# Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution



Ananya Kumar


Aditi Raghunathan


Robbie Jones


Tengyu Ma


Percy Liang

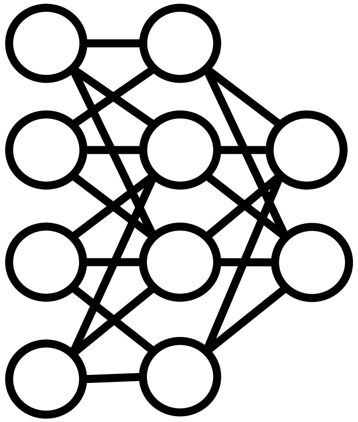
# Motivation: how to use pretrained models?

- Pretrained models like CLIP, SimCLR, BERT, are very useful
- Lots of research on improving these models
- This work: how should we adapt these models properly?

# Setting: Pretrain-Transfer-Test

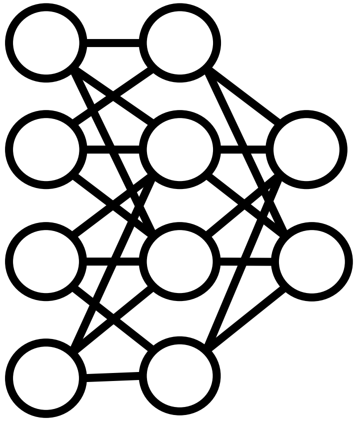
## Pretrained features

Inputs



Features

## Satellite remote sensing



"In-distribution" training data

*North America*

"In-distribution" (ID) test data

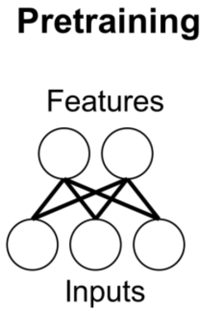
*North America*

"Out-of-distribution" (OOD) test data

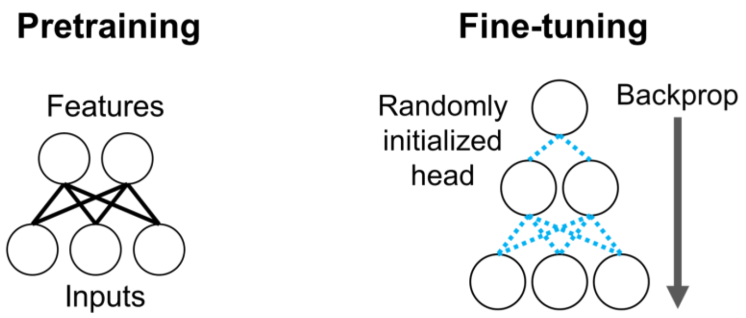
*Africa*

Better than no pretraining (Hendrycks et al 2019, Chen et al 2020, Xie et al 2021, Miller et al 2021)

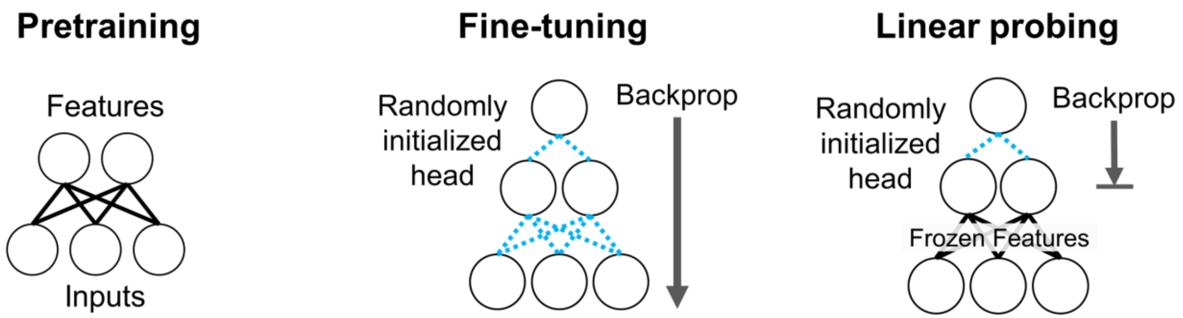
# Linear Probing vs. Fine-tuning



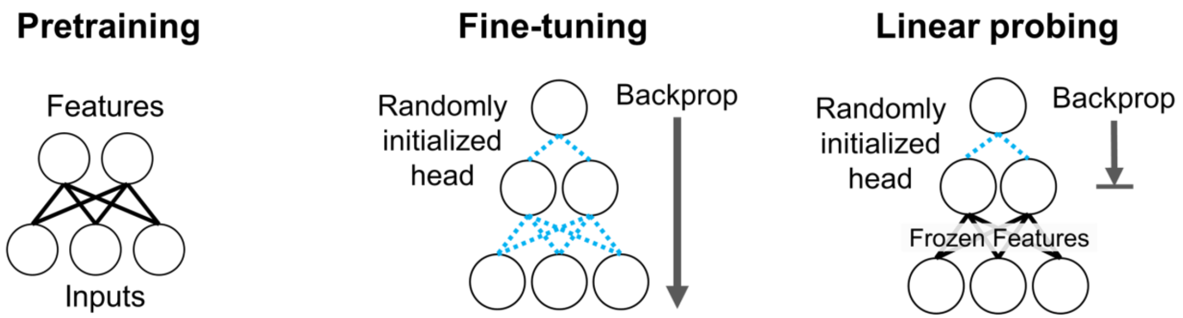
# Linear Probing vs. Fine-tuning



# Linear Probing vs. Fine-tuning



# Linear Probing vs. Fine-tuning




Which method does better?

# Pop Quiz: Background, Living-17

- Breeds Living-17: task is to classify image into animal such as bear (ID contains black bears, sloth bears; OOD has brown bears, polar bears)
- Pretrained model: MoCo-V2, seen *unlabeled* ImageNet images (including various types of bears)

# Pop Quiz: Background, Living-17

| | **Cat** | **Ape** | **Bear** |
|---|---|---|---|
| ID |  |  |  |
| OOD |  |  |  |

# Pop Quiz: Living-17

| Living-17 | ID | OOD |
|---|---|---|
| Scratch | 92.4% | 58.2% |
| Linear Probing | 96.5% | ? |
| Fine-Tuning | 97.1% | |

(a) LP < Scratch (b) Scratch < LP

# Pop Quiz: Living-17

| **Living-17** | **ID** | **OOD** |
|---|---|---|
| Scratch | 92.4% | 58.2% |
| Linear Probing | 96.5% | 82.2% |
| Fine-Tuning | 97.1% | |

(a) LP < Scratch **(b) Scratch < LP**

# Pop Quiz: Living-17

| **Living-17** | **ID** | **OOD** |
| --- | --- | --- |
| Scratch | 92.4% | 58.2% |
| Linear Probing | 96.5% | 82.2% |
| Fine-Tuning | 97.1% | ? |

(a) FT < Scratch (b) Scratch < FT < LP (c) LP < FT

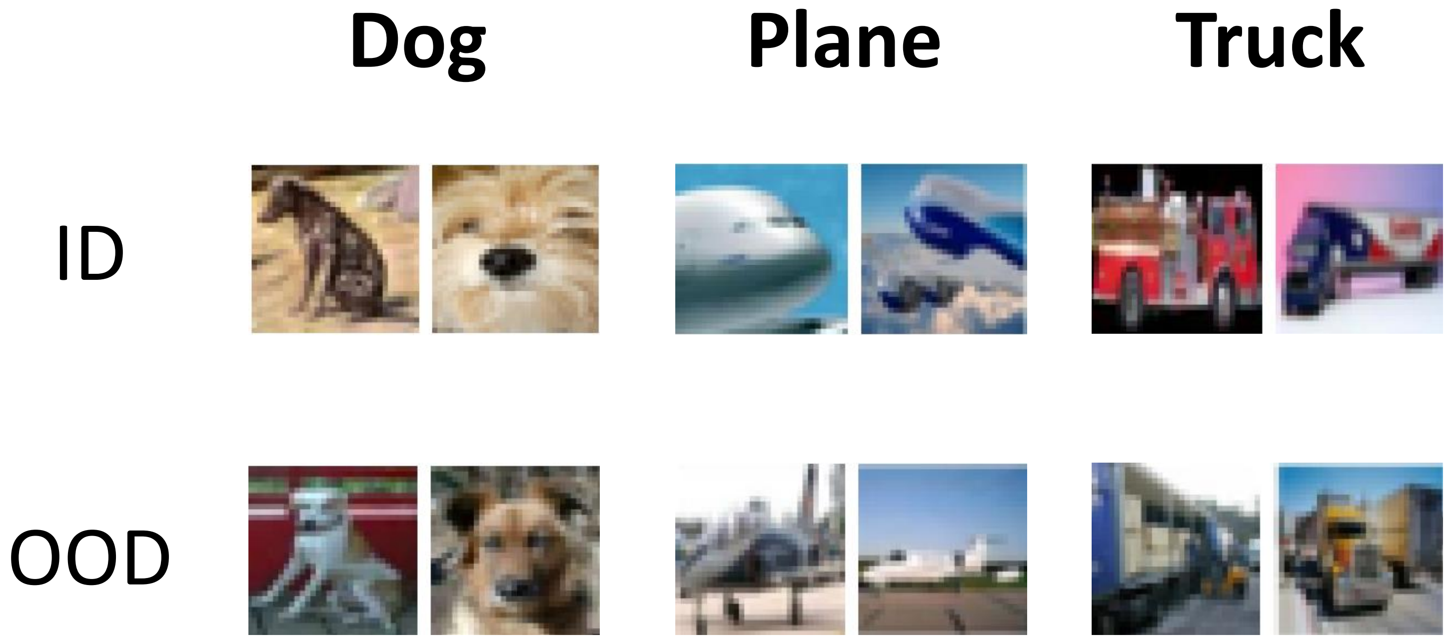
# Pop Quiz: Living-17

| **Living-17** | **ID** | **OOD** |
|---|---|---|
| Scratch | 92.4% | 58.2% |
| Linear Probing | 96.5% | 82.2% |
| Fine-Tuning | 97.1% | 77.7% |

(a) FT < Scratch **(b) Scratch < FT < LP** (c) LP < FT

# Pop Quiz: Background, CIFAR-10.1

- ID = CIFAR-10, OOD = CIFAR-10.1: Dataset collected using a similar protocol to CIFAR-10, “a minute distributional shift”



# Pop Quiz: CIFAR-10.1

| **CIFAR-10.1** | **ID** | **OOD** |
|---|---|---|
| Linear Probing | 91.8% | 82.7% |
| Fine-Tuning | 97.3% | ? |

(a) LP < FT

(b) FT < LP

# Pop Quiz: CIFAR-10.1

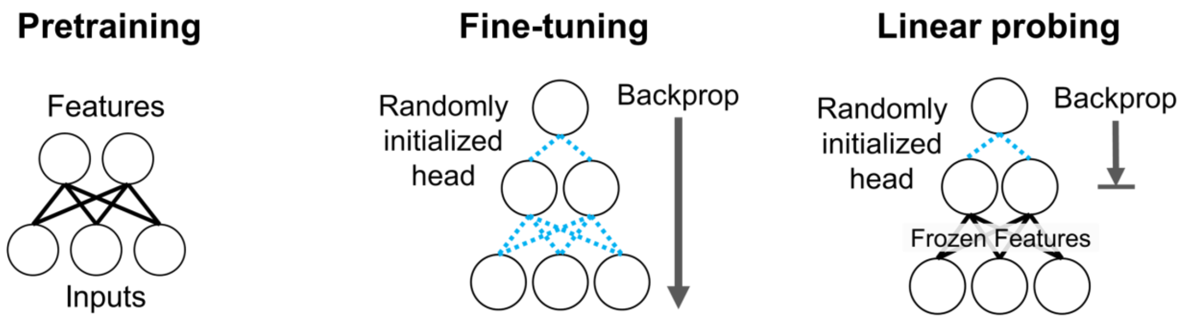
| **CIFAR-10.1** | **ID** | **OOD** |
|---|---|---|
| Linear Probing | 91.8% | 82.7% |
| Fine-Tuning | 97.3% | 92.3% |

