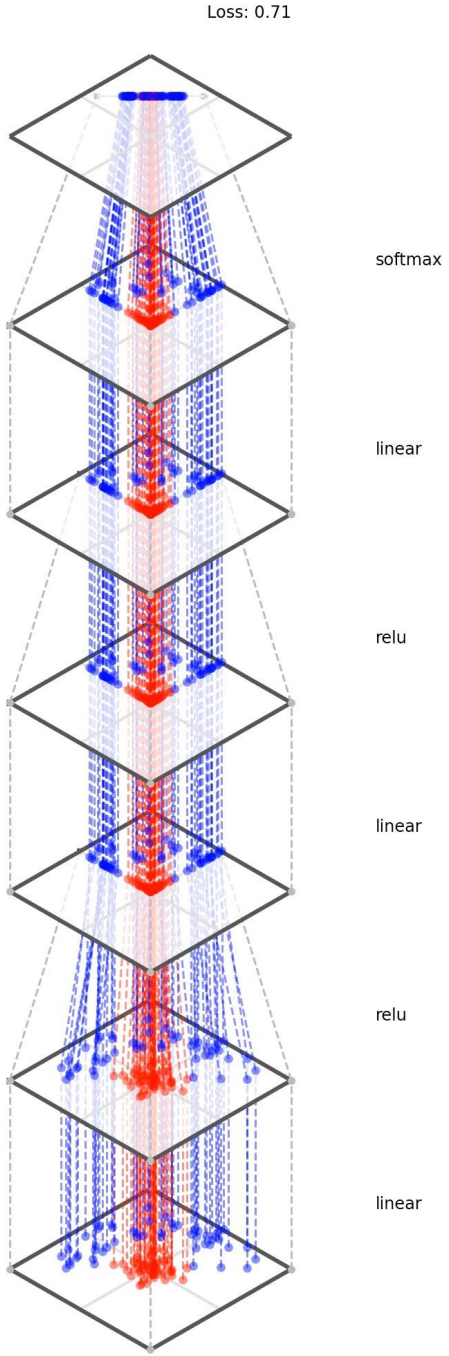
MLP


softmax
linear
relu
linear
relu
linear
Training iteration

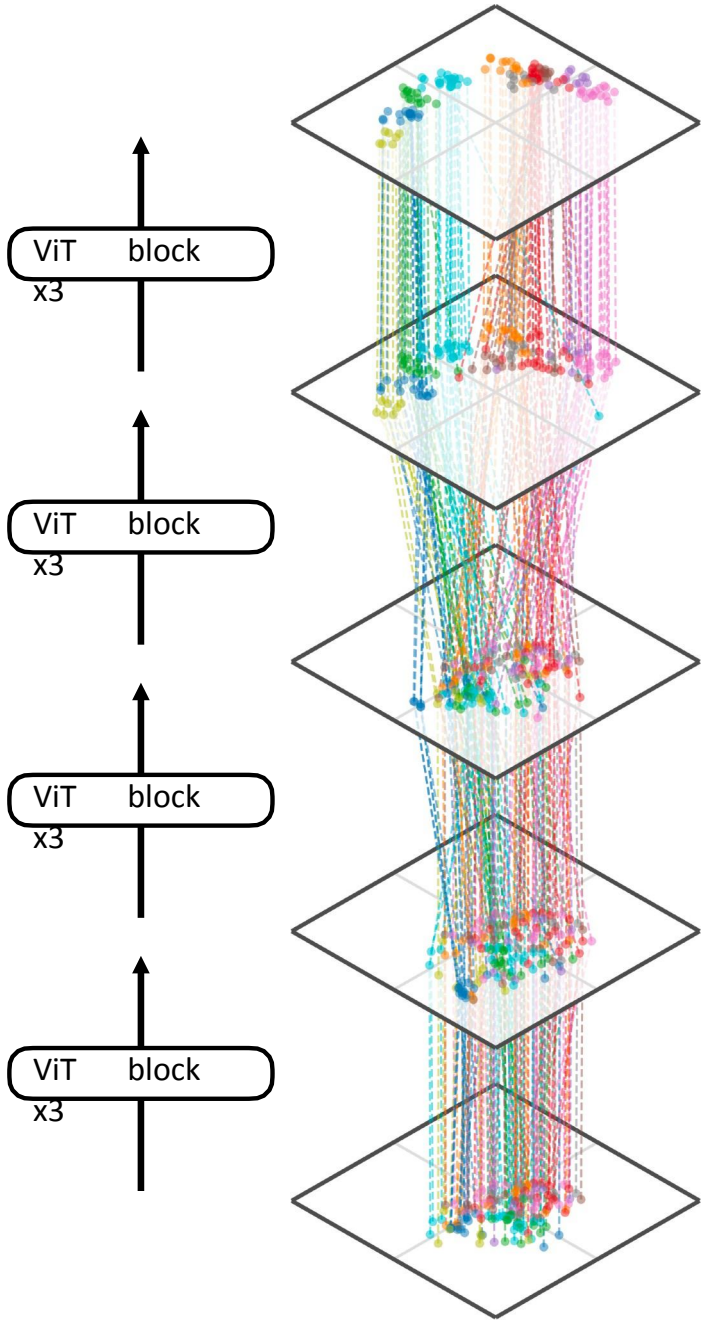
ML P



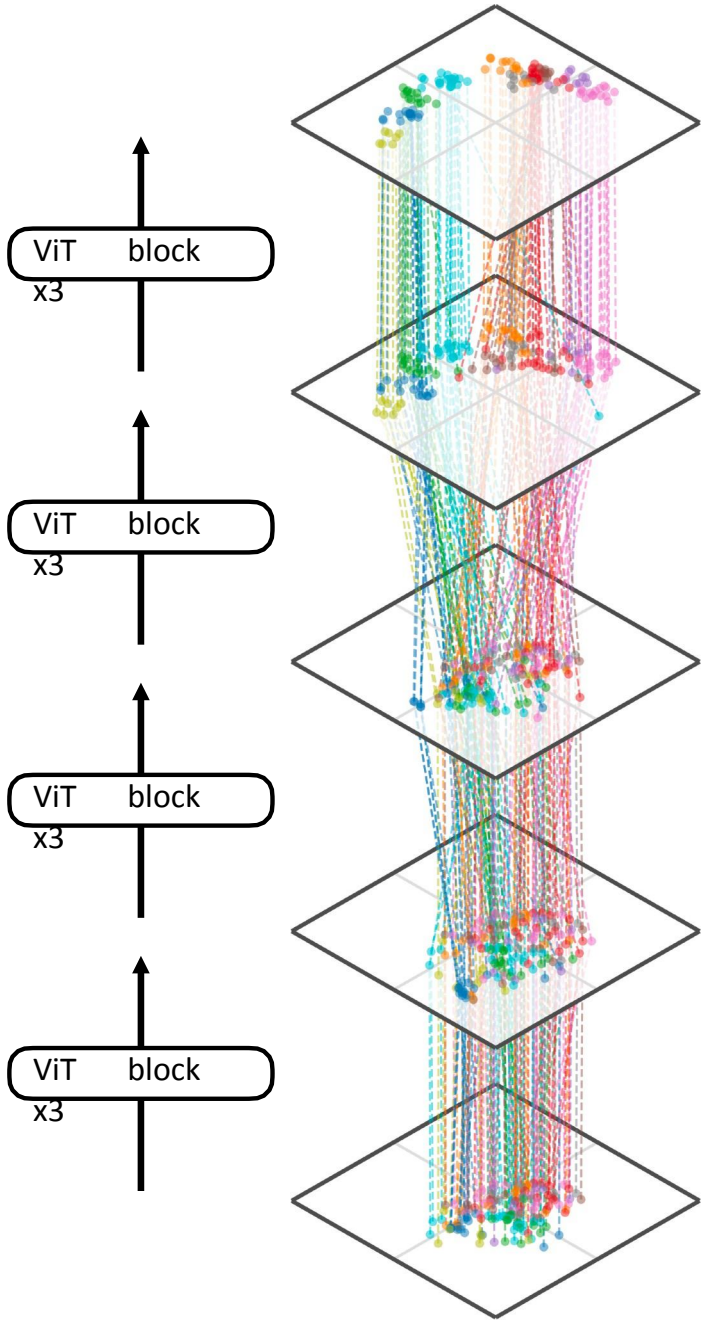
ML P

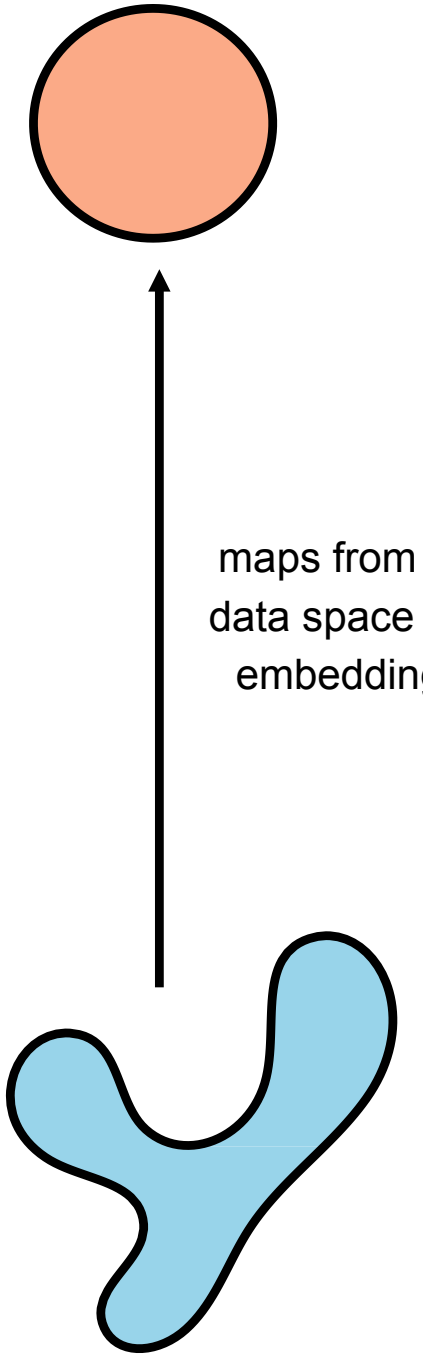


CLI
P



CLI
P

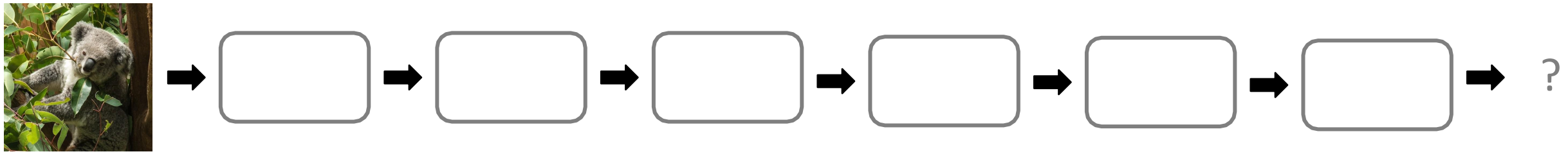

ViT block
x3
ViT block
x3
ViT block
x3
ViT block
x3


maps from complex
data space to simple
embedding space

What Do Deep Networks Learn?

How to Represent Images w/ Deep Learning?

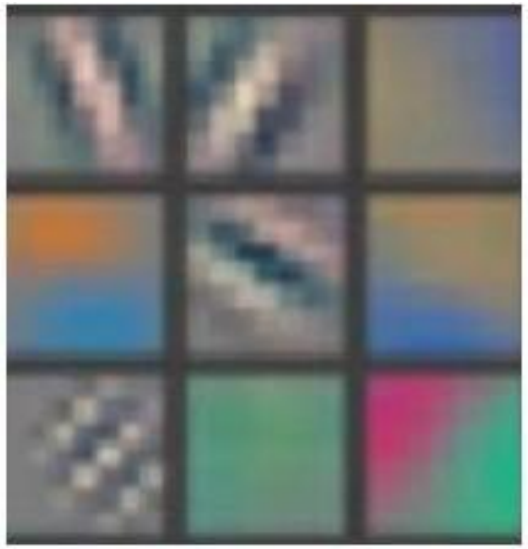
general modules (instead of specialized features)



compose simple modules into complex functions

- build multiple levels of abstractions
- learn by back-prop
- learn from data
- reduce domain knowledge and feature engineering

Multiple Levels of Representations

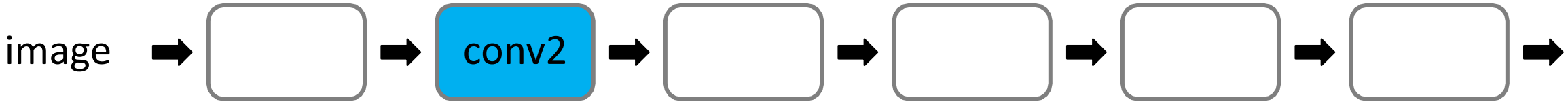


features

stimuli
(patches with the highest
1-hot activations)

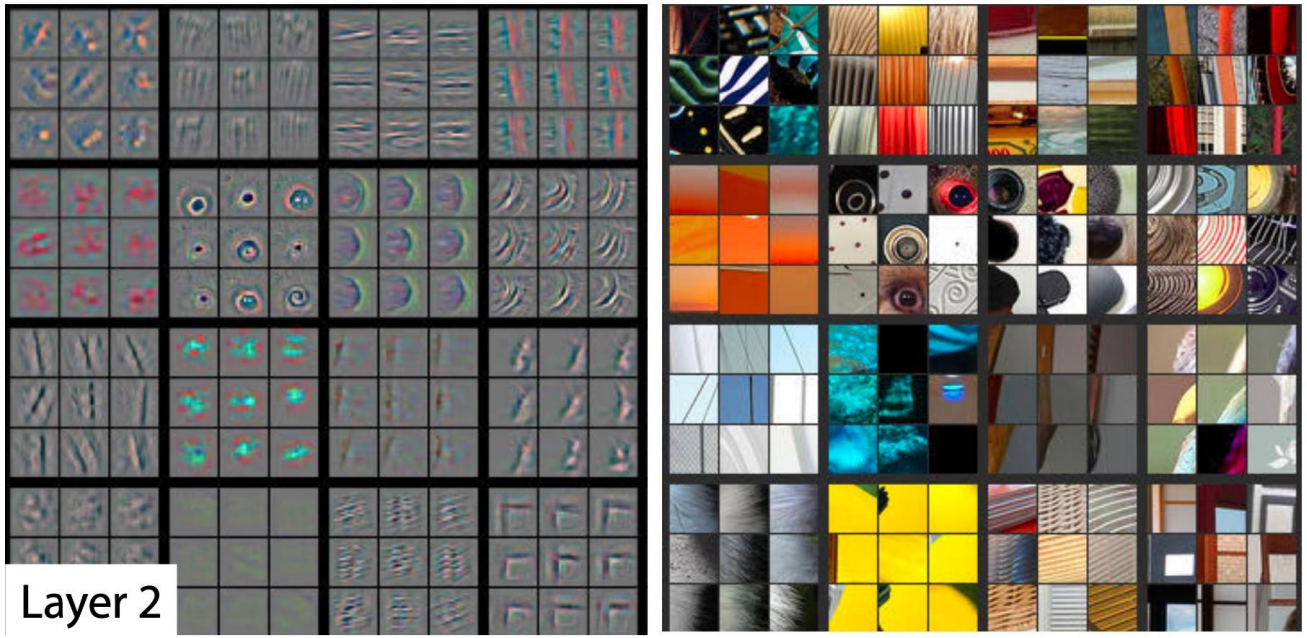
"Visualizing and Understanding Convolutional Networks", Zeiler & Fergus. ECCV 2014

Multiple Levels of Representations



features

stimuli

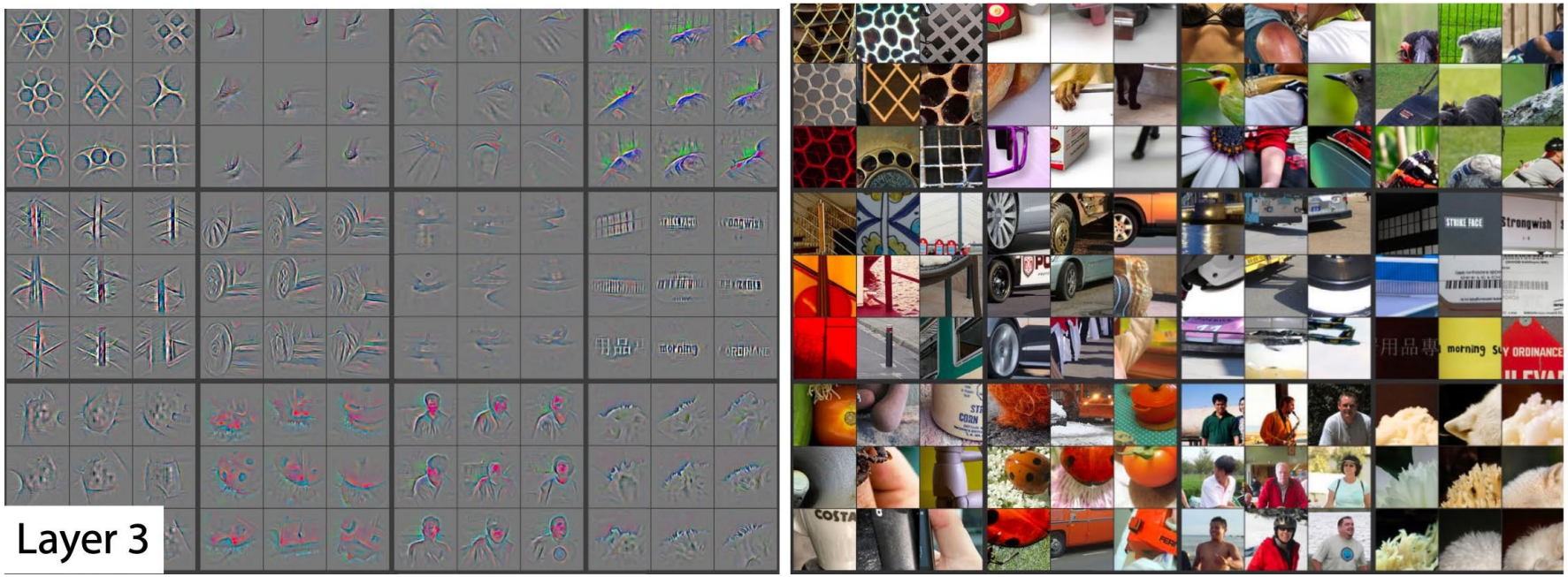


"Visualizing and Understanding Convolutional Networks", Zeiler & Fergus. ECCV 2014

Multiple Levels of Representations

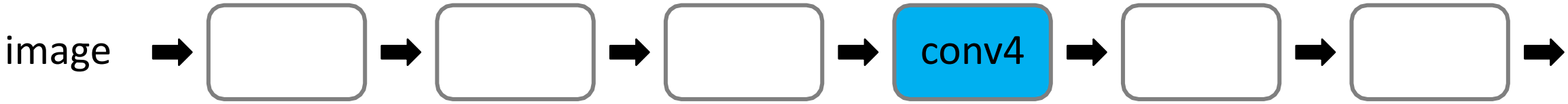
features

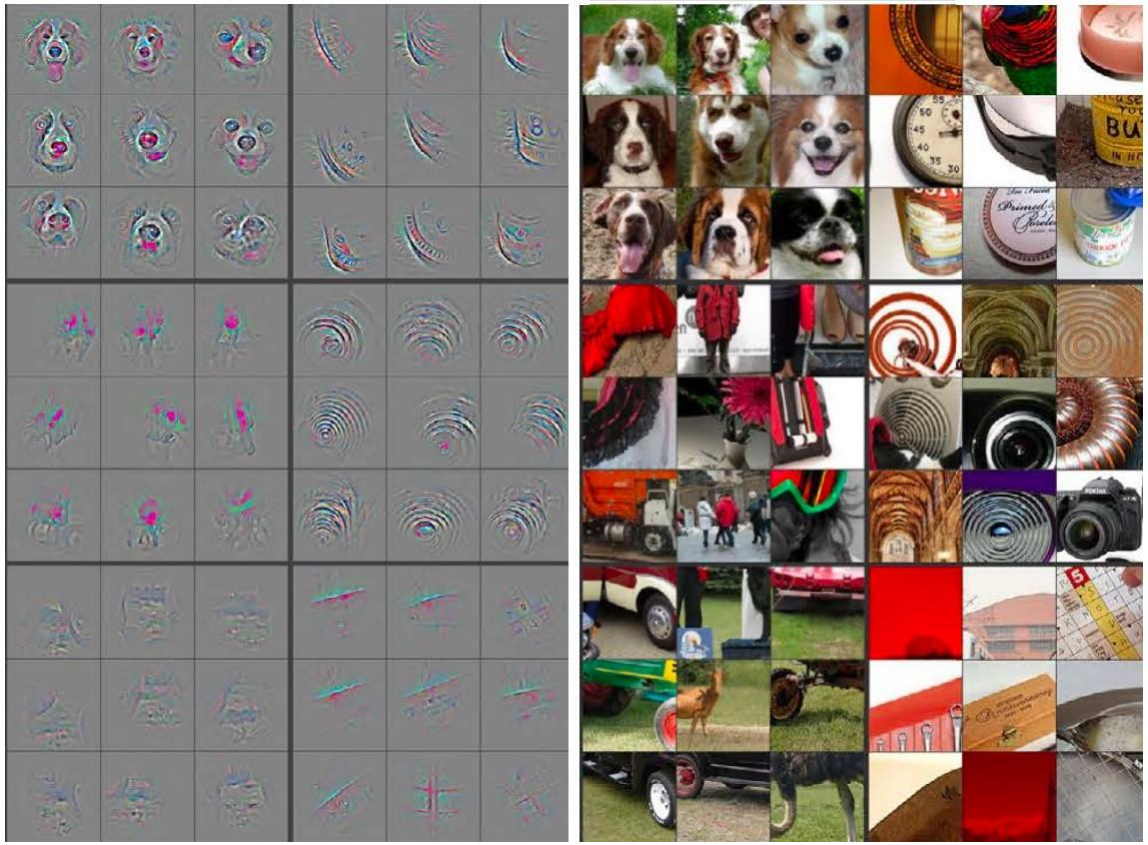
stimuli

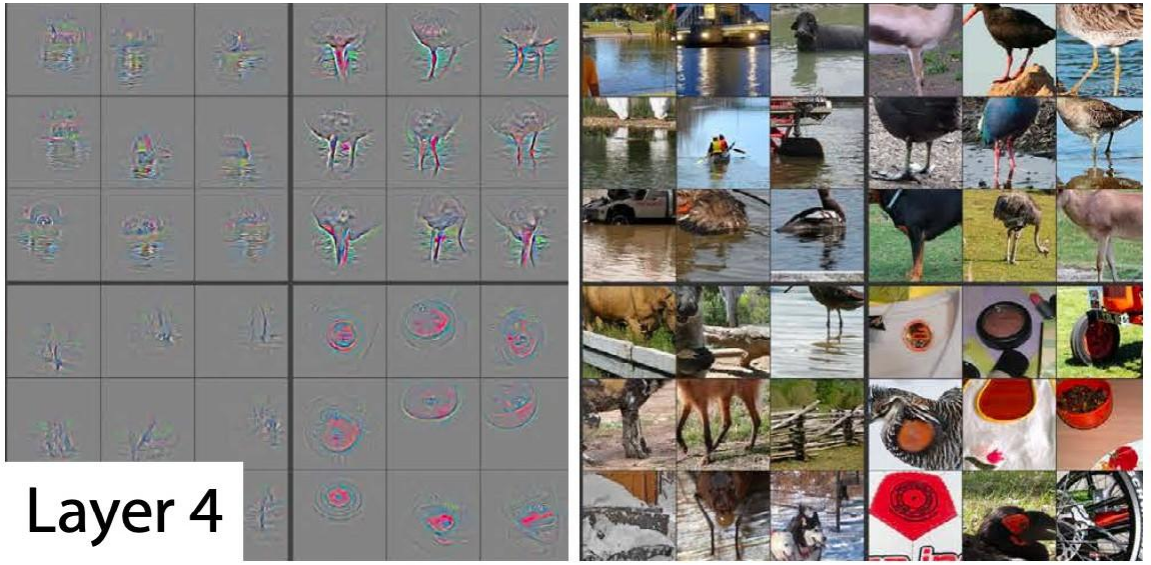


"Visualizing and Understanding Convolutional Networks", Zeiler & Fergus. ECCV 2014

Multiple Levels of Representations

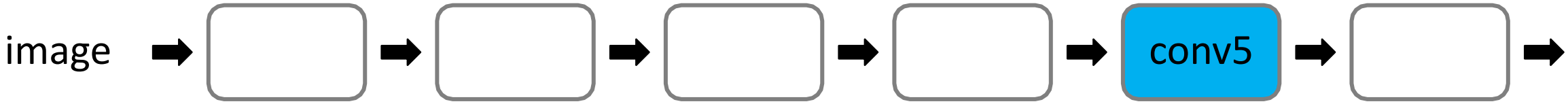


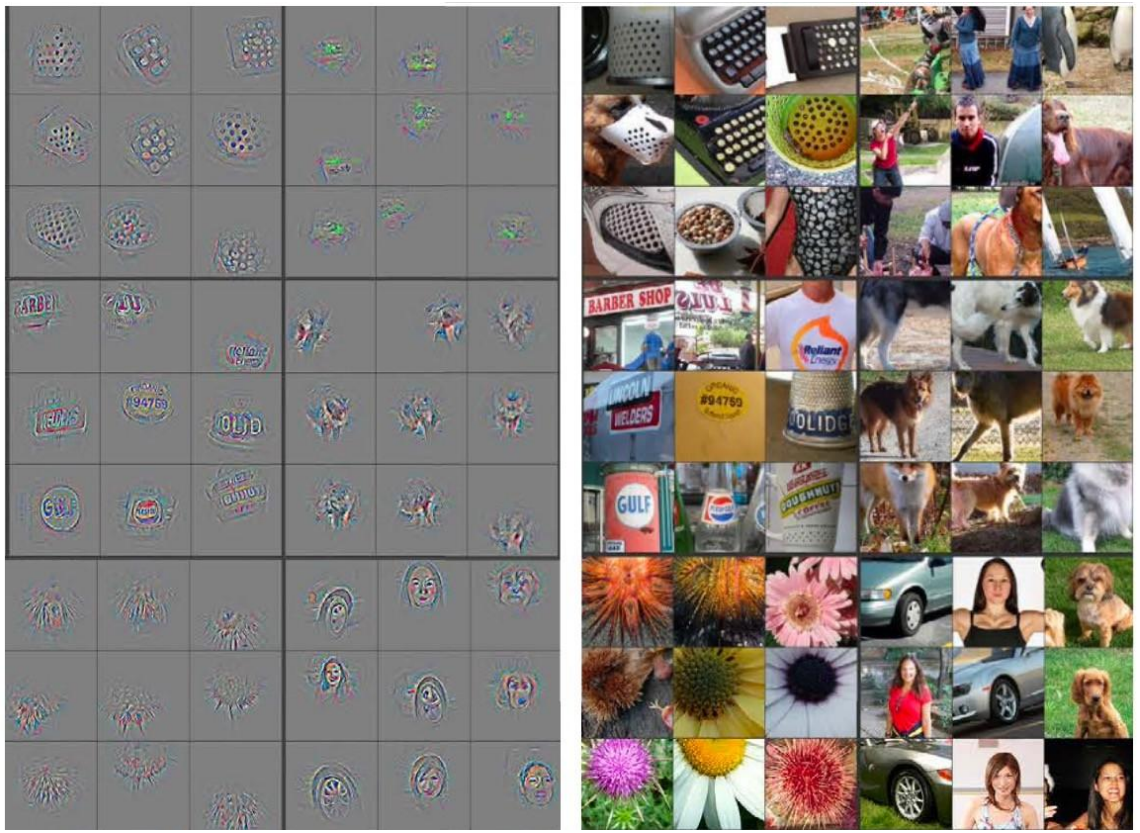


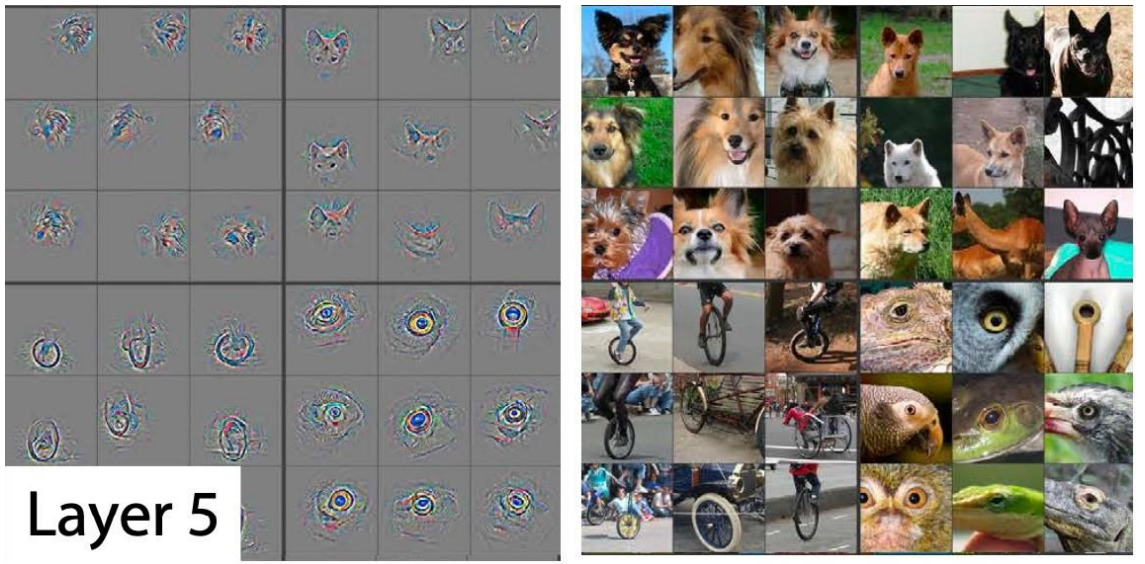


"Visualizing and Understanding Convolutional Networks", Zeiler & Fergus. ECCV 2014

Multiple Levels of Representations



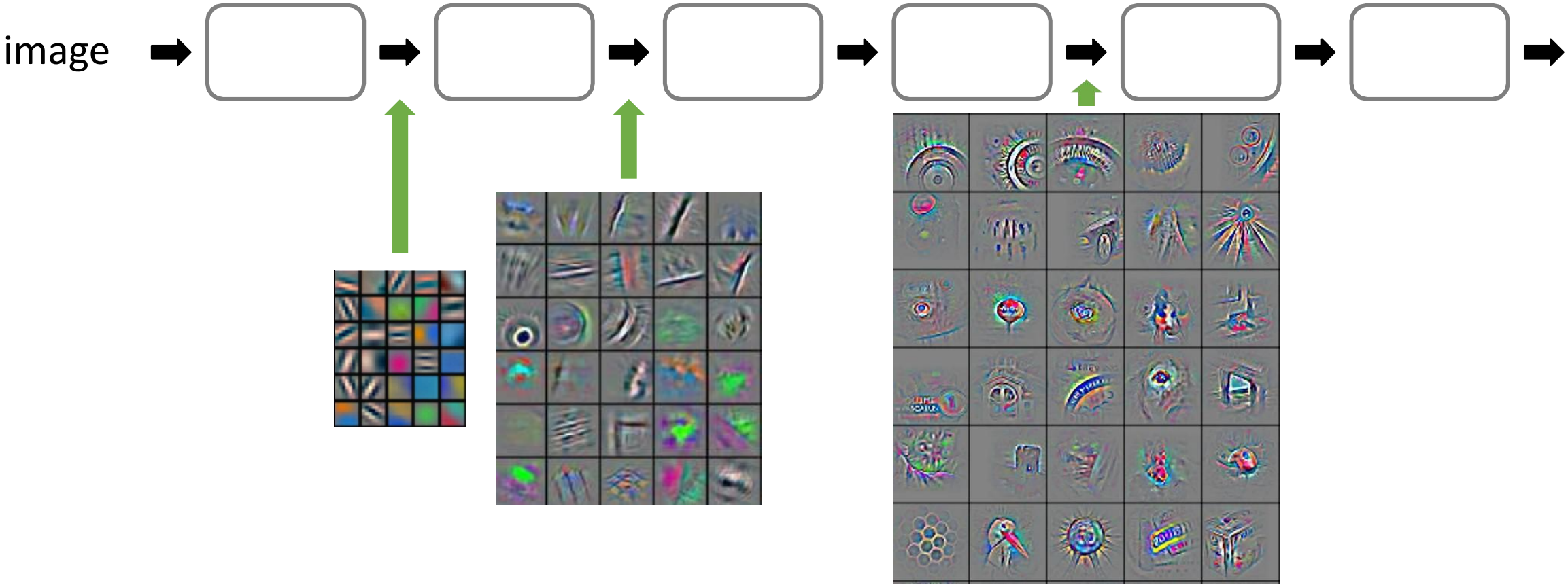




Deeper layers have “higher-level” features.

“Visualizing and Understanding Convolutional Networks”, Zeiler & Fergus. ECCV 2014

Multiple Levels of Representations



Deeper layers have “higher-level” features.

“Visualizing and Understanding Convolutional Networks”, Zeiler & Fergus. ECCV 2014

Deep Representations are Transferrable!

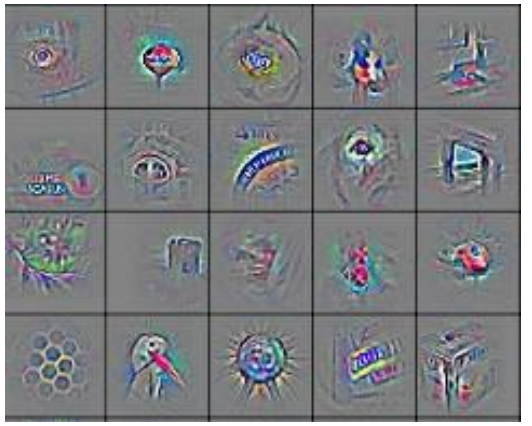
The single most important discovery in DL revolution

Transfer learning:

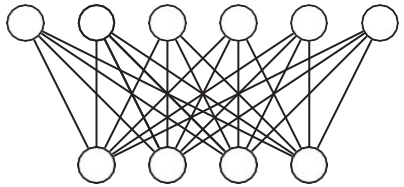
- pre-train on large-scale data
- fine-tune on small-scale data

- enable DL for small datasets
- revolutionize computer vision

- data: engine for general representation
- GPT: a similar principle



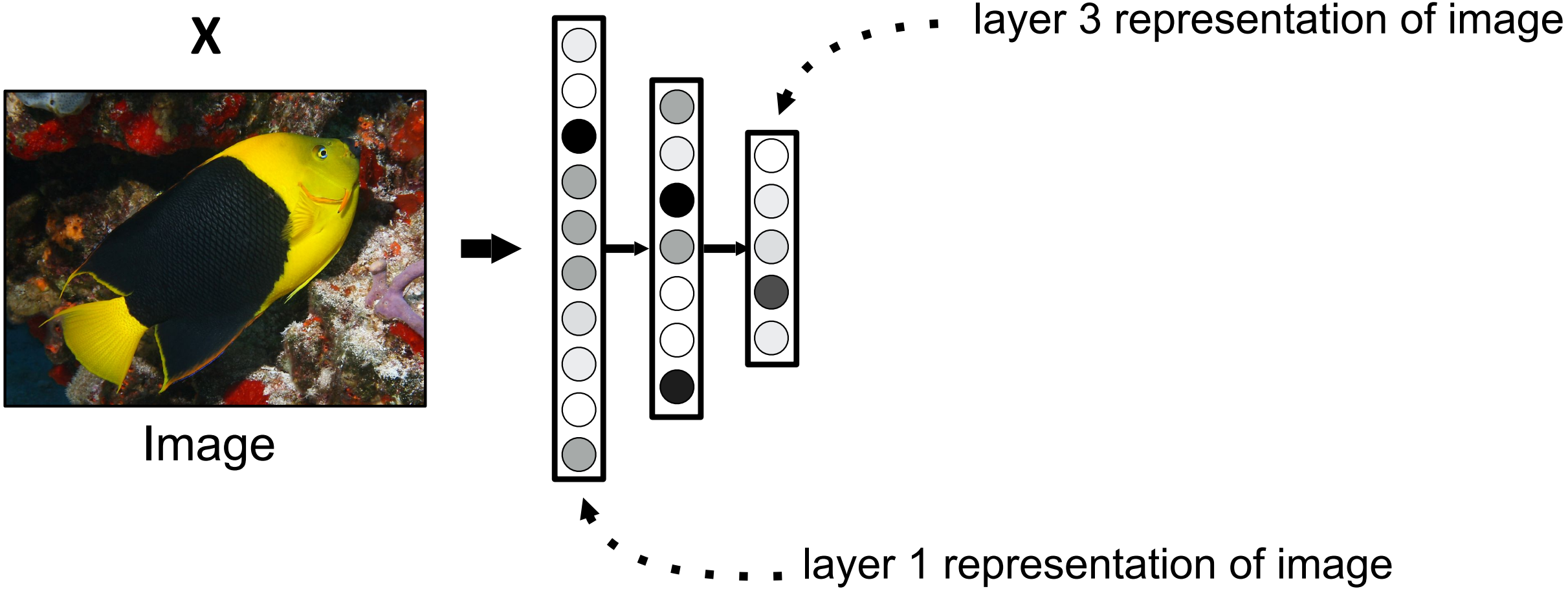
pre-train



fine-tune

"DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", Donahue et al. arXiv 2013
"Visualizing and Understanding Convolutional Networks", Zeiler & Fergus. arXiv 2013 "CNN Features off-the-shelf: an Astounding Baseline for Recognition", Razavian. arXiv 2014

Classification Networks Produce Good Features



Represent data as a neural **embedding** — a vector/tensor of neural activations

(perhaps representing a vector of detected texture patterns or object parts)

What we covered so far?

1. Neural Networks naturally learn representations
2. Good representations are transferable/useful for downstream tasks
3. We have huge unlabeled datasets, small labeled ones

HOW CAN WE LEARN WITHOUT LABAELS?!!!

