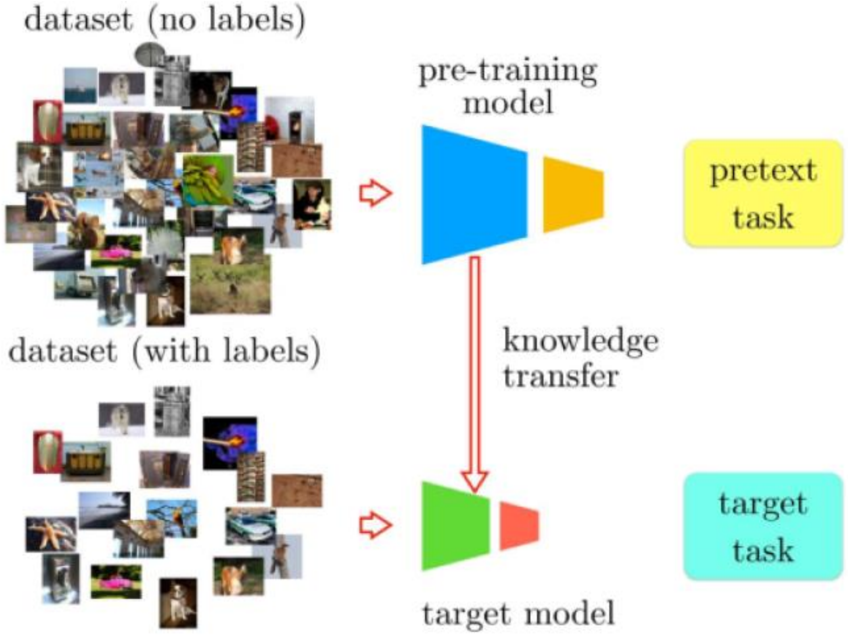
ECE 4252/8803: Fundamentals of Machine Learning (FunML)
Fall 2024

Lecture 26: Self-supervised Learning



Overview

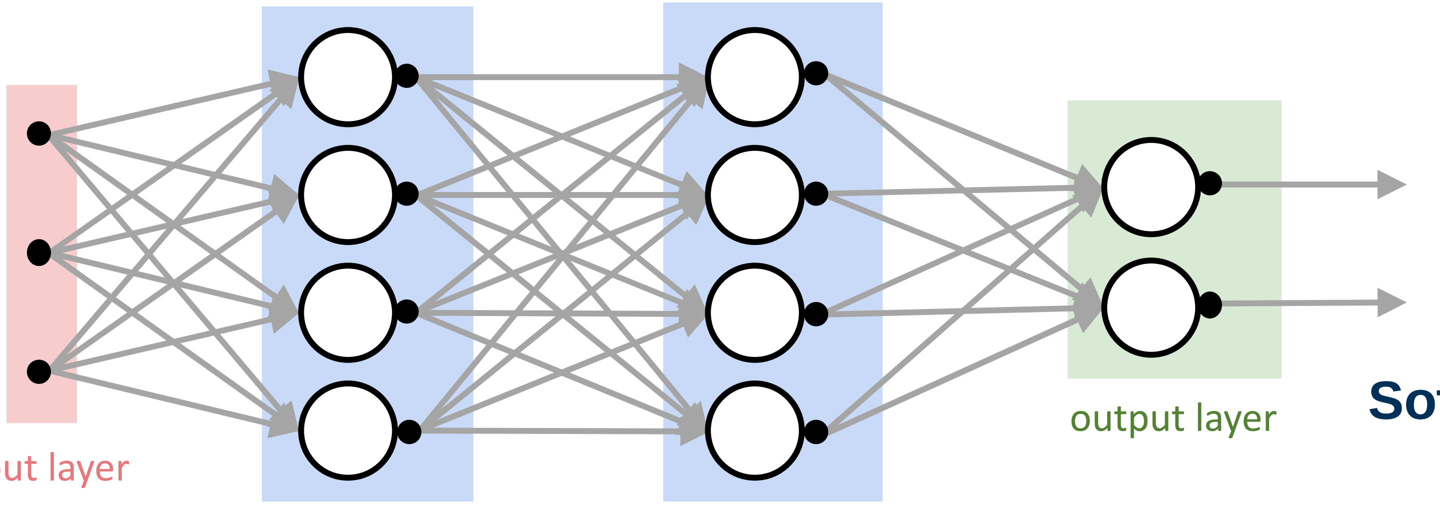
In this Lecture..

- Introduction and Motivation
- Pre-Text Tasks
- Weak Labels
- Contrastive Learning
- Examples of Contrastive Learning

Review

Supervised Learning

Sample x_i



Assume the raw outputs $y_i = [y_{i1}, y_{i2}, y_{i3}] = [3.2, 5.1, -1.7]$

SoftMax activation:

$$f(y_{ij}) = \frac{e^{y_{ij}}}{\sum_k e^{y_{ik}}}$$

Assume **ground-truth class is class 1**

Raw outputs	exp	unnormalized probabilities	normalize	probabilities		One-hot ground-truth label
$\boldsymbol{y_{i1}}$= **3.2**		**24.5**		$\boldsymbol{f(y_{i1})}$ = **0.13**		**1.00**
y_{i2}= 5.1	⇒	164.0	⇒	$f(y_{i2})$ = 0.87	⇒ Loss ⇐	0.00
y_{i3}= -1.7		0.18		$f(y_{i3})$ = 0.00		0.00

Question

Can we always assume that we have enough labeled data for good performance?

Why would we not have fully labeled data?

- Expense interms of Cost and Time
- Requirements of trained experts
- Data privacy laws
- Unreleased datasets

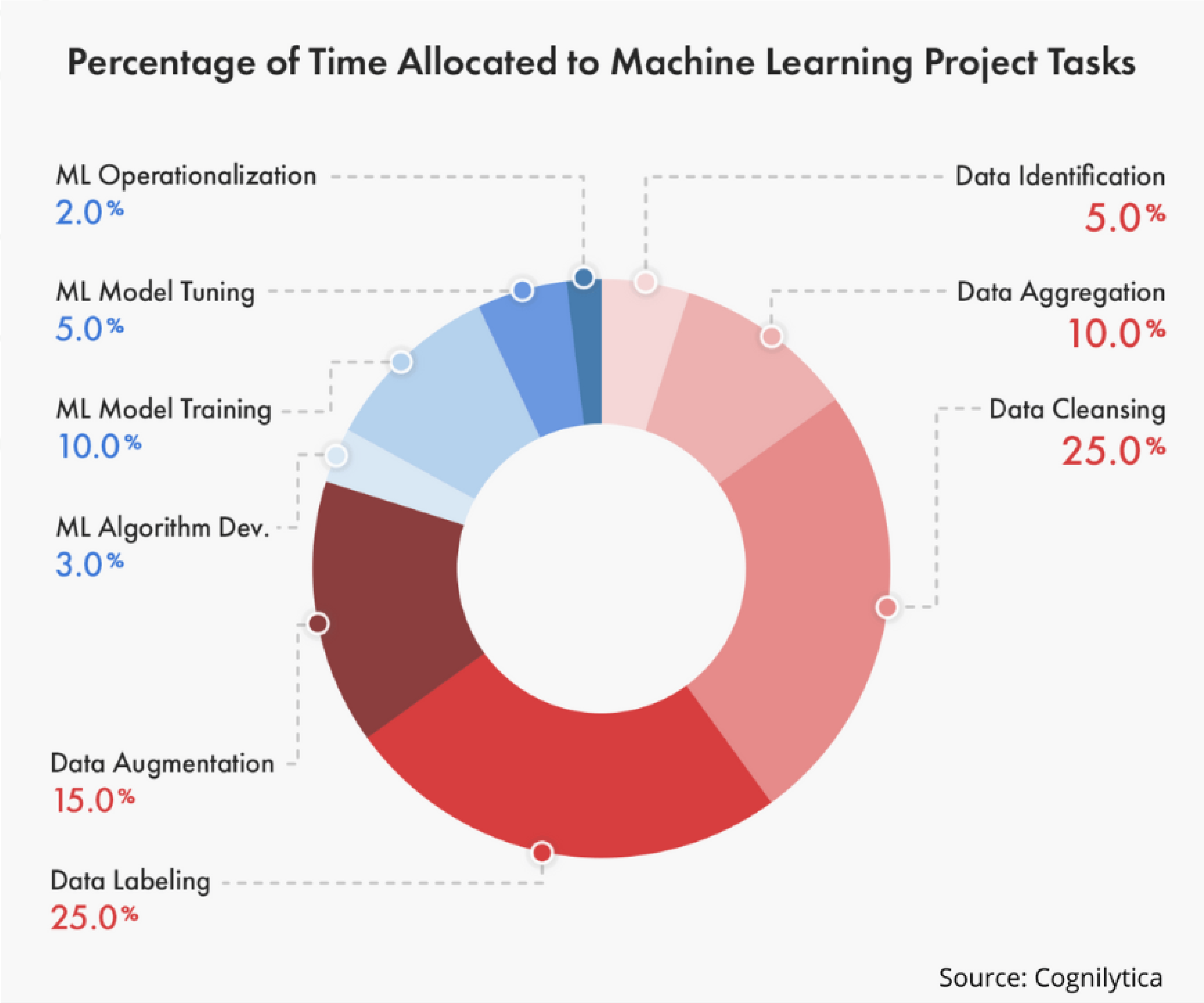
Introduction

Lack of Data: Cost

Labeling large datasets is a billion dollar industry!

Introduction

Lack of Data: Time



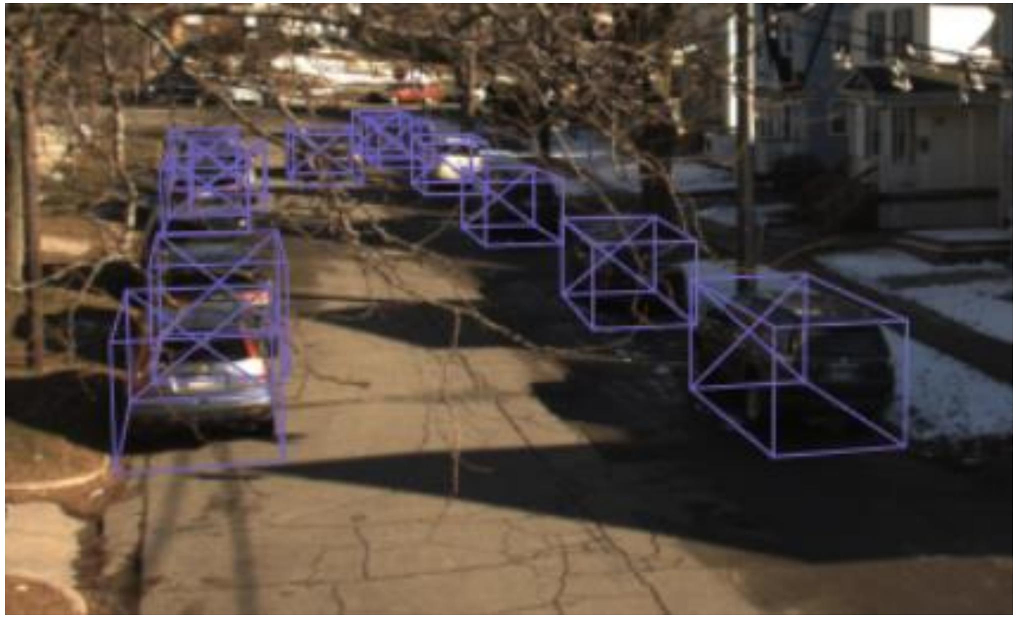
A large part of machine learning work is dedicated to just dealing with the data