**(a) LP < FT** (b) FT < LP

# Datasets

- CIFAR-10 → STL, CIFAR-10.1
- BREEDS Living-17 and BREEDS Entity-30
- DomainNet
- Functional Map of the World
- ImageNet → ImageNet-R, ImageNet-A, ImageNet-V2, ImageNet-Sketch

# Linear Probing vs. Fine-tuning




ID test

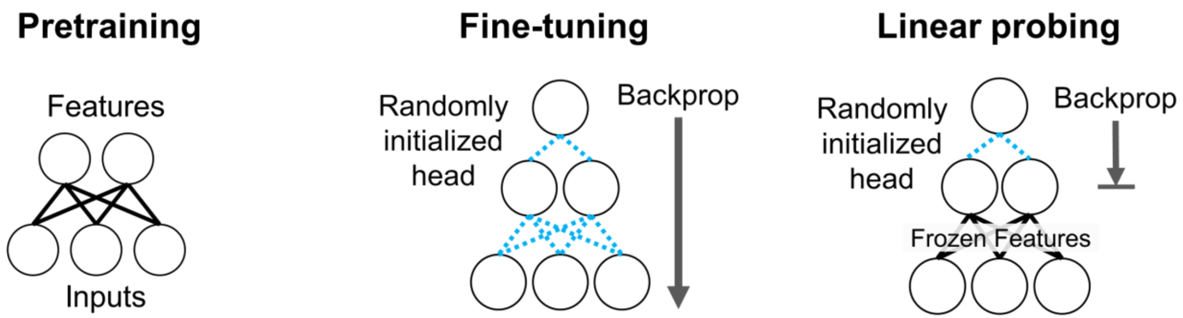
| | |
|---|---|
| 85.1% | 82.9% |

Average accuracies (7 ID datasets)

**Common wisdom is fine-tuning works better than linear probing**

(Kornblith et al 2019, Chen et al 2020, Zhai et al 2020, Chen et al 2021)

# Linear Probing vs. Fine-tuning

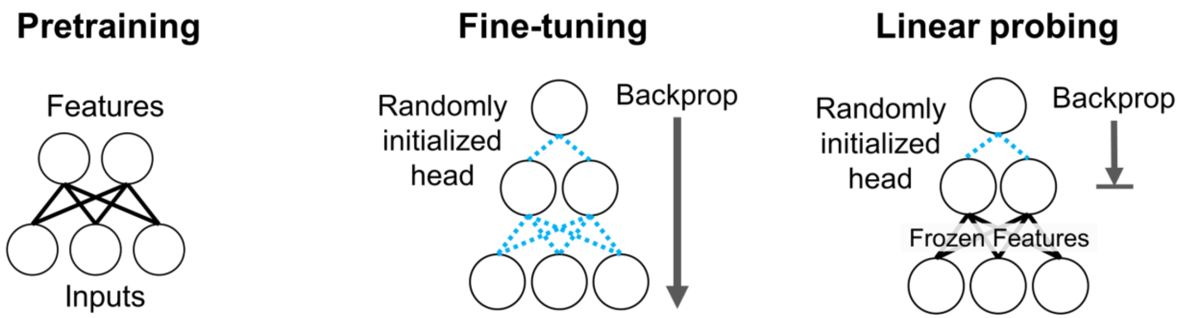



| | | |
|---|---|---|
| ID test  | 85.1% | 82.9% |
| OOD test  | 59.3% | 66.2% |

Average accuracies (10 datasets)

**Fine-tuning worse on 8/10 OOD datasets**

# Linear Probing vs. Fine-tuning




**Fine-tuning can often do worse out-of-distribution**

especially when the pretrained features are high quality and distribution shifts are large

# Outline

1. Fine-tuning can do worse than linear-probing OOD

2. Why fine-tuning can underperform OOD

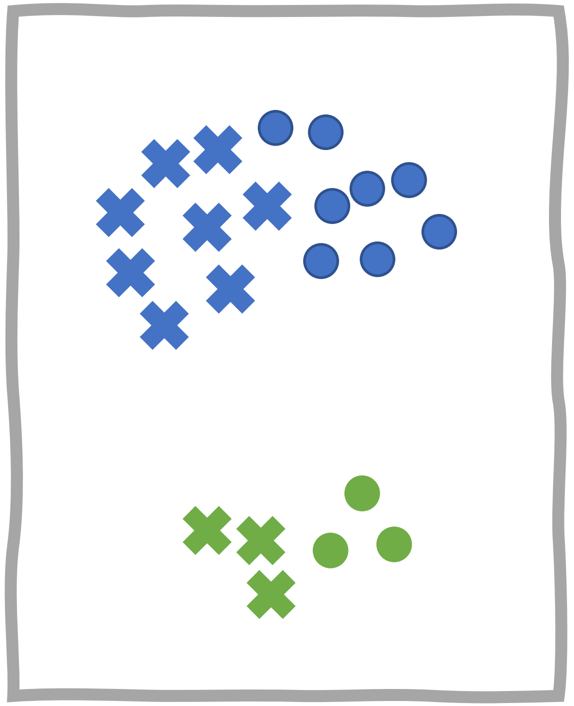
3. Simple change to fine-tuning: improved accuracy on 10 datasets

# Outline

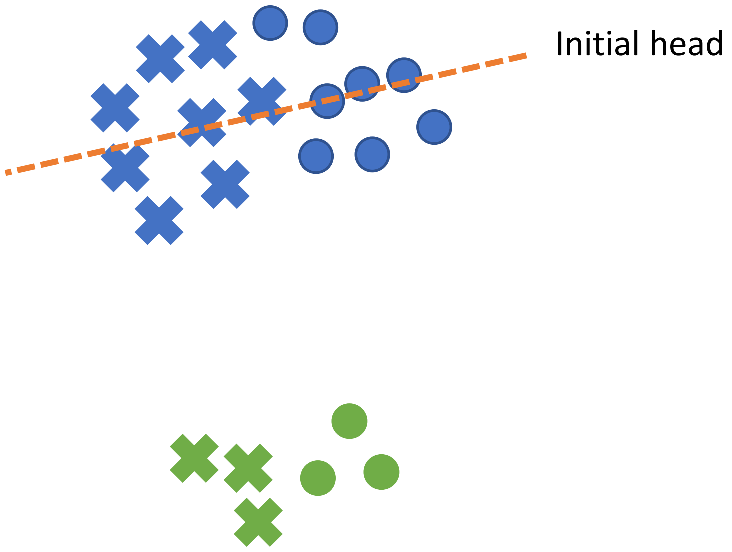
1. Fine-tuning can do worse than linear-probing OOD
2. **Why fine-tuning can underperform OOD**
3. Simple change to fine-tuning: improved accuracy on 10 datasets

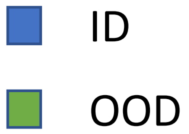
# Feature Distortion Theory

Pretrained Features



Fine-tuning: features for ID examples change in sync with the linear head

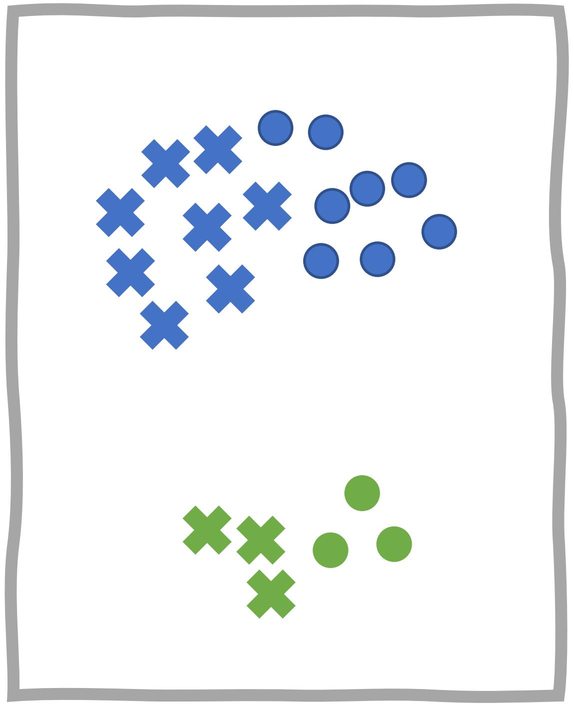


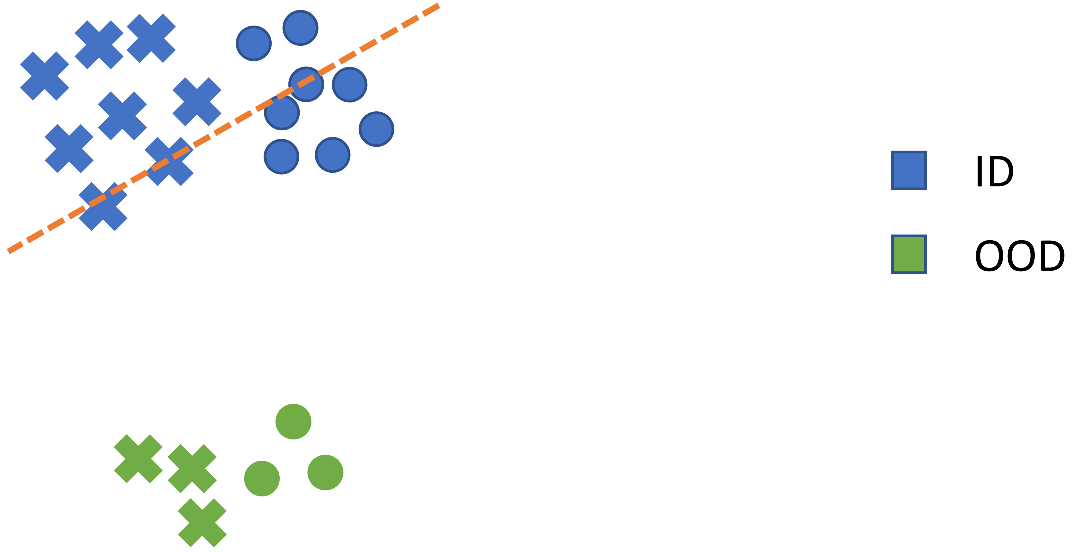

Features for OOD examples change less

# Feature Distortion Theory

Pretrained
Features



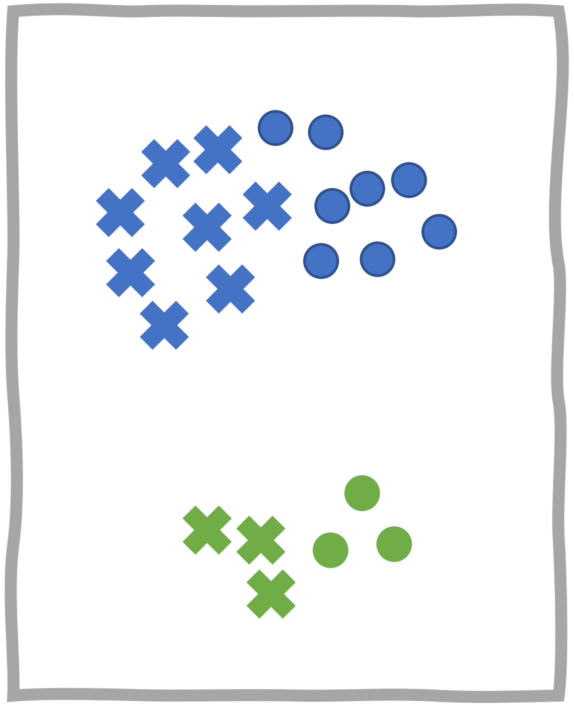
Fine-tuning: features for ID examples change
in sync with the linear head