Supervised Learning

Classification

Object Detection, Segmentation etc

Unsupervised Learning

Dimensionality Reduction

PCA, t-SNE

Clustering

K-Means

No Labels

Self-Supervised Learning (SSL)

Compressive

Predictive

Contrastive

Self-Supervised Learning (SSL)

Compressive

Predictive

Contrastive

Learning via compression

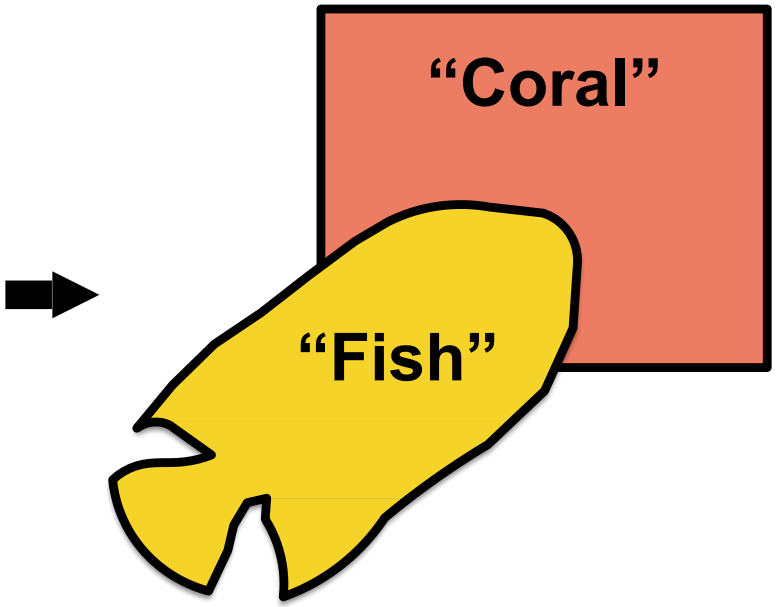
x

Image

Learning via compression

x

Image



Compact mental representation

Learning via compression

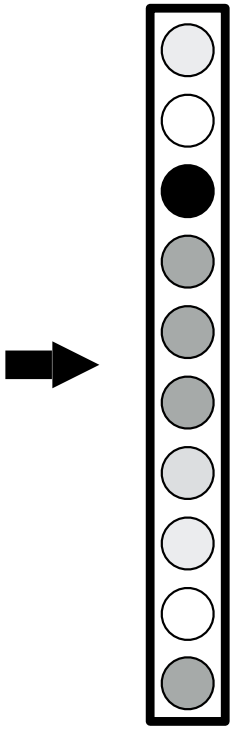
x

Image

Learning via compression

x

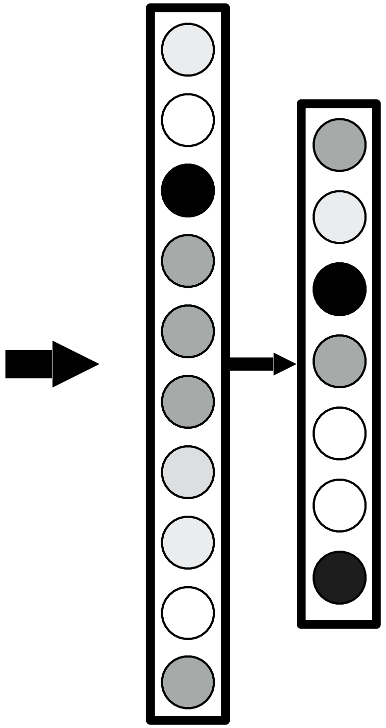
Image



Learning via compression

x

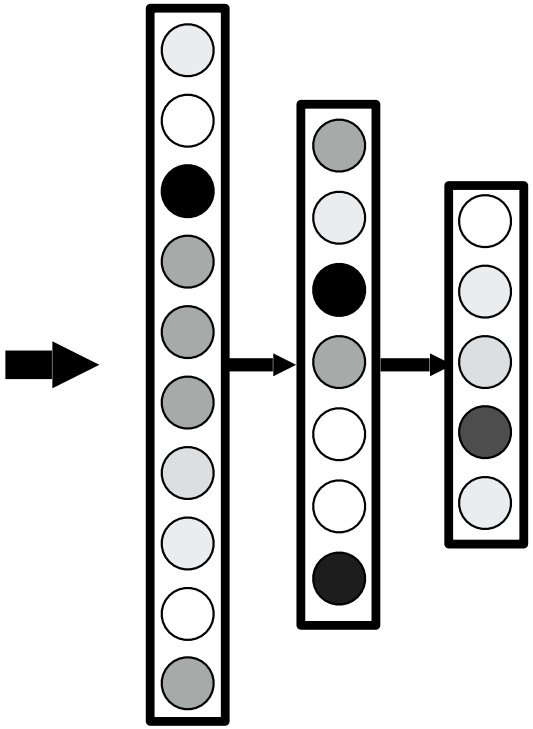
Image



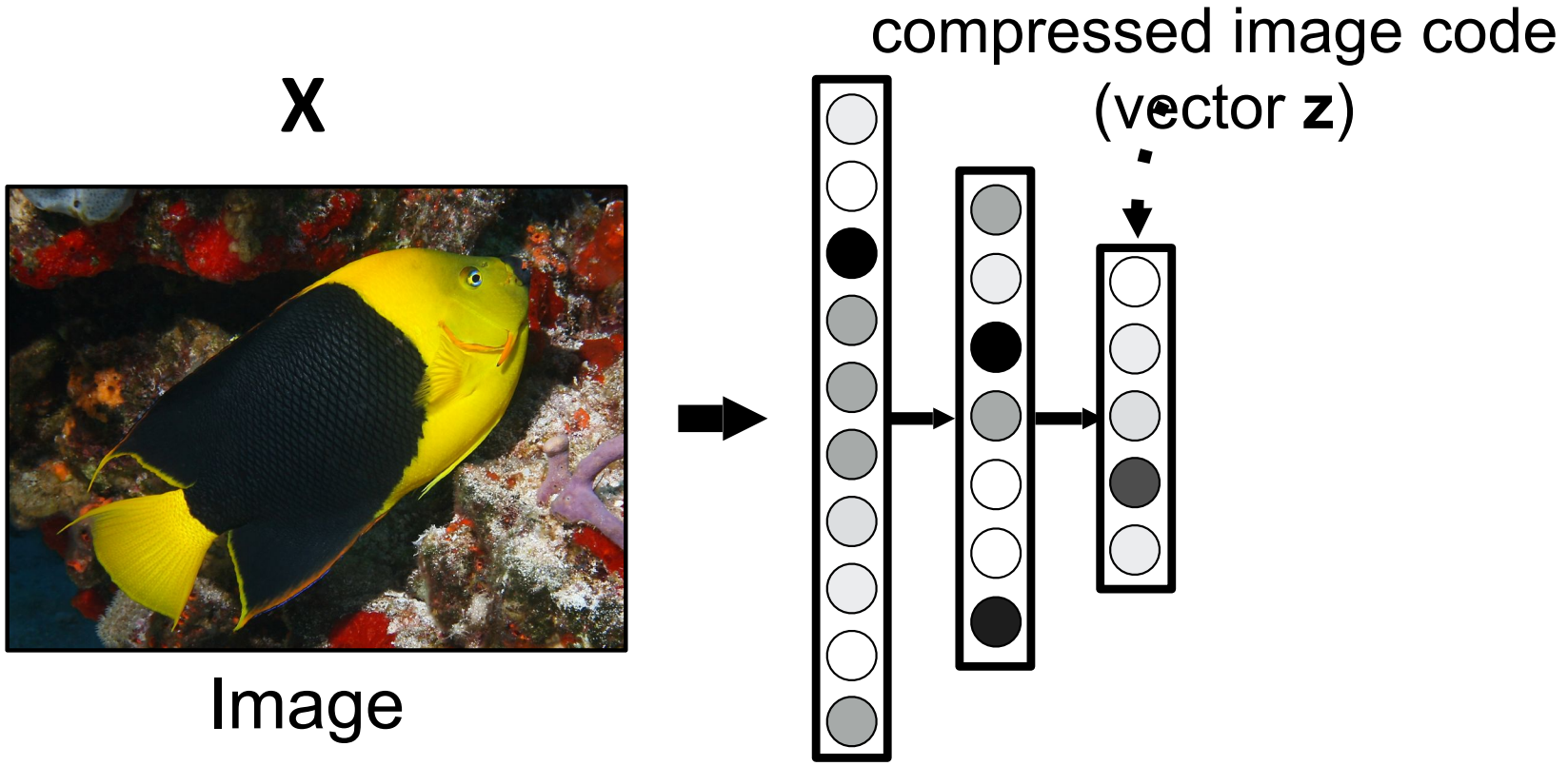
Learning via compression

x

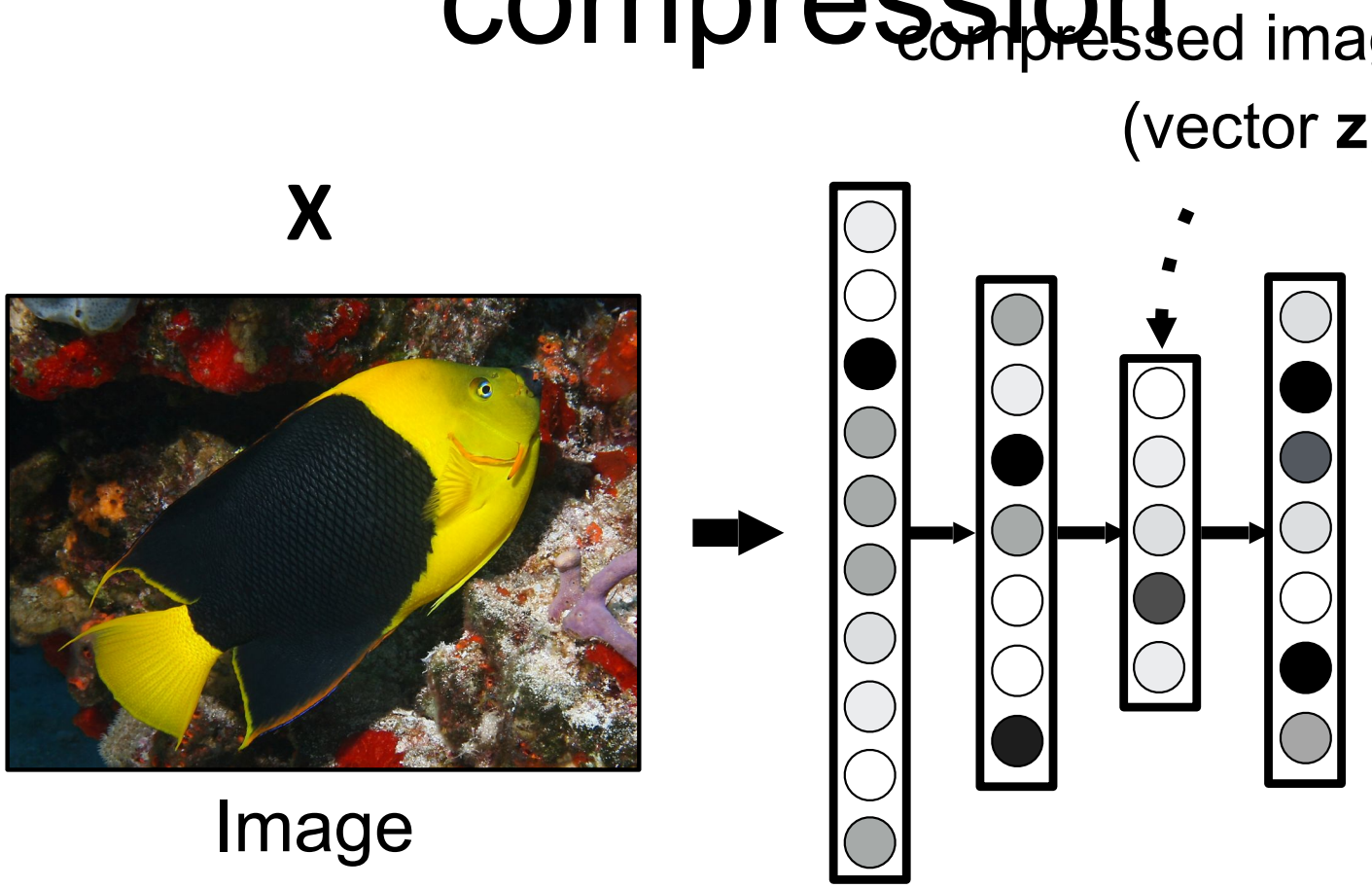
Image



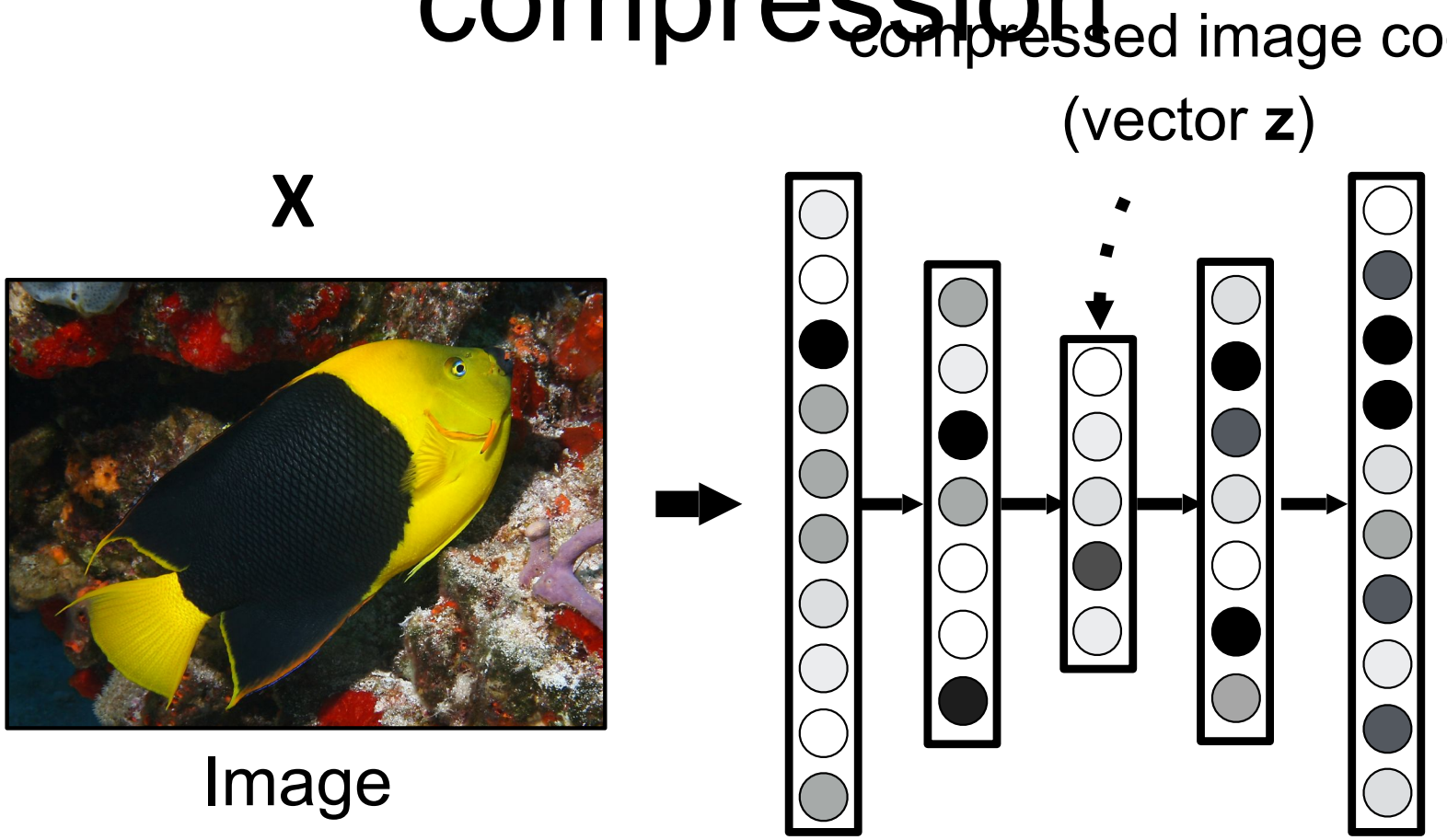
Learning via compression



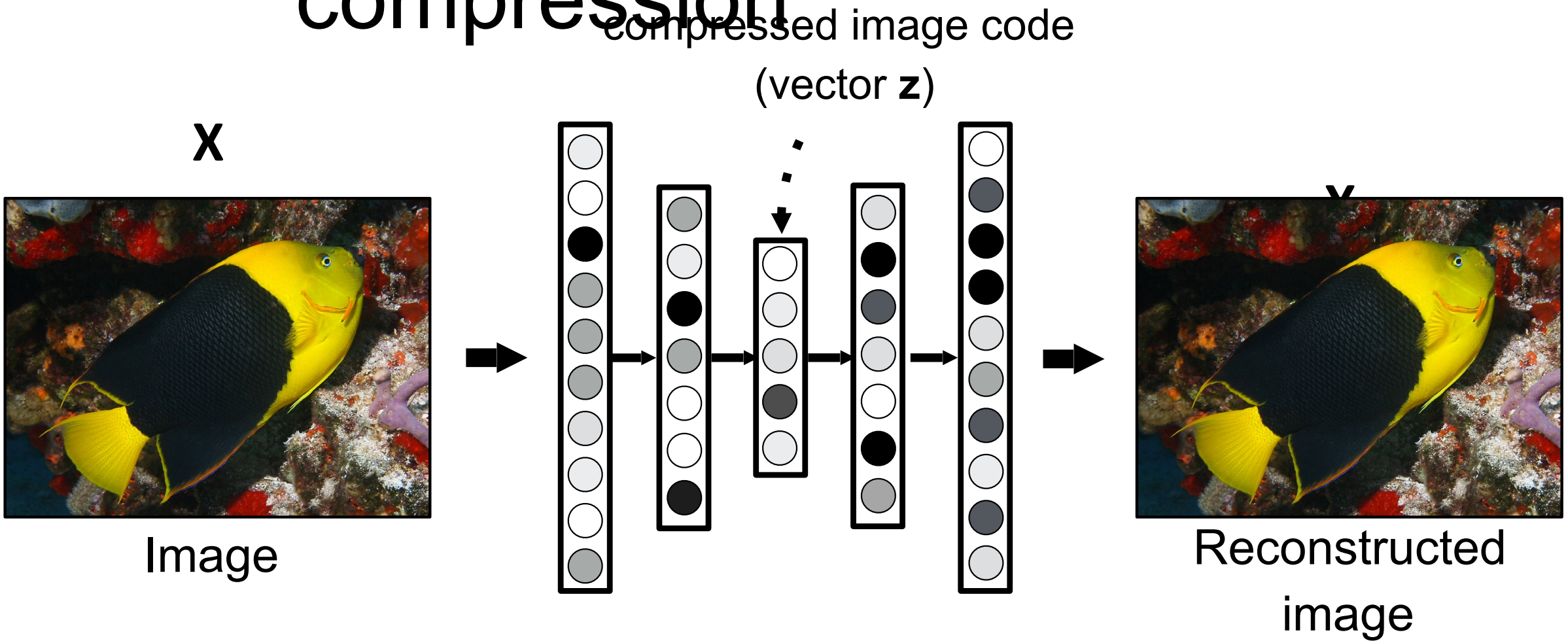
Learning via compression



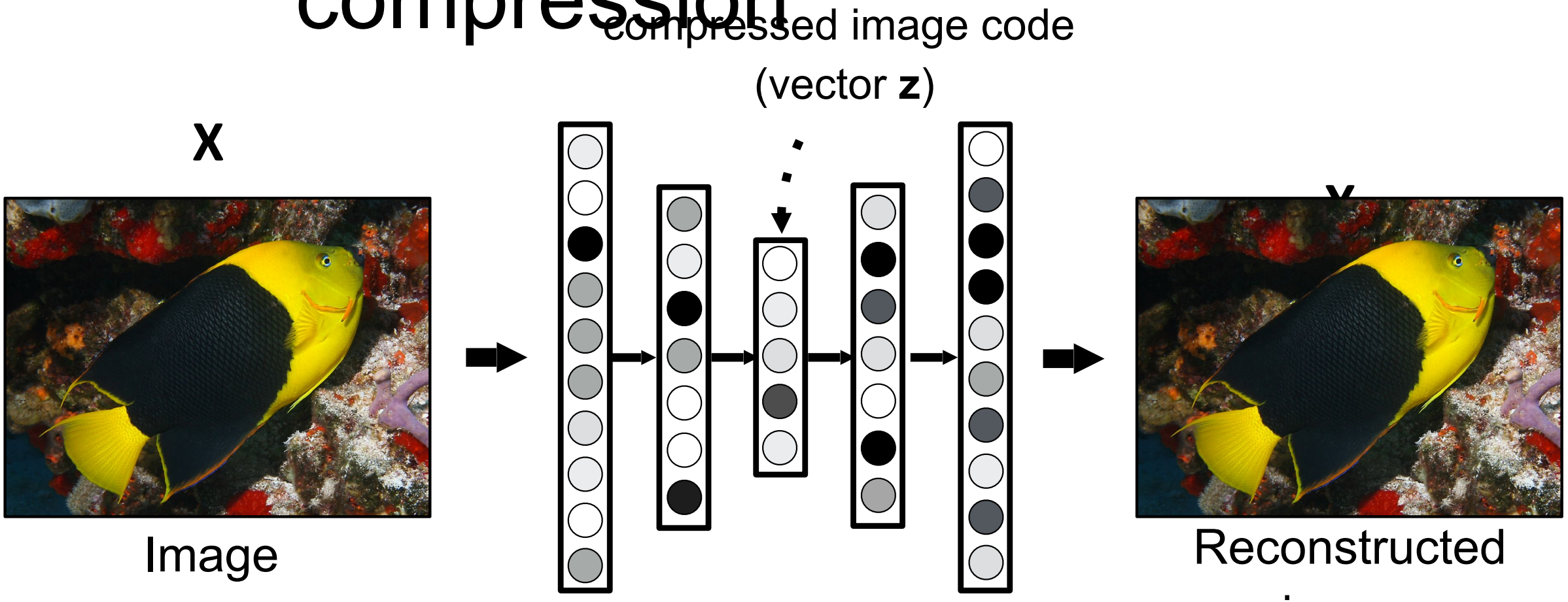
Learning via compression



Learning via compression



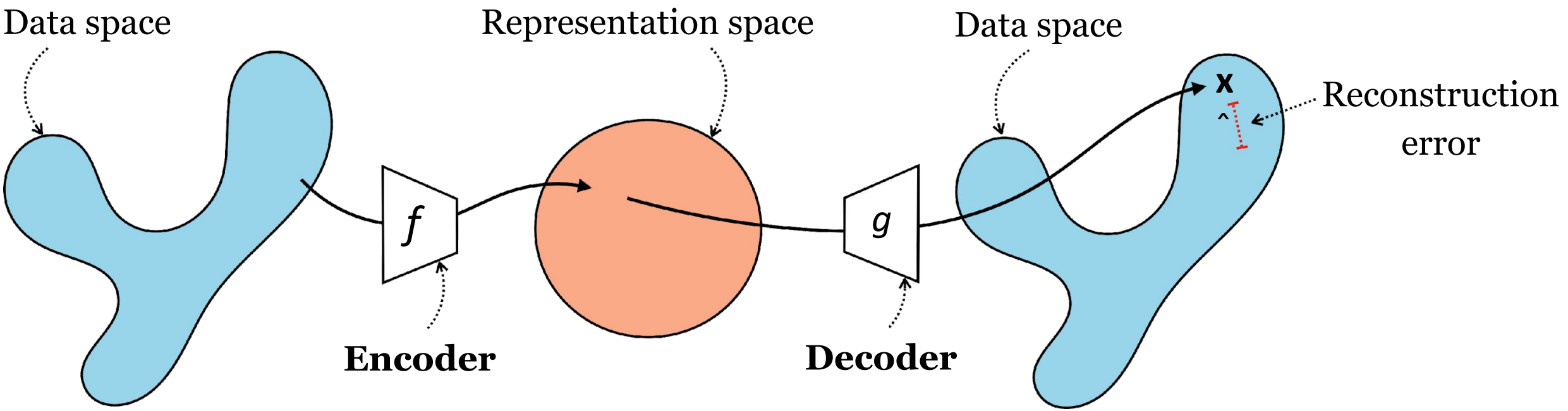
Learning via compression



"Autoencoder"

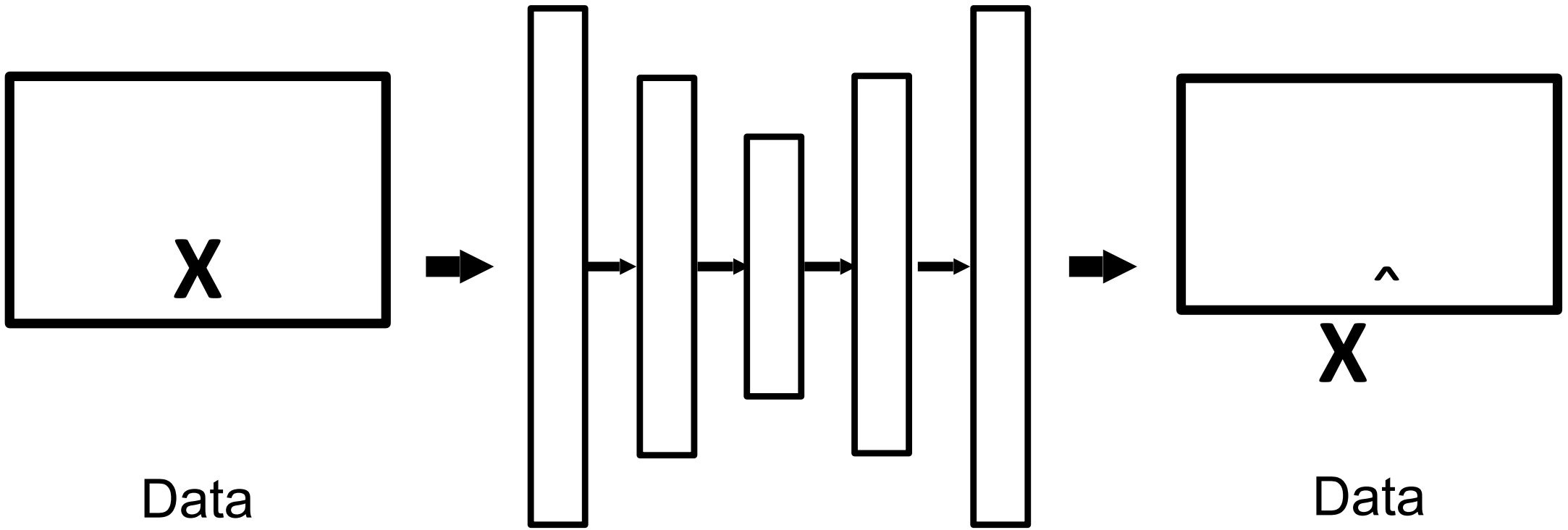
[e.g., Hinton & Salakhutdinov, Science 2006]

Autoencoder



$$f^{\leftarrow}, g^{\leftarrow} = \arg\min_{f,g} \mathsf{E}_{\mathbf{x}} \quad \|\mathbf{x} - g(f(\mathbf{x}))\|_2^2$$

Datacompression



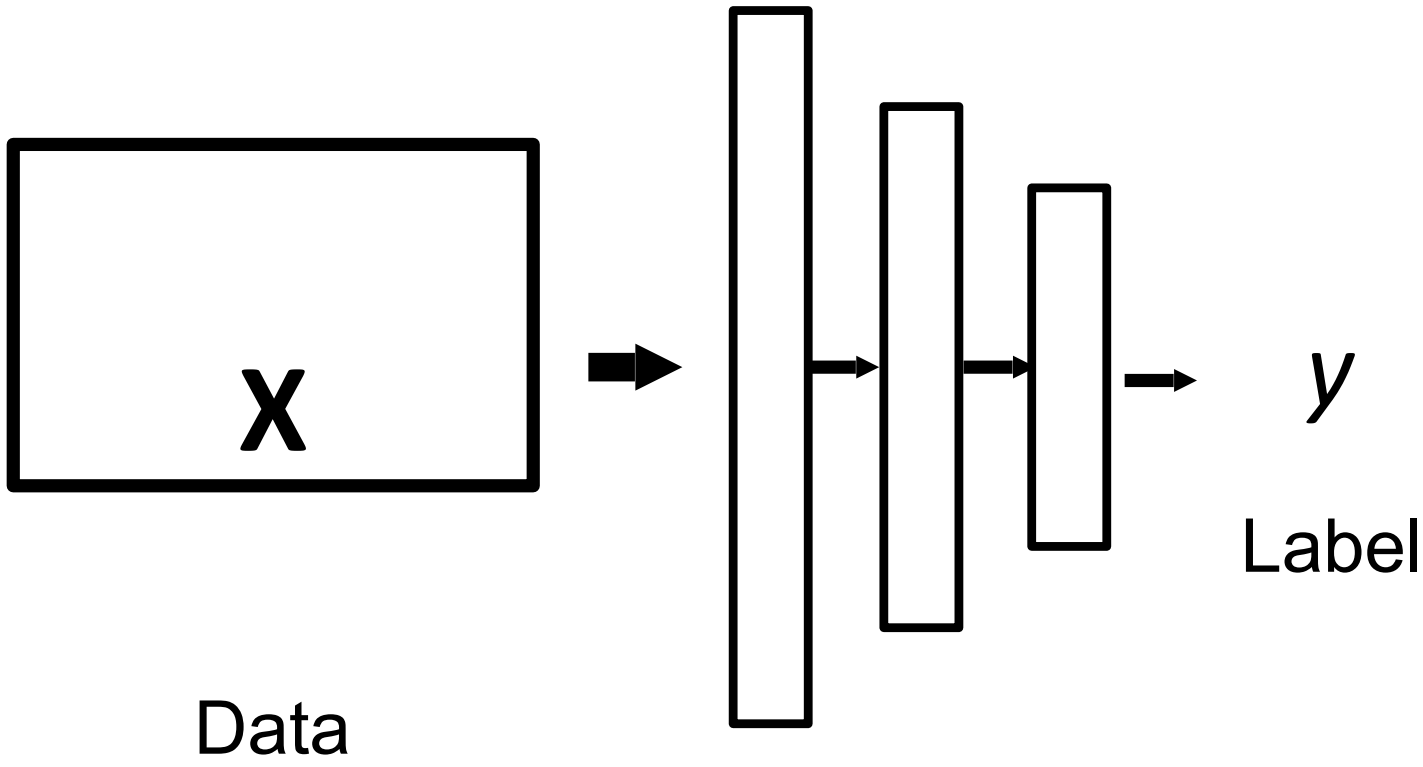
Self-Supervised Learning (SSL)

Compressive

Predictive

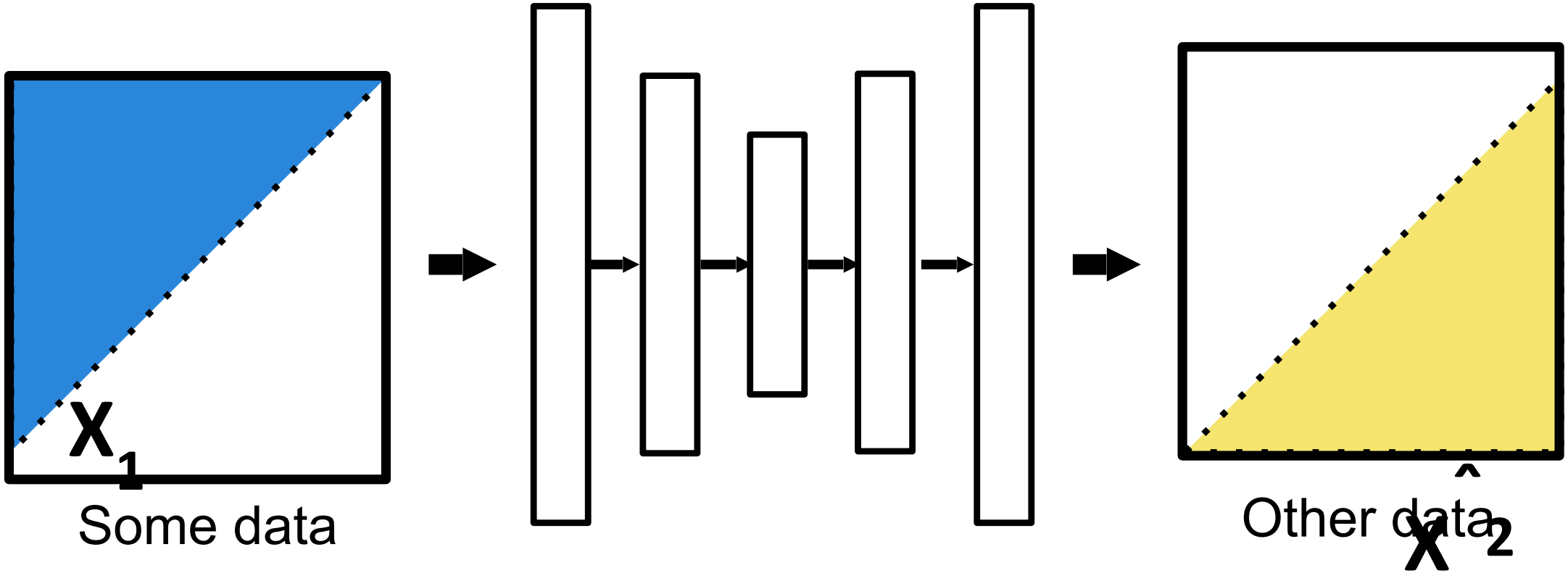
Contrastive

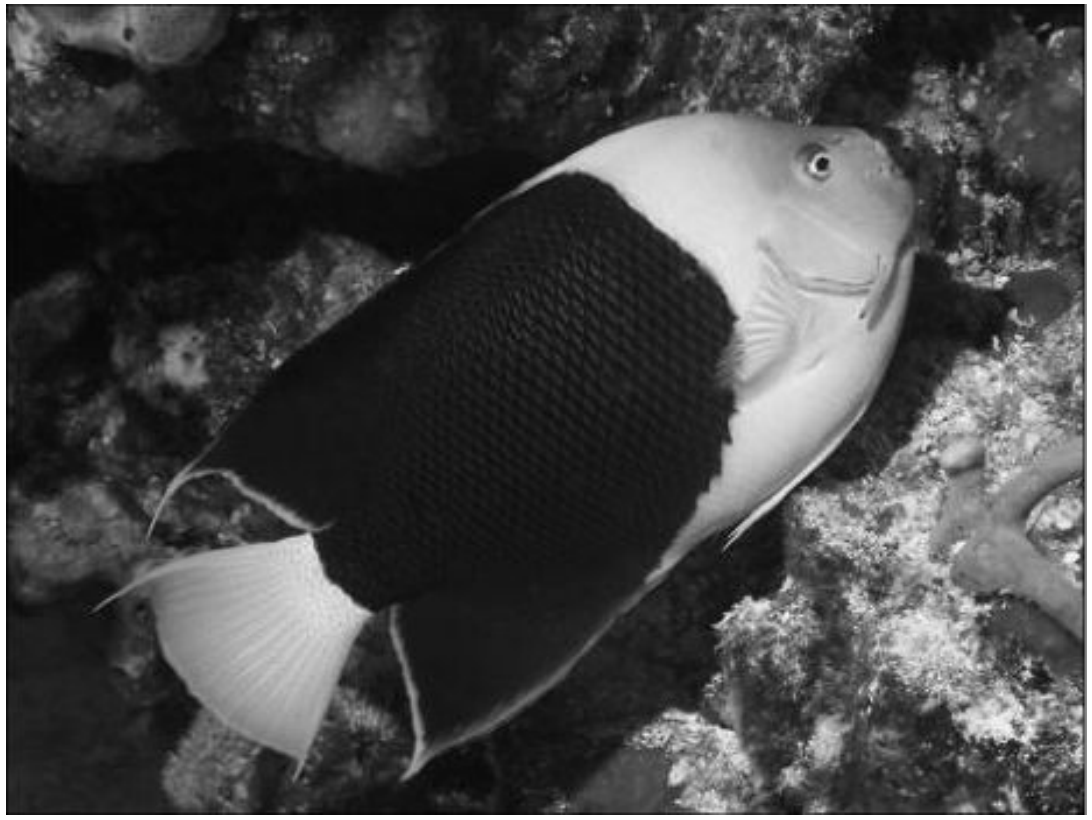
Label prediction



Data prediction

aka “self-supervised learning”





Grayscale image: L chan $\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

L

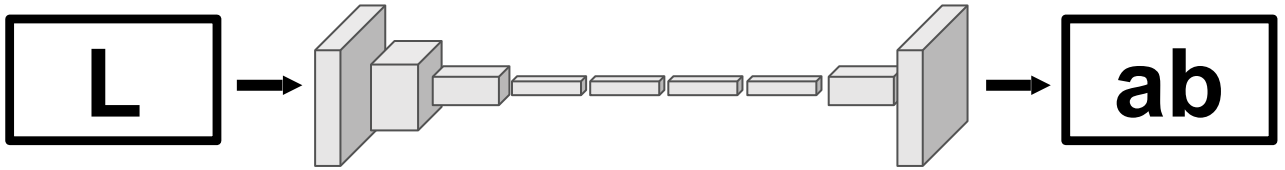
[Zhang, Isola, Efros, ECCV 2016]

$\mathcal{F}$

Grayscale image: L chan$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

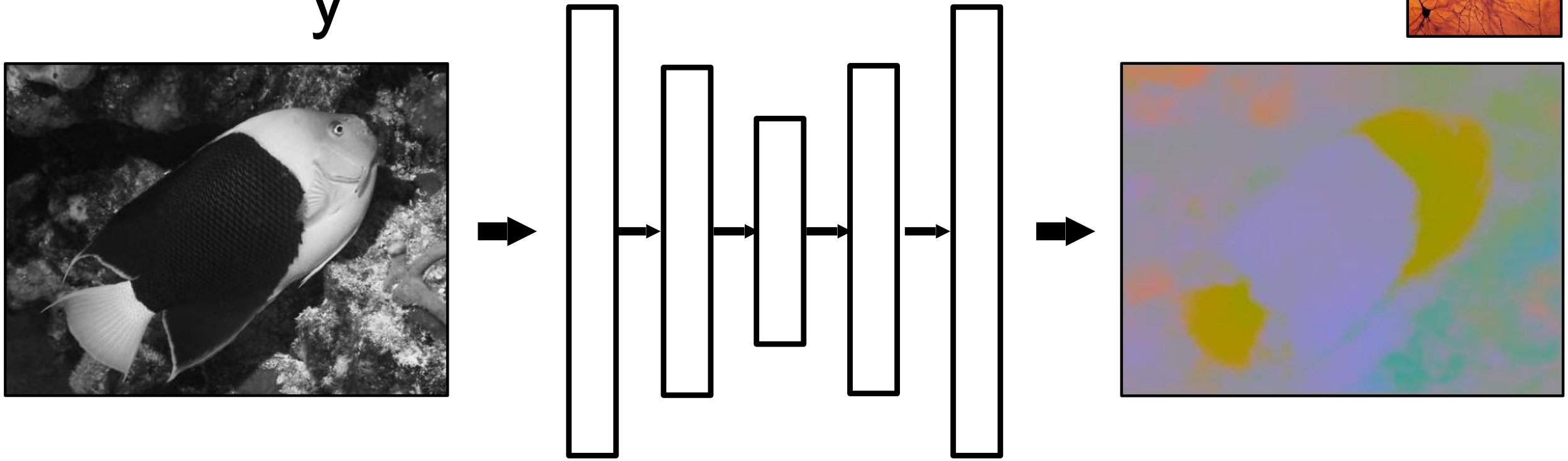
Color information: ab channels

$\hat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$



[Zhang, Isola, Efros, ECCV 2016]

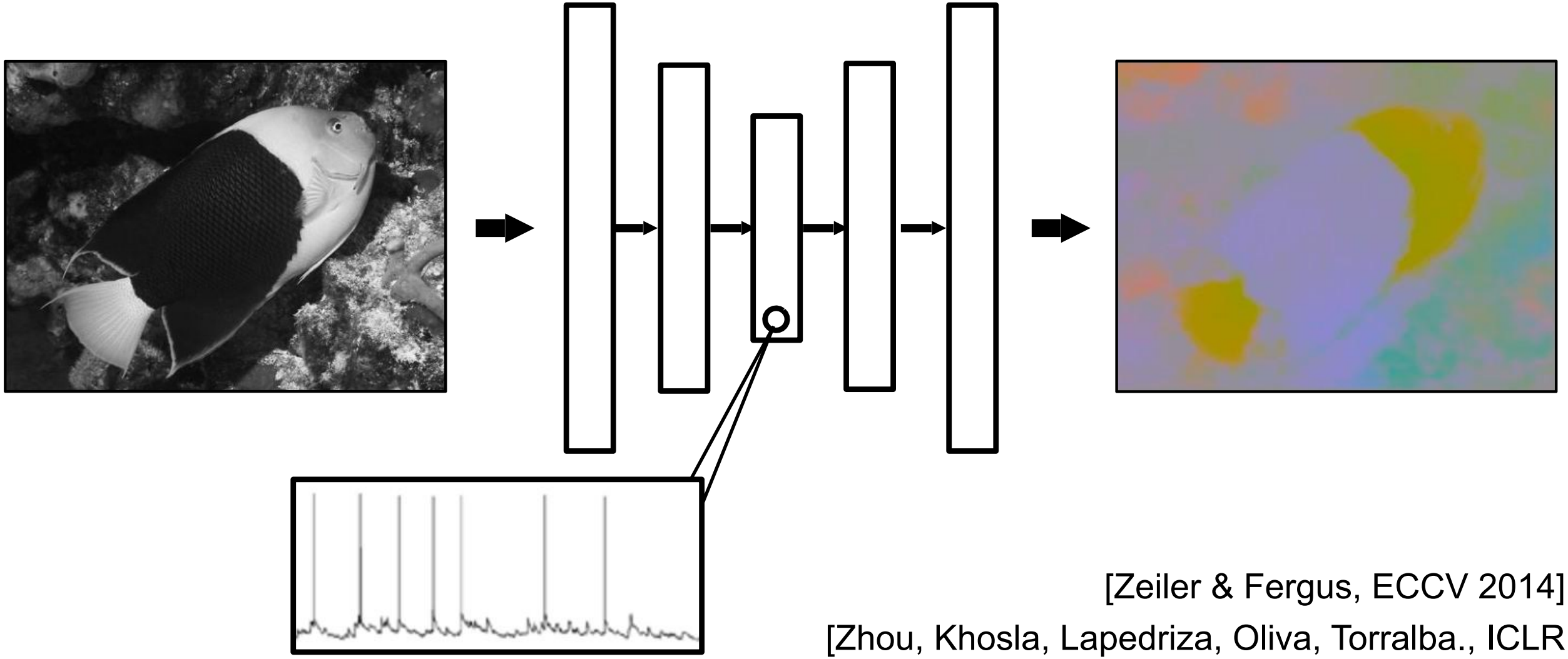
Deep Net “Electrophysiology”



[Zeiler & Fergus, ECCV 2014]

[Zhou, Khosla, Lapedriza, Oliva, Torralba., ICLR 2015]

Deep Net



[Zeiler & Fergus, ECCV 2014]

[Zhou, Khosla, Lapedriza, Oliva, Torralba., ICLR 2015]

Stimuli that drive selected neurons (conv5 layer)

faces

Stimuli that drive selected neurons (conv5 layer)

faces

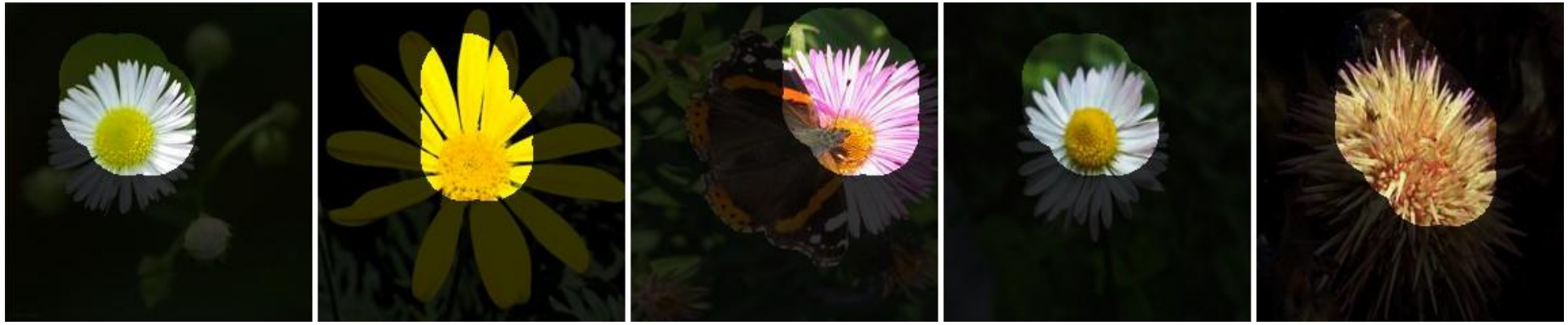
dog faces

Stimuli that drive selected neurons (conv5 layer)

face
s

dog
face
s

flowers



Predictive Learning: Language Models

Next word prediction (GPT)

- Predict the next word (token) given a prefix

Masked language modeling (BERT)

- Predict the masked words (tokens) in a text

Radford, et al., "Improving Language Understanding by Generative Pre-Training", 201

Devlin, et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 201

Predictive Learning: Computer Vision

Masked image modeling (Context Encoders)

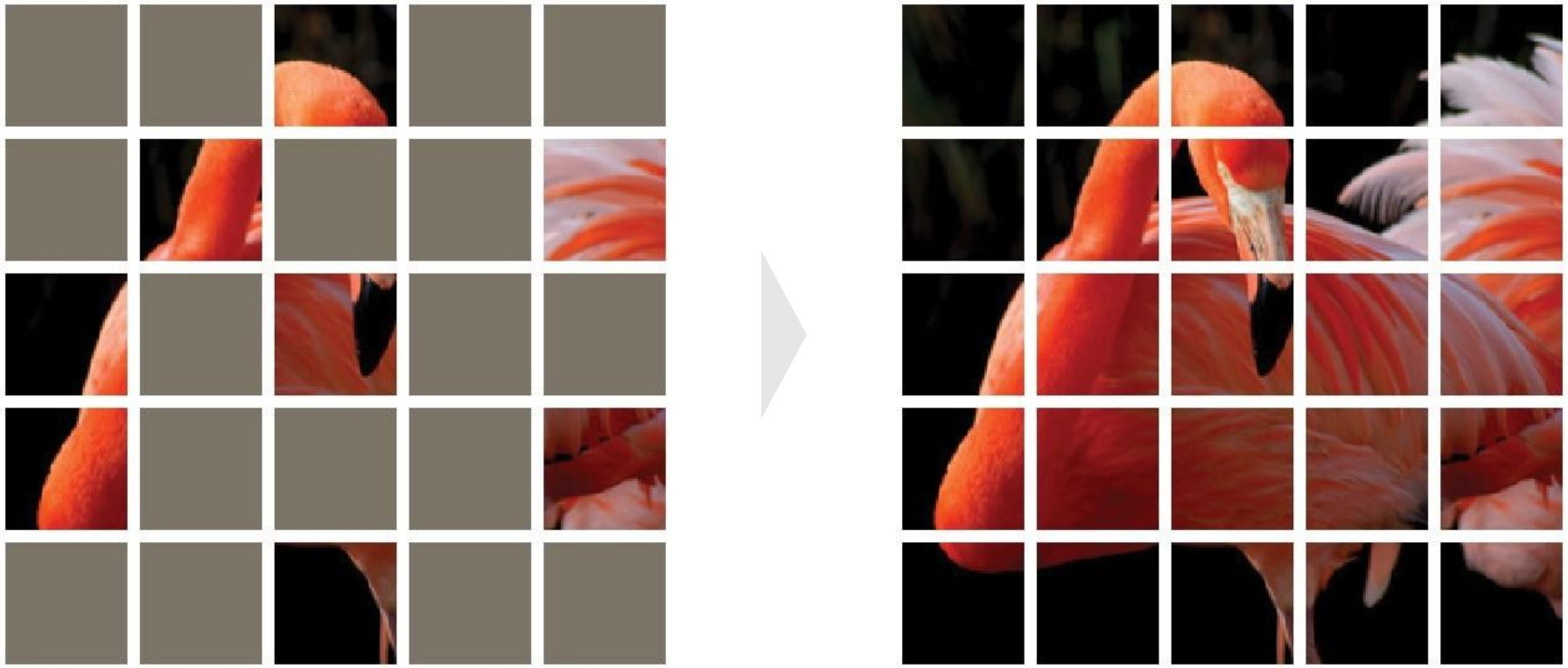
- Predict the masked regions using ConvNets

