Poster Section: Hall C 4-9 #701, 11:30

Latest paper update on the website!

International Conference On Machine Learning

LCA-on-the-Line: Benchmarking Out of Distribution Generalization with Class Taxonomies

Jia Shi, Gautam Gare, Jinjin Tian, Siqi Chai, Zhiqiu Lin, Arun Vasudevan, Di Feng, Francesco Ferroni, Shu Kong

Independent and Identically Distributed (IID) assumption of ML

Training Data

In-Distribution (ID) Testing Data

Machine learning assumes testing data is independent and identically distributed (IID) with the training data.

Models will encounter OOD testing data

Sketch

Illustration

Viewpoint

Training Data

Out-Of-Distribution (OOD) Testing Data

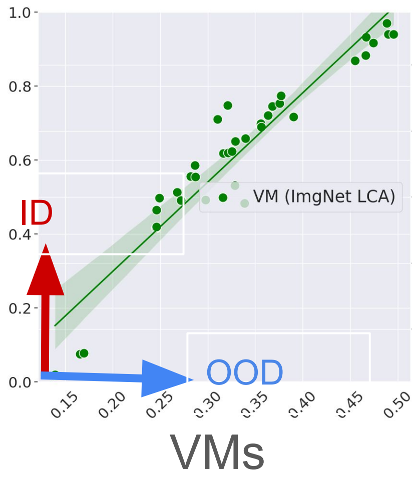
We hope models to generalize to OOD testing data, which has severe visual shift from the training data.

Given a pool of models, how can we predict which model generalizes to OOD testing data better?

Predict OOD performance with ID accuracy

Accuracy-on-the-line [1]: empirically, OOD performance is strongly correlated with ID performance across models and distribution shifts.

This metric predicts the performance of **Vision models (VMs)** only.



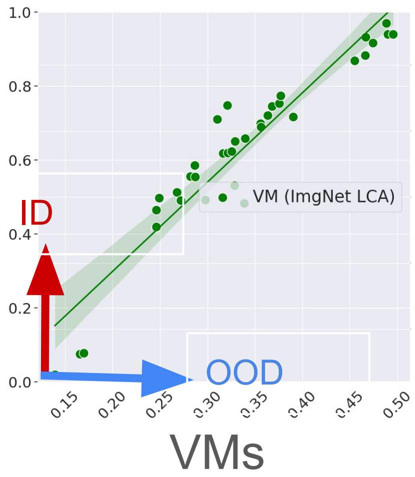
Y-axis: *ImageNet (ID) accuracy*
X-axis: *ObjectNet (OOD) accuracy*

[1] J. Miller, et al., "Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization", ICML, 2021.

Predict OOD performance with ID accuracy

Accuracy-on-the-line [1]: empirically, OOD performance is strongly correlated with ID performance across models and distribution shifts.

This metric predicts the performance of **Vision models (VMs)** only.



Accuracy is on the line!

Y-axis: *ImageNet (ID) accuracy*
X-axis: *ObjectNet (OOD) accuracy*

[1] J. Miller, et al., "Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization", ICML, 2021.

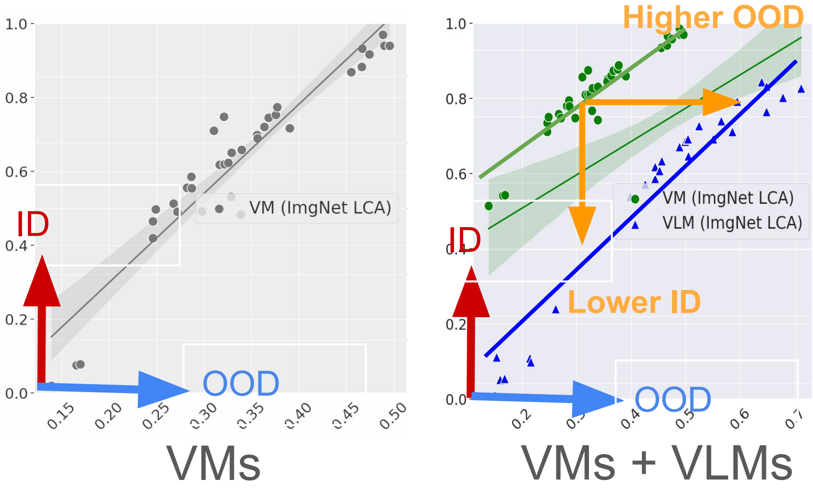
Accuracy is not on the line with VMs + VLMs

Accuracy-on-the-line [1]: empirically, OOD performance is strongly correlated with ID performance across models and distribution shifts.

~~This metric predicts the performance of **Vision models (VMs)** only.~~

This metric ***cannot*** reliably predicts the OOD performance of **Vision models (VMs) + Vision Language models (VLMs)**.

Difference (VMs, VLMs) = modality, training data source/size, loss, etc



Y-axis: ImageNet (ID) accuracy
X-axis: ObjectNet (OOD) accuracy

[1] J. Miller, et al., "Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization", ICML, 2021.

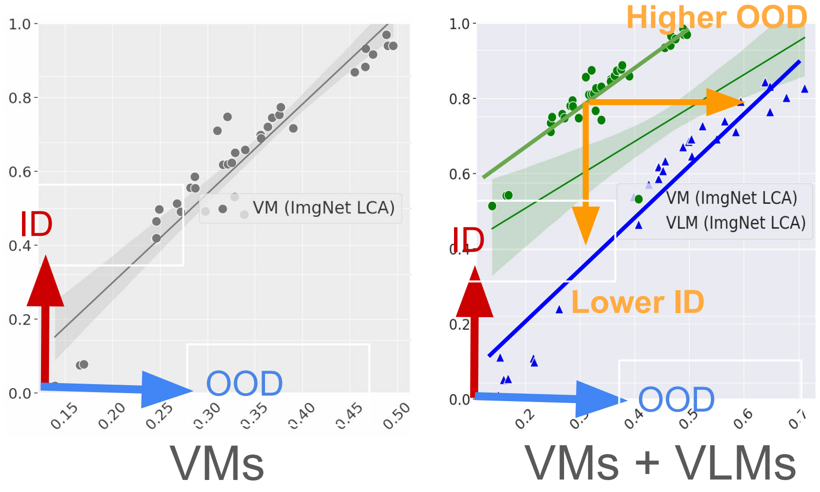
Accuracy is not on the line with VMs + VLMs

Accuracy-on-the-line [1]: empirically, OOD performance is strongly correlated with ID performance across models and distribution shifts.

~~This metric predicts the performance of **Vision models (VMs)** only.~~

This metric ***cannot*** reliably predicts the OOD performance of **Vision models (VMs) + Vision Language models (VLMs)**.

Accuracy is not on the line!



Y-axis: ImageNet (ID) accuracy
X-axis: ObjectNet (OOD) accuracy

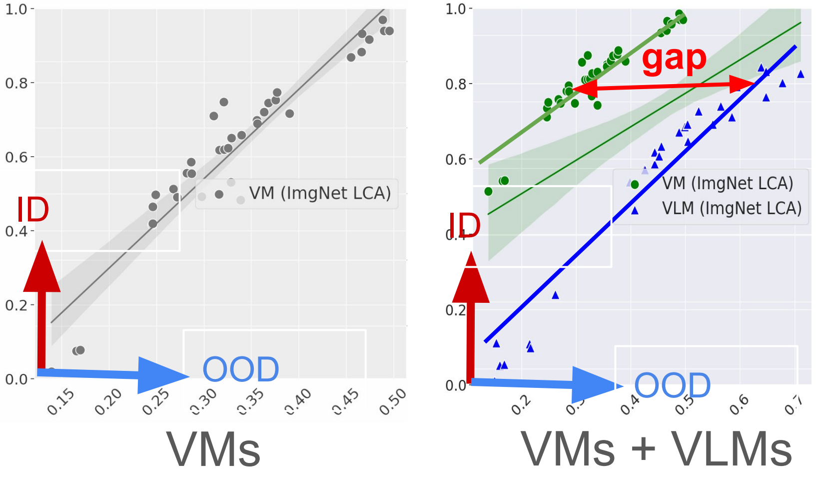
[1] J. Miller, et al., "Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization", ICML, 2021.

Accuracy is not on the line with VMs + VLMs

Accuracy-on-the-line [1]: empirically, OOD performance is strongly correlated with ID performance across models and distribution shifts.

~~This metric predicts the performance of **Vision models (VMs)** only.~~

This metric ***cannot*** reliably predicts the OOD performance of **Vision models (VMs) + Vision Language models (VLMs)**.



In-distribution (ID) accuracy might be ***biased*** by models settings, like modality and training data source.

[1] J. Miller, et al., "Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization", ICML, 2021.
[2] T. Taori, et al., "Measuring Robustness to Natural Distribution Shifts in Image Classification", NeurIPS, 2020.