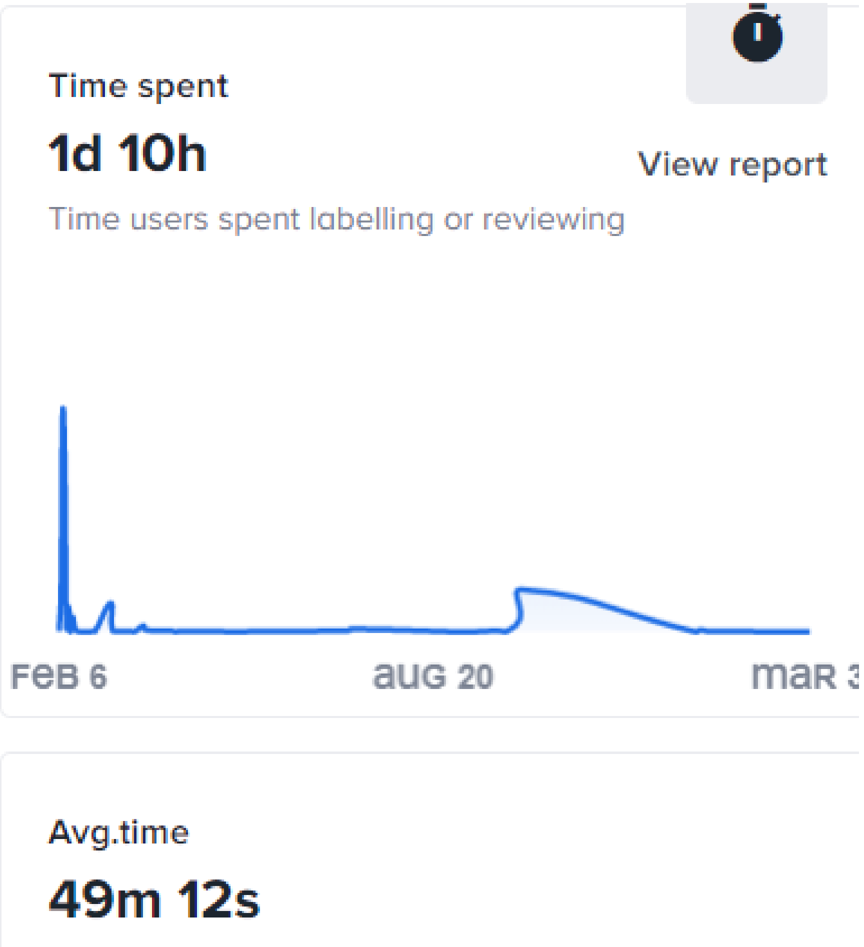
Introduction

Lack of Data: Real-world Example

Sequence

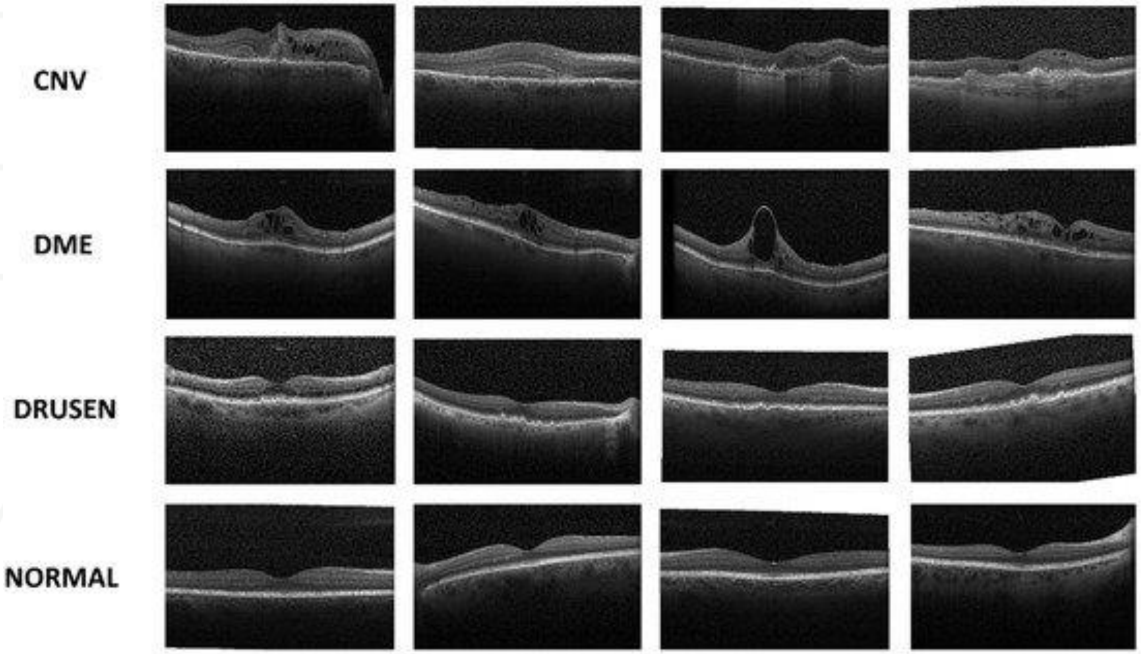
- 578 frames
- Labels for object detection, lidar, agent interactions

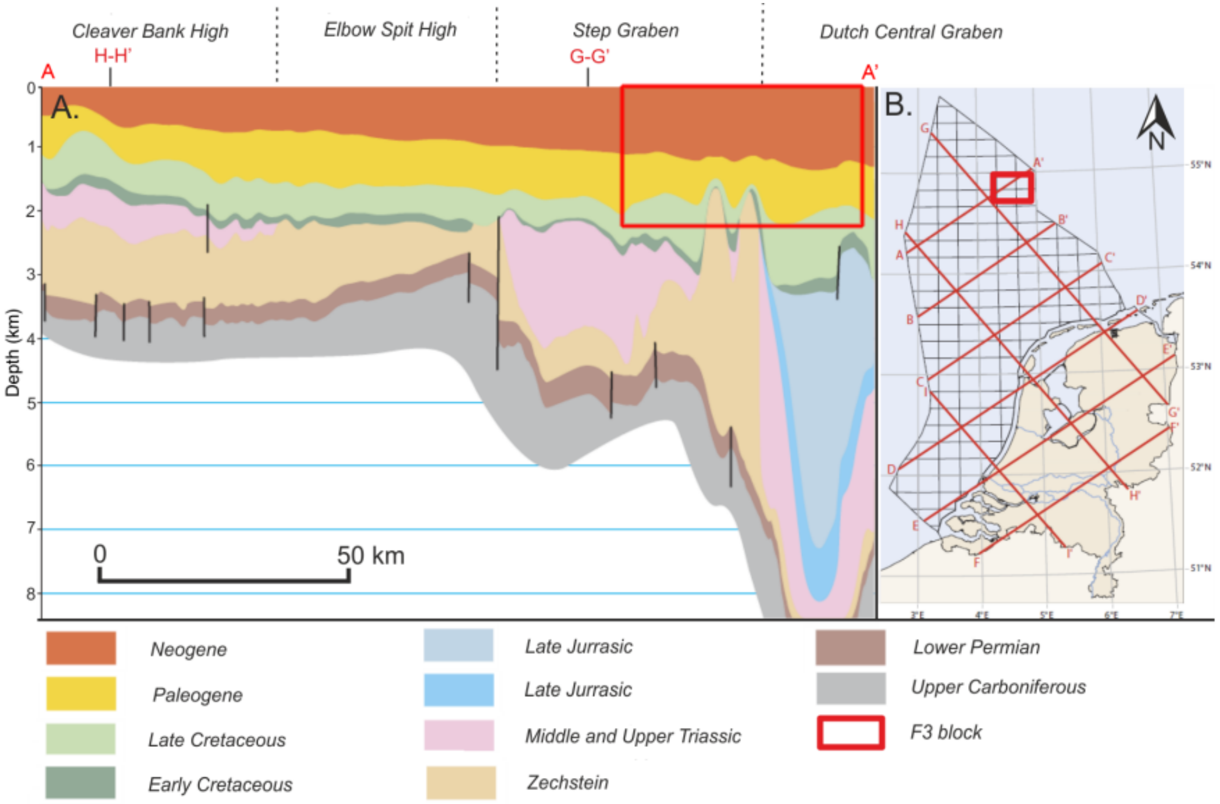




Kokilepersaud, Kiran, et al. "FOCAL: A Cost-Aware Video Dataset for Active Learning." *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023.

Lack of Data: Requirement of Trained Experts



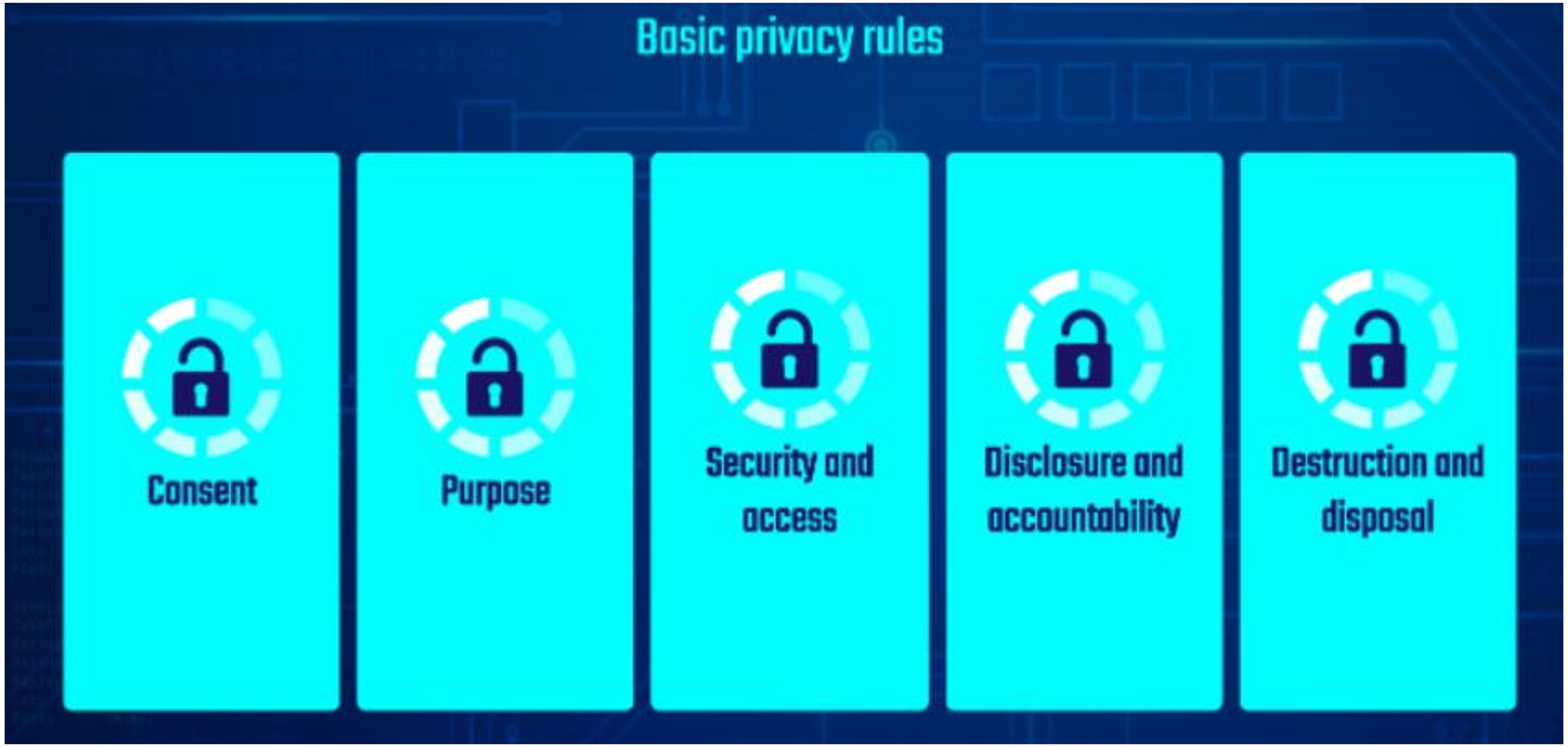


Medical labeling requires doctors/radiologists!

Geophysicists are required for seismic interpretation!

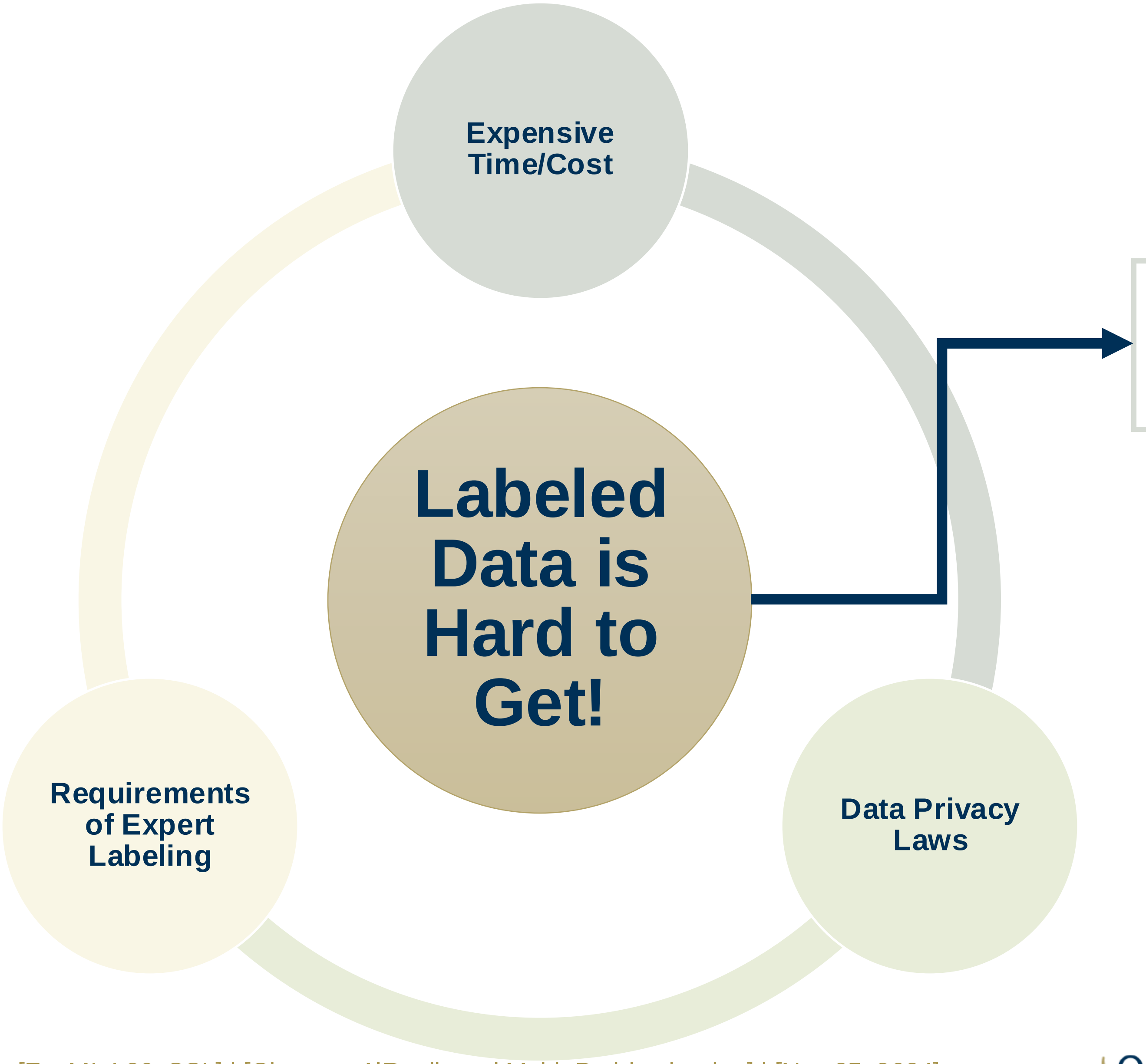
Introduction

Lack of Data: Privacy



Organizations such as HIPPA enforce strict regulations on data sharing.

Efficient Learning
Goal



Can we train a deep network without needing labels?