Pathak, et al., “Context Encoders: Feature Learning by Inpainting”, CVPR 2016

Predictive Learning: Computer Vision

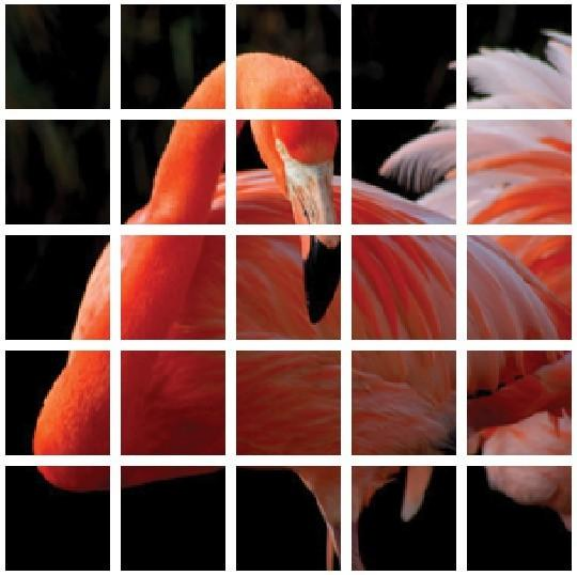
Masked image modeling (Masked Autoencoder)

- Predict the masked patches using Transformers



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

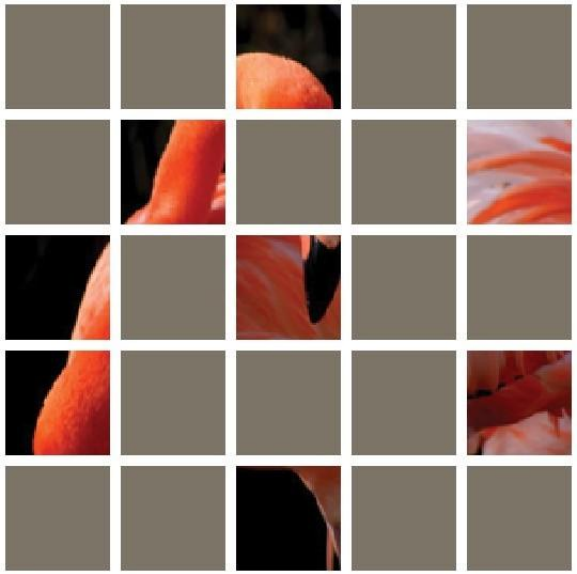
Masked Autoencoder (MAE)



patches as visual tokens (Vision Transformer)

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

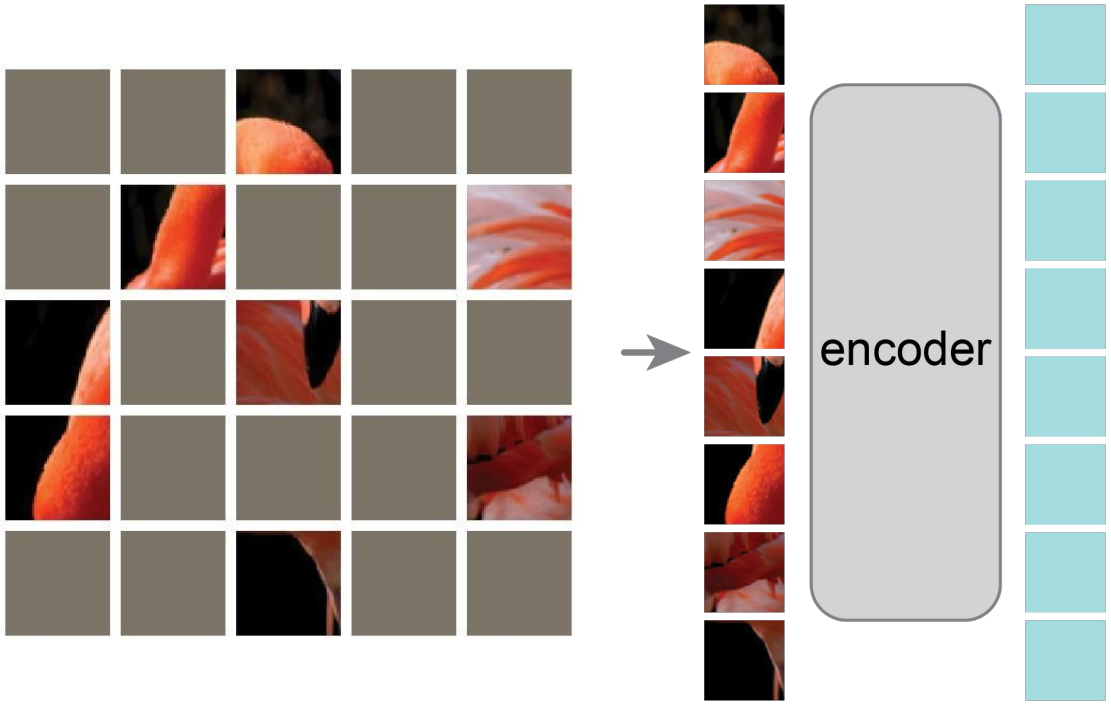
Masked Autoencoder (MAE)



random masking

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

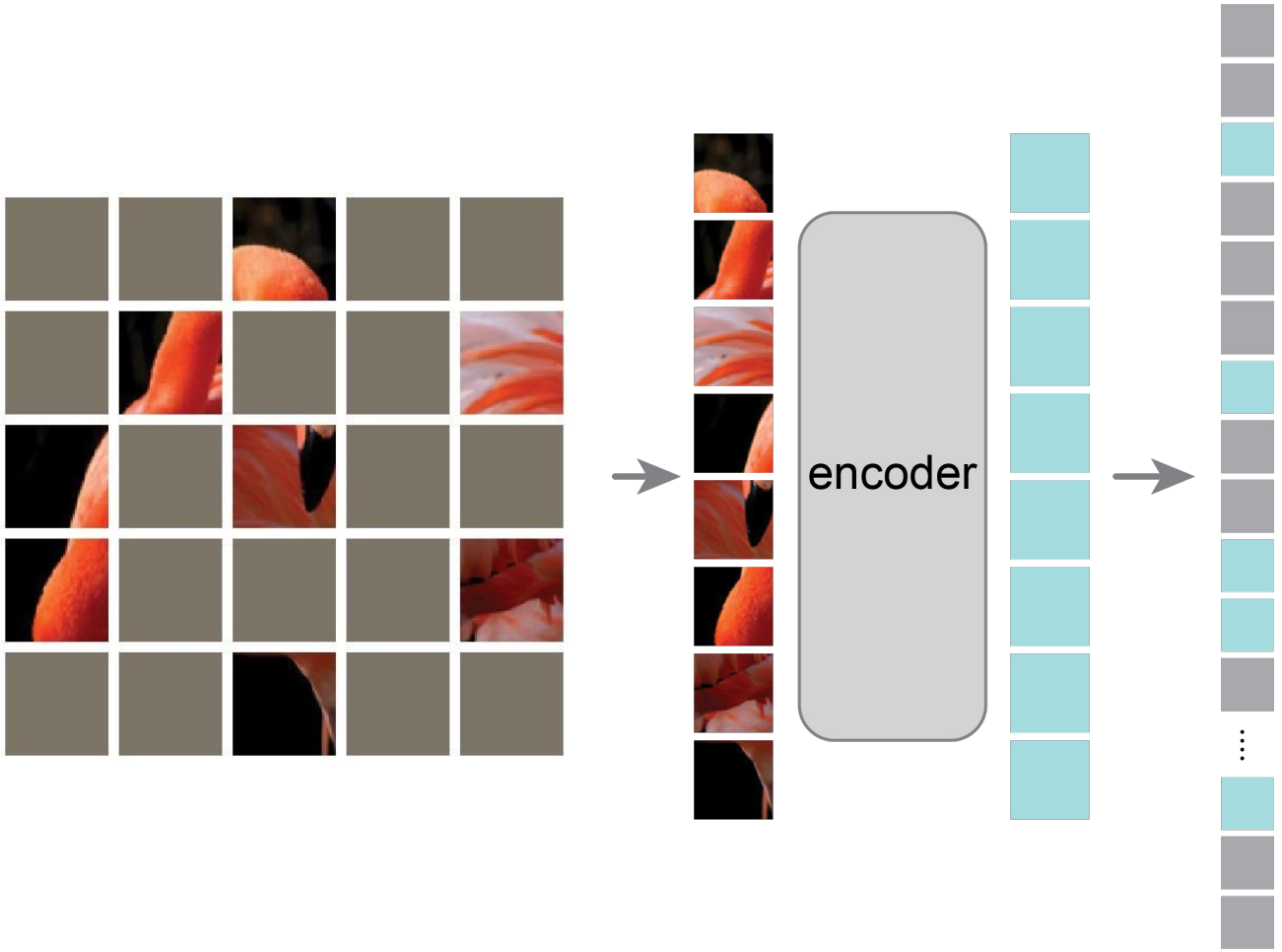
Masked Autoencoder (MAE)



encode visible patches w/ Transformer

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

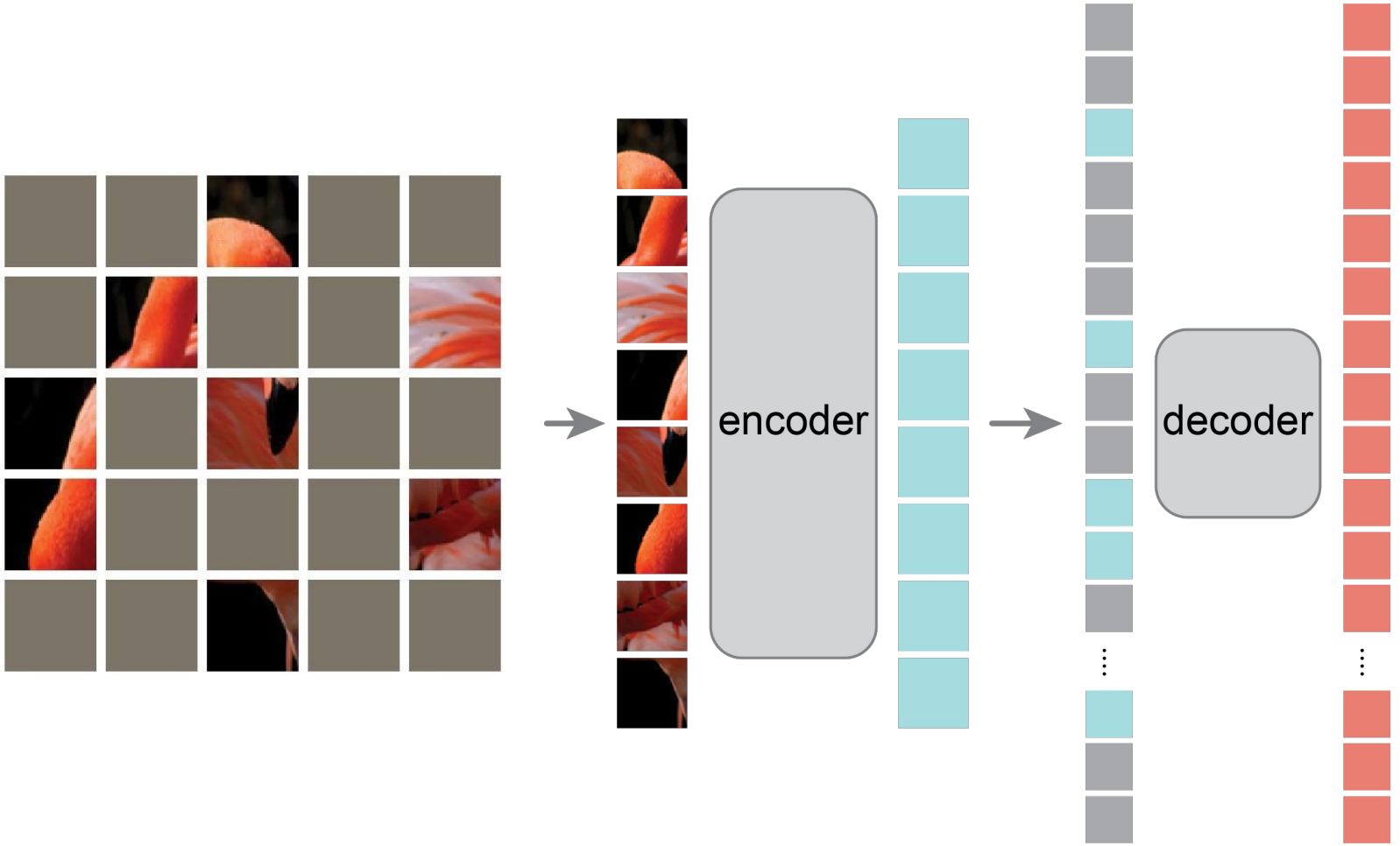
Masked Autoencoder (MAE)



- expand with “mask tokens” (specify where to predict)

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

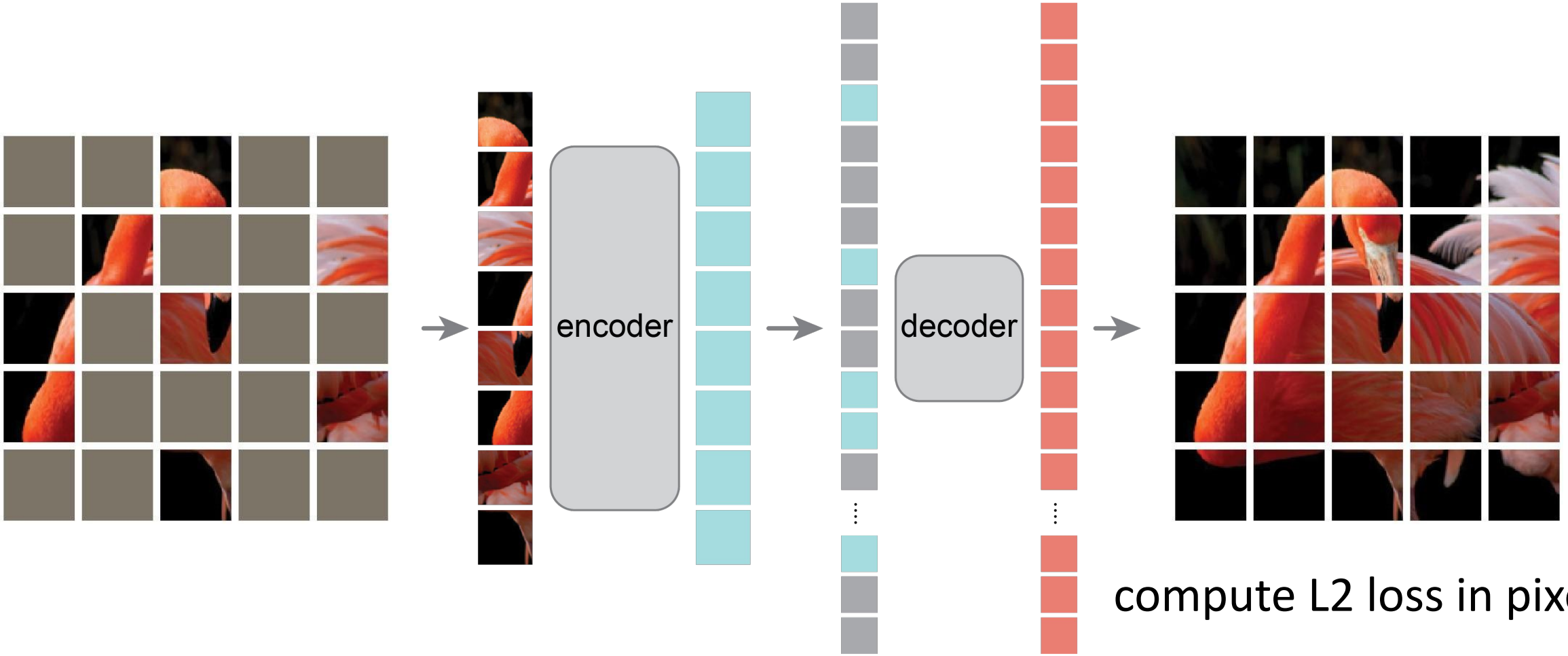
Masked Autoencoder (MAE)



predict the unknown

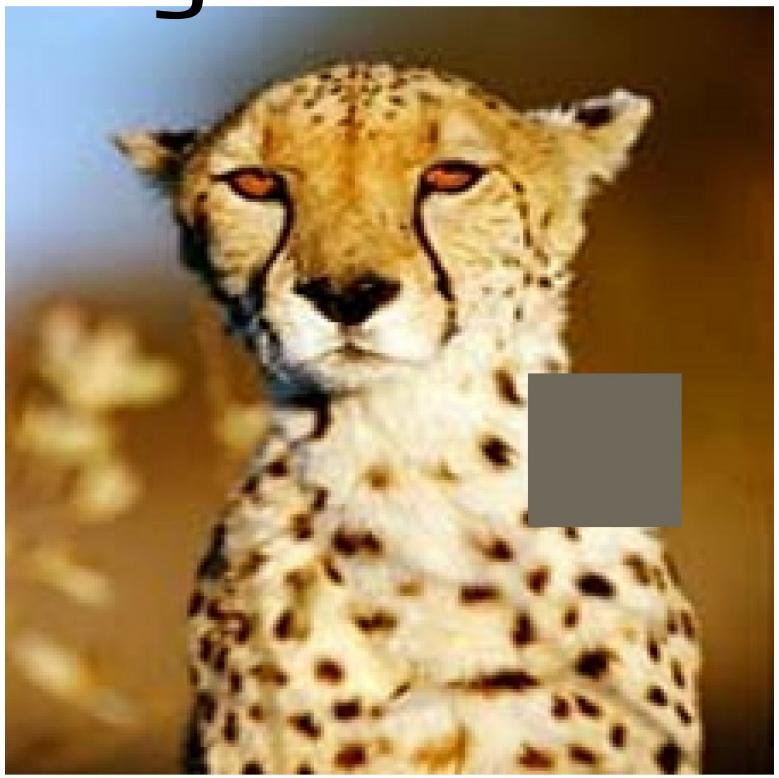
He, et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022

Masked Autoencoder (MAE)

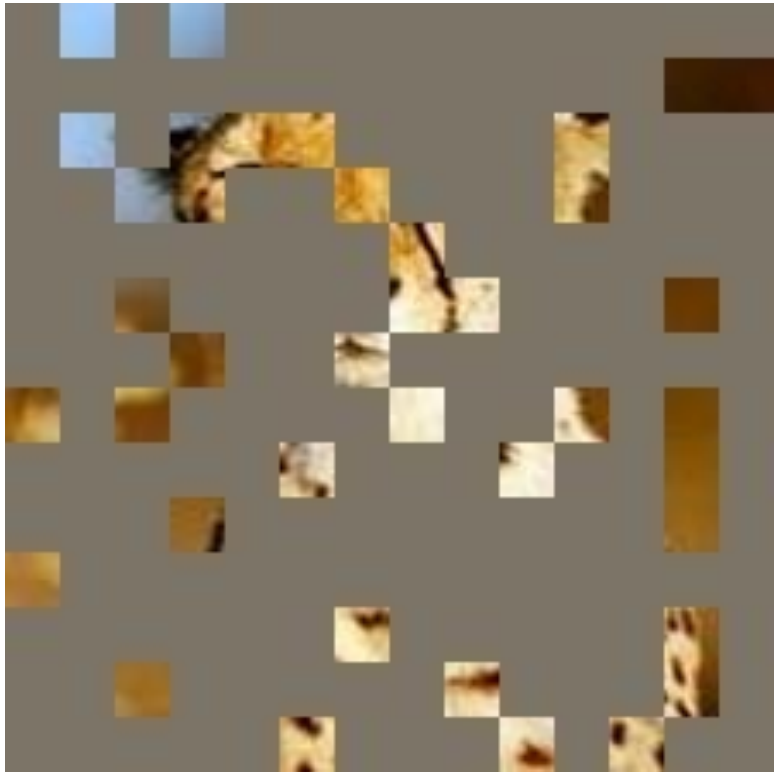


He, et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022

How to learn good representations by predicting?



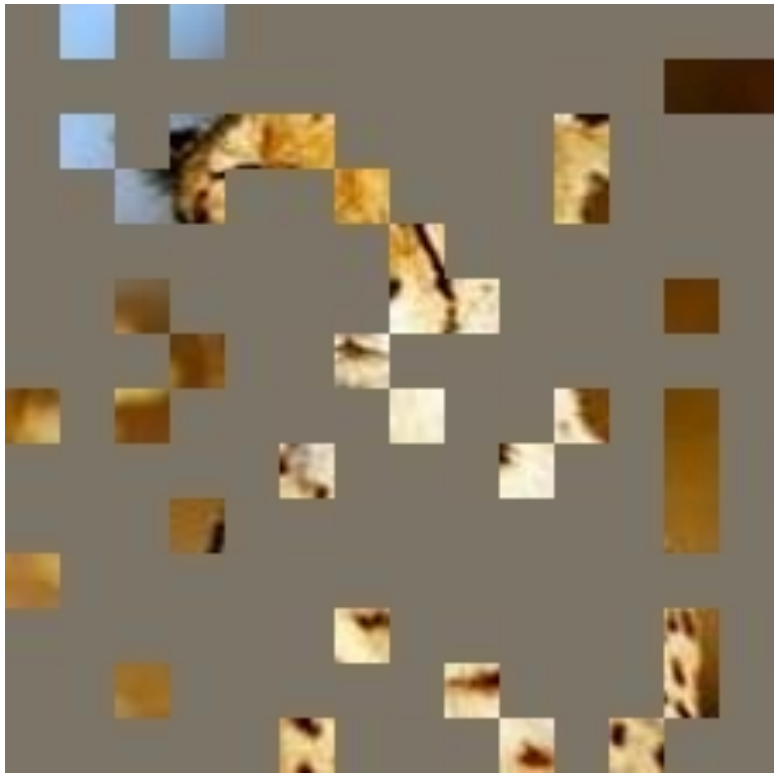
- predicting a small portion may not require high-level understanding



- predicting a large portion of unknown patches encourages to learn semantic features

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

How to learn good representations by predicting?

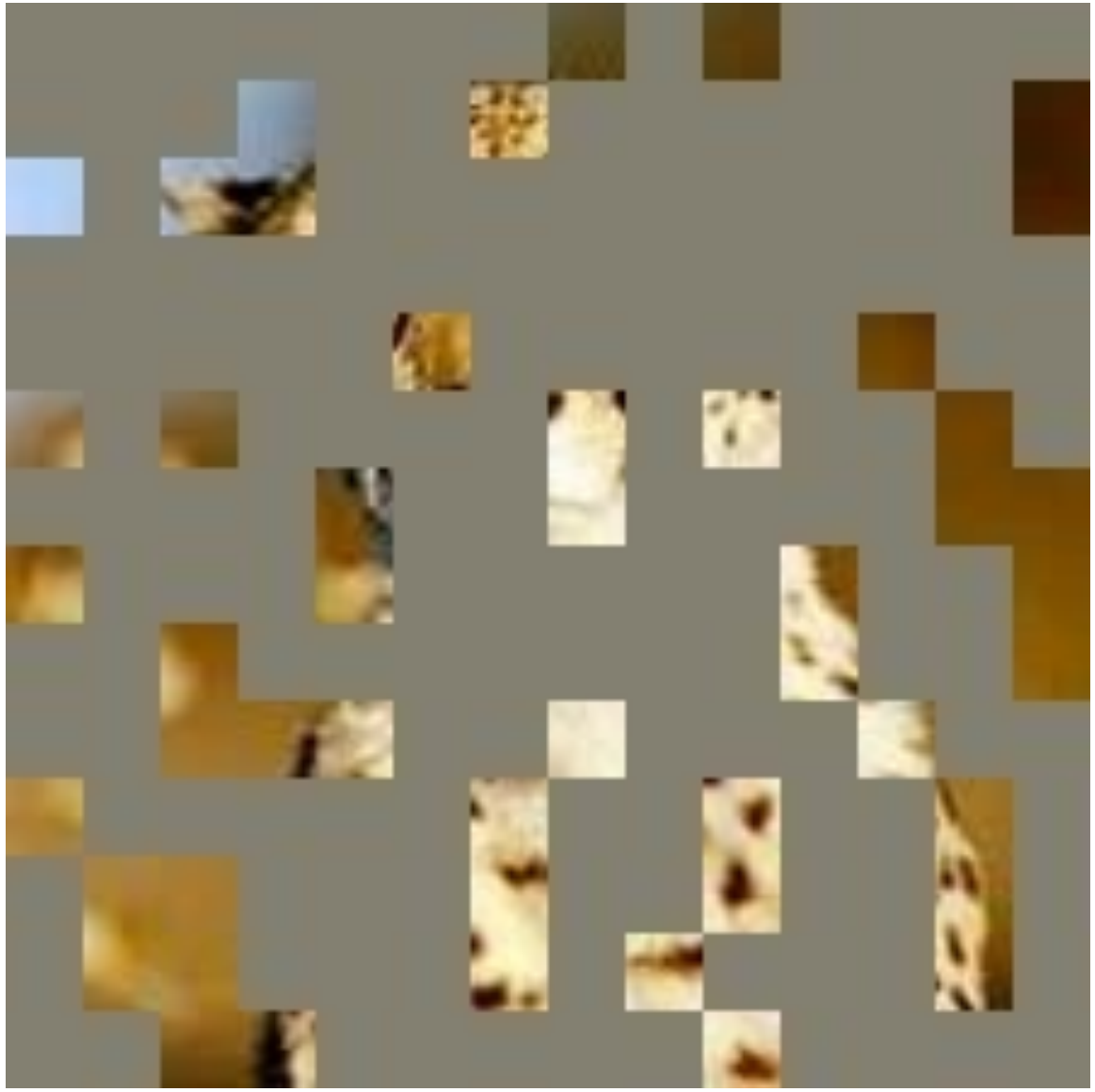


input

MAE prediction

original

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



- **The learning process**: the network gradually makes sense of the semantic patterns

- **The learning process**: the network gradually makes sense of the semantic patterns

Self-Supervised Learning (SSL)

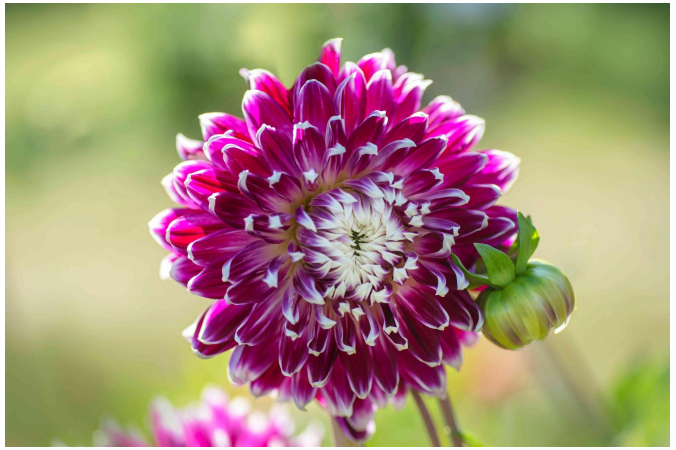
Compressive

Predictive

Contrastive

Contrastive Learning

Motivation: metric learning

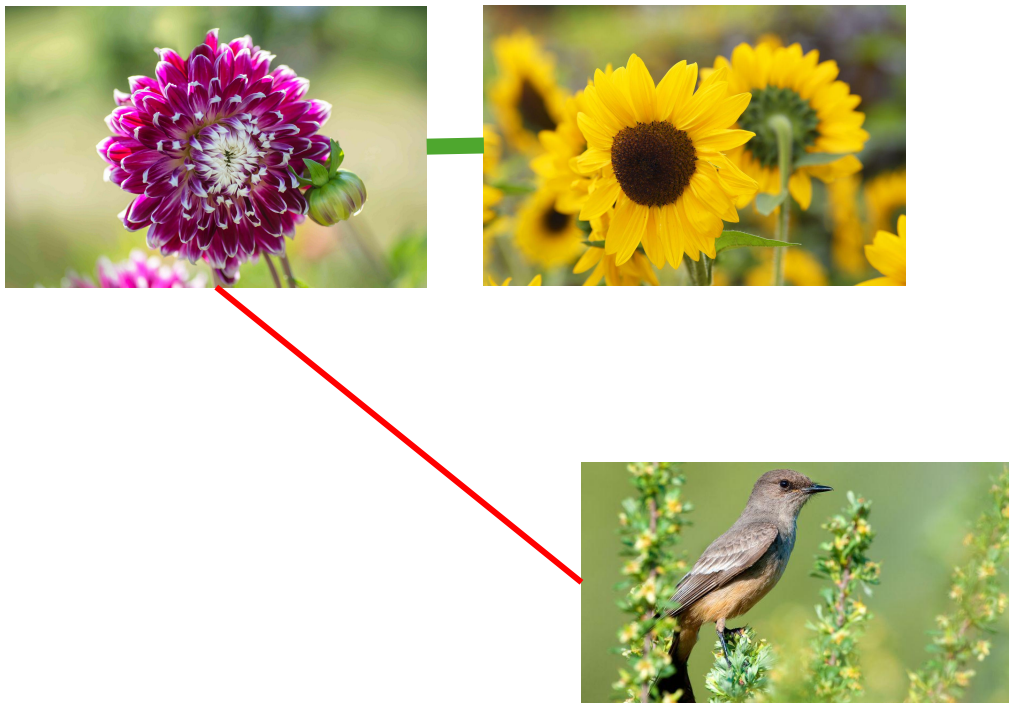


Contrastive Learning

Motivation: metric learning

Distances in Pixel Space

Distances in Human Representation

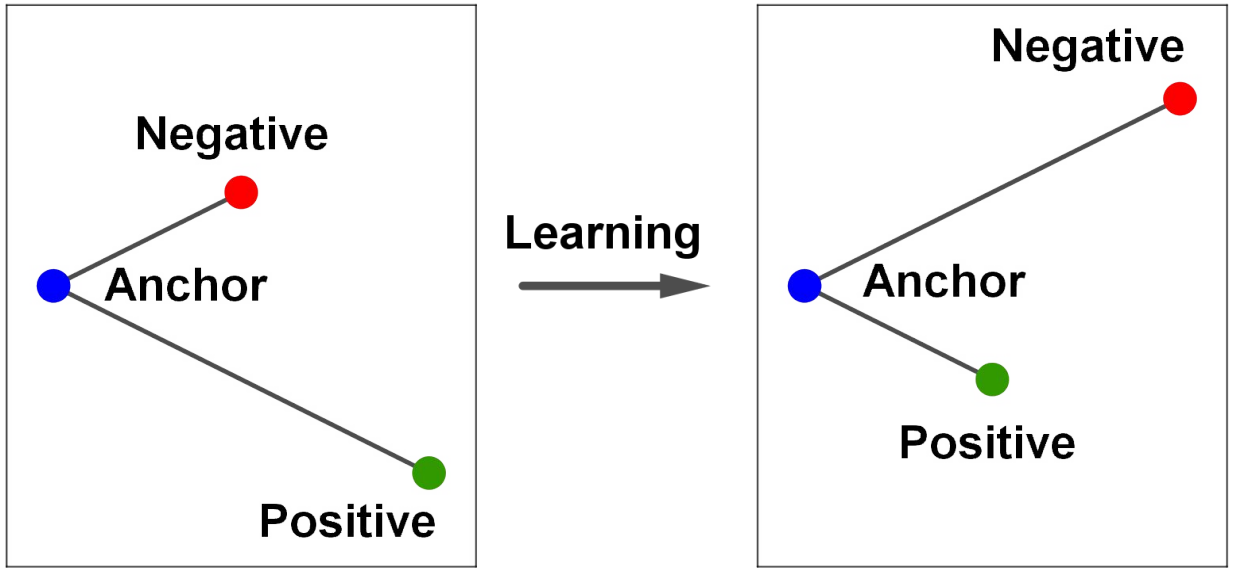


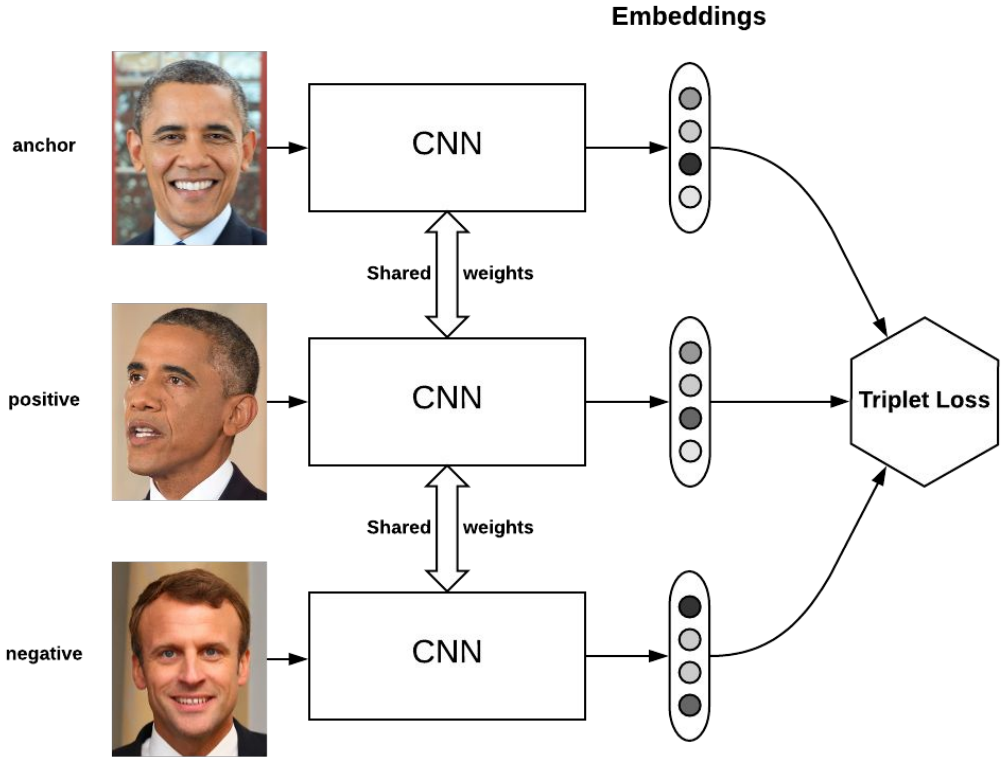
How do we get positive pairs without labels?

crops

augmentations

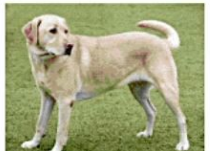
Triplet Loss





$$\mathcal{L}_{\text{triplet}}(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-) = \sum_{\mathbf{x} \in \mathcal{X}} \max\left(0, \|f(\mathbf{x}) - f(\mathbf{x}^+)\|_2^2 - \|f(\mathbf{x}) - f(\mathbf{x}^-)\|_2^2 + \epsilon\right)$$

SimCLR: Simple Framework for Contrastive Learning



. . .

Common setup

- Encoder maps data onto a hypersphere: $f : \mathcal{X} \to \mathbb{S}^{d-1}$
- Cross-entropy for softmax "classifier" to discriminate "classes" defined by similarities

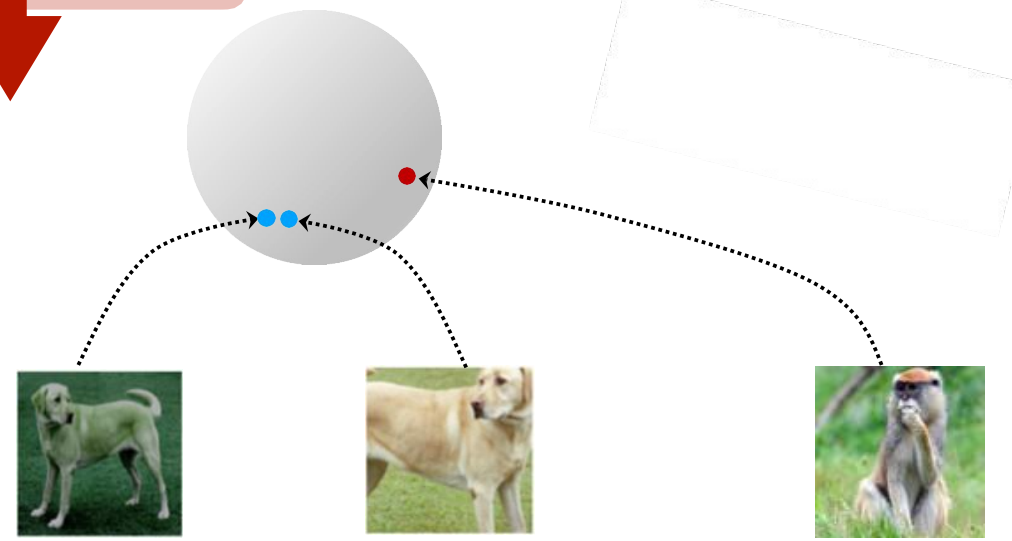
$$\min_f \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i)/\tau}} \right]$$

pull positive pair together

push negative pairs apart

Symmetry: $\forall \mathbf{x}, \mathbf{x}^+, p_{\text{pos}}(\mathbf{x}, \mathbf{x}^+) = p_{\text{pos}}(\mathbf{x}^+, \mathbf{x})$

Matching marginal: $\forall \mathbf{x}, \int p_{\text{pos}}(\mathbf{x}, \mathbf{x}^+) d\mathbf{x}^+ = p_{\text{data}}(\mathbf{x})$

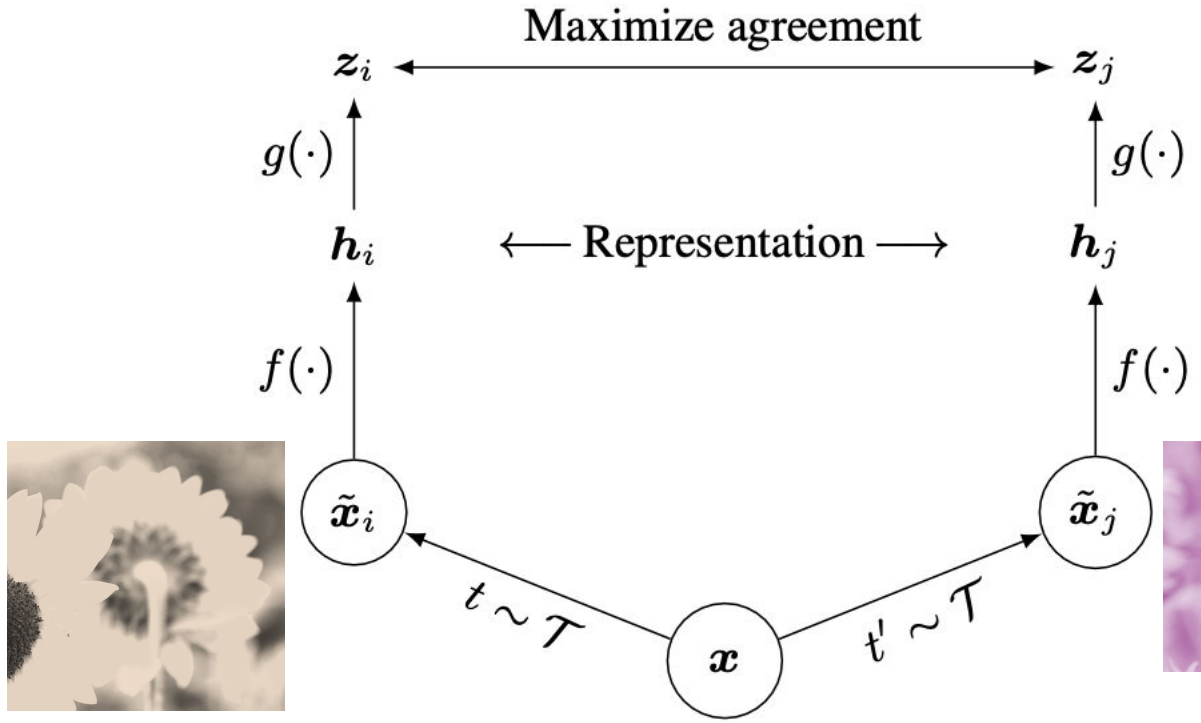


SimCLR: the implementation details

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of $f, g, \mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i, j)$ **as** $\ell(i, j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

Algorithm 1 summarizes the proposed method.