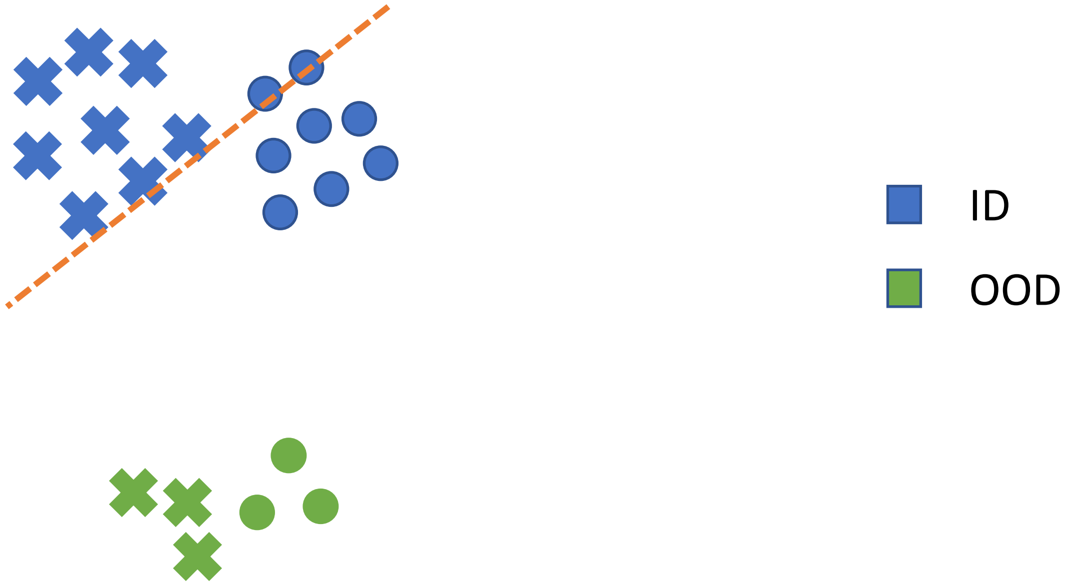

Features for OOD
examples change less

# Feature Distortion Theory

Pretrained
Features



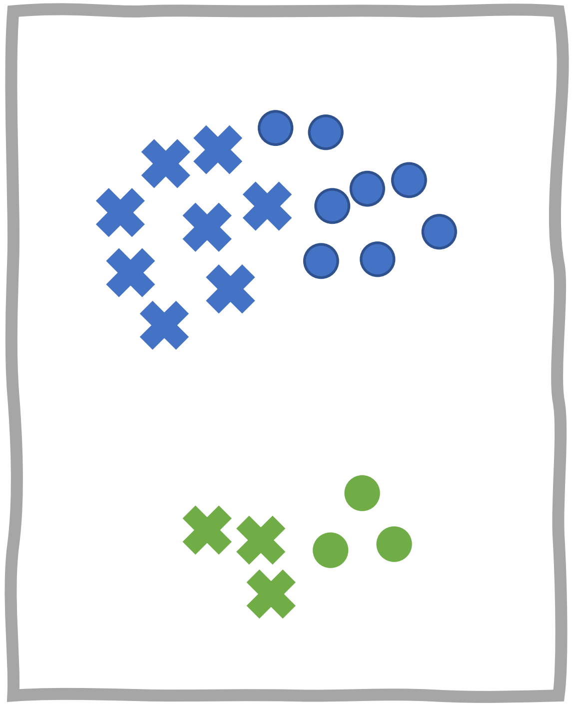
Fine-tuning: features for ID examples change
in sync with the linear head



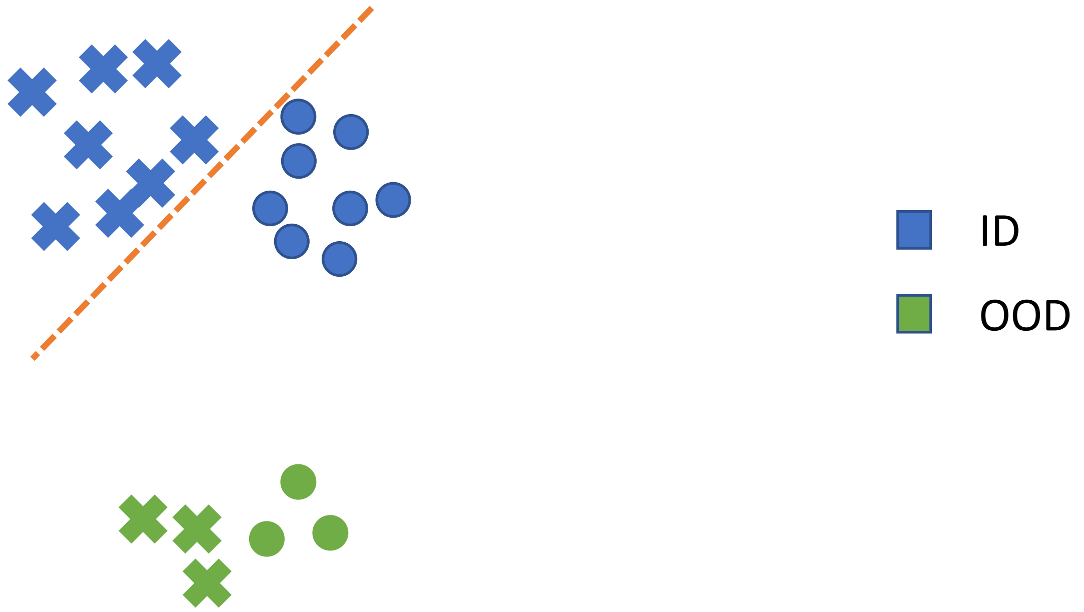

Features for OOD
examples change less

# Feature Distortion Theory

Pretrained
Features



Fine-tuning: features for ID examples change in sync with the linear head

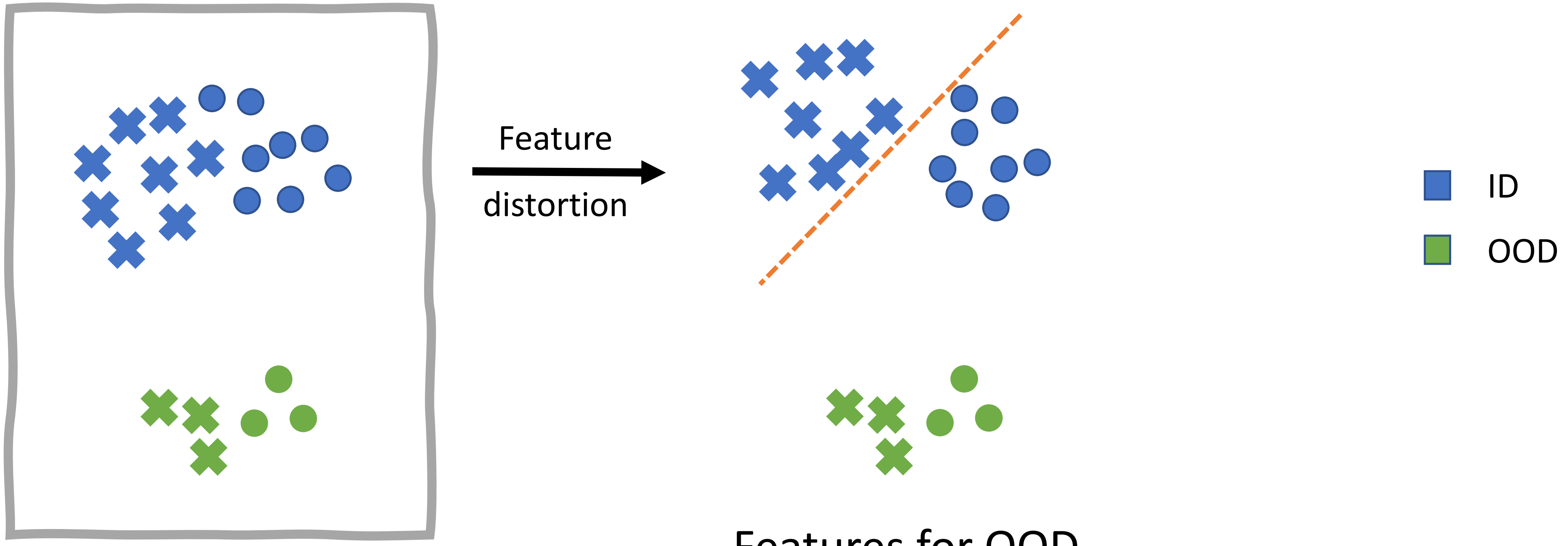



Features for OOD examples change less

# Feature Distortion Theory

Pretrained
Features

Fine-tuning: features for ID examples change
in sync with the linear head

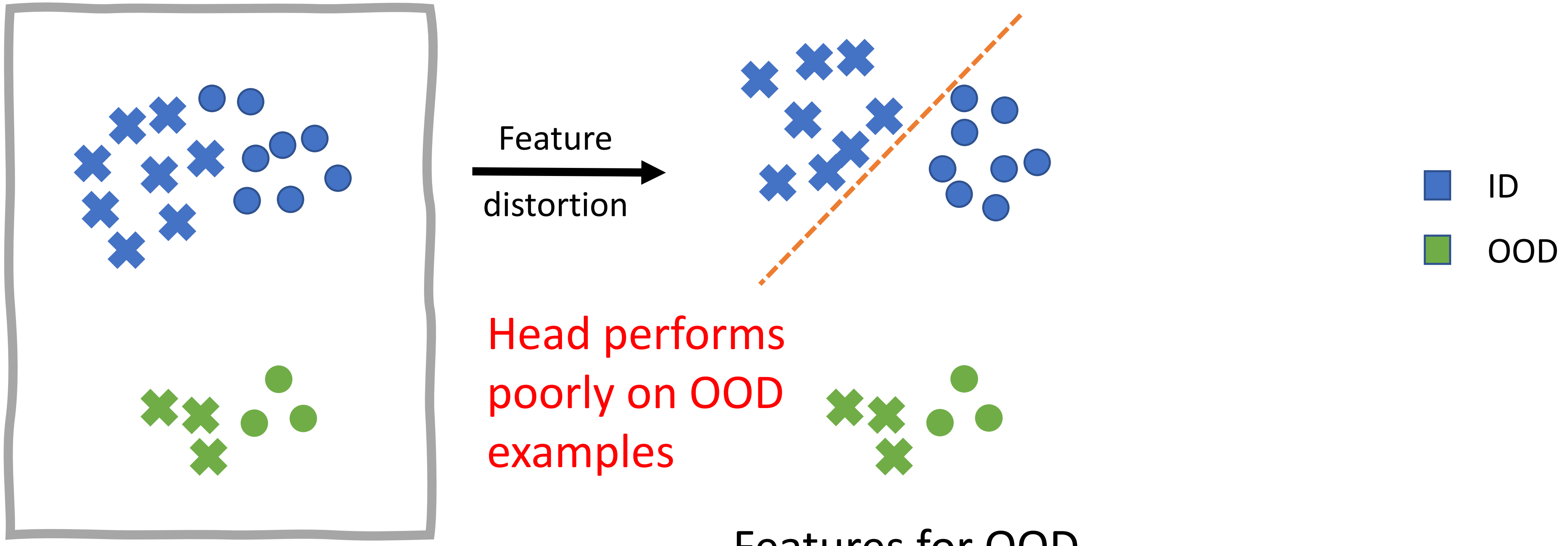



Features for OOD
examples change less

# Feature Distortion Theory

Pretrained
Features

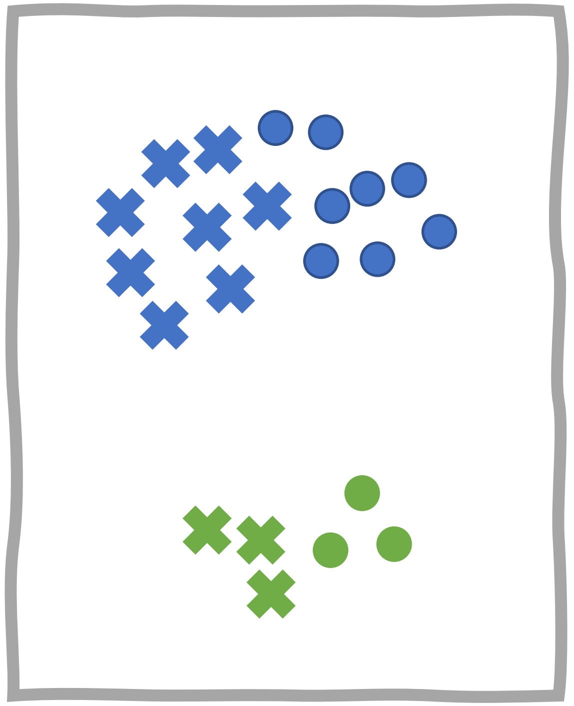
Fine-tuning: features for ID examples change
in sync with the linear head