Efficient Learning

Approaches

Unsupervised	Self-Supervised	Weakly-Supervised	Semi-Supervised	Fully-Supervised
• no labeled data required	• create artificial labels for training base architecture • A small set of true labels used to train only the last layer	• weakly-labeled training data • Weak labels are noisy and less informative, but much easier to obtain	• a subset of the training data is strongly-labeled • the rest is not labeled	• strongly-labeled training data

More supervision

Overview

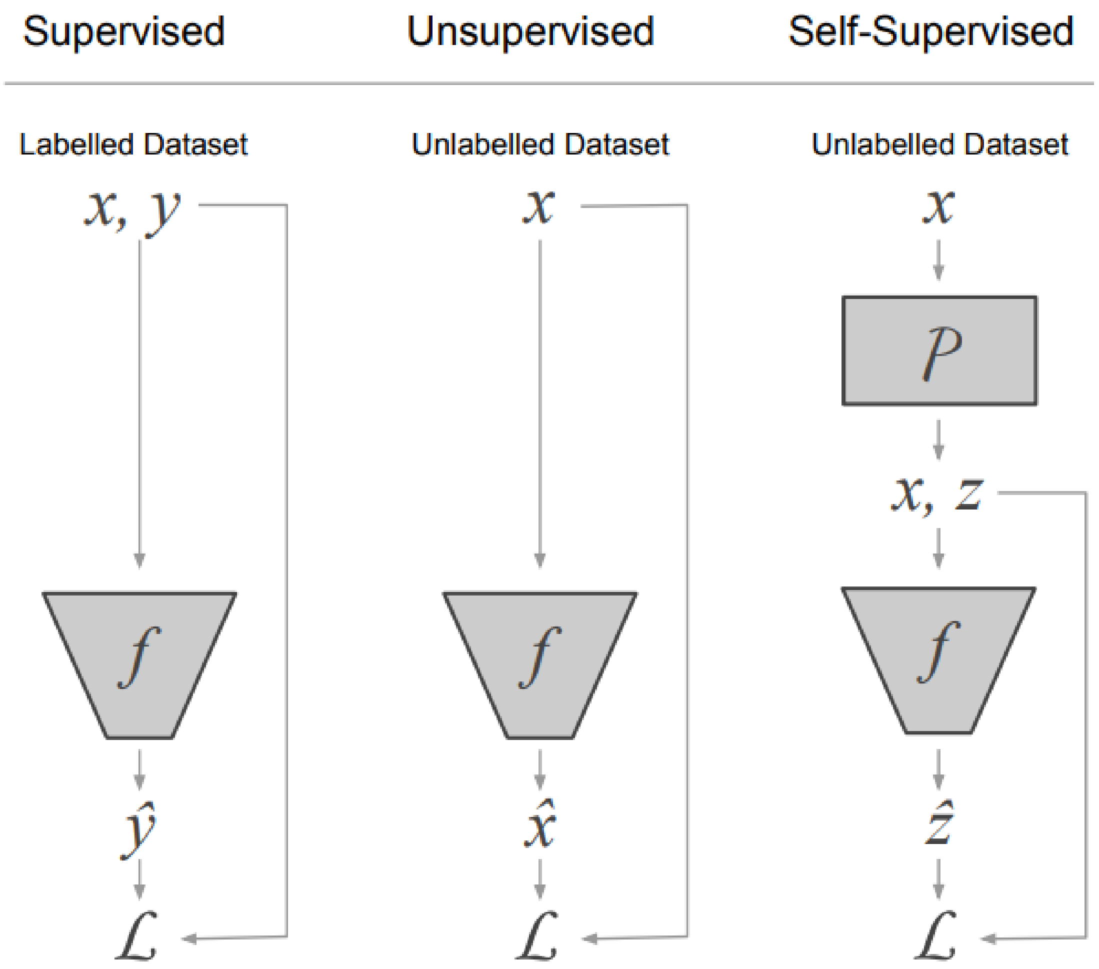
In this Lecture..

- Introduction and Motivation
- Pre-Text Tasks
- Weak Labels
- Contrastive Learning
- Examples of Contrastive Learning

Self-Supervised Learning

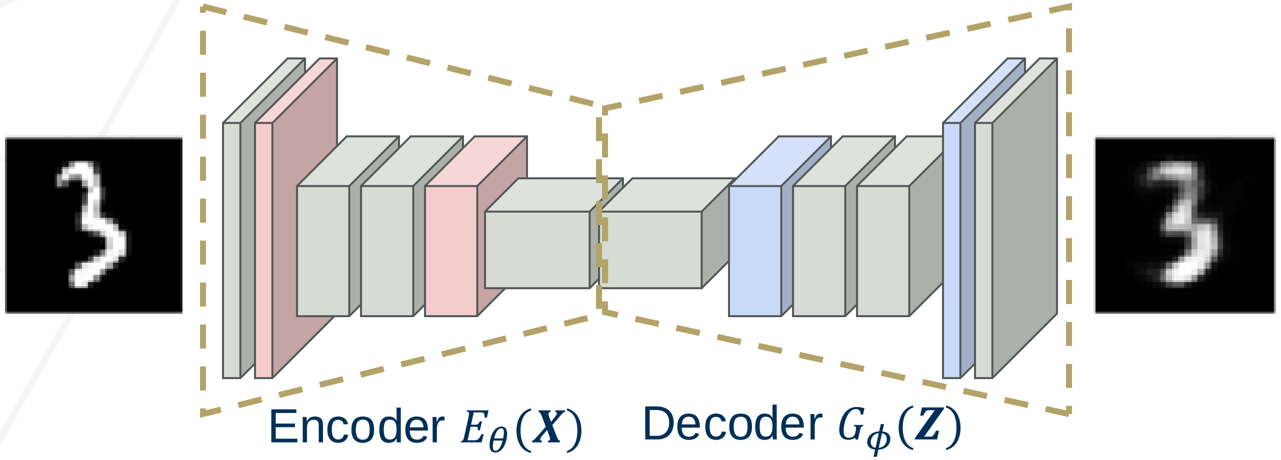
Pre-text tasks

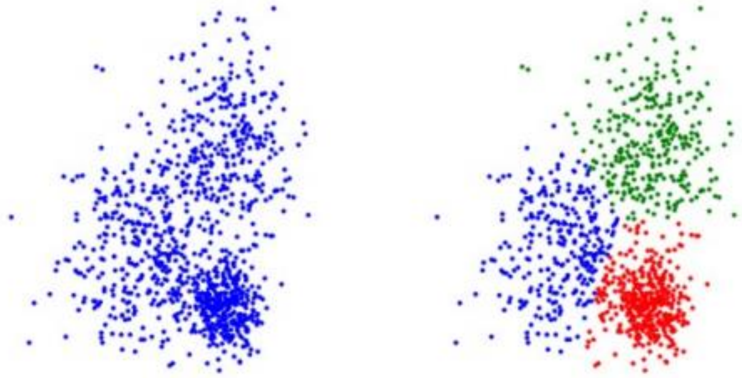
- Type of unsupervised learning
- Primary difference is the introduction of a “**pre-text task.**”
- The pre-text task generates pseudo-labels that are used to train a network.



Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2021). Self-Supervised Representation Learning: Introduction, Advances and Challenges. *arXiv preprint arXiv:2110.09327*.

Self-Supervised Learning
Self-supervised vs unsupervised





Auto-encoders
- No pre-text task

Clustering
- No pre-text task
- Doesn't require deep learning setup

"Pre-text Tasks" are what differentiate self-supervision from other types of unsupervised learning

Self-Supervision

Generate Pseudo Labels

Identify Labeled and Unlabeled Data

Unlabeled Data
$(x_1 \ldots x_N)$

Labeled Data
$(x_1 \ldots x_M)$, $(y_1 \ \ldots y_M)$

1. Generate pseudo-labels with some pre-text task P

Unlabeled Data
$(x_1 \ldots x_N)$

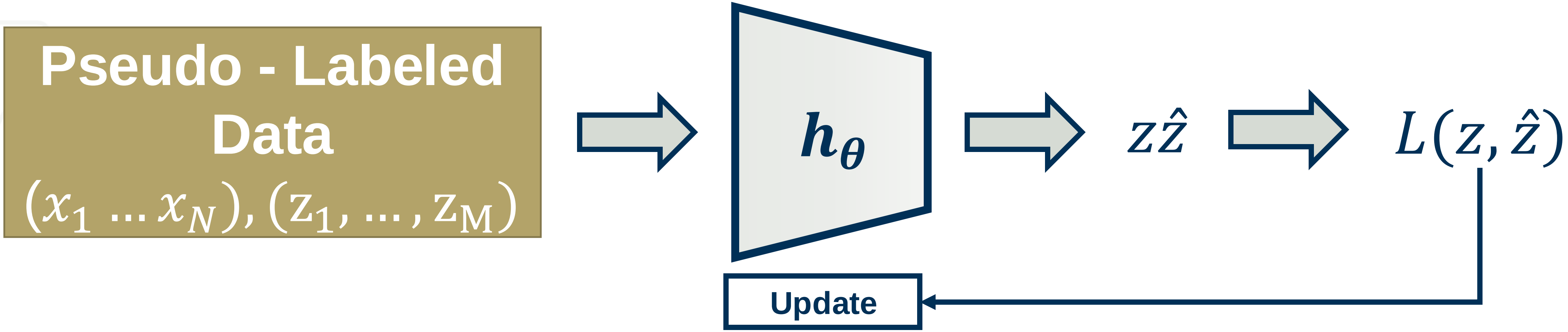
$\Rightarrow$ P $\Rightarrow$

Pseudo - Labeled Data
$(x_1 \ldots x_N), (\mathrm{z}_1, \ldots, \mathrm{z}_\mathrm{M})$

Self-Supervision

Utilize Pseudo Labels to Learn Pre-text Tasks

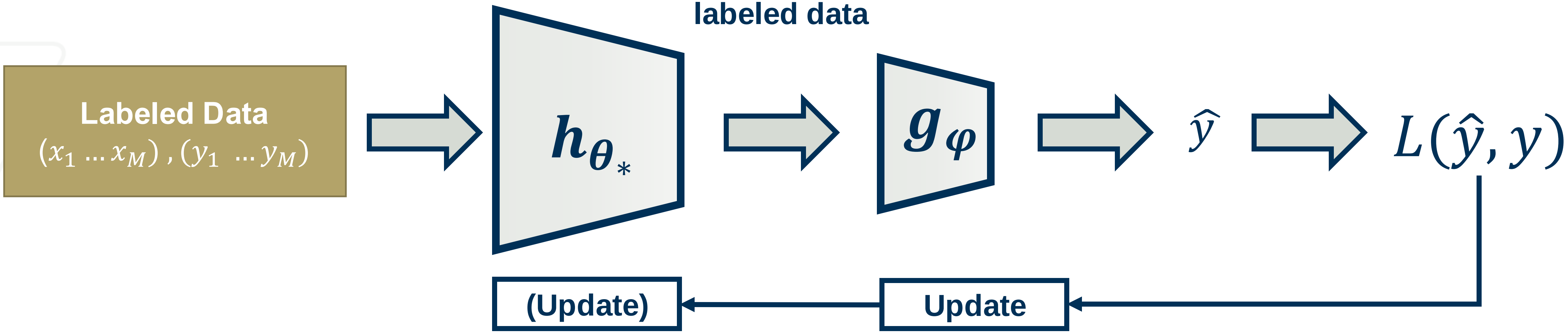
2. Self-supervised Pre-training of network h_θ on pseudo-labels



Self-Supervision

Utilize Pre-text Learned Model

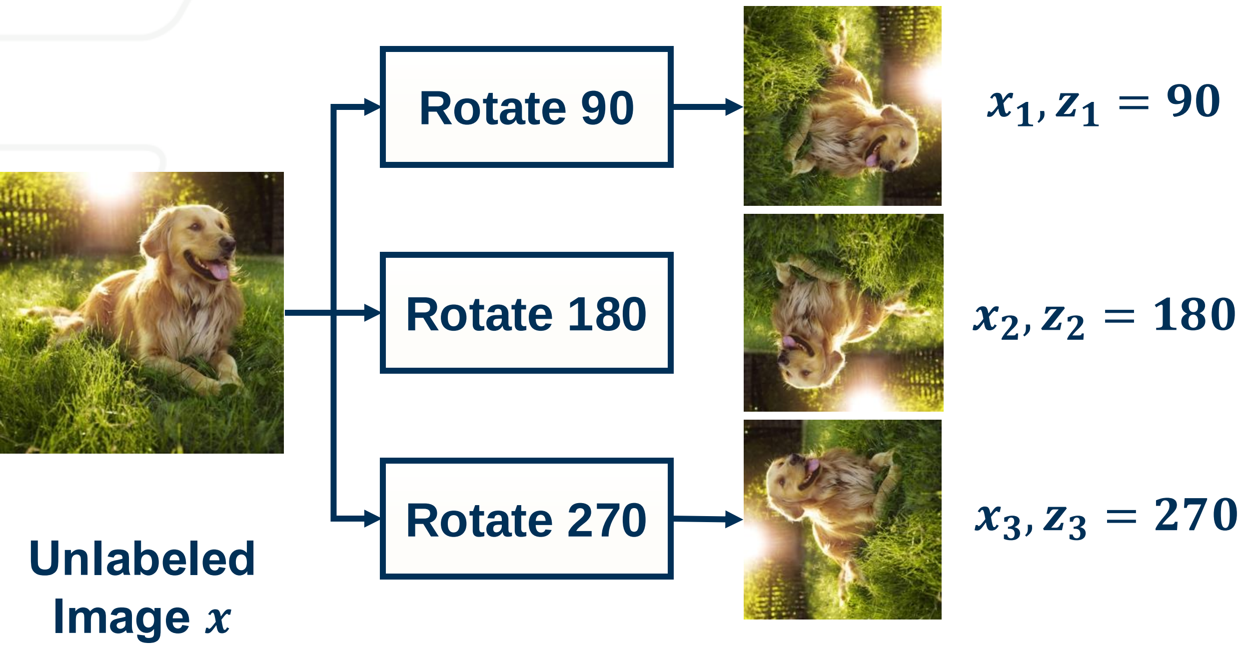
3. Use previously trained network h_{θ_*} and a task adaptation network g_{φ} to train with the labeled data



