Analyzing Gender Biases in Visual Recognition Models

M.S. Thesis Defense By Jaspreet Ranjit

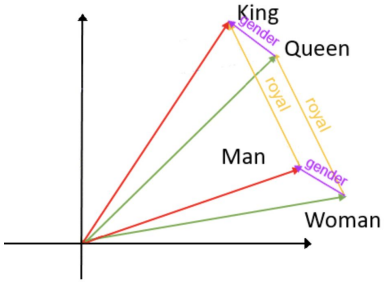
Committee:
Yangfeng Ji (Chair), Vicente Ordóñez (Advisor), Matthew Dwyer

Motivation

Gender Biases in Natural Language Processing

$$X = woman + doctor - man \approx \boldsymbol{nurse}$$

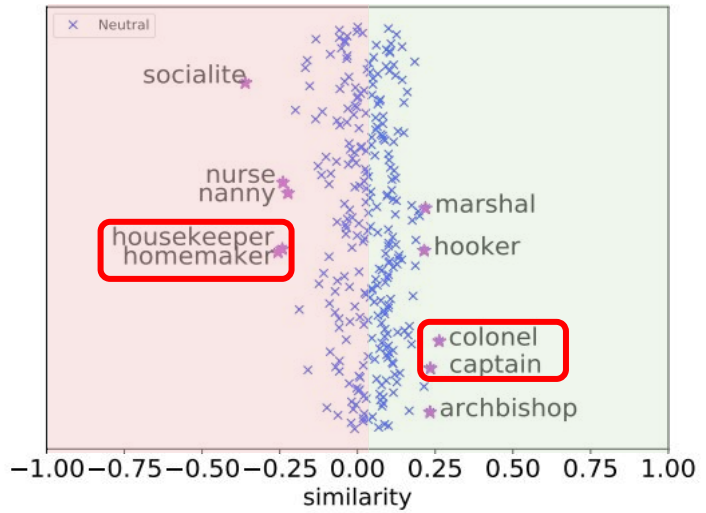
$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$



Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Tolga Bolukbasi et al. NIPS 2016
Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation. Tianlu Wang et al. ACL 2020

Gender Bias in Word Embeddings

- Gender-neutral professions such as ‘nurse’ and ‘captain’ possess no gender association by definition
- GloVe word vectors exhibit strong gender stereotypes in occupations

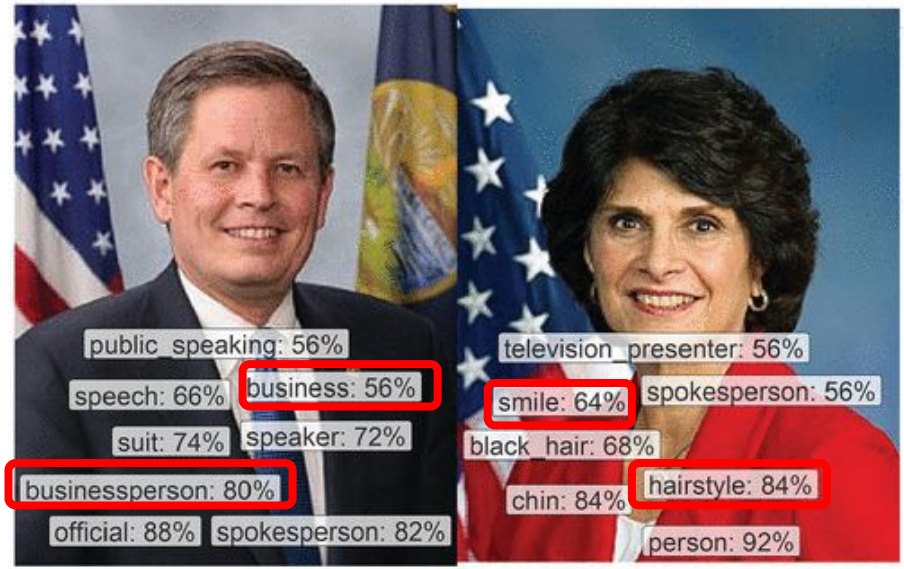


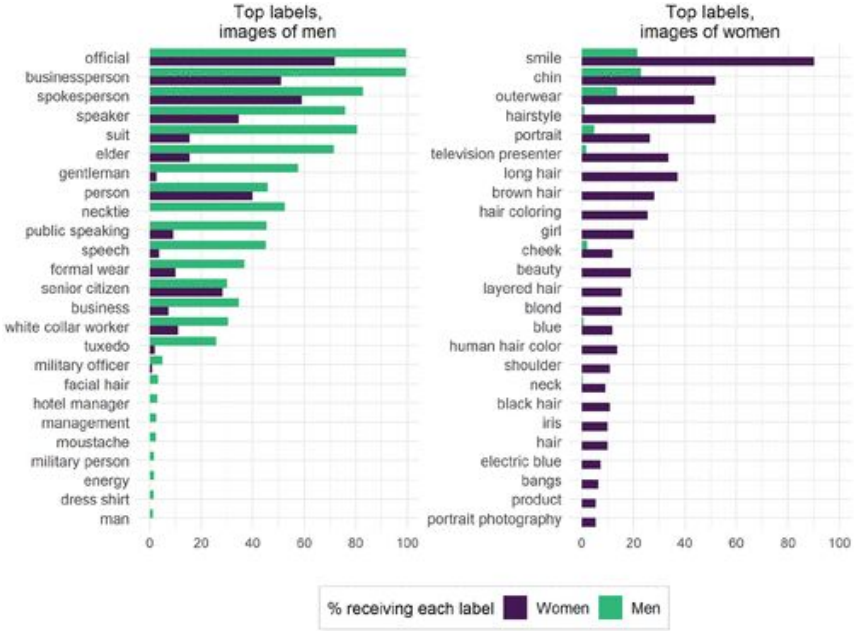
(b) Gender-neutral profession words projected to gender direction in GloVe

Learning Gender Neutral Word Embeddings. Jieyu Zhao et al. EMNLP 2018

Gender Biases in Computer Vision

- Commercial image classifiers such as google cloud vision exhibit gender biases where women receive three times more annotations related to physical appearance





Diagnosing Gender Bias in Image Recognition Systems. Carsten Schwemmer et al. 2020

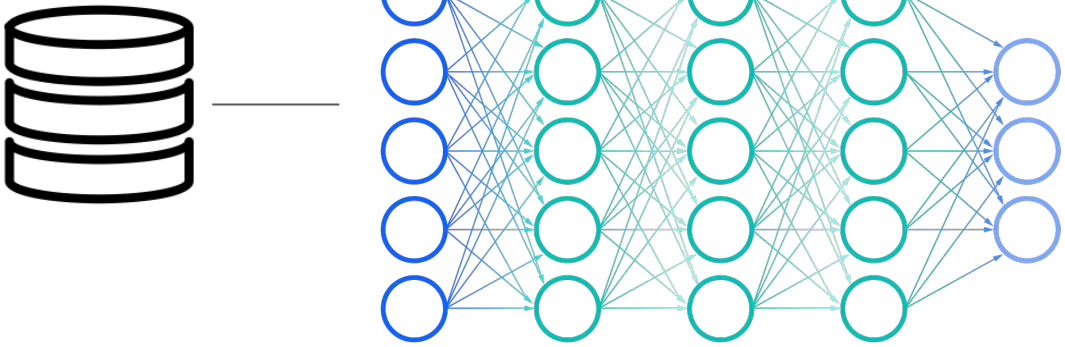
Gender Biases in Image Classifiers

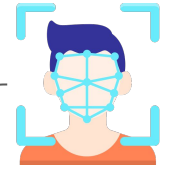
- Biases in datasets (COCO) can be reflected in predictions of image classifiers
- Sports like skiing, snowboarding, and surfboarding are more closely associated with men thus misleading the model's predictions

Figure 11: The model classifies the women in these pictures as men in the COCO dataset.

Testing DNN Image Classifiers for Confusion and Bias Errors. Yuchi Tian et al. ICSE 2020.

Why do we care?





Google's algorithms discriminate against women and people of colour *The Conversation. 2019*

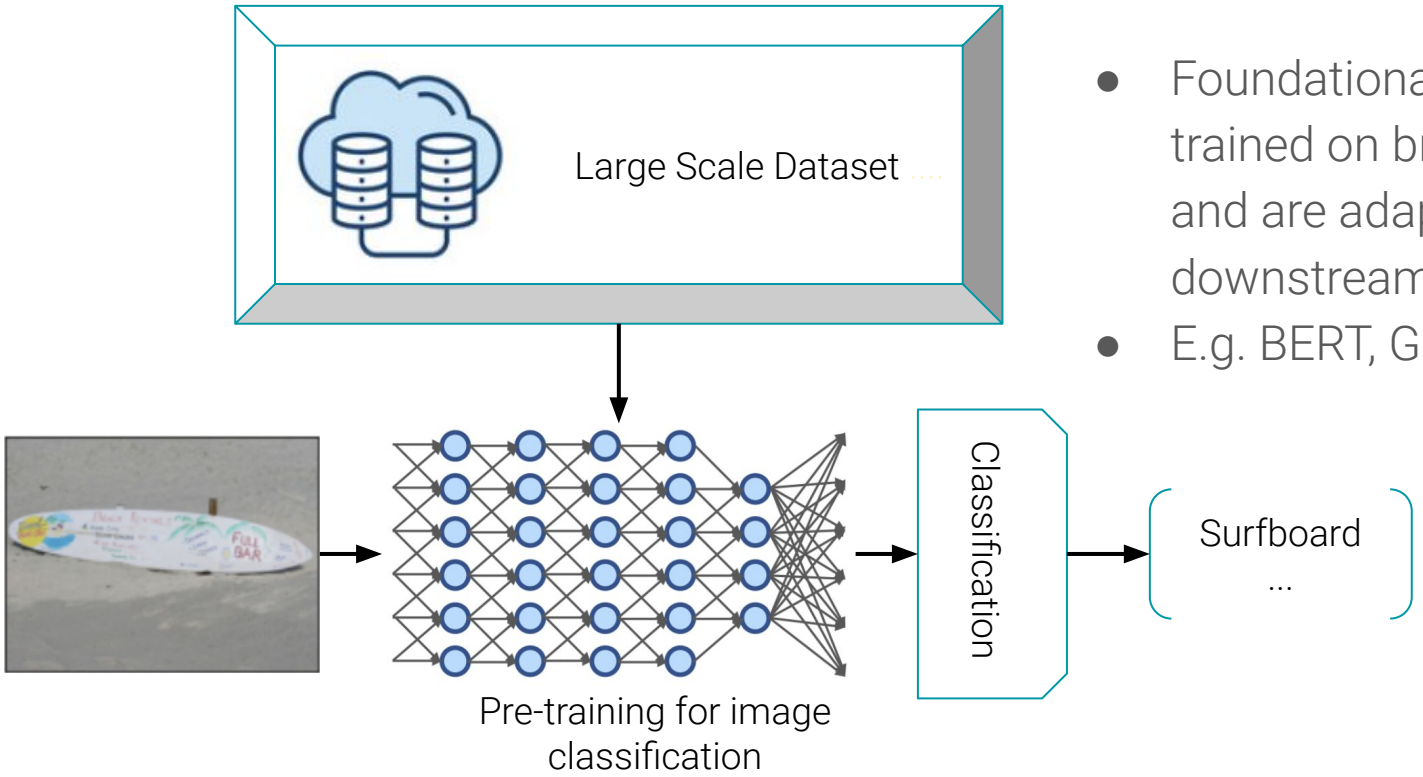
Student proves Twitter algorithm 'bias' toward lighter, slimmer, younger faces *The Guardian. 2021*

How LinkedIn's search engine may reflect a gender bias *The Seattle Times. 2016*

Google apologizes after its Vision AI produced racist results *AlgorithmWatch. 2020*

Background

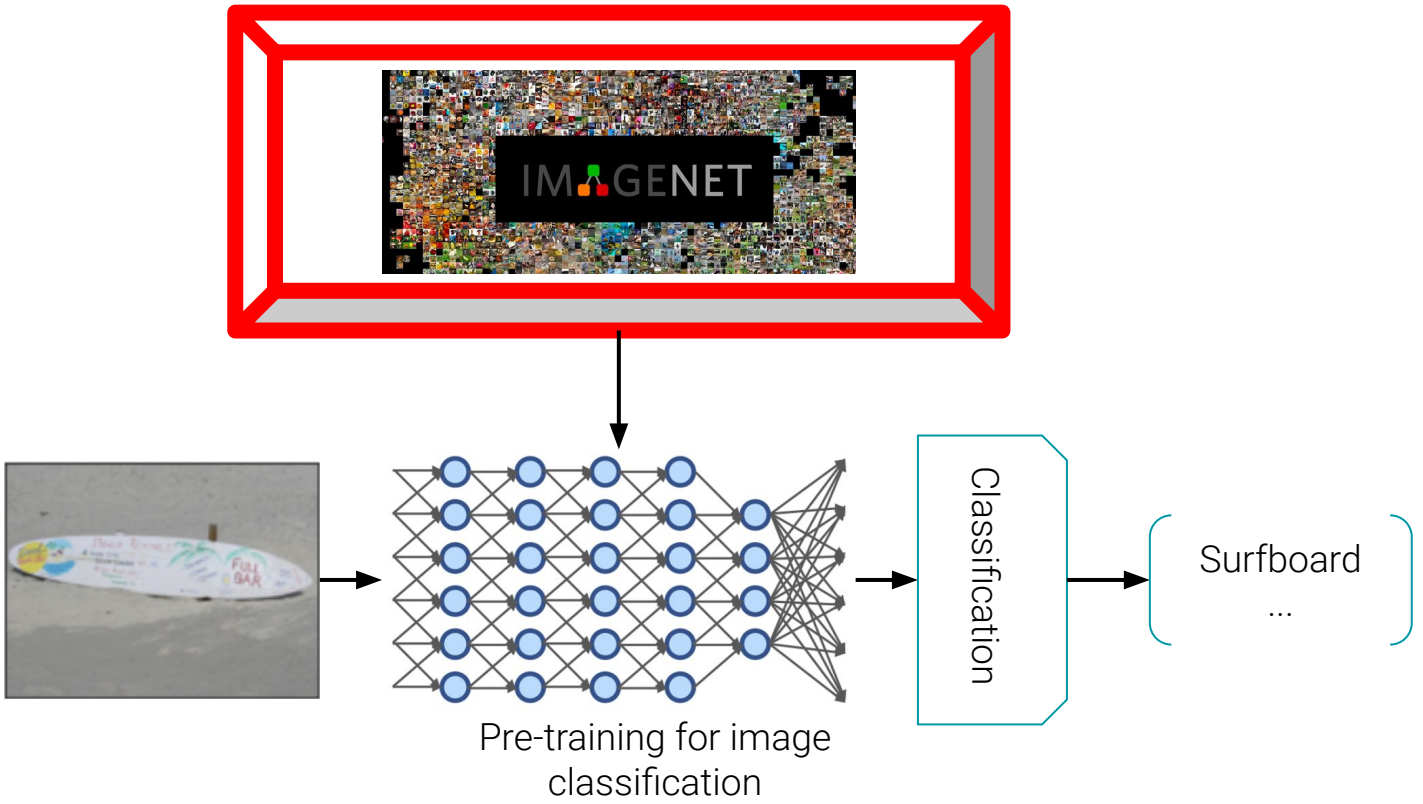
Foundational Models



- Foundational models are trained on broad data at scale and are adaptable to downstream tasks
- E.g. BERT, GPT-3