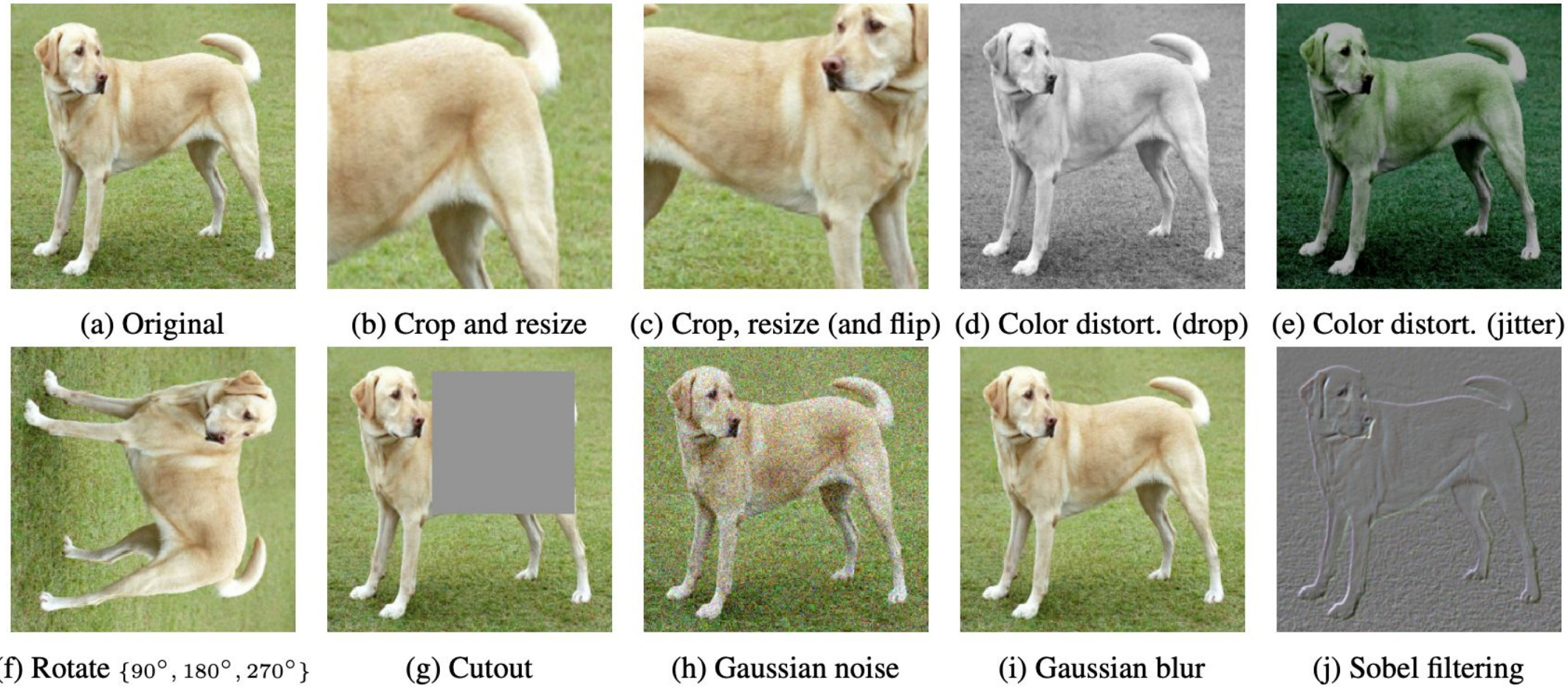
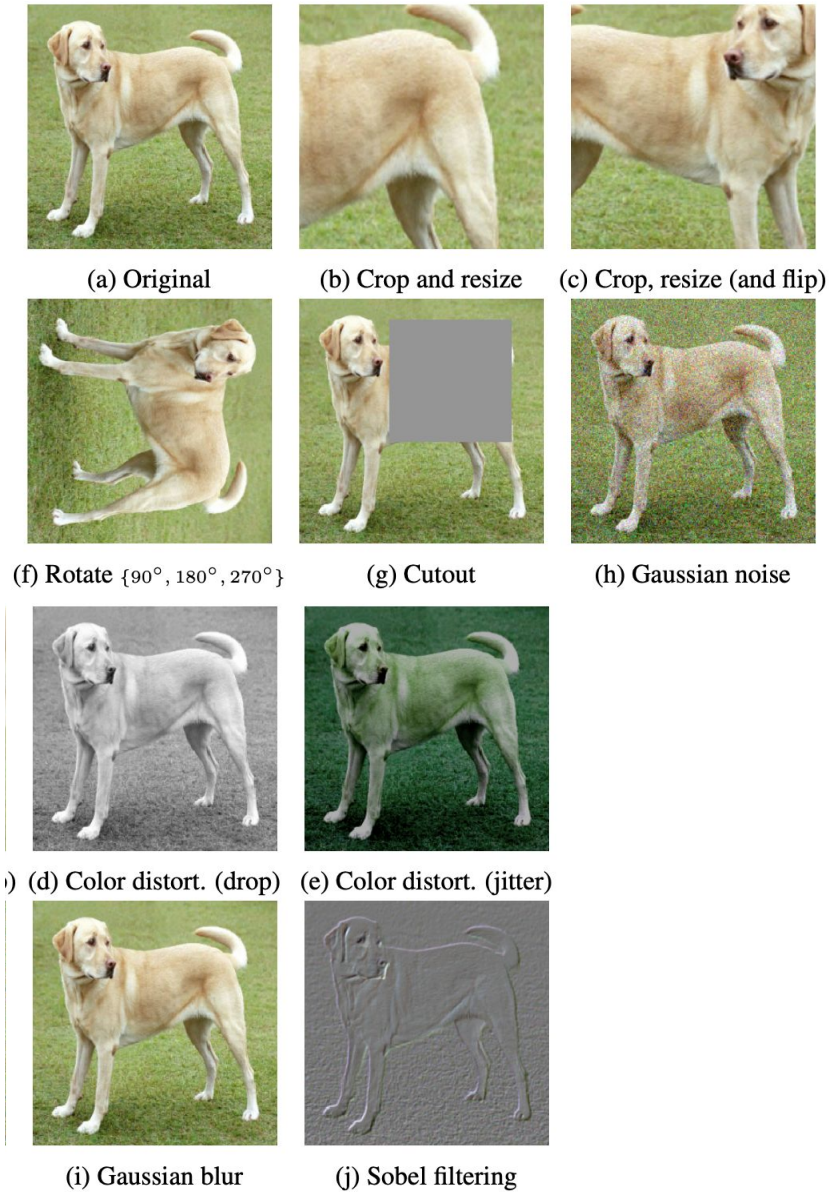


Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

A deeper look into augmentations



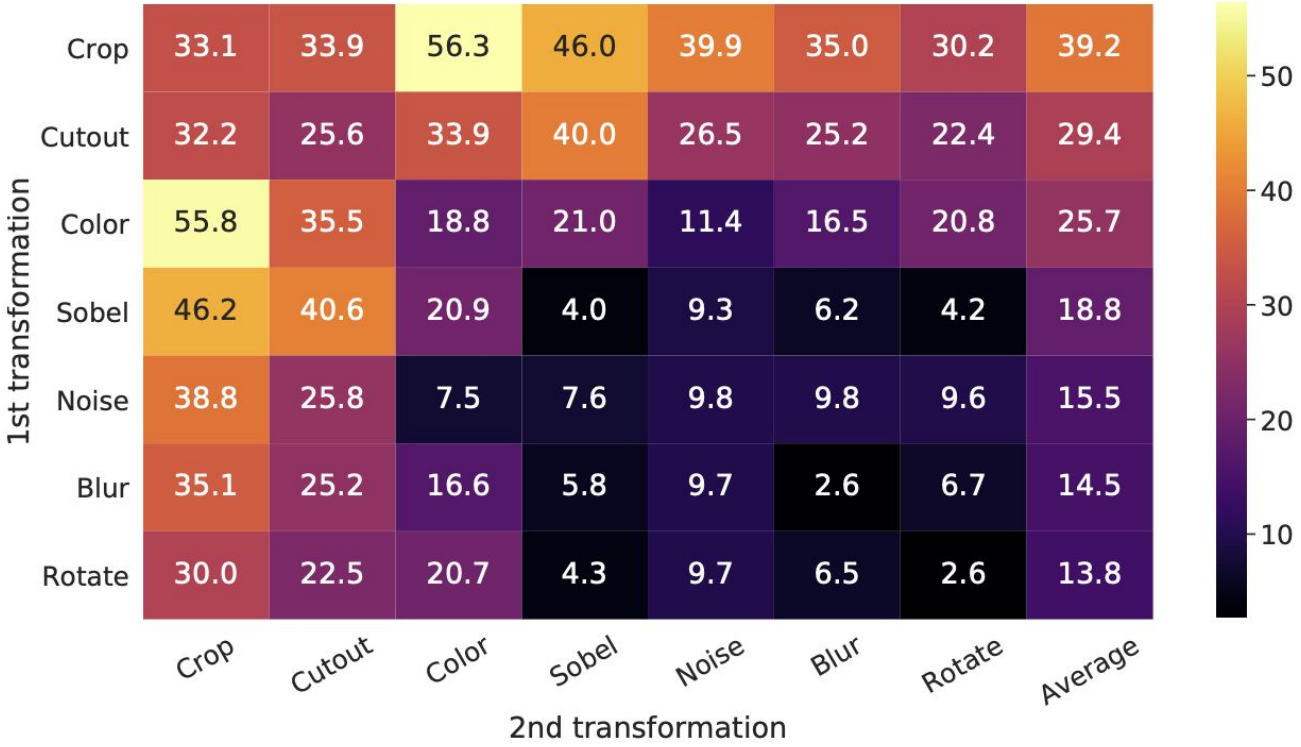
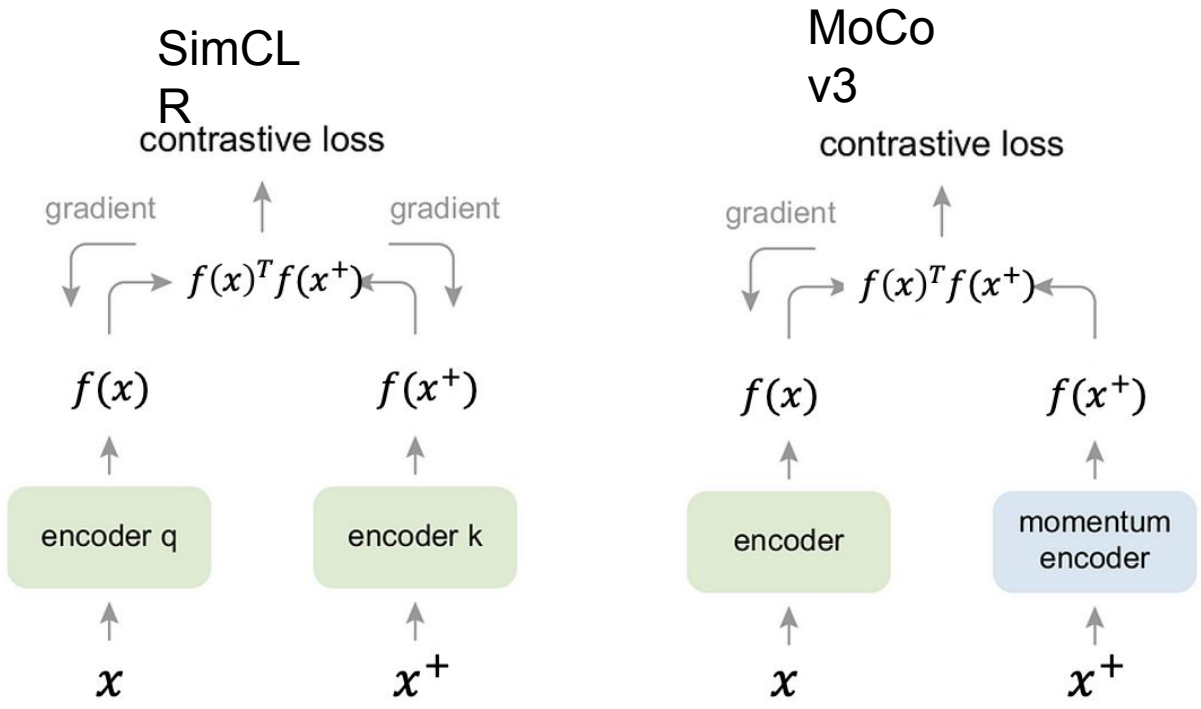


Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.

MoCo v3: **Mo**mentum **Co**ntrast



InfoNCE Loss

$$\min_f \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

pull positive pair together

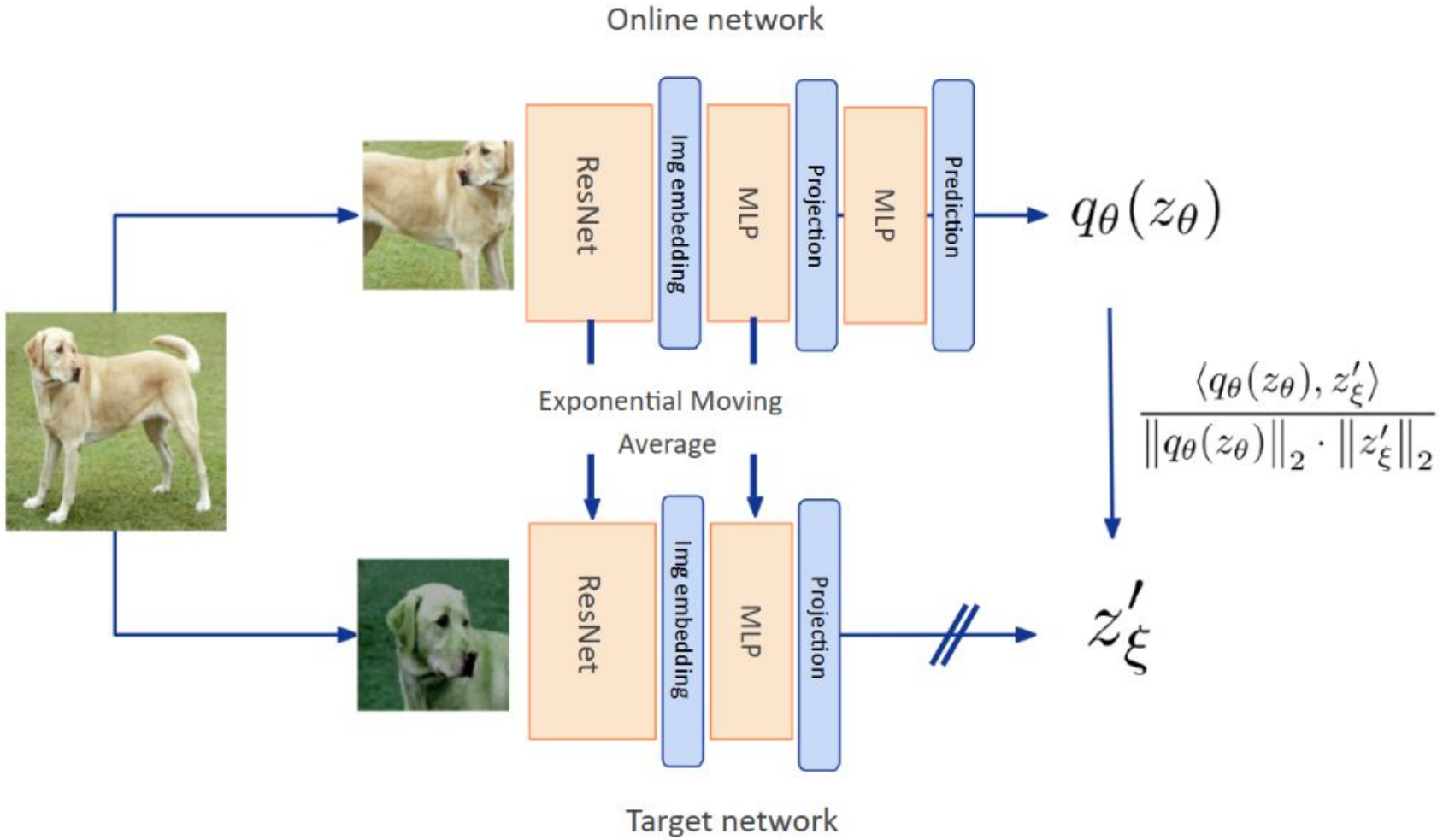
push negative pairs apart

Algorithm 1 MoCo v3: PyTorch-like Pseudocode

```
# f_q: encoder: backbone + proj mlp + pred mlp
# f_k: momentum encoder: backbone + proj mlp
# m: momentum coefficient
# tau: temperature

for x in loader: # load a minibatch x with N samples
   x1, x2 = aug(x), aug(x) # augmentation
   q1, q2 = f_q(x1), f_q(x2) # queries: [N, C] each
   k1, k2 = f_k(x1), f_k(x2) # keys: [N, C] each

   loss = ctr(q1, k2) + ctr(q2, k1) # symmetrized
   loss.backward()

   update(f_q) # optimizer update: f_q
   f_k = m*f_k + (1-m)*f_q # momentum update: f_k

# contrastive loss
def ctr(q, k):
   logits = mm(q, k.t()) # [N, N] pairs
   labels = range(N) # positives are in diagonal
   loss = CrossEntropyLoss(logits/tau, labels)
   return 2 * tau * loss
```

Notes: `mm` is matrix multiplication. `k.t()` is `k`'s transpose. The prediction head is excluded from `f_k` (and thus the momentum update).

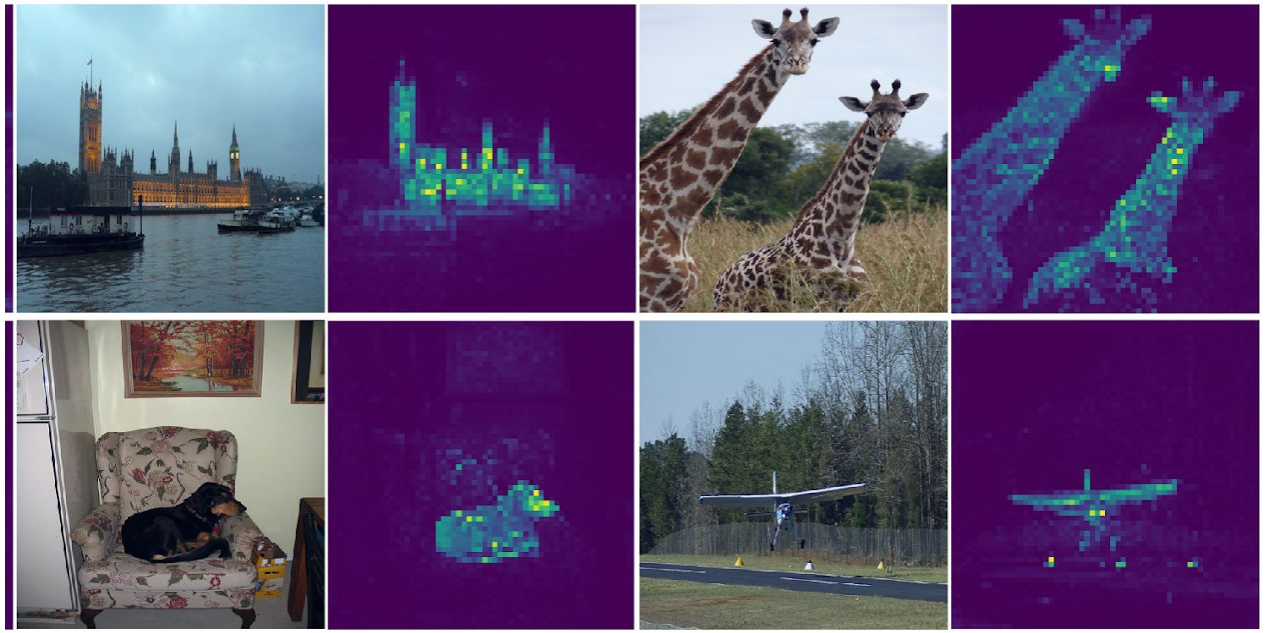
BYOL: **B**ootstrap **Y**our **O**wn **L**atent

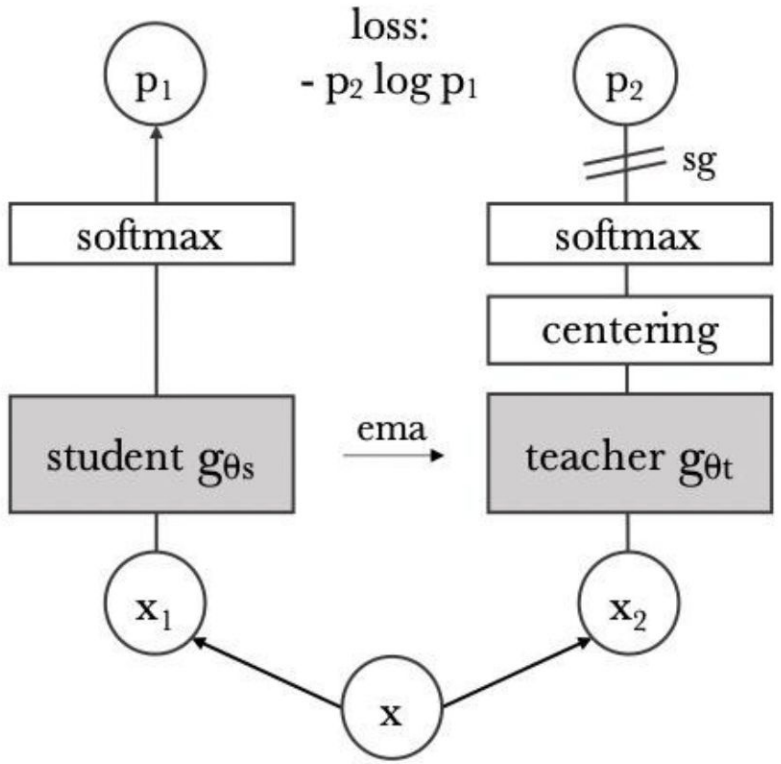


How does BYOL work?

- There are two networks: an **online network** and a **target network**. Both start out the same.
- Each network has an encoder (think of it as a feature extractor).
- The online network has an extra part called the **predictor head**.
- The target network's parameters are updated slowly using the online network's parameters (using something called an exponential moving average).
- The goal is simple: make the online network's prediction as close as possible to the target network's output, using mean squared error between their normalized outputs.

DiNO: **DI**stillation with **NO** labels



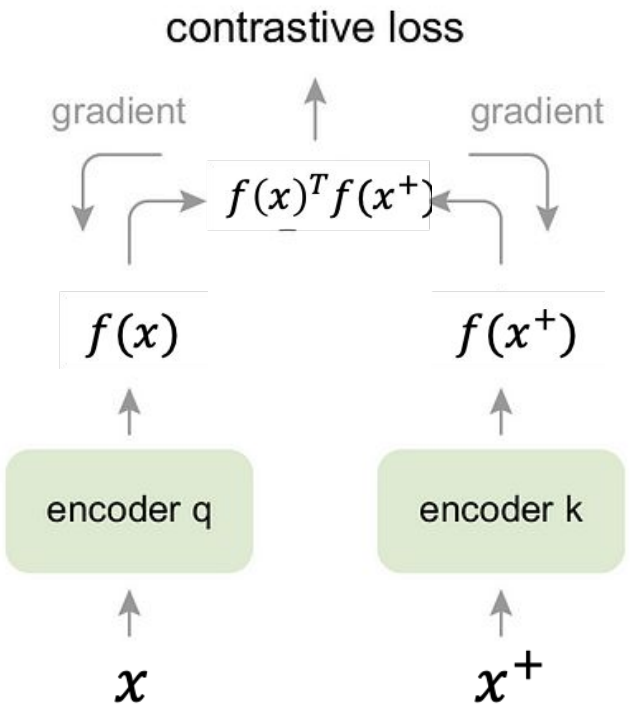


DiNO: **DI**stillation with **NO** labels

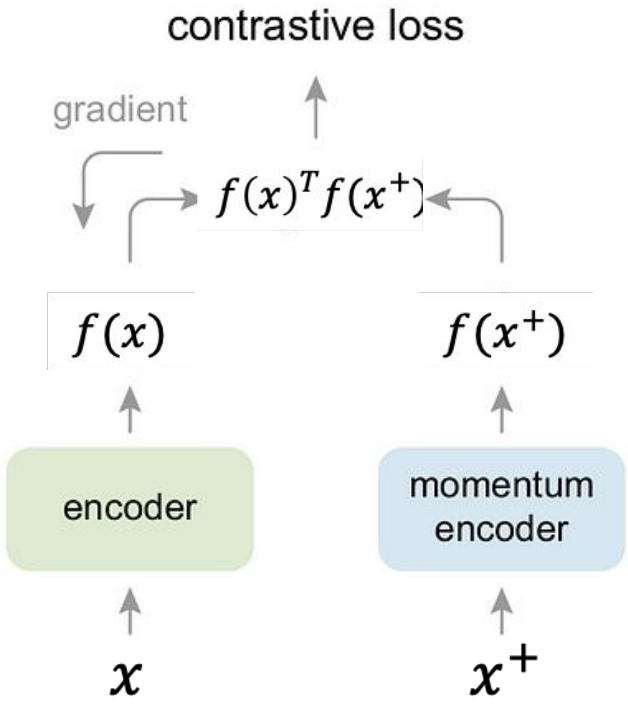
How does DINO work?

- **Student and Teacher Networks:**
  There are two networks with the same design. The teacher is just a slowly updated version of the student (using EMA).
- **Multi-crop Strategy:**
  The model looks at the same image in different ways—two large views and several small crops—to learn features that work at different scales.
- **No Negatives Needed:**
  Instead of comparing with negative samples, DINO uses a cross-entropy loss on soft similarity scores between the student and teacher outputs.
- **Avoiding Collapse:**
  To make sure the model doesn't just output the same thing for every image, DINO centers and sharpens the teacher's outputs.

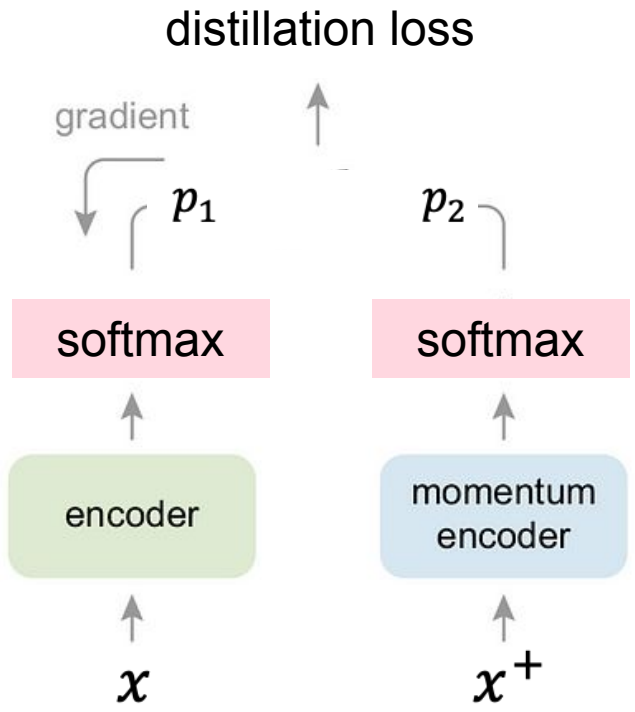
SimCLR



MoCo v3



DiNO



InfoNCE Loss

$$\min_f \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{pos},\{\mathbf{x}_i^-\}_{i=1}^N \sim p_{data}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^N e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right]$$

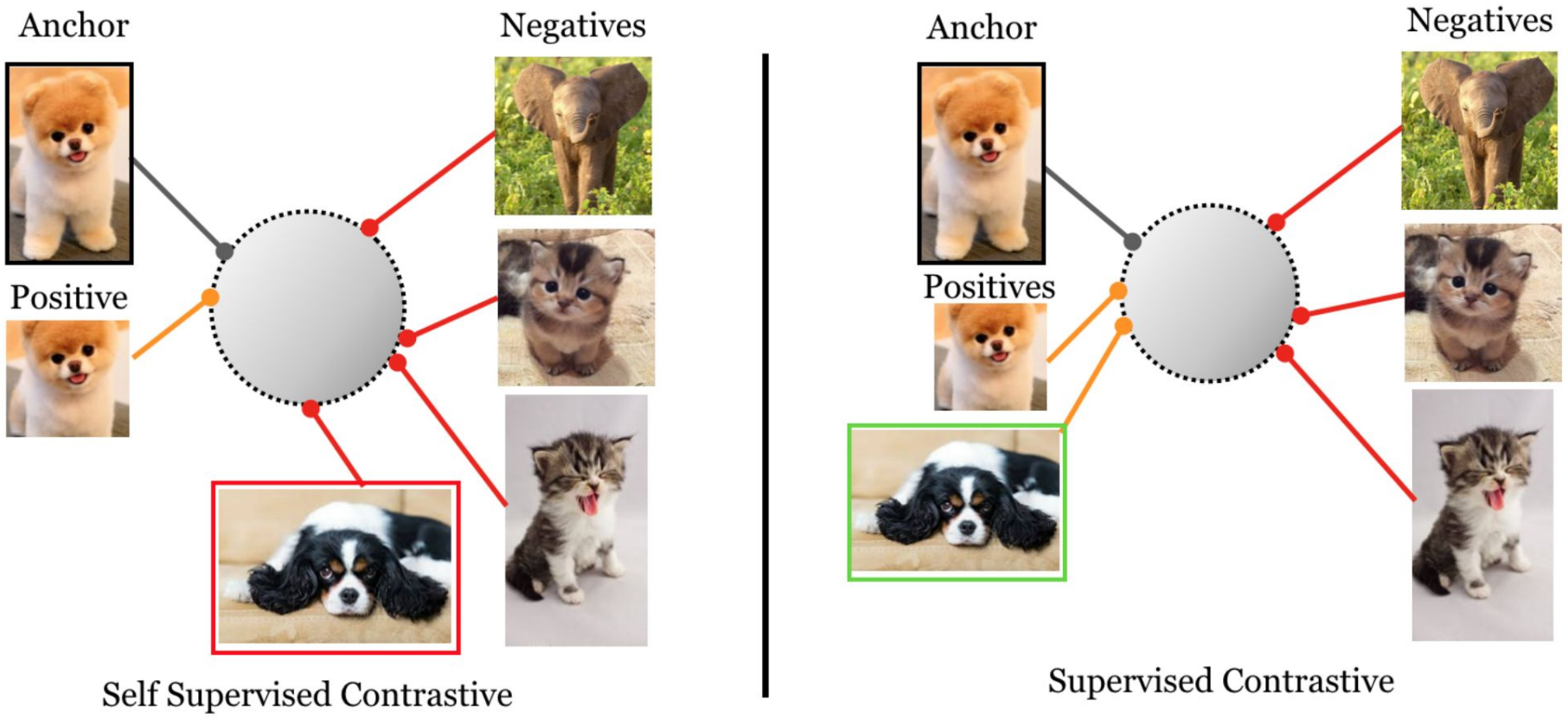
pull positive pair together

push negative pairs apart

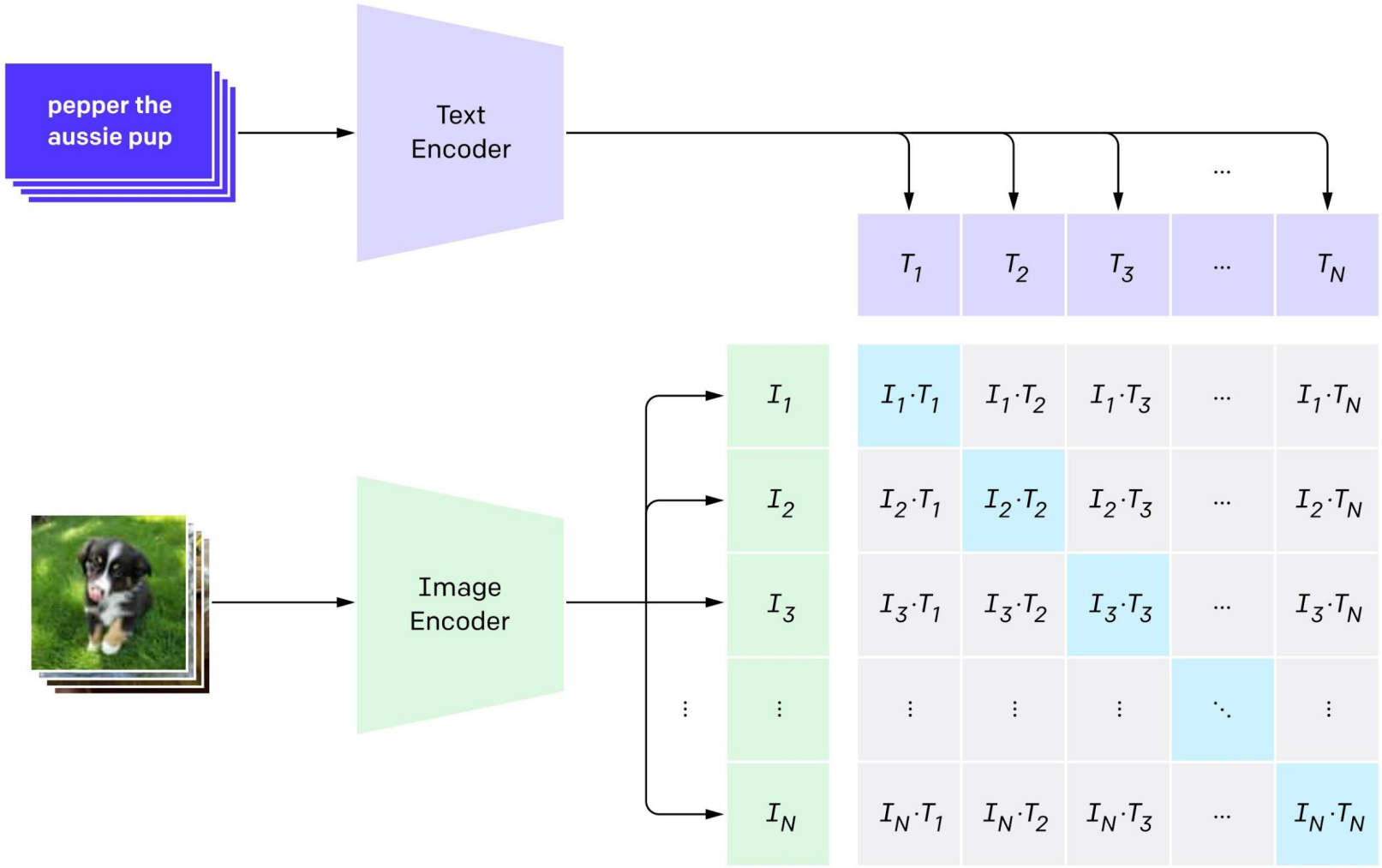
Model Distillation Loss

$$-p_2 \log p_1$$

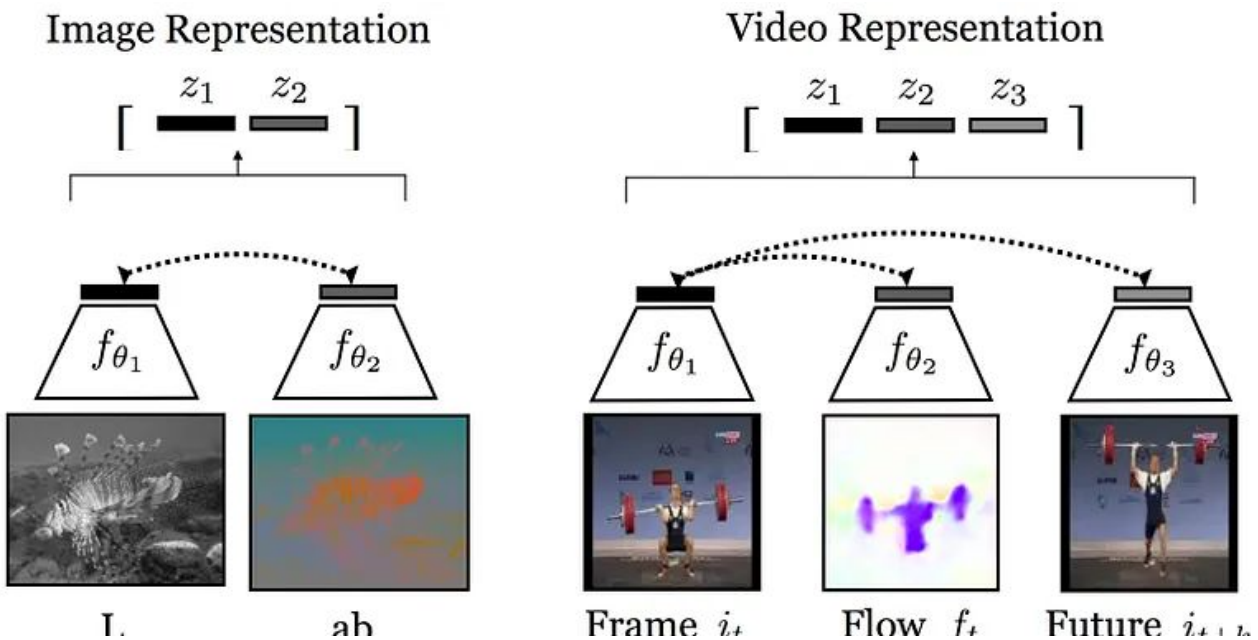
SupCon: Supervised Contrastive Learning



CLIP: **C**ontrastive **L**anguage **I**mage **P**retraining



CMC: **C**ontrastive **M**ultiview **C**oding



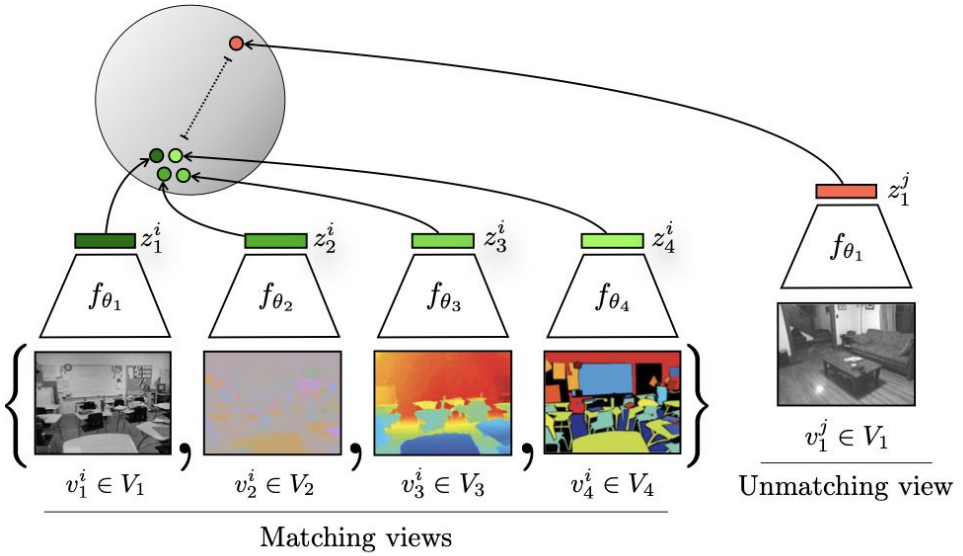


Figure 1: Given a set of sensory views, a deep representation is learnt by bringing views of the *same* scene together in embedding space, while pushing views of *different* scenes apart. Here we show and example of a 4-view dataset (NYU RGBD [53]) and its learned representation. The encodings for each view may be concatenated to form the full representation of a scene.

LGSimCLR: Language SimCLR

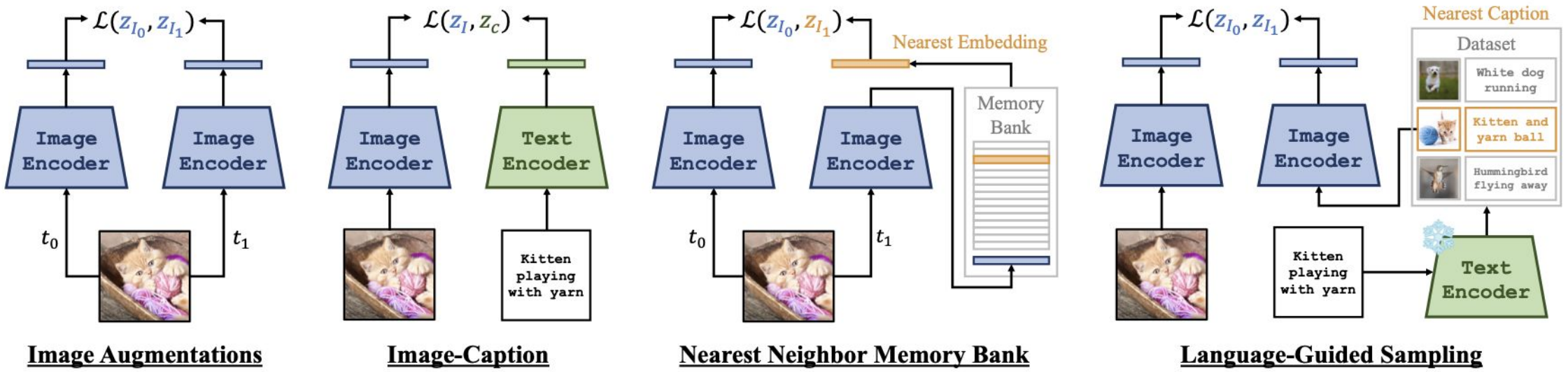


Figure 3. **Contrasting Contrastive Formulations**. While image-only and image-text contrastive learning directly extract views from the instance, nearest-neighbor methods rely on a memory bank of previously extracted features for training. In contrast, our approach samples nearest neighbors in caption embedding space using a pretrained language model and use the associated image for contrastive learning.

Stochastic Neighbor Embedding

Geoffrey Hin
Department of Compute
10 King's College Ro
{hinton,row

We describe a probabilistic app
scribed by high-dimensional ve
low-dimensional space in a wa
Gaussian is centered on each ol
the densities under this Gaussia
to define a probability distribu
the object. The aim of the emb
tion as well as possible when t
low-dimensional "images" of th
sum of Kullback-Leibler diverg
simple gradient for adjusting th
ages. Unlike other dimensional
framework makes it easy to repr

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on

...ts such as ImageNet (Deng et al., 2009). ...training methods which learn directly ...in a similar breakthrough in computer ...s encouraging.

...Mori et al. (1999) explored improving ...e retrieval by training a model to pre...dictives to text documents paired with ... al. (2007) demonstrated it was possi...ata efficient image representations via ...the weight space of classifiers trained ...captions associated with images. Sri...dinov (2012) explored deep represen...raining multimodal Deep Boltzmann ...low-level image and text tag features. ...odernized this line of work and demon...

PMLR 2020

A Simple Framework for Contrastive Learning of Visual Representations

...n [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]

...act

...R: a simple framework ...visual representations. ...posed contrastive self-...ths without requiring ...or a memory bank. In ...enables the contrastive ...useful representations, ...e major components of ...that (1) composition of ...critical role in defining ...(2) introducing a learn...tion between the repre...e loss substantially im...earned representations, ...ng benefits from larger ...ing steps compared to ...mbining these findings, ...y outperform previous ...d and semi-supervised ...linear classifier trained ...tations learned by Sim...1 accuracy, which is a ...over previous state-of-...rmance of a supervised ...ned on only 1% of the ...top-5 accuracy, outper...)× fewer labels. [1]

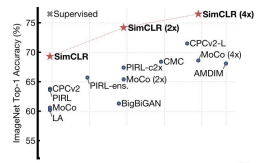


Figure ...on repr... ...ods (pr... ...ResNet

Normalized Cuts and Image Segmentation

Jianbo Shi and Jitendra Malik, *Member, IEEE*

Abstract—We propose a novel approach for solving the perceptual grouping problem in vision. Rather than focusing on local features and their consistencies in the image data, our approach aims at extracting the global impression of an image. We treat image segmentation as a graph partitioning problem and propose a novel global criterion, the *normalized cut*, for segmenting the graph. The *normalized cut* criterion measures both the total dissimilarity between the different groups as well as the total similarity within the groups. We show that an efficient computational technique based on a generalized eigenvalue problem can be used to optimize this criterion. We have applied this approach to segmenting static im... ...encouraging.