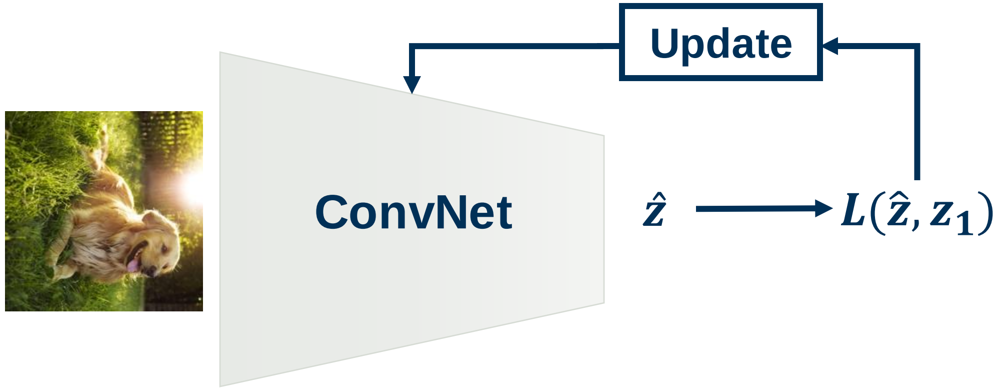
Self-Supervision

Example Training Process

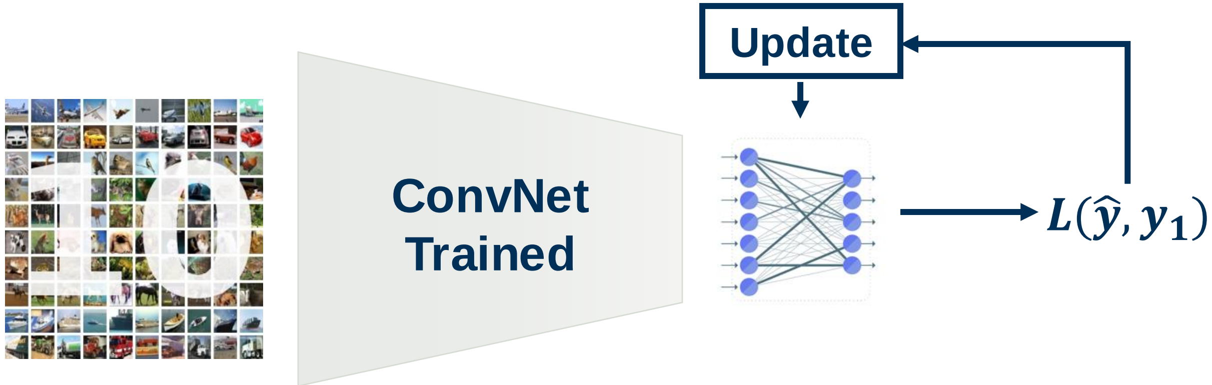
Step 1: Generate pseudo-labels via image rotations



Step 2: Network learns to predict angle image is rotated



Step 3: Attach linear layer and train to classify labels (y) on labeled dataset

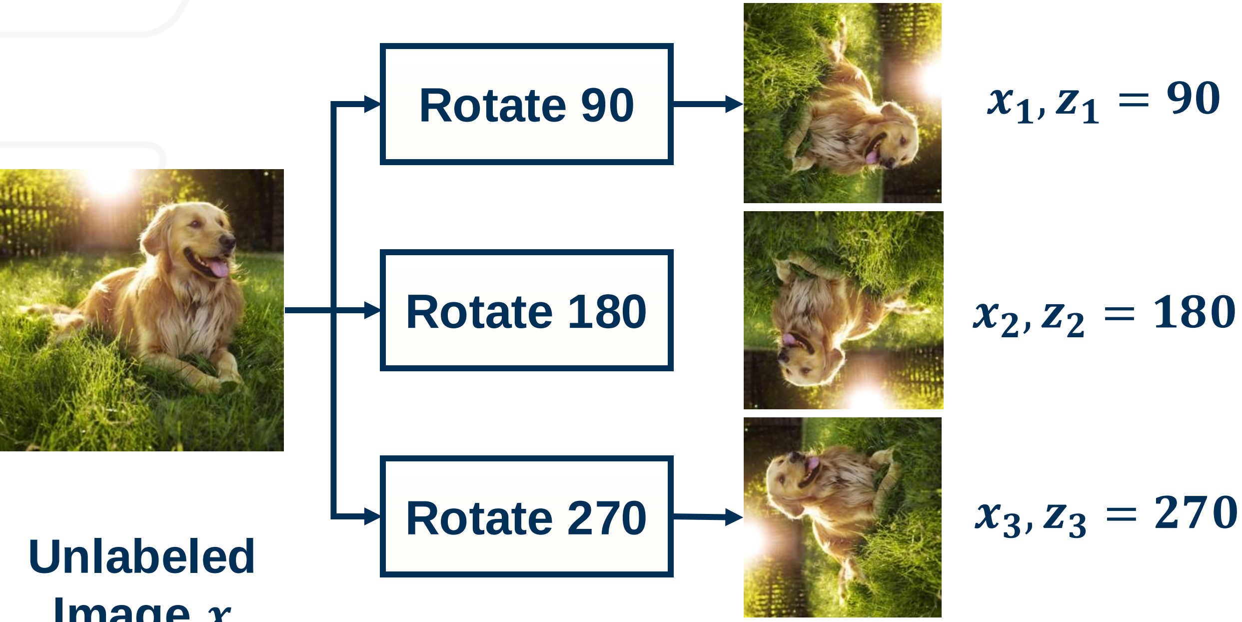


Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Self-Supervision

Motivation

Step 1: Generate pseudo-labels via image rotations



Unlabeled Image x

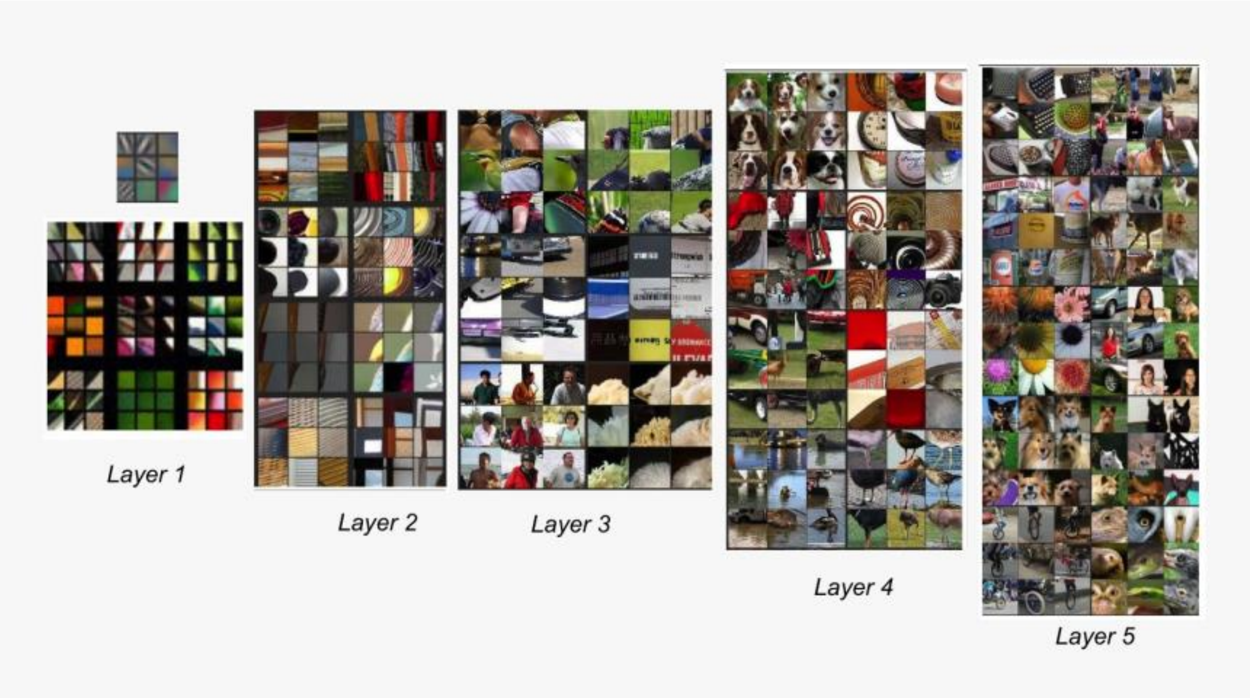
Learning pre-text task will allow network to learn relevant features without needing explicit labels!

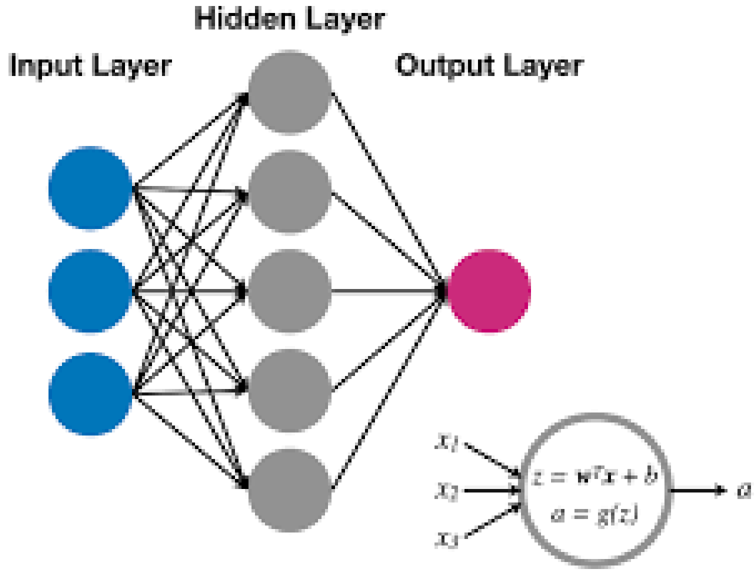
Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Self-Supervision

Motivation

Learning features and learning the task of interest (e.g. classification) does not have to happen at the same time





Learning pre-text task will allow network to learn relevant features without needing explicit labels.

General learnt features are usable for any target task: Classification, Object Detection, Semantic Segmentation

Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Differences in self-supervision are based on the type of pre-text task that is defined

Transformation Prediction

- Pre-text task performs some transformation on data and tasks model with trying to learn nature of transformation.

Masked Prediction

- Pre-text task removes some part of the data and the model is tasked with trying to predict what was removed.

Deep Clustering

- Identify clusters of features and iteratively assign pseudo-labels to train model.

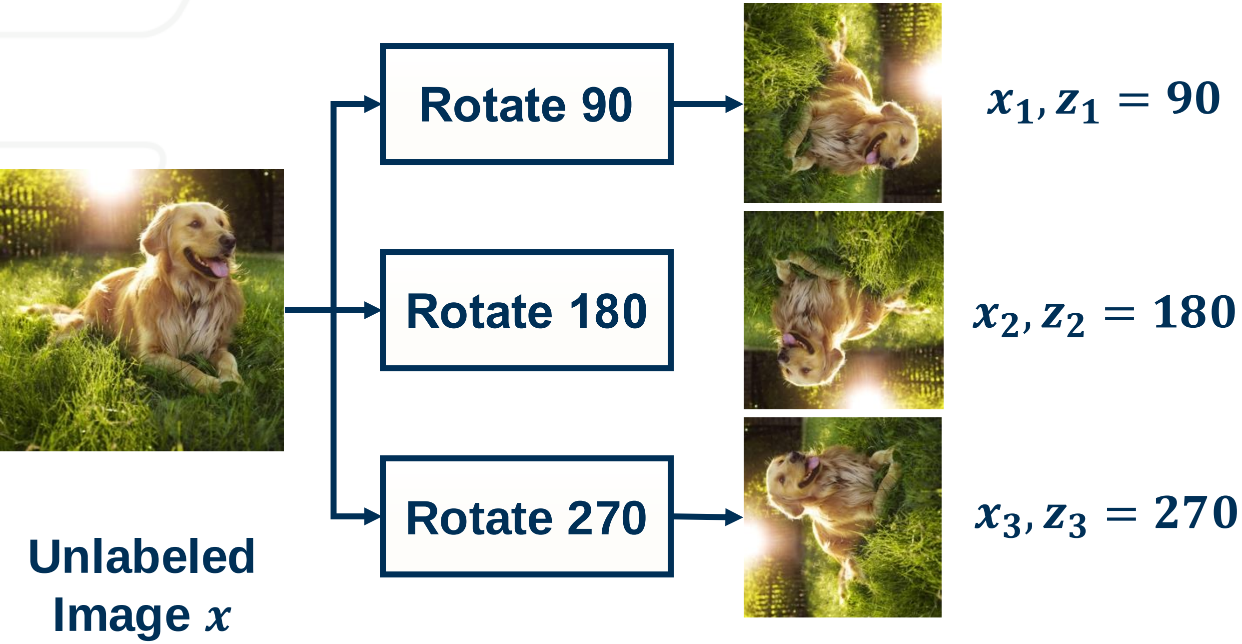
Contrastive Learning

- Pre-text task identifies positive and negative pairs of data and the model is tasked with learning similarities to discriminate between positive and negatives.