LCA distance is a robust generalization indicator

1. What is LCA distance?

2. Why should we use LCA distance?

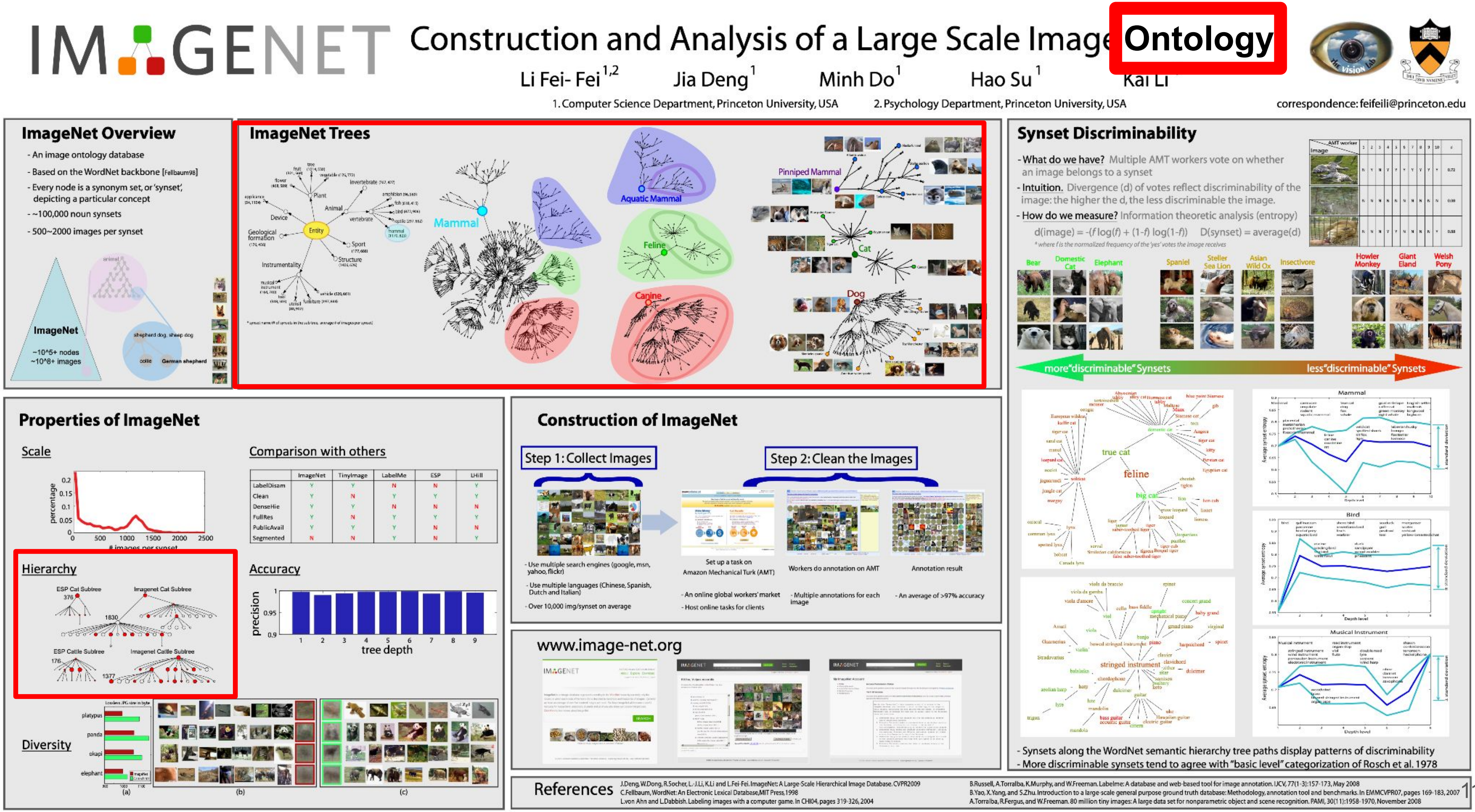
3. How can we use LCA distance to improve model generalization?

LCA distance is a robust generalization indicator

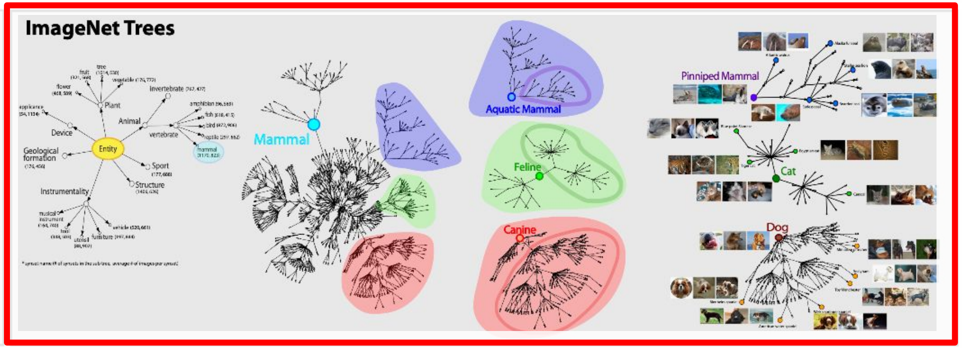
1. **What is LCA distance?**

2. Why should we use LCA distance?

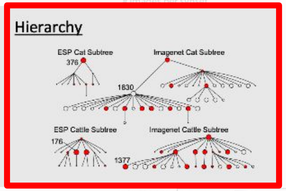
3. How can we use LCA distance to improve model generalization?


IMAGENET
Construction and Analysis of a Large Scale Image Ontology
Li Fei-Fei[1,2] Jia Deng[1] Minh Do[1] Hao Su[1] Kai Li[1]
1. Computer Science Department, Princeton University, USA
2. Psychology Department, Princeton University, USA
correspondence: feifeili@princeton.edu
ImageNet Overview
- An image ontology database
- Based on the WordNet backbone [Fellbaum98]
- Every node is a synonym set, or 'synset', depicting a particular concept
- ~100,000 noun synsets
- 500~2000 images per synset
ImageNet
~10^5+ nodes
~10^8+ images
shepherd dog, sheep dog
collie
German shepherd
ImageNet Trees
Entity
Mammal
Aquatic Mammal
Feline
Canine
Pinniped Mammal
Cat
Dog
Synset Discriminability
- What do we have? Multiple AMT workers vote on whether an image belongs to a synset
- Intuition. Divergence (d) of votes reflect discriminability of the image: the higher the d, the less discriminable the image.
- How do we measure? Information theoretic analysis (entropy)
d(image) = -(f log(f) + (1-f) log(1-f)) D(synset) = average(d)
* where f is the normalized frequency of the 'yes' votes the image receives
Bear
Domestic Cat
Elephant
Spaniel
Steller Sea Lion
Asian Wild Ox
Insectivore
Howler Monkey
Giant Eland
Welsh Pony
more "discriminable" Synsets
less "discriminable" Synsets
true cat
feline
big cat
stringed instrument
Mammal
Bird
Musical Instrument
- Synsets along the WordNet semantic hierarchy tree paths display patterns of discriminability
- More discriminable synsets tend to agree with "basic level" categorization of Rosch et al. 1978
Properties of ImageNet
Scale
percentage
images per synset
Comparison with others
ImageNet
TinyImage
LabelMe
ESP
LHill
LabelDisam
Clean
DenseHie
FullRes
PublicAvail
Segmented
Hierarchy
ESP Cat Subtree
Imagenet Cat Subtree
ESP Cattle Subtree
Imagenet Cattle Subtree
Accuracy
precision
tree depth
Diversity
platypus
panda
okapi
elephant
Construction of ImageNet
Step 1: Collect Images
Step 2: Clean the Images
- Use multiple search engines (google, msn, yahoo, flickr)
- Use multiple languages (Chinese, Spanish, Dutch and Italian)
- Over 10,000 img/synset on average
Set up a task on Amazon Mechanical Turk (AMT)
- An online global workers' market
- Host online tasks for clients
Workers do annotation on AMT
- Multiple annotations for each image
Annotation result
- An average of >97% accuracy
www.image-net.org
References
J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. CVPR2009
C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998
L. von Ahn and L. Dabbish. Labeling images with a computer game. In CHI04, pages 319-326, 2004
B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. IJCV, 77(1-3):157-173, May 2008
B. Yao, X. Yang, and S. Zhu. Introduction to a large-scale general purpose ground truth database: Methodology, annotation tool and benchmarks. In EMMCVPR07, pages 169-183, 2007
A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. PAMI, 30(11):1958-1970, November 2008

Ontology



Semantic concepts are defined w.r.t an ontology, such as WordNet hierarchy [1].



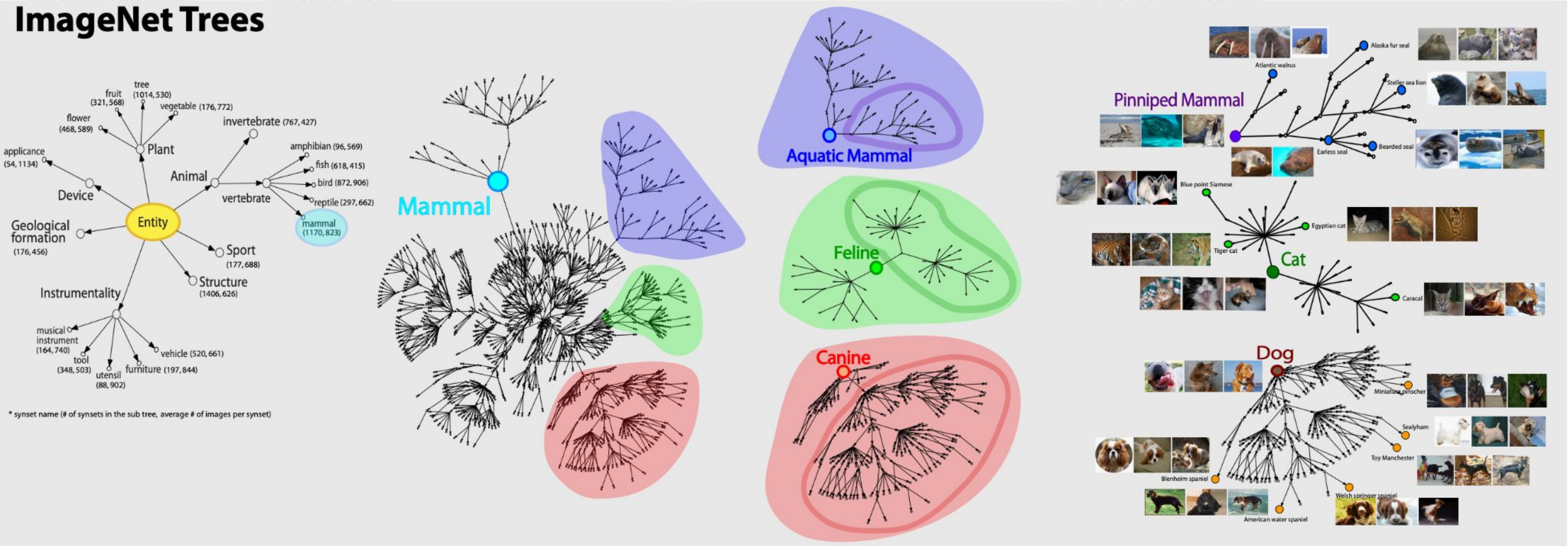
[1] C. Fellbaum. WordNet: An Electronic Lexical Database,1998

