The **Trade-off** between **Universality** and **Label Efficiency** of Representations from Contrastive Learning

Zhenmei Shi*, Jiefeng Chen*, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, Somesh Jha

UW-Madison, Google, XaiPient

MLOPT 2023

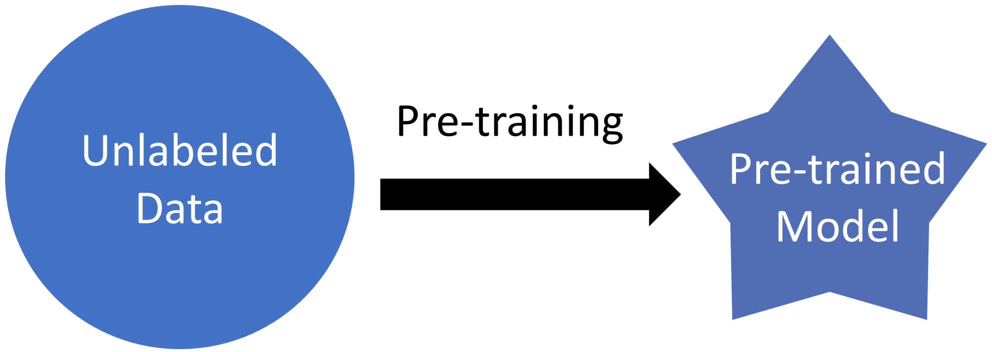
ICLR

New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ➞ pre-training + adaptation

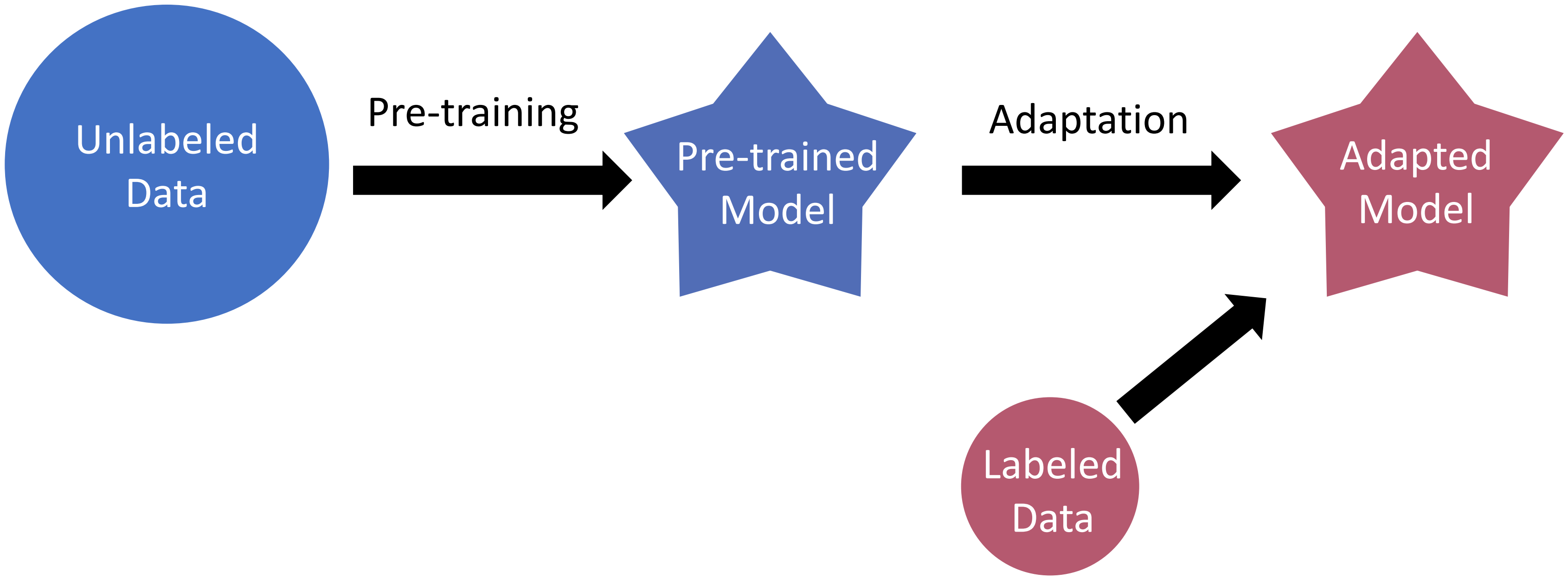
New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ➞ pre-training + adaptation



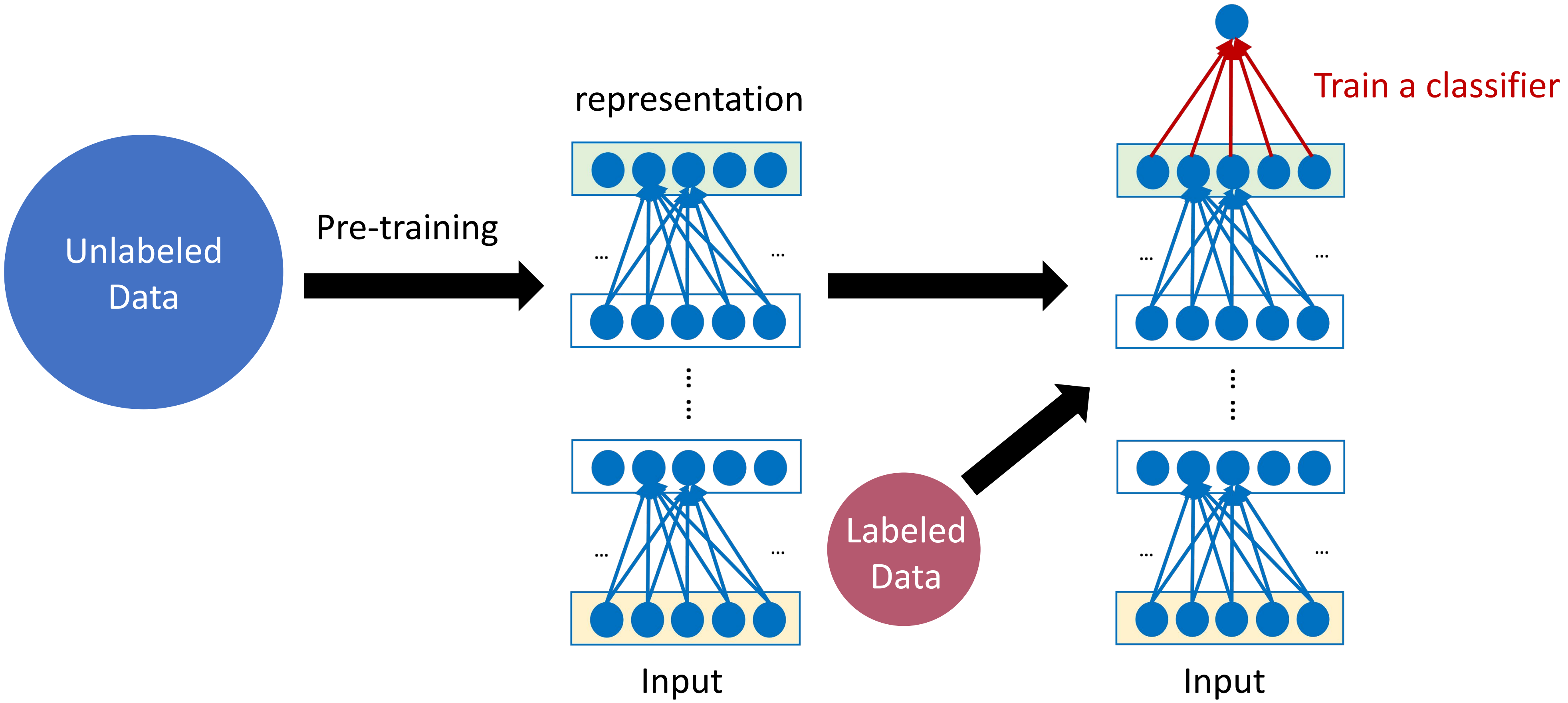
New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning → pre-training + adaptation



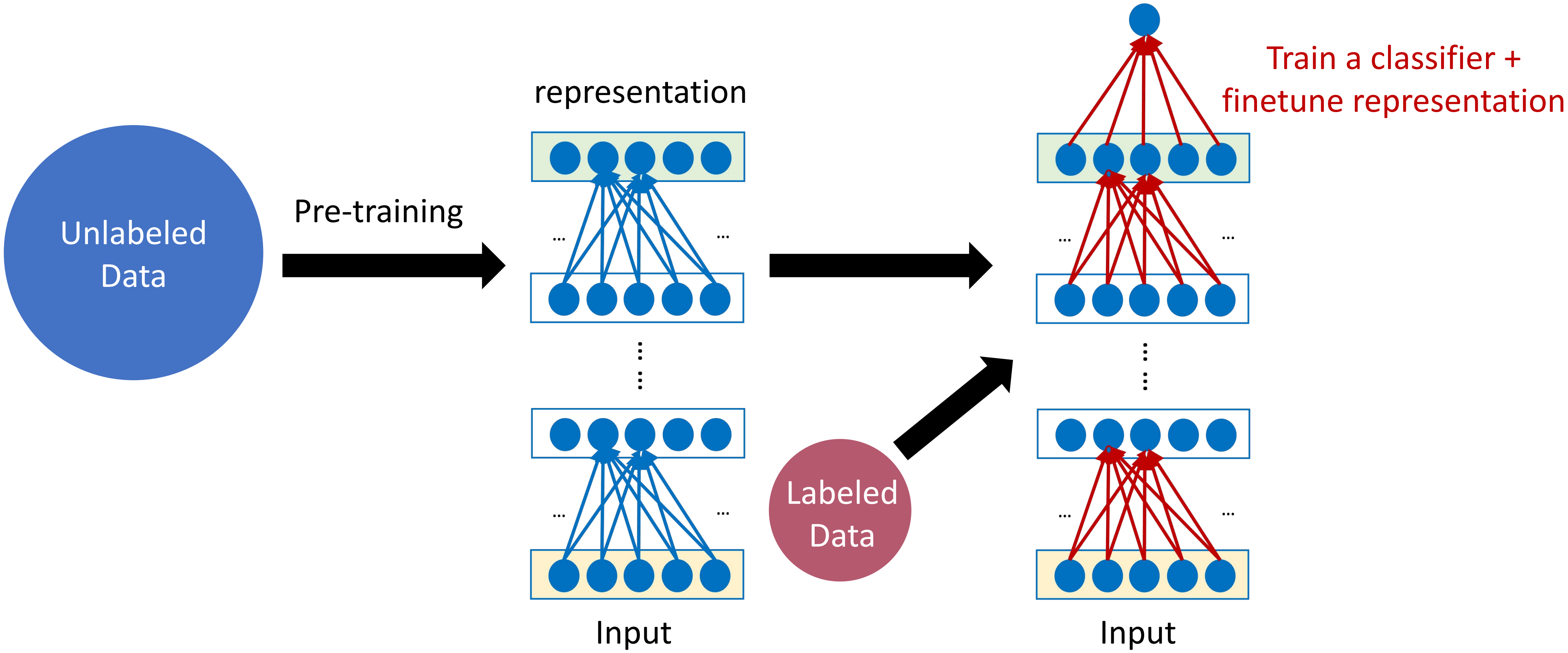
New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ➞ pre-training + adaptation



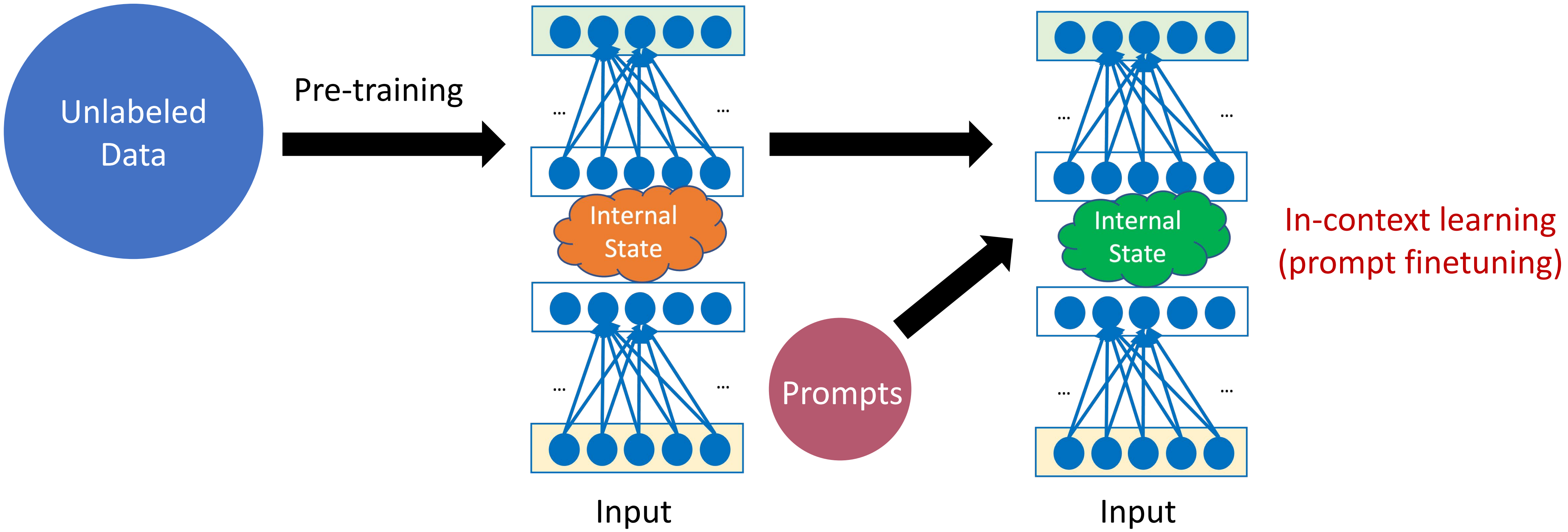
New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ➞ pre-training + adaptation



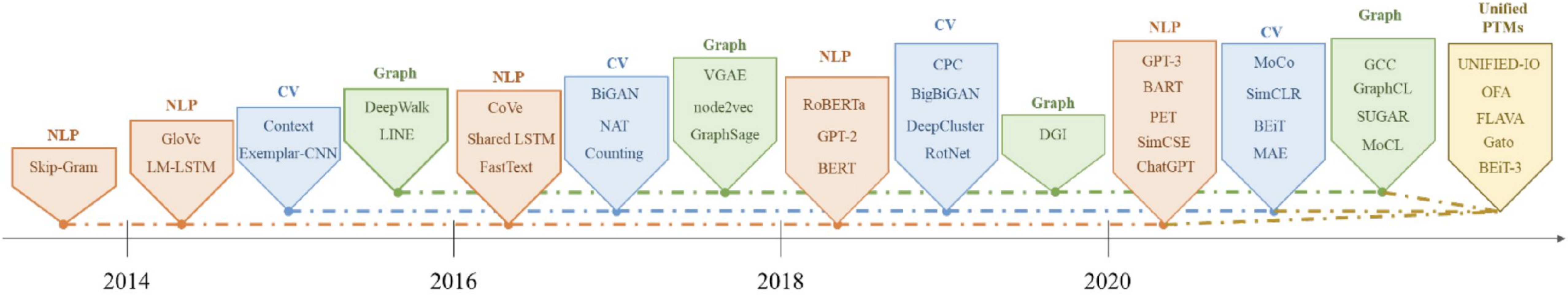
New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ➞ pre-training + adaptation



New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ➞ pre-training + adaptation

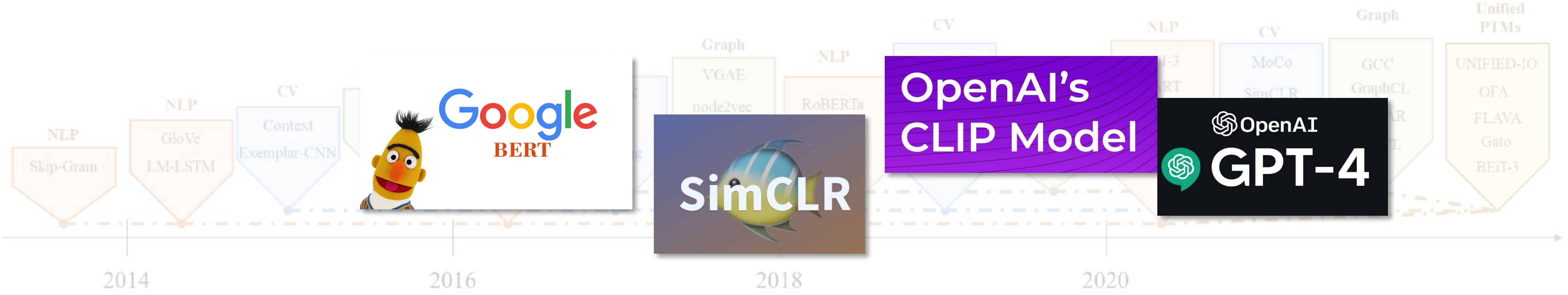


The history and evolution of pre-trained models

Figures from: *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT, 2023.*

New Paradigm: Pre-trained Representations

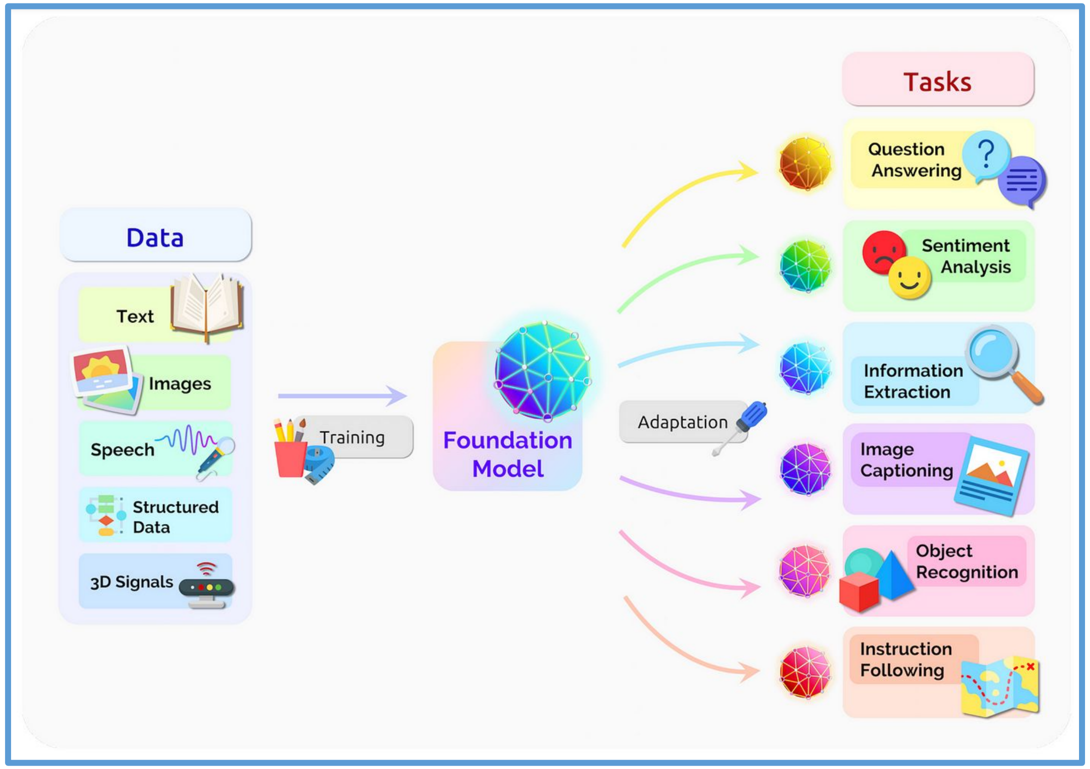
Paradigm shift: supervised learning ➞ pre-training + adaptation