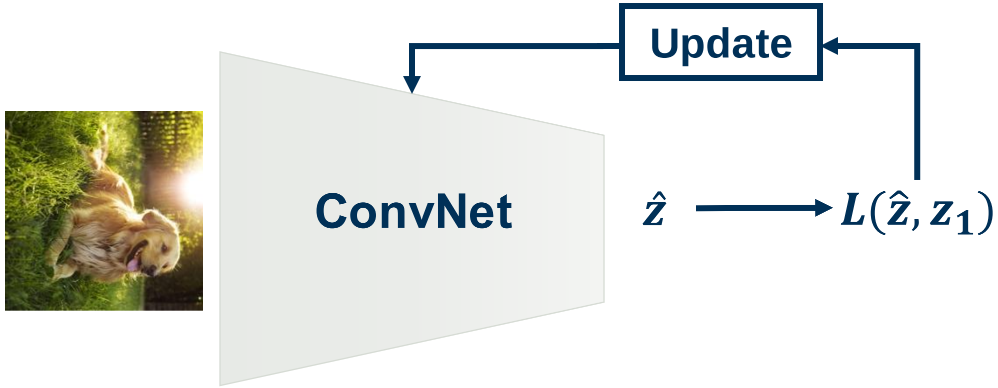
Self-Supervision

Transformation Prediction

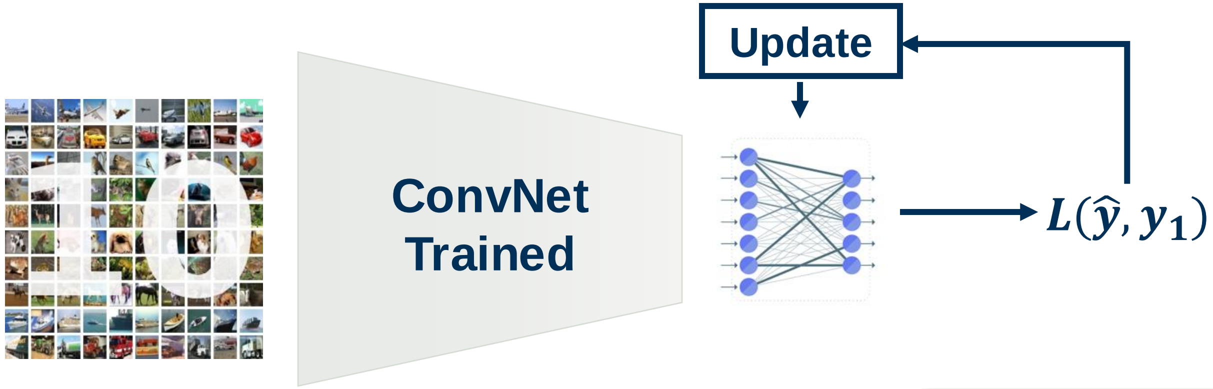
Step 1: Generate pseudo-labels via image rotations



Step 2: Network learns to predict angle image is rotated



Step 3: Attach linear layer and train to classify labels (y) on labeled dataset

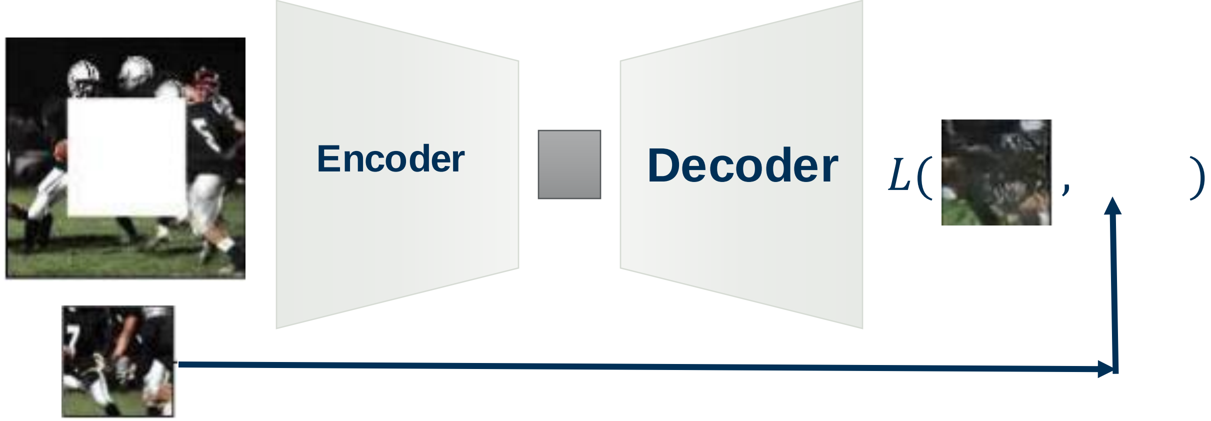


Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

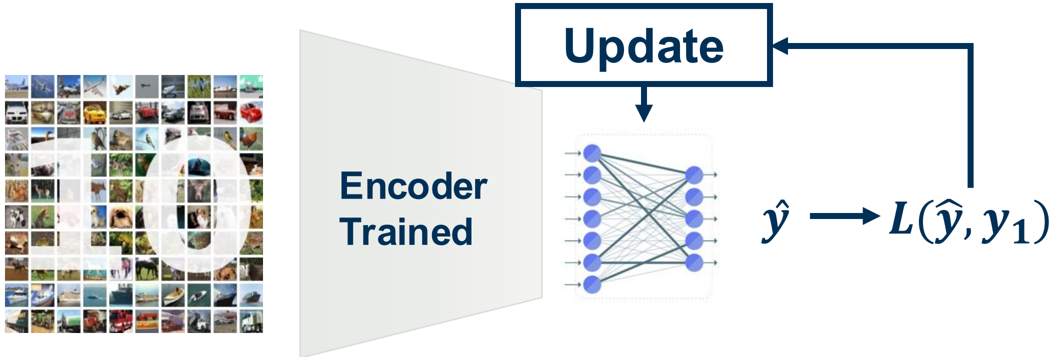
Self-Supervision
Masked Prediction

1. Generate pseudo-labels via Masking

2. Model learns to produce missing region



3. Encoder trained with labels (y) from task of interest

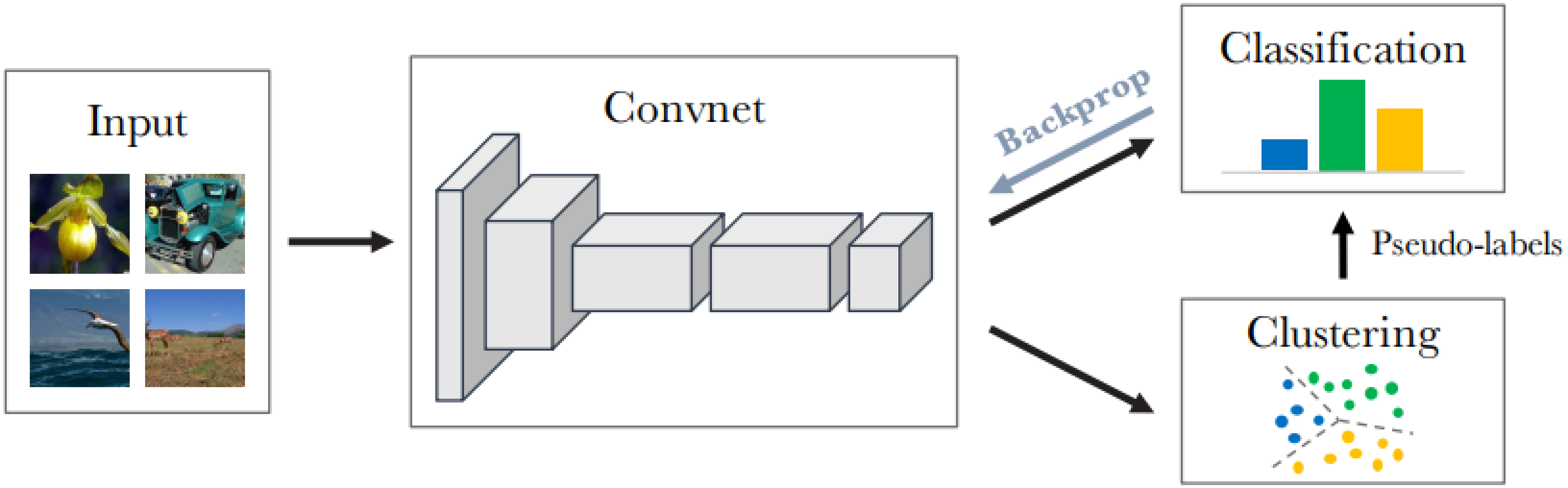


Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2536-2544).

OLIVES @GeorgiaTech
Georgia Tech

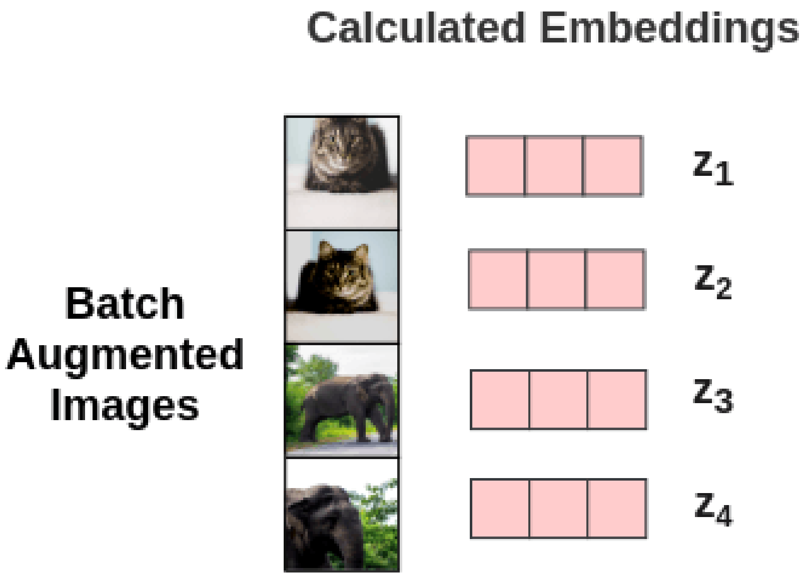
Self-Supervision
Deep Clustering

- Input Batch of Unlabeled Images.
- Cluster features of output to generate pseudo-labels.
- Use pseudo-labels to generate a loss to train the model.
- Take model and fine-tune on some target task



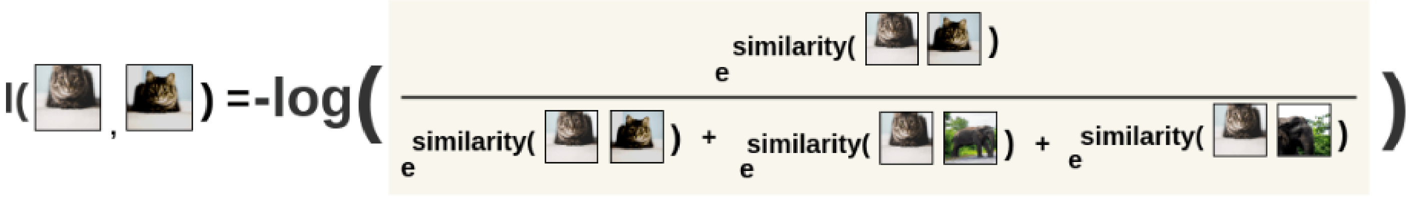
Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

The Pseudo-labels are used to create positive-negative pairs within each batch



Contrastive loss on embeddings

$$l(i,j) = -log \frac{exp(s_{i,j})}{\sum_{k=1}^{2N} l_{[k!=i]} exp(s_{i,k})}$$



Note: The positive pairs are only the augmentations and negative pairs are all other images in the batch

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).

Contrastive Learning

Sim-CLR Framework

The Pseudo-labels are used to create positive-negative pairs within each batch

1. Generate similar pairs in a starting batch by augmenting each image.

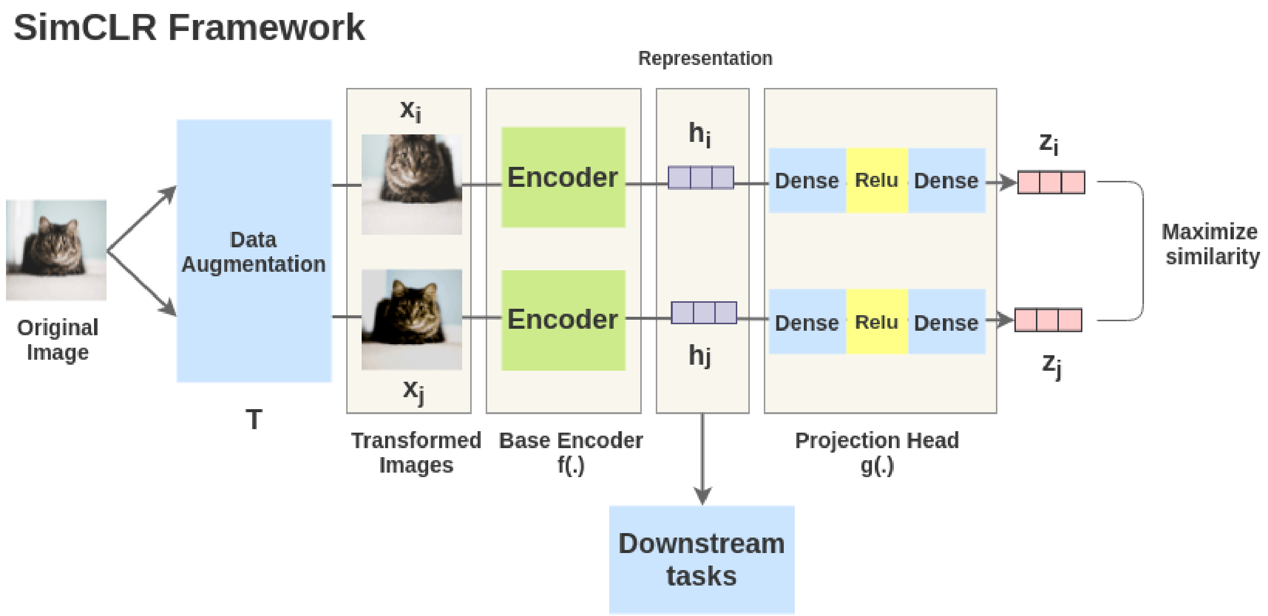
2. Pass image and its generated pair into an encoder to get a lower dimensional representation (h_i and h_j).