IMAGENET Construction and Analysis of a Large Scale Image **Ontology**

Li Fei-Fei[1,2] Jia Deng[1] Minh Do[1] Hao Su[1] Kai Li

1. Computer Science Department, Princeton University, USA 2. Psychology Department, Princeton University, USA

correspondence: feifeili@princeton.edu



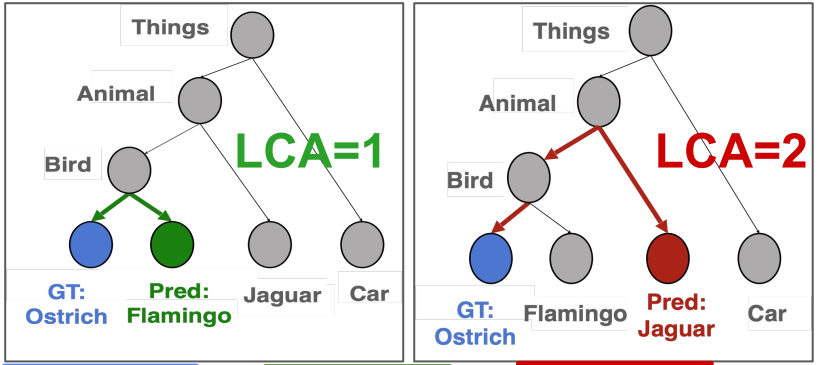
Diversity

- Synsets along the WordNet semantic hierarchy tree paths display patterns of discriminability
- More discriminable synsets tend to agree with "basic level" categorization of Rosch et al. 1978

References
J.Deng, W.Dong, R.Socher, L.-J.Li, K.Li and L.Fei-Fei. ImageNet: A Large Scale Hierarchical Image Database. CVPR2009
C.Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998
L.von Ahn and L.Dabbish. Labeling images with a computer game. In CHI04, pages 319-326, 2004
B.Russell, A.Torralba, K.Murphy, and W.Freeman. Labelme: A database and web-based tool for image annotation. IJCV, 77(1-3):157-173, May 2008
B.Yao, X.Yang, and S.Zhu. Introduction to a large-scale general purpose ground truth database: Methodology, annotation tool and benchmarks. In EMMCVPR07, pages 169-183, 2007
A.Torralba, R.Fergus, and W.Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. PAMI, 30(11):1958-1970, November 2008

LCA (lowest common ancestor) distance

Over an ***ontology***, such as a class hierarchy encoding class relationship, **LCA distance** measures class adjacency.



LCA distance rewards mistakes in prediction that are semantically closer to the ground-truth.



Smaller LCA distance indicate better mistake.

For GT=Ostrich, predicting Flamingo over Jaguar makes better mistakes [1].

[1] L. Bertinetto,, et al., “Leveraging Class Hierarchies with Deep Networks", CVPR, 2020

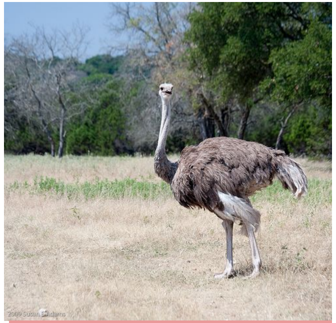
LCA distance is a robust generalization indicator

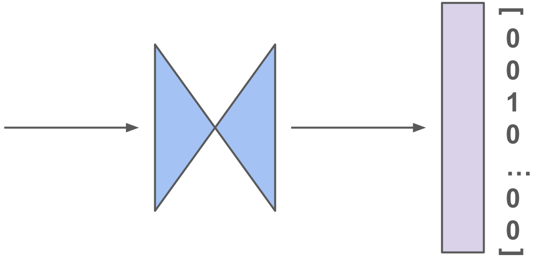
1. What is LCA distance?

2. **Why should we use LCA distance?**

3. How can we use LCA distance to improve model generalization?

What makes a model generalize better?





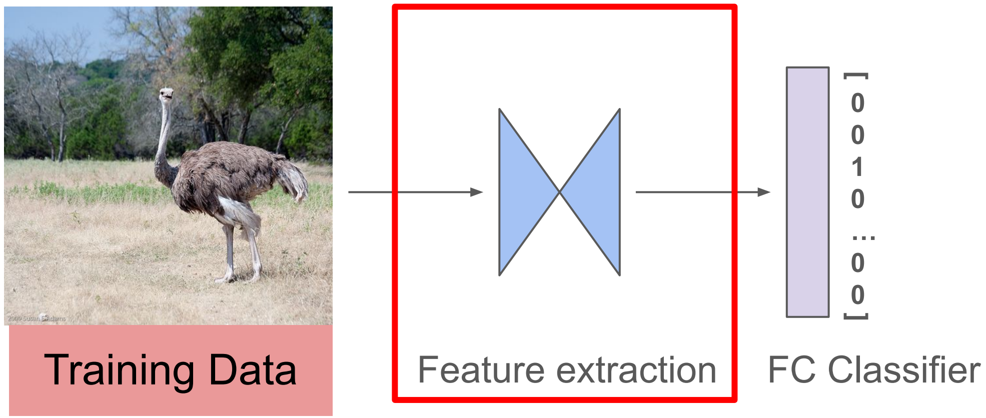
Training Data

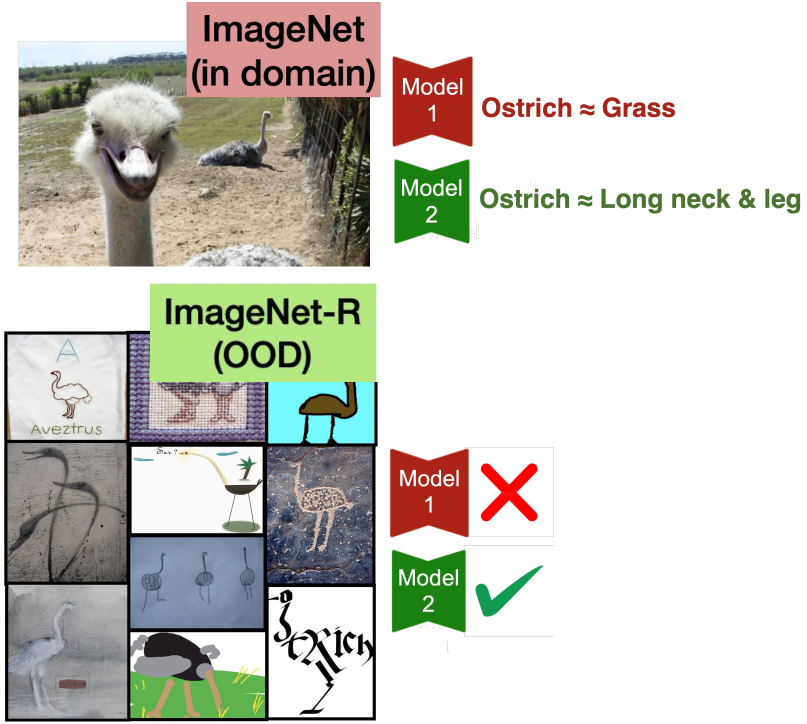
Feature extraction

FC Classifier

A model learns predictive features by likelihood maximization, resulting into an ability to associate input image to target labels.

What makes a model generalize better?

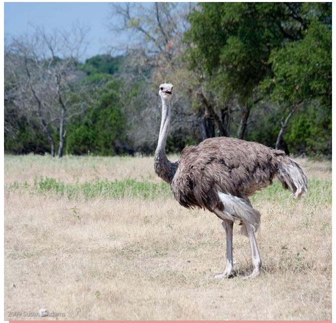




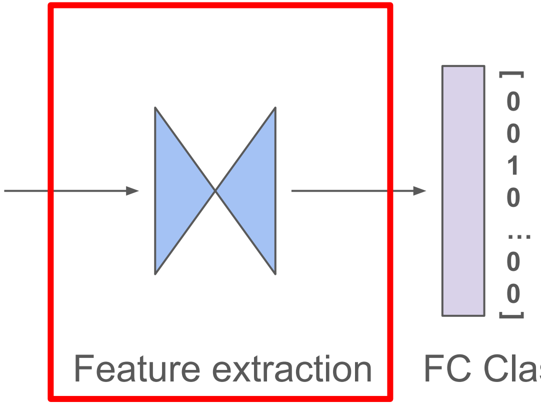
Models learning <u>spurious correlation</u> would fail to generalize to OOD data.