The history and evolution of pre-trained models

Figures from: *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT, 2023.*

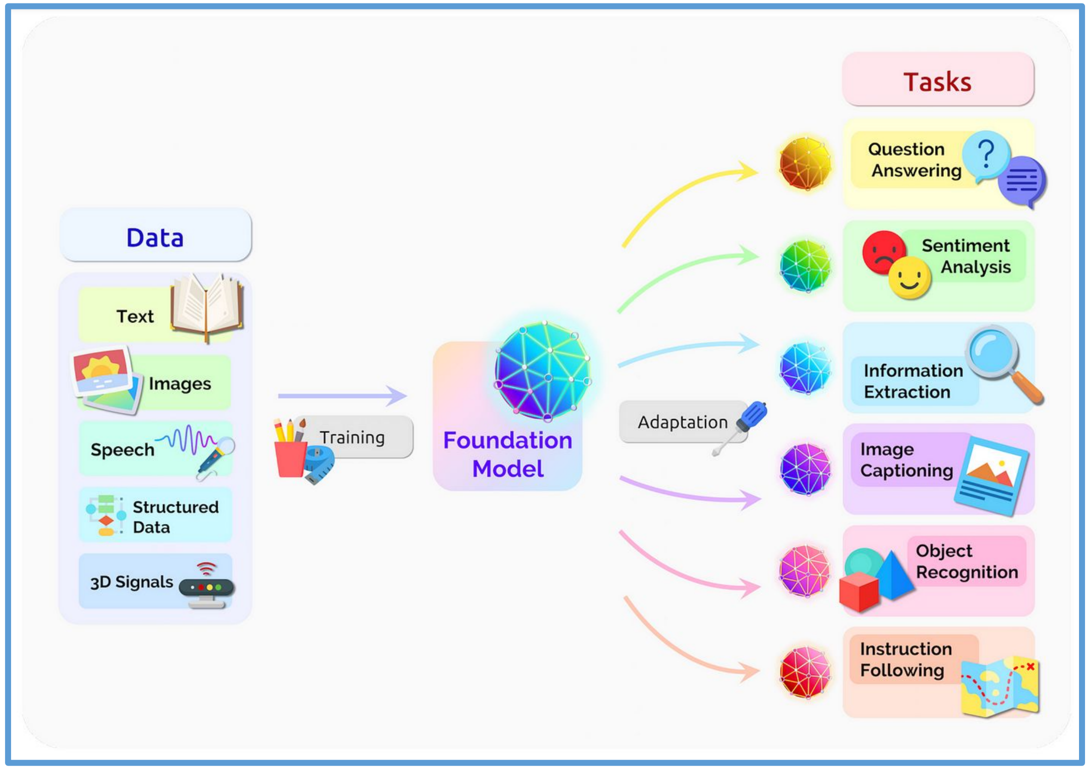
Intro - Foundation Model



Universality

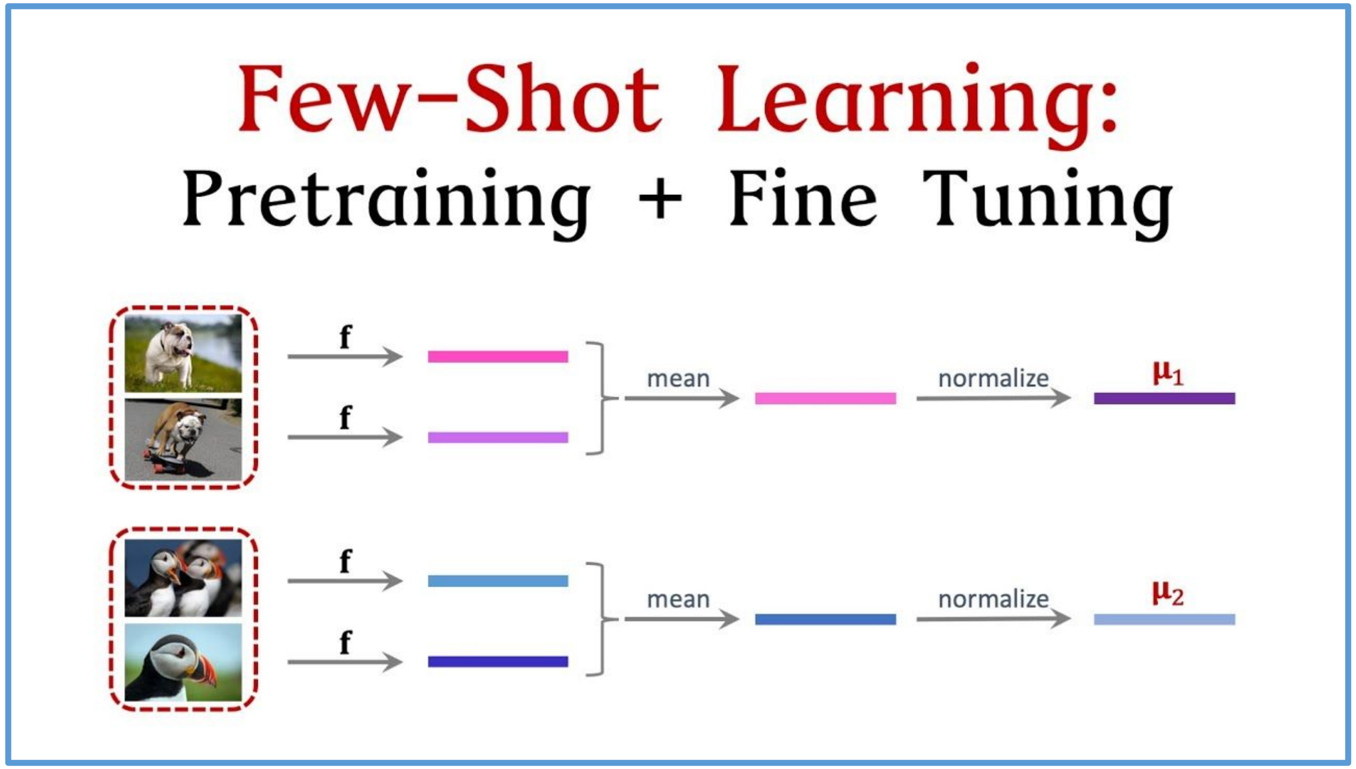
Figures from: *On the opportunities and risks of foundation models, 2021.*

Intro - Foundation Model



Universality

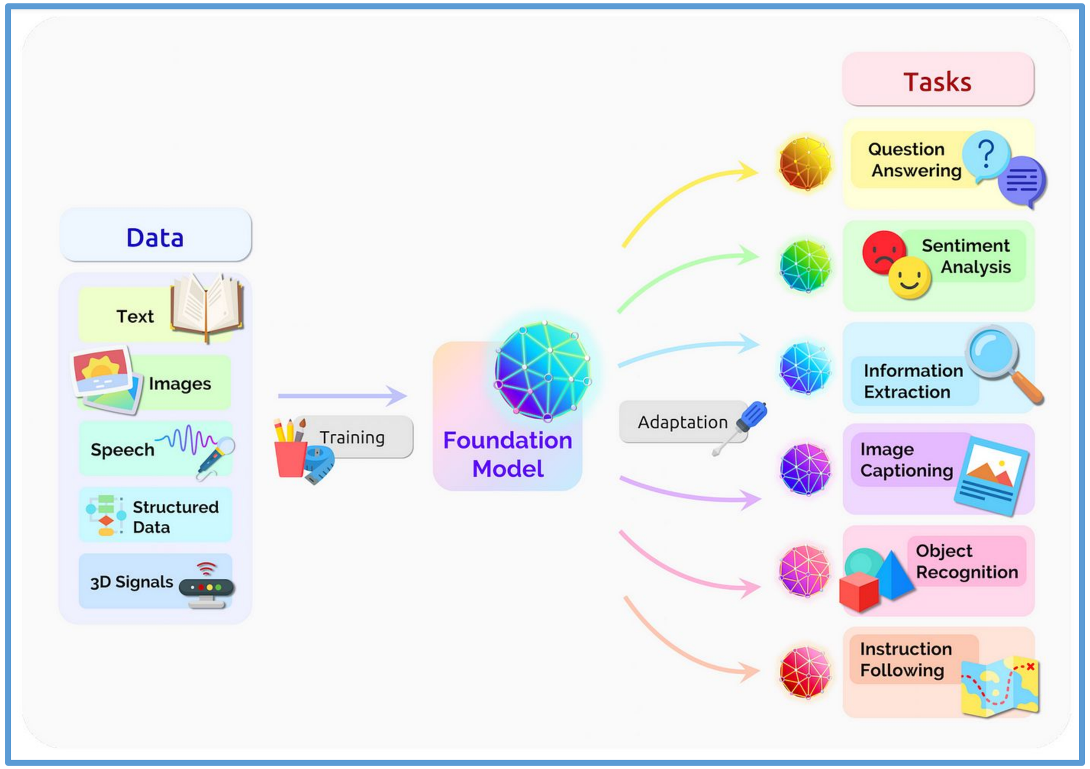
Figures from: *On the opportunities and risks of foundation models, 2021.*



Label Efficiency

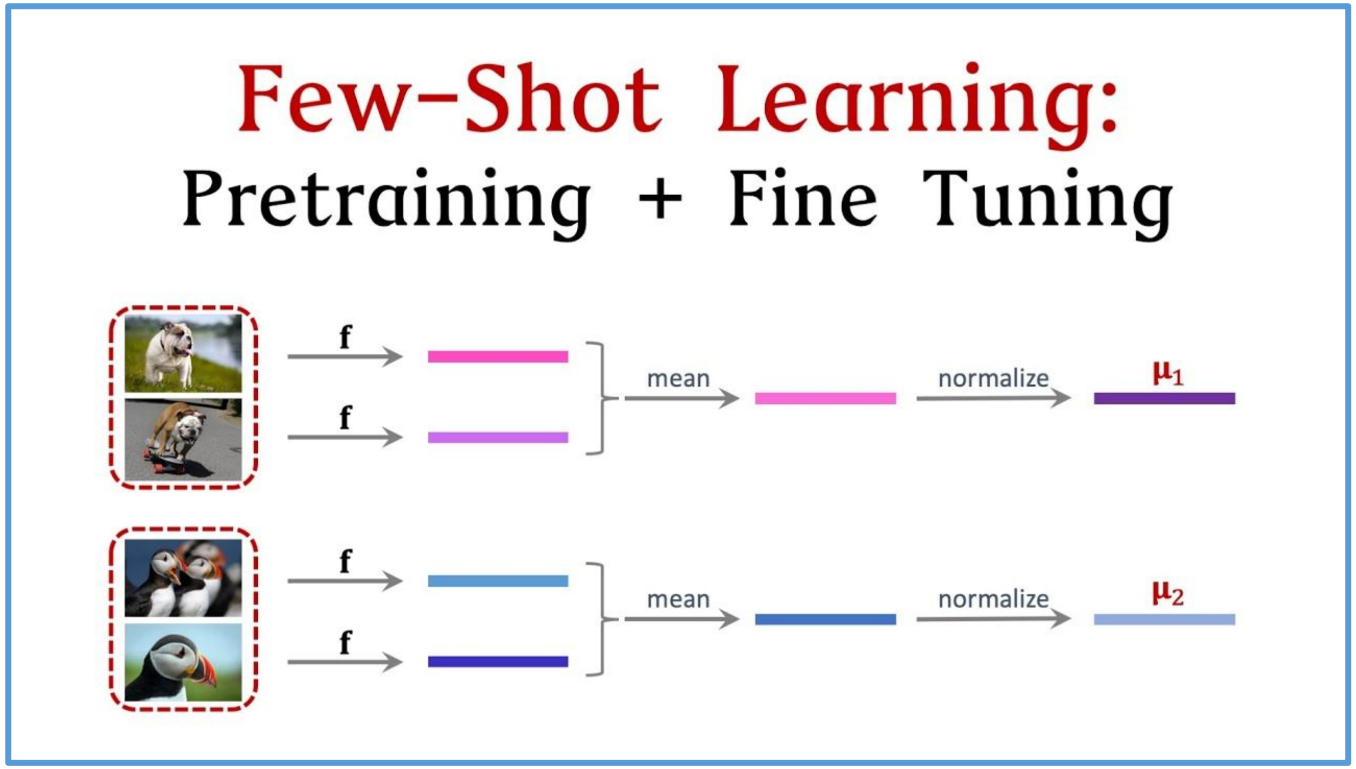
Figures from: *https://www.youtube.com/watch?v=U6uFOIURcD0&ab_channel=ShusenWang, 2020*