3. Compress representation further with projection head to generate an embedding (z_i and z_j) for each image.

4. Calculate similarity of generated embeddings with cosine similarity metric.

5. Calculate the noise contrastive estimation loss using cosine similarity on each pair of images.

6. Compute loss over all pairs in batch and take average.

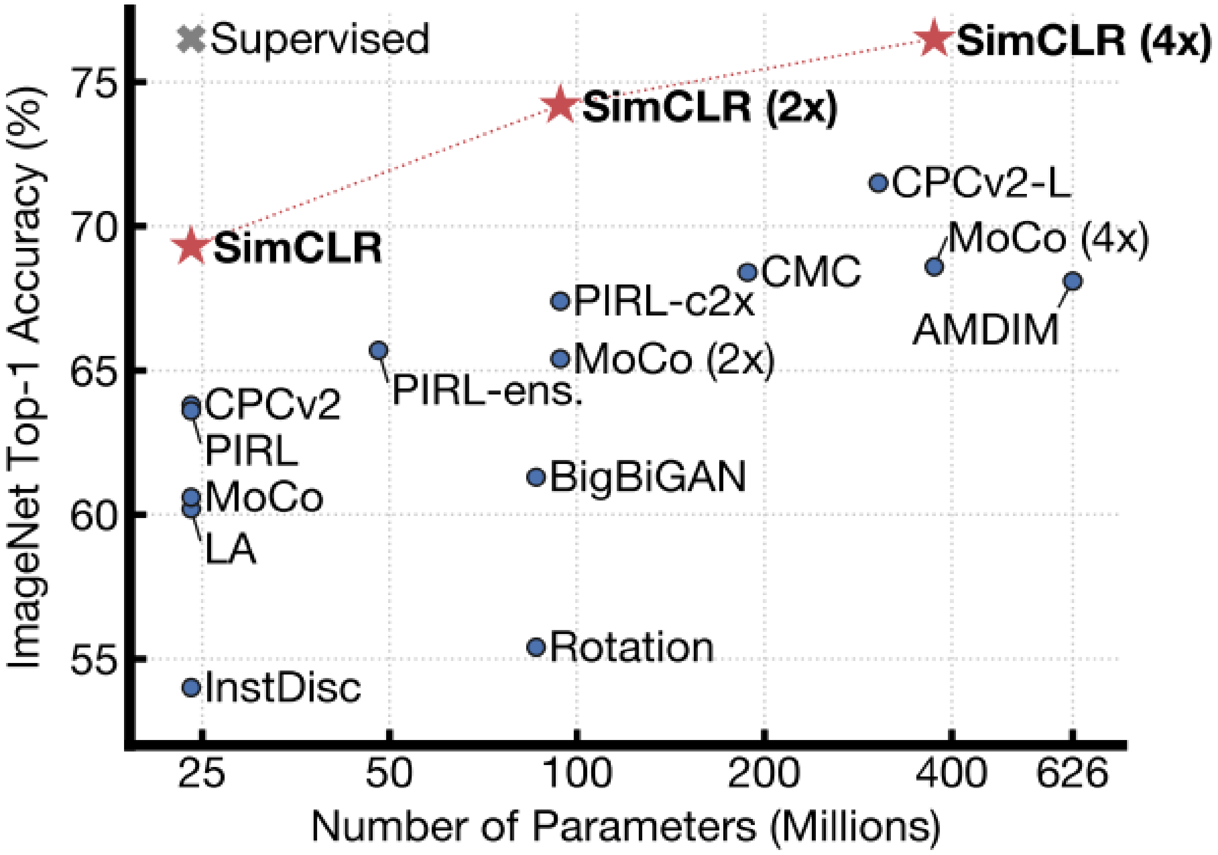


Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).

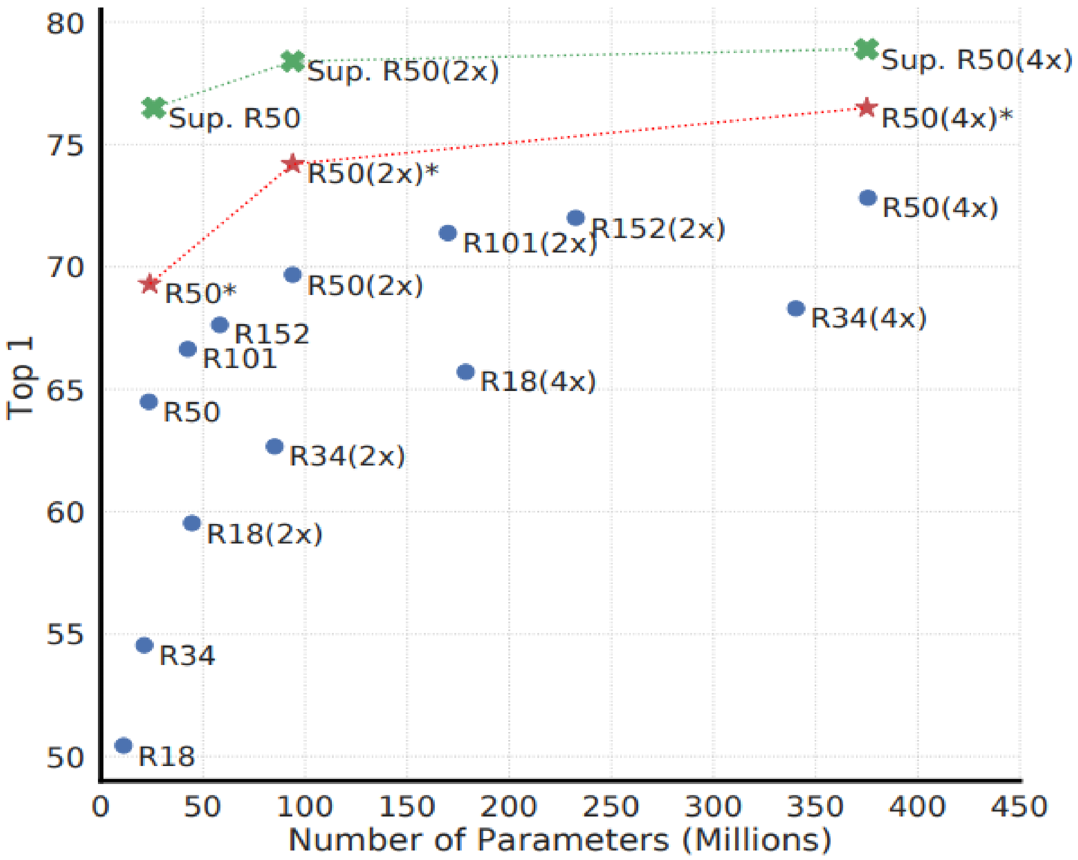
Contrastive Learning

Contrastive Learning vs Supervised Learning

Performance vs Models



Performance vs Parameters



Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).

OLIVES @GeorgiaTech

Georgia Tech

Contrastive Learning other than Sim-CLR

What differentiates other Contrastive Learning methods from Sim-CLR?

The way that similar pairs (positives) and dissimilar pairs (negatives) are generated.

Paper	Short description	Topics of contribution
Becker and Hinton [8]	Maximise MI between two views	Foundation
Bromley et al. [11]	Siamese network in metric learning setting	Foundation
Chopra, Hadsell, and LeCun [20]	Learn similarity metric with contrastive pair loss	Energy-based loss, Application
Hadsell, Chopra, and LeCun [39]	Learn invariant representation from pair loss	Energy-based loss, Application
Weinberger, Blitzer, and Saul [108]	Learn distance metric with triplet loss	Energy-based loss
Collobert and Weston [21]	Learn language model with triplet loss	Application
Chechik et al. [15]	Learn image retrieval model with triplet loss	Application
Noise Contrastive Estimation [38]	Introduce NCE, a general methods to learn unnormalised probabilistic model	Probabilistic loss
Mnih and Teh [71]	Learn language model with NCE-based loss	Application
Mikolov et al. [68]	Learn word embedding with Negative Sampling (NEG), a modified version of NCE	Probabilistic loss, Application
Wang et al. [105]	Learn fine-grained image similarity using deep network and triplet loss	Application
Wang and Gupta [107]	Use video's sequential coherence to learn unsupervised video representation	Similarity, Application
Lifted-structure loss [75]	Extend triplet loss to multiple positive and negative pairs per query	Energy-based loss
N-pair loss [92]	Proposed non-parametric classification loss with multiple negative pairs per query	Probabilistic loss
Wu et al. [109]	Focus on the quality of negative samples through a distance-weighted margin loss	Similarity, Energy-based loss
Hermans, Beyer, and Leibe [45]	State the important of mining hard samples in triplet loss	Similarity
Wu et al. [110]	Self-supervised representation with instance discrimination Memory bank to holds keys for next epoch	Application Encoder
CPC [77]	Mutual Information with the contrastive loss Define similarity with past-future context-instance relationship	Mutual Information loss Similarity
DIM [46]	Evaluate multiple mutual information bound for the contrastive loss Global-local context-instance relationship	Mutual Information Loss Similarity
MoCo [43]	Use momentum encoder to store features to memory queue	Encoder
SimCLR [16]	Simplify and demonstrate large empirical improvement in instance discrimination task Focus on the use of separate heads	Application Transform heads
BYOL [34]	Learning similarity without negative samples	Loss

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).

OLIVES @GeorgiaTech

Georgia Tech

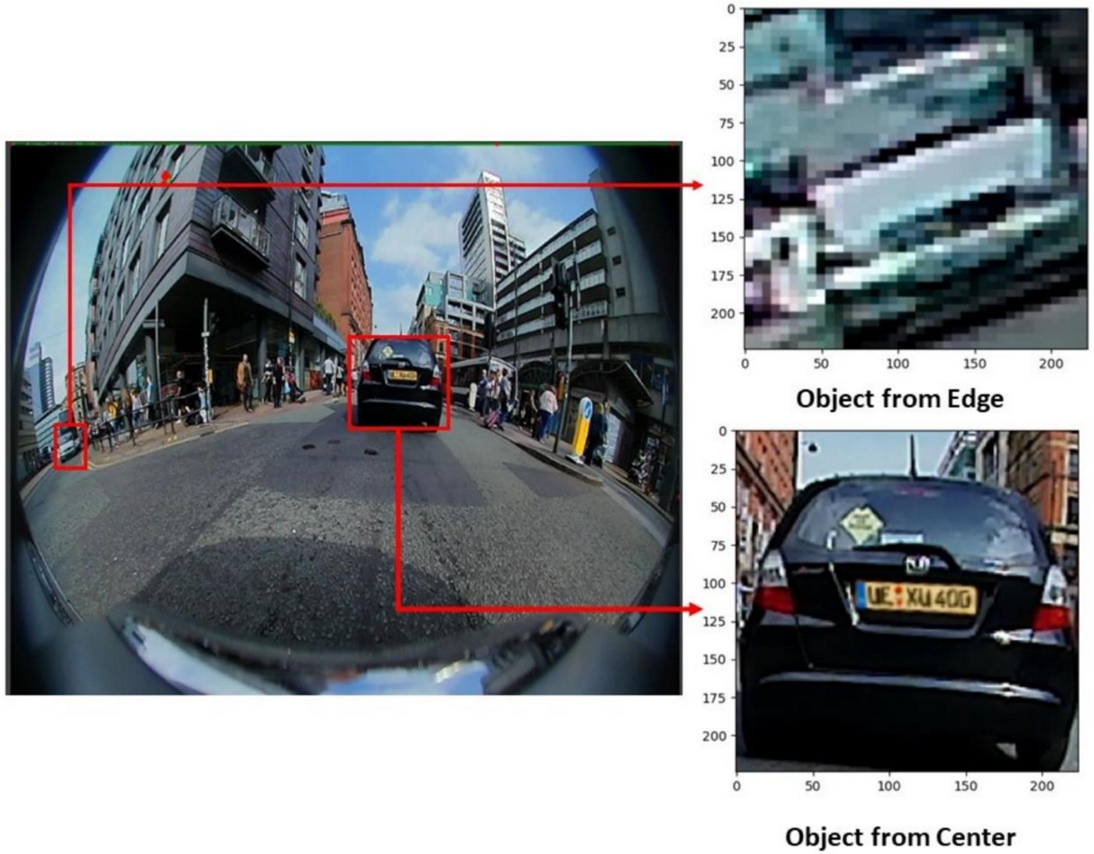
Example 1: Contrastive Learning for Fisheye Images

Positive-negative pairs in Fisheye Images

SCAN ME

Exploiting the Distortion-Semantic Interaction in Fisheye Data

Intuition: Regions within a fisheye image are their own class. Hence, any object within them are positives



Intuition for Loss 1: L_{Class}

All objects from labeled car (be it in the center or the edge) are positives and all other objects (be it in the center or the edge) are negatives

Intuition for Loss 2: $L_{RegionClass}$

All objects from the edge (be it a car, bike, pedestrian) are positives and objects from the center (be it a car, bike, pedestrian) are negatives

Kokilepersaud, Kiran, et al. "Exploiting the distortion-semantic interaction in fisheye data." *IEEE Open Journal of Signal Processing* (2023).