Model learning <u>transferable features</u> would generalize better.

What makes a model generalize better?



Training Data



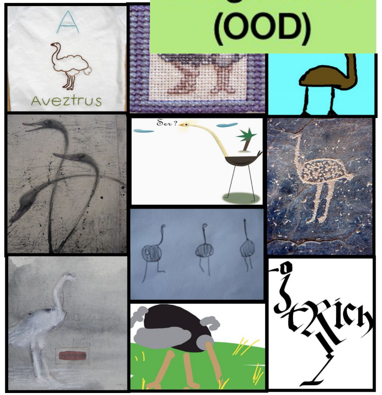
Feature extraction

FC Classifier

ImageNet (in domain)

Model 1 — Ostrich ≈ Grass

Model 2 — Ostrich ≈ Long neck & leg



ImageNet-R (OOD)

Model 1 ✗

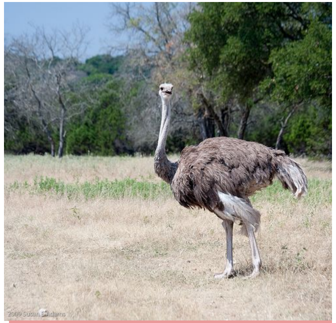
Model 2 ✓

Models learning <u>spurious correlation</u> would fail to generalize to OOD data.

Model learning <u>transferable features</u> would generalize better.

As benchmarks often simulate human-world ontology, the desired transferable features should align with human-defined ontology.

Flashback: Models will encounter OOD testing data



Training Data

Sketch

Illustration

Viewpoint

Out-of-distribution Testing data

We hope models to generalize to OOD testing data, which has severe visual shift from the training data.

Given a random pool of models, how can we predict which model generalizes to OOD testing data better?

Mistake prediction is cue for predictive features

Hypothesis: Transferable features are shared among semantically closer classes.

Ostrich ≈ Grass

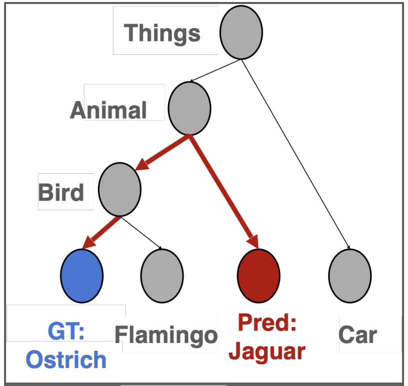
Ostrich ≈ Long neck & leg

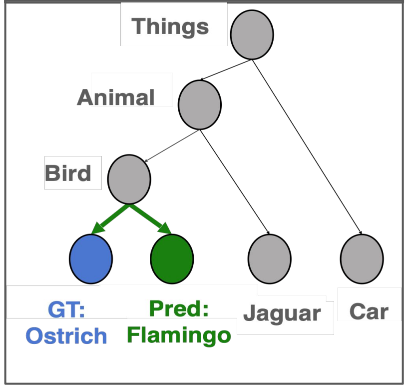
I see **grass**,
maybe it's a **Jaguar**?

High LCA

I see **long neck**,
maybe it's a **flamingo**?

Low LCA





If a model learns such a bird, it should assign high likelihood to other bird classes too.

Lower LCA
→ Models can predict semantically closer classes.
→ Models establish less spurious correlation.
→ Models can learn more transferable features.
→ Models generalize better.

LCA-on-the-Line is a robust indicator of generalization

LCA distance is a general metric, only depending on the relative ranking among class predictions. It is

- agnostic to model modality
- agnostic to training- and testing-sets attributes
- agnostic to the amount of training data
- easy to calculate and requires only one-time inference.

Experiments

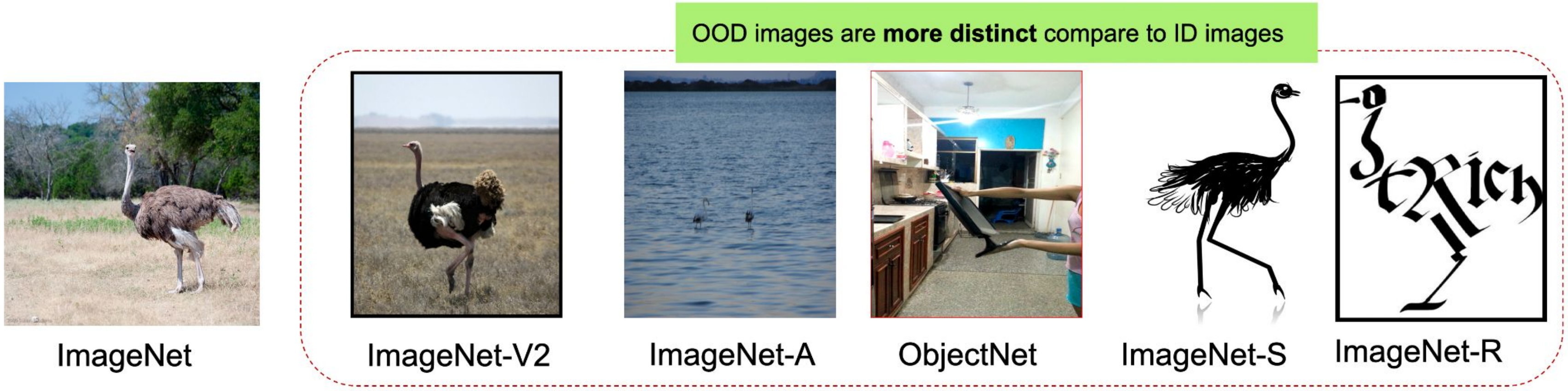
Experiment Settings

ID dataset / Source datasets: ImageNet

OOD datasets / Target datasets:

ImageNet v2 / Sketch / Rendition / Adversarial / ObjectNet

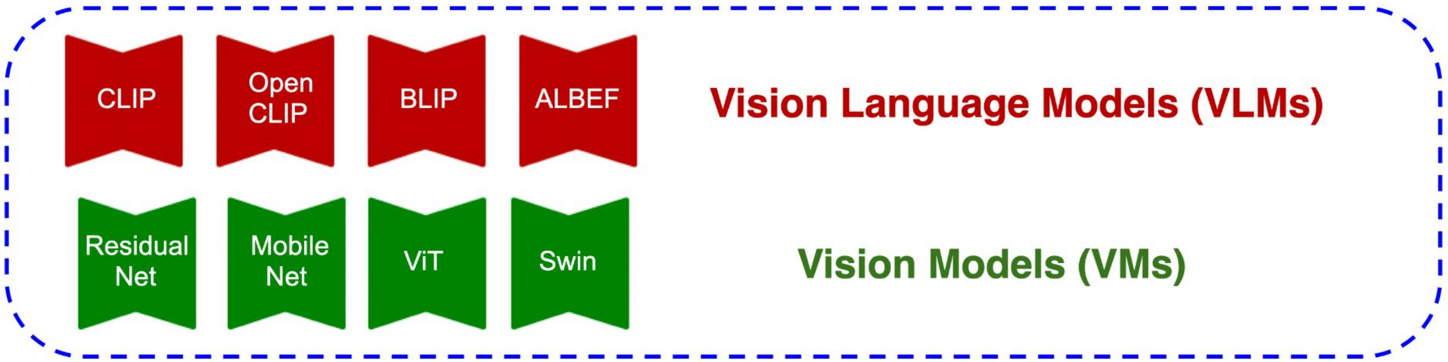
LCA-on-the-Line evaluates on severe visual shift datasets



Experiment Settings

75 models:

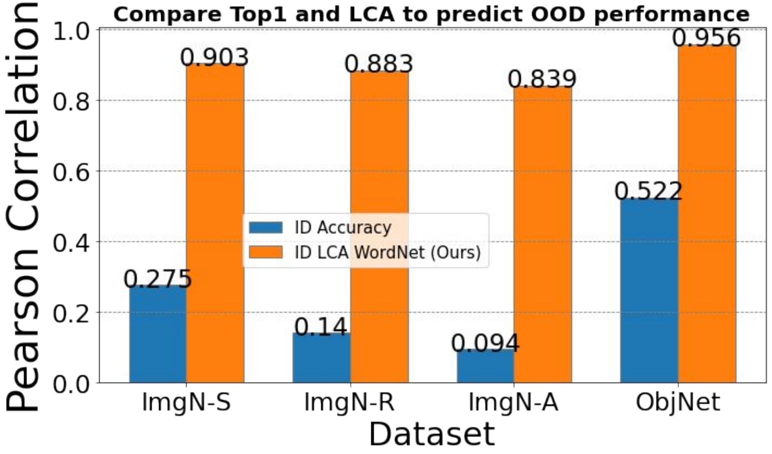
- 36 pre-trained *Vision Models (VMs)* on ImageNet
  - [AlexNet, …., SwinTransformer]
- 39 pre-trained *Vision-Language Models (VLMs)* using internet data
  - [ALBEF, BLIP, CLIP*7, OpenCLIP*30]