Index Terms—Grouping, image segmentation, graph partitioning

1 Introduction

...EARLY 75 years ago, Wertheimer [24] pointed out the importance of perceptual grouping and organization in vision and listed several key factors, such as similarity, proximity, and good continuation, which lead to visual

Contrastive Multiview Coding

Yonglong Tian
MIT CSAIL
yonglong@mit.edu

Dilip Krishnan
Google Research
dilipkay@google.com

Phillip Isola
MIT CSAIL
phillipi@mit.edu

Abstract

Humans view the world through many sensory channels, e.g., the long-wavelength light channel, viewed by the left eye, or the high-frequency vibrations channel, heard by the right ear. Each view is noisy and incomplete, but important factors, such as physics, geometry, and semantics, tend to be shared between all views (e.g., a "dog" can be seen, heard, and felt). We investigate the classic hypothesis that a powerful representation is one that models view-invariant factors. We study this hypothesis under the framework of multiview contrastive learning, where we learn a representation that aims to maximize mutual information between different views of the same scene but is otherwise compact. Our approach scales to any number of views, and is view-agnostic. We analyze key properties of the approach that make it work, finding that the contrastive loss outperforms

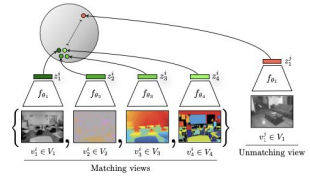


Figure 1: Given a set of sensory views, a deep representation is learnt by bringing views of the *same* scene together in embedding space, while pushing views of *different* scenes apart. Here we show and example of a 4-view dataset (NYU RGBD [53]) and its learned representation. The encodings for each view may be concatenated to form the full representation of a scene.

...conditions or thermal noise in a camera's sensor.

We revisit the classic hypothesis that the good bits are the

Published as a conference paper at ICLR 2022

Learning Visual Representations via Language-Guided Sampling

Mohamed El Banani Karan Desai Justin Johnson
University of Michigan
{mbanani,kdexd,justincj}@umich.edu

Abstract

...bject may appear in numerous contexts, we ...n a limited number of ways. Language al...ct away visual variation to represent and ...cepts. Building on this intuition, we pro...e approach to visual representation learn...ge similarity to sample semantically sim...for contrastive learning. Our approach ...age-based contrastive learning by sam...xing language similarity instead of hand...tions or learned clusters. Our approach ...image-text contrastive learning by relying ...guage models to guide the learning rather ...imizing a cross-modal loss. Through a se...ts, we show that language-guided learning ...res than image-based and image-text rep...ing approaches.

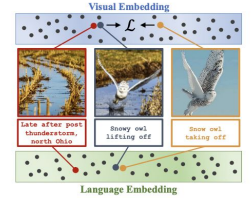


Figure 1. Language allows us to find conceptually similar image pairs even if they are visually dissimilar. We use those pairs for contrastive learning to learn generalizable visual features.

...ister structure and ...

CVPR 2023

Unsupervised visualization of image datasets learning

Harmonic Loss Trains Interpretable AI Models

David D. Baek [*1] Ziming Liu [*1] Riya Tyagi [1] Max Tegmark [1]

Abstract

In this paper, we introduce **harmonic loss** as an alternative to the standard cross-entropy loss for training neural networks and large language models (LLMs). Harmonic loss enables improved interpretability and faster convergence, owing to its scale invariance and finite convergence point by design, which can be interpreted as a class center. We first validate the performance of harmonic models across algorithmic, vision, and language datasets. Through extensive experiments, we demonstrate that models trained with harmonic loss outperform standard models by: (a) enhancing interpretability, (b) requiring less data for generalization, and (c) reducing grokking. Moreover, we compare a GPT-2 model trained with harmonic loss to the standard GPT-2, illustrating that the harmonic model develops more interpretable representations. Looking forward, we believe harmonic loss has the potential to become a valuable tool in domains with limited data availability or in high-stakes applications where interpretability and reliability are paramount, paving the way for more robust and efficient neural network models.

...comes to generalization:

(1) Lack of interpretability: Neural networks ...interpretability. This opacity limits our unders... ...how these models arrive at their conclusions, ...critical issue in high-stakes applications like hea... ...nance, and autonomous systems. While multipl... ...efforts have advanced our insight into the inner w... ...LLMs (Bereska & Gavves, 2024), we are still far ...explaining their outputs. Ultimately, we believe i... ...to design systems that are interpretable by desig... ...interpretability, it becomes challenging to diagn... ...ensure fairness, or build trust in a model's decisi...

(2) Low data efficiency: Generalization often re... ...and diverse training data. This raises a critical qu... ...models generalize effectively with less data? This ...pecially relevant in domains where data availabilit... ...such as rare disease diagnosis, low-resource lan... ...cessing, or specialized scientific fields like materi... ...or drug discovery. Previous approaches for impr... ...ral network generalization include efficient data... ...(Li et al., 2024a) and modifications to the training ...to accelerate training (Wang et al., 2024). How... ...methods focus on optimizing existing training p... ...rather than addressing the core issues in model d...

CVPR 2023 ... IC...

Your Contrastive Learning Is Secretly Doing Stochastic Neighbor Embedding

Tianyang Hu[1] Zhili Liu[1,2] Fengwei Zhou[1] Wenjia Wang[2,3] Weiran Huang[4]*

[1] Huawei Noah's Ark Lab [2] Hong Kong University of Science and Technology
[3] Hong Kong University of Science and Technology (Guangzhou)
[4] Qing Yuan Research Institute, Shanghai Jiao Tong University

Abstract

Contrastive learning, especially self-supervised contrastive learning (SSCL), ha... achieved great success in extracting powerful features from unlabeled data. In th... work, we contribute to the theoretical understanding of SSCL and uncover its conne... tion to the classic data visualization method, stochastic neighbor embedding (SNE) (Hinton & Roweis, 2002), whose goal is to preserve pairwise distances. From the pe... spective of preserving neighboring information, SSCL can be viewed as a special ca... of SNE with the input space pairwise similarities specified by data augmentation. Th... established correspondence facilitates deeper theoretical understanding of learne... features of SSCL, as well as methodological guidelines for practical improvemen... Specifically, through the lens of SNE, we provide novel analysis on domain-agnost... augmentations, implicit bias and robustness of learned features. To illustrate the pra... tical advantage, we demonstrate that the modifications from SNE to t-SNE (Van d... Maaten & Hinton, 2008) can also be adopted in the SSCL setting, achieving signi... icant improvement in both in-distribution and out-of-distribution generalization.

...hypotheses. The main ... ...tables, with the margin... ...of the rth order"). The ...order interaction." Sign... ...rth-order or higher-ord... ...order or higher-order i... ...Bartlett and by Roy a... ...rth-order restraints are ...order and higher-order ...if the restraints are con... ...generalizes a recent co... ...maximum entropy and ...between maximum entr...

ICLR 2023

...ATE OBSE...

...lacQUEEN
...University of California, Los A...

1. Introduction

The main purpose of this paper is to describe a pr... N-dimensional population into k sets on the basis of ... which is called 'k-means,' appears to give partitio... efficient in the sense of within-class variance. That is, if ... function for the population, $S = \{S_1, S_2, \cdots, S_k\}$ is ... $i = 1, 2, \cdots, k$, is the conditional mean of p over ... $\sum_{i=1}^{k} \int_{S_i} |x - u_i|^2 dp(x)$ tends to be low for the partit... method. We say 'tends to be low,' primarily because o... corroborated to some extent by mathematical analysi... tional experience. Also, the k-means procedure is es... computationally economical, so that it is feasible to p... on a digital computer. Possible applications include ... grouping, nonlinear prediction, approximating multiv... nonparametric tests for independence among several v...

In addition to suggesting practical classification meth... has proved to be theoretically interesting. The k-me... generalization of the ordinary sample mean, and one is naturally led to study the pertinent asymptotic behavior, the object being to establish some sort of law of large numbers for the k-means. This problem is sufficiently interesting, in fact, for us to devote a good portion of this paper to it. The k-means are defined in section 2.1, and the main results which have been obtained on the asymptotic behavior are given there. The rest of section 2 is devoted to the proofs of these results. Section 3 describes several specific possible applications, and reports some preliminary results from computer experiments conducted to explore the possibilities inherent in the k-means idea. The extension to general metric spaces is indicated briefly in section 4.

The original point of departure for the work described here was a series of

Clustering by Low-Rank Doubly Stochastic Matrix Decomposition

Zhirong Yang ZHIRONG.YANG@AALTO.FI
Department of Information and Computer Science, Aalto University, 00076, Finland

Erkki Oja ERKKI.OJA@AALTO.FI
Department of Information and Computer Science, Aalto University, 00076, Finland

Abstract

Clustering analysis by nonnegative low-rank approximations has achieved remarkable progress in the past decade. However, most approximation approaches in this direction are still restricted to matrix factorization. We propose a new low-rank learning method to improve the clustering performance, which is beyond matrix factorization. The approximation is based on a two-step bipartite random walk through virtual cluster nodes, where the approximation is formed by only cluster assigning probabilities. Minimizing the approximation error measured by Kullback-Leibler divergence is equivalent to maximizing the likelihood of a discriminative model, which endows our method with a solid probabilistic interpretation. The optimization is implemented by a relaxed Majorization-Minimization algorithm that is advantageous in finding good local minima. Furthermore, we point out that the regularized algorithm with Dirichlet prior only serves as initialization. Experimental results show that the new method has strong performance in clustering purity for various datasets, especially for large-scale manifold data.

negativity constraint, together with various low-rank matrix approximation objectives, has widely been used for the relaxation purpose in the past decade.

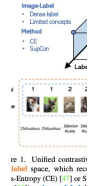
The most popular nonnegative low-rank approximation method is *Nonnegative Matrix Factorization* (NMF). It finds a matrix that approximates the similarities and can be factorized into several nonnegative low-rank matrices. NMF was originally applied to vectorial data, where Ding et al. (2010) have shown that NMF is equivalent to the classical k-means method. Later NMF was applied to the (weighted) graph given by the pairwise similarities. For example, Ding et al. (2008) presented Nonnegative Spectral Cuts by using a multiplicative algorithm; Arora et al. (2011) proposed Left Stochastic Decomposition that approximates a similarity matrix based on Euclidean distance and a left-stochastic matrix. Another stream in the same direction is topic modeling. Hofmann (1999) gave a generative model in *Probabilistic Latent Semantic Indexing* (PLSI) for counting data, which is essentially equivalent to NMF using Kullback-Leibler (KL) divergence and Tri-factorizations. Bayesian treatment of PLSI by using Dirichlet prior was later introduced by Blei et al. (2001). Symmetric PLSI with the same Bayesian treatment is called *Interaction Component Model* (ICM) (Sinkkonen et al., 2008).

Despite remarkable progress, the above relaxation approaches are still not fully satisfactory in all of the fol...

ECC...

...Image-Text-La...

...huan Zhang[1]* Bi...
...unfeng Gao[1]
...[2]Microsoft Cloud ...
...Luyuan, jfgac...



re 1. Unified contrasti... label space, which rec... s-Entropy (CE) [47] or S... [30]) on image-label d... ing (e.g., CLIP [48] or A...

...d with clean and larg... l data, *e.g.*, ImageNe... n decent visual recog... gories [23,35,50] an... ties [14, 33]. Neverth... l data can be a laboric... its difficulty to scale ... he other hand, languag... recently emerged as a promising approach by leveraging

2008 JMLR 2014

VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised L...

Adrien Bardes[1,2] **Jean Ponce**[2,4]
[1]Facebook AI Research
[2]Inria, École normale supérieure, CNRS, PSL Research University
[3]Courant Institute, New York University
[4]Center for Data Science, New York University

Abstract

Recent self-supervised methods for image representation learning ... agreement between embedding vectors produced by encoders fed ... views of the same image. The main challenge is to prevent a *col...* the encoders produce constant or non-informative vectors. We intr... (Variance-Invariance-Covariance Regularization), a method that e... the collapse problem with two regularizations terms applied to bo... separately: (1) a term that maintains the variance of each embedd... above a threshold, (2) a term that decorrelates each pair of vari... most other approaches to the same problem, VICReg does *not* requ... such as: weight sharing between the branches, batch normalization, feature-wise normalization, output quantization, stop gradient, memory banks, etc., and achieves results on par with the state of the art on several downstream tasks. In addition, we show that our variance regularization term stabilizes the training of other methods and leads to performance improvements.

CVPR 2022 ICLR 2022

Published as a conference paper at ICLR 2023

X-Sample Contrastive Loss: Improving Contrastive Learning with Sample Similarity Graphs

Vlad Sobal[1,2] **Mark Ibrahim**[1] **Randall Balestriero**[3] **Vivien Cabannes**[1]
Diane Bouchacourt[1] **Pietro Astolfi**[1] **Kyunghyun Cho**[2,4,5] **Yann LeCun**[1,2]
[1]Meta FAIR [2]New York Univeristy [3]Brown University [4]Genentech [5]CIFAR

Abstract

Learning good representations involves capturing the diverse ways in which data samples relate. Contrastive loss—an objective matching related samples—underlies methods from self-supervised to multimodal learning. Contrastive losses, however, can be viewed more broadly as modifying a similarity graph to indicate how samples should relate in the embedding space. This view reveals a shortcoming in contrastive learning: the similarity graph is binary, as only one sample is the related positive sample. Crucially, similarities *across* samples are ignored. Based on this observation, we revise the standard contrastive loss to explicitly encode how a sample relates to others. We experiment with this new objective, called X-Sample Contrastive, to train vision models based on similarities in class or text caption descriptions. Our study spans three scales: ImageNet-1k with 1 million, CC3M with 3 million, and CC12M with 12 million samples. The representations learned via our objective outperform both contrastive self-supervised and vision-language models trained on the same data across a range of tasks. When training on CC12M, we outperform CLIP by 0.6% on both ImageNet and ImageNet Real. Our objective appears to work particularly well in lower-data regimes, with gains over CLIP of 17.2% on ImageNet and 18.0% on ImageNet Real when training with CC3M. Finally, our objective seems to encourage the model to learn representations that separate objects from their attributes and backgrounds, with gains of 3.3-5.6% over CLIP on ImageNet9. We hope the proposed solution takes a small step towards developing richer learning objectives for understanding sample relations in foundation models.

ICLR 2025

...*the best pub...*
...*embeddings,*
...*different ver...*
...*ent networks)*
compatible to each other and allow for direct comparison between each other.

ICV 2023

...rXiv 2025

...of Training Self-Supervised Vision Transf...

...ei Chen* Saining Xie* Kaiming He
Facebook AI Research (FAIR)

...tps://github.com/facebookresearch/moco-v3

...*el method. Instead,*
...*ntal, yet must-know*
...*omputer vision (ViT). While*
...*volutional networks*
...*e recipes for ViT are*
...*upervised scenarios*
...*ing. In this work, we*
...*ffects of several fun-*
...*-supervised ViT. We*
...*that degrades accu-*
...*ly good results. We*
...*tial failure, and they*
...*de more stable. We*
...*d several other self-*
...*in various aspects.*
...*nce as well as chal-*
...*at this work will pro-*
...*for future research.*

framework	model
linear probing:	
iGPT [9]	iGPT-L
iGPT [9]	iGPT-XL
MoCo v3	ViT-B
MoCo v3	ViT-L
MoCo v3	ViT-H
MoCo v3	ViT-BN-H
MoCo v3	ViT-BN-L/7
end-to-end fine-tuning:	
masked patch pred. [16]	ViT-B
MoCo v3	ViT-B
MoCo v3	ViT-L

Table 1. **State-of-the-art Self-supervi...** ImageNet classification, evaluated by line... or end-to-end fine-tuning (bottom panel)... masked patch prediction [16] belong to the... paradigm. MoCo v3 is a contrastive learn... pares two (224×224) crops. ViT-B, -L, -H... formers proposed in [16]. ViT-BN is mod... and "/7" denotes a patch size of 7×7. [1]: pr...

...ion and P...
...g standing ...
...utput dista...
...nd a differ...
...lumination combinations. A d... identical, 4.0 corresponds to t... identities. You can see that a th... pair correctly.

CVPR 2021

faces of the same person have small distances and faces of distinct people have large distances.

...ed, then the afore-

CVPR 2015

A Unified Framework for Representation Learning

Stochastic Neig

A Simple Framework for Contrastive Learning of Visual Representations

Normalized Cuts and Image Segmentation

Jianbo Shi and Jitendra Malik, *Member, IEEE*

Published as a conference paper at ICLR 2023

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on crowd-labeled datasets such as ImageNet (Deng et al., 2009).

Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]

Supervised
SimCLR (4x)
SimCLR (2x)
SimCLR
CPCv2-L
MoCo (4x)
CMC
PIRL-c2x
MoCo (2x)
AMDIM
PIRL-ens.
CPCv2
PIRL
MoCo
LA
BigBiGAN
ImageNet Top-1 Accuracy (%)

Contrastive Multiview Coding

Yonglong Tian
MIT CSAIL
yonglong@mit.edu

Dilip Krishnan
Google Research
dilipkay@google.com

Phillip Isola
MIT CSAIL
phillipi@mit.edu

Abstract

Humans view the world through many sensory channels, e.g., the long-wavelength light channel, viewed by the left eye, or the high-frequency vibrations channel, heard by the right ear. Each view is noisy and incomplete, but important factors, such as physics, geometry, and semantics, tend to be shared between all views (e.g., a "dog" can be seen, heard, and felt). We investigate the classic hypothesis that a powerful representation is one that models view-invariant factors. We study this hypothesis under the framework of multiview contrastive learning, where we learn a representation that aims to maximize mutual information between different views of the same scene but is otherwise compact. Our approach scales to any number of views, and is view-agnostic. We analyze key properties of the approach that make it work, finding that the contrastive loss outperforms

Matching views
Unmatching view

Figure 1: Given a set of sensory views, a deep representation is learnt by bringing views of the *same* scene together in embedding space, while pushing views of *different* scenes apart. Here we show and example of a 4-view dataset (NYU RGBD [53]) and its learned representation. The encodings for each view may be concatenated to form the full representation of a scene.

conditions or thermal noise in a camera's sensor.

We revisit the classic hypothesis that the good bits are the

Learning Visual Representations via Language-Guided Sampling

Mohamed El Banani Karan Desai Justin Johnson
University of Michigan
{mbanani,kdexd,justincj}@umich.edu

Abstract

Visual Embedding
Language Embedding
Late after post thunderstorm, north Ohio
Snowy owl lifting off
Snow owl taking off

Figure 1. Language allows us to find conceptually similar image pairs even if they are visually dissimilar. We use those pairs for contrastive learning to learn generalizable visual features.

rich cluster structure and highl

UNSUPERVISED VISUALIZATION OF IMAGE DATASETS
E LEARNING

Harmonic Loss Trains Interpretable AI Models

David D. Baek [*1] Ziming Liu [*1] Riya Tyagi [1] Max Tegmark [1]

Abstract

In this paper, we introduce **harmonic loss** as an alternative to the standard cross-entropy loss for training neural networks and large language models (LLMs). Harmonic loss enables improved interpretability and faster convergence, owing to its scale invariance and finite convergence point by design, which can be interpreted as a class center. We first validate the performance of harmonic models across algorithmic, vision, and language datasets. Through extensive experiments, we demonstrate that models trained with harmonic loss outperform standard models by: (a) enhancing interpretability, (b) requiring less data for generalization, and (c) reducing grokking. Moreover, we compare a GPT-2 model trained with harmonic loss to the standard GPT-2, illustrating that the harmonic model develops more interpretable representations. Looking forward, we believe harmonic loss has the potential to become a valuable tool in domains with limited data availability or in high-stakes applications where interpretability and reliability are paramount, paving the way for more robust and efficient neural network models.

comes to generalization:

(1) Lack of interpretability: Neural networks

interpretability. This opacity limits our understa

how these models arrive at their conclusions, w

critical issue in high-stakes applications like heal

nance, and autonomous systems. While multiple

efforts have advanced our insight into the inner w

LLMs (Bereska & Gavves, 2024), we are still far

explaining their outputs. Ultimately, we believe it

to design systems that are interpretable by design

interpretability, it becomes challenging to diagnc

ensure fairness, or build trust in a model's decisic

(2) Low data efficiency: Generalization often req

and diverse training data. This raises a critical que

models generalize effectively with less data? This i

pecially relevant in domains where data availabilit

such as rare disease diagnosis, low-resource lang

cessing, or specialized scientific fields like materia

or drug discovery. Previous approaches for impro

ral network generalization include efficient data

(Li et al., 2024a) and modifications to the training

to accelerate training (Wang et al., 2024). How

methods focus on optimizing existing training pr

rather than addressing the core issues in model de

YOUR CONTRASTIVE LEARNING IS SECRETLY DOING STOCHASTIC NEIGHBOR EMBEDDING

Tianyang Hu [1] Zhili Liu [1,2] Fengwei Zhou [1] Wenjia Wang [2,3] Weiran Huang [4*]

[1] Huawei Noah's Ark Lab [2] Hong Kong University of Science and Technology
[3] Hong Kong University of Science and Technology (Guangzhou)
[4] Qing Yuan Research Institute, Shanghai Jiao Tong University

ABSTRACT

Contrastive learning, especially self-supervised contrastive learning (SSCL), ha achieved great success in extracting powerful features from unlabeled data. In thi work, we contribute to the theoretical understanding of SSCL and uncover its conne tion to the classic data visualization method, stochastic neighbor embedding (SNE (Hinton & Roweis, 2002), whose goal is to preserve pairwise distances. From the pe spective of preserving neighboring information, SSCL can be viewed as a special cas of SNE with the input space pairwise similarities specified by data augmentation. Th established correspondence facilitates deeper theoretical understanding of learne features of SSCL, as well as methodological guidelines for practical improvemen Specifically, through the lens of SNE, we provide novel analysis on domain-agnosti augmentations, implicit bias and robustness of learned features. To illustrate the pra tical advantage, we demonstrate that the modifications from SNE to t-SNE (Van de Maaten & Hinton, 2008) can also be adopted in the SSCL setting, achieving signi icant improvement in both in-distribution and out-of-distribution generalization.

Clustering by Low-Rank Doubly Stochastic Matrix Decomposition

Zhirong Yang ZHIRONG.YANG@AALTO.FI
Department of Information and Computer Science, Aalto University, 00076, Finland

Erkki Oja ERKKI.OJA@AALTO.FI
Department of Information and Computer Science, Aalto University, 00076, Finland

Abstract

Clustering analysis by nonnegative low-rank approximations has achieved remarkable progress in the past decade. However, most approximation approaches in this direction are still restricted to matrix factorization. We propose a new low-rank learning method to improve the clustering performance, which is beyond matrix factorization. The approximation is based on a two-step bipartite random walk through virtual cluster nodes, where the approximation is formed by only cluster assigning probabilities. Minimizing the approximation error measured by Kullback-Leibler divergence is equivalent to maximizing the likelihood of a discriminative model, which endows our method with a solid probabilistic interpretation. The optimization is implemented by a relaxed Majorization-Minimization algorithm that is advantageous in finding good local minima. Furthermore, we point out that the regularized algorithm with Dirichlet prior only serves as initialization. Experimental results show that the new method has strong performance in clustering purity for various datasets, especially for large-scale manifold data.

negativity constraint, together with various low-rank matrix approximation objectives, has widely been used for the relaxation purpose in the past decade.

The most popular nonnegative low-rank approximation method is *Nonnegative Matrix Factorization* (NMF). It finds a matrix that approximates the similarities and can be factorized into several nonnegative low-rank matrices. NMF was originally applied to vectorial data, where Ding et al. (2010) have shown that NMF is equivalent to the classical k-means method. Later NMF was applied to the (weighted) graph given by the pairwise similarities. For example, Ding et al. (2008) presented Nonnegative Spectral Cuts by using a multiplicative algorithm; Arora et al. (2011) proposed Left Stochastic Decomposition that approximates a similarity matrix based on Euclidean distance and a left-stochastic matrix. Another stream in the same direction is topic modeling. Hofmann (1999) gave a generative model in *Probabilistic Latent Semantic Indexing* (PLSI) for counting data, which is essentially equivalent to NMF using Kullback-Leibler (KL) divergence and Tri-factorizations. Bayesian treatment of PLSI by using Dirichlet prior was later introduced by Blei et al. (2001). Symmetric PLSI with the same Bayesian treatment is called *Interaction Component Model* (ICM) (Sinkkonen et al., 2008).

Despite remarkable progress, the above relaxation approaches are still not fully satisfactory in all of the fol-

CLASSIFICATION AND A
OF MULTIVARIATE OBSE

J. MacQUEEN
University of California, Los A

1. Introduction

The main purpose of this paper is to describe a pr
N-dimensional population into k sets on the basis of
which is called 'k-means,' appears to give partition
efficient in the sense of within-class variance. That is, if
function for the population, $S = \{S_1, S_2, \cdots, S_k\}$ is a
$i = 1, 2, \cdots, k$, is the conditional mean of p over t
$\sum_{i=1}^{k} \int_{S_i} |x - u_i|^2 dp(x)$ tends to be low for the partit
method. We say 'tends to be low,' primarily because of
corroborated to some extent by mathematical analysis
tional experience. Also, the k-means procedure is ea
computationally economical, so that it is feasible to pr
on a digital computer. Possible applications include
grouping, nonlinear prediction, approximating multiv
nonparametric tests for independence among several v

In addition to suggesting practical classification meth
has proved to be theoretically interesting. The k-me
generalization of the ordinary sample mean, and one is naturally led to study the
pertinent asymptotic behavior, the object being to establish some sort of law of
large numbers for the k-means. This problem is sufficiently interesting, in fact,
for us to devote a good portion of this paper to it. The k-means are defined in
section 2.1, and the main results which have been obtained on the asymptotic
behavior are given there. The rest of section 2 is devoted to the proofs of these
results. Section 3 describes several specific possible applications, and reports
some preliminary results from computer experiments conducted to explore the
possibilities inherent in the k-means idea. The extension to general metric spaces

Published as a conference paper at ICLR 2022

VICREG: VARIANCE-INVARIANCE-COVARIANCE REGULARIZATION FOR SELF-SUPERVISED LEARNING

Adrien Bardes [1,2] Jean Ponce [2,4] Yann LeCun [1,3,4]
[1]Facebook AI Research
[2]Inria, École normale supérieure, CNRS, PSL Research University
[3]Courant Institute, New York University
[4]Center for Data Science, New York University

ABSTRACT

Recent self-supervised methods for image representation learning maximize the agreement between embedding vectors produced by encoders fed with different views of the same image. The main challenge is to prevent a *collapse* in which the encoders produce constant or non-informative vectors. We introduce VICReg (Variance-Invariance-Covariance Regularization), a method that explicitly avoids the collapse problem with two regularizations terms applied to both embeddings separately: (1) a term that maintains the variance of each embedding dimension above a threshold, (2) a term that decorrelates each pair of variables. Unlike most other approaches to the same problem, VICReg does *not* require techniques such as: weight sharing between the branches, batch normalization, feature-wise normalization, output quantization, stop gradient, memory banks, etc., and achieves results on par with the state of the art on several downstream tasks. In addition, we show that our variance regularization term stabilizes the training of other methods and leads to performance improvements.

X-Sample Contrastive Loss: Improving Contrastive Learning with Sample Similarity Graphs

Mark Ibrahim [1] Randall Balestriero [3] Vivien Cabannes [1]
acourt [1] Pietro Astolfi [1] Kyunghyun Cho [2,4,5] Yann LeCun [1,2]
[2]New York Univeristy [3]Brown University [4]Genentech [5]CIFAR

Abstract

od representations involves capturing the diverse ways in which data te. Contrastive loss—an objective matching related samples—underlies m self-supervised to multimodal learning. Contrastive losses, however, ed more broadly as modifying a similarity graph to indicate how uld relate in the embedding space. This view reveals a shortcoming e learning: the similarity graph is binary, as only one sample is the ive sample. Crucially, similarities *across* samples are ignored. Based rvation, we revise the standard contrastive loss to explicitly encode how ates to others. We experiment with this new objective, called X-Sample to train vision models based on similarities in class or text caption . Our study spans three scales: ImageNet-1k with 1 million, CC3M on, and CC12M with 12 million samples. The representations learned ctive outperform both contrastive self-supervised and vision-language ed on the same data across a range of tasks. When training on CC12M, m CLIP by 0.6% on both ImageNet and ImageNet Real. Our objective work particularly well in lower-data regimes, with gains over CLIP ImageNet and 18.0% on ImageNet Real when training with CC3M. objective seems to encourage the model to learn representations that jects from their attributes and backgrounds, with gains of 3.3-5.6% on ImageNet9. We hope the proposed solution takes a small step eloping richer learning objectives for understanding sample relations n models.

curacy over k-means on 17 and CIFAR100, respectively. ve a clustering accuracy of ithub.com/HHU-MMBS/

We also introduce the concept of harmonic *embeddings, and a* harmonic *triplet loss, which describe different versions of face embeddings (produced by different networks) that are compatible to each other and allow for direct comparison between each other.*

cognition and Clustering

study of Training Self-Supervised Vision Transf

Xinlei Chen* Saining Xie* Kaiming He
Facebook AI Research (FAIR)

de: https://github.com/facebookresearch/moco-v3

framework	model
linear probing:	
iGPT [9]	iGPT-L
iGPT [9]	iGPT-XL
MoCo v3	ViT-B
MoCo v3	ViT-L
MoCo v3	ViT-H
MoCo v3	ViT-BN-H
MoCo v3	ViT-BN-L/7
end-to-end fine-tuning:	
masked patch pred. [16]	ViT-B
MoCo v3	ViT-B
MoCo v3	ViT-L

Table 1. **State-of-the-art Self-supervi**
ImageNet classification, evaluated by line
or end-to-end fine-tuning (bottom panel).
masked patch prediction [16] belong to the
paradigm. MoCo v3 is a contrastive learn
pares two (224×224) crops. ViT-B, -L, -H
formers proposed in [16]. ViT-BN is mod
and "/7" denotes a patch size of 7×7. †: pr

faces of the same person have small distances and faces of distinct people have large distances.

Once this embedding has been produced, then the afore-

A Unified Framework for Representation Learning

- **I-Con**: A single information-theoretic loss that unifies several major classes of representation learning.
- **Periodic Table of Methods**: I-Con enables a structured organization of diverse learning approaches.
- **Filling the Gaps**: I-Con helps derive a new method that boosts unsupervised ImageNet-1K accuracy by **+8%**.

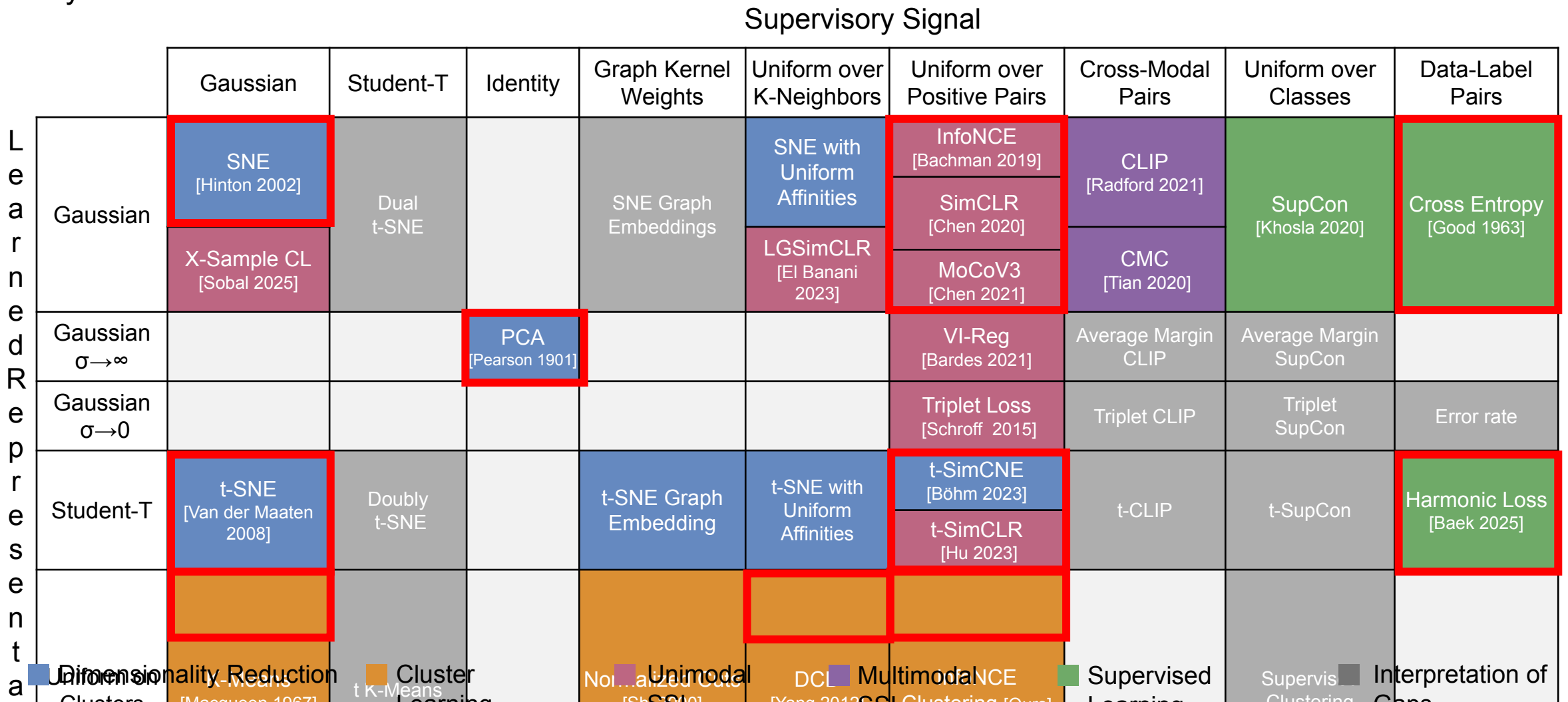
Supervisory Signal

Learned Representation	Gaussian	Student-T	Identity	Graph Kernel Weights	Uniform over K-Neighbors	Uniform over Positive Pairs	Cross-Modal Pairs	Uniform over Classes	Data-Label Pairs
Gaussian	SNE [Hinton 2002] X-Sample CL [Sobal 2025]	Dual t-SNE		SNE Graph Embeddings	SNE with Uniform Affinities LGSimCLR [El Banani 2023]	InfoNCE [Bachman 2019] SimCLR [Chen 2020] MoCoV3 [Chen 2021]	CLIP [Radford 2021] CMC [Tian 2020]	SupCon [Khosla 2020]	Cross Entropy [Good 1963]
Gaussian σ→∞			PCA [Pearson 1901]			VI-Reg [Bardes 2021]	Average Margin CLIP	Average Margin SupCon	
Gaussian σ→0						Triplet Loss [Schroff 2015]	Triplet CLIP	Triplet SupCon	Error rate
Student-T	t-SNE [Van der Maaten 2008]	Doubly t-SNE		t-SNE Graph Embedding	t-SNE with Uniform Affinities	t-SimCNE [Böhm 2023] t-SimCLR [Hu 2023]	t-CLIP	t-SupCon	Harmonic Loss [Baek 2025]
Uniform on Clusters	K-Means [Macqueen 1967]	t K-Means		Normalized Cuts [Shi 2000]	DCL [Yang 2023]	SogCLR [Qiu 2023]		Supervised Clustering	

Dimensionality Reduction · Cluster Learning · Unimodal SSL · Multimodal SSL · Supervised Learning · Interpretation of Gaps

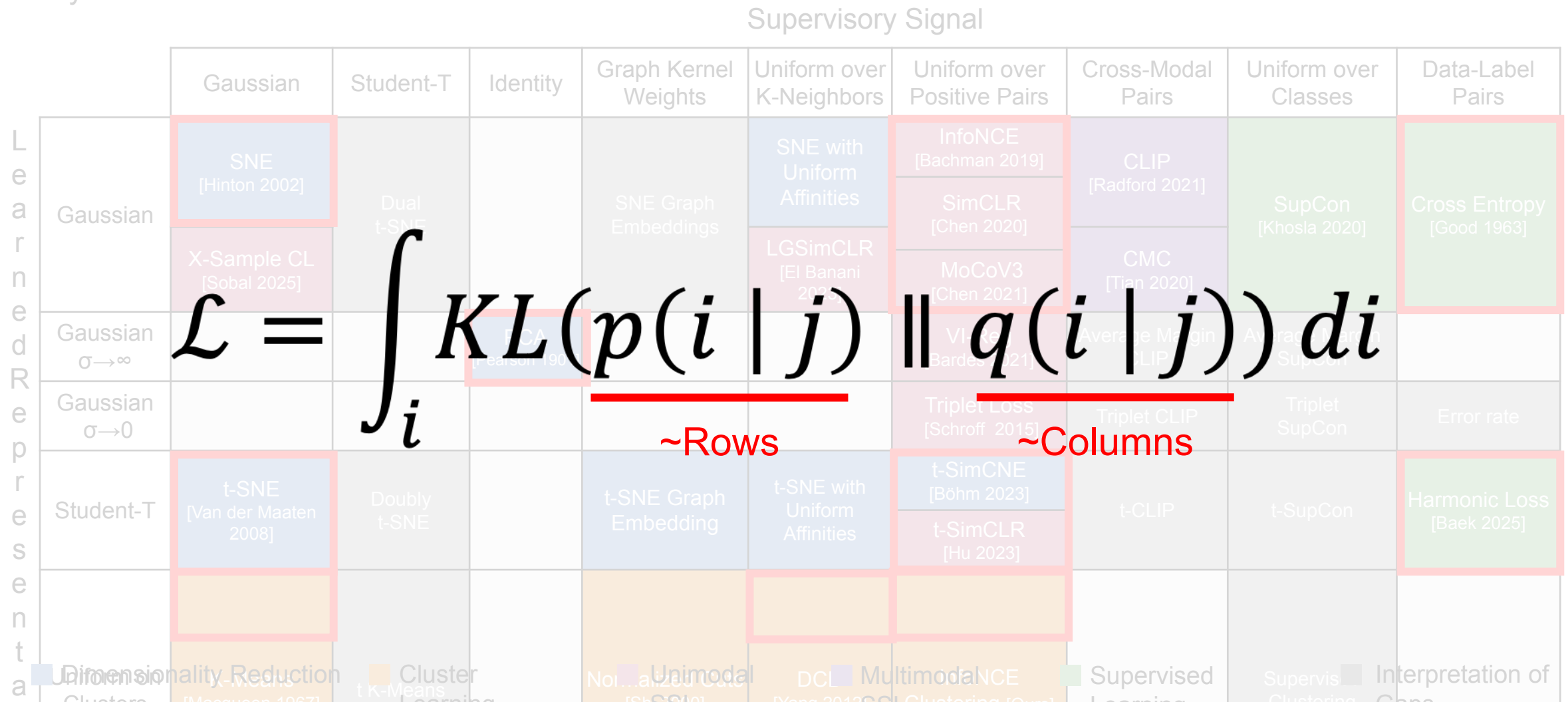
A Unified Framework for Representation Learning

- **I-Con**: A single information-theoretic loss that unifies several major classes of representation learning.
- **Periodic Table of Methods**: I-Con enables a structured organization of diverse learning approaches.
- **Filling the Gaps**: I-Con helps derive a new method that boosts unsupervised ImageNet-1K accuracy by **+8%**.



A Unified Framework for Representation Learning

- **I-Con**: A single information-theoretic loss that unifies several major classes of representation learning.
- **Periodic Table of Methods**: I-Con enables a structured organization of diverse learning approaches.
- **Filling the Gaps**: I-Con helps derive a new method that boosts unsupervised ImageNet-1K accuracy by **+8%**.