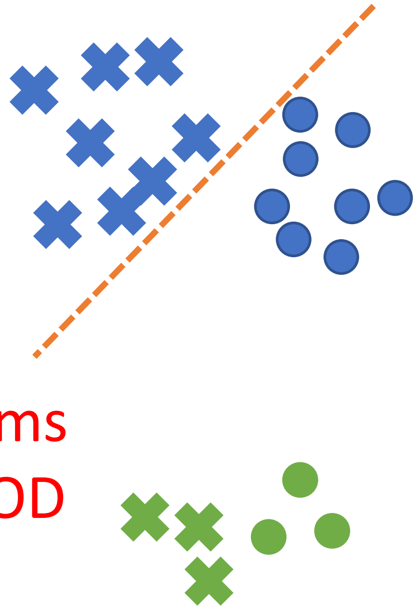

Features for OOD
examples change less

# Feature Distortion Theory

Pretrained Features

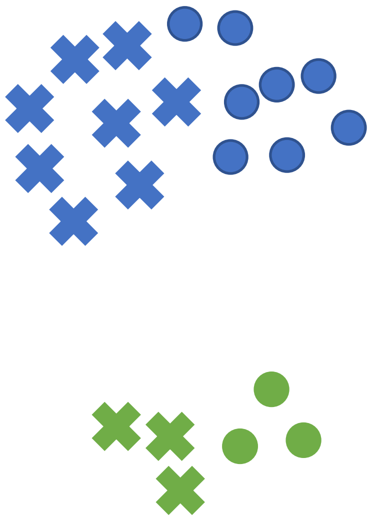


Fine-tuning



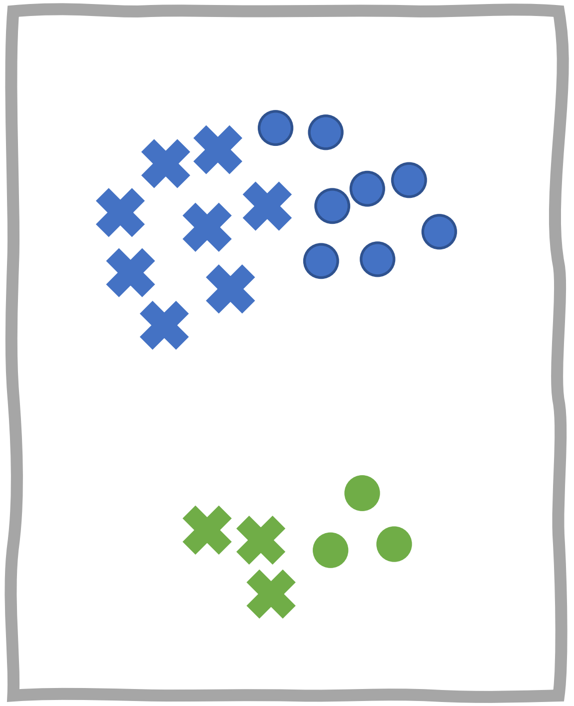
Head performs poorly on OOD examples

Linear probing: freezes pretrained features

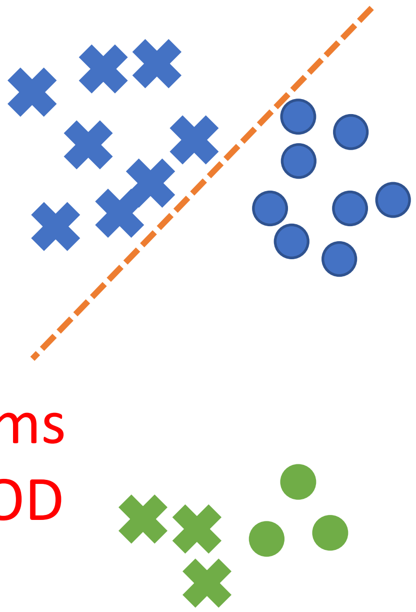


# Feature Distortion Theory

Pretrained Features

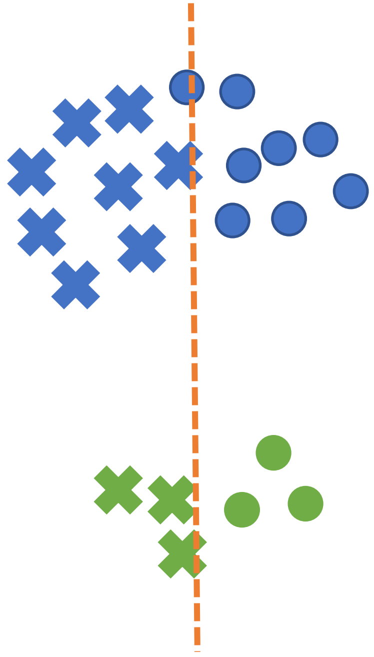


Fine-tuning



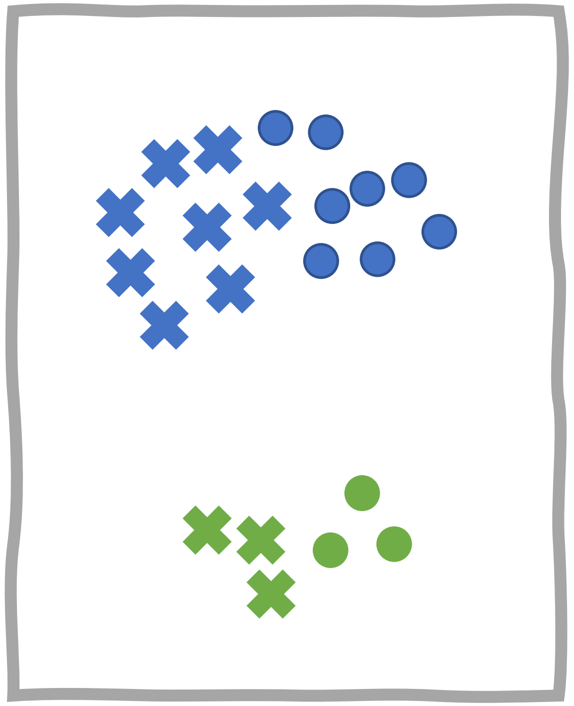
Head performs poorly on OOD examples

Linear probing: freezes pretrained features

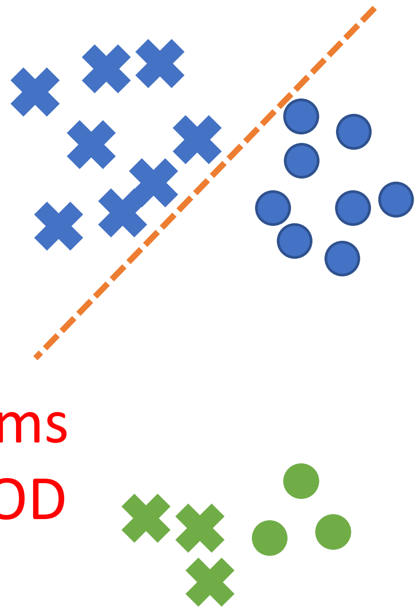


# Feature Distortion Theory

Pretrained
Features

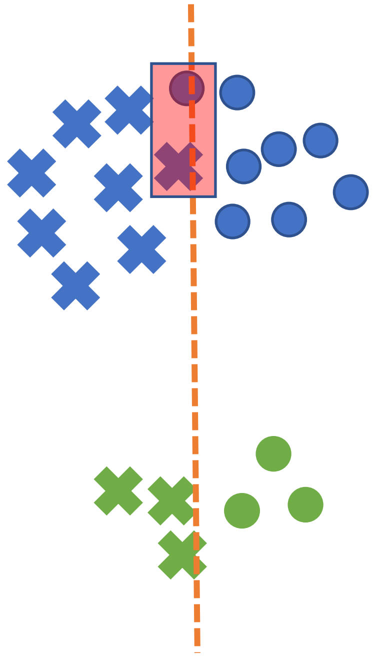


Fine-tuning



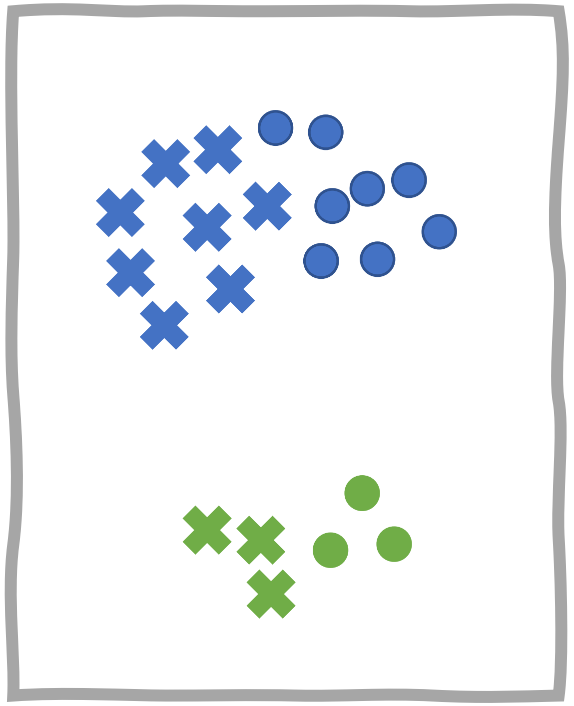
Head performs poorly on OOD examples

Linear probing: freezes pretrained features

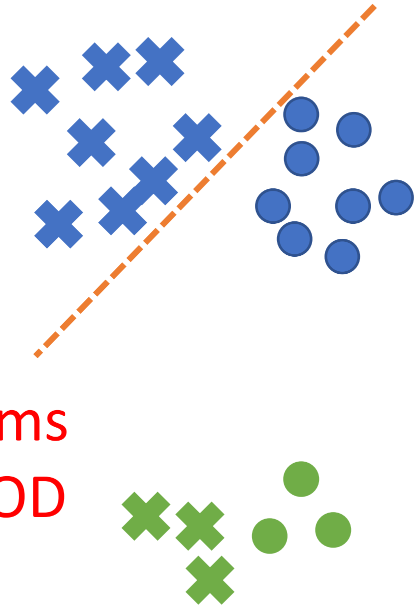


# Feature Distortion Theory

Pretrained Features

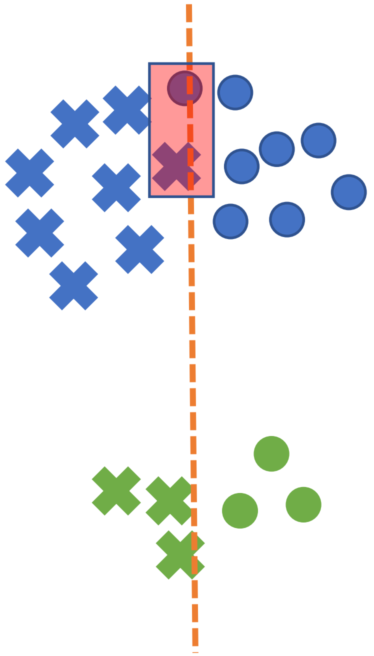


Fine-tuning



Head performs poorly on OOD examples

Linear probing: freezes pretrained features



Head is decent on OOD examples

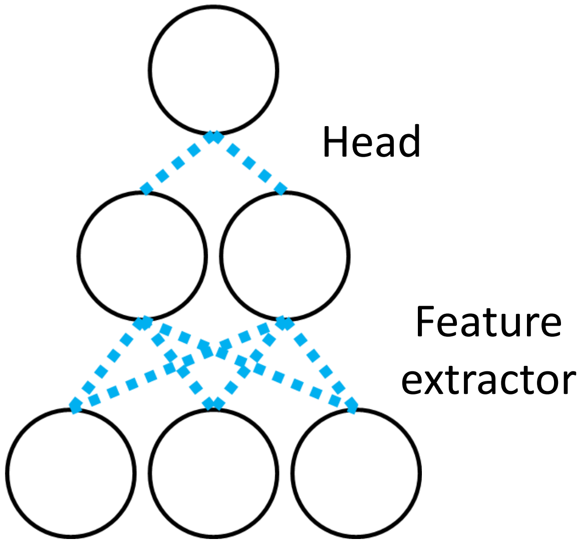
# Feature Distortion Theory (Overview)

- Prior work: linear probing. Fine-tuning: challenging to analyze
- Two-layer linear networks: we prove that fine-tuning distorts features

<u>Theorem 3.3 (Informal)</u>

$$L_{\text{ood}}(\text{fine−tuning}) \geq f(\epsilon, \sigma)$$

- $\sigma$: “amount” of distribution shift
- $\epsilon$: “quality” of pretrained features



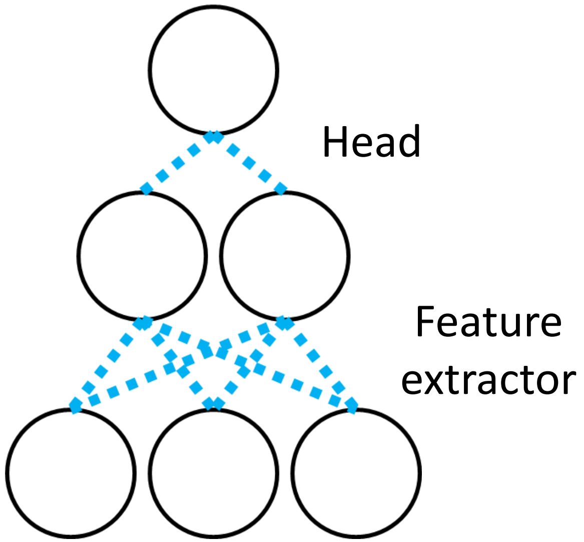
# Feature Distortion Theory (Overview)

Theorem 3.5 (Informal)

$$\frac{L_{\text{ood}}(\text{linear}-\text{probing})}{L_{\text{ood}}(\text{fine}-\text{tuning})} \xrightarrow{p} 0, \qquad \text{as } B_0 \rightarrow B_* \text{ (up to rotational symmetries)}$$

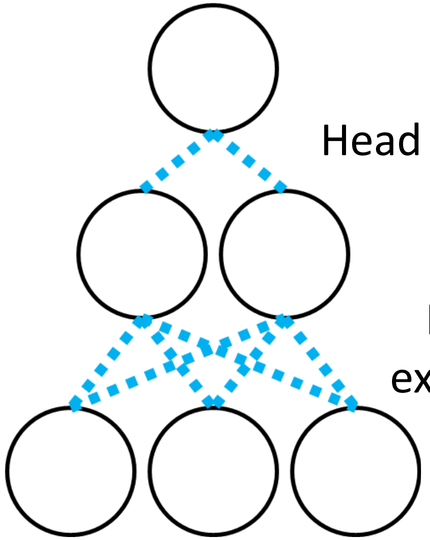
- $B_0$: pretrained feature extractor, $B_*$: optimal feature extractor