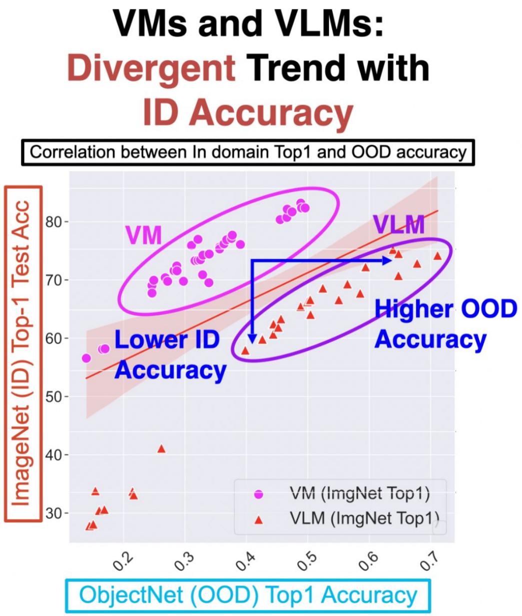


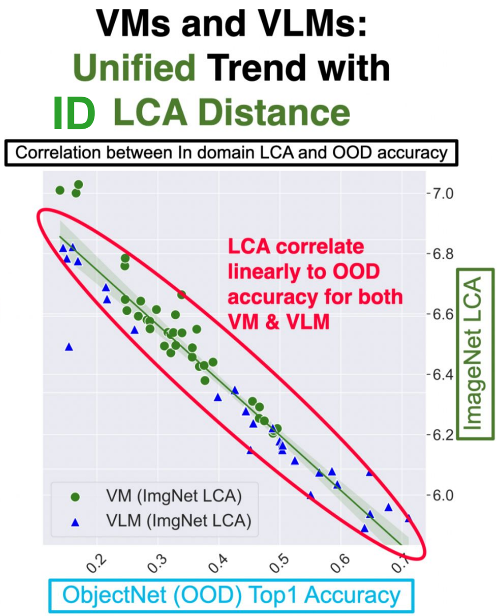
Experiment 1: Predict OOD from ID metric

Correlation comparison against OOD accuracy.

- **Baseline**: Accuracy-on-the-line [1] (ID accuracy)
- **Ours**: LCA-on-the-line (ID LCA distance)





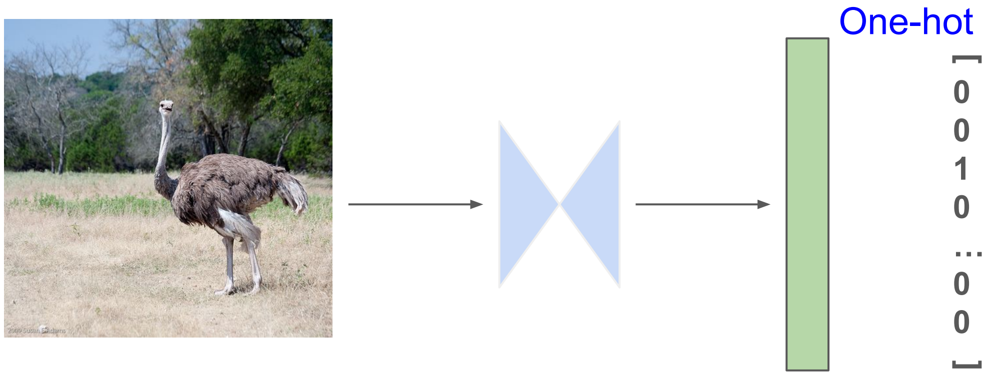


LCA distance restores the 'on-the-line' relationship across VMs & VLMs, *displaying a strong correlation.*

[1] J. Miller, et al., "Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization", ICML, 2021

LCA distance is a robust generalization indicator

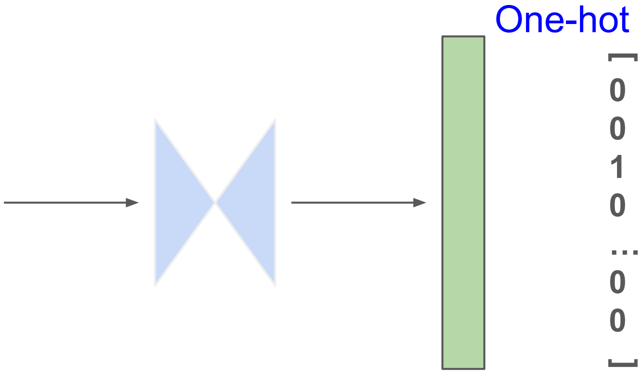
1. What is LCA distance?
2. Why should we use LCA distance?
3. **How can we use LCA distance to improve model generalization?**



Training Data

Feature extraction

FC Classifier

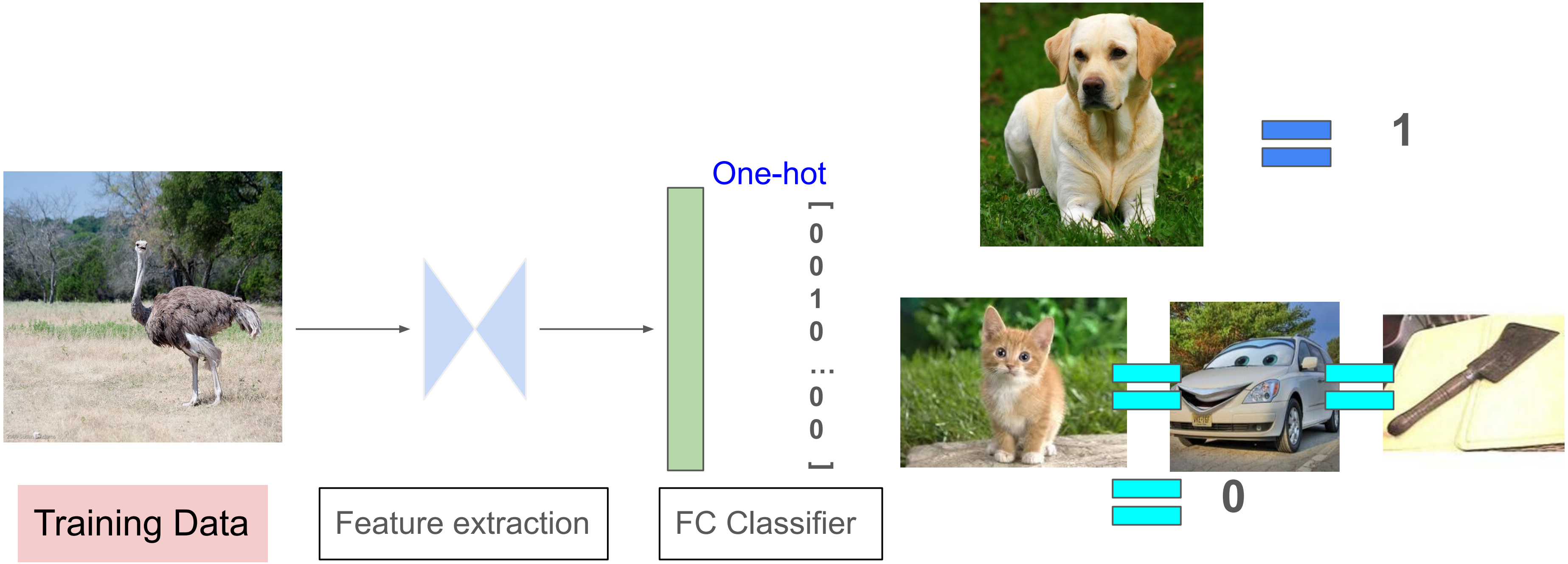


Training Data

Feature extraction

FC Classifier

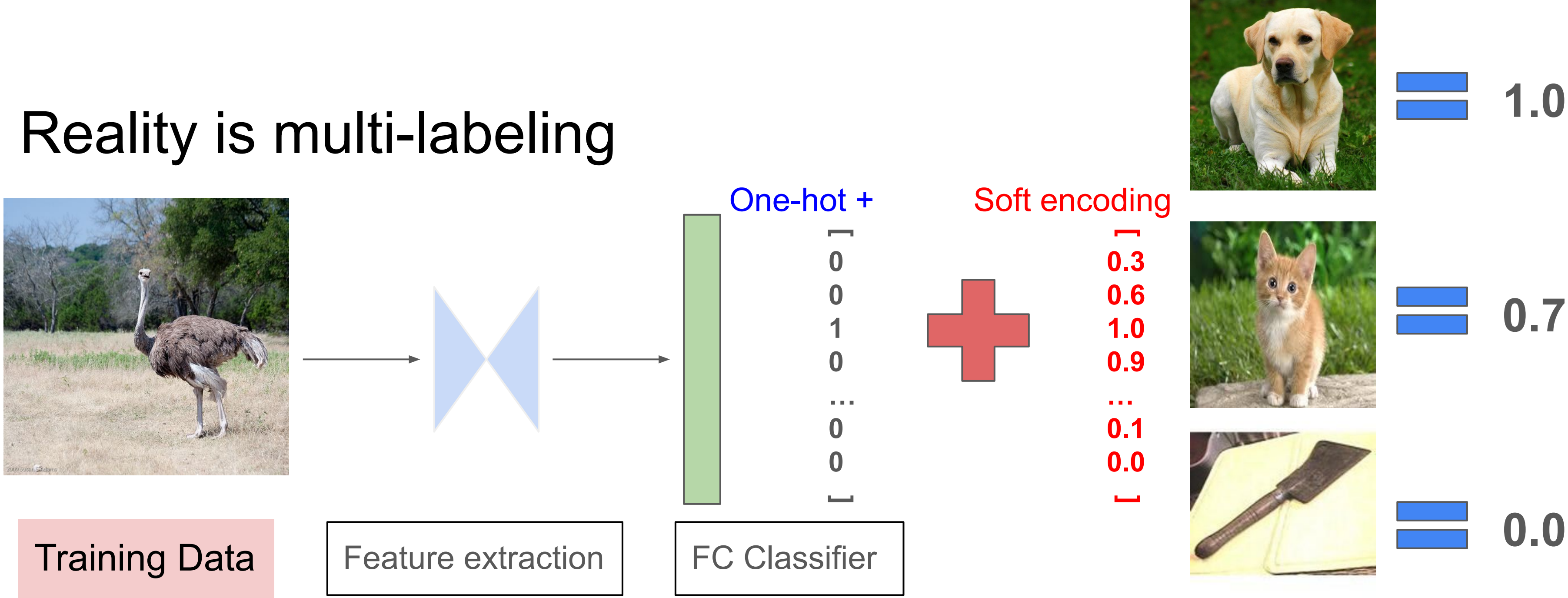
Only adopting one-hot-encoding is vulnerable to spurious correlation during training.