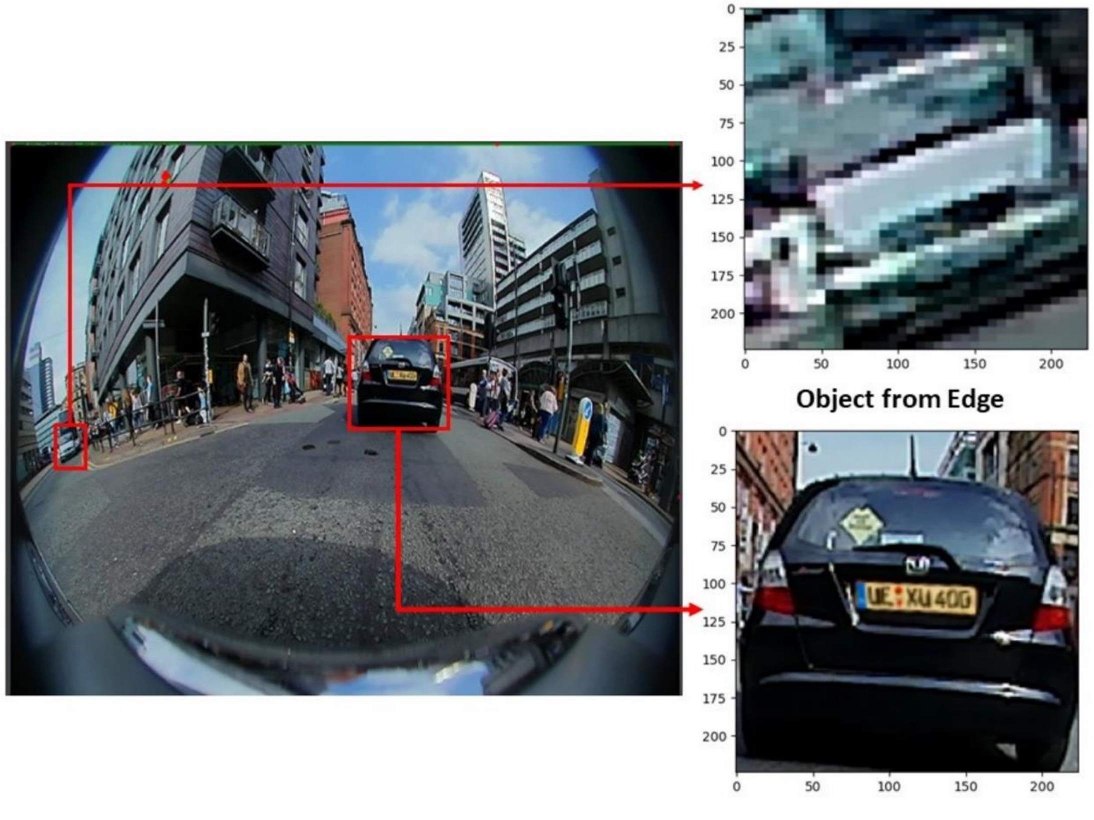
Example 1: Contrastive Learning for Fisheye Images

Positive-negative pairs in Fisheye Images

Exploiting the Distortion-Semantic Interaction in Fisheye Data

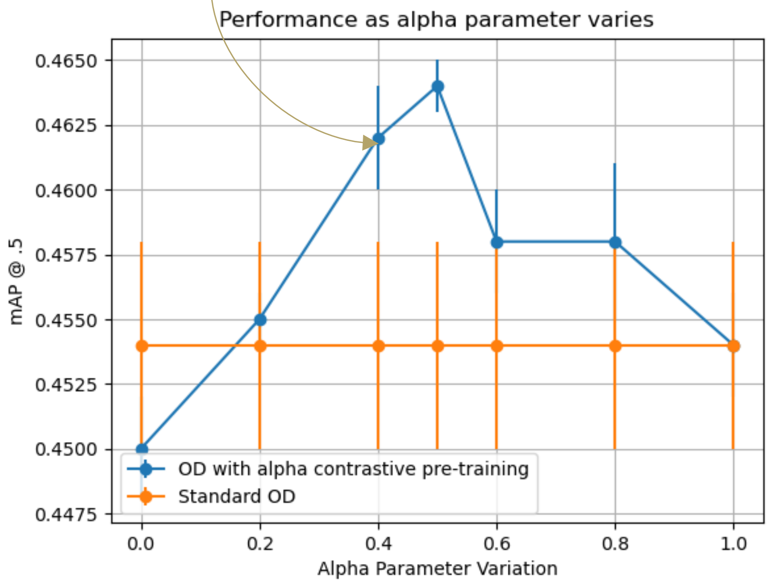
SCAN ME

Intuition: Regions within a fisheye image are their own class. Hence, any object within them are positives



$$\alpha L_{class} + (1-\alpha) L_{RegionClass}$$

α controls the level of unsupervised contrastive learning



Kokilepersaud, Kiran, et al. "Exploiting the distortion-semantic interaction in fisheye data." *IEEE Open Journal of Signal Processing* (2023).

OLIVES @GeorgiaTech

Georgia Tech

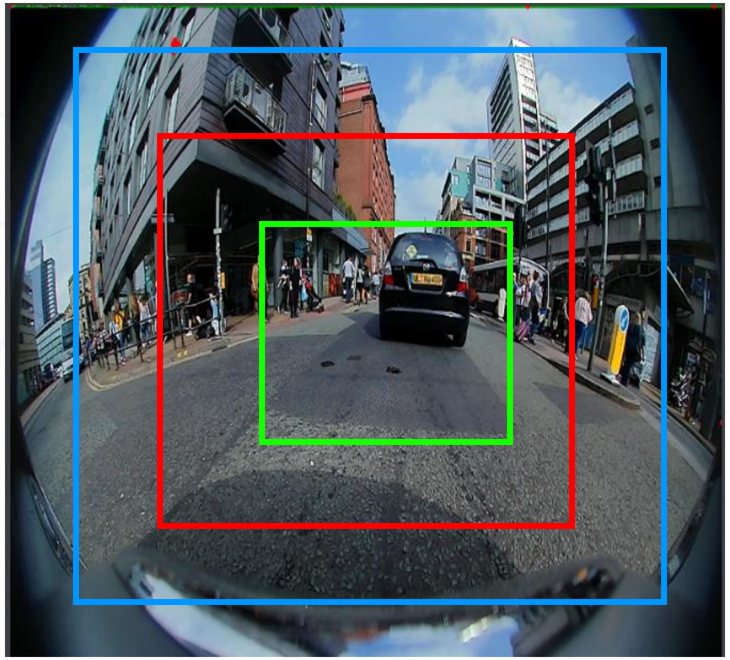
Example 1: Contrastive Learning for Fisheye Images

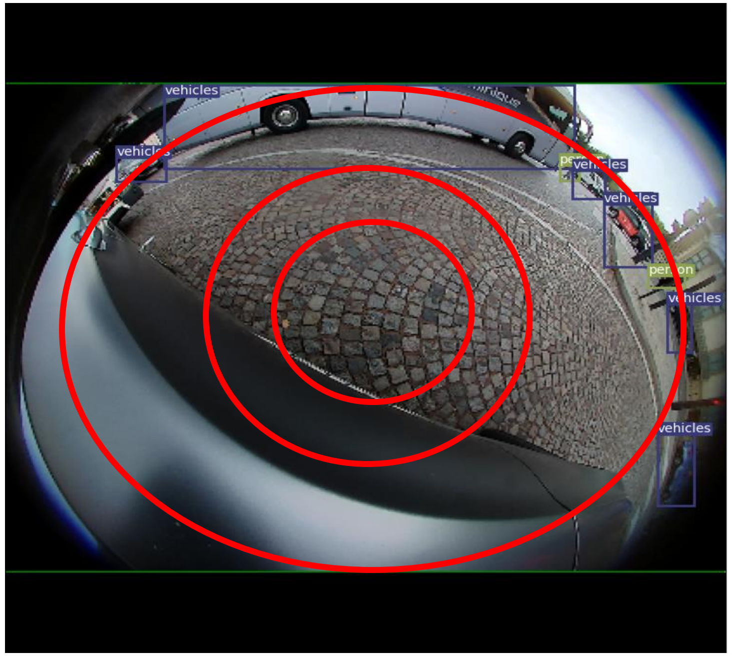
Positive-negative pairs in Fisheye Images

Exploiting the Distortion-Semantic Interaction in Fisheye Data

SCAN ME

Are there alternative ways of partitioning the regions?





Defining the positive-negative pairs is application dependent

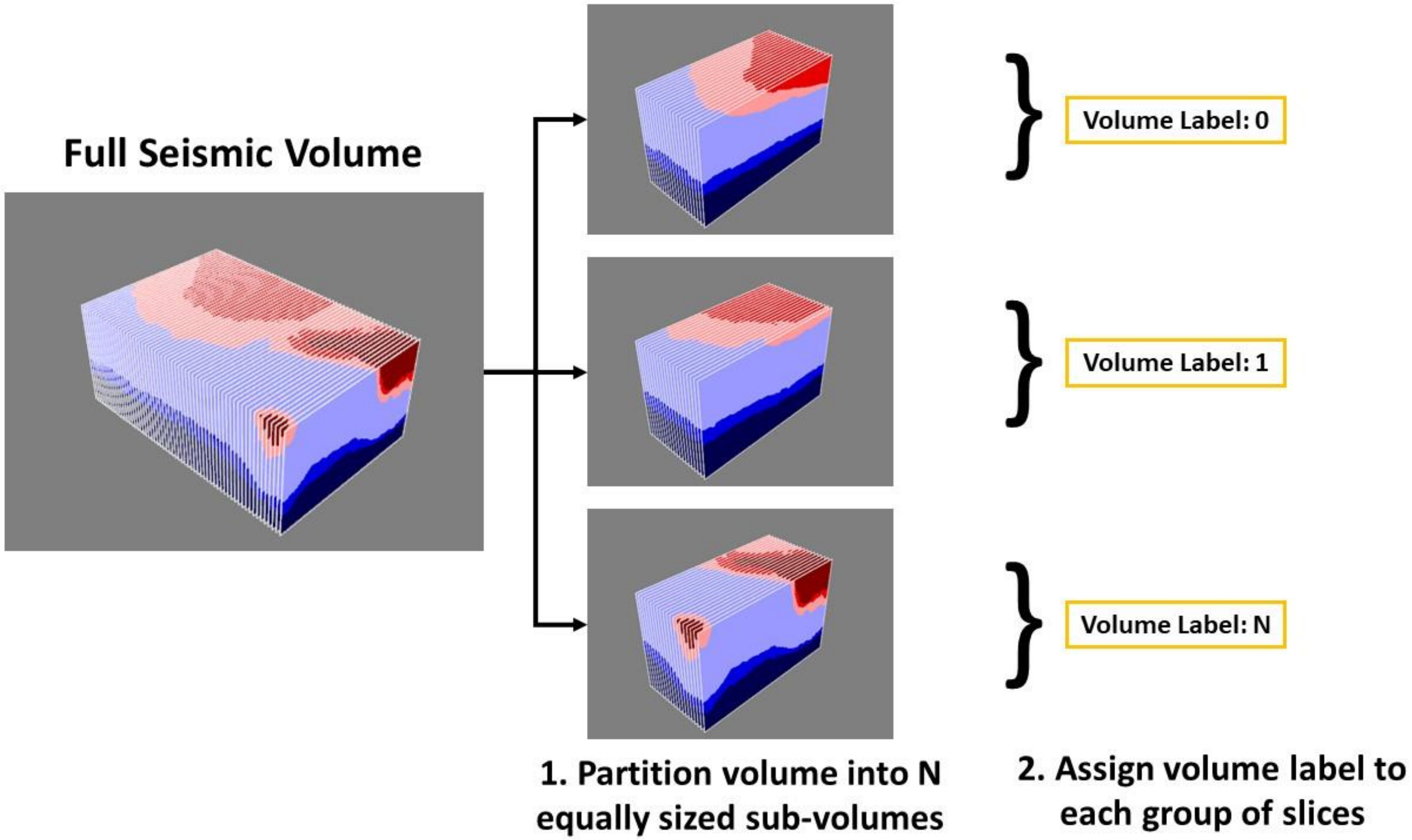
OLIVES @GeorgiaTech

Georgia Tech

Example 2: Contrastive Learning for Seismic Images

Positive-negative pairs in Seismic Images

Pre-text task is to classify volume label: Positive-negative pair is volume number

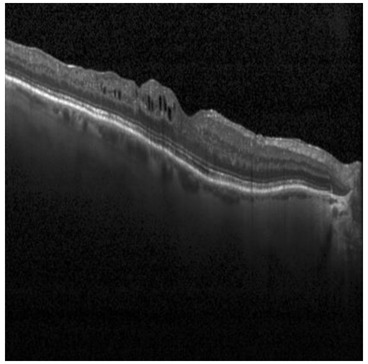


Kokilepersaud, Kiran, Mohit Prabhushankar, and Ghassan AlRegib. "Volumetric supervised contrastive learning for seismic semantic segmentation." *Second International Meeting for Applied Geoscience & Energy*. Society of Exploration Geophysicists and American Association of Petroleum Geologists, 2022.

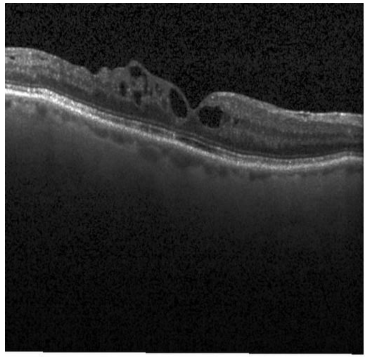
Example 3: Contrastive Learning for Medical Images

Positive-negative pairs in Medical Images

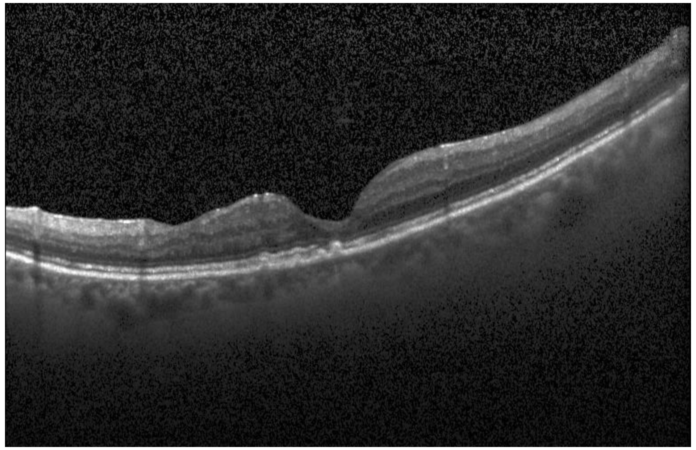
Pre-text task is to classify Patient ID/Clinical labels: Positive-negative pair is Patient ID/Clinical labels