- OOD: fine-tuning worse than linear probing
  - If pretrained features good, OOD shift large
  - Throughout the process of fine-tuning
- ID: fine-tuning better than linear probing



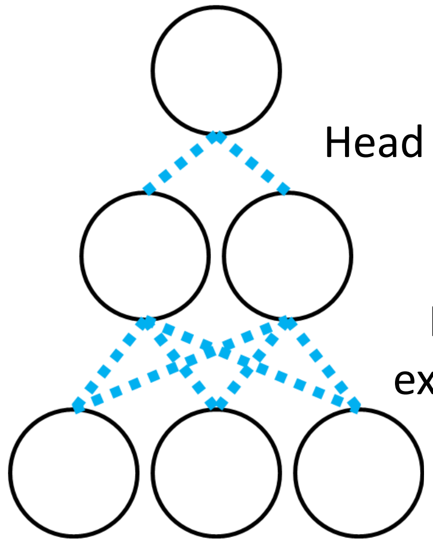
# Feature Distortion Theory (Details)

- Two-layer linear networks
  - High dimensional input: $x$
  - Lower dimensional features: $B_* x$
  - Ground truth outputs: $y = v_*^{\mathrm{T}} B_* x$ (both ID and OOD)
- From prior work on pretraining, suppose we have $B_0$ close to $B_*$ (e.g., in operator norm)
- $P_{\mathrm{id}}$ and $P_{\mathrm{ood}}$ define different distributions on $x$



# Feature Distortion Theory (Details)

- $y = v_*^{\mathrm{T}} B_* x$ (both ID and OOD)
- Have $B_0$ close to $B_*$ (from pretraining)

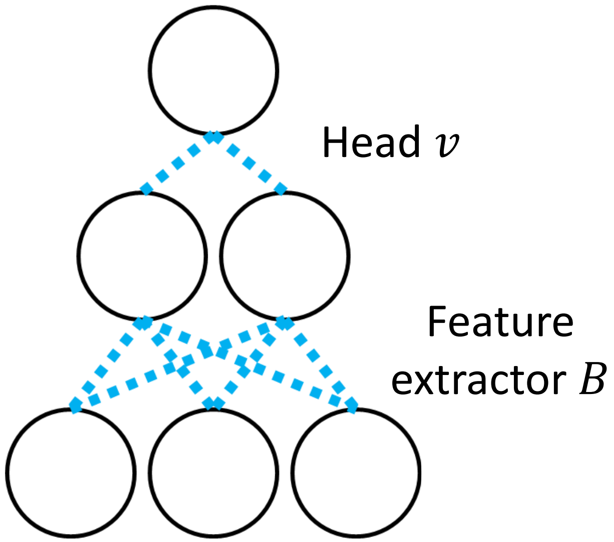



- Distributions:
  - $P_{\mathrm{id}}$ supported on subspace $S$
  - $P_{\mathrm{ood}}$ includes directions outside of $S$ (unseen directions)
- Overparameterized setting
  - Both fine-tuning and training from scratch fit train loss, but have different test losses
- FT updates $v$ and $B$, LP only updates head $v$

# Feature Distortion Theory (Details)

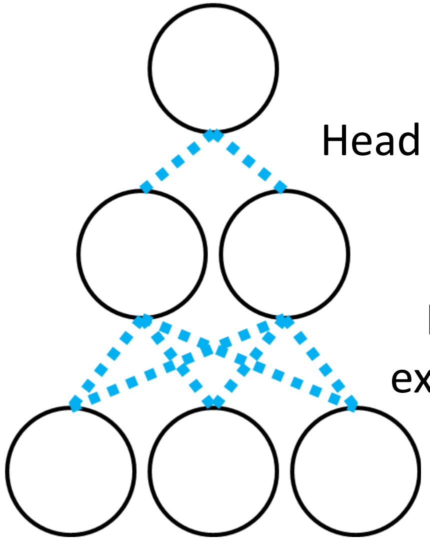
- $y = v_*^{\mathrm{T}} B_* x$ (both ID and OOD)
- Have $B_0$ close to $B_*$ (from pretraining)

- Squared loss: $L(v, B) = E_{x,y \sim P_{\mathrm{id}}}\left[(y - v^{\mathrm{T}} B x)^2\right]$
- Algorithms
  - LP initializes random head $v_0$, optimizes $\min_{v} L(v, B)$
  - FT initializes random head $v_0$, optimizes $\min_{v,B} L(v, B)$
  - LP-FT optimizes $\min_{v,B} L(v, B)$ but initializing $v$ from LP

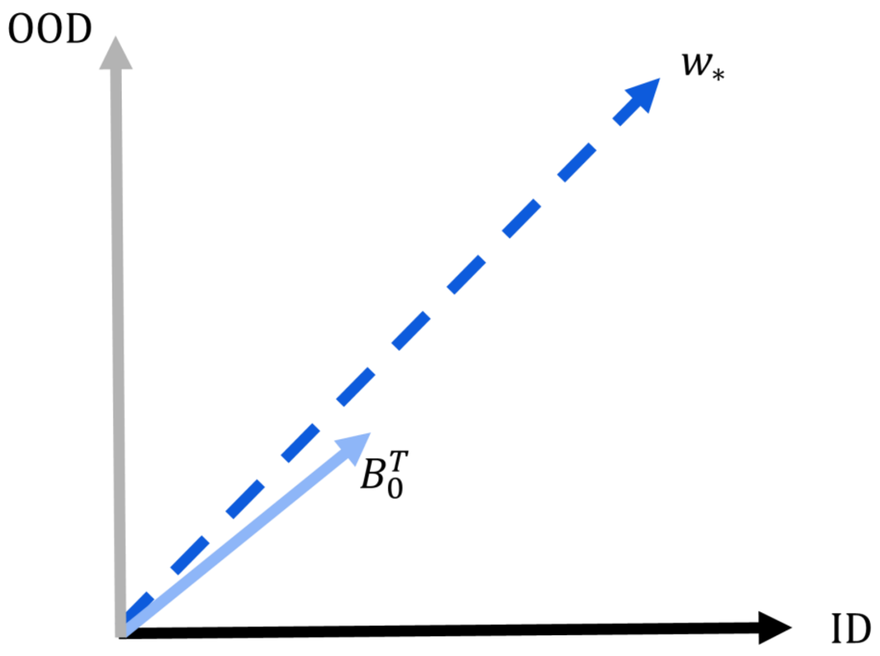


# Feature Distortion Theory (Details)

- Challenges
  - Prior work studies linear probing (fitting linear head on features)
  - Fine-tuning is non-convex, trajectory is complicated and has no closed form
  - Tool: leverage invariants that hold throughout process of fine-tuning