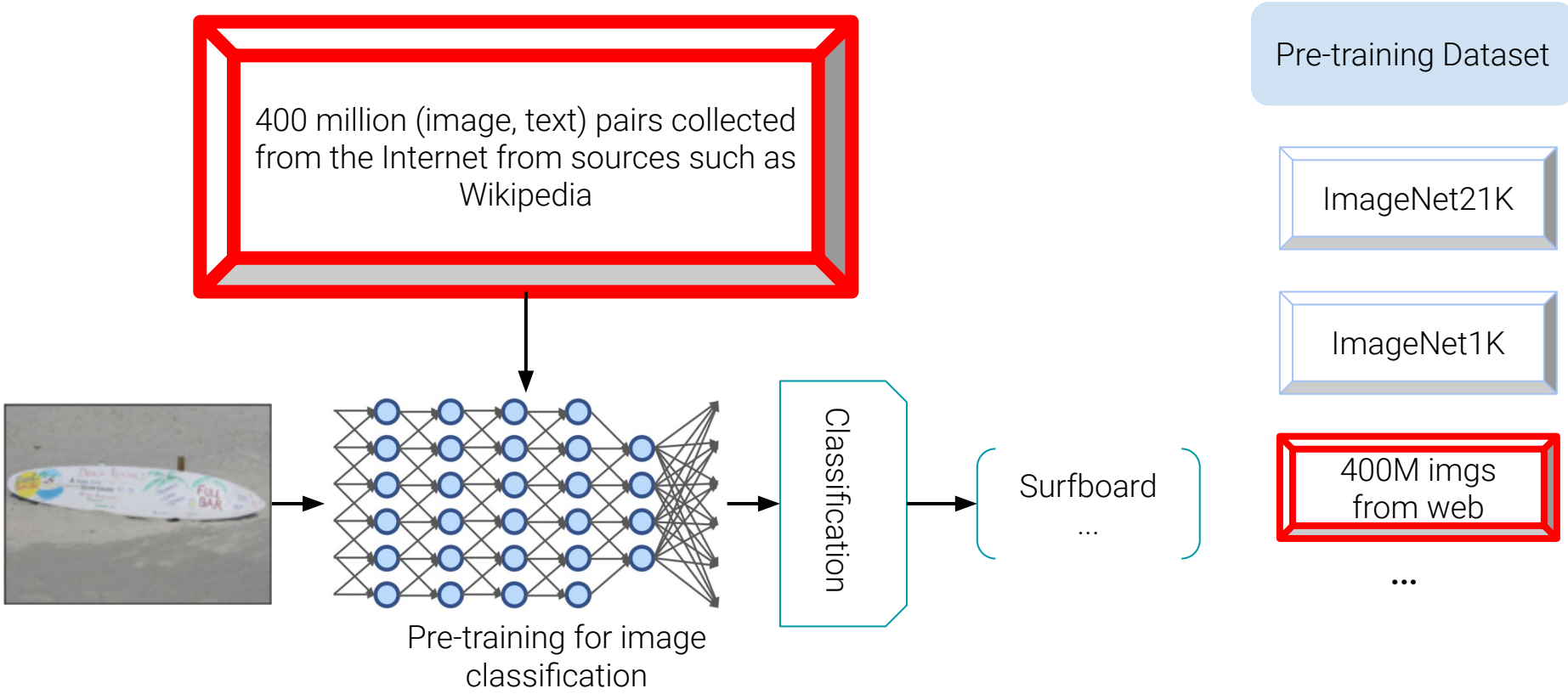
On the Opportunities and Risks of Foundational Models. Rishi Bommasani et al. CoRR 2021.

Current Works - Pre-training Dataset



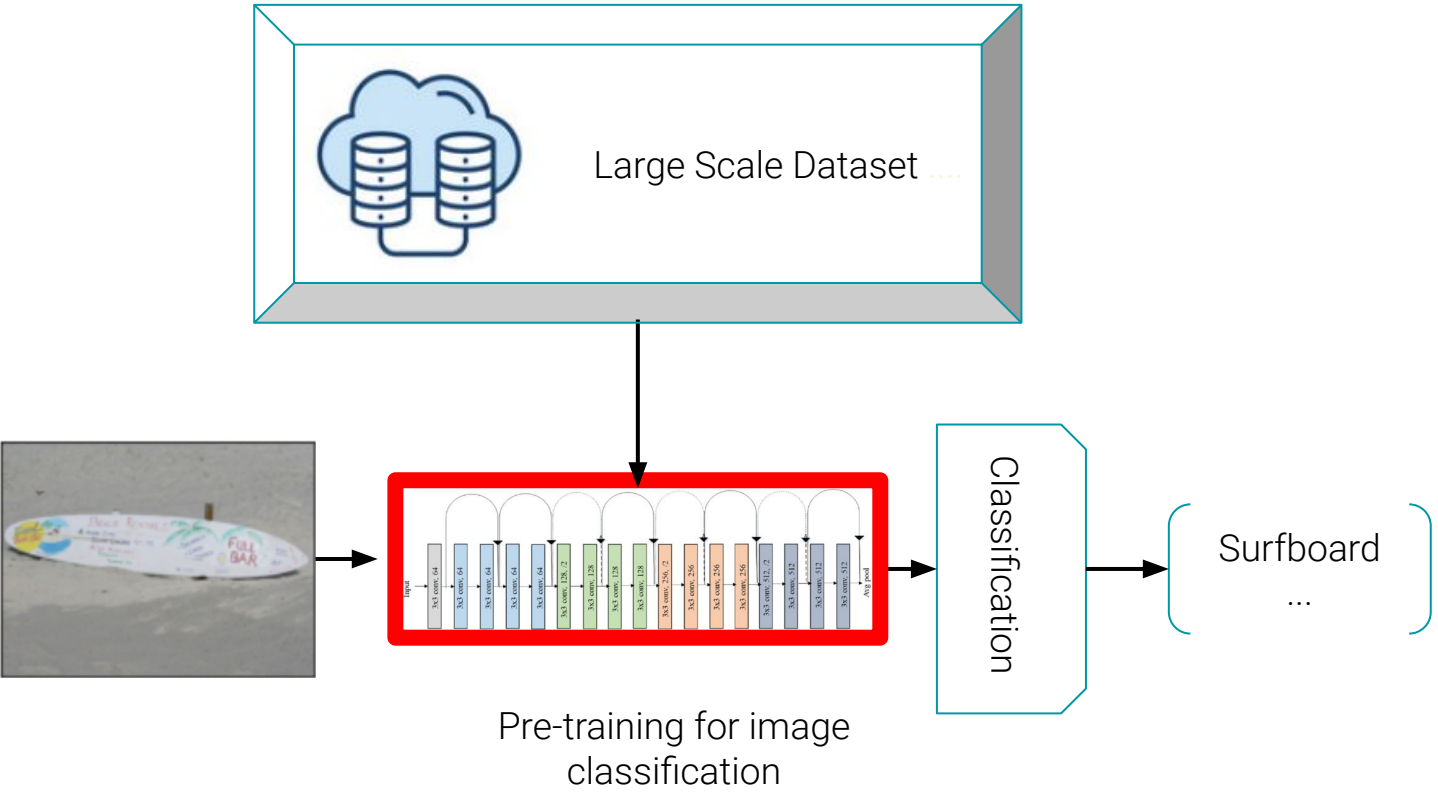
On the Opportunities and Risks of Foundational Models. Rishi Bommasani et al. CoRR 2021.

Current Works - Pre-training Dataset



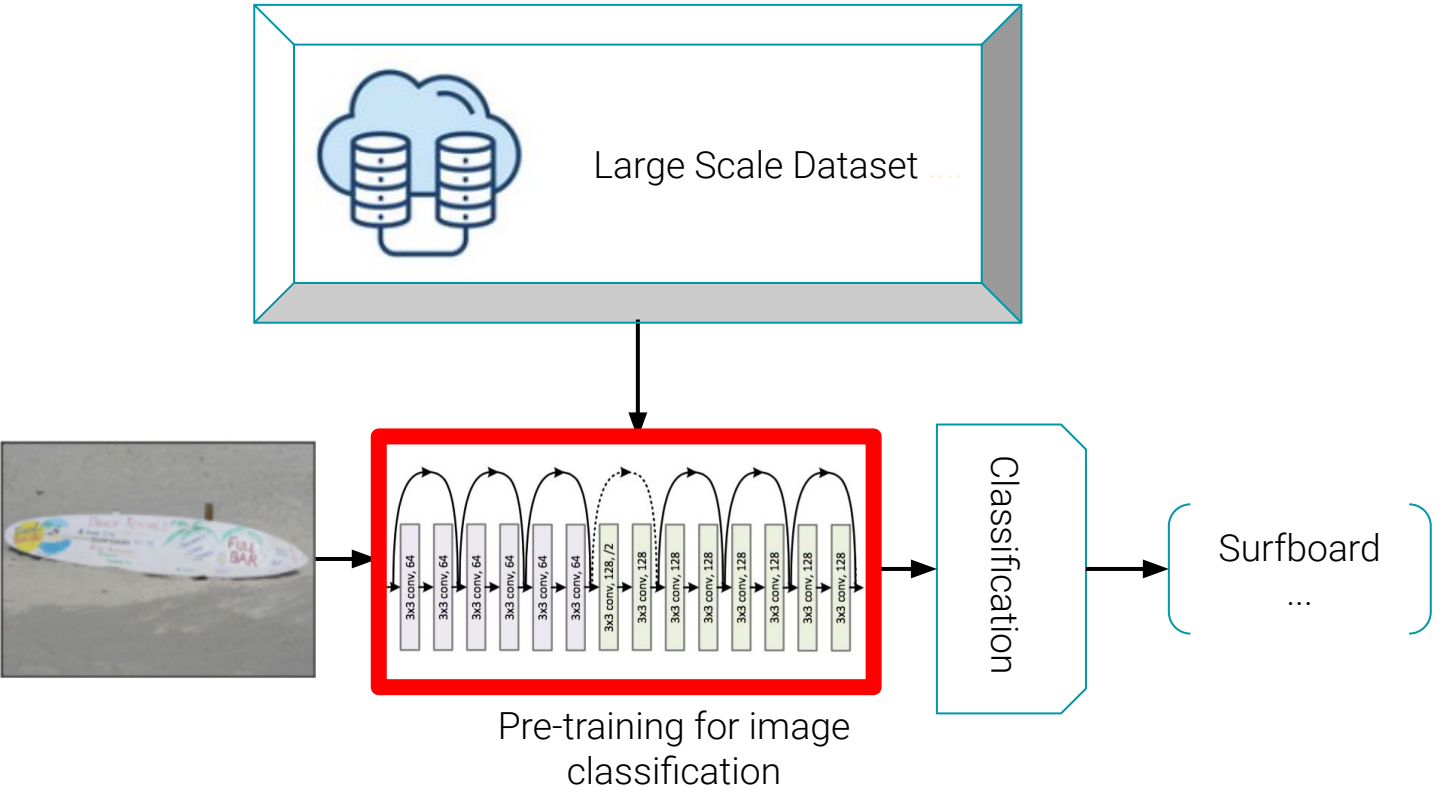
On the Opportunities and Risks of Foundational Models. Rishi Bommasani et al. CoRR 2021.

Current Works - Network Architecture



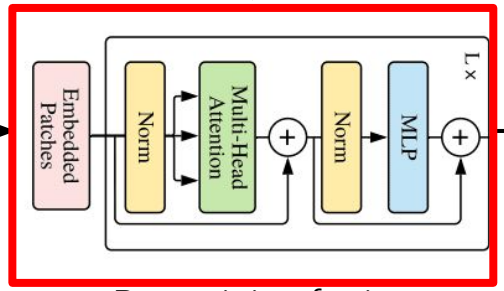
On the Opportunities and Risks of Foundational Models. Rishi Bommasani et al. CoRR 2021.

Current Works - Network Architecture



On the Opportunities and Risks of Foundational Models. Rishi Bommasani et al. CoRR 2021.

Current Works - Network Architecture



Pre-training for image classification

Classification

Surfboard
...