Only adopting one-hot-encoding is vulnerable to spurious correlation during training.

One-hot encoding assumes that the likelihood of all the non-GT classes are *created equal*.
Discrimination between semantic closer class will force model ignore shared feature, which is more transferable.

Reality is multi-labeling



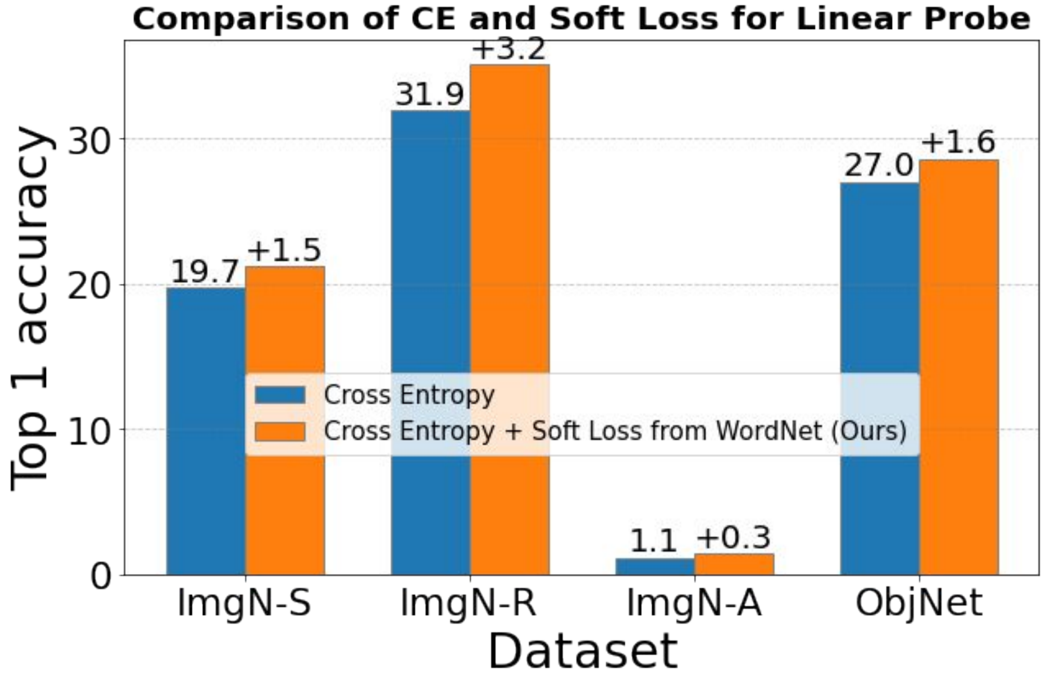
Only adopting one-hot-encoding is vulnerable to spurious correlation during training.

One-hot encoding assumes that the likelihood of all the non-GT classes are *created equal*.
Discrimination between semantic closer class will force model ignore shared feature, which is more transferable.

Adopting soft labels (constructed from the ontology) can better regularize the training, resulting into a more generalizable model to OOD data.

Experiment 2: Linear Probing Experiment

- Baseline: Trained with cross entropy loss
- Ours: Trained with cross entropy loss + soft label loss from hierarchy



Adopting hierarchy as soft labels *boosts OOD performance* without affecting ID accuracy!

LCA distance as robust generalization indicator

1. What is LCA distance?

2. Why should we use LCA distance?

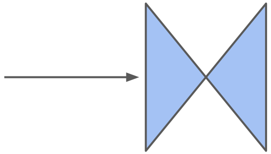
3. How can we use LCA distance to improve model generalization?

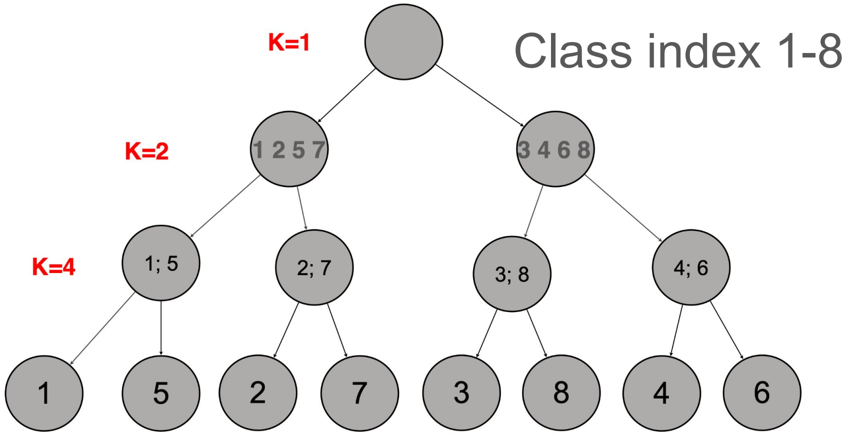
Wait! My dataset doesn't have a predefined hierarchy?

Latent hierarchy(class distance) on any datasets with clustering

- WordNet hierarchy is manually designed.
- We can also construct a hierarchy by clustering per-class features.

Pretrained model
(e.g., CLIP)





Step1: Extract per-class mean features over all classes

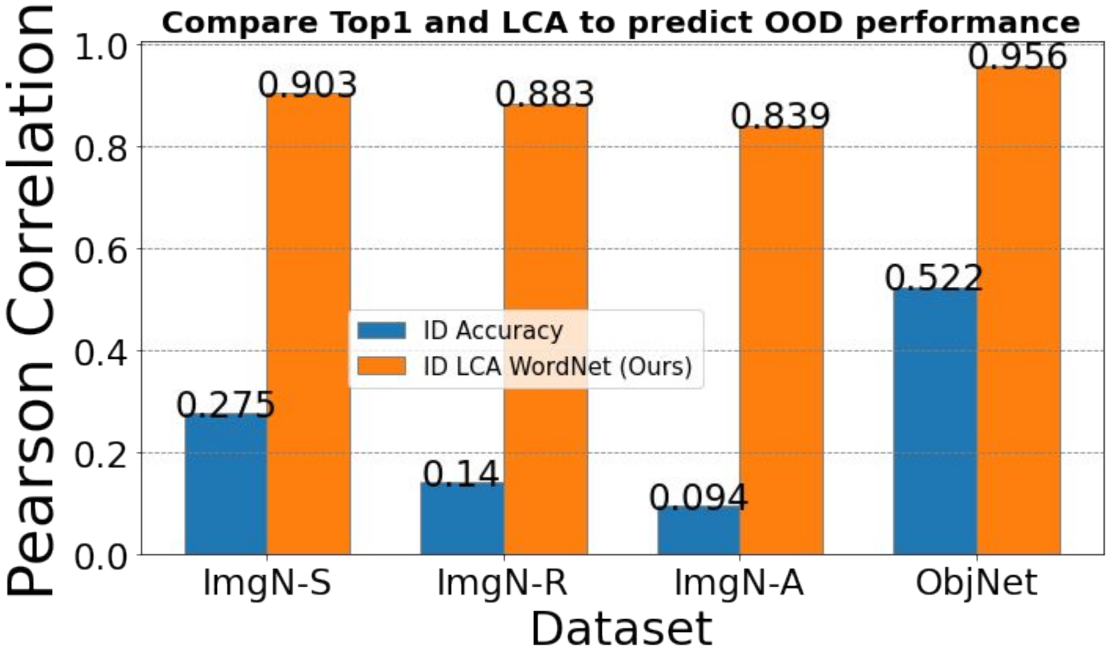
Step2: Cluster them towards a hierarchy

Does Latent hierarchy helps as well as WordNet ?

Experiment 1: Predict OOD from ID

Correlation comparison against OOD accuracy.