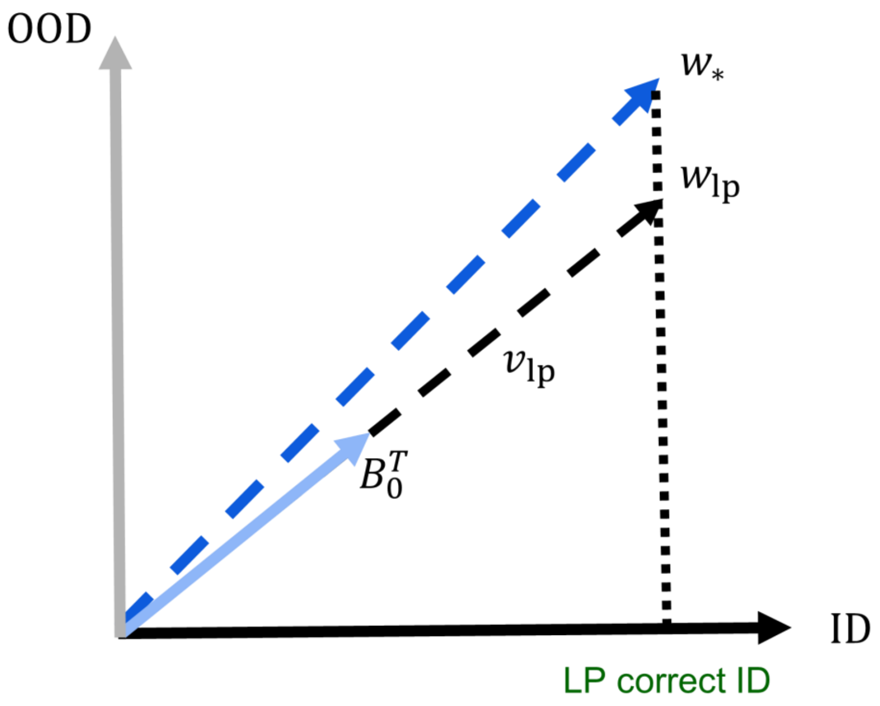


# Feature Distortion (Toy Example)



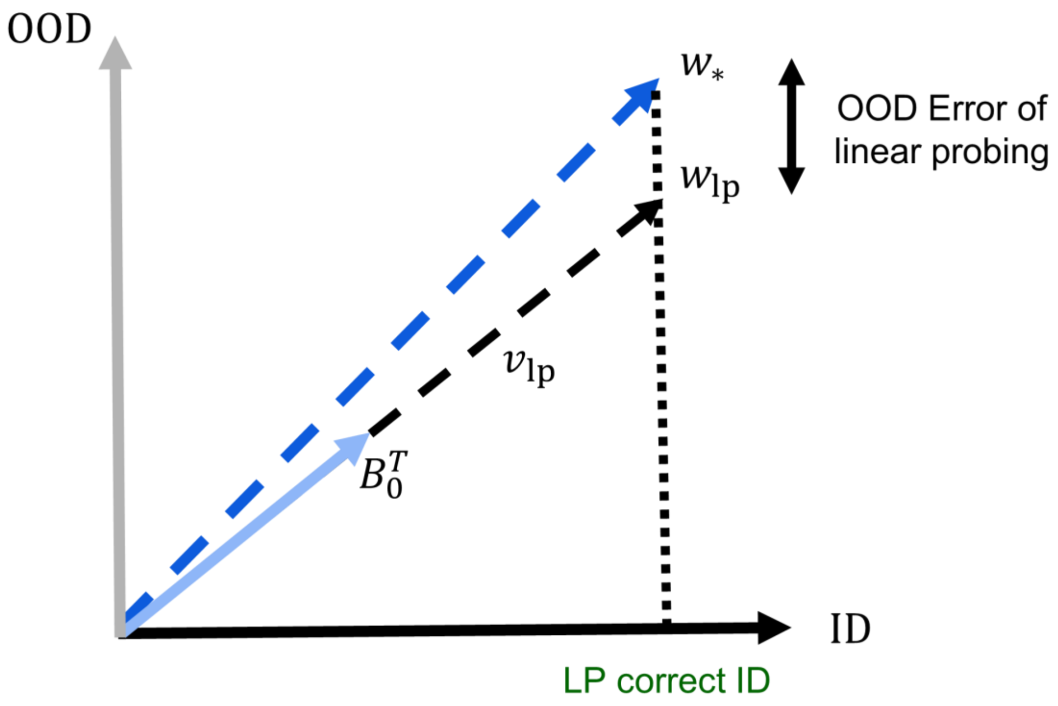

(a) Toy example (Linear probing)

# Feature Distortion (Toy Example)



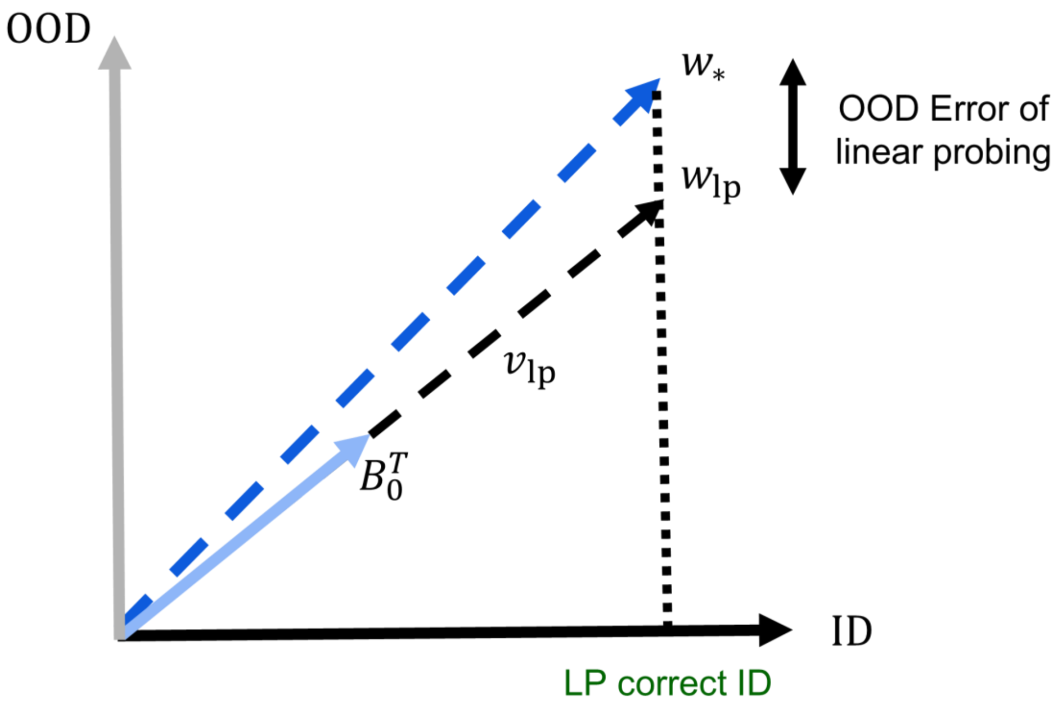

(a) Toy example (Linear probing)

# Feature Distortion (Toy Example)

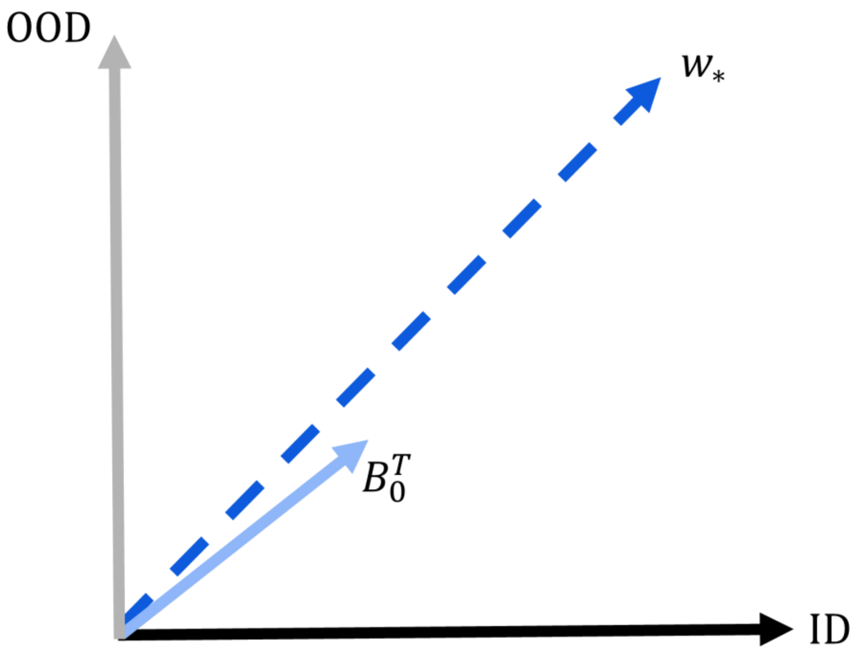



(a) Toy example (Linear probing)

# Feature Distortion (Toy Example)

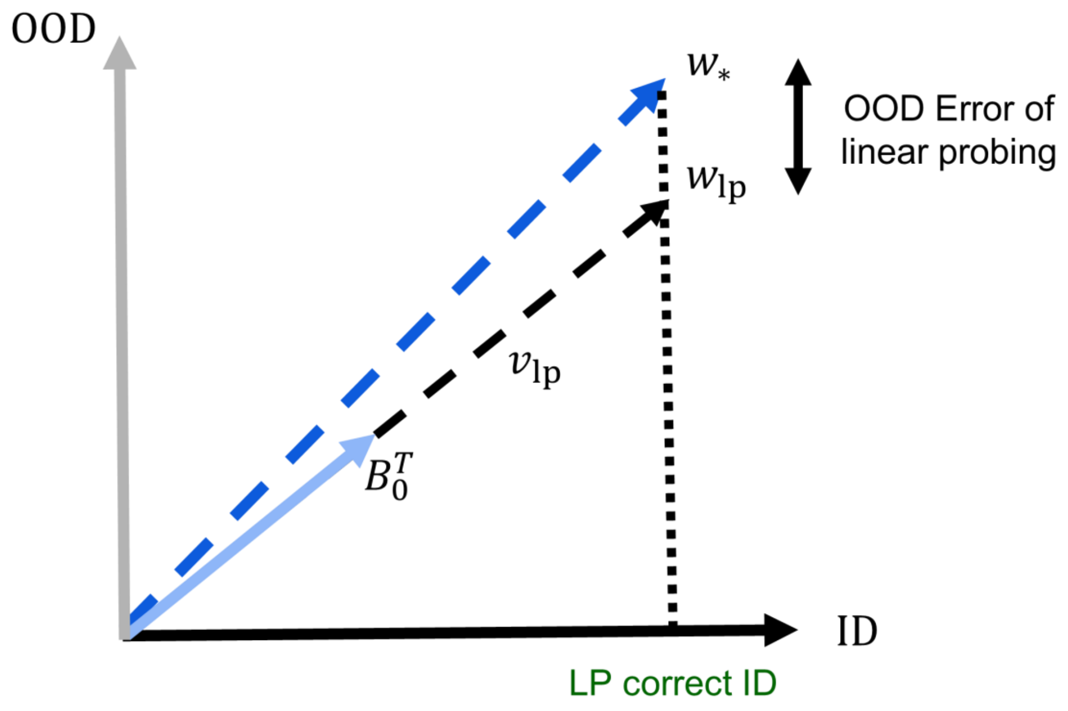



(a) Toy example (Linear probing)