Network Architecture

ResNet18

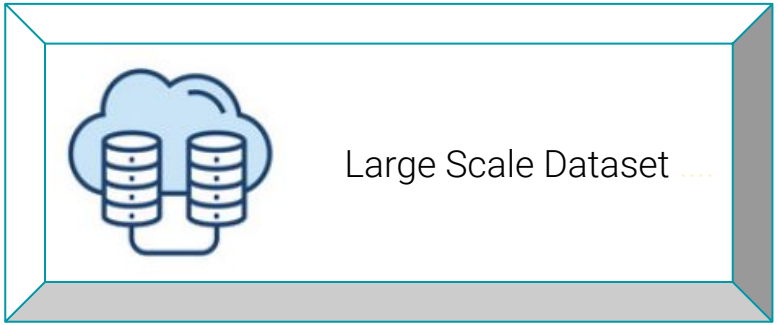
ResNet50

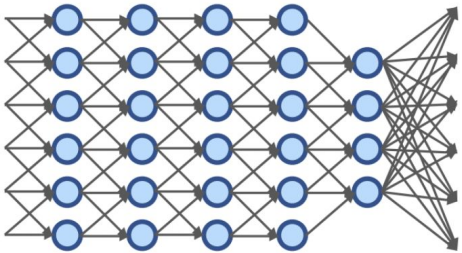
CLIP: ViT/B-32

...

On the Opportunities and Risks of Foundational Models. Rishi Bommasani et al. CoRR 2021.

Current Works - Training Setting





Pre-training for image classification

Classification

Surfboard
...

Training Setting

Supervised

Self-Supervised

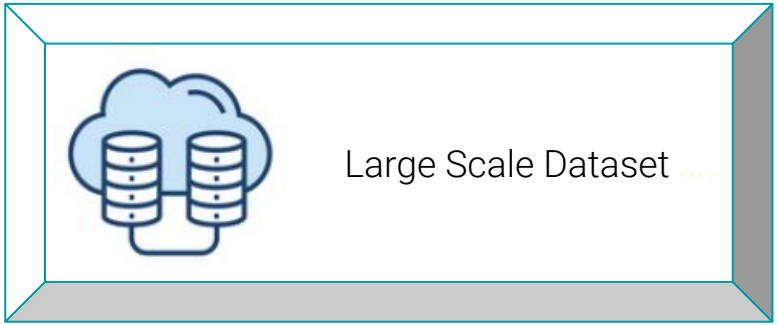
Weakly Supervised

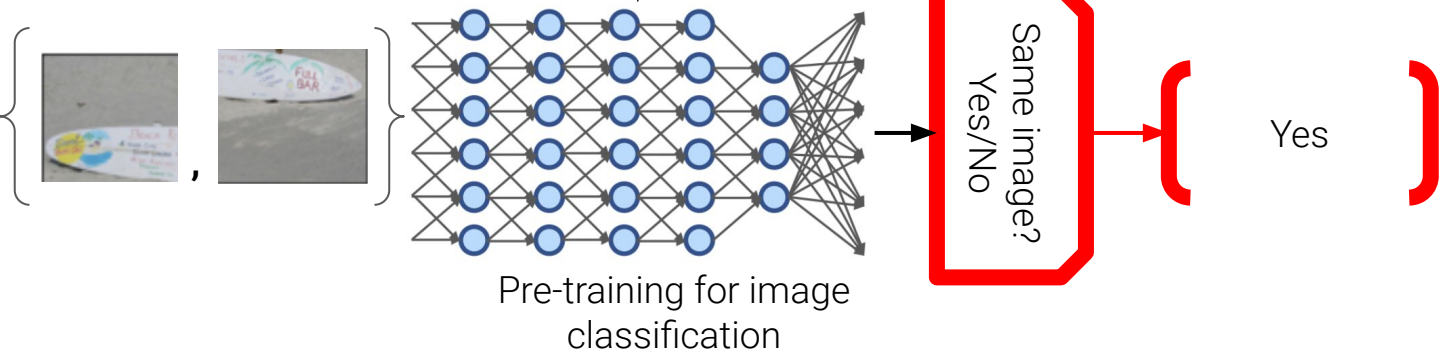
Semi Supervised

...

On the Opportunities and Risks of Foundational Models. Rishi Bommasani et al. CoRR 2021.

Current Works - Training Setting



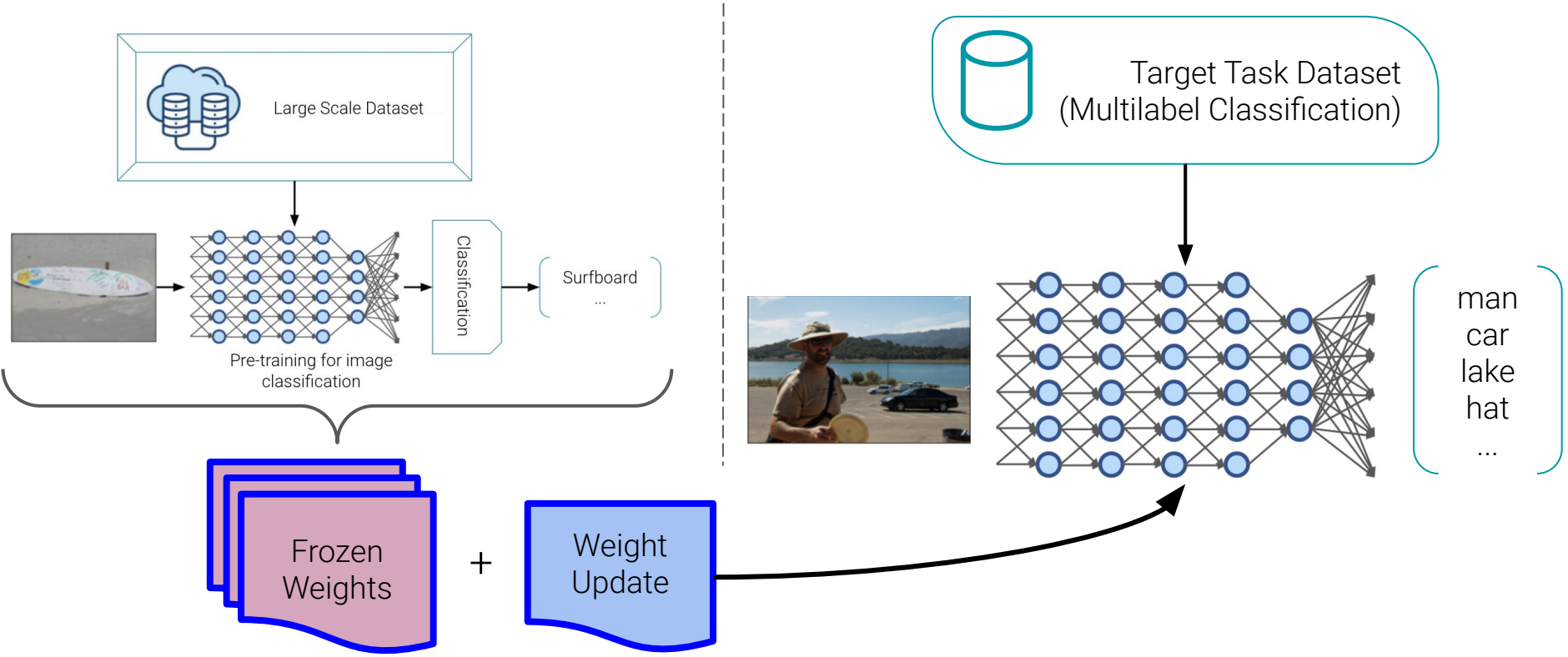


Training Setting

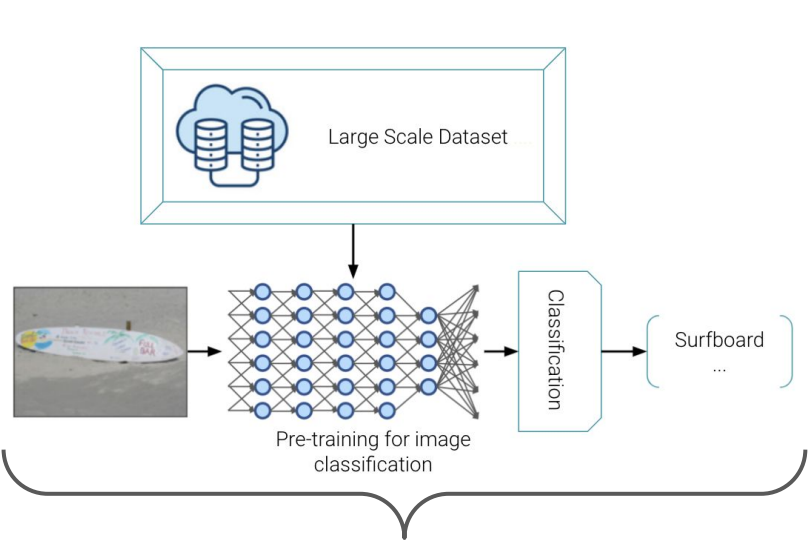
- Supervised
- **Self-Supervised**
- Weakly Supervised
- Semi Supervised
- ...

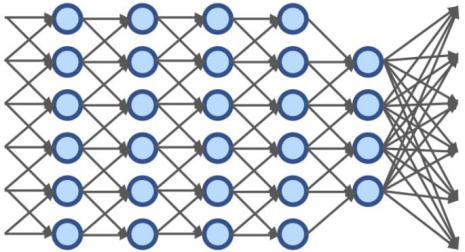
On the Opportunities and Risks of Foundational Models. Rishi Bommasani et al. CoRR 2021.

Model Application



Model Application - Finetuning Datasets





man
car
lake
hat
...