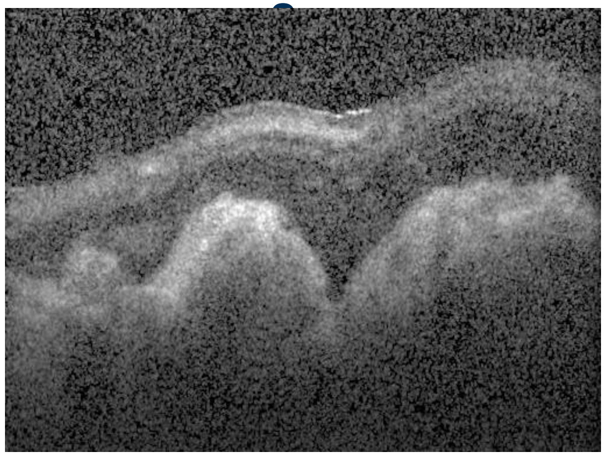
Patient 30521, Image



Patient 30521, Image



Patient 66861, Image 1



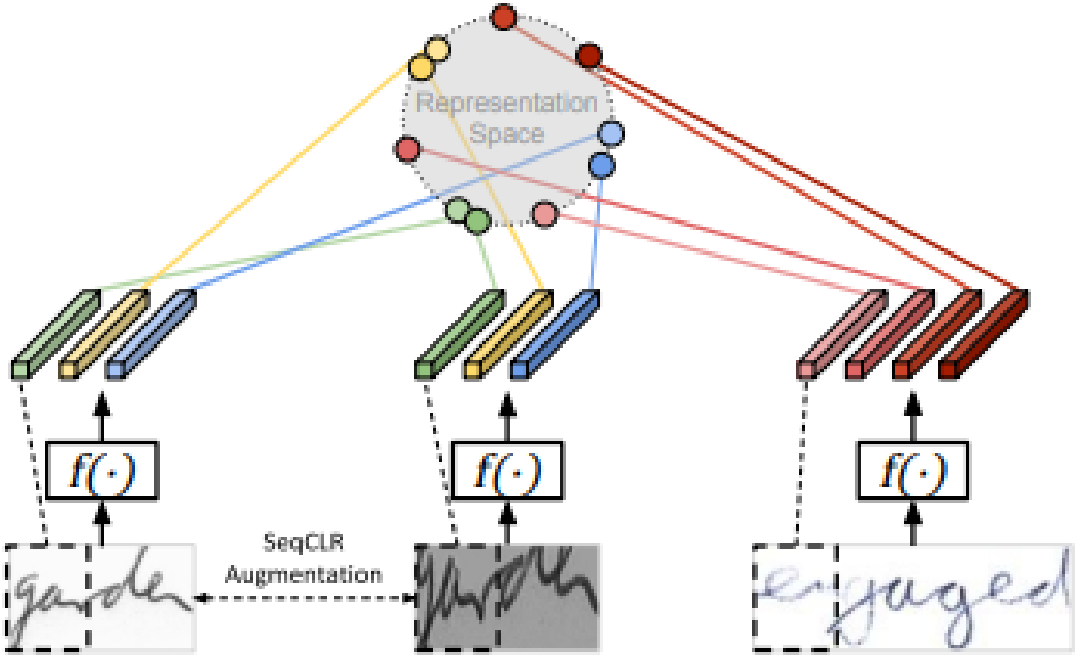
Patient 53264, Image 1

Kokilepersaud, Kiran, et al. "Clinically labeled contrastive learning for oct biomarker classification." *IEEE Journal of Biomedical and Health Informatics* (2023).

Contrastive Learning

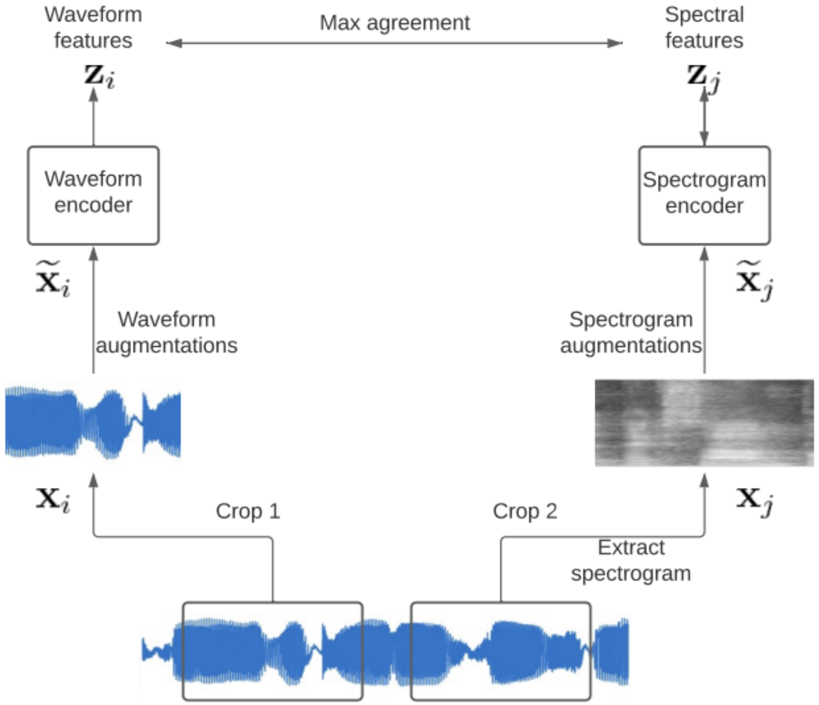
Contrastive Learning in Other Modalities

Textual Models



(b) Sequence-level Contrastive Learning (SeqCLR).

Audio Processing Models

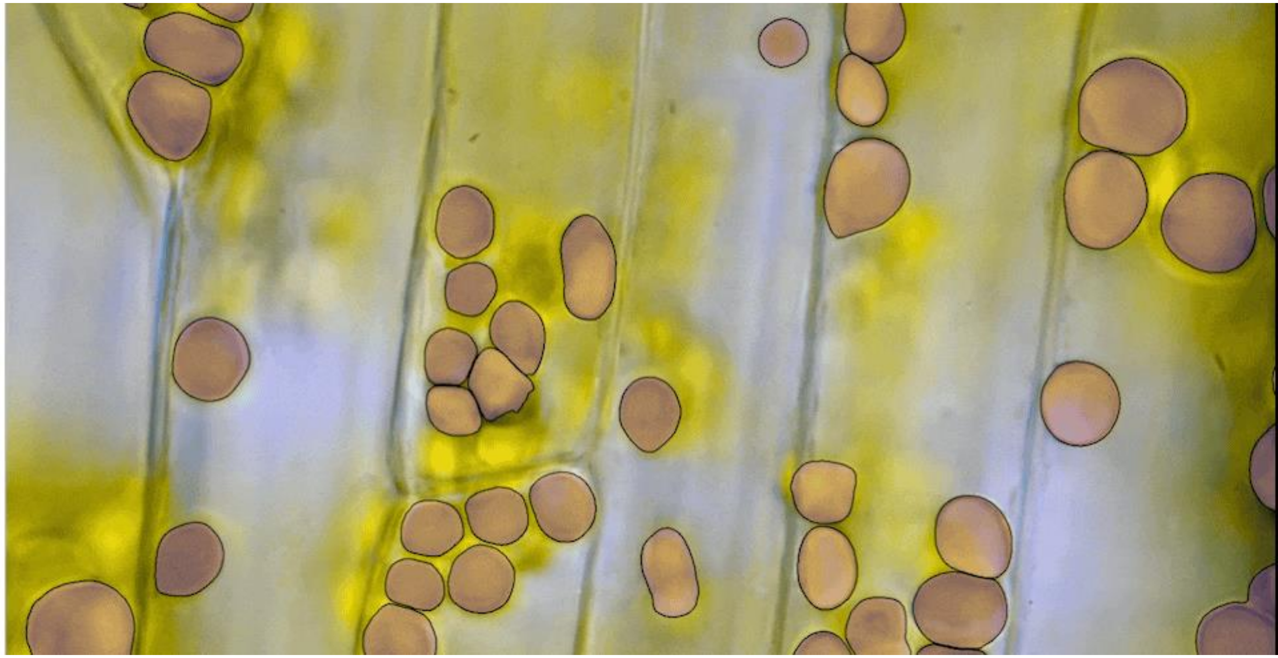


Xiong, L., Xiong, C., Li, Y., Tang, K. F., Liu, J., Bennett, P., ... & Overwijk, A. (2020). Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Wang, L., & Oord, A. V. D. (2021). Multi-format contrastive learning of audio representations. *arXiv preprint arXiv:2103.06508*.

As an Aside: Foundation Models

Segment Anything Model



Segment Anything Model (SAM) released by Meta on April 5, 2023 was trained on Segment Anything 1 Billion dataset with 1.1 billion high-quality segmentation masks from 11 million images

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao et al. "Segment anything." *arXiv preprint arXiv:2304.02643* (2023).

As an Aside: Foundation Models

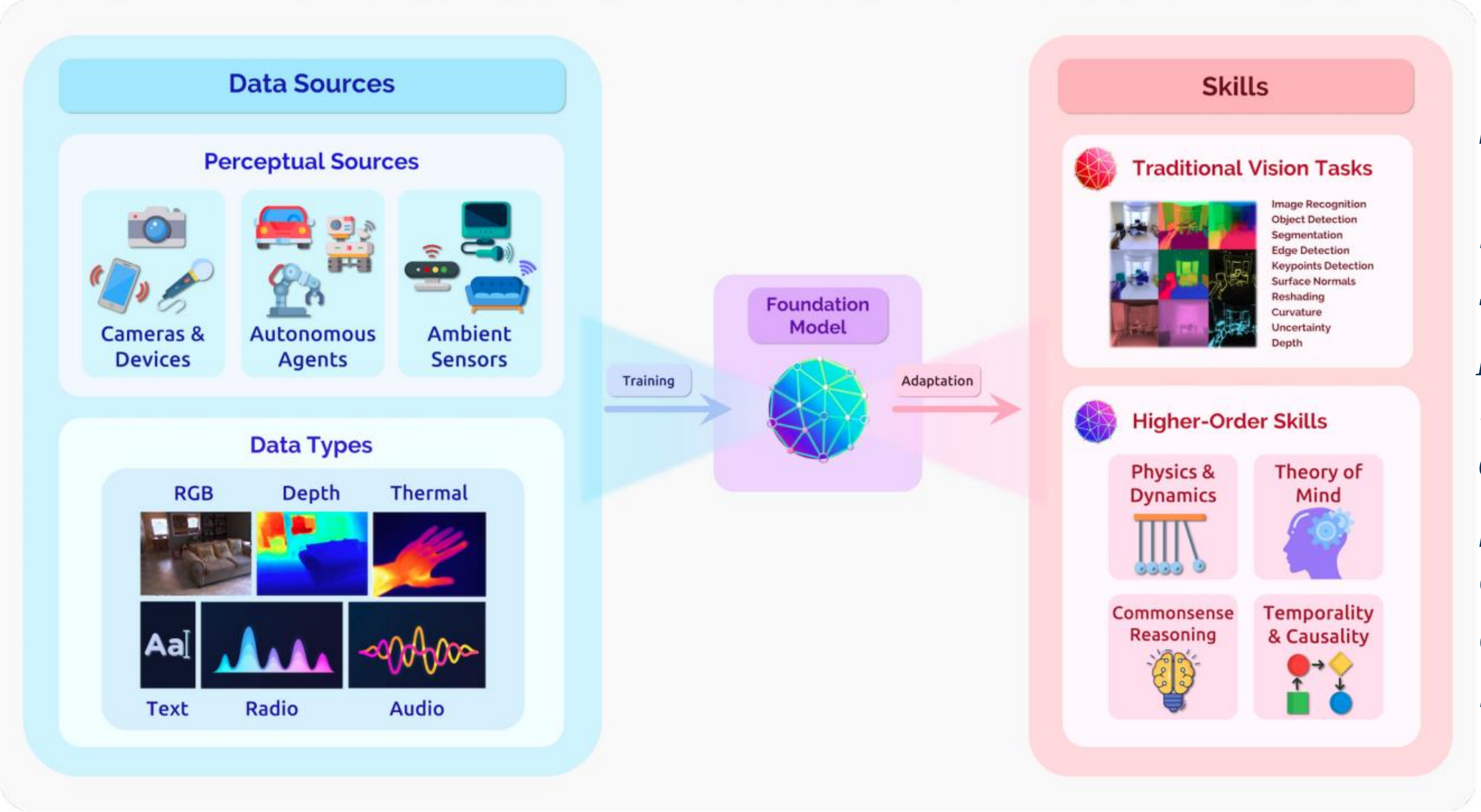
Origin of the term Foundation Models

- **Foundation models** are like any other deep network that have employed **transfer learning**, except **at *scale***
- ***Scale*** brings about ***emergent* properties** that are common between tasks
- **Before 2019:** Base architectures that powered multiple neural networks were **ResNets, VGG** etc.
- **Since 2019: BERT, DALL-E, GPT, Flamingo**
- Changes since 2019: **Transformer architectures and Self-Supervision**

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

As an Aside: Foundation Models

Origin of the term Foundation Models



'By harnessing ***self-supervision at scale****, foundation models for vision have the potential to* ***distill raw, multimodal sensory information into visual knowledge****, which may effectively support traditional* ***perception tasks*** *and possibly enable new progress on challenging higher-order skills like* ***temporal and commonsense reasoning*** *These inputs can come from a* ***diverse range of data sources*** *and application domains, suggesting promise for applications in* ***healthcare and embodied, interactive perception settings****'*

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

Terminology

- *Distribution*: (sample space) the set of all possible samples
- *Dataset*: a set of samples drawn from a distribution
- *Batch*: a subset of samples drawn from the dataset
- *Sample*: a single data object represented as a set of features
- *Feature*: value of a single attribute, property, in a sample. Could be numeric or categorical.

Appendix A: Notations

- x_i: a single feature
- $\boldsymbol{x}_i$: feature vector (a data sample)
- $\boldsymbol{x}_{:,i}$: feature vector of all data samples
- $\boldsymbol{X}$: matrix of feature vectors (dataset)
- $\boldsymbol{W}$: weight matrix
- $\boldsymbol{Z}$: latent representation
- E_θ: encoding function
- G_ϕ: decoding function
- $\widehat{\boldsymbol{X}}$: reconstruction of data
- N: number of data samples
- P: number of features in a feature vector
- $P^{(k)}$: the number of neurons in layer k
- α: learning rate
- Bold letter/symbol: vector
- Bold capital letters/symbol: matrix
- h_i and h_j representation space vectors
- z_i and z_j embedding space vectors