Intro - Foundation Model



Universality

Figures from: *On the opportunities and risks of foundation models, 2021.*

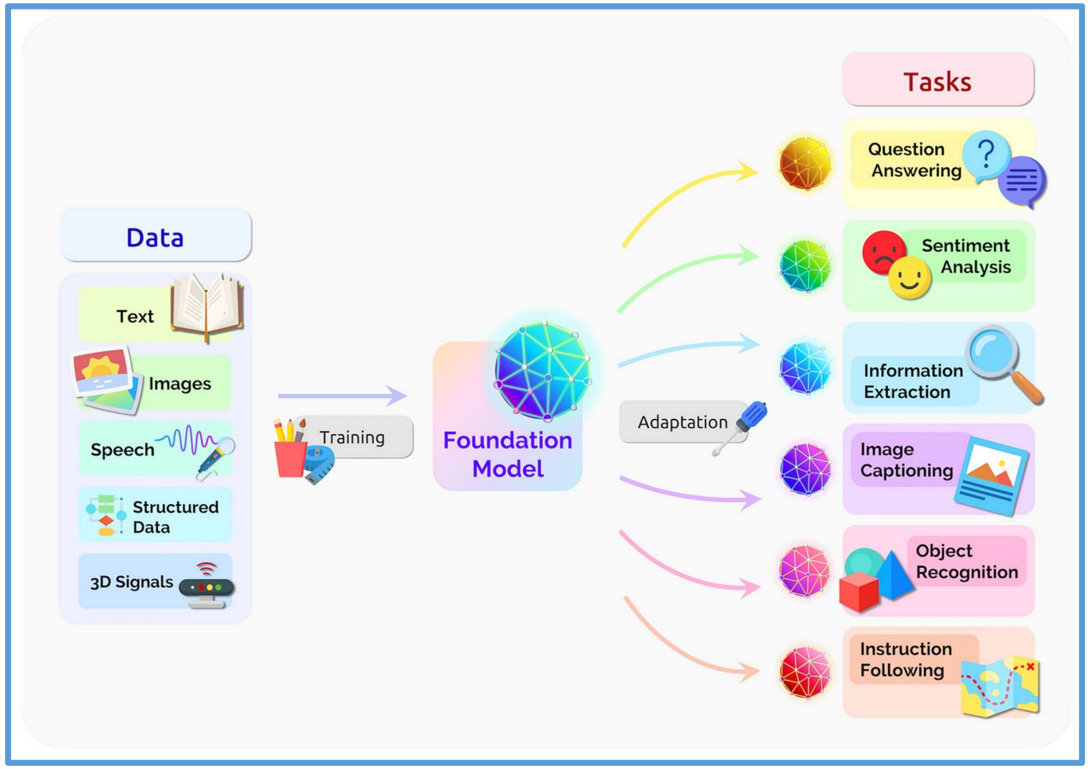


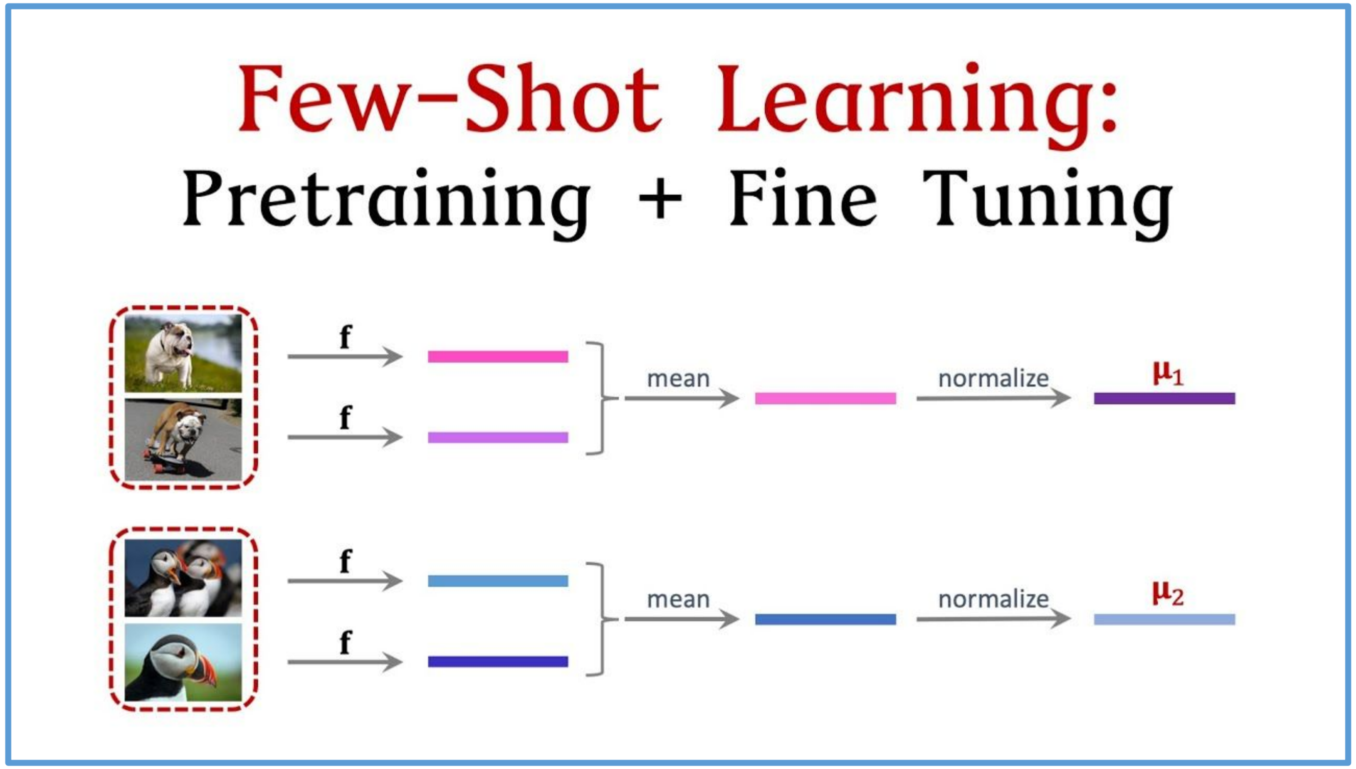
Label Efficiency

Figures from: *https://www.youtube.com/watch?v=U6uFOIURcD0&ab_channel=ShusenWang, 2020*

Q: Can we gain two key properties simultaneously?

Intro - Foundation Model





Universality

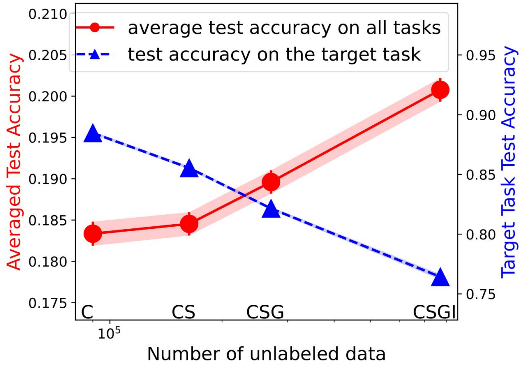
Figures from: *On the opportunities and risks of foundation models, 2021.*

Label Efficiency

Figures from: *https://www.youtube.com/watch?v=U6uFOIURcD0&ab_channel=ShusenWang, 2020*

Q: Can we gain two key properties simultaneously?

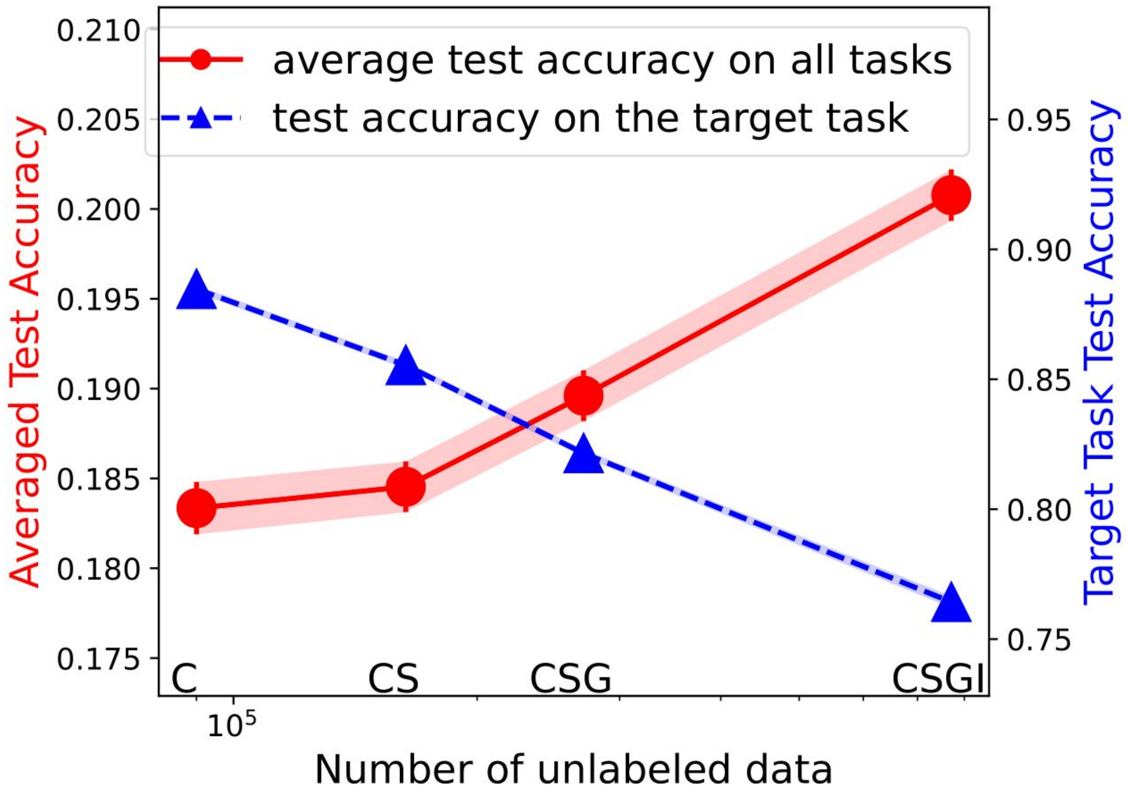
A: A **trade-off** exists at least for **contrastive learning**!



Trade-off of Label Efficiency and Universality

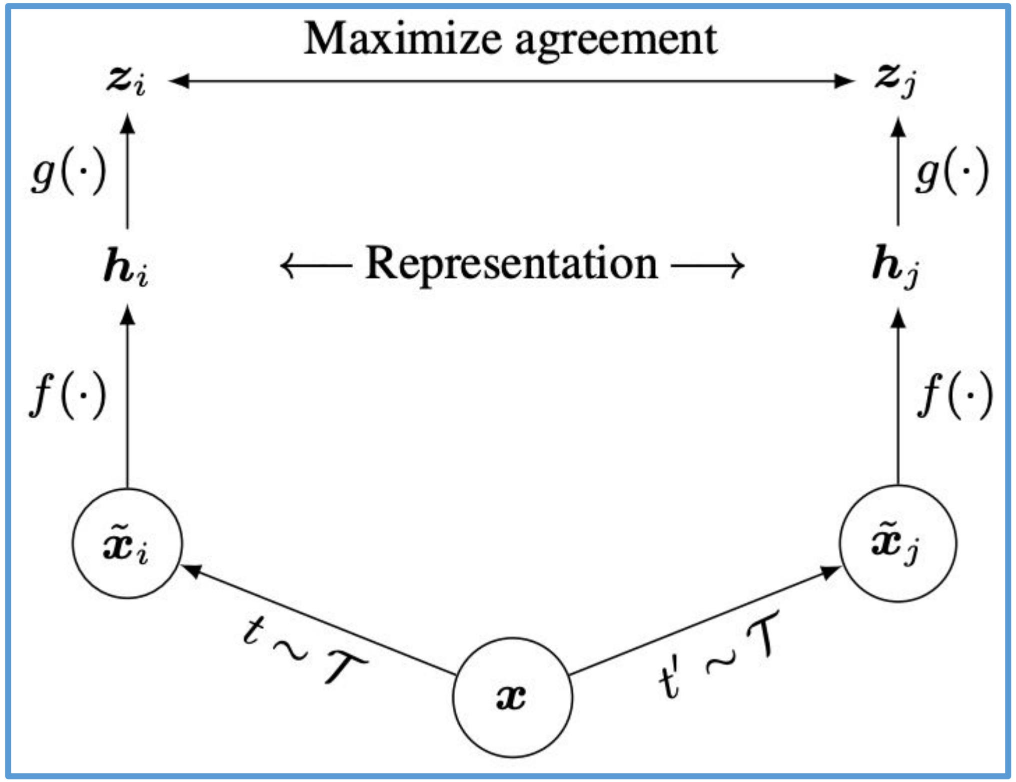
Contrastive learning ResNet18 backbone via MoCo, then classify on CIFAR10.

From left to right, incrementally add to pre-training: CINIC-10 (C), SVHN (S), GTSRB (G), and ImageNet32 (I)



[SCL+23] Shi, Chen, Li, Raghuram, Wu, Liang, Jha. The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning. ICLR’2023.

Intro - Contrastive Learning



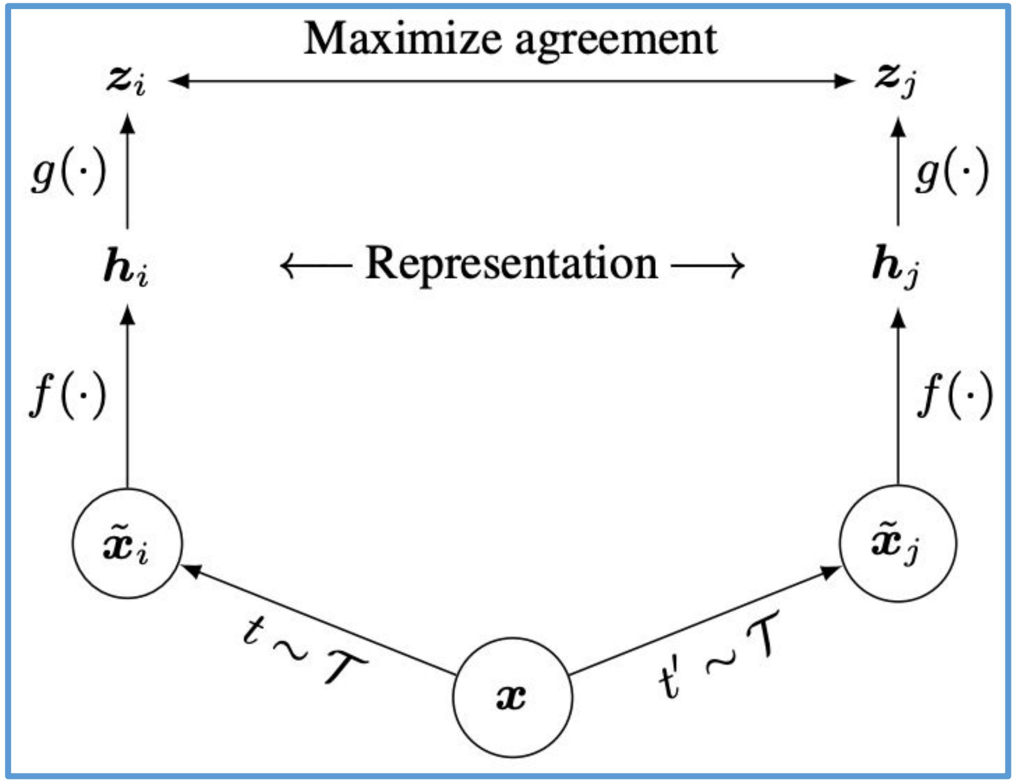
$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

SimCLR - (Image, Image)

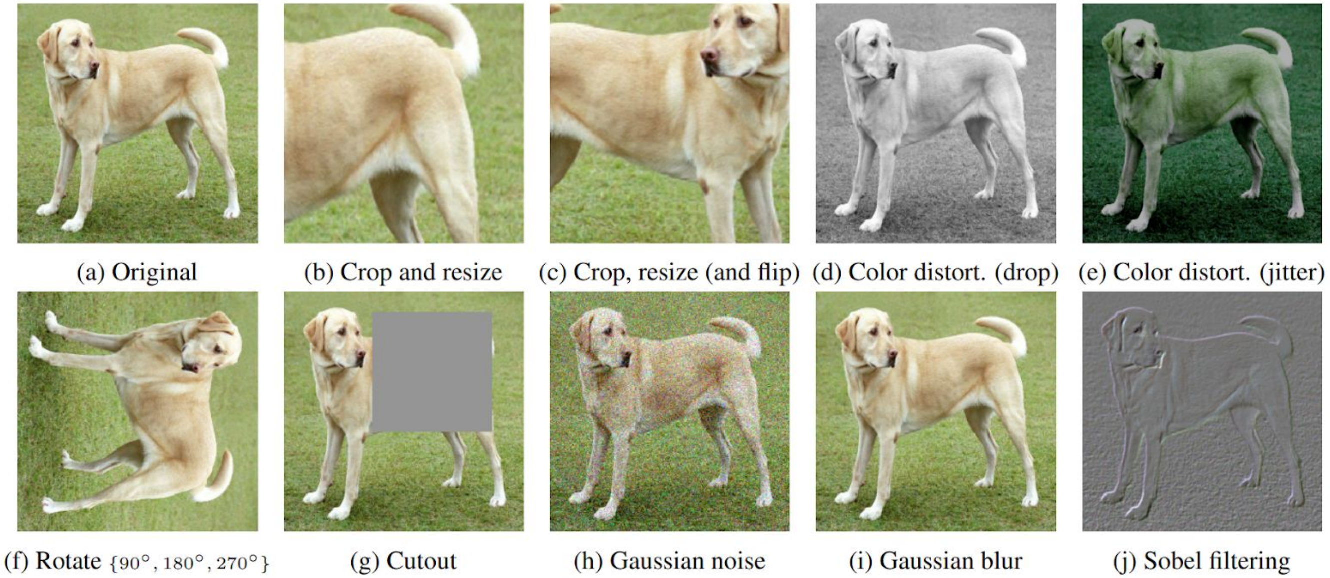
No need labels

Figures from: *A Simple Framework for Contrastive Learning of Visual Representations, 2020*

Intro - Contrastive Learning



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



SimCLR - (Image, Image)