Frozen Weights

+

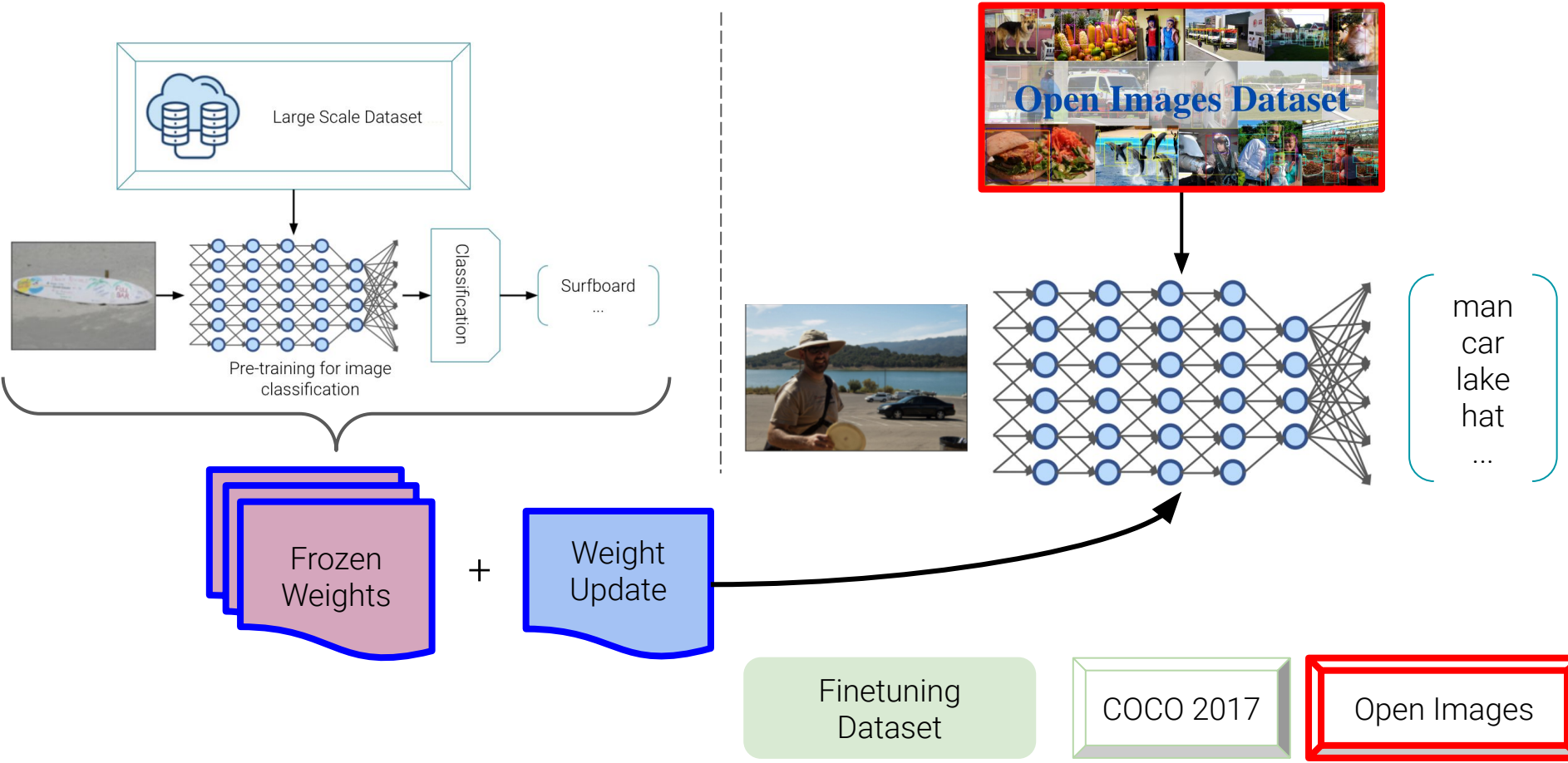
Weight Update

Finetuning Dataset

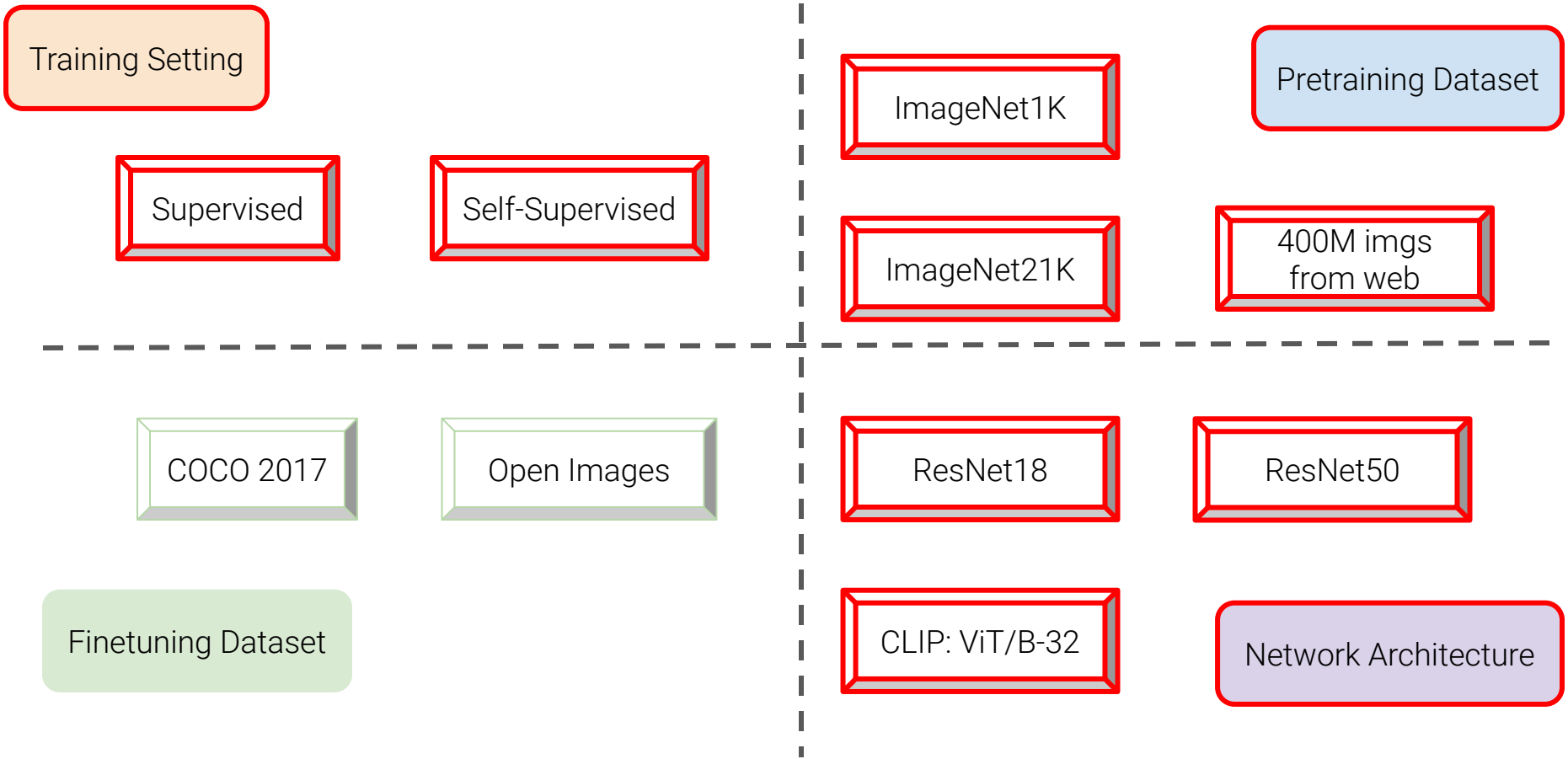
COCO 2017

Open Images

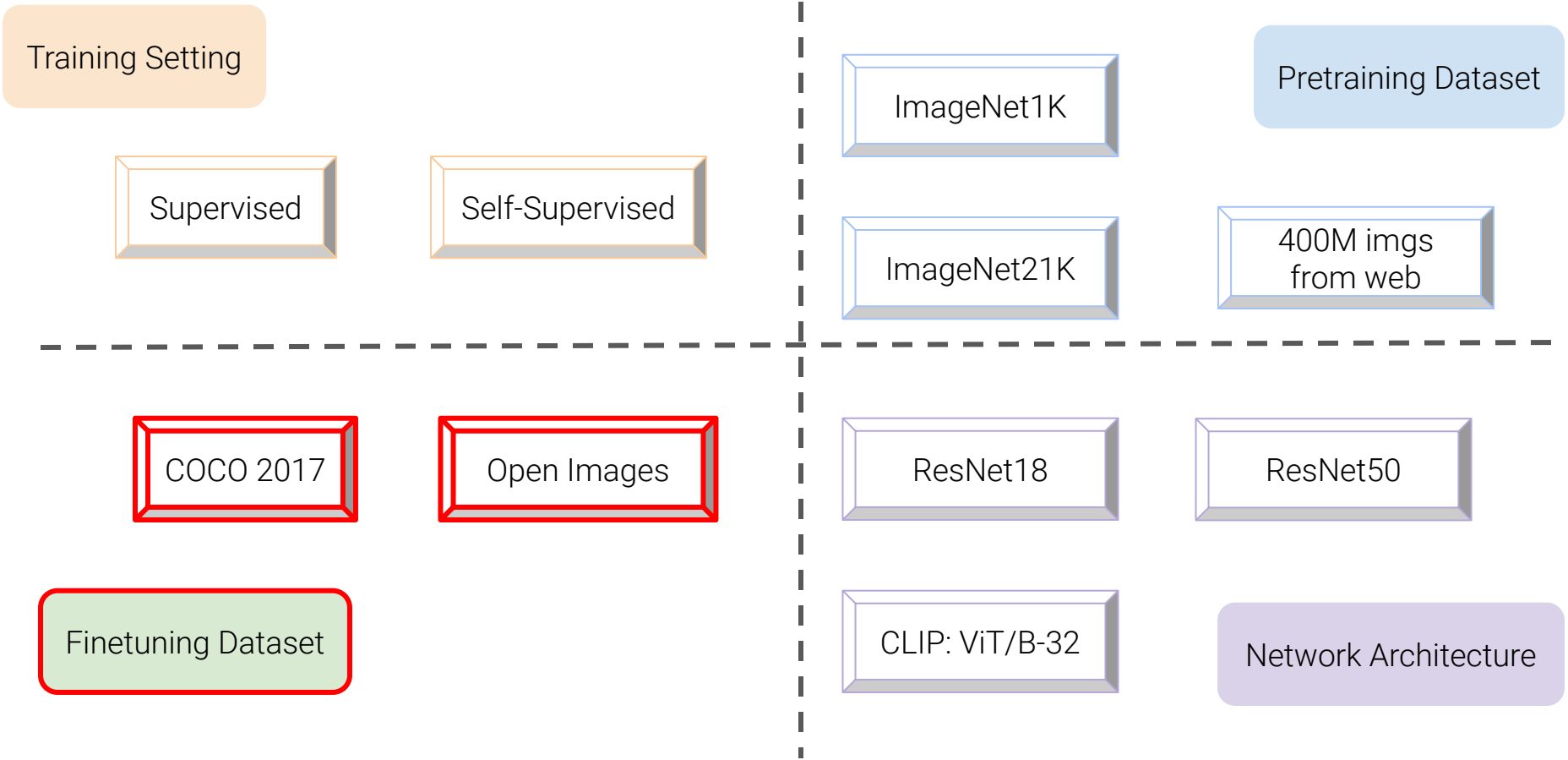
Model Application - Finetuning Datasets



Pretraining Setting: Variables



Finetuning Setting: Variables



My Research: Analyzing Biases in Visual Recognition Models

Training Setting

Pretrained Dataset

ImageNet21K

400M imgs from web

Gathering labeled data for bias exploration

Identifying metrics to represent biases at the feature level

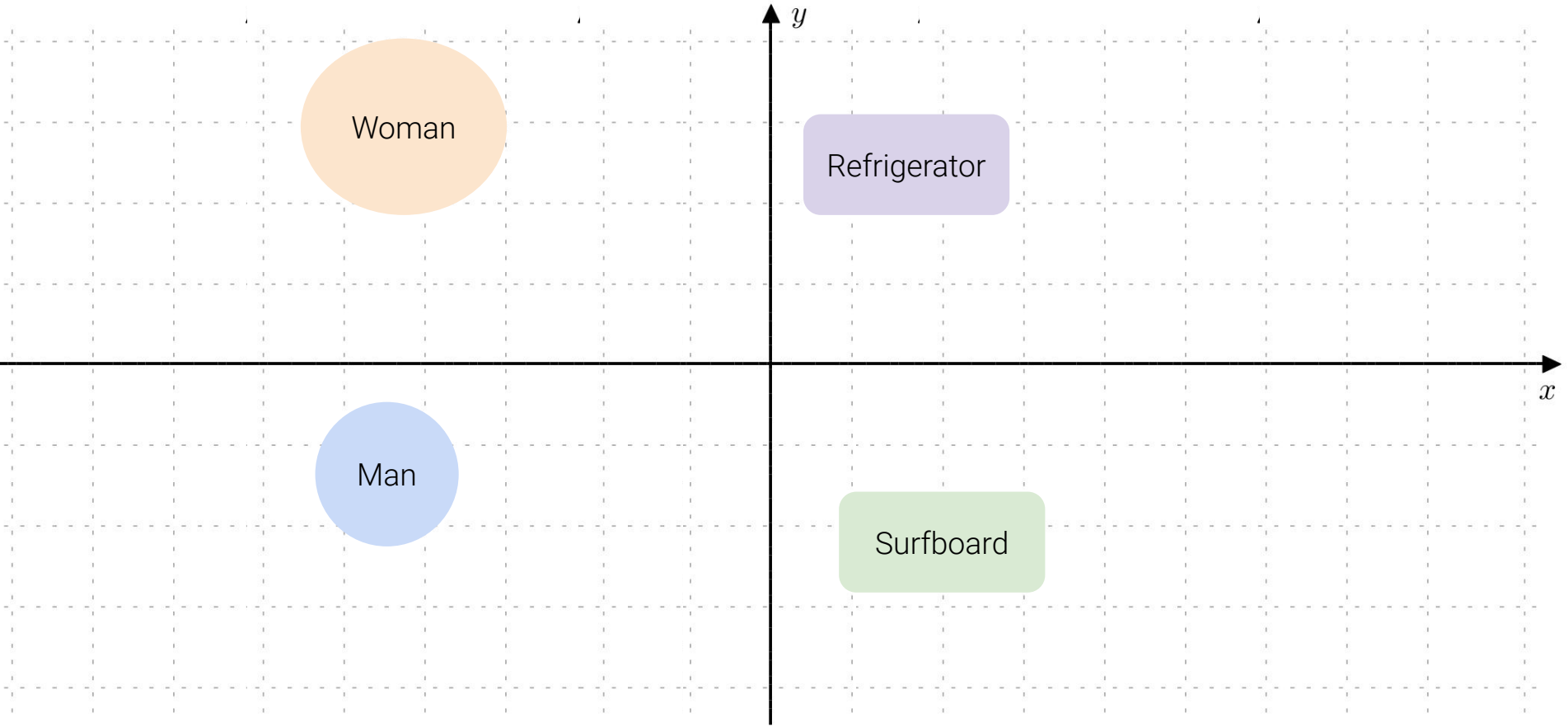
Finetuned Dataset

CLIP: ViT/B-32

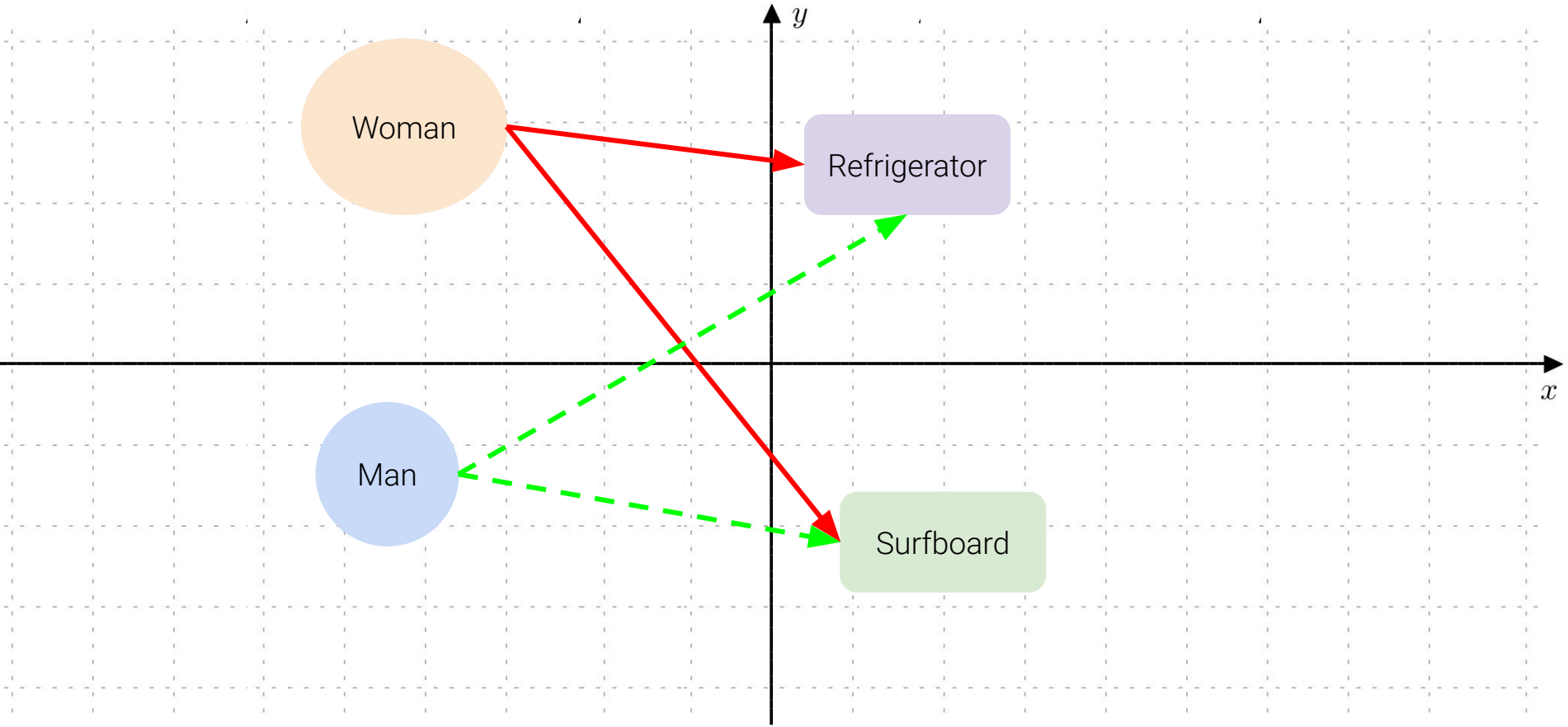
Network Architecture

Problem Definition

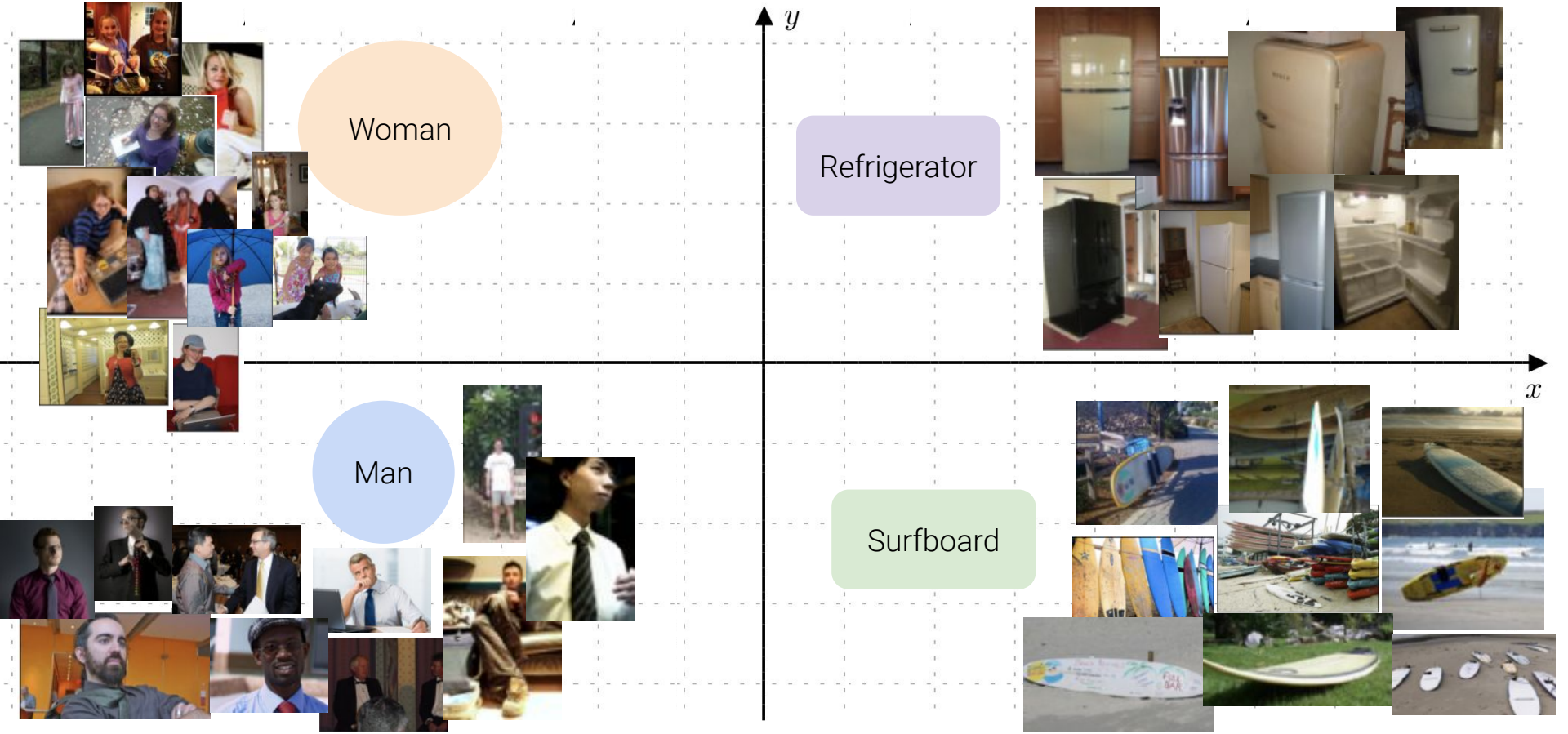
Measuring Biases in Text through Analogies



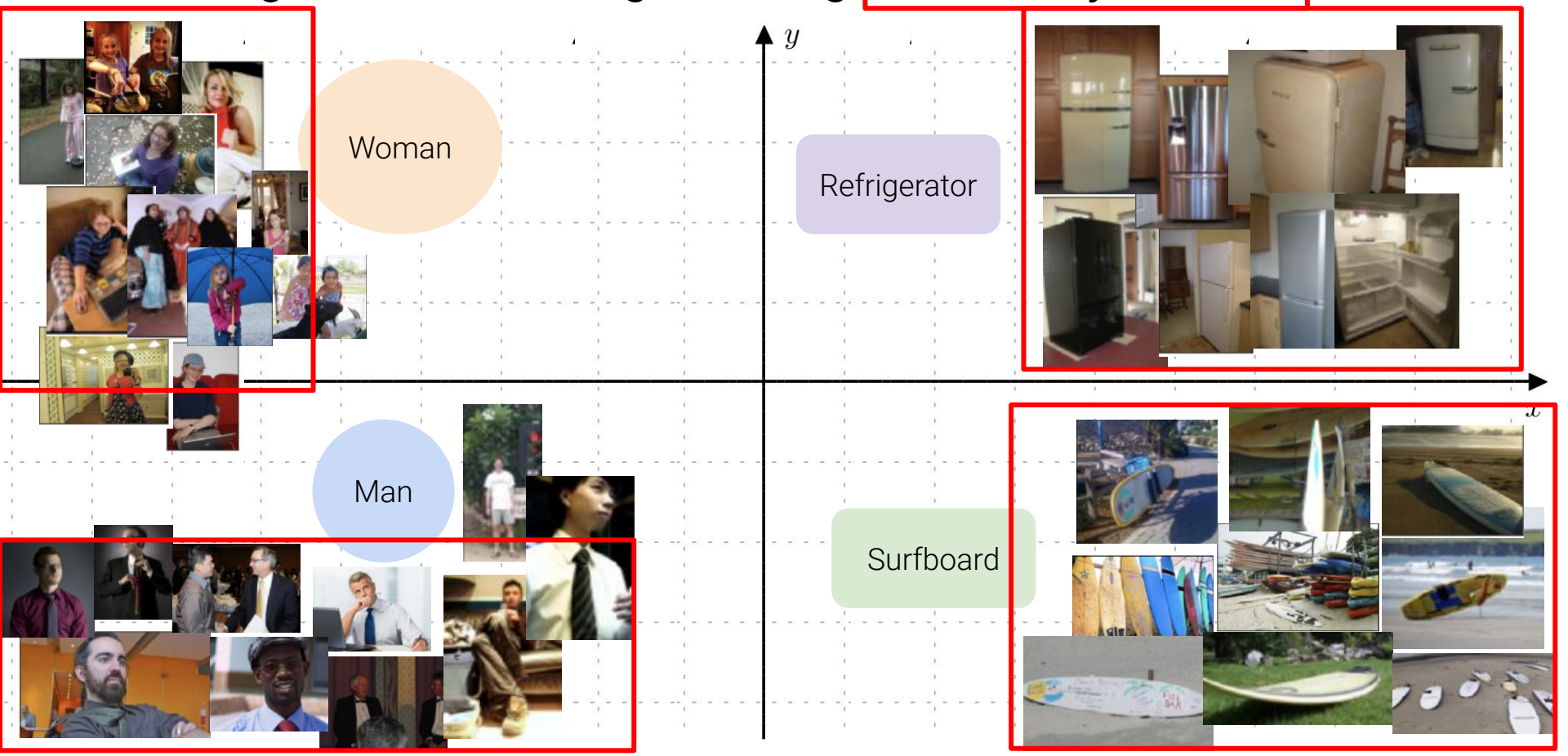