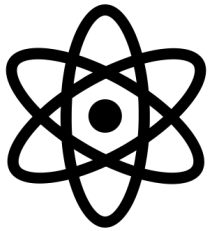
$$\mathcal{L} = \int_i KL(p(i \mid j) \,\|\, q(i \mid j))\, di$$

~Rows

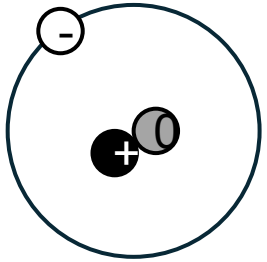
~Columns

Chemical Periodic Table

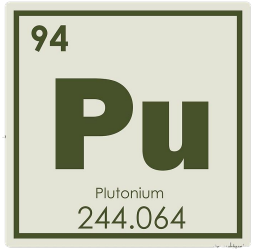
Element



Protons and Neutrons

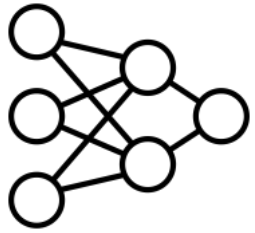


Gaps in the Periodic Table

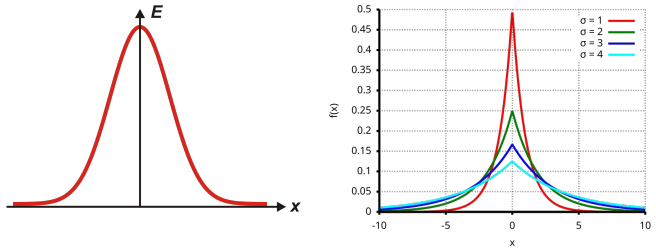


Machine Learning Periodic Table

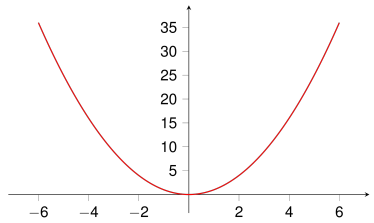
Algorithm



Distributions



New Loss Functions



Overview

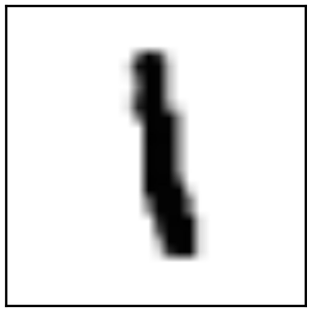
- Intro
  - What's Representation Learning?
  - The Periodic Table
- Methods:
  - **Exploring some Examples**
  - Generalizing ML Methods with a single Equation
  - Building the Periodic Table
- Experiments
  - Building an Image Classifier that doesn't need human labels
- Future Directions

Example 1: Dimensionality Reduction

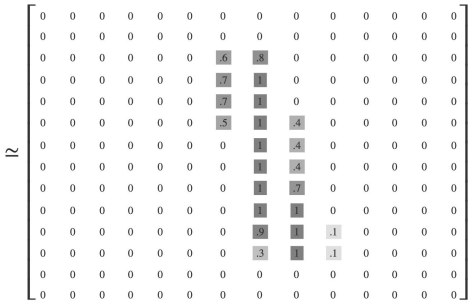
High Dimensional Data

RGB Images

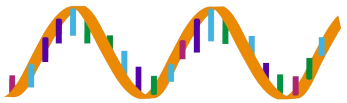
x_j



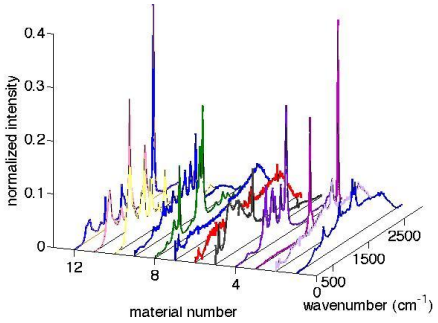
x_i



mRNA-seq



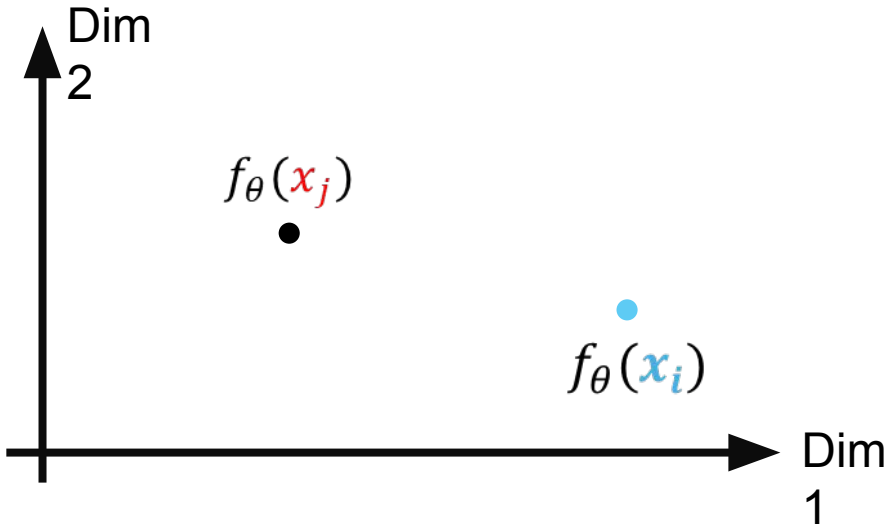
Spectroscopy Data



Mapper/ LookUp

f_θ

Low Dimensional Data



PCA

MDS

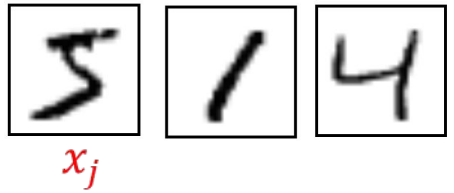
tSNE

Example 1: Dimensionality Reductio

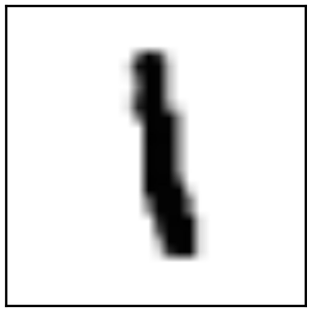
PCA

High Dimensional Data

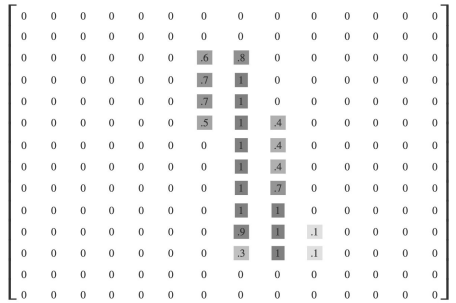
RGB Images



x_j



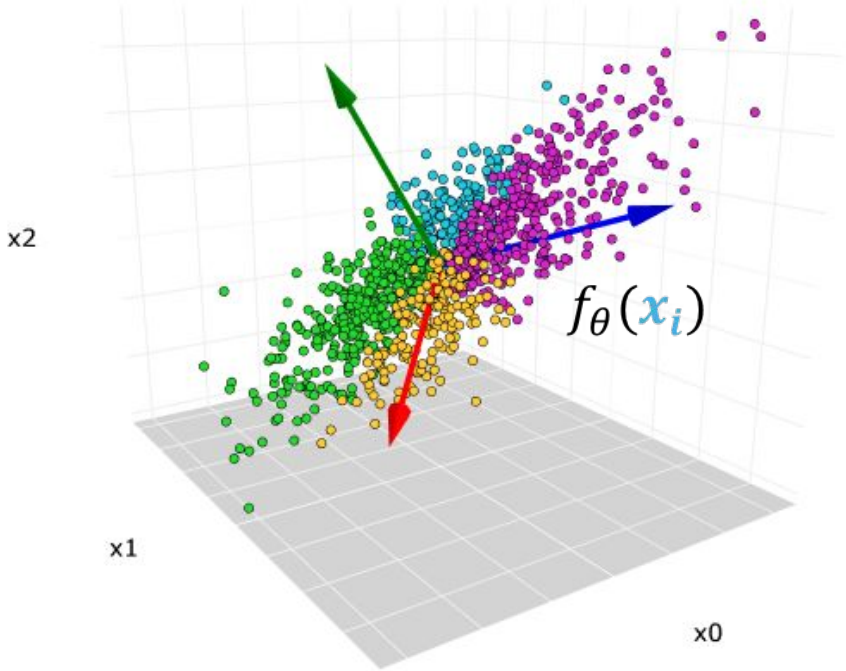
x_i



Projection Map

f_θ

Low Dimensional Data



$$\mathcal{L} = -Var(f_\theta(X)) = -\sum_i \left(f_\theta(x_i) - f_\theta(x_j)\right)^2$$

where f_θ is a linear projection map

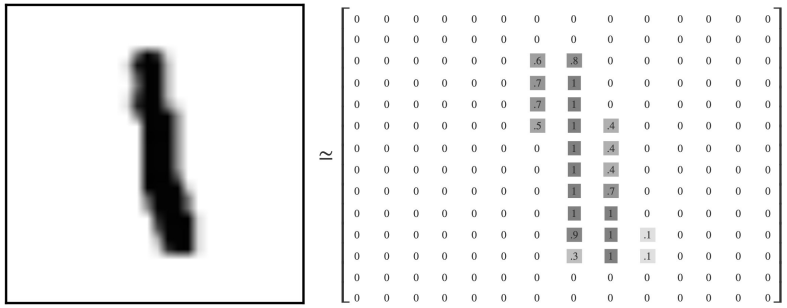
Example 1: Dimensionality Reductio

PCA

High Dimensional Data

RGB Images

x_j

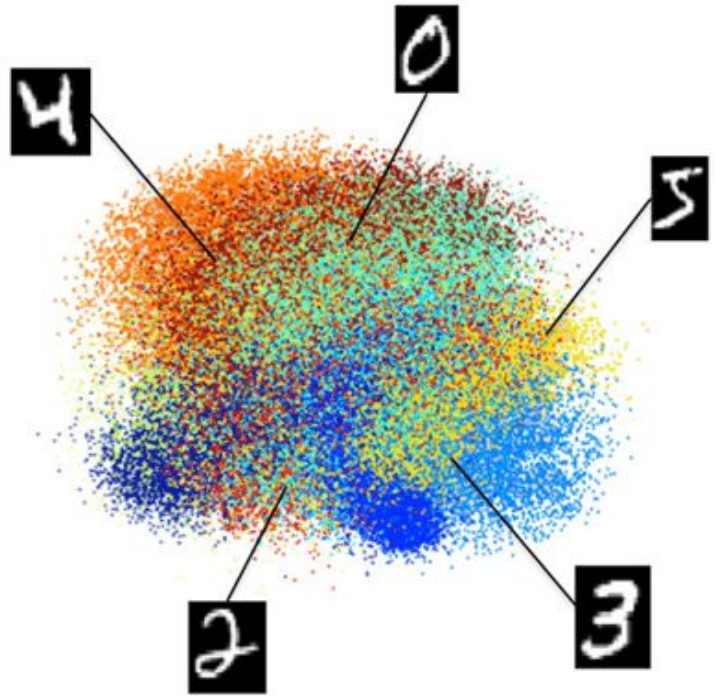


x_i

Projection Map

f_θ

Low Dimensional Data



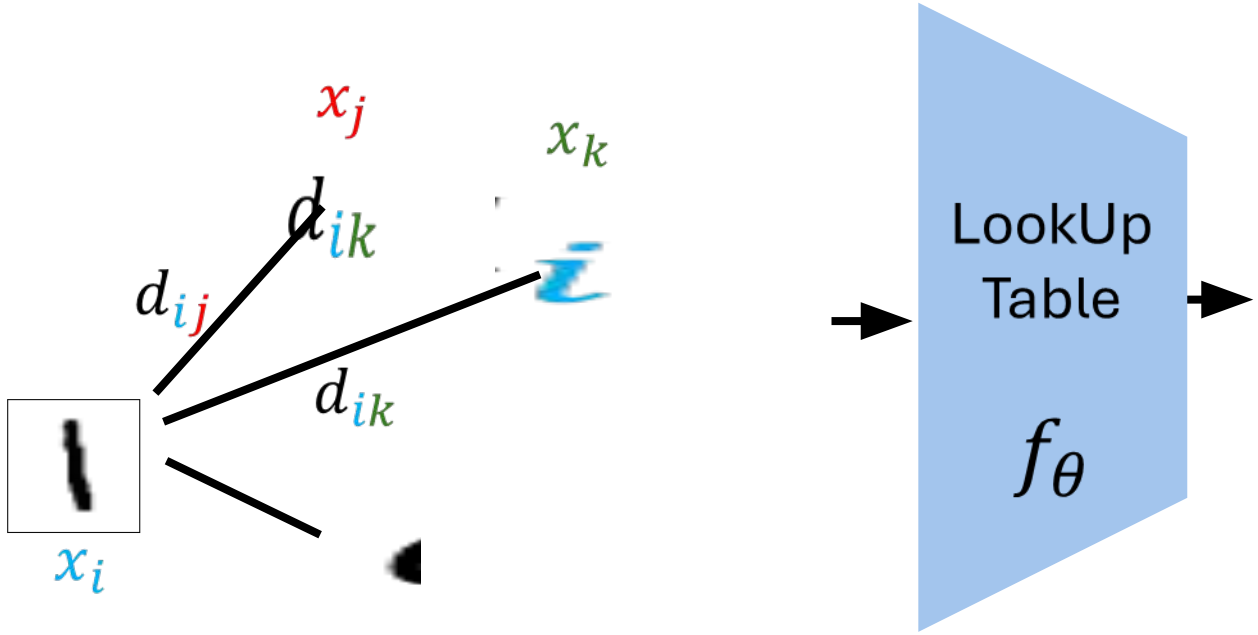
$$\mathcal{L} = -Var(f_\theta(X)) = -\sum_i \left(f_\theta(x_i) - f_\theta(x_j)\right)^2$$

where f_θ is a linear projection map

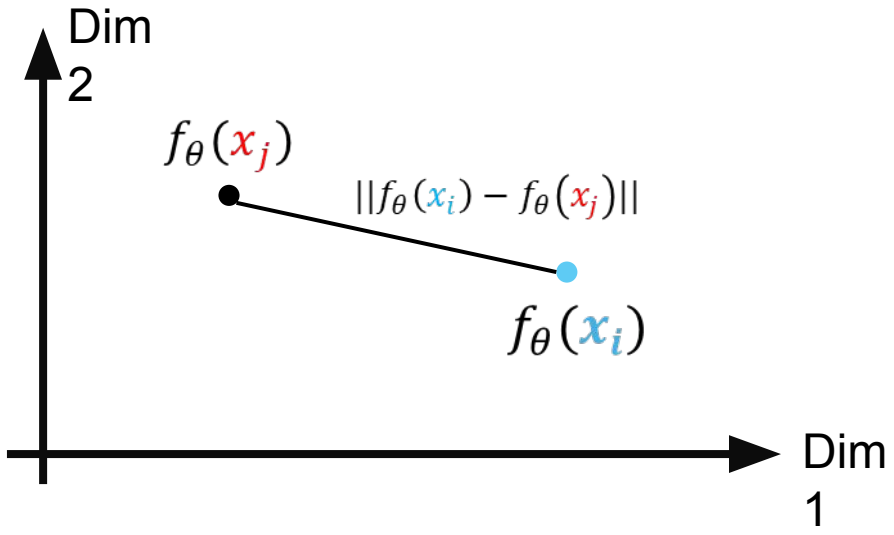
Example 1: Dimensionality Reductio

MDS

High Dimensional Data



Low Dimensional Data



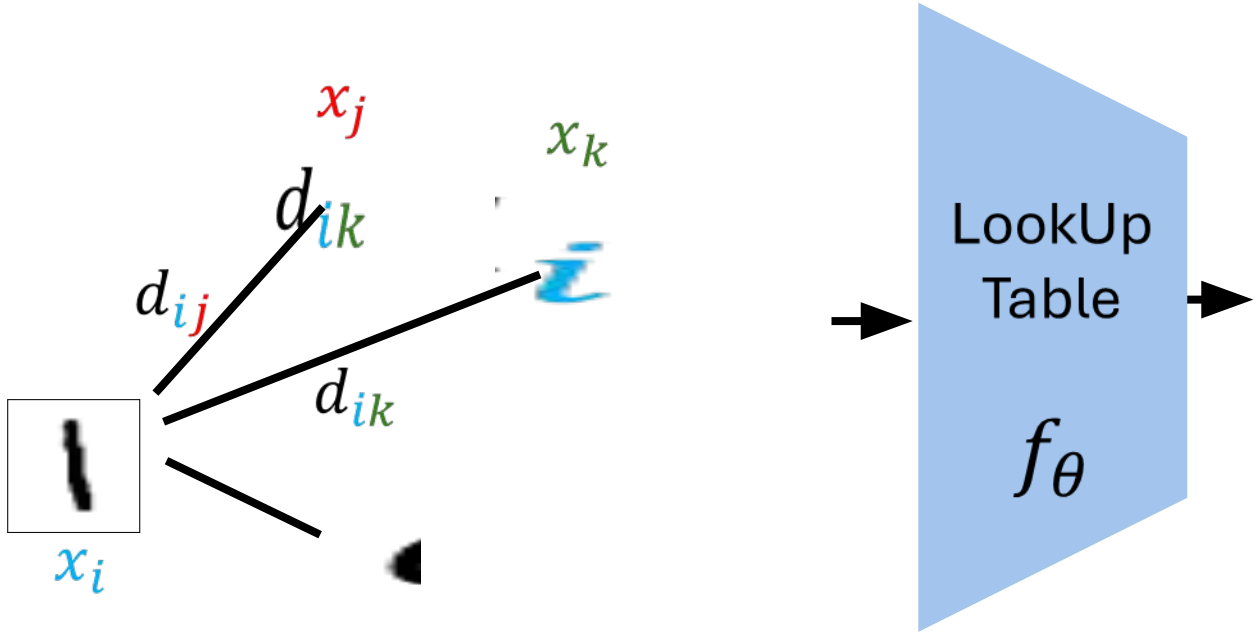
$$\mathcal{L} = \sum_i \left(||f_\theta(x_i) - f_\theta(x_j)|| - d_{ij} \right)^2$$

where f_θ is a lookup table

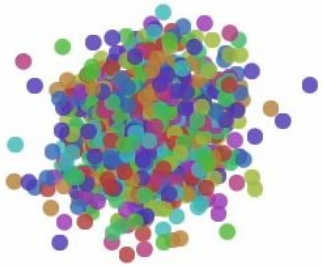
Example 1: Dimensionality Reductio

MDS

High Dimensional Data



Low Dimensional Data



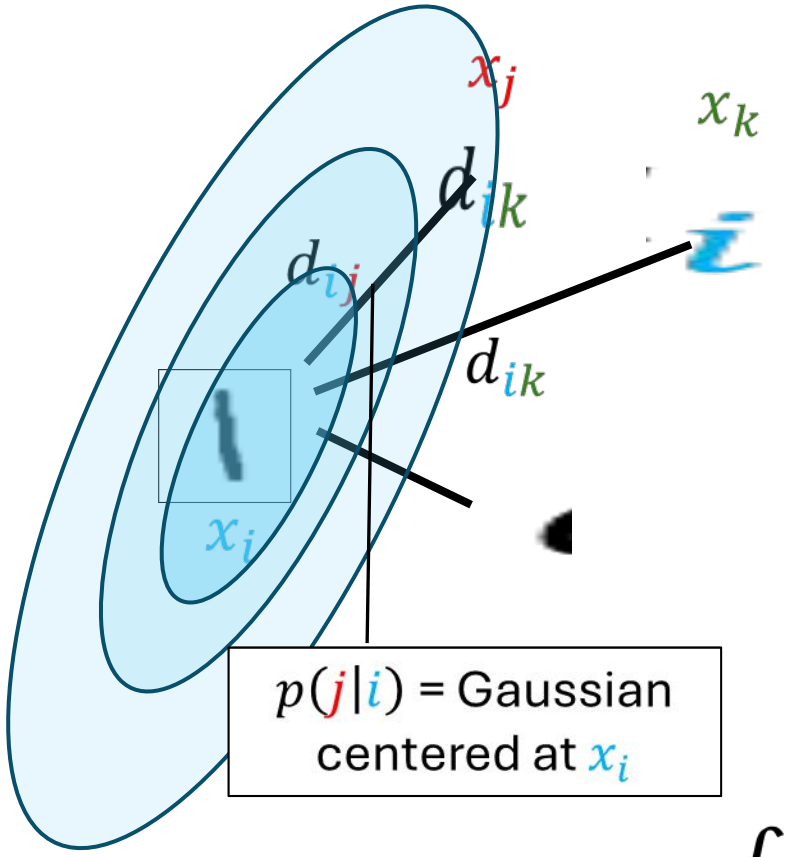
$$\mathcal{L} = \sum_i \left(||f_\theta(x_i) - f_\theta(x_j)|| - d_{ij} \right)^2$$

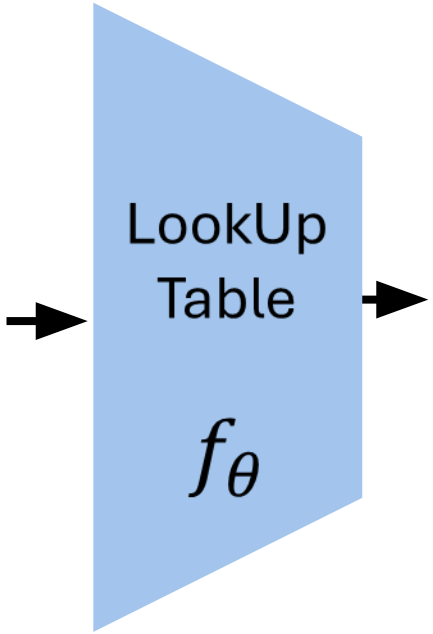
where f_θ is a lookup table

Example 1: Dimensionality Reductic

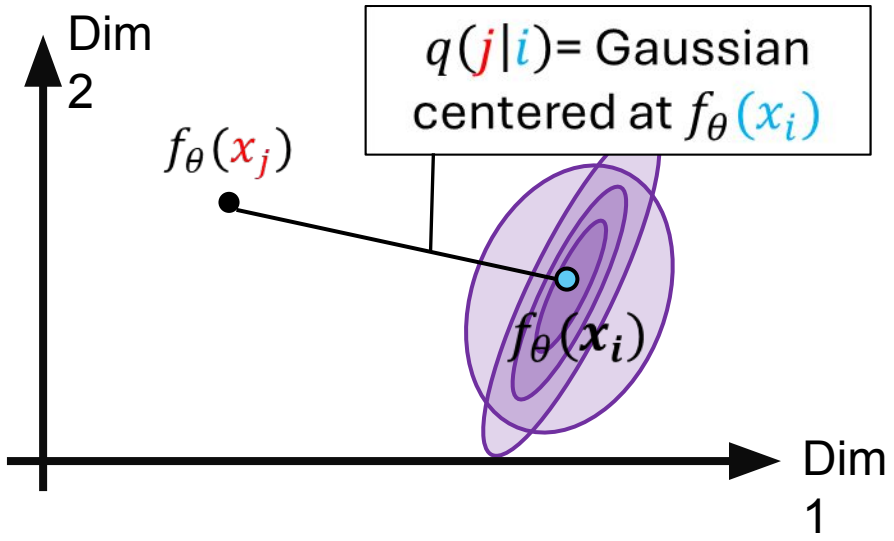
SNE

High Dimensional Data





Low Dimensional Data

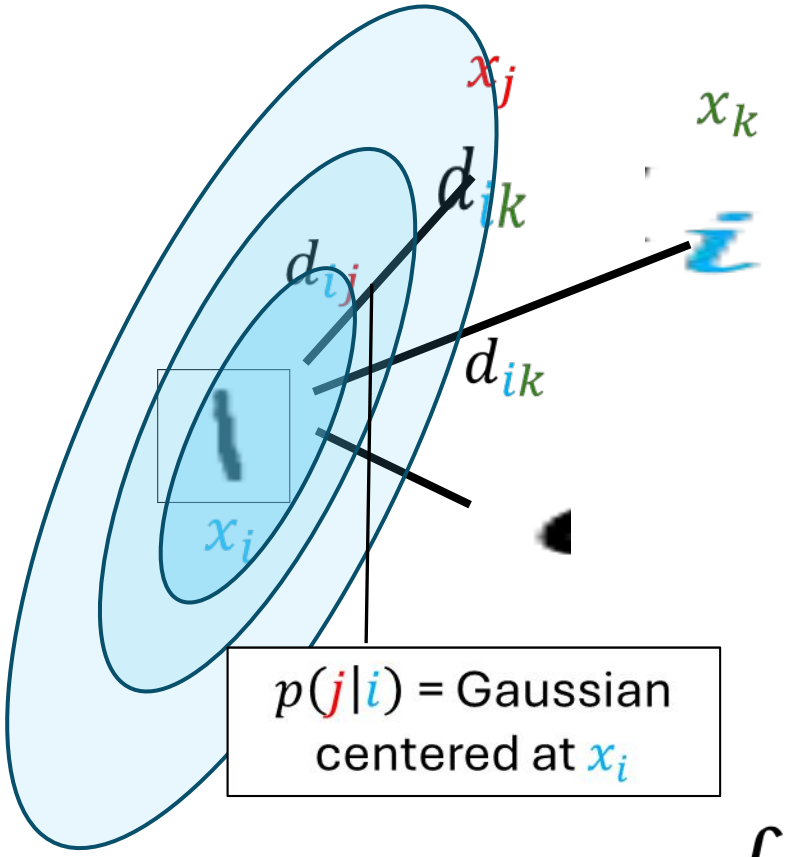


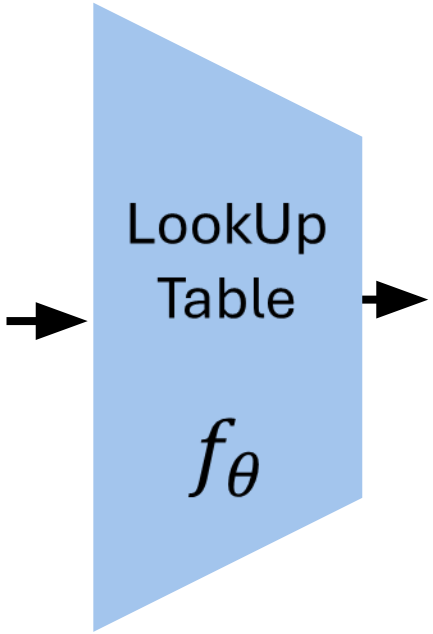
$$\mathcal{L} = \sum_i D_{KL}(p(*|i) \parallel q(*|i))$$

Example 1: Dimensionality Reductio

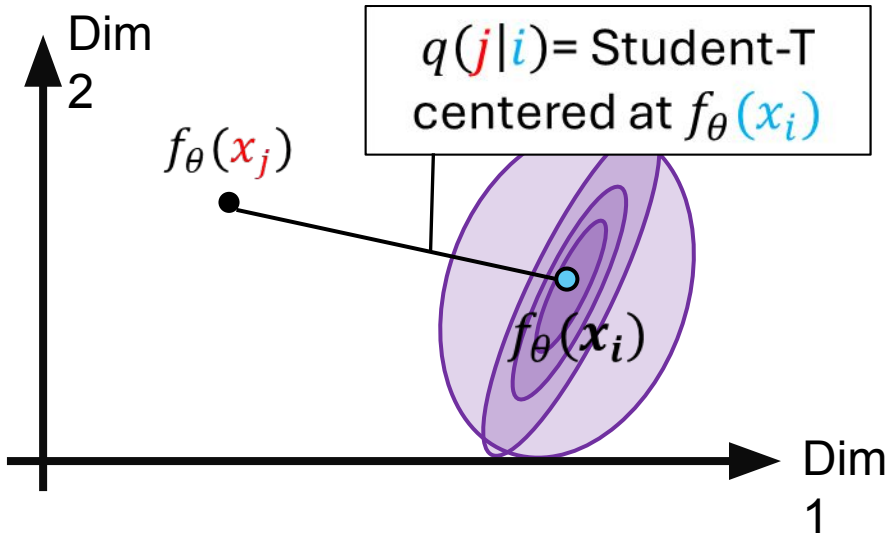
t-SNE

High Dimensional Data





Low Dimensional Data

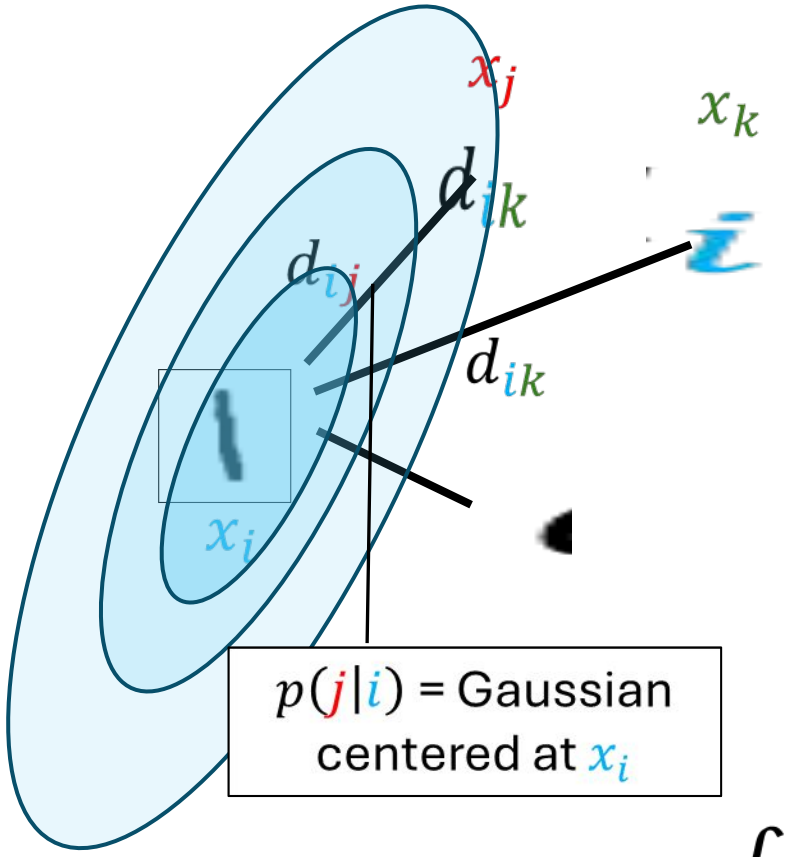


$$\mathcal{L} = \sum_i D_{KL}(p(*|i) \parallel q(*|i))$$

Example 1: Dimensionality Reductio

t-SNE

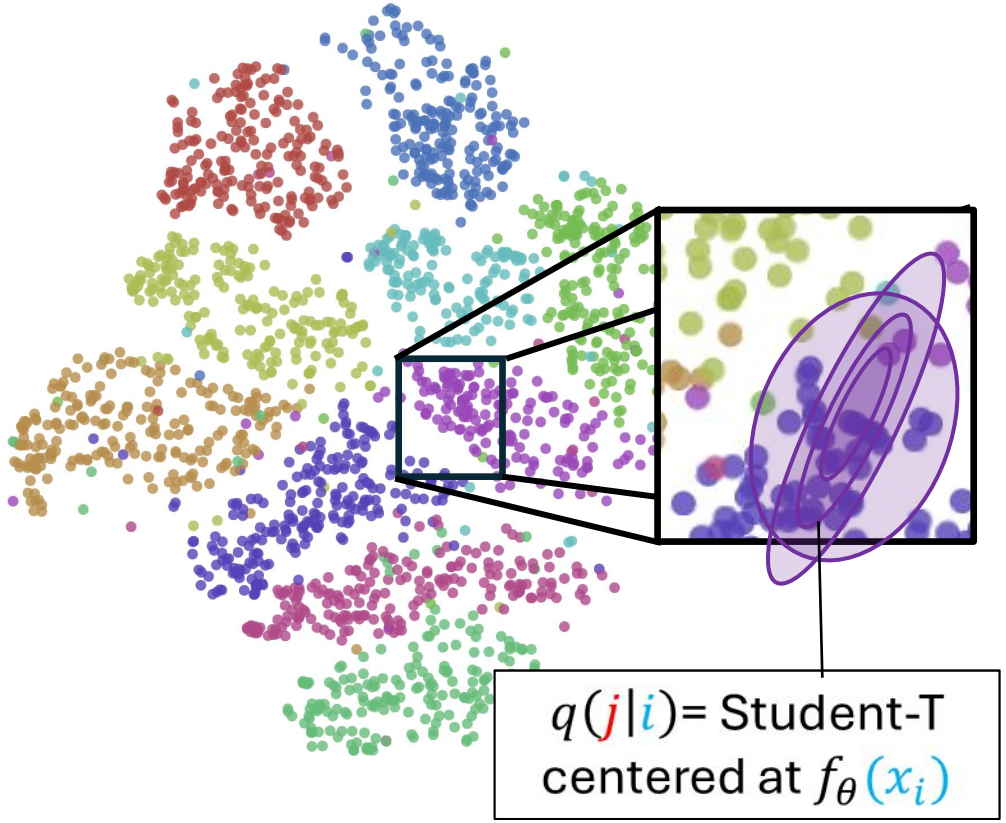
High Dimensional Data



LookUp Table

f_θ

Low Dimensional Data



$$\mathcal{L} = \sum_i D_{KL}(p(*|i) \parallel q(*|i))$$

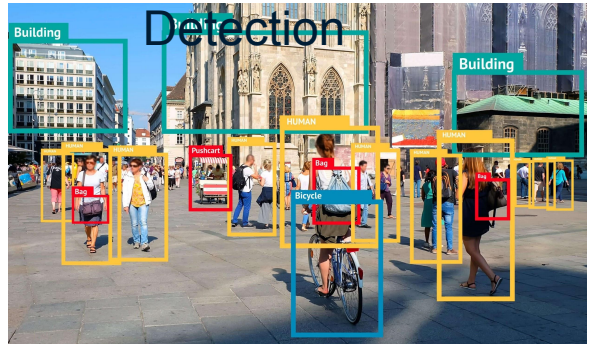
Example 2: Contrastive Learning

SimCLR

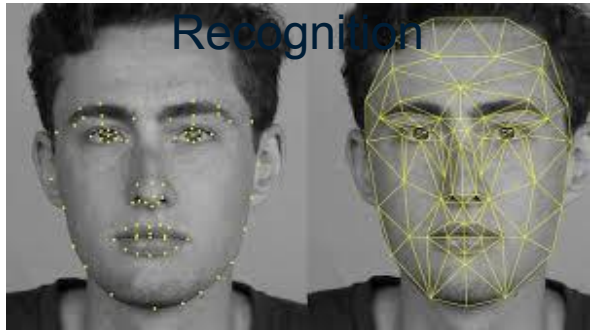
Downstream Tasks

Segmentation

Object Detection



Face Recognition



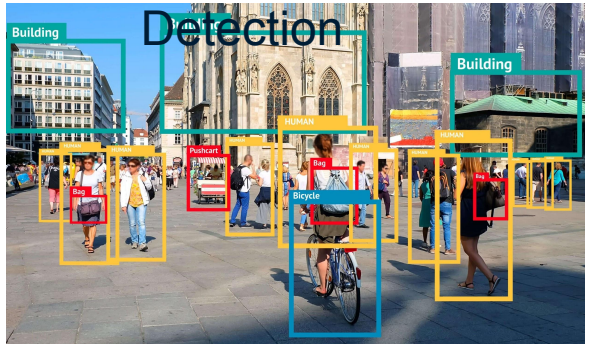
Example 2: Contrastive Learning

SimCLR

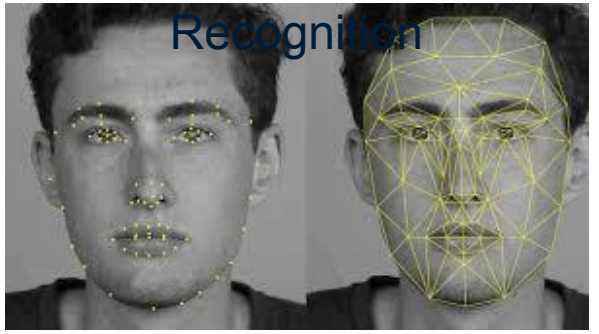
Downstream Tasks

Segmentation

Object Detection



Face Recognition



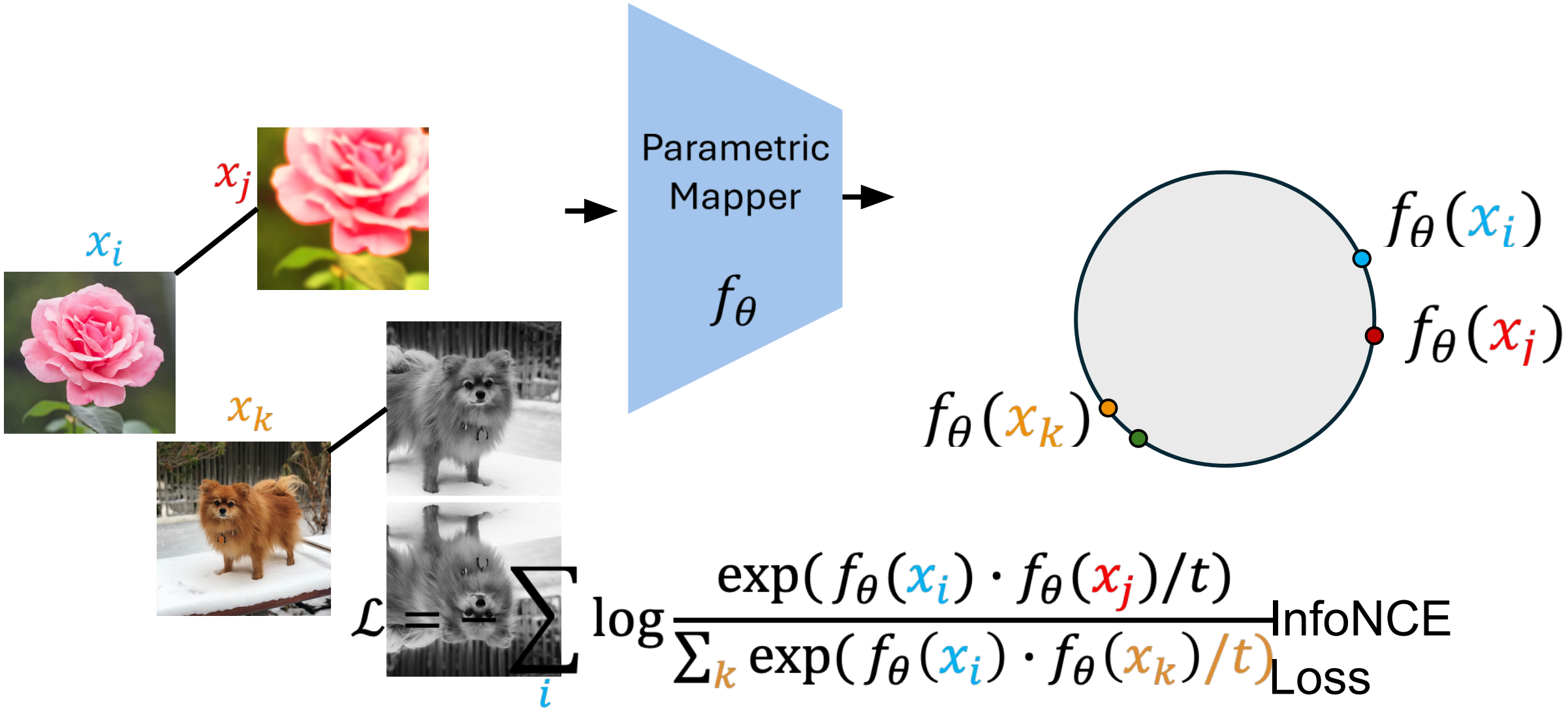
. . .