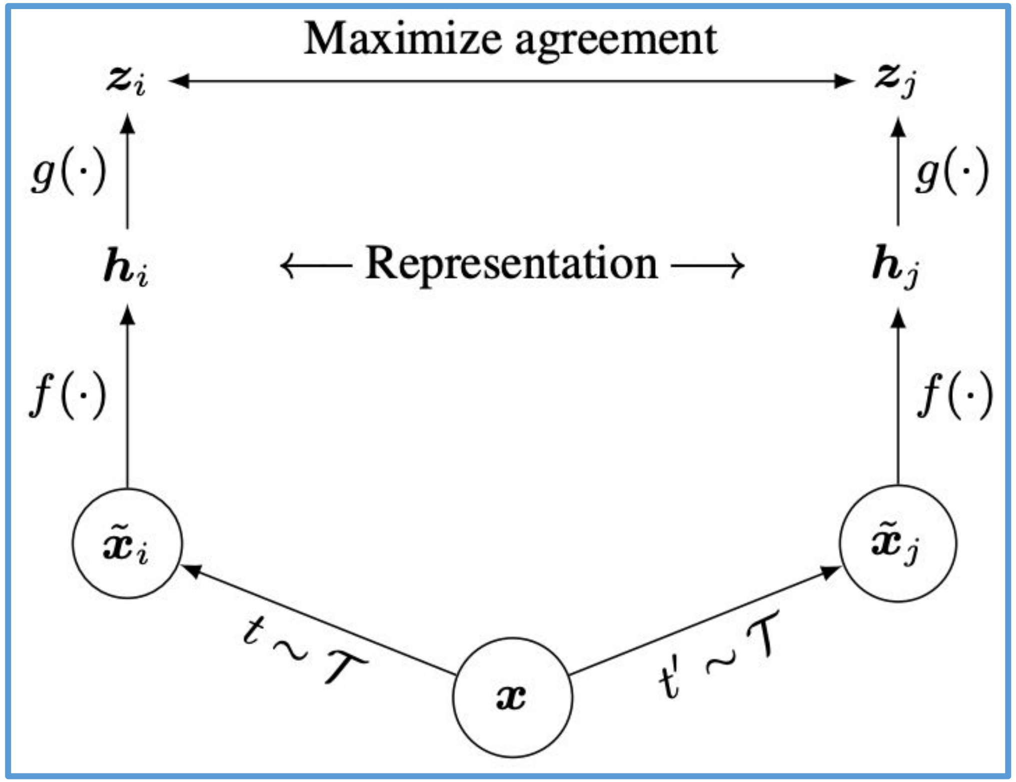
No need labels

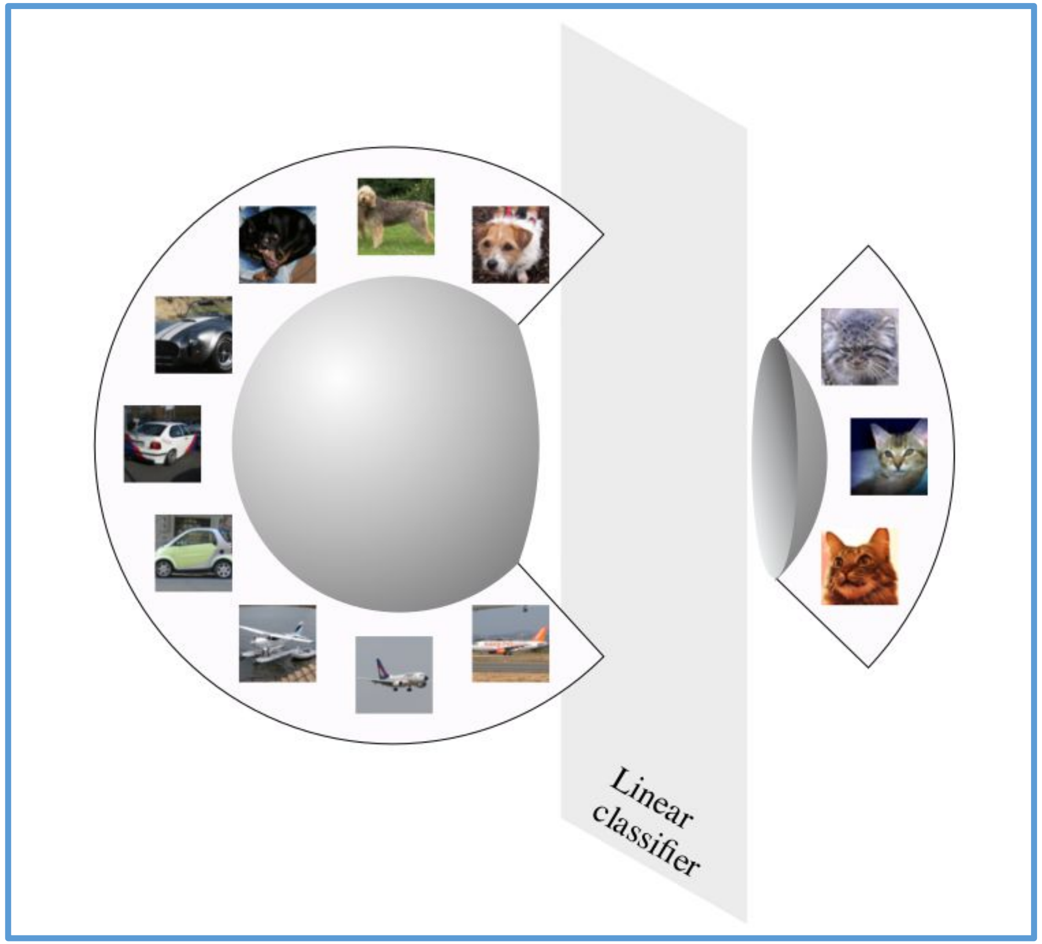
Figures from: *A Simple Framework for Contrastive Learning of Visual Representations, 2020*

Image Data Augmentation

Figures from: *A Simple Framework for Contrastive Learning of Visual Representations, 2020*

Intro - Contrastive Learning





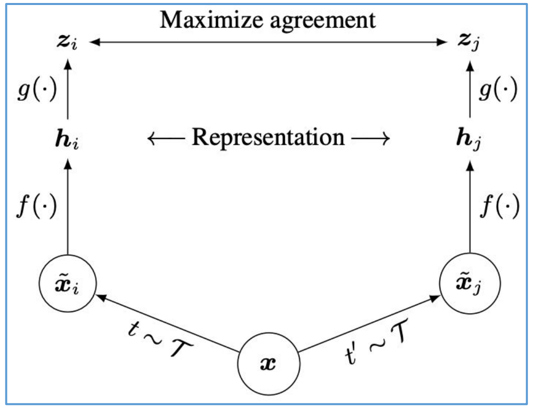
SimCLR - (Image, Image)
No need labels

Figures from: *A Simple Framework for Contrastive Learning of Visual Representations, 2020*

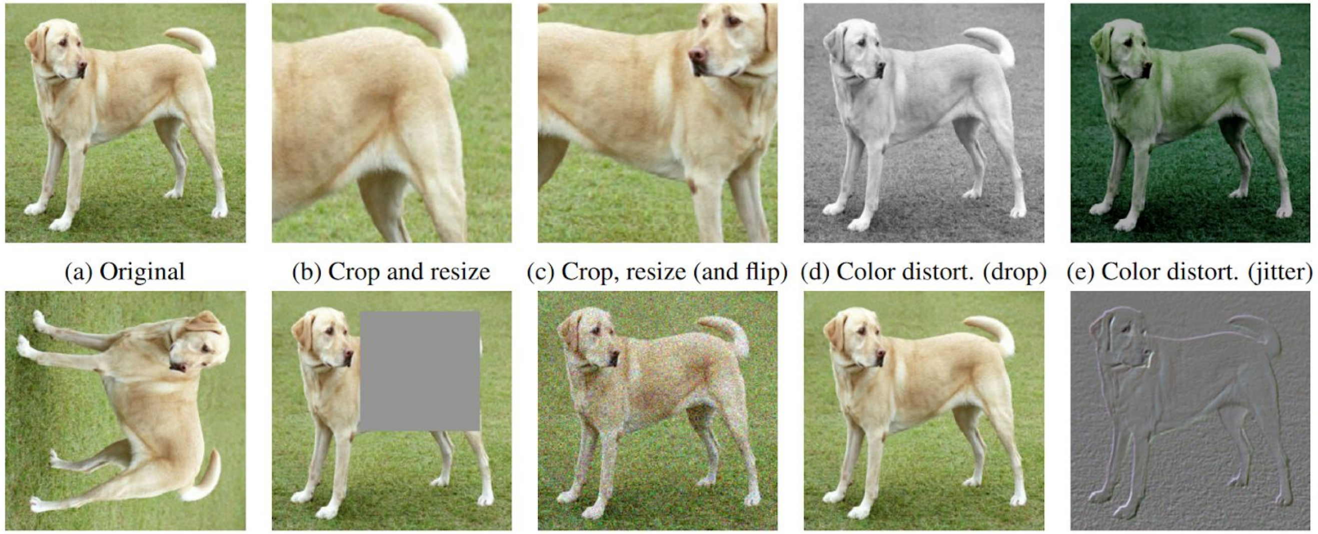
Linear Probing

Figures from: *Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere, 2021.*

Intro - Invariant/Spurious Feature



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



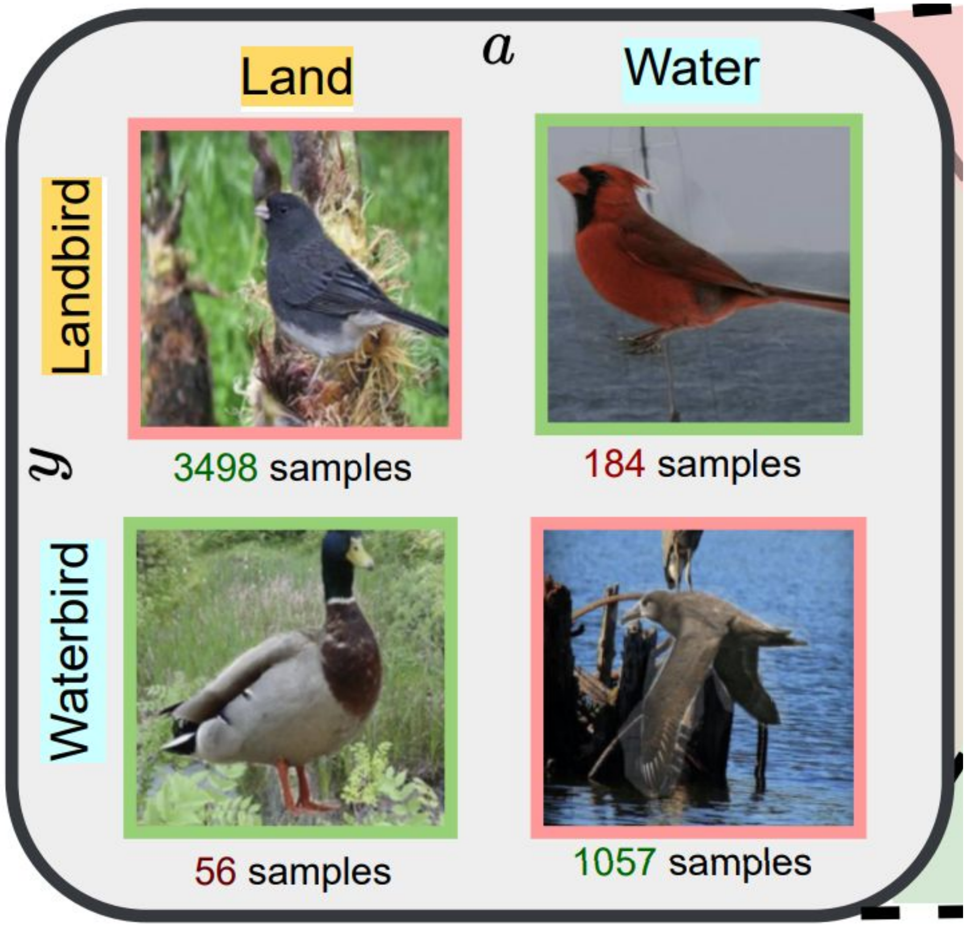
(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

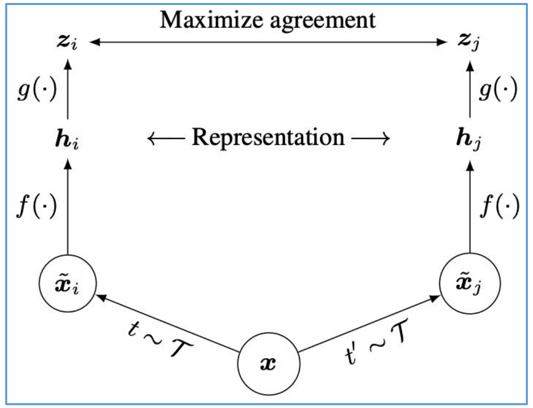
Image Data Augmentation

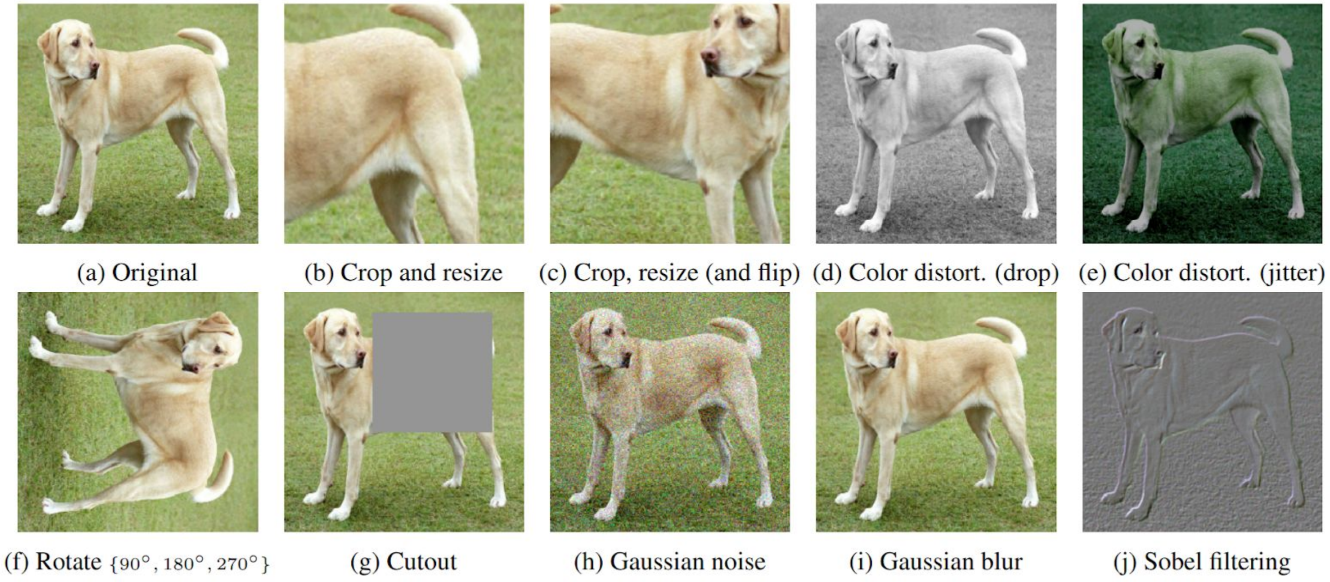
Figures from: *A Simple Framework for Contrastive Learning of Visual Representations, 2020*

Intro - Invariant/Spurious Feature



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





Waterbirds

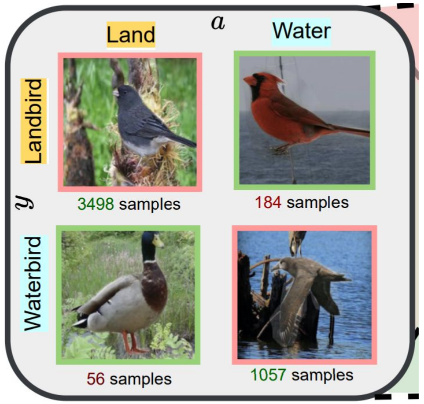
Invariant - Birds; Spurious - Background

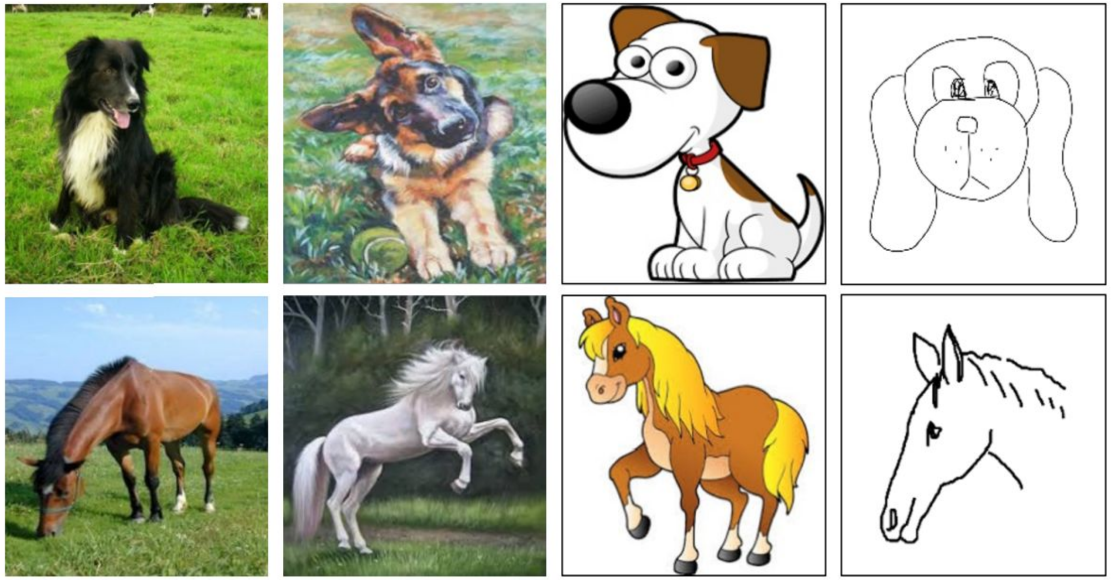
Figures from: *Avoiding spurious correlations via logit correction, 2023*

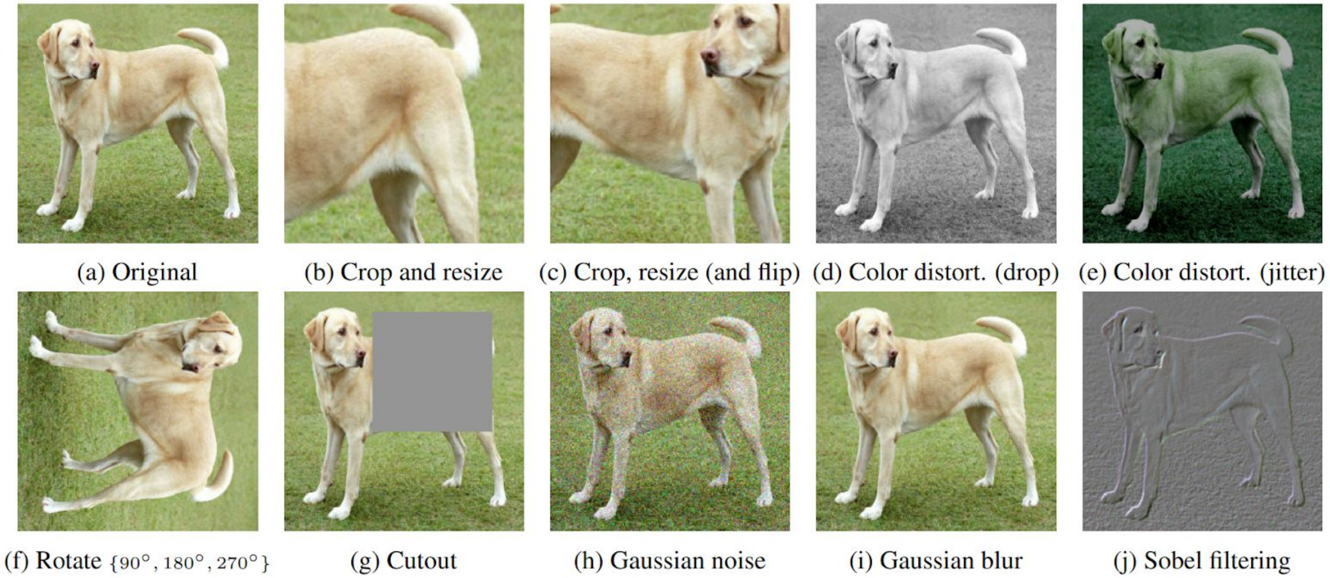
Image Data Augmentation

Figures from: *A Simple Framework for Contrastive Learning of Visual Representations, 2020*