Measuring Biases in Text through Analogies



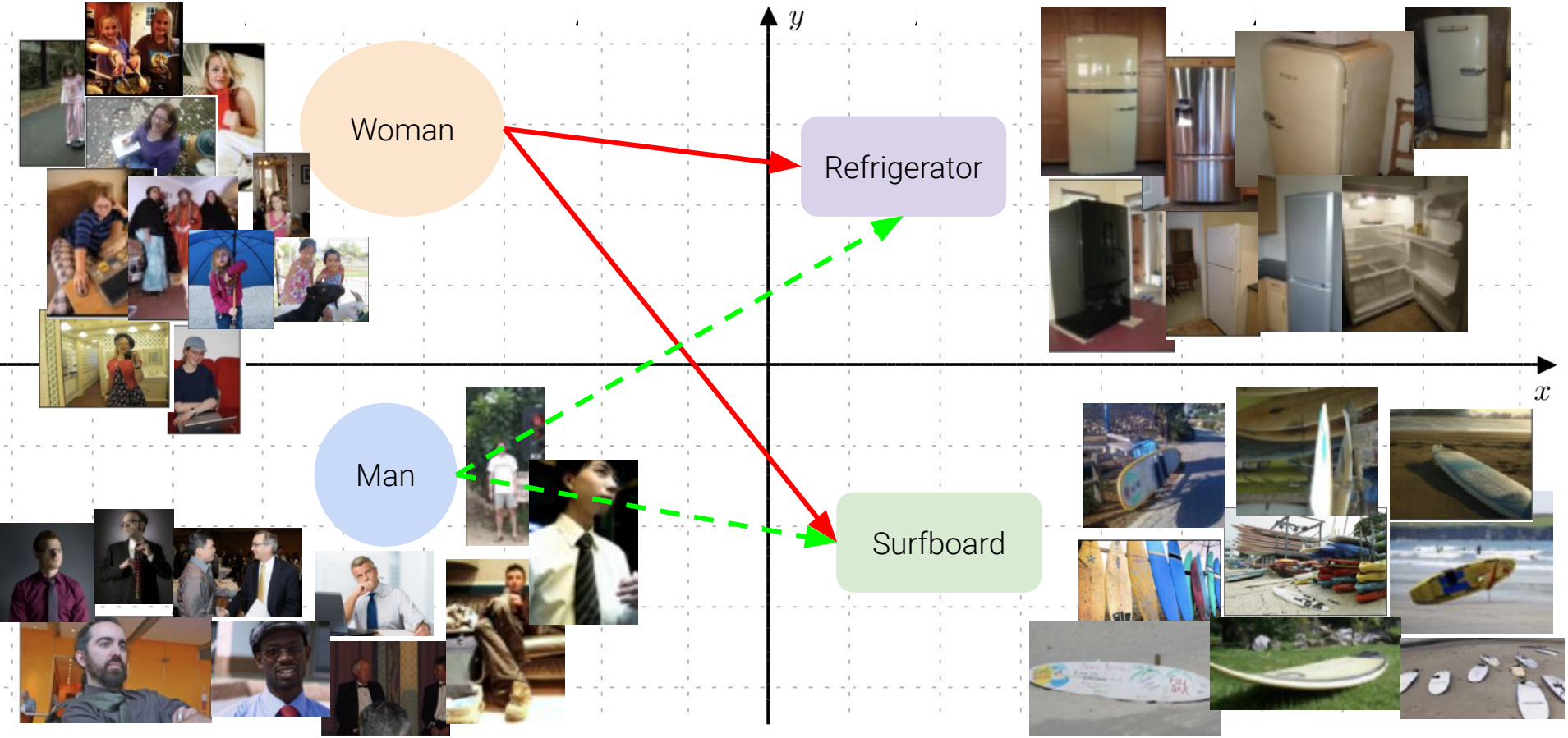
Measuring Biases in Images using Bias Analysis Sets



Measuring Biases in Images using Bias Analysis Sets



Measuring Biases in Images using Bias Analysis Sets



Bias Analysis Sets

COCO 2017 Analysis Set

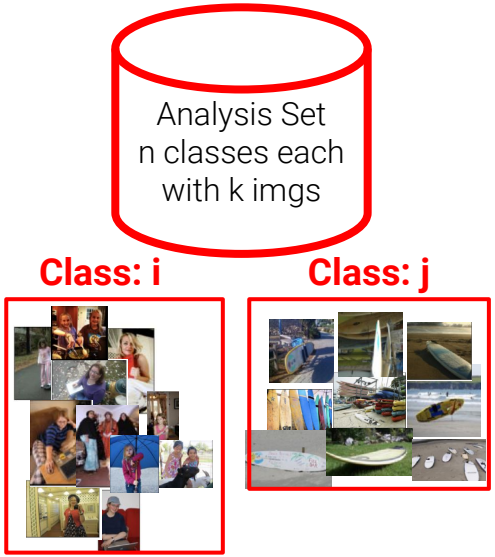
Class	Number of Examples
man	12
woman	15
random	20
stopsign	44
car	12
car+man	9
car+woman	6
refrigerator	18
refrigerator+man	8
refrigerator+woman	9
surfboard	14
surfboard+man	23
surfboard+woman	16

Open Images Analysis Set

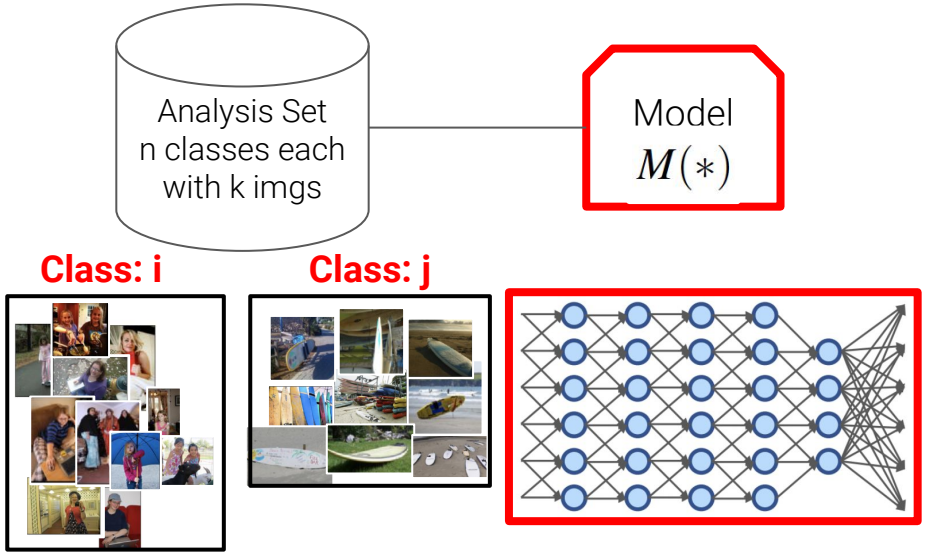
Class	Number of Examples
man	150
woman	150
random	150
stopsign	22
car	150
car+man	150
car+woman	49
sports	150
sports+man	120
sports+woman	32
fashion	150
fashion+man	52
fashion+woman	150
mammal	150
mammal+man	150
mammal+woman	150