- **Baseline**: Accuracy-on-the-line[1] (ID accuracy)
- **Ours**: LCA-on-the-line (ID LCA distance on **WordNet**)

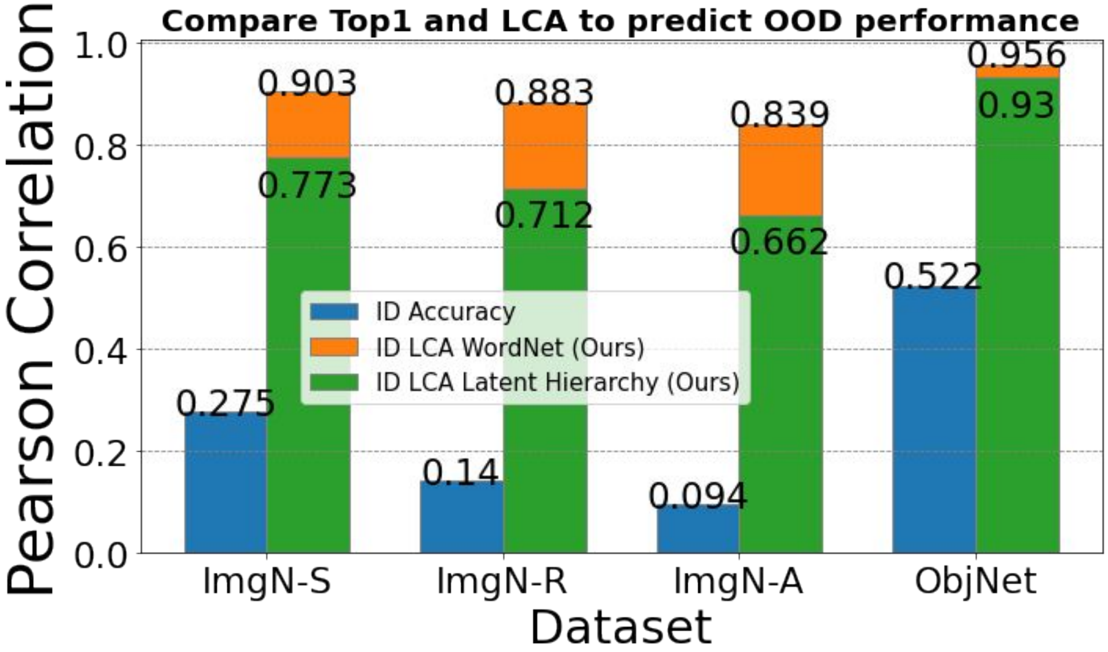


Does Latent hierarchy helps as well as WordNet ?

Experiment 1: Predict OOD from ID

Correlation comparison against OOD accuracy.

- **Baseline**: Accuracy-on-the-line[1] (ID accuracy)
- **Ours**: LCA-on-the-line (ID LCA distance on **WordNet**)
- **Ours**: LCA-on-the-line (ID LCA distance on **Latent Hierarchy**)

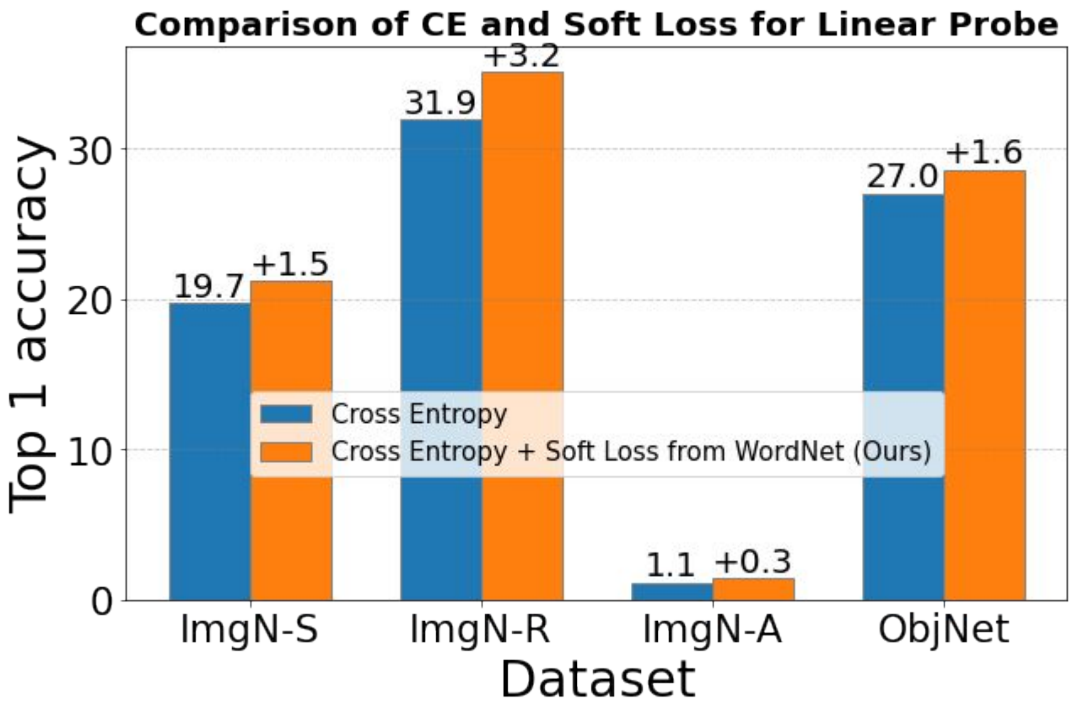


Constructed latent hierarchies similarly shows strong correlation to OOD performance.

Does Latent hierarchy helps as well as WordNet ?

Experiment 2: Linear Probing over Res18

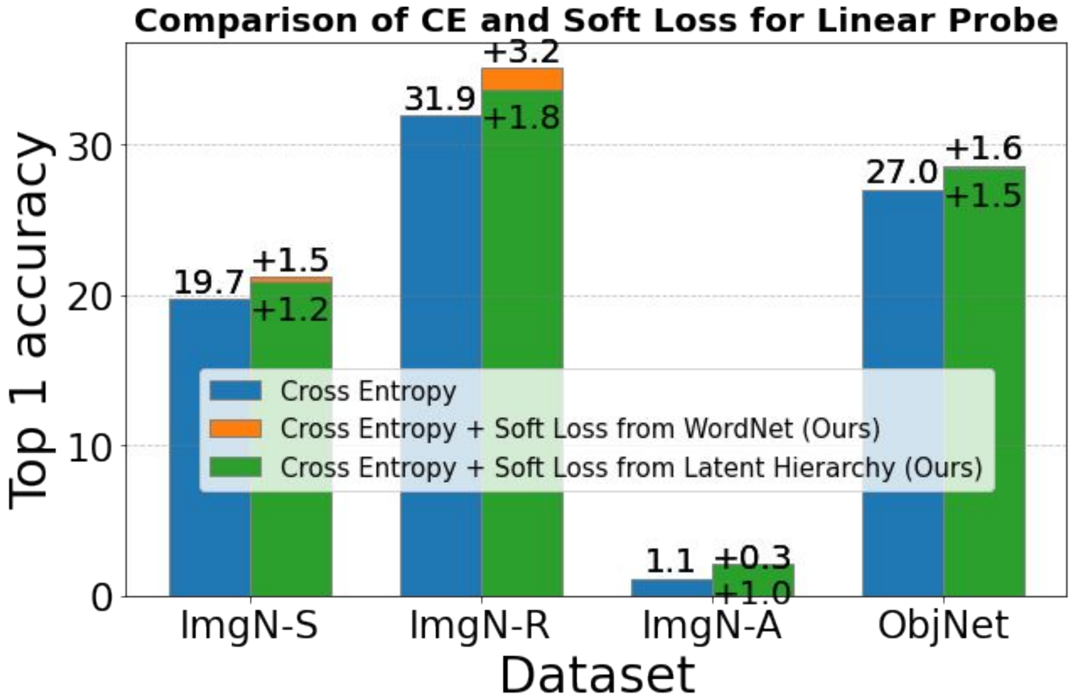
- **Baseline**: training with cross entropy loss
- **Ours**: training with cross entropy loss + soft label loss **(WordNet)**



Does Latent hierarchy helps as well as WordNet ?

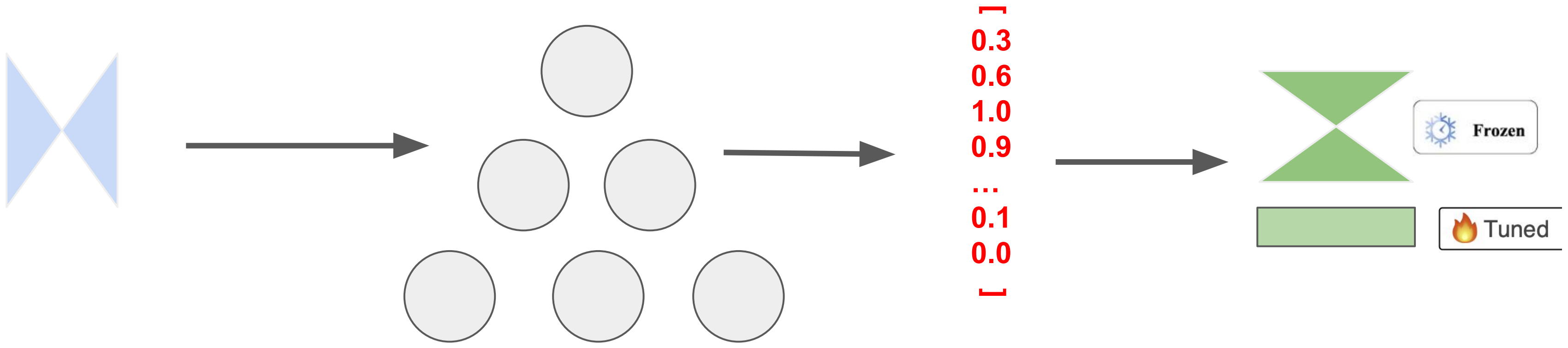
Experiment 2: Linear Probing over Res18

- **Baseline**: training with cross entropy loss
- **Ours**: training with cross entropy loss + soft label loss **(WordNet)**
- **Ours**: training with cross entropy loss + soft label loss **(Latent)**



Learning with a constructed latent hierarchy consistently boosts OOD performance.