Intro - Invariant/Spurious Feature







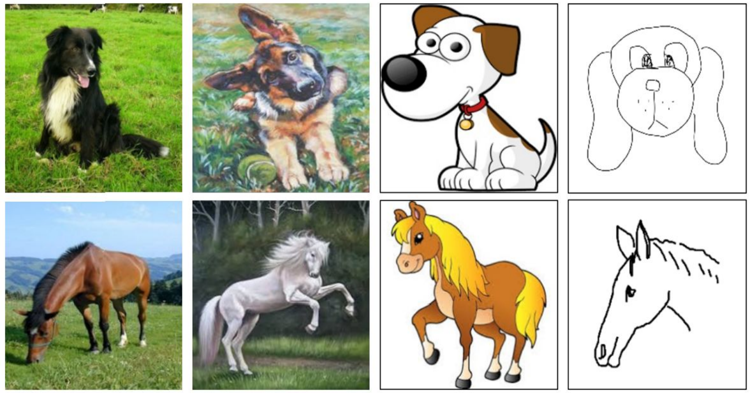
PACS: Photo, Painting, Cartoon, Sketch

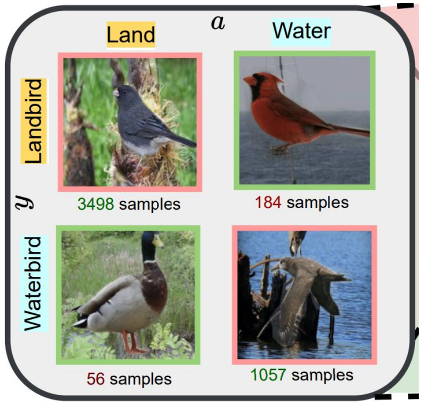
Invariant - Object; Spurious - Image Style

Image Data Augmentation

Figures from: *A Simple Framework for Contrastive Learning of Visual Representations, 2020*

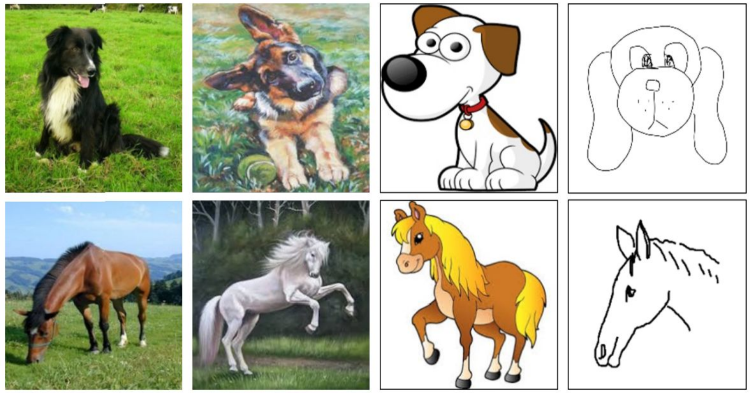
Question?

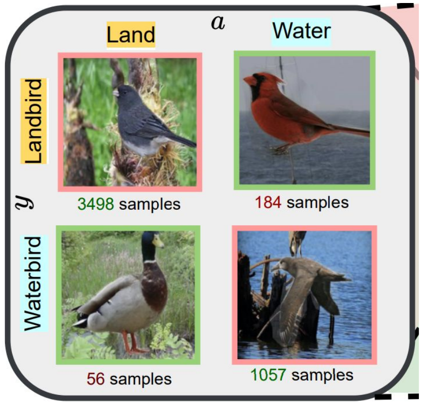




Q1: What features are learned by contrastive learning?

Question?

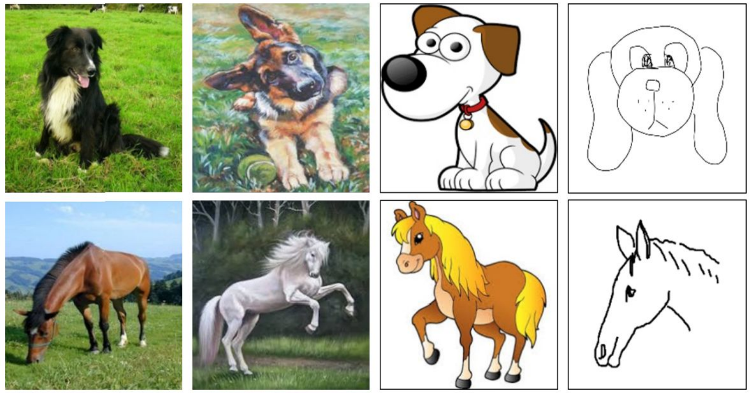


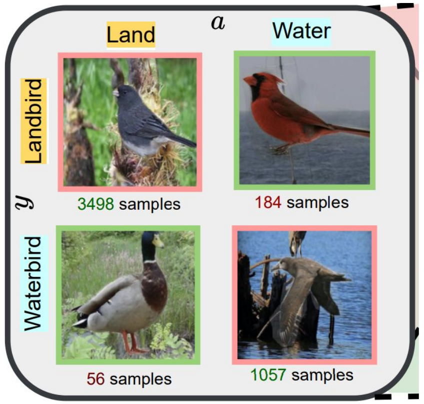


Q1: What features are learned by contrastive learning?

A1: Contrastive Learning = Generalized Nonlinear PCA. Encodes almost all invariant features but removes the others.

Question?



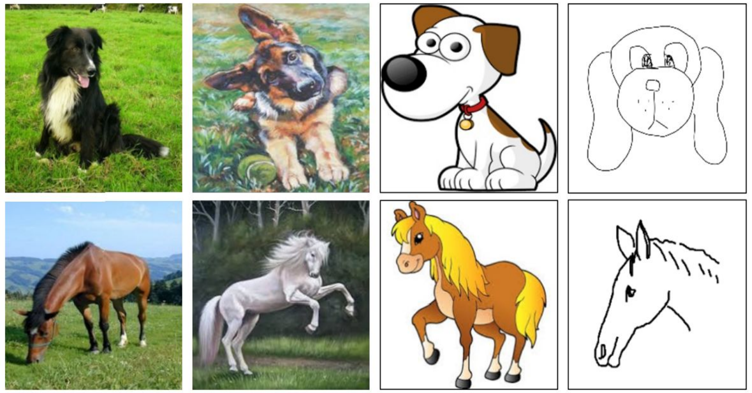


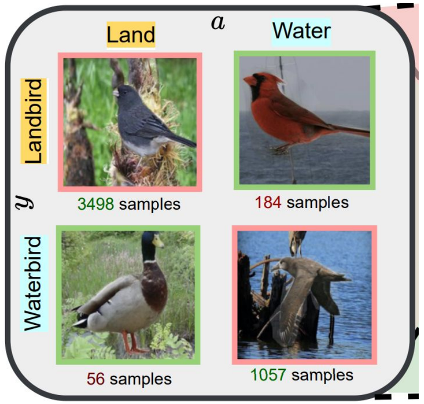
Q1: What features are learned by contrastive learning?

A1: Contrastive Learning = Generalized Nonlinear PCA. Encodes almost all invariant features but removes the others.

Q2: Is it always good to encode almost **all** invariant features?

Question?





Q1: What features are learned by contrastive learning?

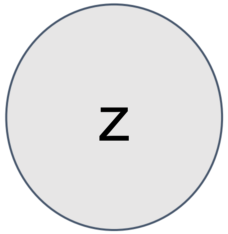
A1: Contrastive Learning = Generalized Nonlinear PCA. Encodes almost all invariant features but removes the others.

Q2: Is it always good to encode almost **all** invariant features?

A2: No! More Diverse data ➞ Target task features down-weighted ➞ Worse target task performance

Problem Setup - Hidden representation data model

- Hidden representation space $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ over distribution $\mathcal{D}_z$

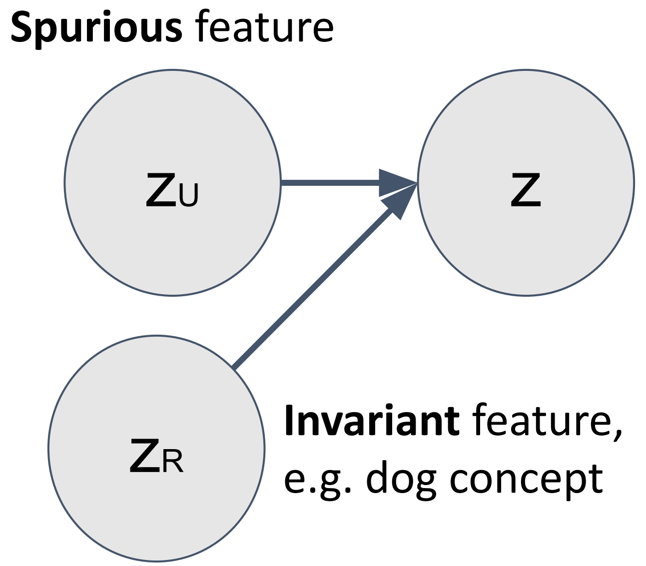


Data Model

Figures from: *Expanding Small-Scale Datasets with Guided Imagination, 2023*

Problem Setup - Hidden representation data model

- Hidden representation space $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ over distribution $\mathcal{D}_z$
- Invariant feature R, Spurious feature U, $R \cup U = [d]$, $R \cap U = \emptyset$

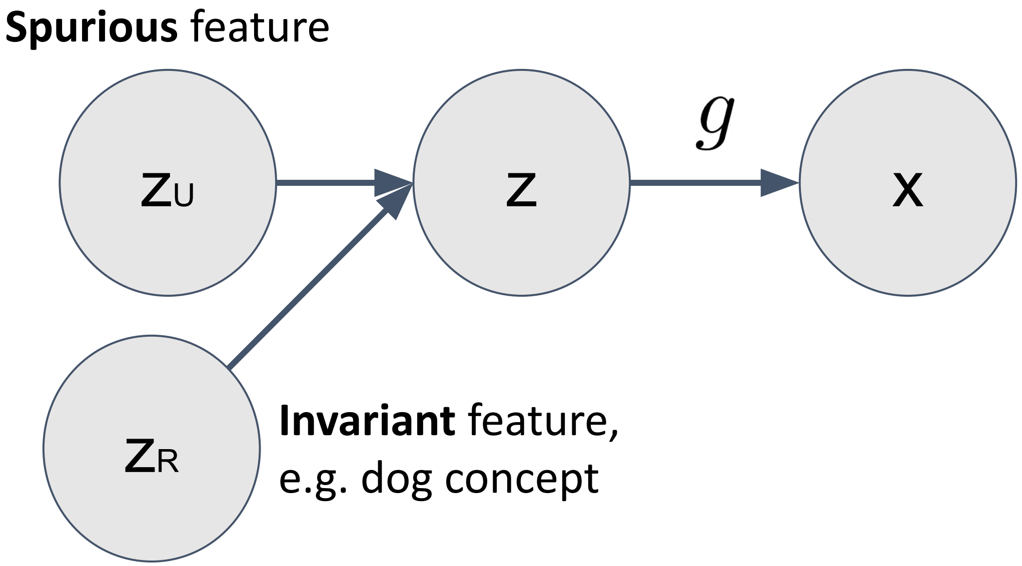


Data Model

Figures from: *Expanding Small-Scale Datasets with Guided Imagination, 2023*

Problem Setup - Hidden representation data model

- Hidden representation space $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ over distribution $\mathcal{D}_z$
- Invariant feature R, Spurious feature U, $R \cup U = [d]$, $R \cap U = \emptyset$
- $x = g(z)$, g is a generative function; y depends on z as well

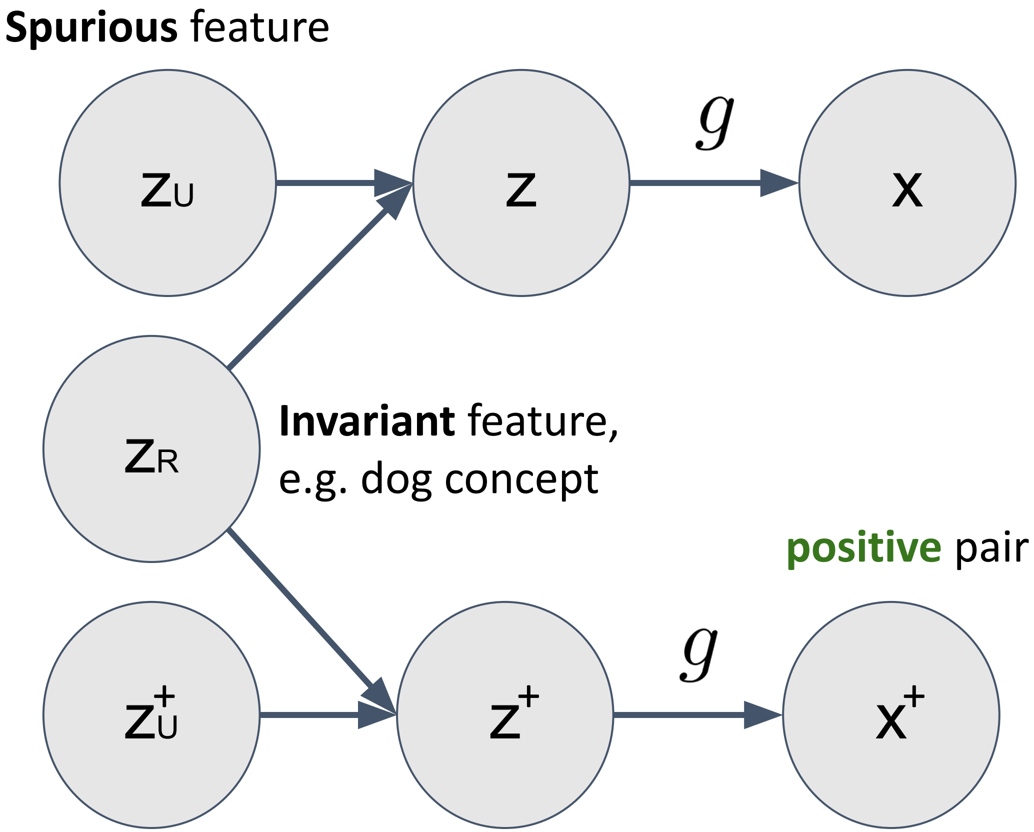


Data Model

Figures from: *Expanding Small-Scale Datasets with Guided Imagination, 2023*

Problem Setup - Hidden representation data model

- Hidden representation space $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ over distribution $\mathcal{D}_z$
- Invariant feature R, Spurious feature U, $R \cup U = [d]$, $R \cap U = \emptyset$
- $x = g(z)$, g is a generative function; y depends on z as well

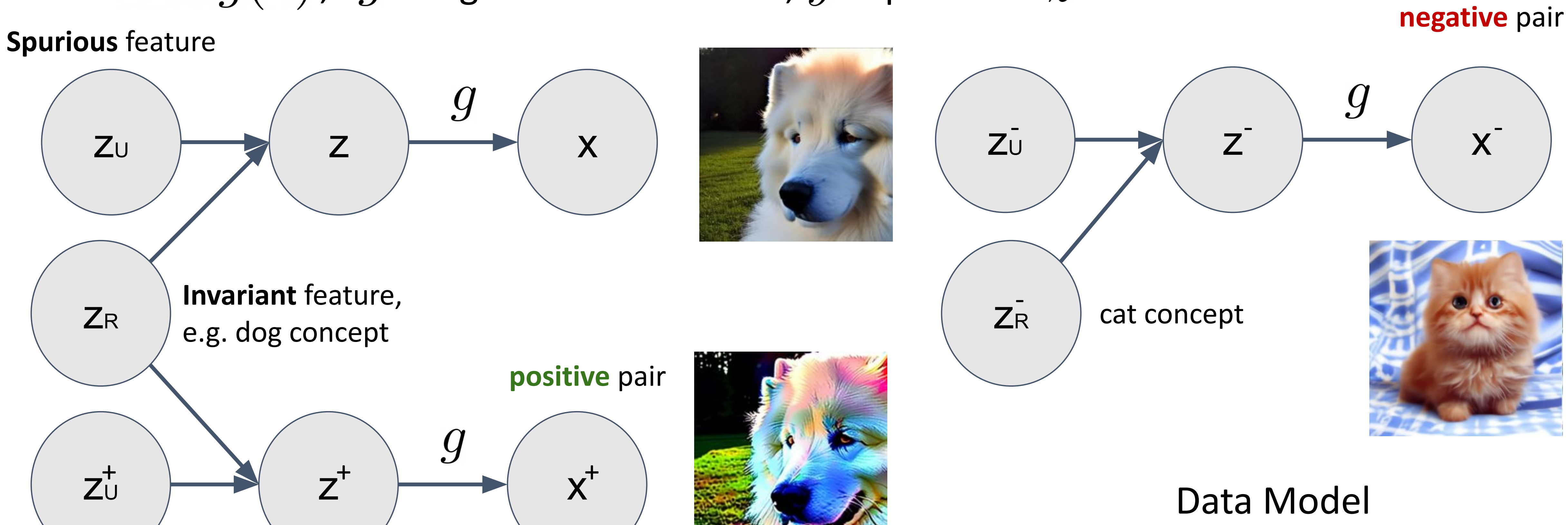


Data Model

Figures from: *Expanding Small-Scale Datasets with Guided Imagination, 2023*

Problem Setup - Hidden representation data model

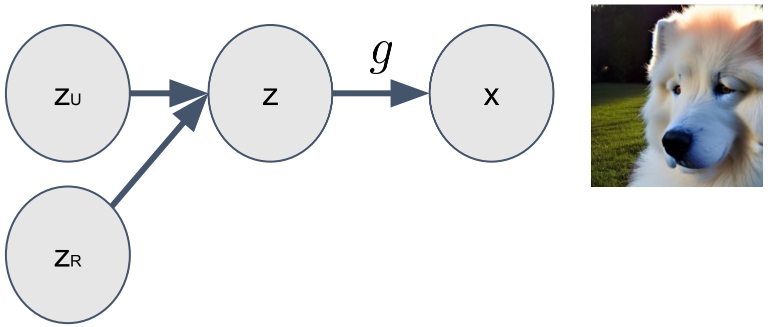
- Hidden representation space $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ over distribution $\mathcal{D}_z$
- Invariant feature R, Spurious feature U, $R \cup U = [d]$, $R \cap U = \emptyset$
- $x = g(z)$, g is a generative function; y depends on z as well