Bias Metric Definition

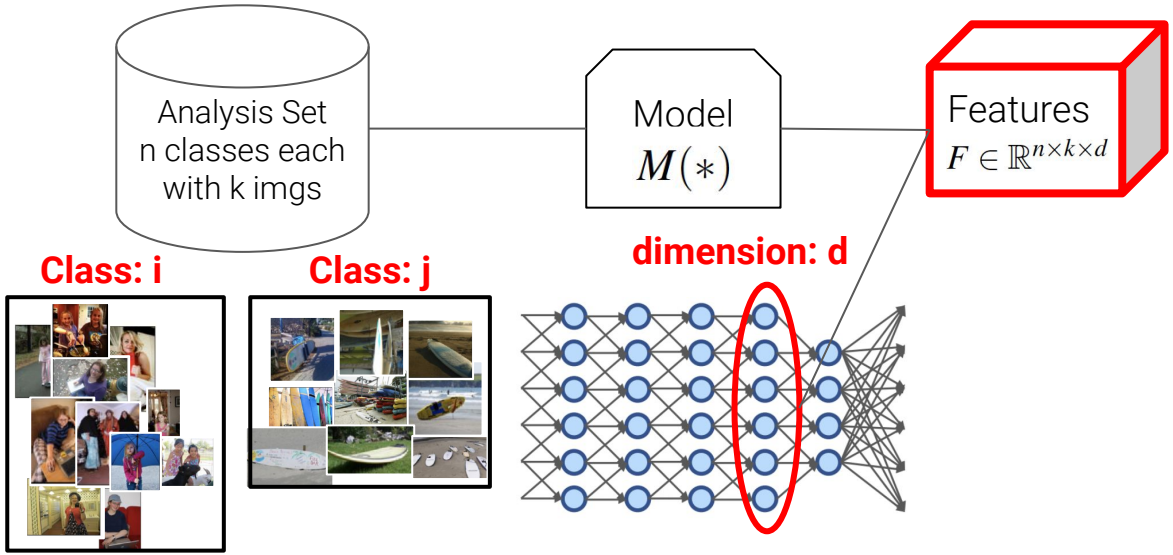
Bias Definition: Quantifying Bias at the Feature Level: Intra-Class



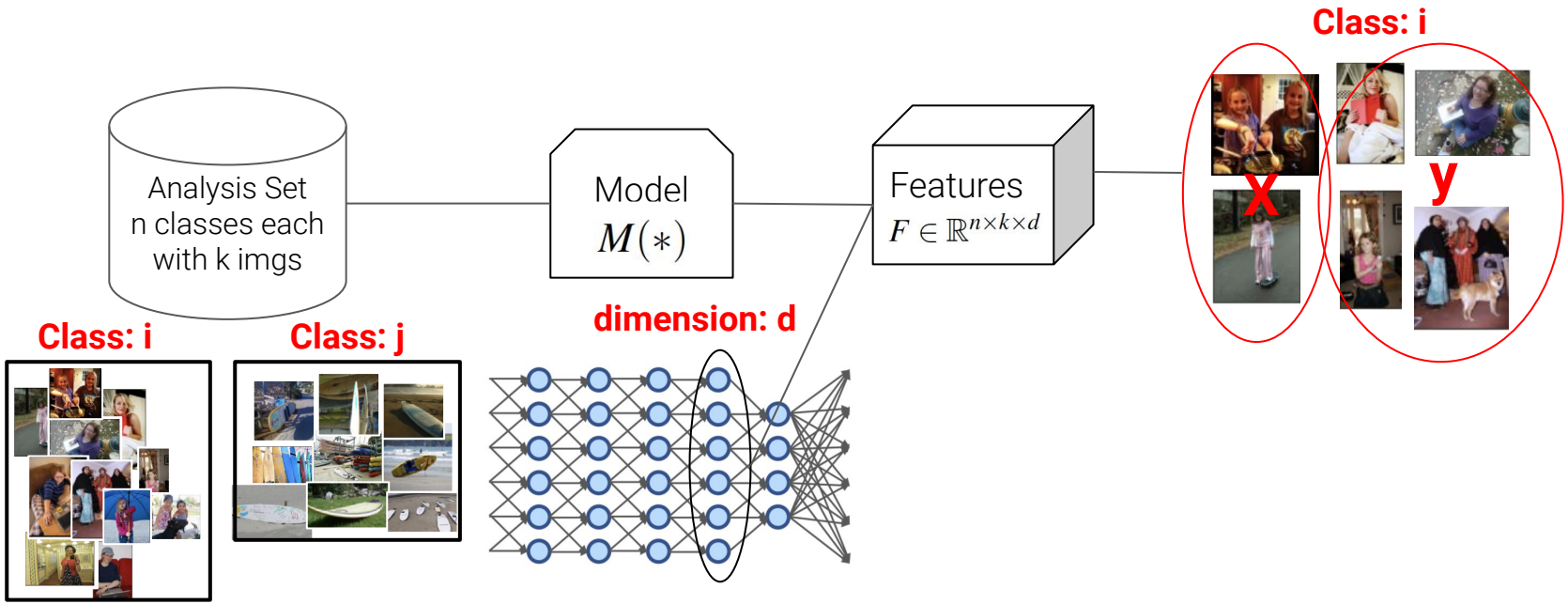
Bias Definition: Quantifying Bias at the Feature Level: Intra-Class



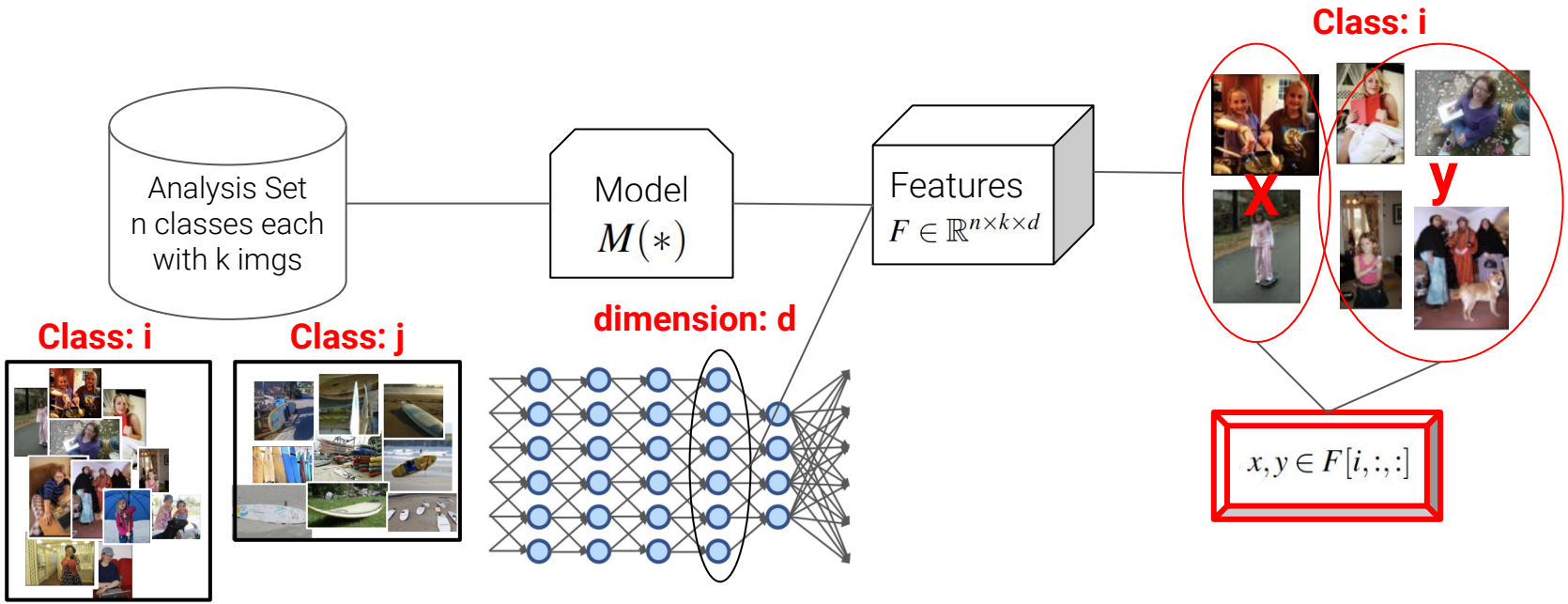
Bias Definition: Quantifying Bias at the Feature Level: Intra-Class



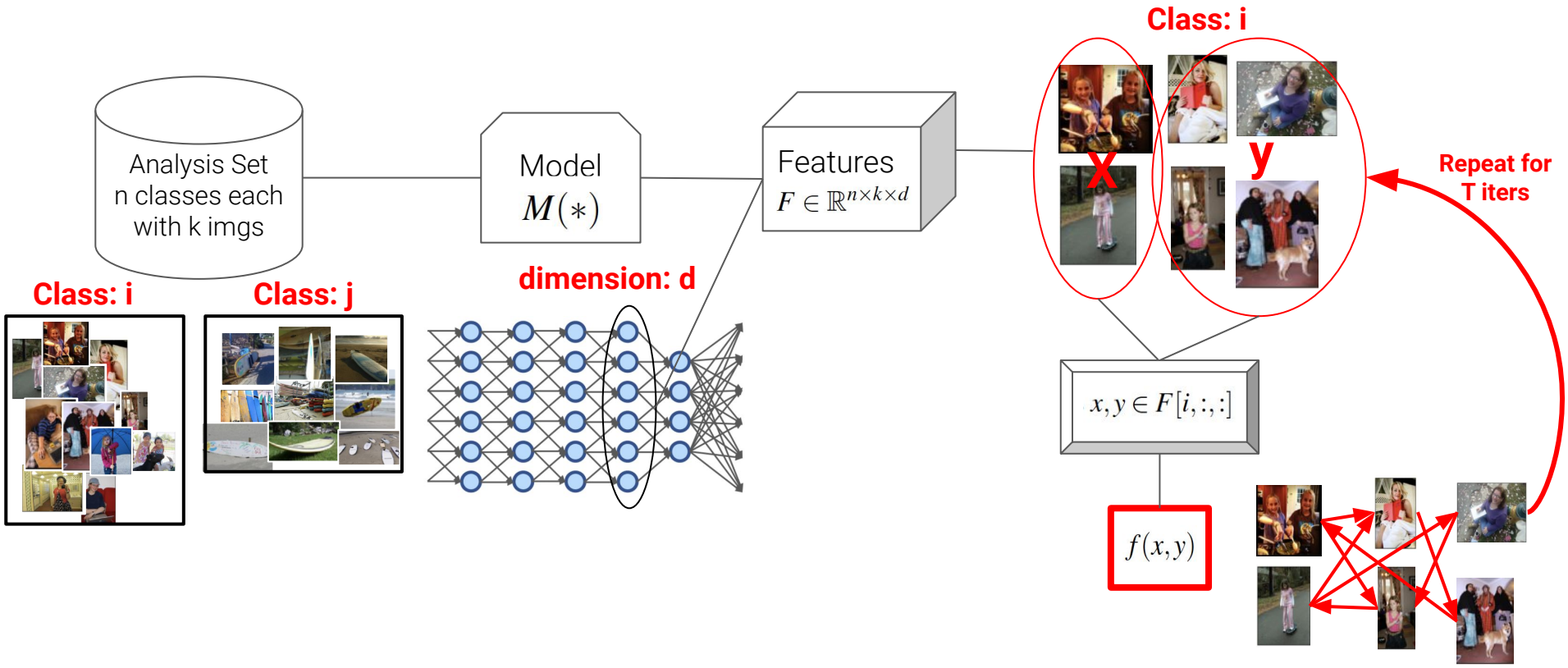
Bias Definition: Quantifying Bias at the Feature Level: Intra-Class