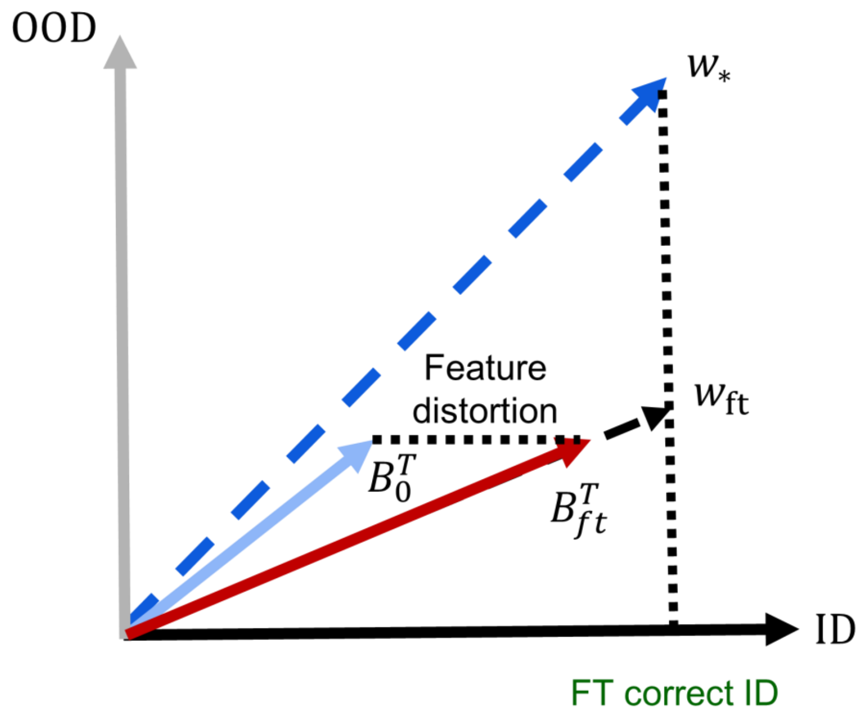



(b) Toy example (fine-tuning)

# Feature Distortion (Toy Example)

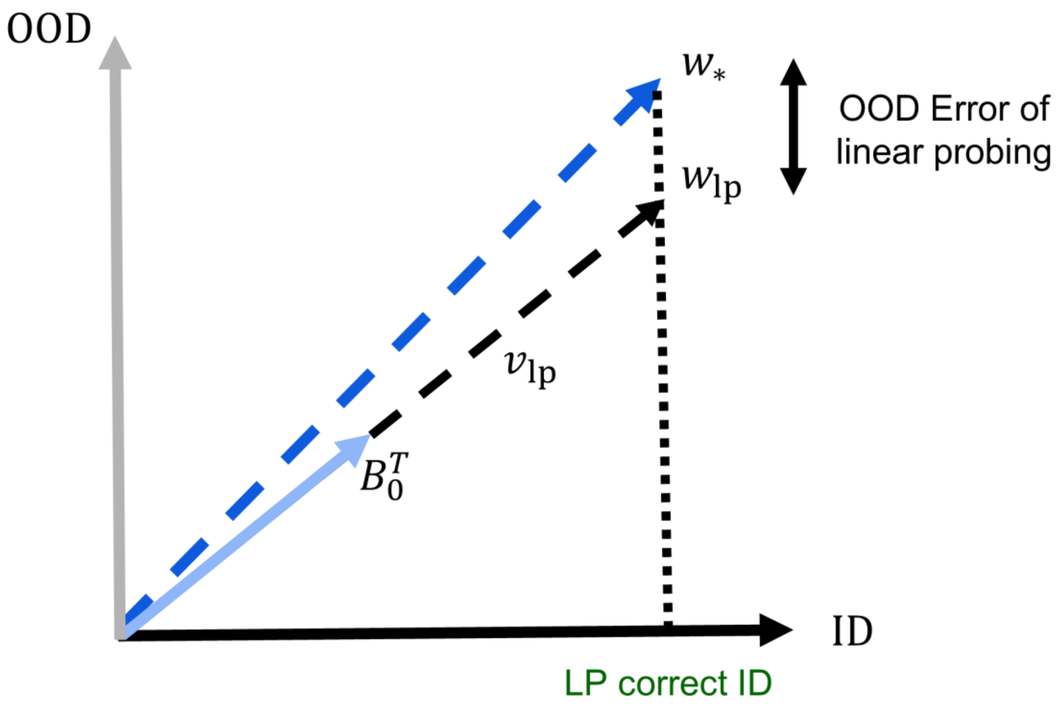



(a) Toy example (Linear probing)

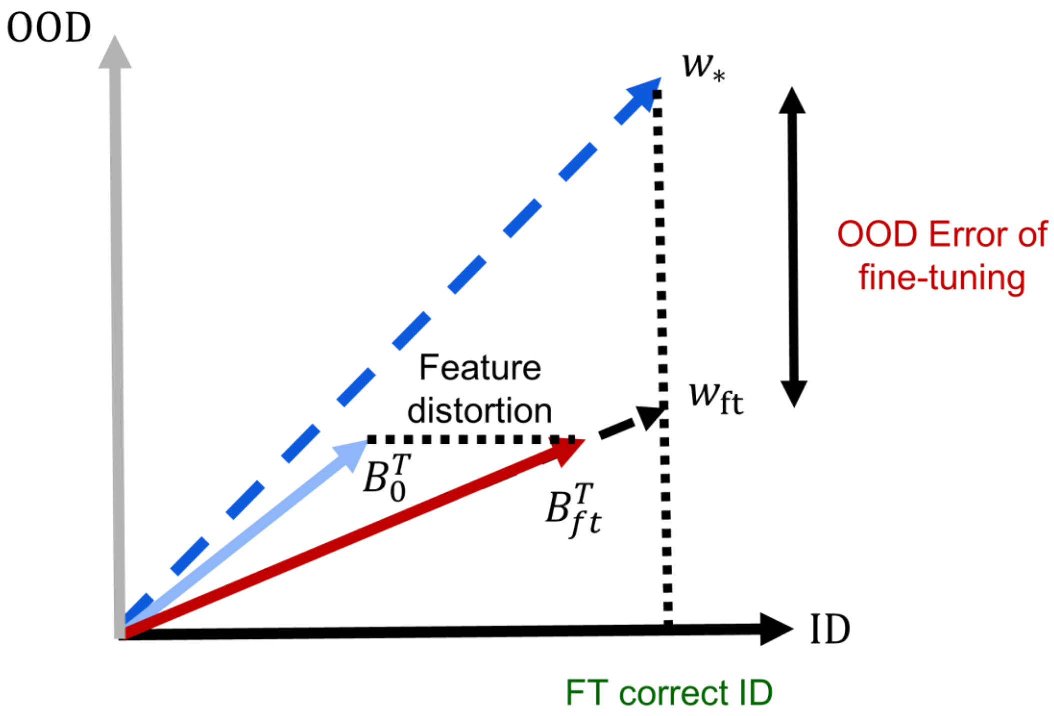



(b) Toy example (fine-tuning)

# Feature Distortion (Toy Example)




(a) Toy example (Linear probing)




(b) Toy example (fine-tuning)

# How to learn pretrained features

- Learn good features for *both* ID and OOD
- Auxiliary information
  - In-N-Out: Pre-Training and Self-Training using Auxiliary Information for Out-of-Distribution Robustness. SMX*, **AK***, RJ*, FK, TM, PL. ICLR 2020.
- Contrastive learning
  - Connect, Not Collapse: Explaining Contrastive Learning for Unsupervised Domain Adaptation. KS*, RJ*, **AK***, SMX*, JZH, TM, PL. Preprint.

# Outline

1. Fine-tuning can do worse than linear-probing OOD
2. Why fine-tuning can underperform OOD
3. **Simple change to fine-tuning: improved accuracy on 10 datasets**

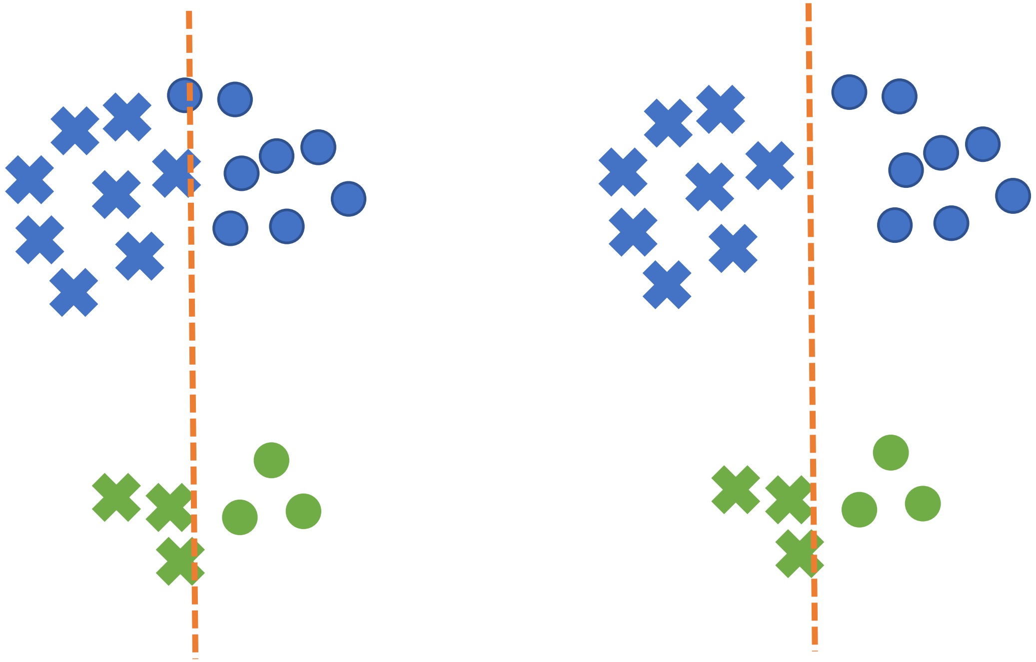
# Improving fine-tuning

- Fine-tuning works better on ID test; linear probing works better on OOD test
- Reason: start with random head, changes a lot → features get distorted

**Can we refine features without distorting them too much?**

**LP-FT**

Step 1: Linear probe
Step 2: Fine-tune



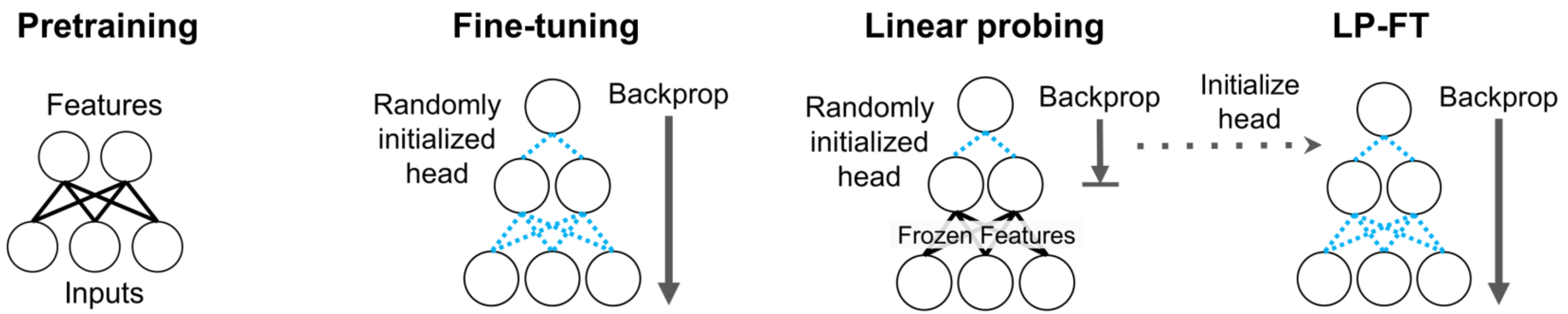
Prove this intuition in a simple setting

(Levine et al 2016, Kanavati & Tsuneki, 2021)

# Improving fine-tuning: experiments

- Datasets: standard datasets like CIFAR, ImageNet, DomainNet, BREEDS, satellite remote sensing
- Models: conv nets (ResNet-50) and Vision Transformers (ViT-B/16)
- Protocols:
  - Rigorous protocol for tuning hyperparameters on ID validation data
  - Ensure that LP-FT and fine-tuning use the same computation