Figures from: *Expanding Small-Scale Datasets with Guided Imagination, 2023*

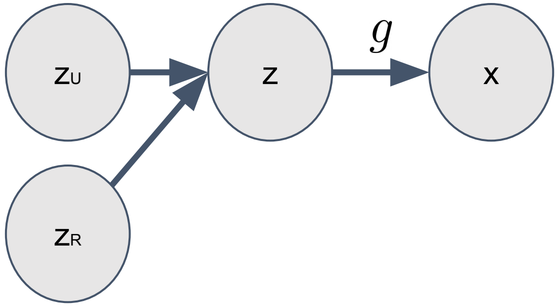
Problem Setup - Hidden representation data model

- Hidden representation space $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ over distribution $\mathcal{D}_z$
- Invariant feature R, Spurious feature U, $R \cup U = [d]$, $R \cap U = \emptyset$
- $x = g(z)$, g is a generative function; y depends on z as well
- $\phi \in \Phi$ hypothesis class of representation functions, e.g, ResNet, ViT



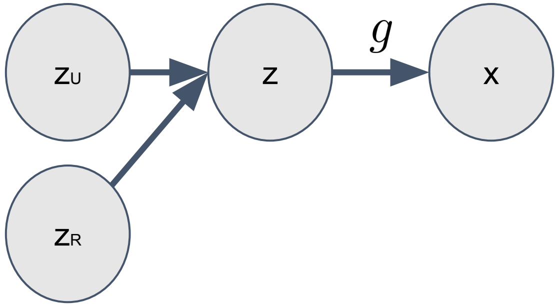
Problem Setup - Hidden representation data model

- Hidden representation space $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ over distribution $\mathcal{D}_z$
- Invariant feature R, Spurious feature U, $R \cup U = [d]$, $R \cap U = \emptyset$
- $x = g(z)$, g is a generative function; y depends on z as well
- $\phi \in \Phi$ hypothesis class of representation functions, e.g, ResNet, ViT
- Contrastive Loss $\min_{\phi \in \Phi} \mathbb{E}_{(x,x^+,x^-)\sim\mathcal{D}_{pre}}\left[\ell\left(\phi(x)^\top(\phi(x^+) - \phi(x^-))\right)\right]$



Problem Setup - Hidden representation data model

- Hidden representation space $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ over distribution $\mathcal{D}_z$
- Invariant feature R, Spurious feature U, $R \cup U = [d]$, $R \cap U = \emptyset$
- $x = g(z)$, g is a generative function; y depends on z as well
- $\phi \in \Phi$ hypothesis class of representation functions, e.g, ResNet, ViT
- Contrastive Loss $\min_{\phi \in \Phi} \mathbb{E}_{(x,x^+,x^-)\sim\mathcal{D}_{pre}} \left[\ell\left(\phi(x)^\top(\phi(x^+) - \phi(x^-))\right)\right]$
- In SimCLR, we have multiple negative pairs and $\ell(t) = \log(1 + \exp(-t))$

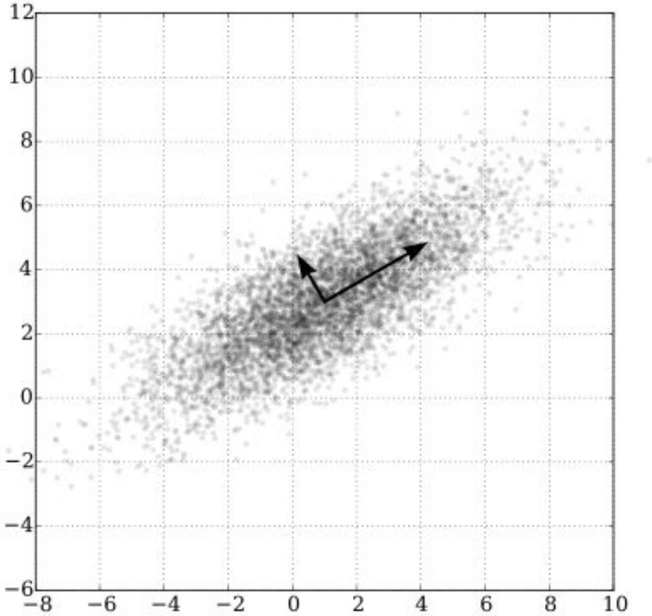


Q1: What features are learned by contrastive learning?

- Contrastive Loss $\min_{\phi \in \Phi} \mathbb{E}_{(x,x^+,x^-)\sim\mathcal{D}_{pre}} \left[\ell\left(\phi(x)^\top(\phi(x^+) - \phi(x^-))\right)\right]$

Q1: What features are learned by contrastive learning?

- Contrastive Loss $\min_{\phi \in \Phi} \mathbb{E}_{(x,x^+,x^-)\sim\mathcal{D}_{pre}} \left[\ell\left(\phi(x)^\top(\phi(x^+) - \phi(x^-))\right)\right]$
- PCA on $\phi(x)$ $\min_{\phi \in \Phi} -\mathbb{E}_{x\sim\mathcal{D}}[\|\phi(x) - \mathbb{E}_{x'\sim\mathcal{D}}[\phi(x')]\|^2] = -\mathbb{E}_{x\sim\mathcal{D}}[\|\phi(x) - \phi_0\|^2]$
- $\phi_{z_R} := \mathbb{E}[\phi(x) \mid z_R] = \mathbb{E}[\phi(g(z)) \mid z_R]$



Principal Component Analysis

Figures from: *Wikipedia*

Q1: What features are learned by contrastive learning?

- Contrastive Loss $\min_{\phi\in\Phi} \mathbb{E}_{(x,x^+,x^-)\sim\mathcal{D}_{pre}}\left[\ell\left(\phi(x)^\top(\phi(x^+)-\phi(x^-))\right)\right]$
- PCA on $\phi(x)$ $\min_{\phi\in\Phi} -\mathbb{E}_{x\sim\mathcal{D}}[\|\phi(x)-\mathbb{E}_{x'\sim\mathcal{D}}[\phi(x')]\|^2] = -\mathbb{E}_{x\sim\mathcal{D}}[\|\phi(x)-\phi_0\|^2]$
- $\phi_{z_R} := \mathbb{E}[\phi(x) \mid z_R] = \mathbb{E}[\phi(g(z)) \mid z_R]$

Theorem (Contrastive Learning is Generalized Nonlinear PCA)
If $\ell(t) = -t$, Contrastive Learning is equivalent to PCA on ϕ_{z_R}.
Moreover, if ϕ is linear function, it is equivalent to linear PCA on ϕ_{z_R}.

Q1: What features are learned by contrastive learning?

- Contrastive Loss $\min_{\phi \in \Phi} \mathbb{E}_{(x,x^+,x^-)\sim\mathcal{D}_{pre}} \left[\ell\left(\phi(x)^\top(\phi(x^+) - \phi(x^-))\right)\right]$
- PCA on $\phi(x)$ $\min_{\phi \in \Phi} -\mathbb{E}_{x\sim\mathcal{D}}[\|\phi(x) - \mathbb{E}_{x'\sim\mathcal{D}}[\phi(x')]\|^2] = -\mathbb{E}_{x\sim\mathcal{D}}[\|\phi(x) - \phi_0\|^2]$
- $\phi_{z_R} := \mathbb{E}[\phi(x) \mid z_R] = \mathbb{E}[\phi(g(z)) \mid z_R]$

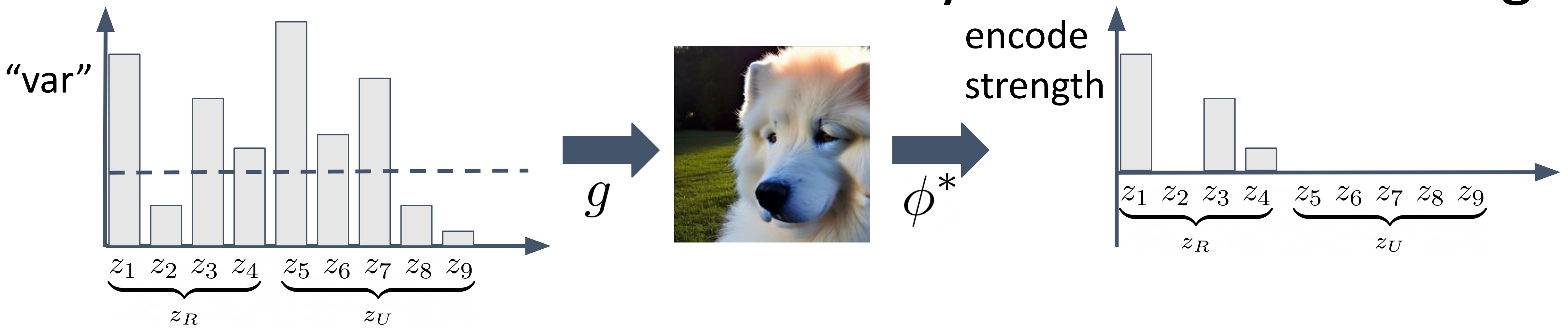
Theorem (Encode Invariant Feature; Remove Spurious Feature)

If $\ell(t)$ is convex, decrease, lower-bound, and $z_R \to x$ is one-to-one, with regular assumption, the optimal representation ϕ^* satisfies:

(1) ϕ^* does not encode spurious feature: $\phi^* \circ g(z) \perp z_U$

(2) ϕ^* only encodes invariant feature whose "variance" large enough, and encoding strength increases when "variance" becomes larger.

Q1: What features are learned by contrastive learning?



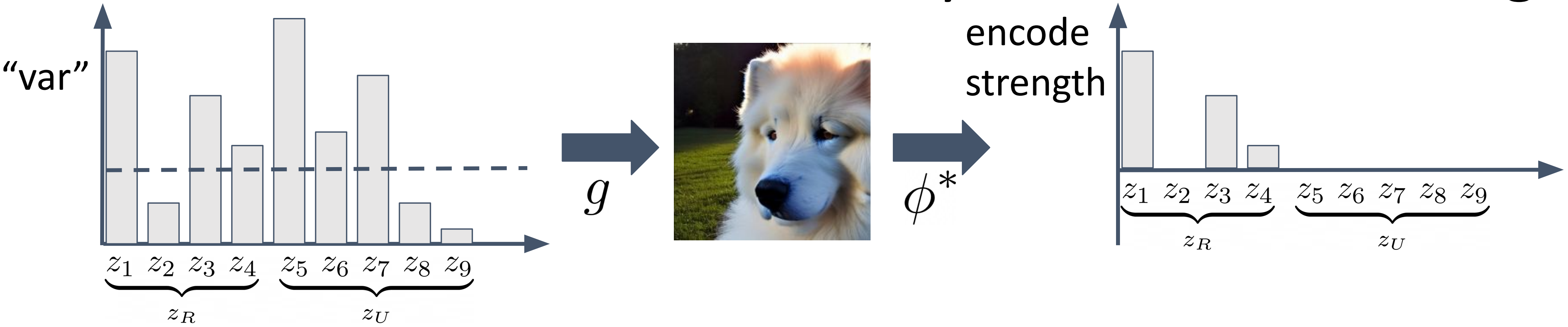
Theorem (Encode Invariant Feature; Remove Spurious Feature)

If $\ell(t)$ is convex, decrease, lower-bound, and $z_R \to x$ is one-to-one, with regular assumption, the optimal representation ϕ^* satisfies:

(1) ϕ^* does not encode spurious feature: $\phi^* \circ g(z) \perp z_U$

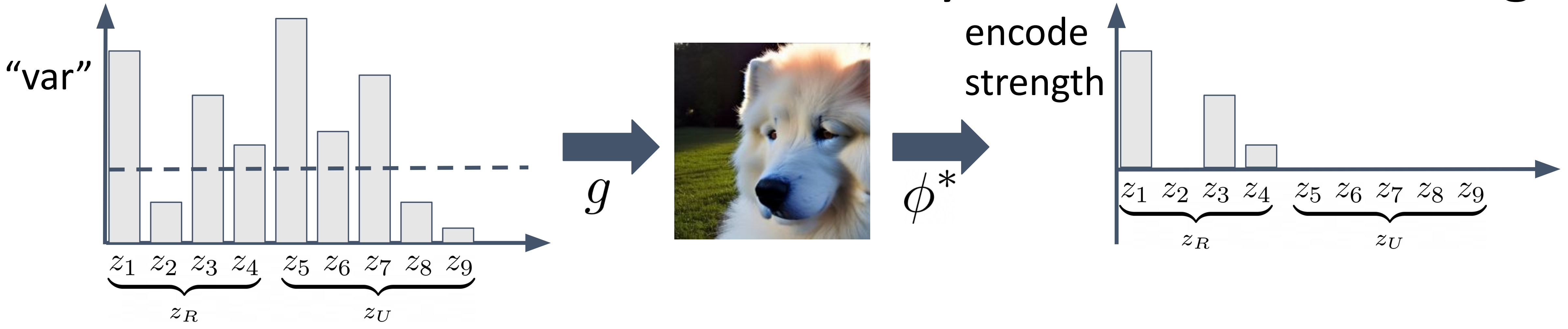
(2) ϕ^* only encodes invariant feature whose "variance" large enough, and encoding strength increases when "variance" becomes larger.

Q1: What features are learned by contrastive learning?

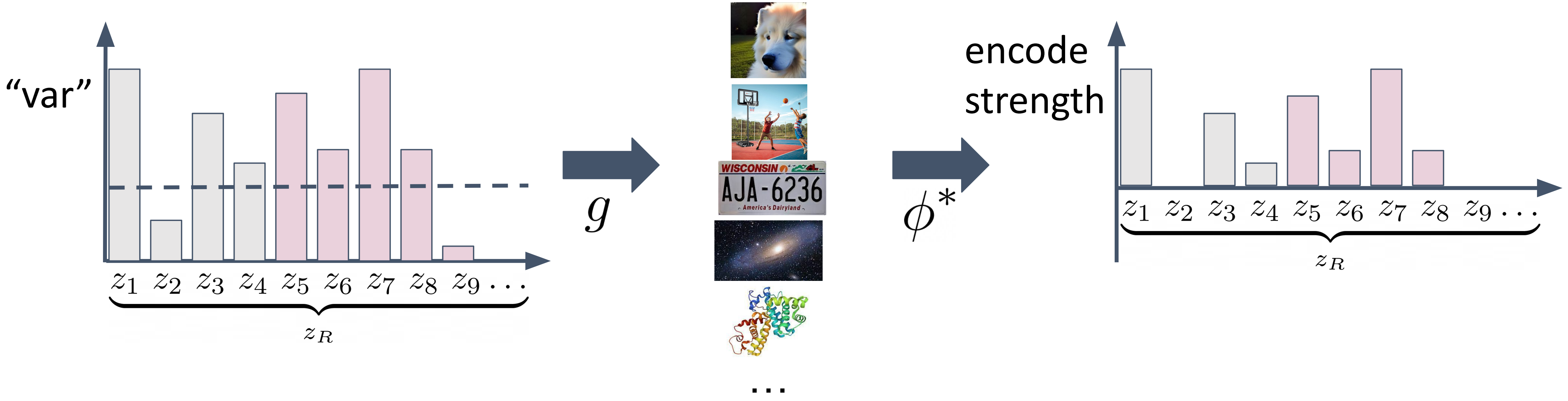


Q2: Is it always good to encode almost **all** invariant features?

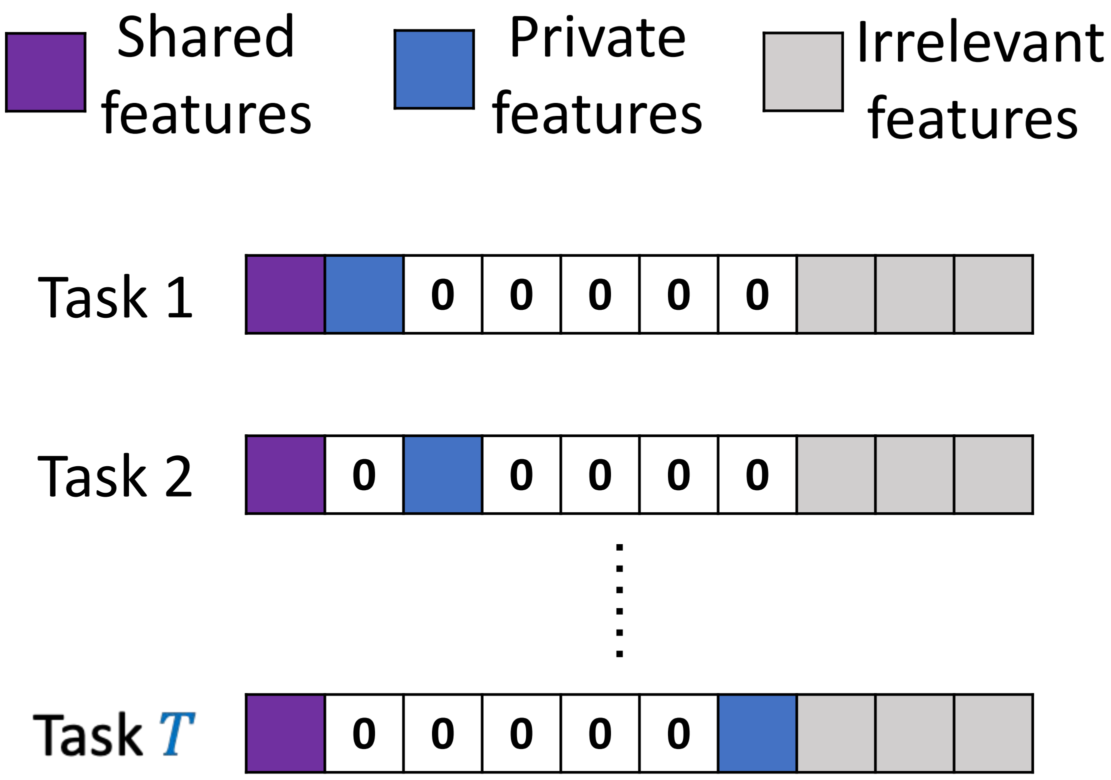
Q1: What features are learned by contrastive learning?