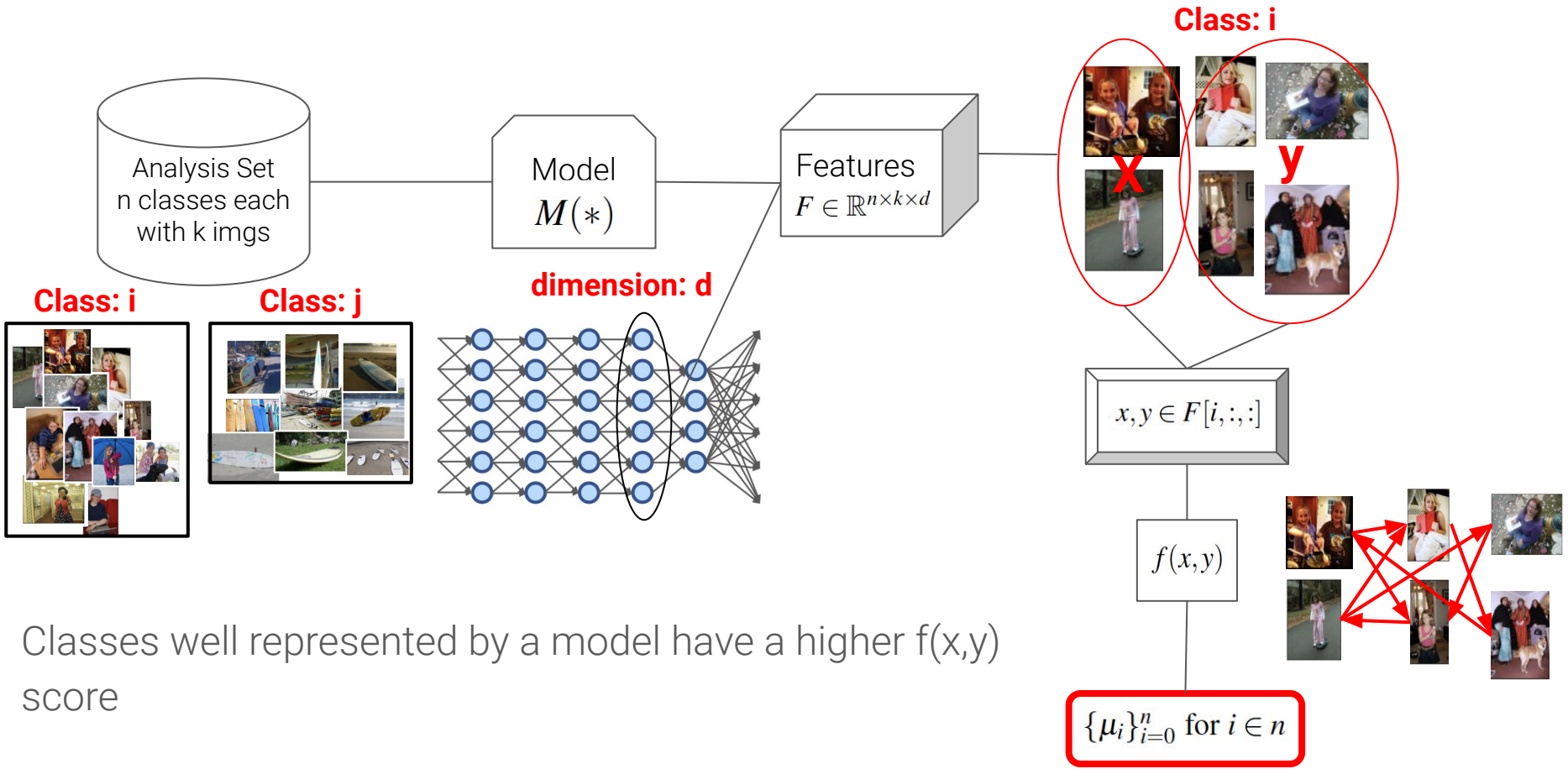
Bias Definition: Quantifying Bias at the Feature Level: Intra-Class



Bias Definition: Quantifying Bias at the Feature Level: Intra-Class

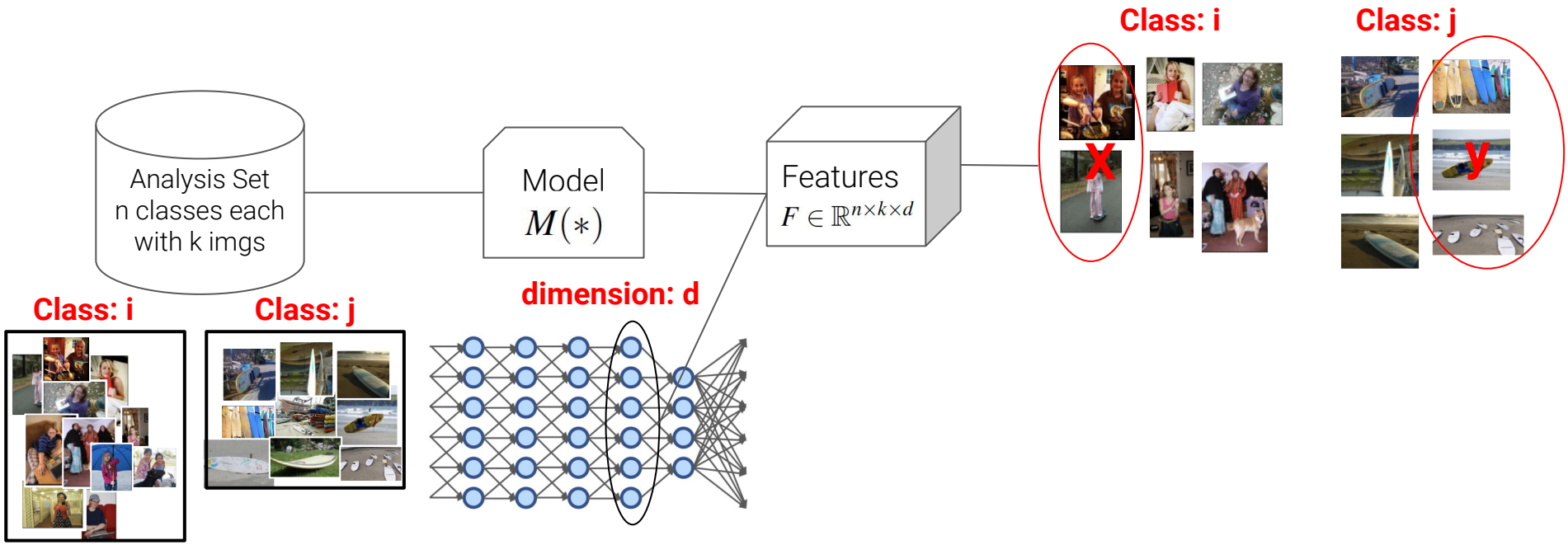


Bias Definition: Quantifying Bias at the Feature Level: Intra-Class

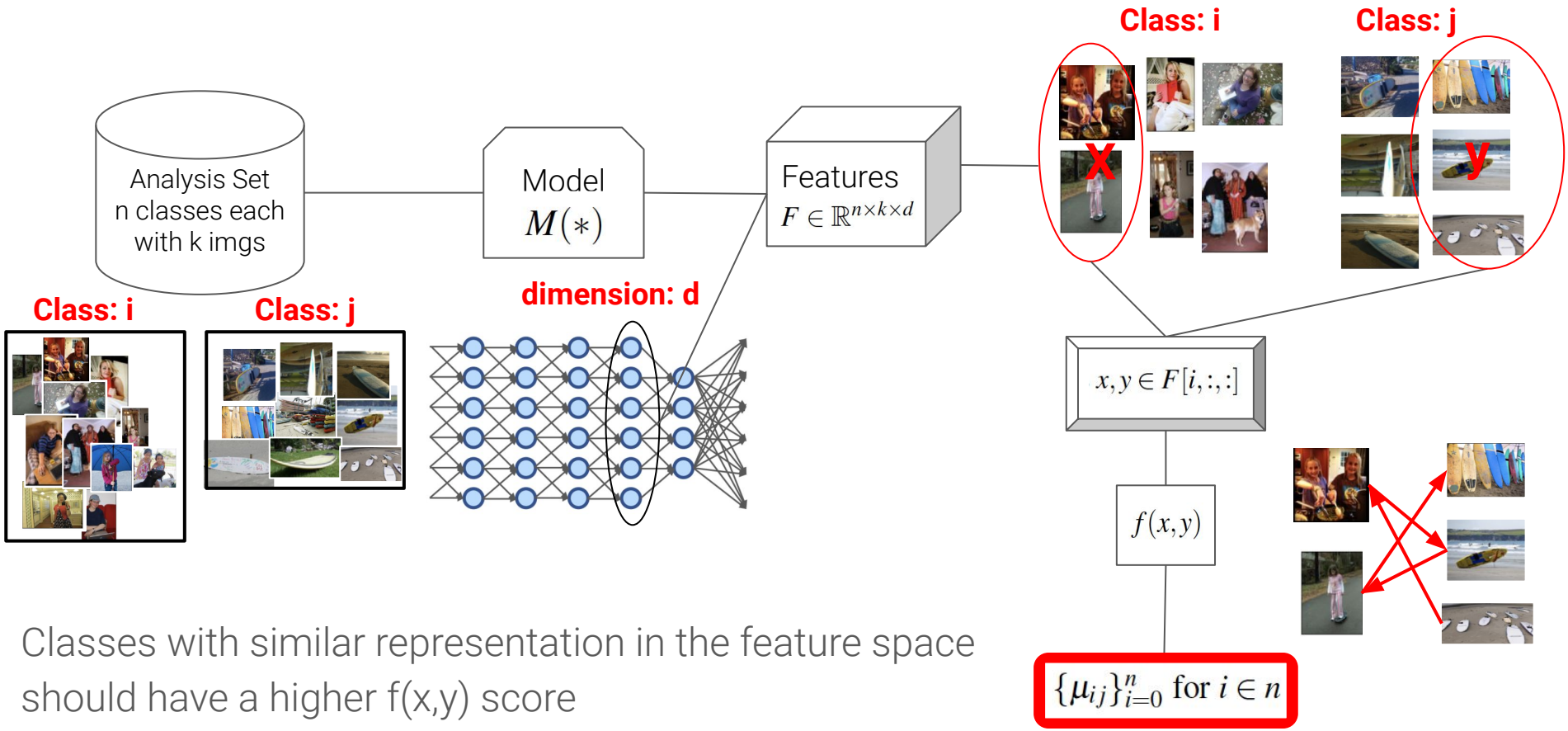


Classes well represented by a model have a higher f(x,y) score

Bias Definition: Quantifying Bias at the Feature Level: Inter-Class



Bias Definition: Quantifying Bias at the Feature Level: Inter-Class



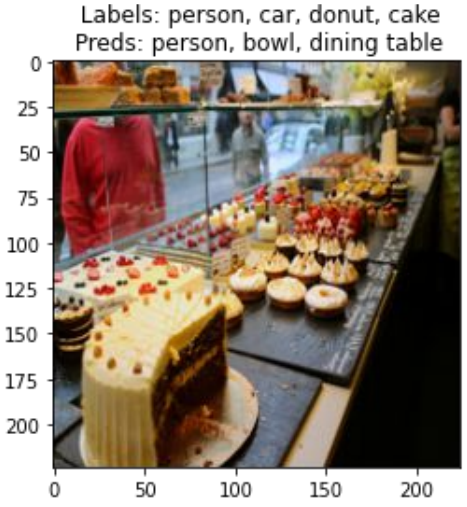
Classes with similar representation in the feature space should have a higher f(x,y) score

Experimental Setup

Finetuning on COCO 2017: Results (Multilabel Classification)

Model	Pretraining Dataset	Pretraining Setting	Epochs	Learning Rate	Optimizer	mAP	Micro F1
BiT-M-R50x1	ImageNet-21k (20M)	Supervised	15	0.003	SGD, m: 0.9	0.7527	0.827
ResNet50	ImageNet-1k (1M)	Supervised	15	0.001	SGD, m: 0.9	0.7024	0.8363
ResNet18	ImageNet-1k (1M)	Supervised	40	0.1: reduce on plateau	SGD, m: 0.9, wd:1e-5	0.7443	0.7307
CLIP: ViT-B/32	400M images from web	Supervised	20	0.001	SGD, m: 0.9	0.7053	0.7929
MoCo ResNet50	ImageNet-1k (1M)	Self Supervised	20	0.1: reduce on plateau	SGD, m: 0.9, wd:1e-5	0.6268	0.6460

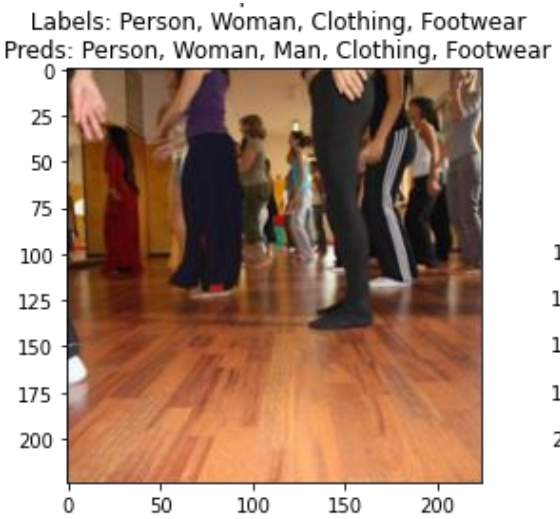
MoCo Finetuning

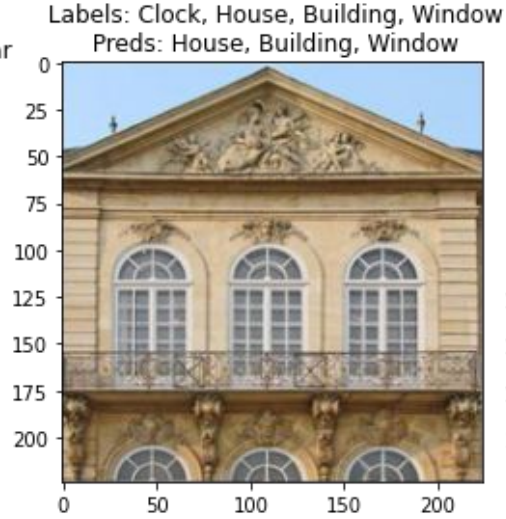


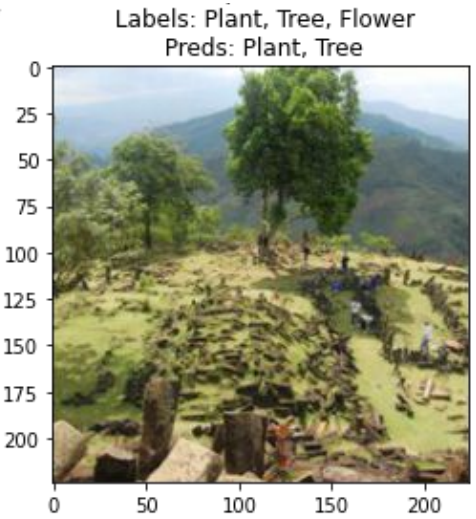
Finetuning on Open Images v4: Results (Multilabel Classification)