Example 2: Contrastive Learning

SimCLR

RGB Images

Embeddings

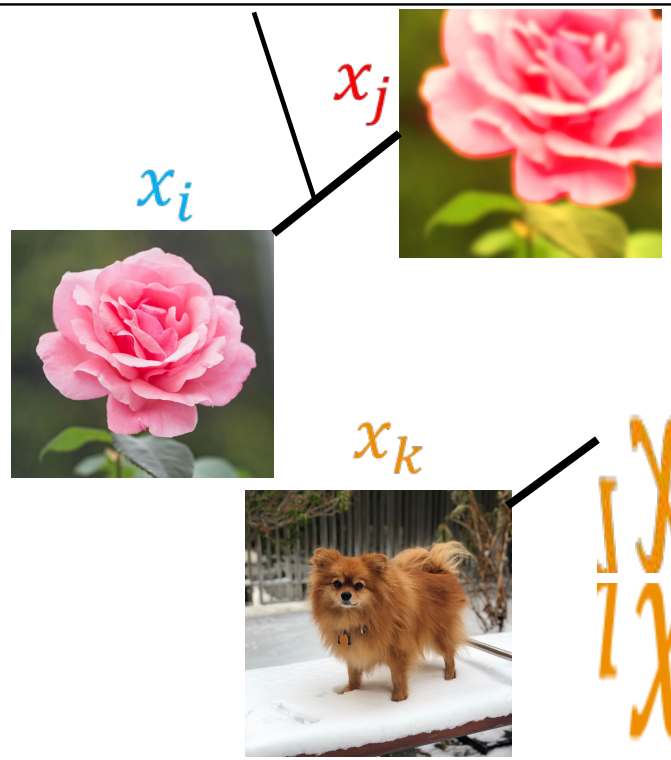


Example 2: Contrastive Learning

SimCLR

RGB Images

$p(j|i)) = \mathbb{1}[\ i$ and j are augmentations of the same image]

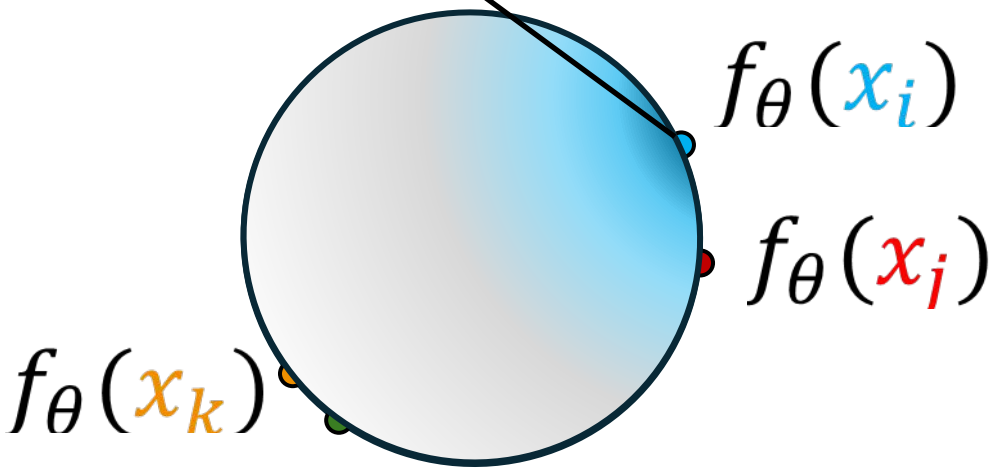


Parametric Mapper

f_θ

Embeddings

$q(j|i)$= Gaussian centered at $f_\theta(x_i)$

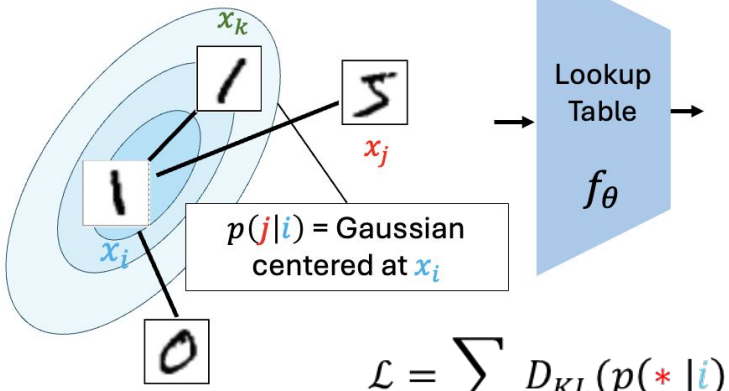


$$\mathcal{L} = \sum_i D_{KL}(p(*|i) \parallel q(*|i))$$

Strong Common Pattern

SNE

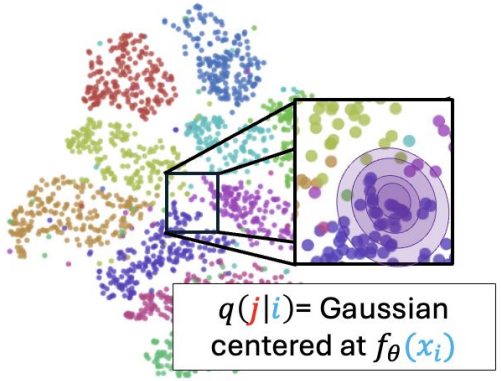
High Dimensional Data



Lookup Table

f_θ

Low Dimensional Data

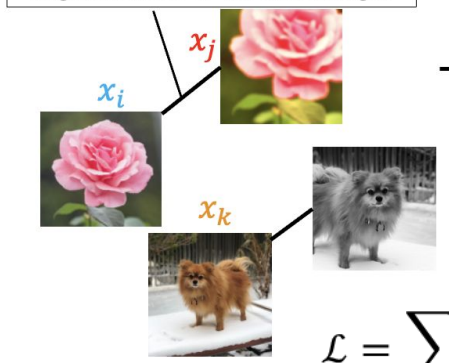


$$\mathcal{L} = \sum D_{KL}(p(*|i) \parallel q(*|i))$$

SimCLR

RGB Images

$p(j|i)) = \mathbb{1}[\ i$ and j are augmentations of the same image]

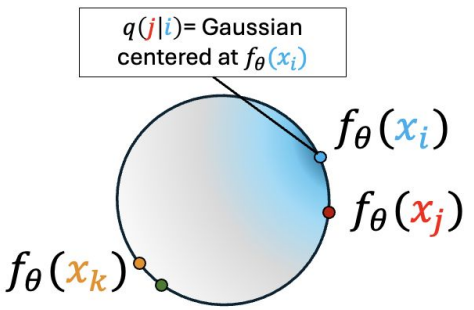


Parametric Mapper

f_θ

Embeddings

$q(j|i)$= Gaussian centered at $f_\theta(x_i)$

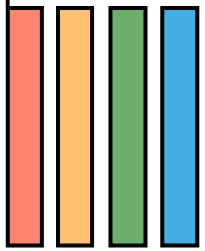


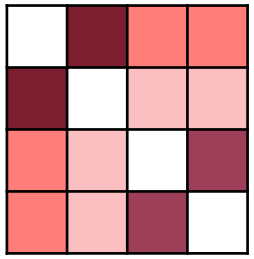
$$\mathcal{L} = \sum D_{KL}(p(*|i) \parallel q(*|i))$$

I-Con Framework Overview


Mapper $f_\cdot$

Learned Representations





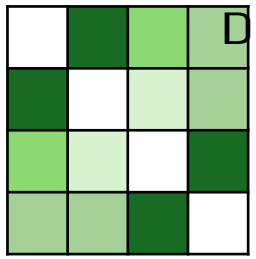
Learned Conditional Distribution

$q(i|$ 

$\int D_{KL}(p(i|$  $||q(i|$  $)$

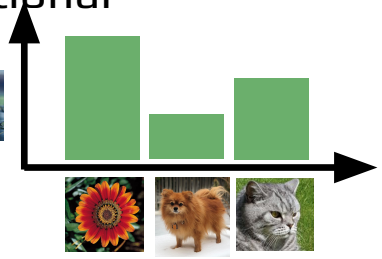
Supervisory Signal

e.g. Gaussian, Paired Inputs and Classes, Augmentation Pairs



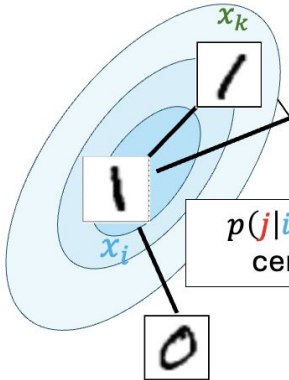
Supervisory Conditional Distribution

$p(i|$ 



SNE

High Dimensional Data

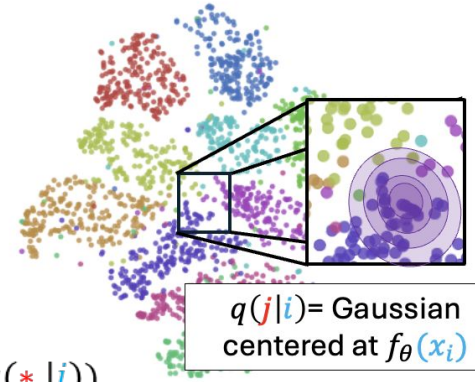


$p(j|i)$ = Gaussian centered at x_i

Lookup Table

f_θ

Low Dimensional Data



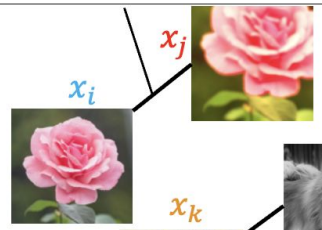
$q(j|i)$= Gaussian centered at $f_\theta(x_i)$

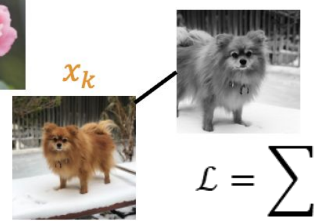
$$\mathcal{L} = \sum D_{KL}(p(*|i) \parallel q(*|i))$$

SimCLR

RGB Images

$p(j|i)) = \mathbb{1}[\ i$ and j are augmentations of the same image$]$



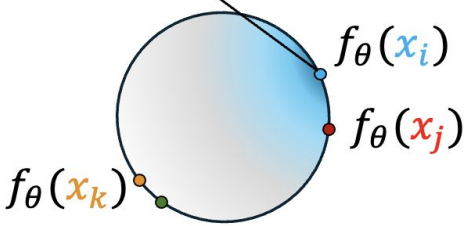


Parametric Mapper

f_θ

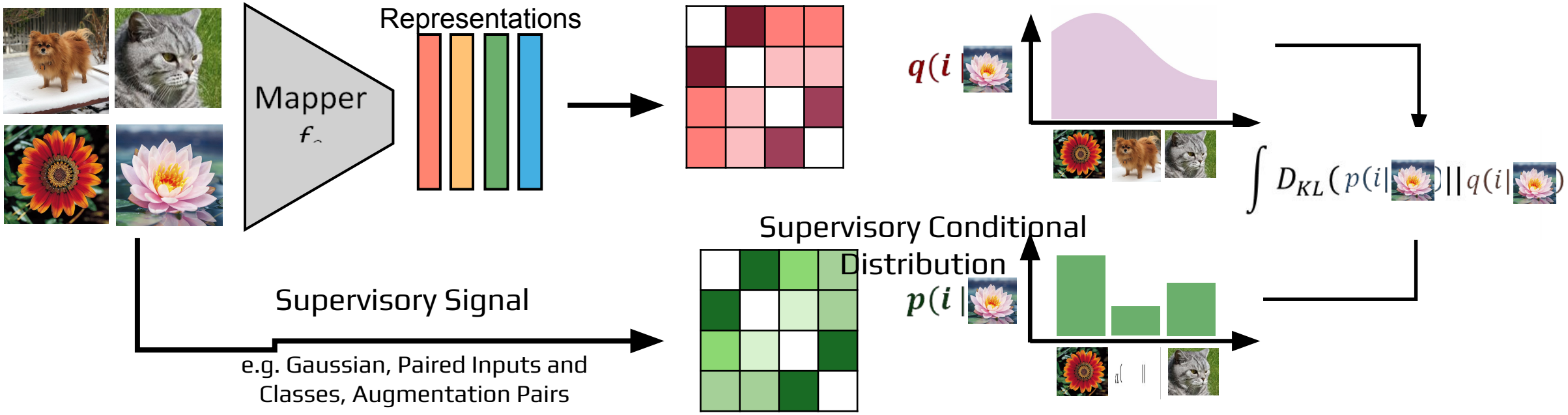
Embeddings

$q(j|i)$= Gaussian centered at $f_\theta(x_i)$



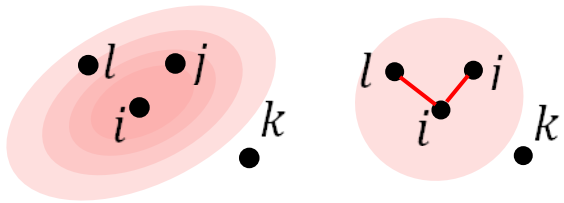
$$\mathcal{L} = \sum D_{KL}(p(*|i) \parallel q(*|i))$$

Framework Overview



Examples for types of p and q

A) Spatial

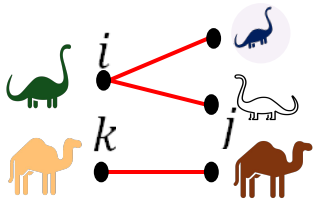


e.g. Gaussian, Studen-T

k-NN

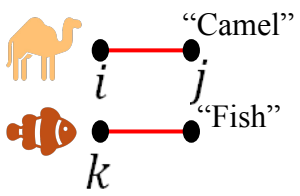
Neighbors are based on distance

B) Discrete



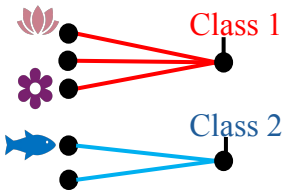
Positive Pairs

e.g., data augmentations or same-label pairs



Cross-Modal

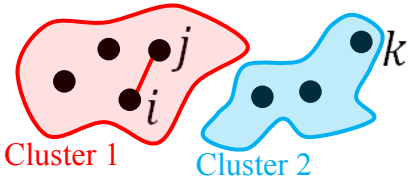
e.g. image-text



Labels

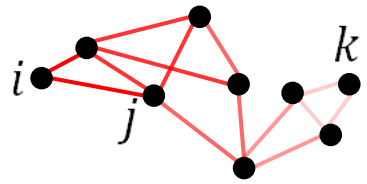
e.g., images linked to category prototypes

C) Cluster



Neighbors grouped by shared cluster membership

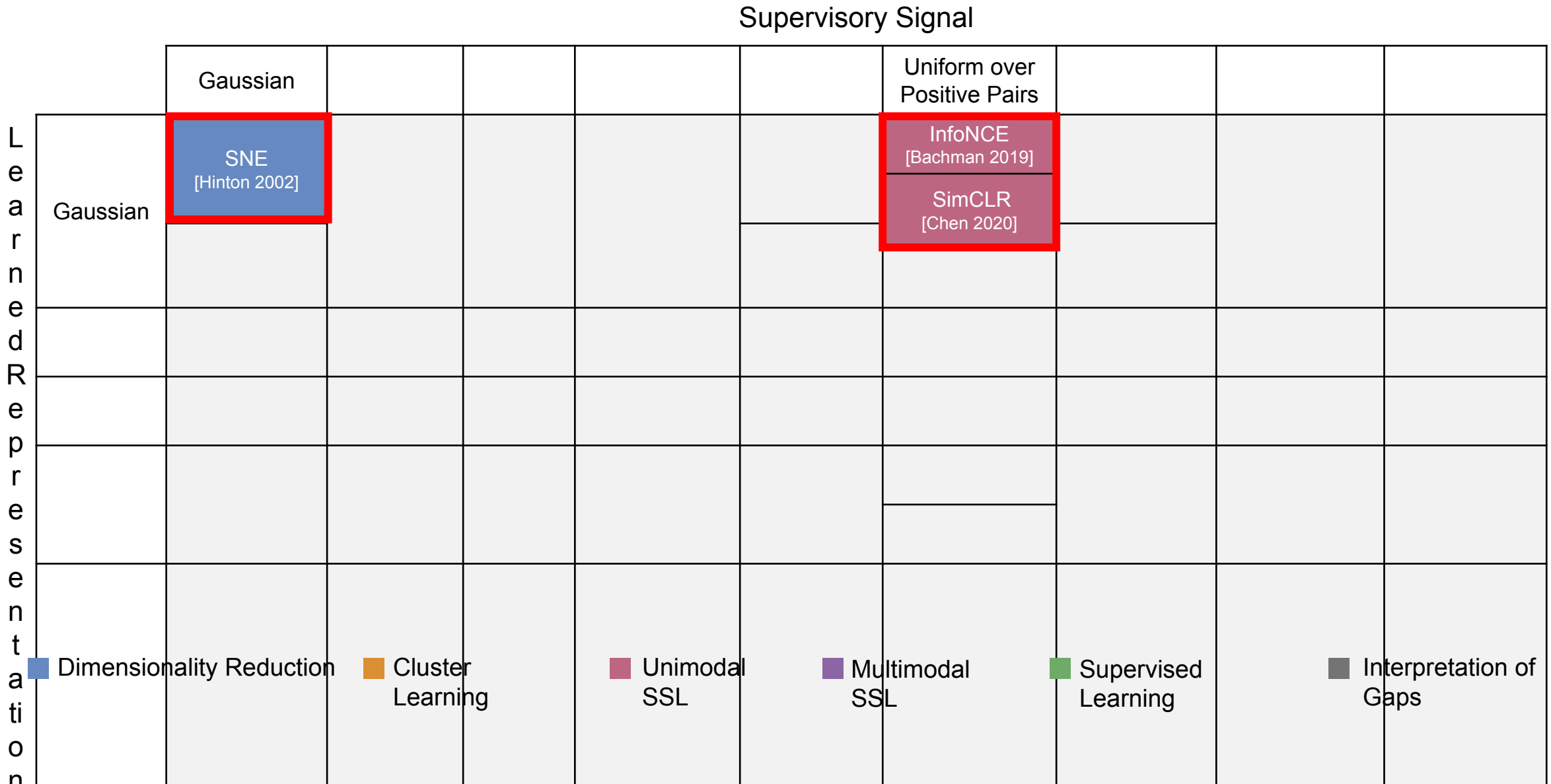
D) Graph



Neighbors determined by graph connectivity

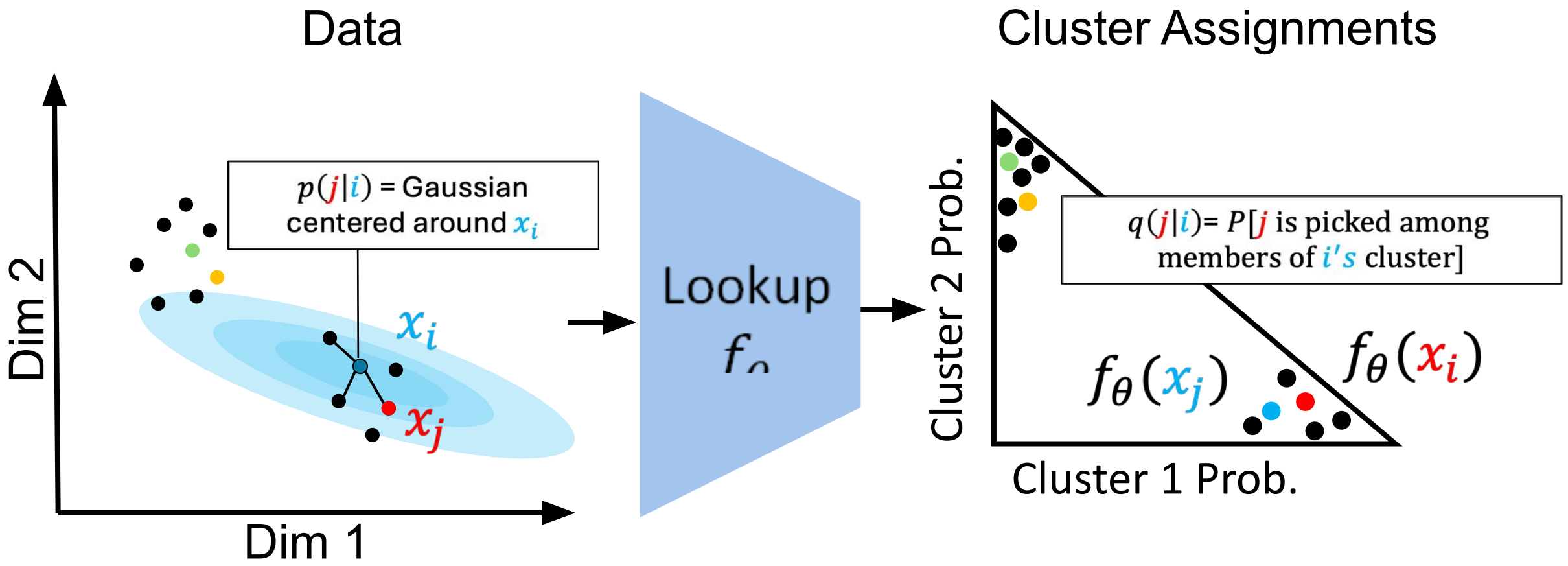
Overview

- Intro
  - Welcome to the Zoo
  - The Periodic Table
- Methods:
  - Exploring some Examples
  - Generalizing ML Methods with a single Equation
  - **Building the Periodic Table**
- Experiments
  - Building an Image Classifier that doesn’t need human labels
- Future Directions


Supervisory Signal
Gaussian
Uniform over Positive Pairs
Learned Representation
Gaussian
SNE
[Hinton 2002]
InfoNCE
[Bachman 2019]
SimCLR
[Chen 2020]
Dimensionality Reduction
Cluster Learning
Unimodal SSL
Multimodal SSL
Supervised Learning
Interpretation of Gaps

Example 3: Clustering

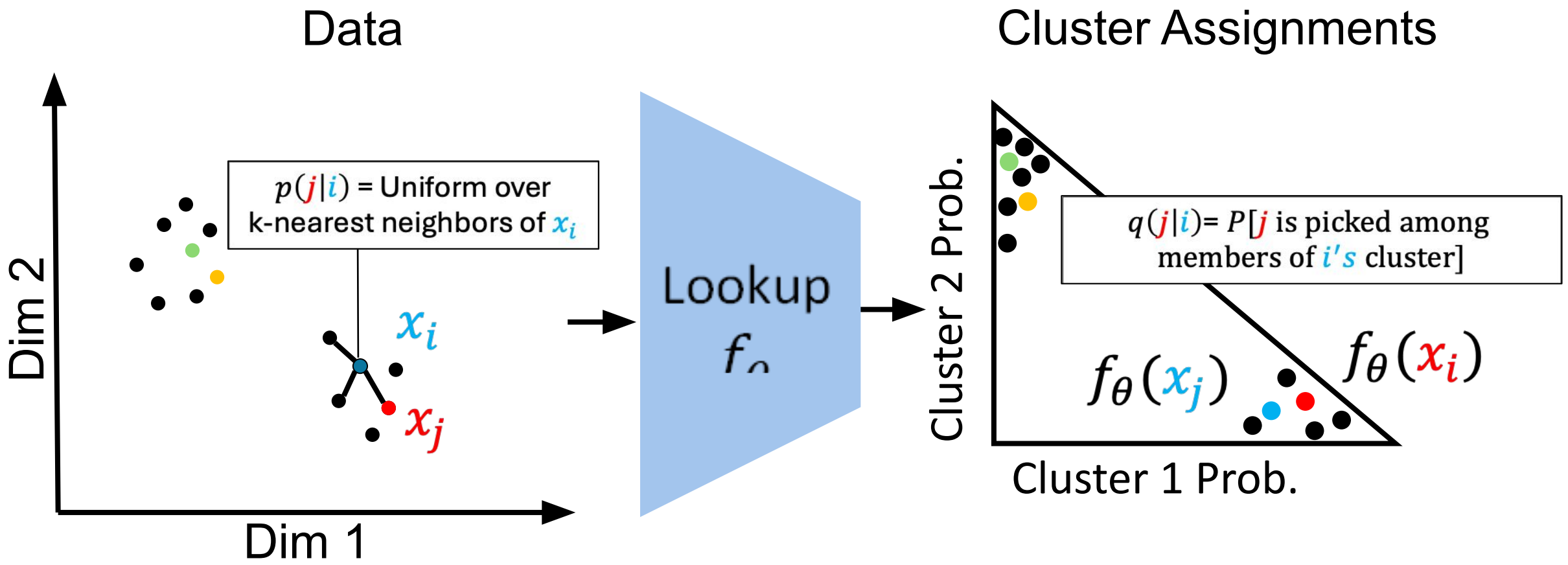
K-Means



$$\mathcal{L} = \sum_i D_{KL}(p(*|i) \parallel q(*|i))$$

Example 3: Clustering

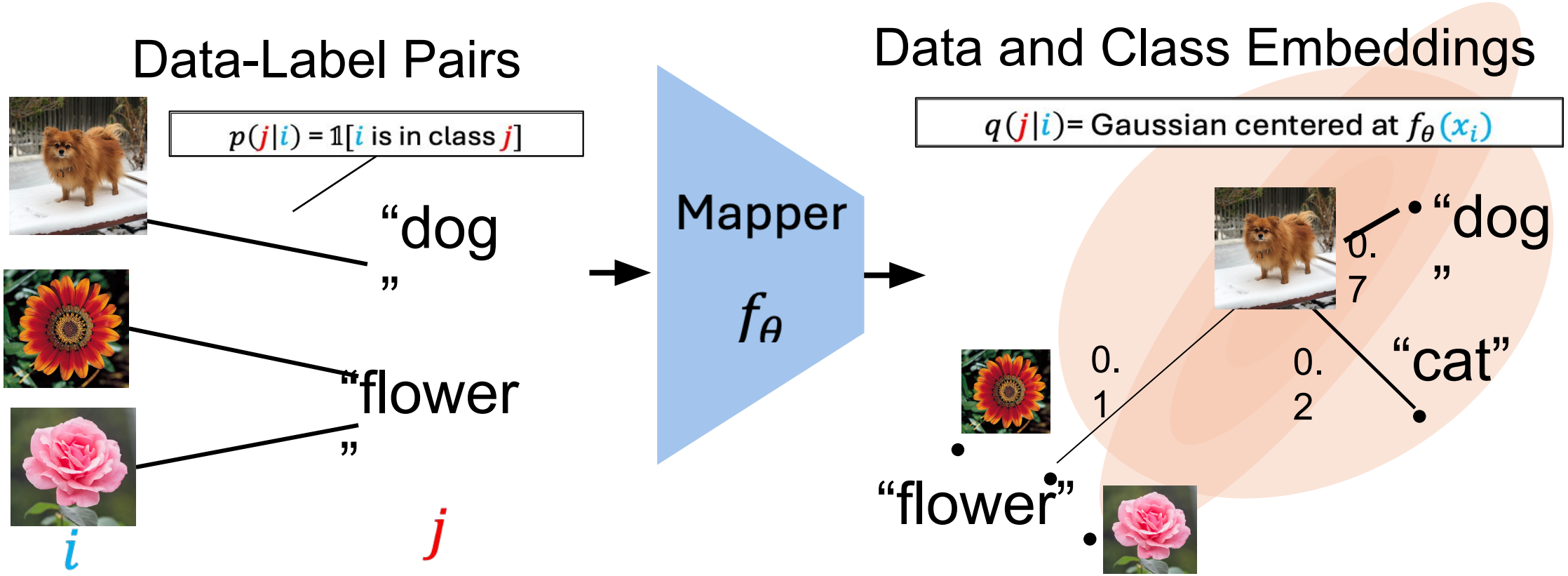
DCD



$$\mathcal{L} = \sum_i D_{KL}(p(*|i) \parallel q(*|i))$$

Example 4: Supervised Learning

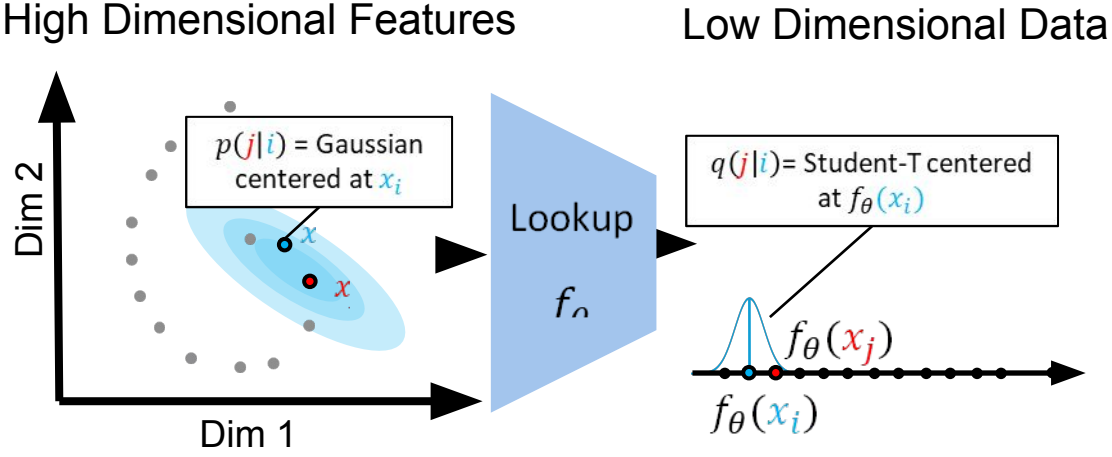
CrossEntropy



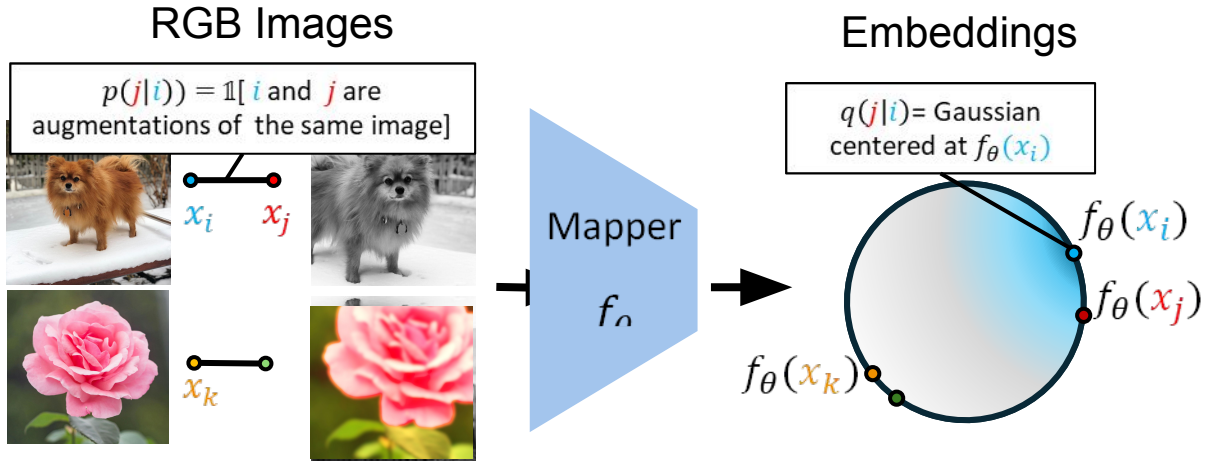
$$\mathcal{L} = \sum_i D_{KL}(p(*|i) \parallel q(*|i))$$

Examples of methods as special cases of I-Con via different choices of p, q, and f_θ

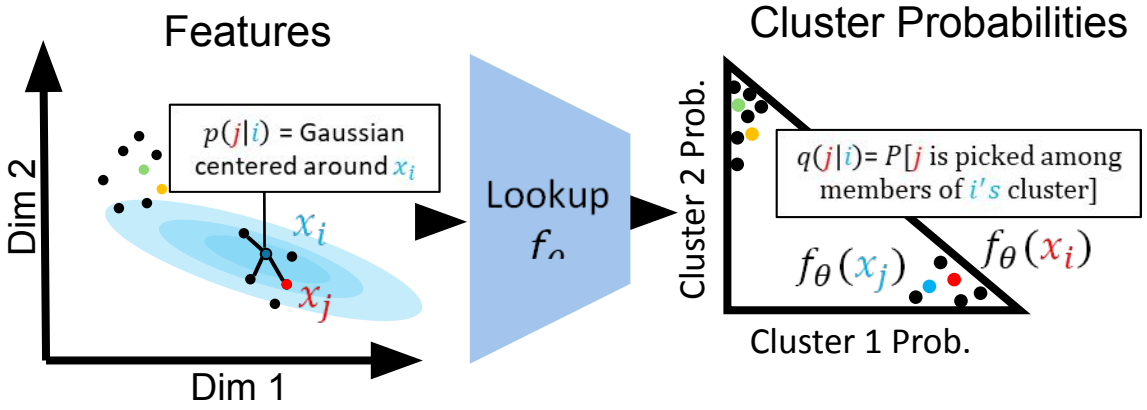
t-SNE (Dimensionality Reduction)



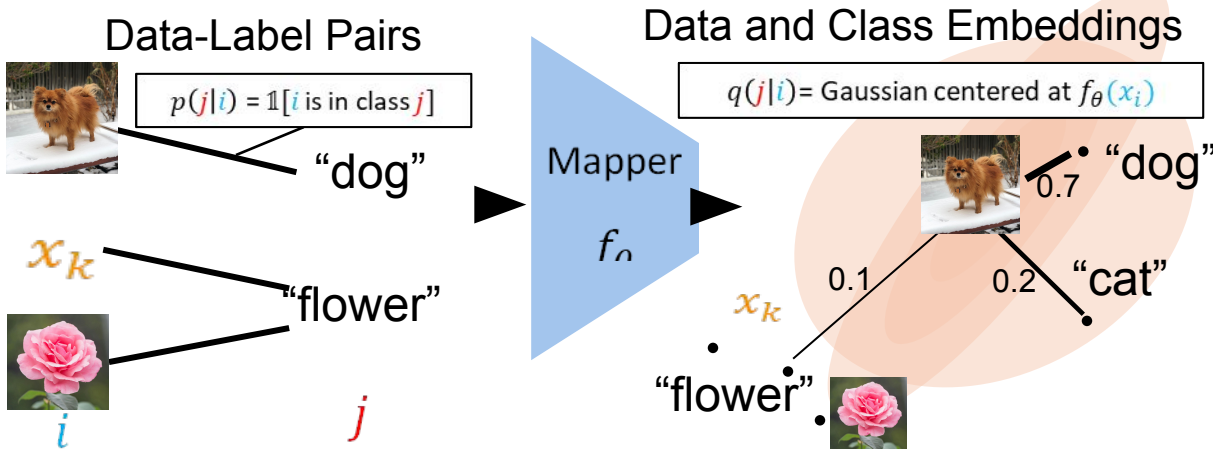
SimCLR (Self-Supervised Learning)



K-Means (Clustering)

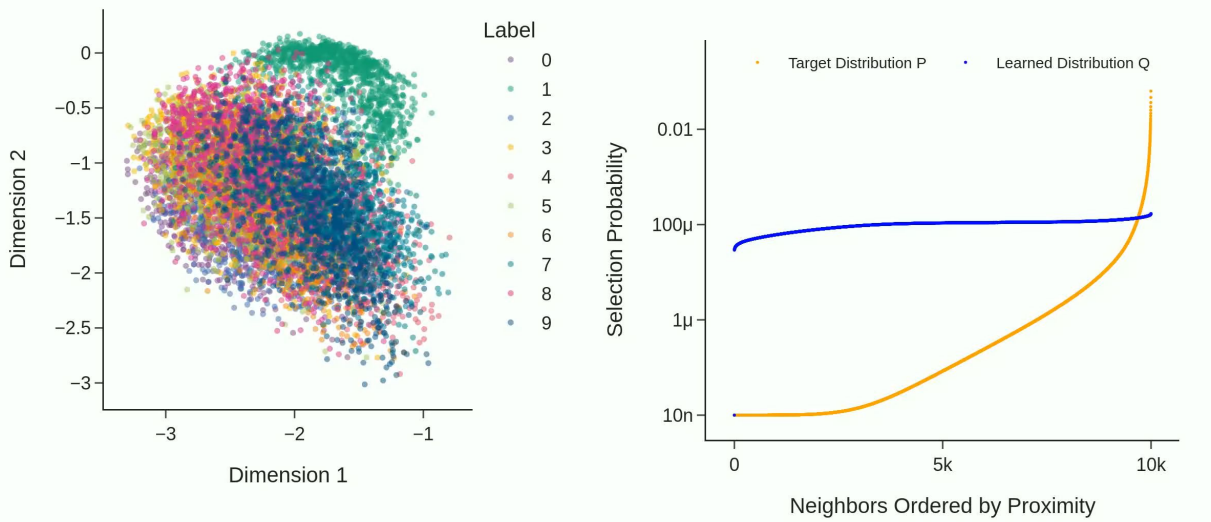


Supervised Cross Entropy (Classification)

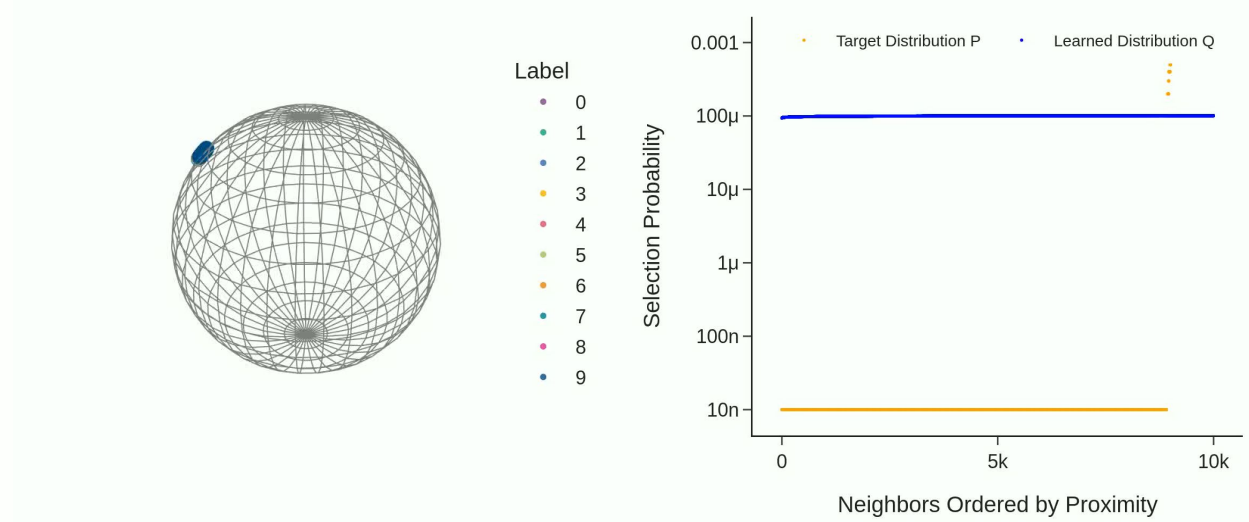


Examples of methods as special cases of I-Con via different choices of p, q, and f_θ on MNIST

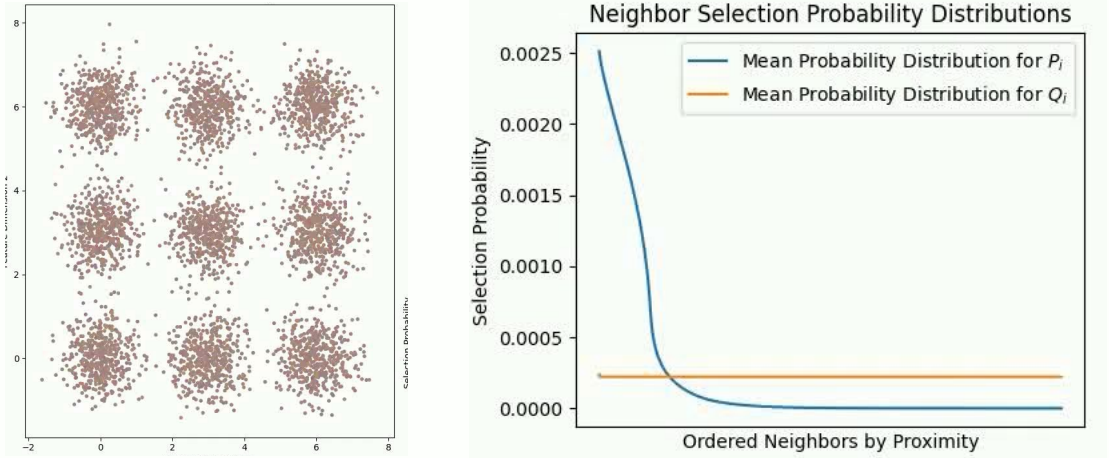
t-SNE (Dimensionality Reduction)



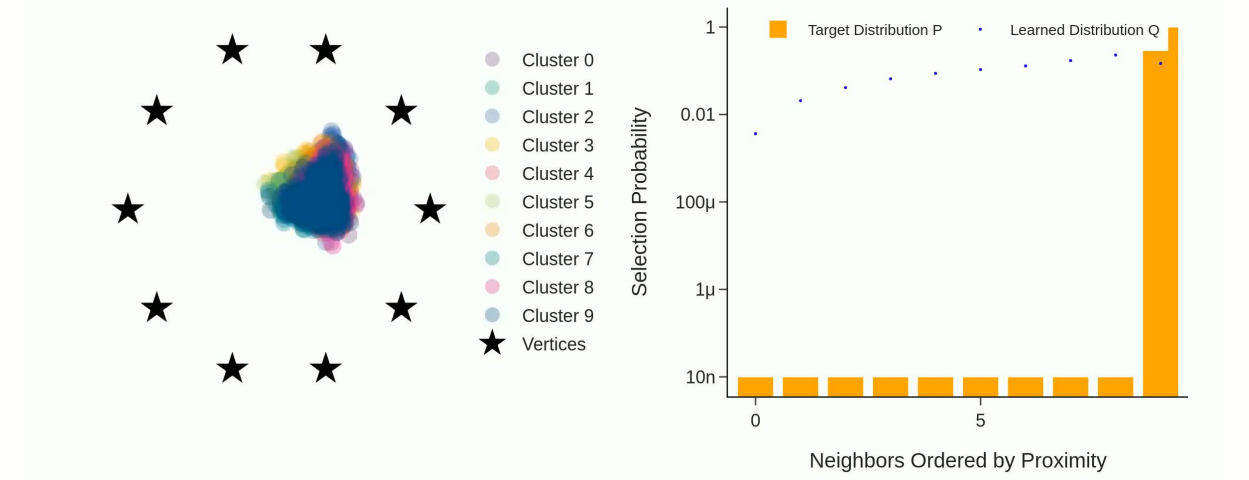
Supervised Contrastive Learning

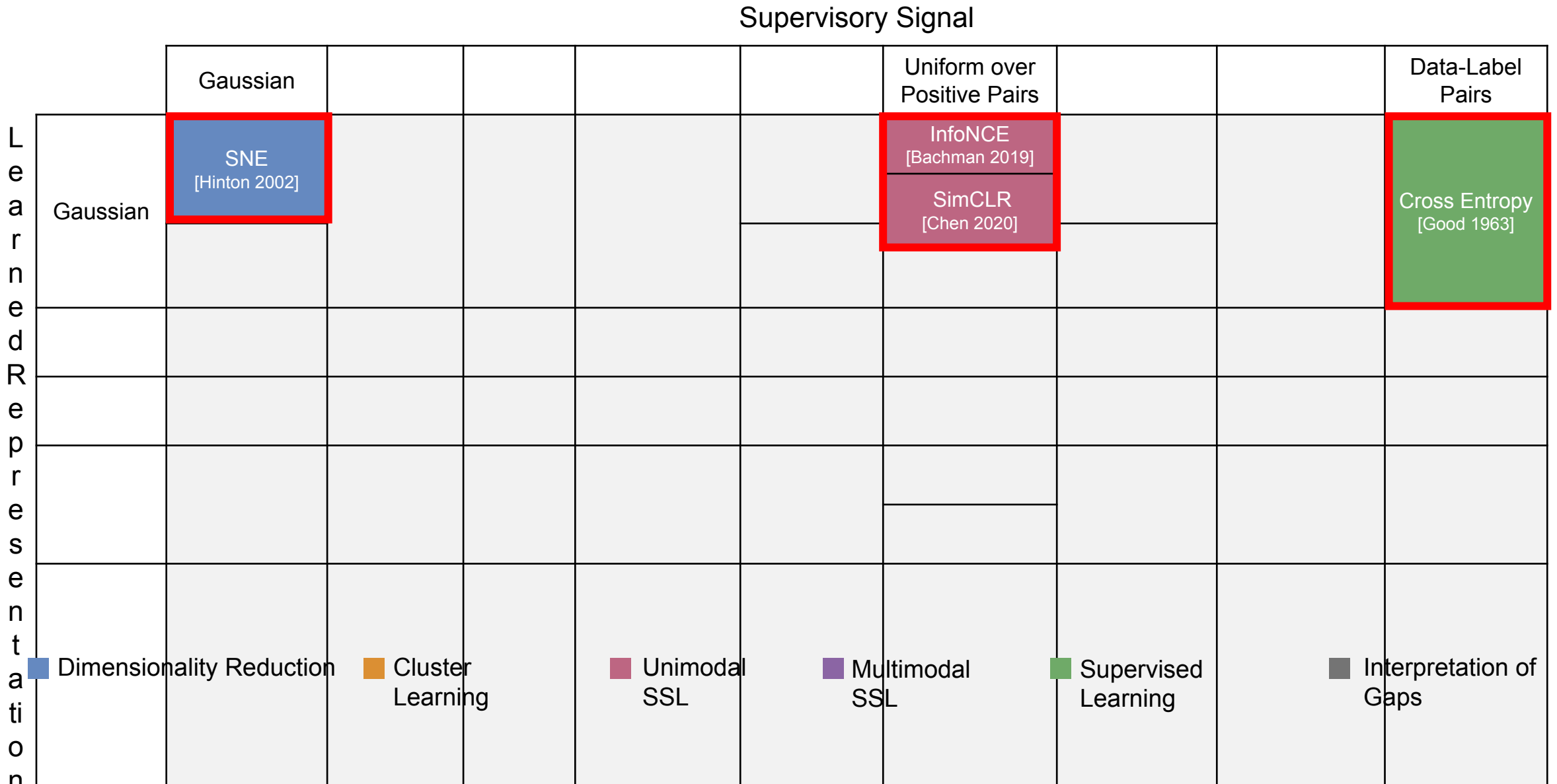


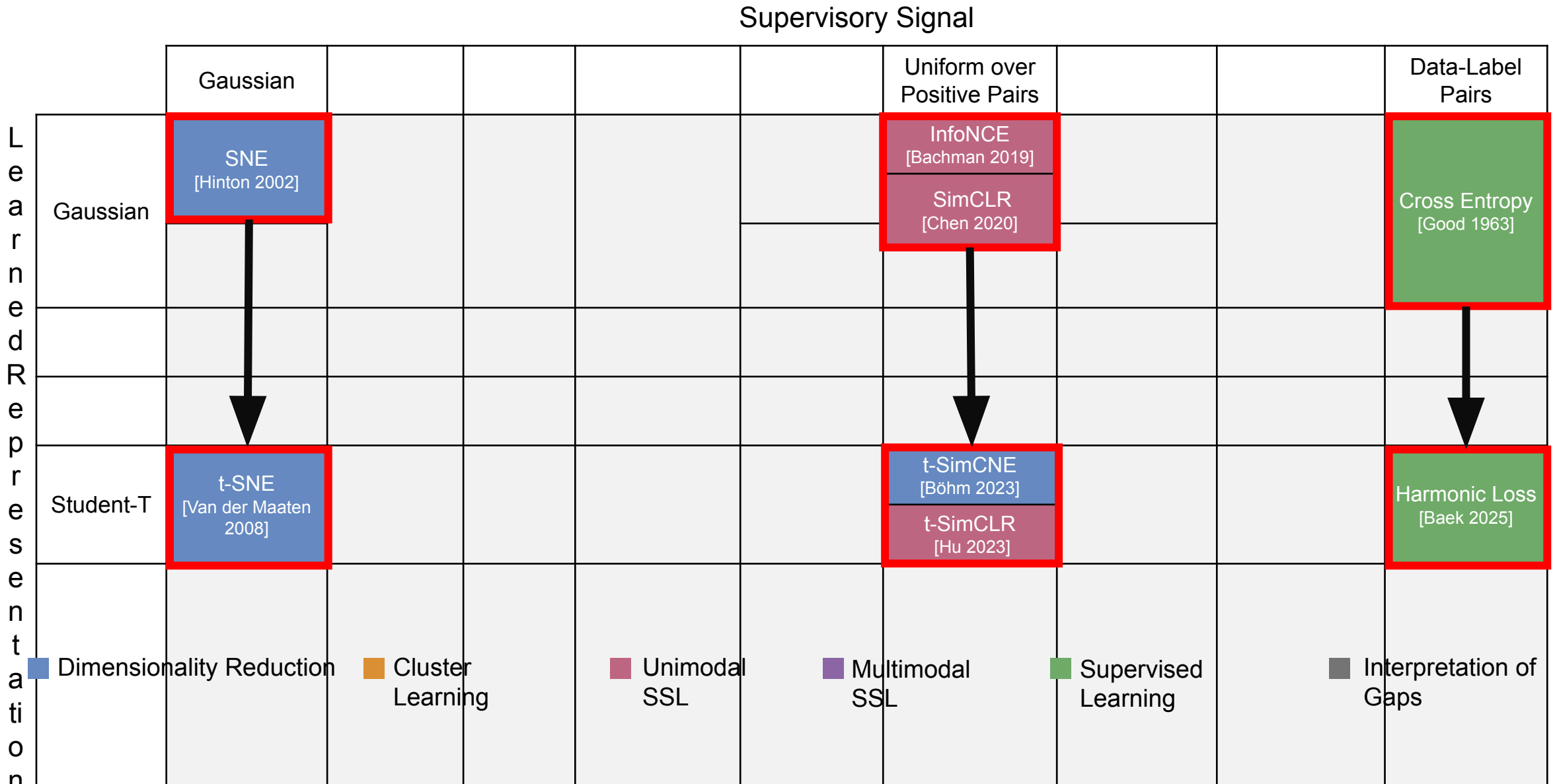
K-Means or DCD (Clustering)