Q2: Is it always good to encode almost **all** invariant features?



A2: Trade-off Comes from Feature Weighting

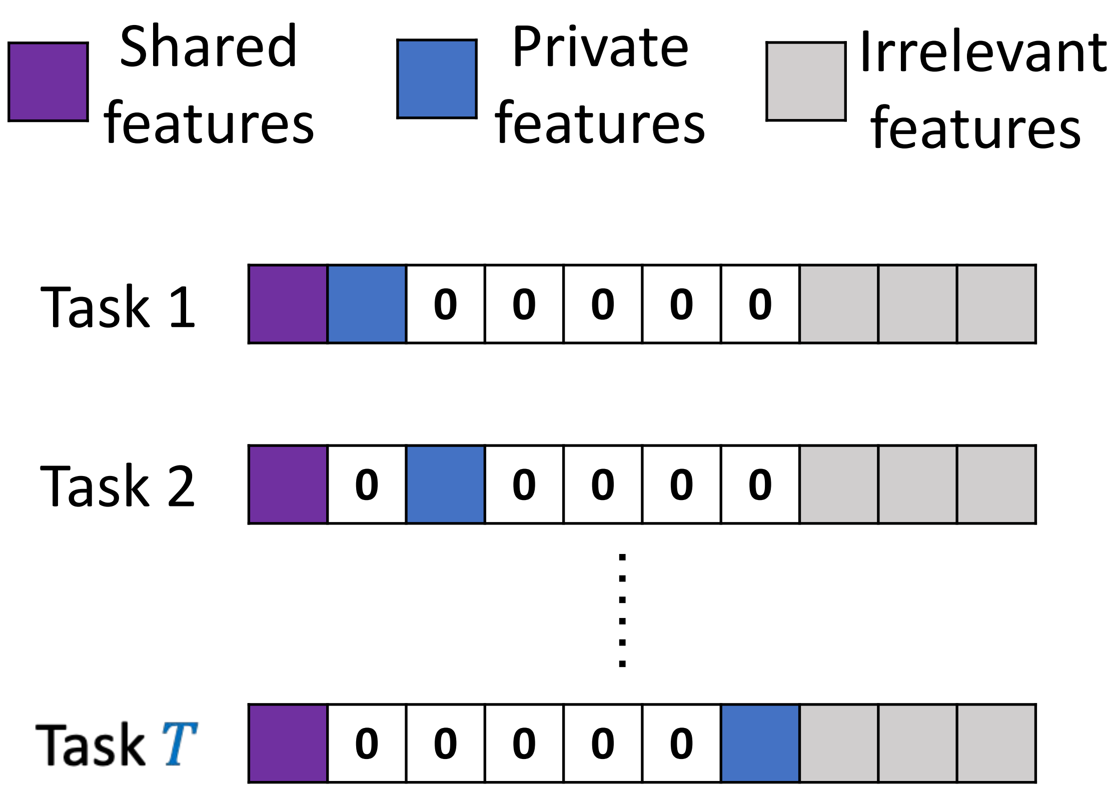


- Input: linearly generated from features
- Label: linear on shared/private features

- Pre-train a linear representation and then learn linear classifiers
- Best representation: weight shared/private features equally
- Pre-trained on only Task 1 v.s. Pre-trained on mixture of all tasks

[SCL+23] Shi, Chen, Li, Raghuram, Wu, Liang, Jha. The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning. ICLR'2023.

A2: Trade-off Comes from Feature Weighting



- Input: linearly generated from features
- Label: linear on shared/private features

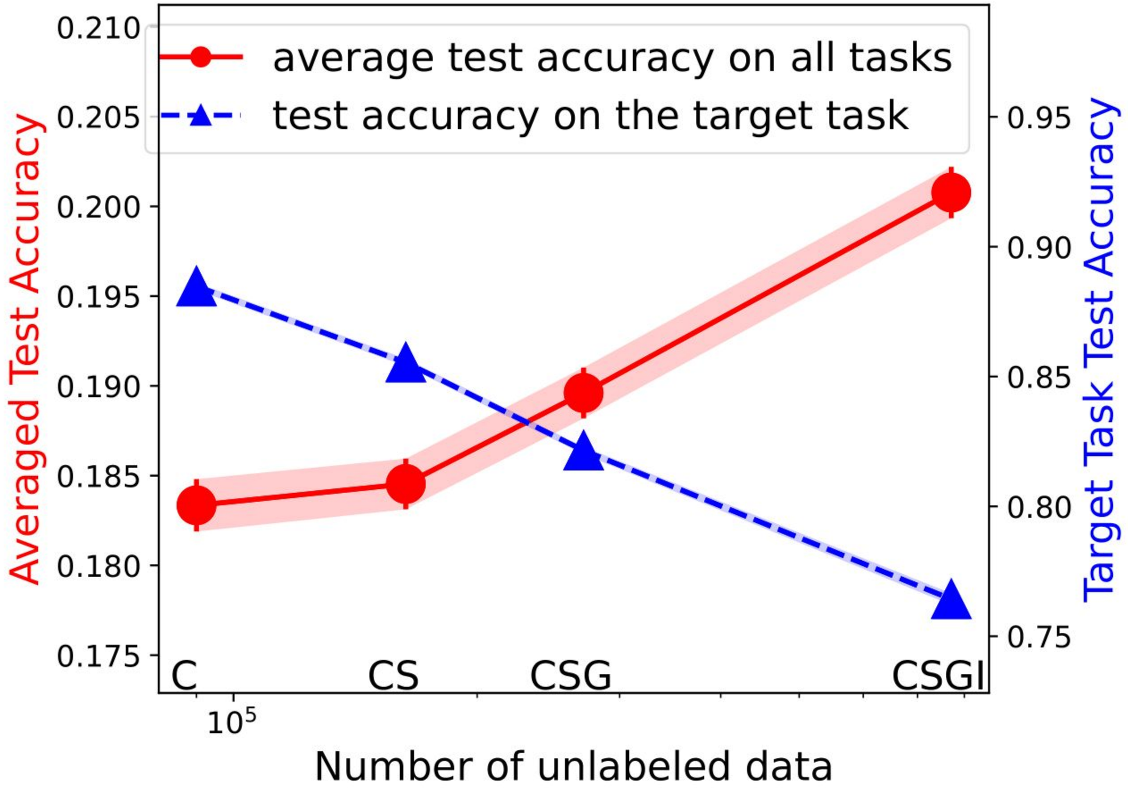
- Pre-trained on Task 1:
  - Recover features for Task 1 but not for others
  - Good prediction on Task 1 but not on others
- Pre-trained on mixture of all tasks:
  - Recover all shared/private features
  - Up-weights the shared features by $O(\sqrt{T})$
  - $O(\sqrt{T})$ worse on Task 1 but better on average

[SCL+23] Shi, Chen, Li, Raghuram, Wu, Liang, Jha. The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning. ICLR'2023.

Trade-off of Label Efficiency and Universality

Contrastive learning ResNet18 backbone via MoCo, then classify on CIFAR10.

From left to right, incrementally add to pre-training: CINIC-10 (C), SVHN (S), GTSRB (G), and ImageNet32 (I)

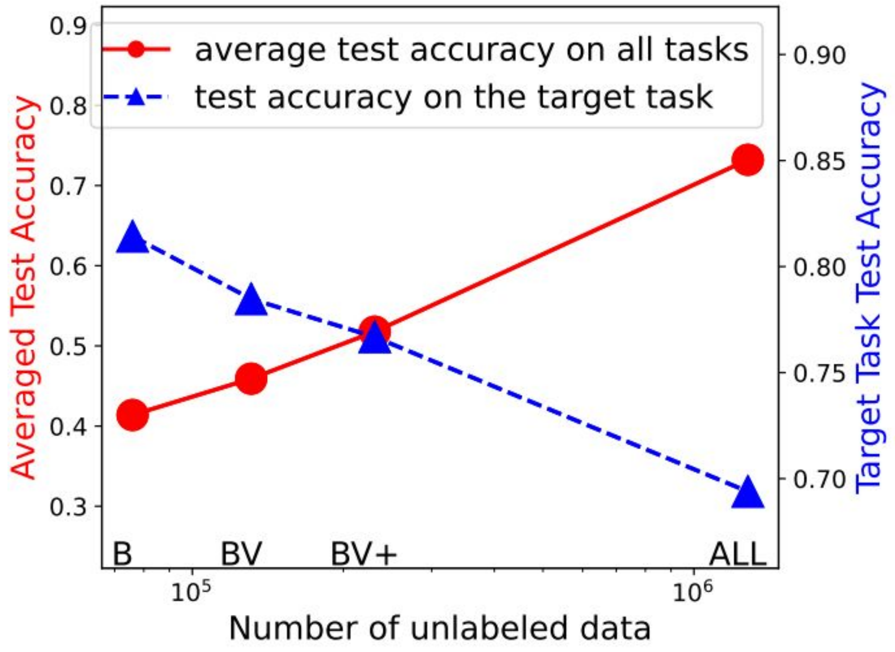


[SCL+23] Shi, Chen, Li, Raghuram, Wu, Liang, Jha. The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning. ICLR'2023.

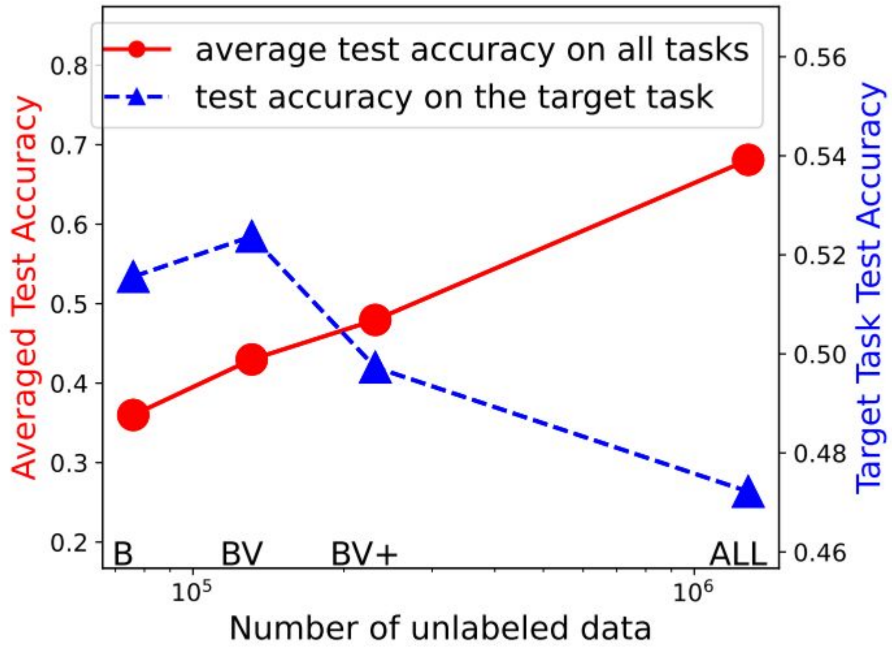
Trade-off of Label Efficiency and Universality

Contrastive learning, then classify on ImageNet-Bird (B).

From left to right, incrementally add to pre-training: ImageNet-Bird (B), ImageNet-Vehicle (V), ImageNet-Cat/Ball/Shop/Clothing/Fruit (+), and ImageNet (ALL)



(a) MoCo v3 (backbone ViT-S)



(b) SimSiam (backbone ResNet50)

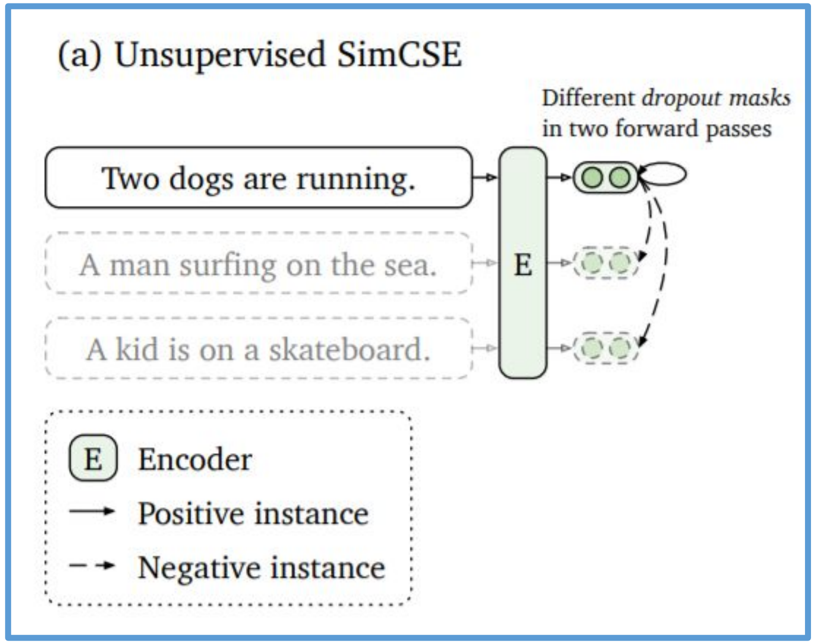
[SCL+23] Shi, Chen, Li, Raghuram, Wu, Liang, Jha. The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning. ICLR'2023.

Solution 1 - Contrastive Regularization

$$\ell_c(f(\phi(x)), y) \;+\; \frac{\lambda}{|\mathcal{R}|} \sum_{(\tilde{x}, \tilde{x}^+, \tilde{x}^-) \in \mathcal{R}} \ell\left(\phi(\tilde{x})^\top (\phi(\tilde{x}^+) - \phi(\tilde{x}^-))\right)$$

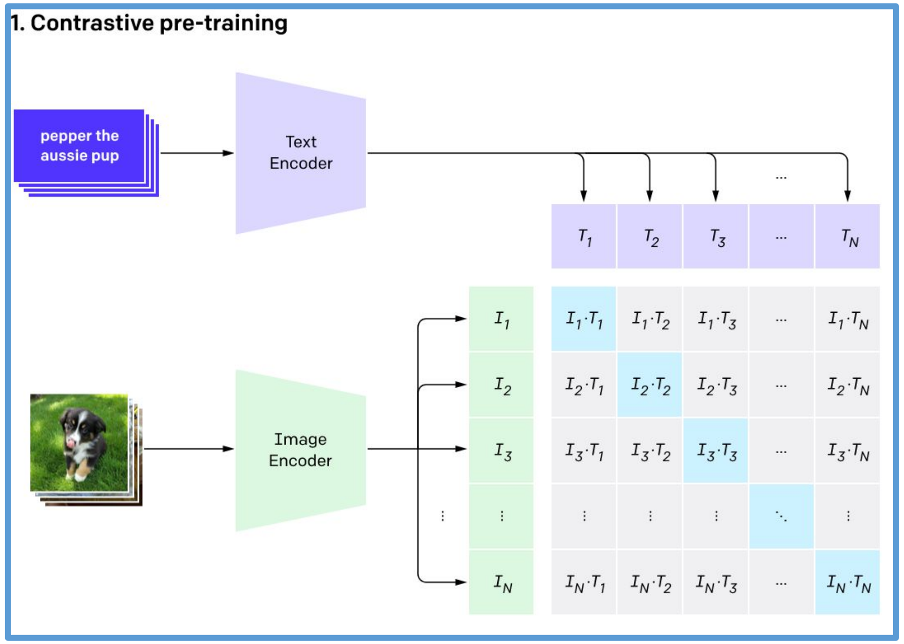
Solution 1 - Contrastive Regularization

$$\ell_c(f(\phi(x)), y) \; + \; \frac{\lambda}{|\mathcal{R}|} \sum_{(\tilde{x}, \tilde{x}^+, \tilde{x}^-) \in \mathcal{R}} \ell\left(\phi(\tilde{x})^\top (\phi(\tilde{x}^+) - \phi(\tilde{x}^-))\right)$$



SimCSE - (Text, Text)

Figures from: *SimCSE: Simple Contrastive Learning of Sentence Embeddings, 2021.*