# Improving fine-tuning




| | Fine-tuning | Linear probing | LP-FT |
|---|---|---|---|
| ID test  | 85.1% | 82.9% | **85.7%** |
| OOD test  | 59.3% | 66.2% | **68.8%** |

Average accuracies (10 datasets)

**+10% over fine-tuning!**

# In-Distribution Accuracies

| | CIFAR-10 | Ent-30 | Liv-17 | DomainNet | FMoW | ImageNet | Average |
|---|---|---|---|---|---|---|---|
| FT | **97.3 (0.2)** | **93.6 (0.2)** | 97.1 (0.2) | 84.5 (0.6) | **56.5 (0.3)** | **81.7 (-)** | 85.1 |
| LP | 91.8 (0.0) | 90.6 (0.2) | 96.5 (0.2) | 89.4 (0.1) | 49.1 (0.0) | 79.7 (-) | 82.9 |
| LP-FT | **97.5 (0.1)** | **93.7 (0.1)** | **97.8 (0.2)** | **91.6 (0.0)** | 51.8 (0.2) | **81.7 (-)** | **85.7** |

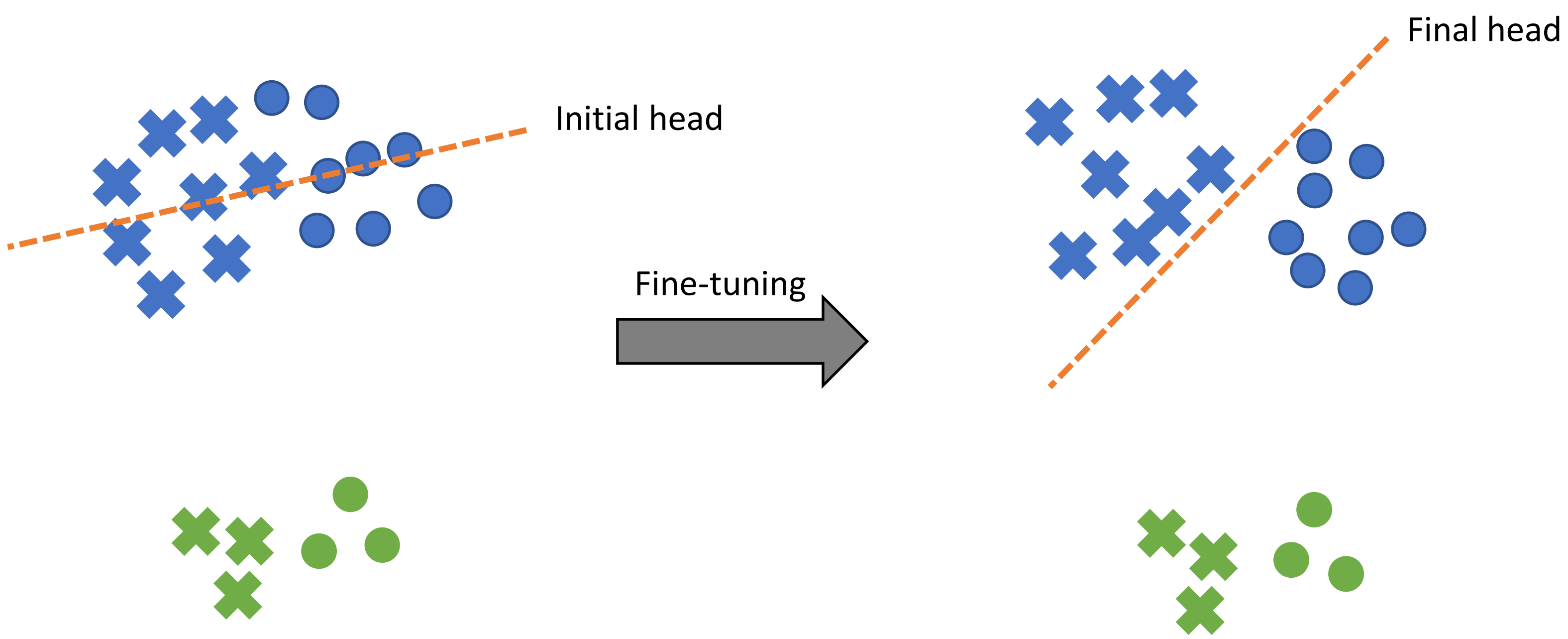
# Out-of-Distribution Accuracies

| | STL | CIFAR-10.1 | Ent-30 | Liv-17 | DomainNet | FMoW |
|---|---|---|---|---|---|---|
| FT | 82.4 (0.4) | 92.3 (0.4) | 60.7 (0.2) | 77.8 (0.7) | 55.5 (2.2) | 32.0 (3.5) |
| LP | 85.1 (0.2) | 82.7 (0.2) | **63.2 (1.3)** | 82.2 (0.2) | 79.7 (0.6) | **36.6 (0.0)** |
| LP-FT | **90.7 (0.3)** | **93.5 (0.1)** | **62.3 (0.9)** | **82.6 (0.3)** | **80.7 (0.9)** | **36.8 (1.3)** |

| | ImNetV2 | ImNet-R | ImNet-Sk | ImNet-A | Average |
|---|---|---|---|---|---|
| FT | **71.5 (-)** | 52.4 (-) | 40.5 (-) | 27.8 (-) | 59.3 |
| LP | 69.7 (-) | 70.6 (-) | 46.4 (-) | 45.7 (-) | 66.2 |
| LP-FT | **71.6 (-)** | **72.9 (-)** | **48.4 (-)** | **49.1 (-)** | **68.9** |

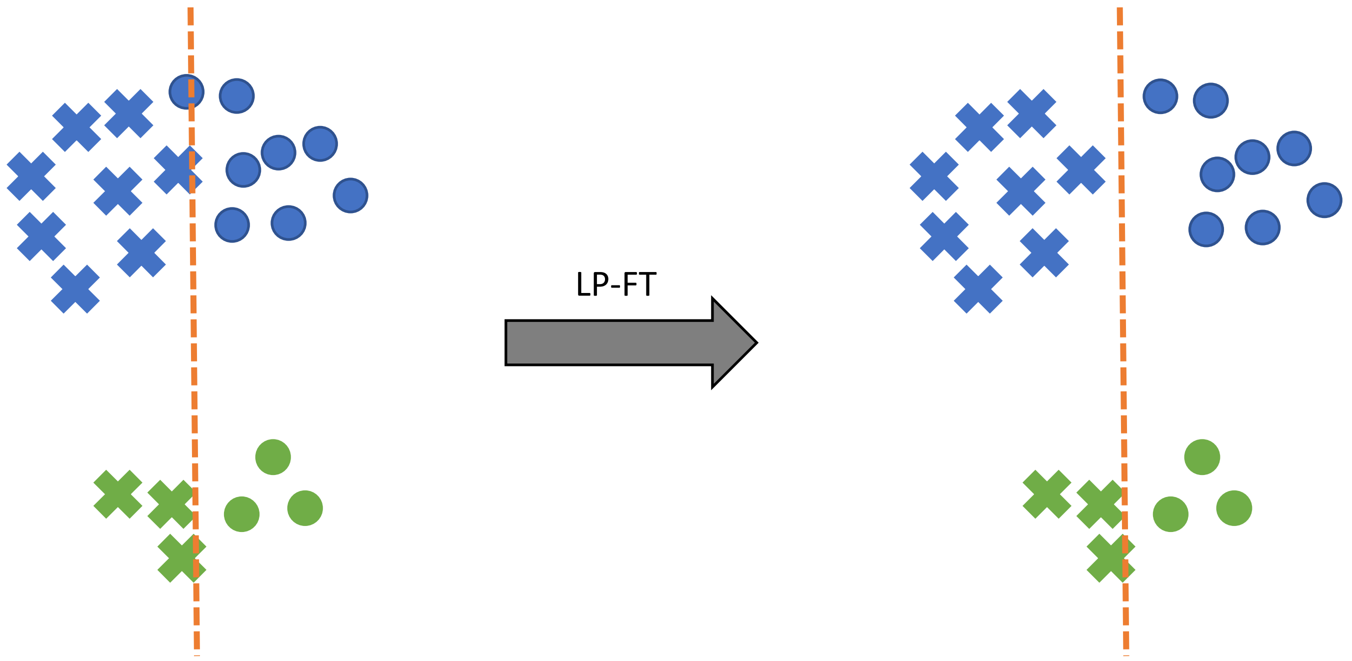
# Does feature distortion happen?

- ID features change more than OOD features



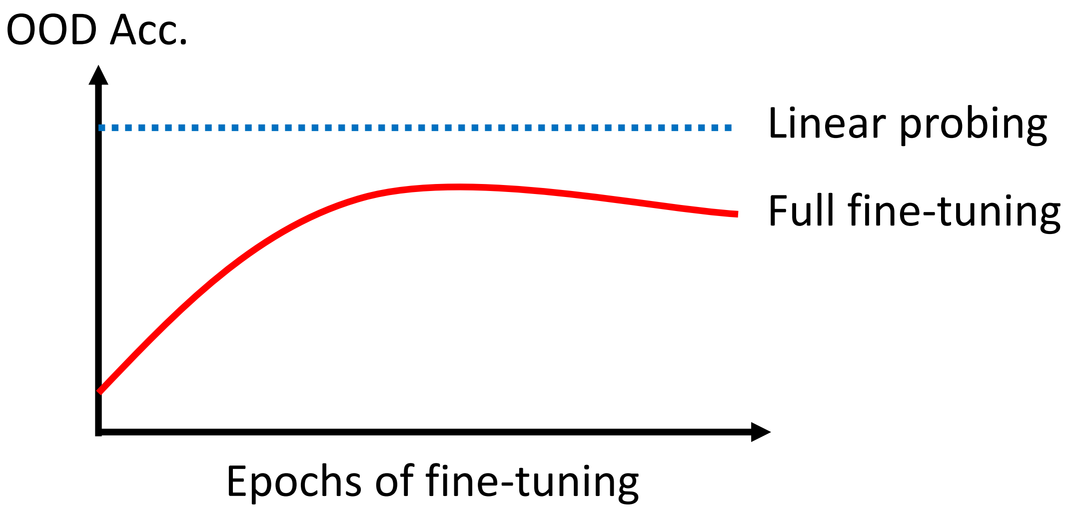
# Does feature distortion happen?

- Features change orders of magnitude less with LP-FT



# Does feature distortion happen?

- Early stopping does not solve the problem with fine-tuning



# Important conditions for LP vs. FT

- Theory says fine-tuning does worse than linear probing **if** features good, distribution shift large
- CIFAR-10.1, ImageNetV2: small shift, FT does better
- Use MoCo-V1 instead of MoCo-V2: worse features, FT does better

# Discussion

- Pretrained models give large improvements in accuracy, but how we fine-tune them is key
- LP-FT is just a starting point
- What to do when linear probing not so good?

# Related Work

- Lightweight fine-tuning
  - Can often improve OOD accuracy, we give one explanation
  - Increasingly important as pretrained feature quality improves
  - Adapter tuning, prefix tuning, composed fine-tuning
- Linear probing then fine-tuning
  - Sometimes used as a heuristic for ID, e.g. ULMFit
  - Just a starting point

# Summary

1. Fine-tuning can do worse than linear-probing OOD
2. Why fine-tuning can underperform OOD
3. Simple change to fine-tuning: improved accuracy on 10 datasets
    1. Linear probe to learn good head initialization
    2. Fine-tune to refine features

# Summary

1. Fine-tuning can do worse than linear-probing OOD
2. Why fine-tuning can underperform OOD
3. Simple change to fine-tuning: improved accuracy on 10 datasets
    1. Linear probe to learn good head initialization
    2. Fine-tune to refine features