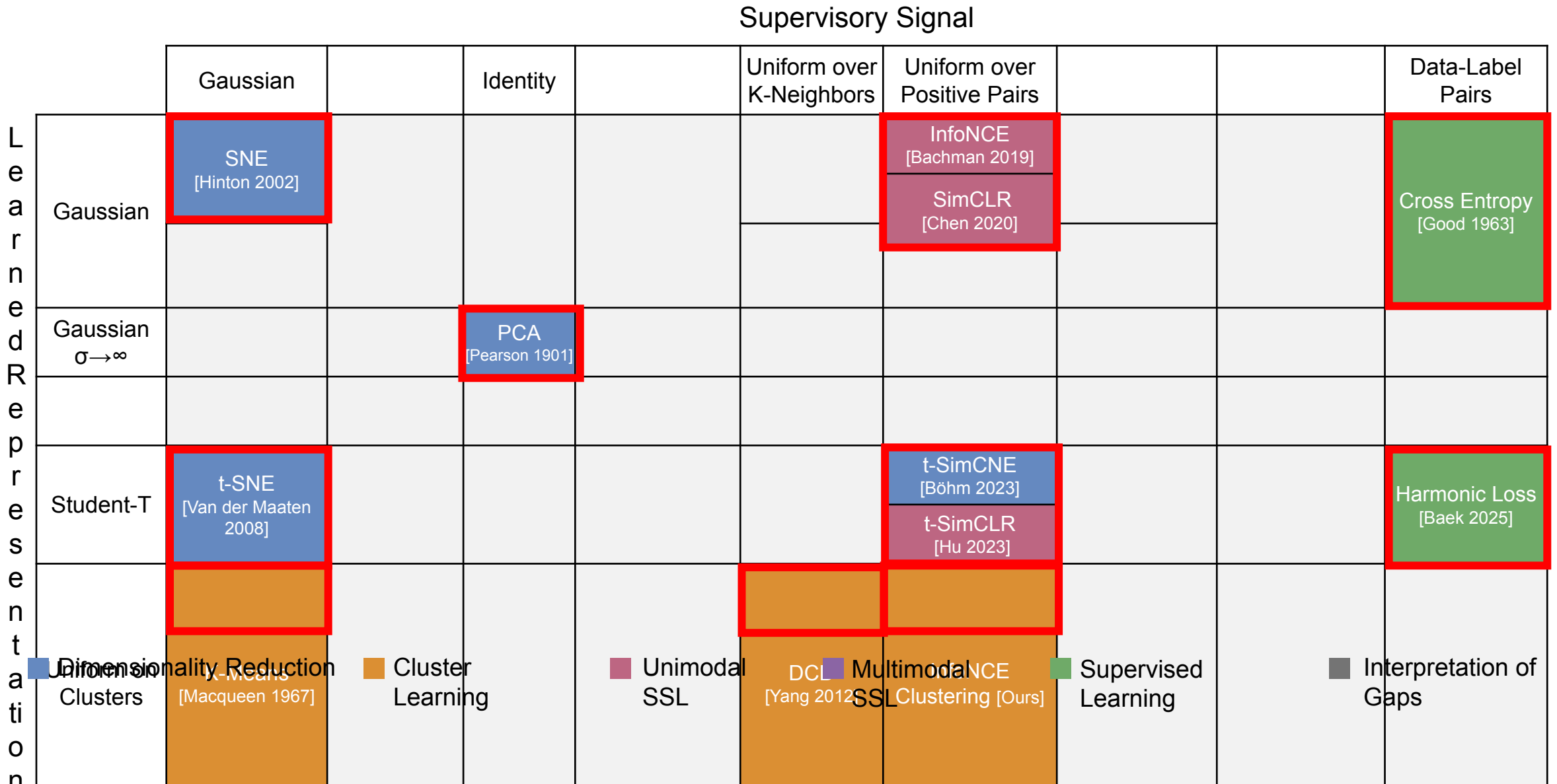


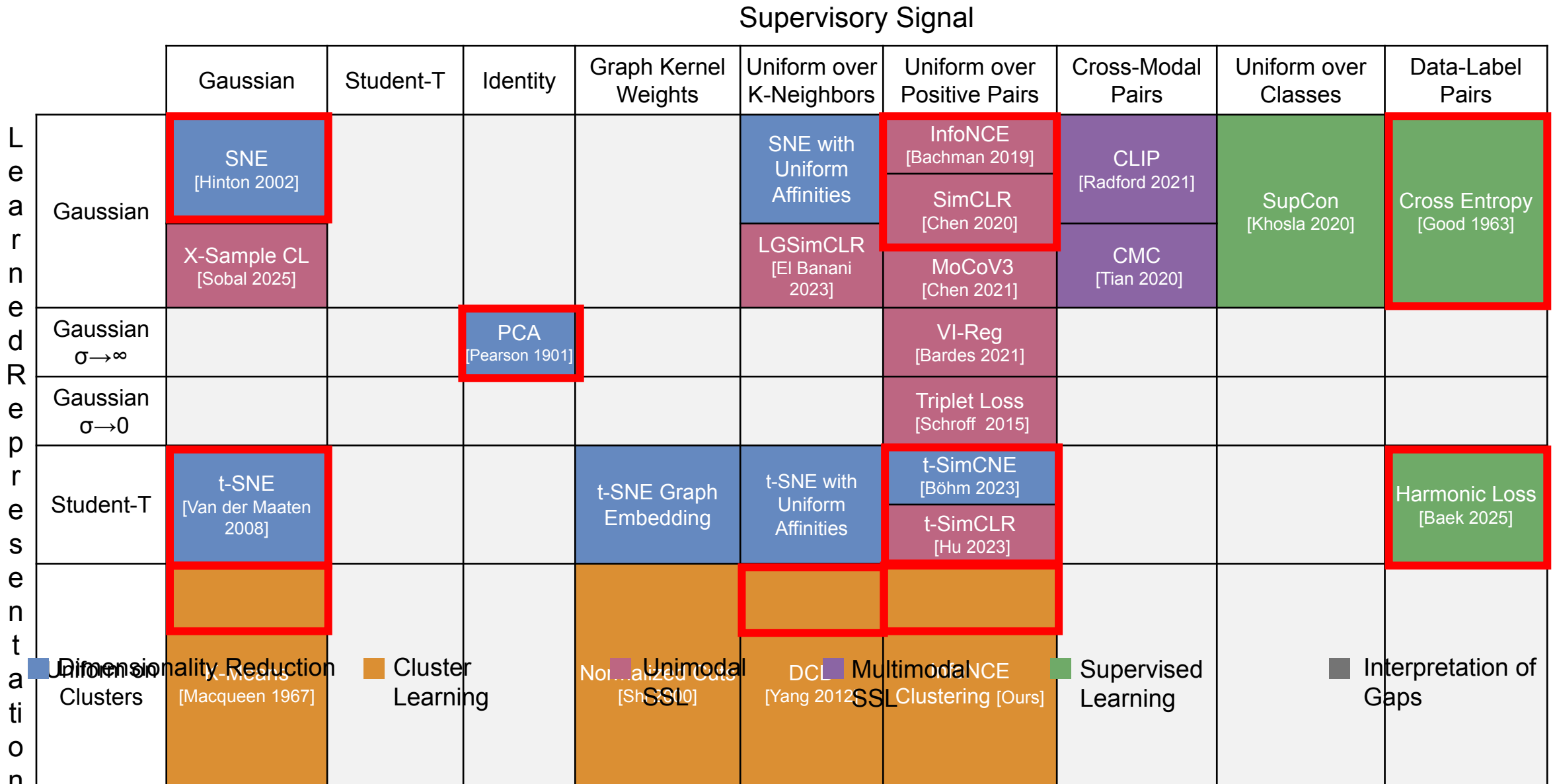
Supervised Cross Entropy (Classification)

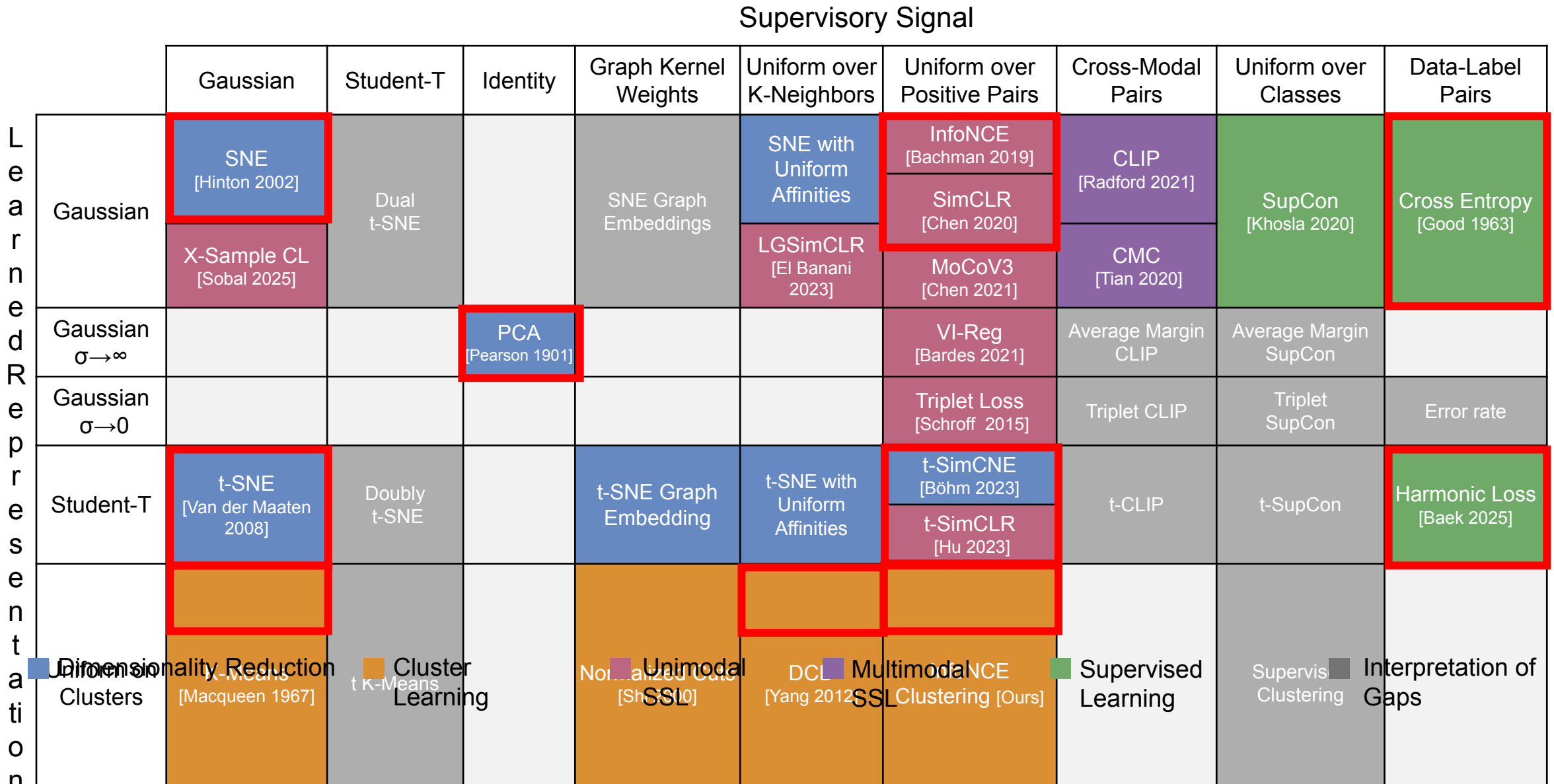



Supervisory Signal
Gaussian
Uniform over Positive Pairs
Data-Label Pairs
Learned Representation
Gaussian
SNE
[Hinton 2002]
InfoNCE
[Bachman 2019]
SimCLR
[Chen 2020]
Cross Entropy
[Good 1963]
Dimensionality Reduction
Cluster Learning
Unimodal SSL
Multimodal SSL
Supervised Learning
Interpretation of Gaps


Supervisory Signal
Gaussian
Uniform over Positive Pairs
Data-Label Pairs
Learned Representation
Gaussian
Student-T
SNE
[Hinton 2002]
InfoNCE
[Bachman 2019]
SimCLR
[Chen 2020]
Cross Entropy
[Good 1963]
t-SNE
[Van der Maaten 2008]
t-SimCNE
[Böhm 2023]
t-SimCLR
[Hu 2023]
Harmonic Loss
[Baek 2025]
Dimensionality Reduction
Cluster Learning
Unimodal SSL
Multimodal SSL
Supervised Learning
Interpretation of Gaps


Supervisory Signal
Gaussian
Identity
Uniform over K-Neighbors
Uniform over Positive Pairs
Data-Label Pairs
Learned Representation
Gaussian
SNE [Hinton 2002]
InfoNCE [Bachman 2019]
SimCLR [Chen 2020]
Cross Entropy [Good 1963]
Gaussian σ→∞
PCA [Pearson 1901]
Student-T
t-SNE [Van der Maaten 2008]
t-SimCNE [Böhm 2023]
t-SimCLR [Hu 2023]
Harmonic Loss [Baek 2025]
Uniform on Clusters
k-Means [Macqueen 1967]
DCL [Yang 2012]
CUNCE Clustering [Ours]
Dimensionality Reduction
Cluster Learning
Unimodal SSL
Multimodal SSL
Supervised Learning
Interpretation of Gaps


Supervisory Signal
Learned Representation
Gaussian
Student-T
Identity
Graph Kernel Weights
Uniform over K-Neighbors
Uniform over Positive Pairs
Cross-Modal Pairs
Uniform over Classes
Data-Label Pairs
Gaussian
SNE [Hinton 2002]
X-Sample CL [Sobal 2025]
SNE with Uniform Affinities
LGSimCLR [El Banani 2023]
InfoNCE [Bachman 2019]
SimCLR [Chen 2020]
MoCoV3 [Chen 2021]
CLIP [Radford 2021]
CMC [Tian 2020]
SupCon [Khosla 2020]
Cross Entropy [Good 1963]
Gaussian σ→∞
PCA [Pearson 1901]
VI-Reg [Bardes 2021]
Gaussian σ→0
Triplet Loss [Schroff 2015]
Student-T
t-SNE [Van der Maaten 2008]
t-SNE Graph Embedding
t-SNE with Uniform Affinities
t-SimCNE [Böhm 2023]
t-SimCLR [Hu 2023]
Harmonic Loss [Baek 2025]
Uniform on Clusters
k-Means [Macqueen 1967]
Normalized Cuts [Shi 2000]
DCL [Yang 2012]
Clustering [Ours]
Dimensionality Reduction
Cluster Learning
Unimodal SSL
Multimodal SSL
Supervised Learning
Interpretation of Gaps


Supervisory Signal
Learned Representation
Gaussian
Student-T
Identity
Graph Kernel Weights
Uniform over K-Neighbors
Uniform over Positive Pairs
Cross-Modal Pairs
Uniform over Classes
Data-Label Pairs
Gaussian
SNE [Hinton 2002]
X-Sample CL [Sobal 2025]
Dual t-SNE
SNE Graph Embeddings
SNE with Uniform Affinities
LGSimCLR [El Banani 2023]
InfoNCE [Bachman 2019]
SimCLR [Chen 2020]
MoCoV3 [Chen 2021]
CLIP [Radford 2021]
CMC [Tian 2020]
SupCon [Khosla 2020]
Cross Entropy [Good 1963]
Gaussian σ→∞
PCA [Pearson 1901]
VI-Reg [Bardes 2021]
Average Margin CLIP
Average Margin SupCon
Gaussian σ→0
Triplet Loss [Schroff 2015]
Triplet CLIP
Triplet SupCon
Error rate
Student-T
t-SNE [Van der Maaten 2008]
Doubly t-SNE
t-SNE Graph Embedding
t-SNE with Uniform Affinities
t-SimCNE [Böhm 2023]
t-SimCLR [Hu 2023]
t-CLIP
t-SupCon
Harmonic Loss [Baek 2025]
Uniform on Clusters
K-Means [Macqueen 1967]
t K-Means
Normalized Cuts [Shi 2000]
DCL [Yang 2012]
SimCNE Clustering [Ours]
Supervised Clustering
Dimensionality Reduction
Cluster Learning
Unimodal SSL
Multimodal SSL
Supervised Learning
Interpretation of Gaps

Overview

- Intro
  - Welcome to the Zoo
  - The Periodic Table
- Methods:
  - Exploring some Examples
  - Generalizing ML Methods with a single Equation
  - Building the Periodic Table
- Experiments
  - **Building an Image Classifier that doesn't need human labels**
- Future Directions

Building New Representation Learners

Given the I-Con framework, we developed state-of-the-art clustering methods by:

- Filling a gap in the table
- Transferring ideas from other fields
- Adapting the neighborhood width

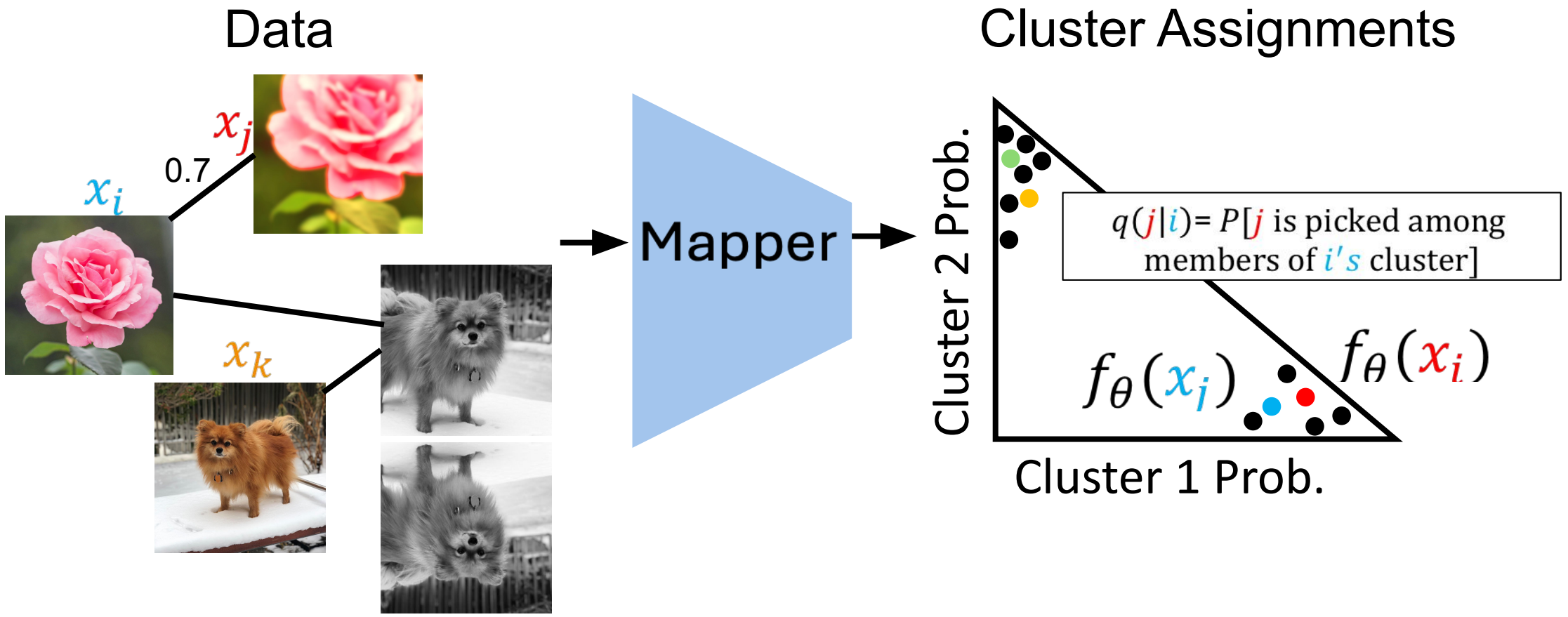
Supervisory Signal

Learned Representation	Gaussian	Student-T	Identity	Graph Kernel Weights	Uniform over K-Neighbors	Uniform over Positive Pairs	Cross-Modal Pairs	Uniform over Classes	Data-Label Pairs
Gaussian	SNE [Hinton 2002] X-Sample CL [Sobal 2025]	Dual t-SNE		SNE Graph Embeddings	SNE with Uniform Affinities LGSimCLR [El Banani 2023]	InfoNCE [Bachman 2019] SimCLR [Chen 2020] MoCoV3 [Chen 2021]	CLIP [Radford 2021] CMC [Tian 2020]	SupCon [Khosla 2020]	Cross Entropy [Good 1963]
Gaussian $\sigma \to \infty$			PCA [Pearson 1901]			VI-Reg [Bardes 2021]	Average Margin CLIP	Average Margin SupCon	
Gaussian $\sigma \to 0$						Triplet Loss [Schroff 2015]	Triplet CLIP	Triplet SupCon	Error rate
Student-T	t-SNE [Van der Maaten 2008]	Doubly t-SNE		t-SNE Graph Embedding	t-SNE with Uniform Affinities	t-SimCNE [Böhm 2023] t-SimCLR [Hu 2023]	t-CLIP	t-SupCon	Harmonic Loss [Baek 2025]
Uniform									

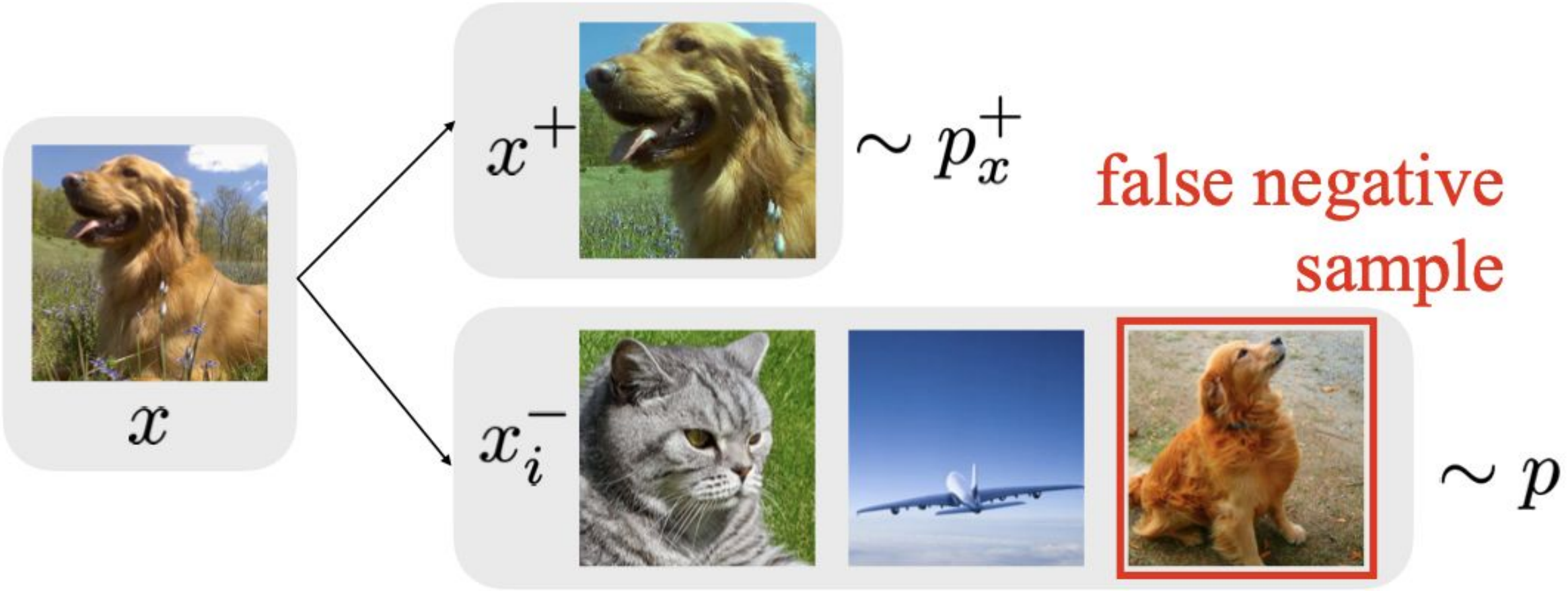
Building New Representation Learners

Given the I-Con framework, we developed state-of-the-art clustering methods by:

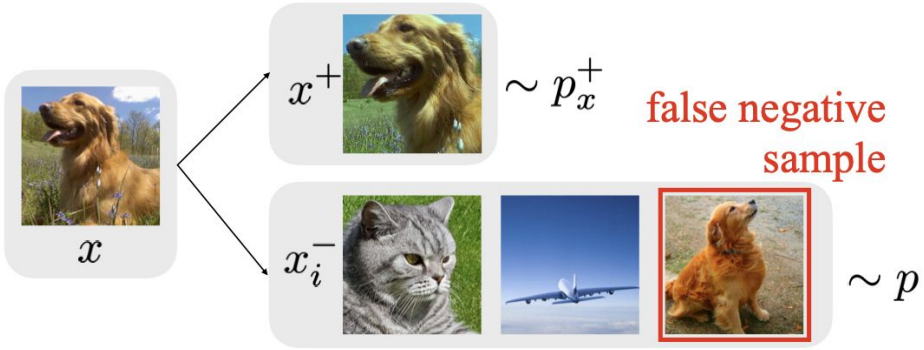
- Filling a gap in the table
- Transferring ideas from other fields
- Adapting the neighborhood width



Debiased Contrastive Learning (Chuang et al, NeurIPS 2020)



Debiased Contrastive Learning (Chuang et al, NeurIPS 2020)



doesn't represent a conditional probability

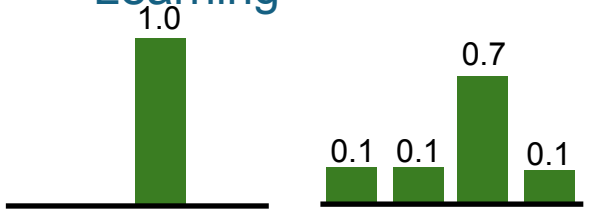
$$L_{\text{Debiased}}^{N,M}(f) = \mathbb{E}_{\substack{x\sim p;\, x^+\sim p_x^+ \\ \{u_i\}_{i=1}^N \sim p^N \\ \{v_i\}_{i=1}^N \sim p_x^{+M}}} \left[-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + N g\left(x, \{u_i\}_{i=1}^N, \{v_i\}_{i=1}^M\right)} \right]$$

where

$$g(x, \{u_i\}_{i=1}^N, \{v_i\}_{i=1}^M) = \max\left\{ \frac{1}{\tau^-}\left(\frac{1}{N}\sum_{i=1}^N e^{f(x)^T f(u_i)} - \tau^+ \frac{1}{M}\sum_{i=1}^M e^{f(x)^T f(v_i)} \right),\ e^{-1/t} \right\}$$

A more general approach for debiasing

Supervised Learning

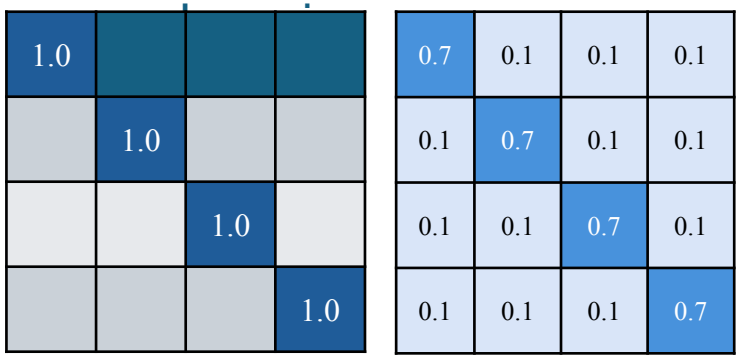


$$\tilde{p}(c|i) = (1-\alpha)p(c|i) + \frac{\alpha}{N}$$

Reflect the uncertainty in the labels, not the softmax

$$\mathcal{L} = \sum \tilde{p}(c|i) \log q(c|i)$$

Contrastive

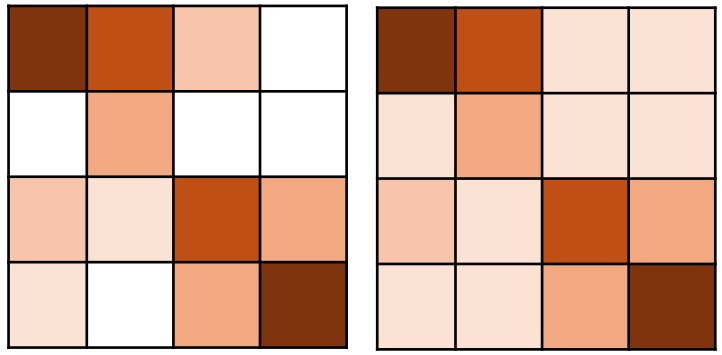


$$\tilde{p}(j|i) = (1-\alpha)p(j|i) + \frac{\alpha}{N}$$

Soften target distribution, but learned distribution unchanged

$$\mathcal{L} = \mathbb{E}_{\tilde{p}}\left[-\log \frac{\exp(f_\theta(i)\cdot f_\theta(j))}{\sum_k \exp(f_\theta(i)\cdot f_\theta(k))} \right]$$

Clustering

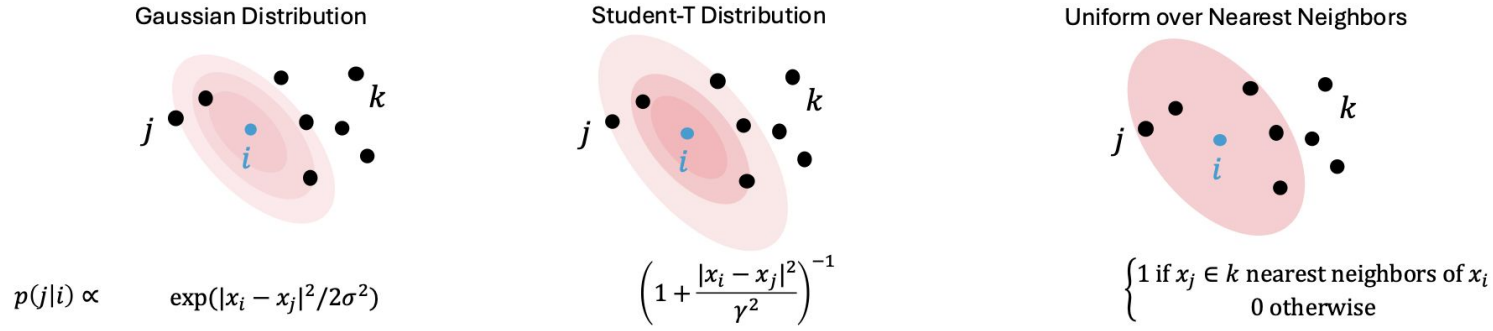


$$\tilde{p}(j|i) = (1-\alpha)p(j|i) + \frac{\alpha}{N}$$

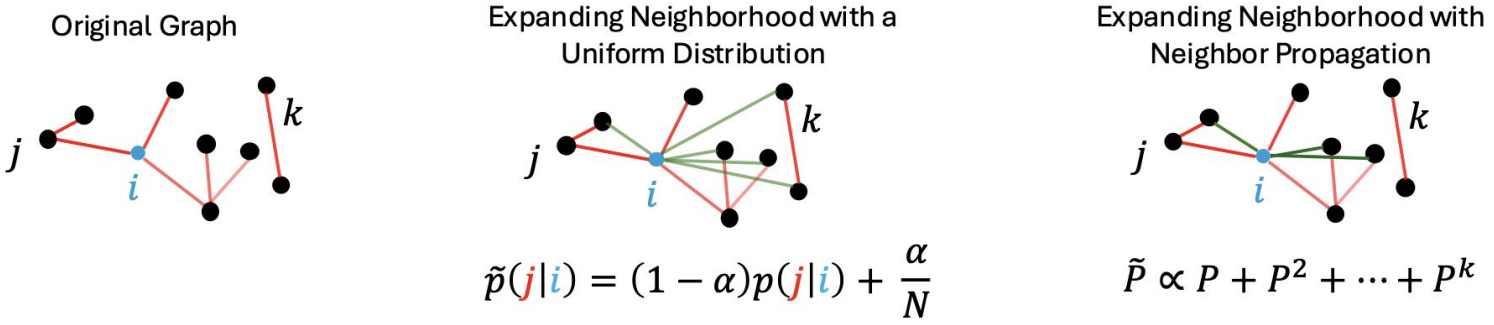
Learned distribution unchanged

$$\mathcal{L} = \sum_i KL\left(\tilde{p}(j|i) \,\|\, q\left(f_\theta(x_j)|f_\theta(x_i)\right)\right)$$

Adapting the neighborhood width



(a) Continuous distance-based distributions control neighborhood width via hyperparameters.



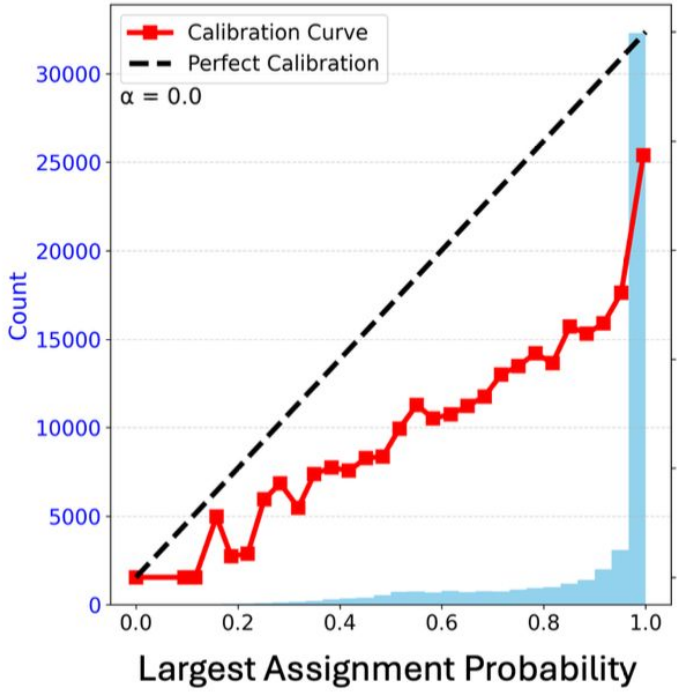
(b) Graph-based distributions expand neighborhoods through structural strategies.

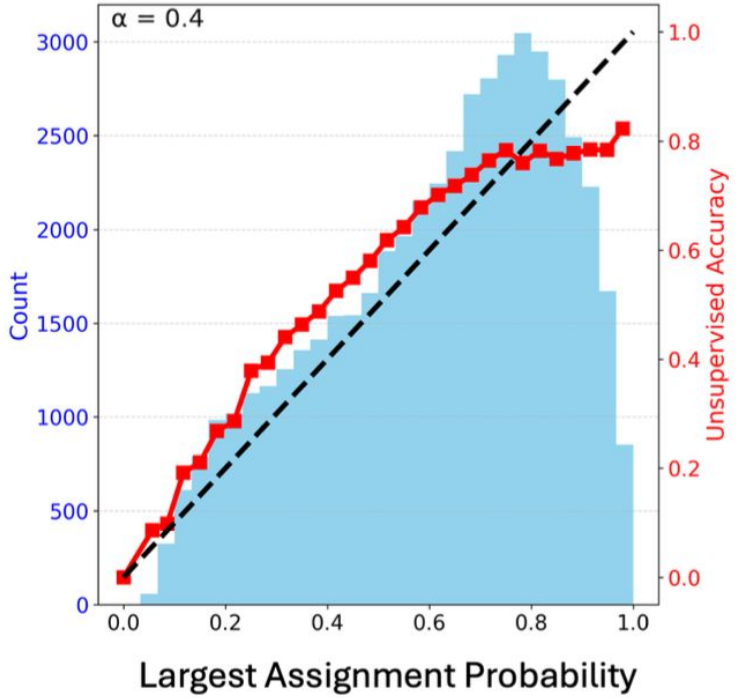
Method	DiNO ViT-S/14	DiNO ViT-B/14	DiNO ViT-L/14
Baseline	55.51	63.03	65.72
+ KNNs	56.43	64.26	65.70
+ 1-walks on KNN	**58.09**	**64.29**	65.97
+ 2-walks on KNN	57.84	64.27	**67.26**
+ 3-walks on KNN	57.82	64.15	67.02

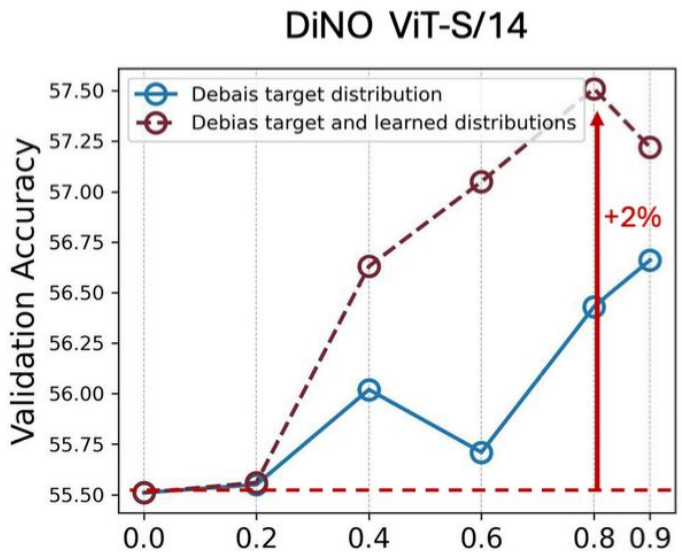
Results for our Debiasing in **Cluster Learning**

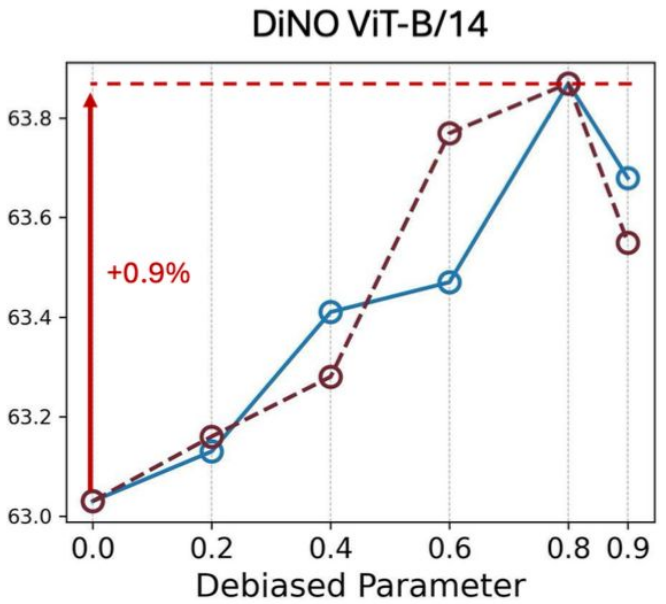
$$\mathcal{L} = \sum_i KL\left(\tilde{p}(j|i) \parallel q\left(f_\theta(x_j)\middle|f_\theta(x_i)\right)\right)$$

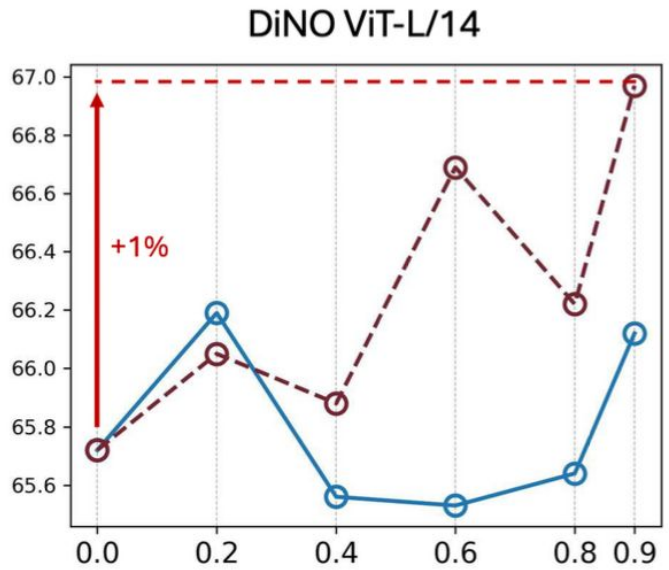
where $\tilde{p}(j|i) = (1-\alpha)p(j|i) + \frac{\alpha}{N}$











More Results in **Cluster Learning**

Method	DiNO ViT-S/14	DiNO ViT-B/14	DiNO ViT-L/14
Baseline	55.51	63.03	65.70
+ Debiasing	57.27± 0.07	63.72 ± 0.09	66.87 ± 0.07
+ KNN Propagation	**58.45** ± 0.23	64.87 ± 0.19	67.25 ± 0.21
+ EMA	57.8 ± 0.26	**64.75** ± 0.18	**67.52** ± 0.28

Method	DiNO ViT-S/14	DiNO ViT-B/14	DiNO ViT-L/14
k-Means	51.84	52.26	53.36
Contrastive Clustering	47.35	55.64	59.84
SCAN	49.20	55.60	60.15
TEMI	56.84	58.62	–
Debiased InfoNCE Clustering (Ours)	**57.8** ± 0.26	**64.75** ± 0.18	**67.52** ± 0.28

Slides Credits

- 6.S898 Deep Learning, MIT EECS https://phillipi.github.io/6.s898/
- 6.8300/6.8301: Advances in Computer Vision, MIT CSAIL, https://advances-in-vision.github.io