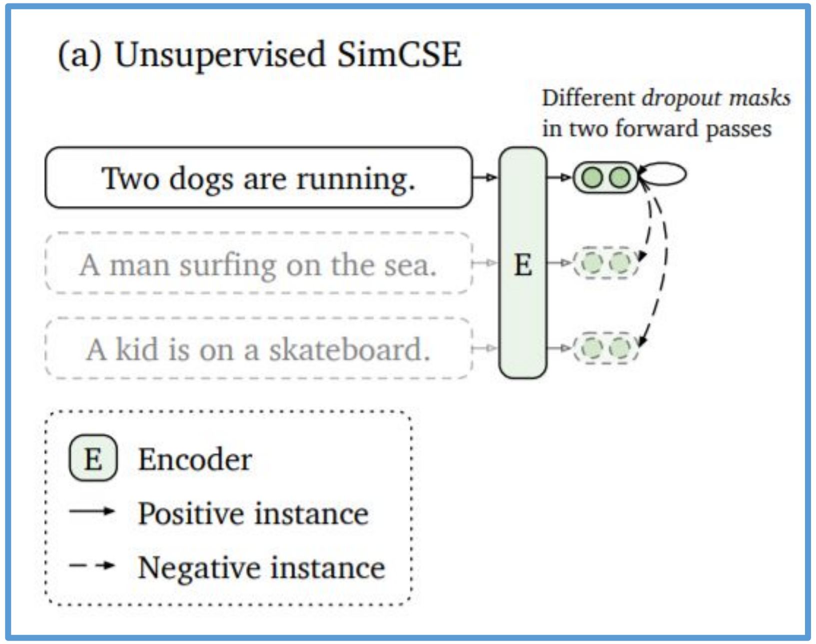


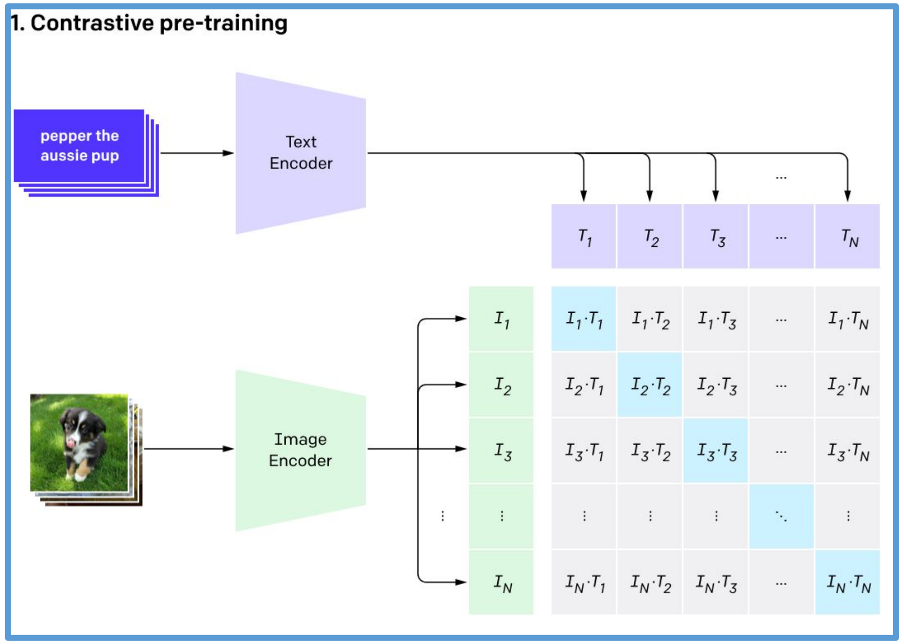
CLIP - (Image, Text)

Figures from: *Learning Transferable Visual Models From Natural Language Supervision, 2021.*

Solution 1 - Contrastive Regularization

$$\ell_c(f(\phi(x)), y) \; + \; \frac{\lambda}{|\mathcal{R}|} \sum_{(\tilde{x}, \tilde{x}^+, \tilde{x}^-) \in \mathcal{R}} \ell\left(\phi(\tilde{x})^\top (\phi(\tilde{x}^+) - \phi(\tilde{x}^-))\right)$$





SimCSE - (Text, Text)

Figures from: *SimCSE: Simple Contrastive Learning of Sentence Embeddings, 2021.*

CLIP - (Image, Text)

Figures from: *Learning Transferable Visual Models From Natural Language Supervision, 2021.*

Method	CLIP			MoCo v3			SimCSE	
	ImageNet	SVHN	GTSRB	CIFAR-10	SVHN	GTSRB	IMDB	AGNews
LP	77.84±0.02	63.44±0.01	86.56±0.01	95.82±0.01	61.92±0.01	75.37±0.01	86.49±0.16	87.76±0.66
FT	83.65±0.01	78.22±0.18	90.74±0.06	96.17±0.12	65.36±0.33	76.45±0.29	92.31±0.26	93.57±0.23
Ours	**84.94**±0.09	**78.72**±0.37	**92.01**±0.28	**96.71**±0.10	**66.29**±0.20	**81.28**±0.10	**92.85**±0.03	**93.94**±0.02

Solution 2 - Few-Shot Multitask Finetuning

- CLIP pretraining model.
- Few-Shot: 5 training sample for each target class.
- Few-Shot Multitask finetuning on mini-Imagenet training classes.
- Few-Shot evaluate on mini-Imagenet target classes.

[XSW+23] Xu, Shi, Wei, Li, Liang. Improving Foundation Models for Few-Shot Learning via Multitask Finetuning. ICLR'2023 workshop.

Solution 2 - Few-Shot Multitask Finetuning

- CLIP pretraining model.
- Few-Shot: 5 training sample for each target class.
- Few-Shot Multitask finetuning on mini-Imagenet training classes.
- Few-Shot evaluate on mini-Imagenet target classes.

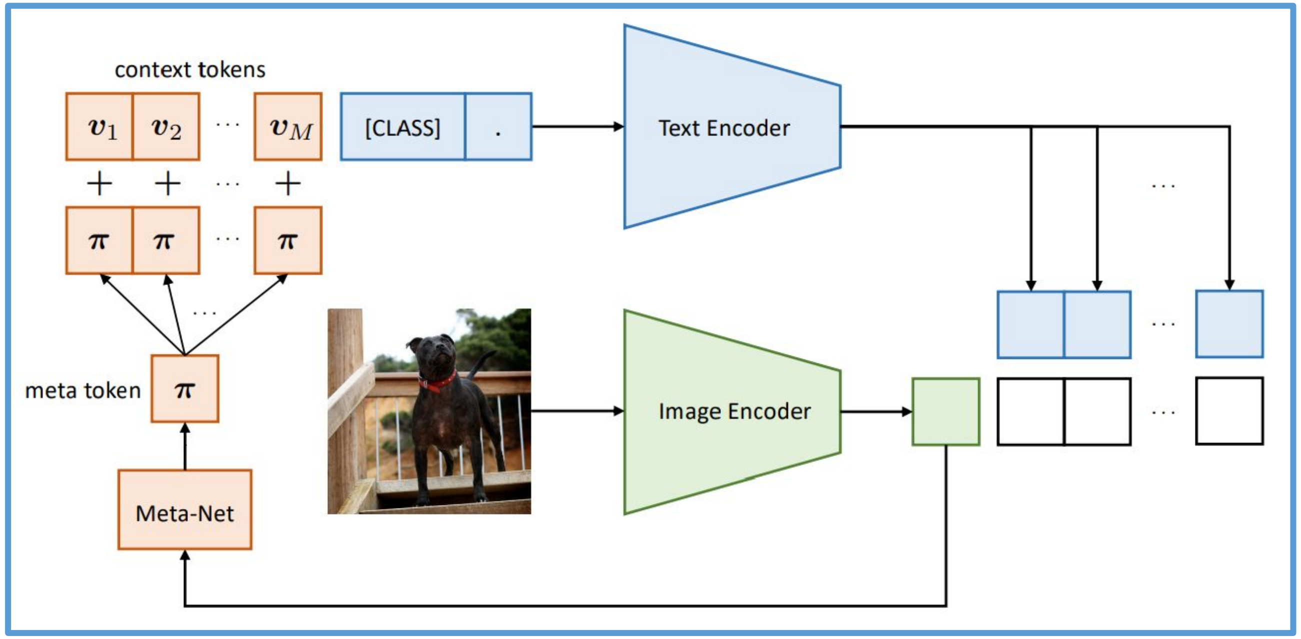
Backbone	Direct Adaptation	Finetuning
ViT-B32	83.03 ± 0.24	$\mathbf{89.07} \pm 0.20$
ResNet50	78.36 ± 0.25	$\mathbf{81.19} \pm 0.25$

Table 1: Effects of multitask finetuning.

[XSW+23] Xu, Shi, Wei, Li, Liang. Improving Foundation Models for Few-Shot Learning via Multitask Finetuning. ICLR'2023 workshop.

Future Work

- Any other better paradigm than Contrastive Learning + Linear Probing? May be Prompt?

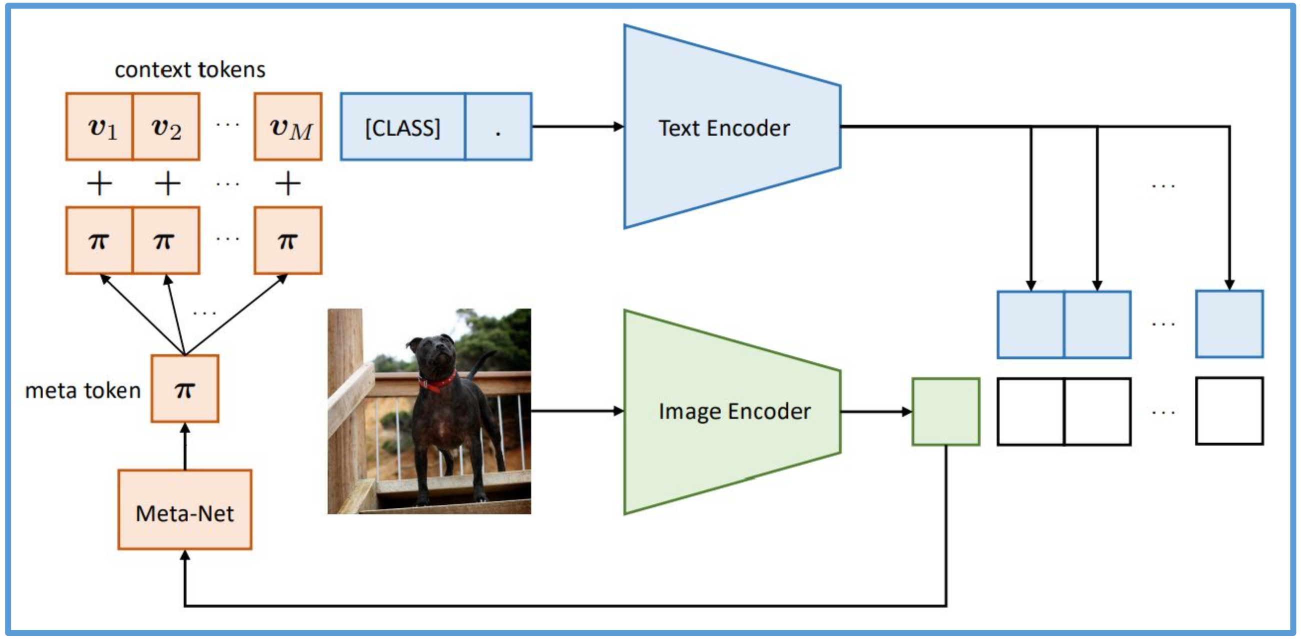


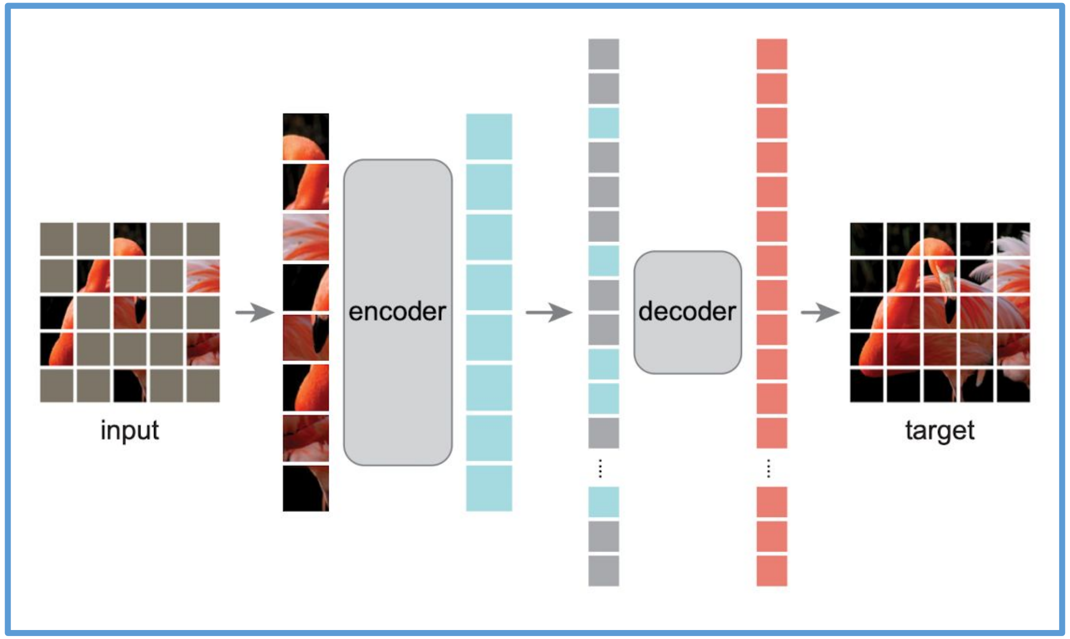
CoCoOp

Figures from: *Conditional Prompt Learning for Vision-Language Models, 2022.*

Future Work

- Any other better paradigm than Contrastive Learning + Linear Probing? May be Prompt?
- Do the other self-supervised learning methods have a similar trade-off, e.g., MAE, GPT4?





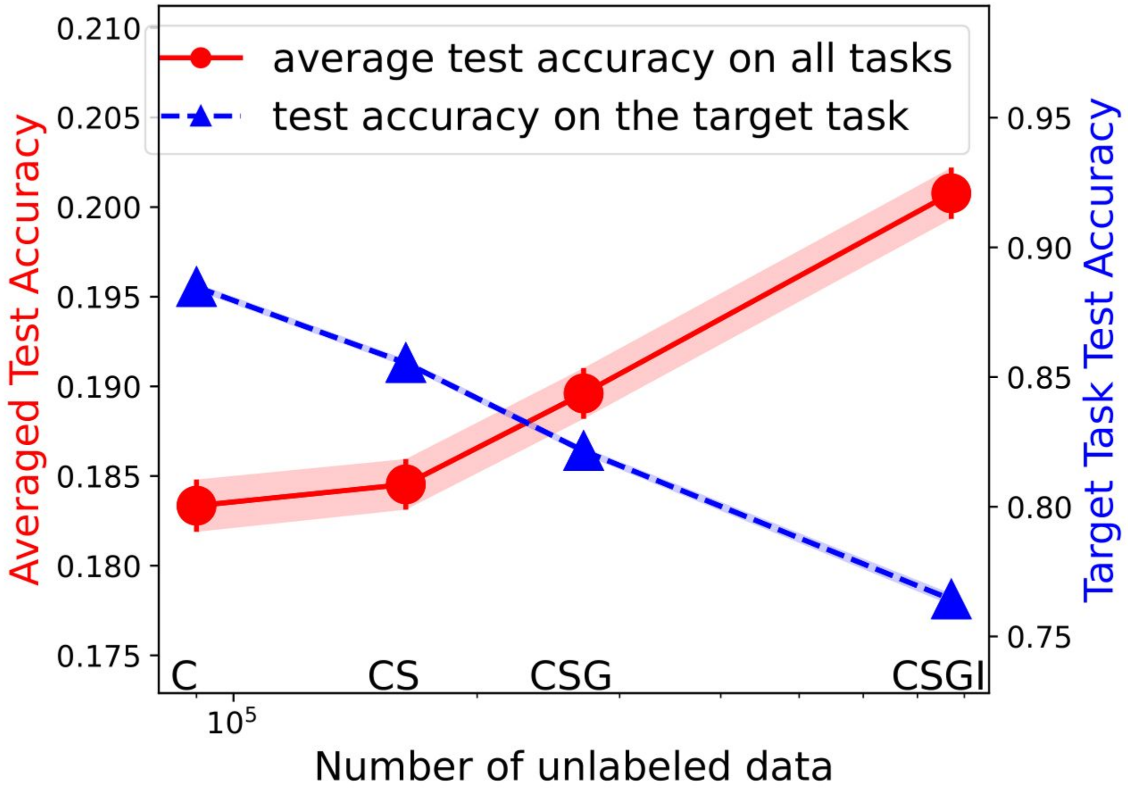
CoCoOp

Figures from: *Conditional Prompt Learning for Vision-Language Models, 2022.*

Mask Autoencoder

Figures from: *Masked Autoencoders Are Scalable Vision Learners, 2021.*

Take Home Message



Thanks!

MLOPT 2023

Theorem (Encode Invariant Feature; Remove Spurious Feature)
If $\ell(t)$ is convex, decrease, lower-bound, and $z_R \to x$ is one-to-one, with regular assumption, the optimal representation ϕ^* satisfies:
(1) ϕ^* does not encode spurious feature: $\phi^* \circ g(z) \perp z_U$
(2) ϕ^* only encodes invariant feature whose "variance" large enough, and encoding strength increases when "variance" becomes larger.

$$
\begin{aligned}
&\mathbb{E}_{(x,x^+,x^-)}\left[\ell\left(\phi(x)^\top[\phi(x^+)-\phi(x^-)]\right)\right] \\
&= \mathbb{E}_{(z,z^+,z^-)}\left[\ell\left((\phi\circ g(z))^\top(\phi\circ g(z^+)-\phi\circ g(z^-))\right)\right] \\
&= \mathbb{E}_{(z_R,z_R^-)}\left[\mathbb{E}\left[\ell\left((\phi\circ g(z))^\top(\phi\circ g(z^+)-\phi\circ g(z^-))\right) \mid z_R, z_R^-\right]\right] \\
&\geq \mathbb{E}_{(z_R,z_R^-)}\left[\ell\left(\mathbb{E}\left[(\phi\circ g(z))^\top(\phi\circ g(z^+)-\phi\circ g(z^-)) \mid z_R, z_R^-\right]\right)\right] \\
&= \mathbb{E}_{(z_R,z_R^-)}\left[\ell\left(\mathbb{E}[\phi\circ g(z)\mid z_R]^\top\left(\mathbb{E}[\phi\circ g(z^+)\mid z_R]-\mathbb{E}[\phi\circ g(z^-)\mid z_R^-]\right)\right)\right] \\
&= \mathbb{E}_{(z_R,z_R^-)}\left[\ell\left(\phi_{z_R}^\top\phi_{z_R}-\phi_{z_R}^\top\phi_{z_R^-}\right)\right],
\end{aligned}
$$