New Paradigm: Pre-trained Representations

- Paradigm shift: supervised learning $\rightarrow$ pre-training + adaptation
New Paradigm: Pre-trained Representations

• Paradigm shift: supervised learning → pre-training + adaptation
New Paradigm: Pre-trained Representations

• Paradigm shift: supervised learning $\rightarrow$ pre-training + adaptation
New Paradigm: Pre-trained Representations

- Paradigm shift: supervised learning $\rightarrow$ pre-training + adaptation
New Paradigm: Pre-trained Representations

- Paradigm shift: supervised learning $\rightarrow$ pre-training + adaptation

The history and evolution of pre-trained models

New Paradigm: Pre-trained Representations

• Paradigm shift: supervised learning $\rightarrow$ pre-training + adaptation

Label Efficiency

• Great performance with limited labeled data in downstream tasks

How to quantify the benefit of knowledge transfer?

• Pre-train $h \in \mathcal{H}$, then learn a classifier $f \in \mathcal{F}$ to get final model $f \circ h$

• Pre-train minimizes an unsupervised loss to $\leq \epsilon_{pre}$

• Without pre-train: $\mathcal{F} \circ \mathcal{H}$

Label Efficiency

• Great performance with limited labeled data in downstream tasks

How to quantify the benefit of knowledge transfer?

• Pre-train $h \in \mathcal{H}$, then learn a classifier $f \in \mathcal{F}$ to get final model $f \circ h$

• Pre-train minimizes an unsupervised loss to $\leq \epsilon_{pre}$

• Without pre-train: $\mathcal{F} \circ \mathcal{H}$

• With pre-train: $\mathcal{F} \circ \mathcal{H}_{\leq \epsilon_{pre}}$

Universality

- Generally applicable to different tasks

Figure credit: James John Williams, digt.com
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).

![Graph showing trade-off between universality and label efficiency](image)

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