Recall: Construct soft labels from latent hierarchy



Pretrained **source model** (e.g., ResNet/CLIP)

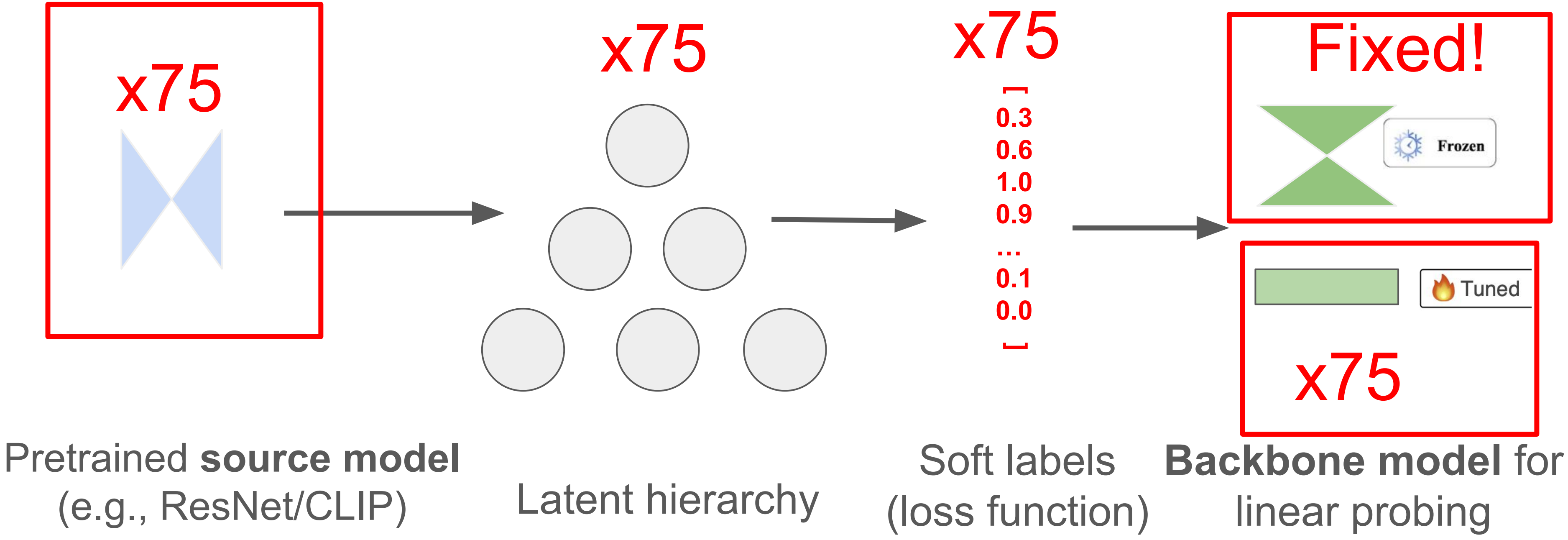
Latent hierarchy

Soft labels (loss function)

Backbone model for linear probing

75 pre-trained model can construct 75 groups soft labels

Do better soft labels emerge in more generalizable models?

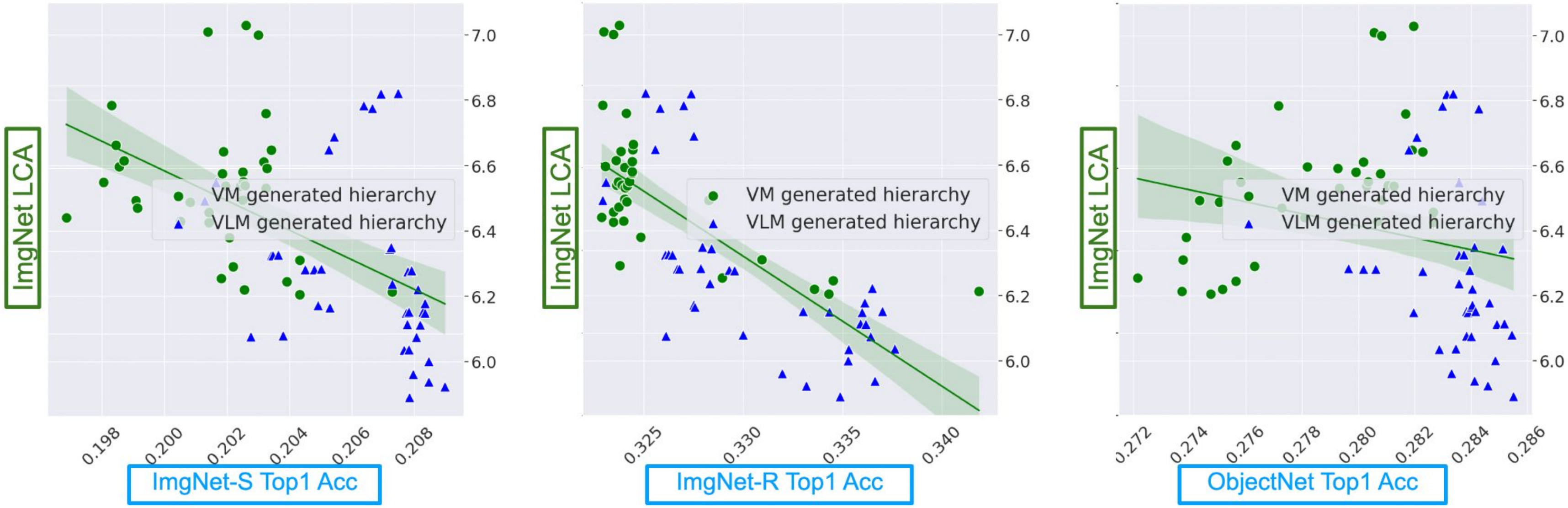


Do more generalizable models form better soft labels??

Do better soft labels emerge in more generalizable models?

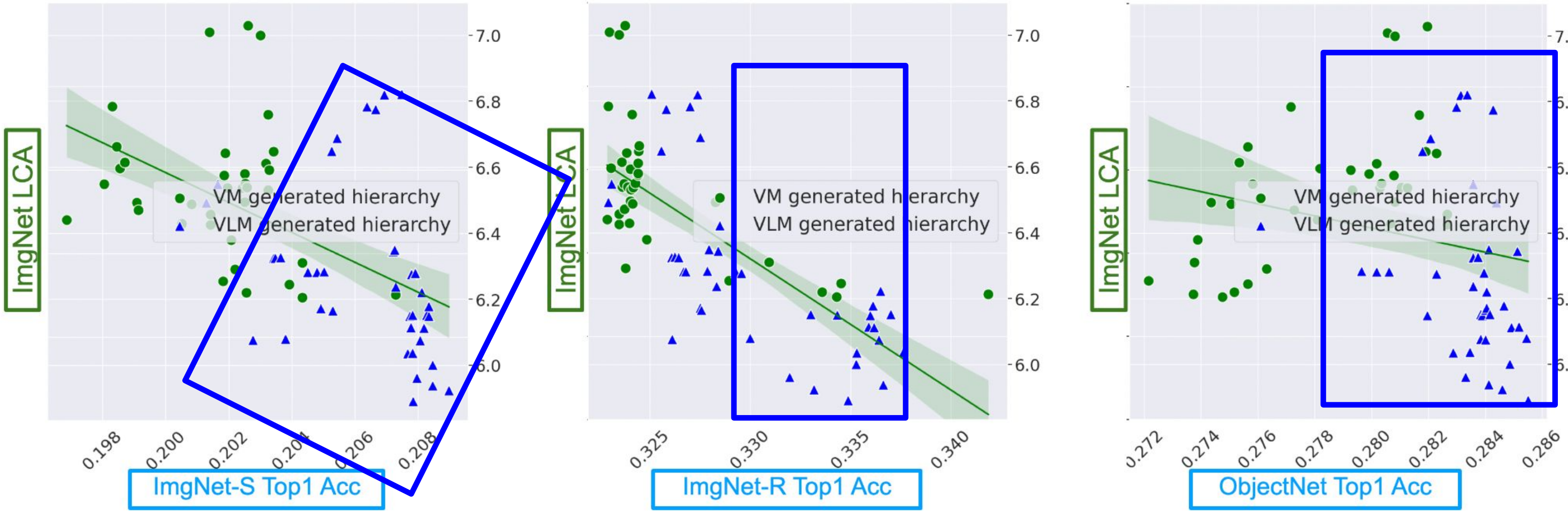
Yes!

- ***y*-axis**: LCA distance on ImageNet (ID dataset) between WordNet hierarchy and each of the source pretrained models (that generate hierarchies).
- ***x*-axis**: top-1 accuracy on an OOD dataset by linear probing over each of the generated hierarchies.



Alternative view behind VLM’s generalization

- Soft labels generated by VLMs help more for OOD generalization than VMs (cf. better LCA and better OOD top-1).
- Note that benchmarks often simulate human-world ontology (e.g., top-1 accuracy on OOD data). That said, VLM's high-level perceptual understanding better aligns with human-world ontology.



Conclusion

1. LCA distance robustly predict models' OOD performance.
2. LCA distance suggests how to improve models' generalization.
3. LCA distance offers insights why VLMs generalize so well.

Paper updated after camera ready!

Our Project Page

Conclusion

1. LCA distance robustly predict models' OOD performance.
2. LCA distance suggests how to improve models' generalization.
3. LCA distance offers insights why VLMs generalize so well.

Paper updated after camera ready!

***LCA-on-the-Line*: Benchmarking Out of Distribution Generalization with Class Taxonomies**

Poster Section: Hall C 4-9 #701, 11:30

Our Project Page