Model	Pretraining Dataset	Pretraining Setting	Epochs	Learning Rate	Optimizer	Micro F1
BiT-M-R50x1	ImageNet-21k (20M)	Supervised	15	0.1:reduce on plateau	SGD, m: 0.9, wd:1e-5	0.3237
ResNet50	ImageNet-1k (1M)	Supervised	15	0.1:reduce on plateau	SGD, m: 0.9, wd:1e-5	0.4039
ResNet18	ImageNet-1k (1M)	Supervised	15	0.1: reduce on plateau	SGD, m: 0.9, wd:1e-5	0.338
CLIP: ViT-B/32	400M images from web	Supervised	20	0.001	SGD, m: 0.9, wd:1e-5	0.3139
MoCo ResNet50	ImageNet-1k (1M)	Self Supervised	10	0.1: reduce on plateau	SGD, m: 0.9, wd:1e-5	0.3132

ResNet50 Finetuning

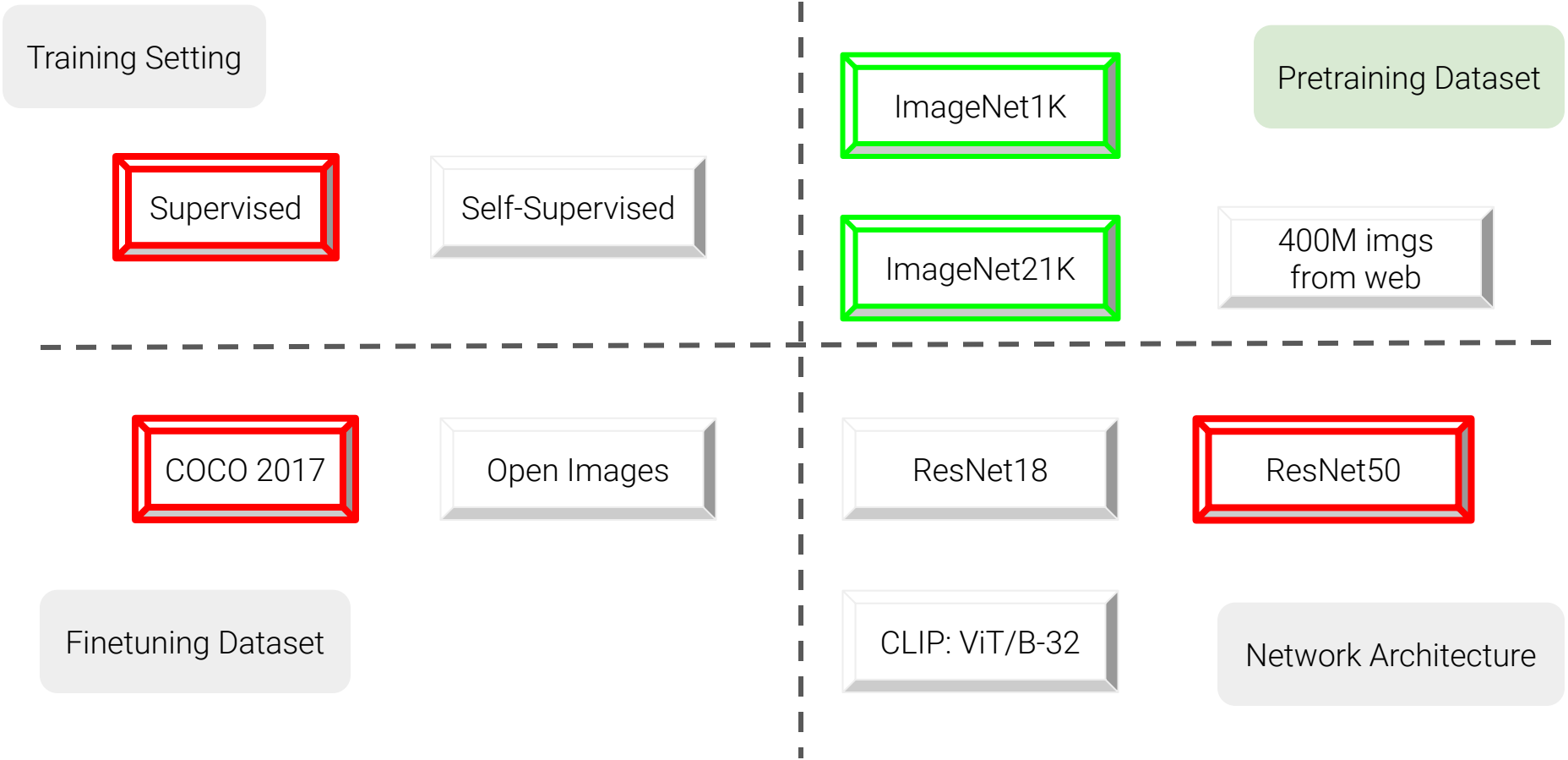




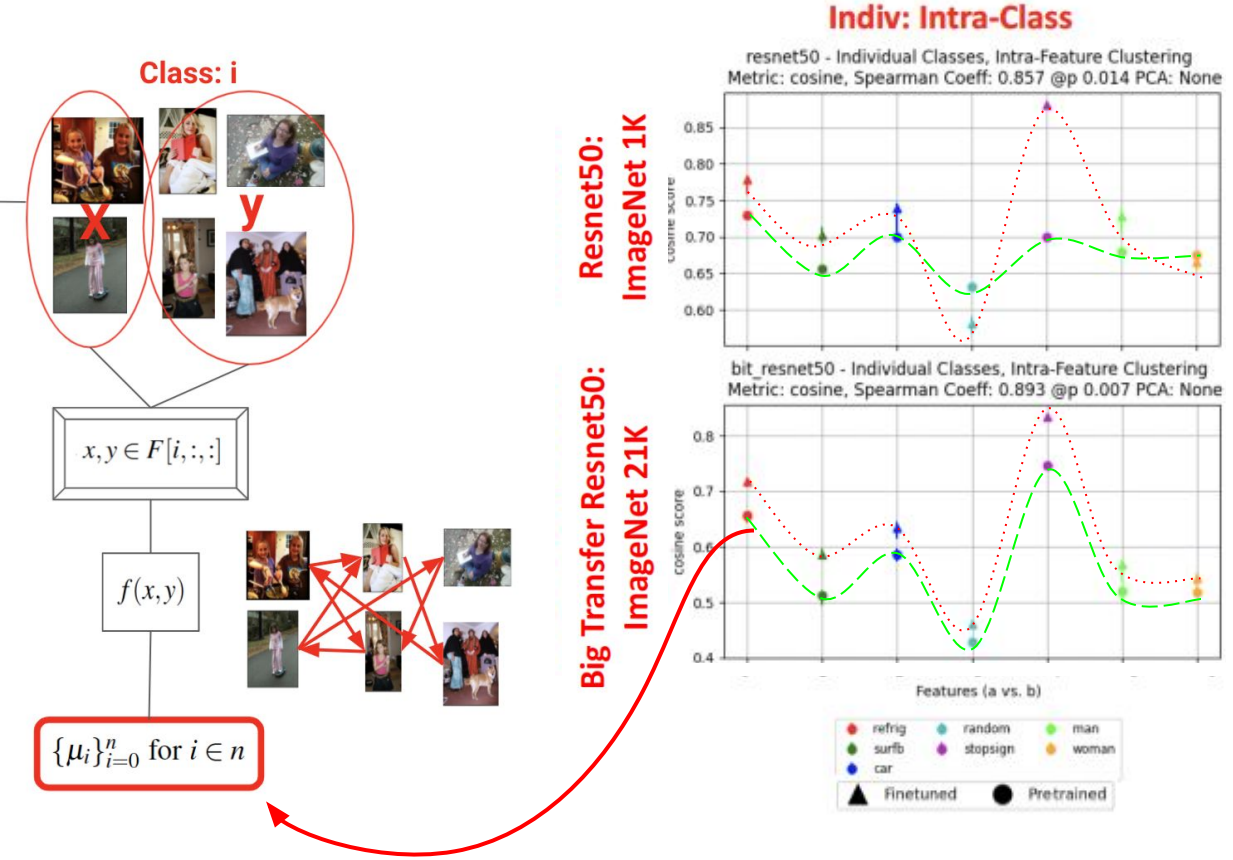


Results: Pretraining Dataset (COCO)

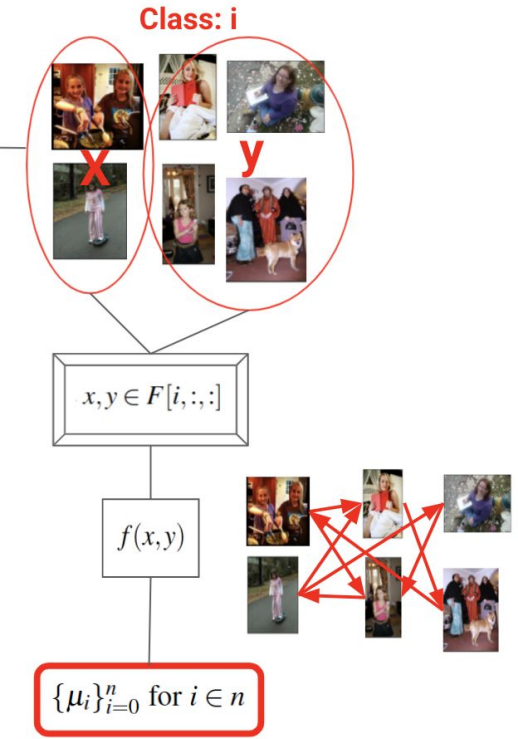
Pretraining Dataset



Big Transfer ResNet50 and ResNet50 - Intra-Class

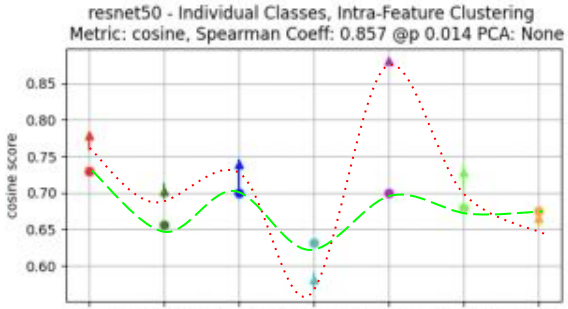


Big Transfer ResNet50 and ResNet50 - Intra-Class

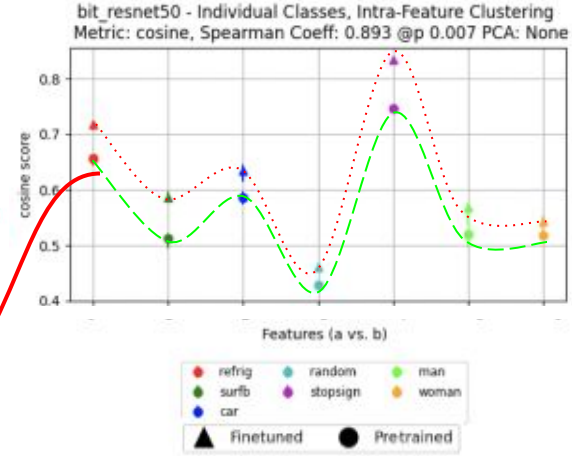


Indiv: Intra-Class

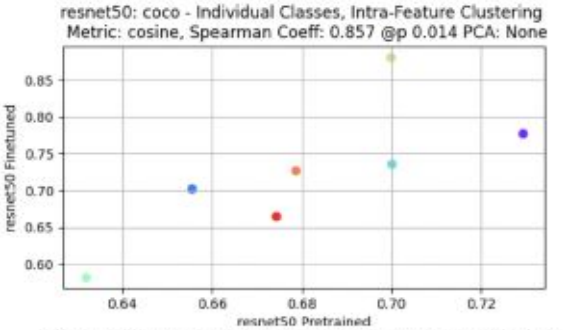
Resnet50: ImageNet 1K

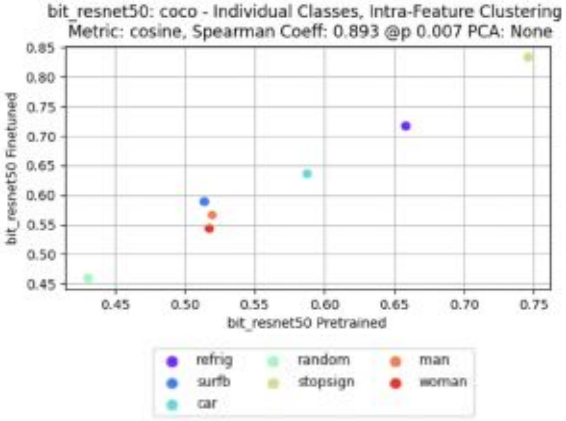


Big Transfer Resnet50: ImageNet 21K



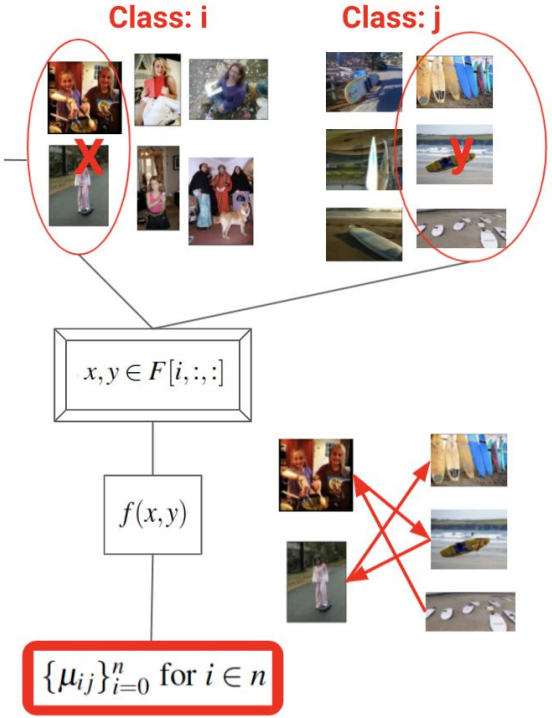
Indiv: Intra-Class





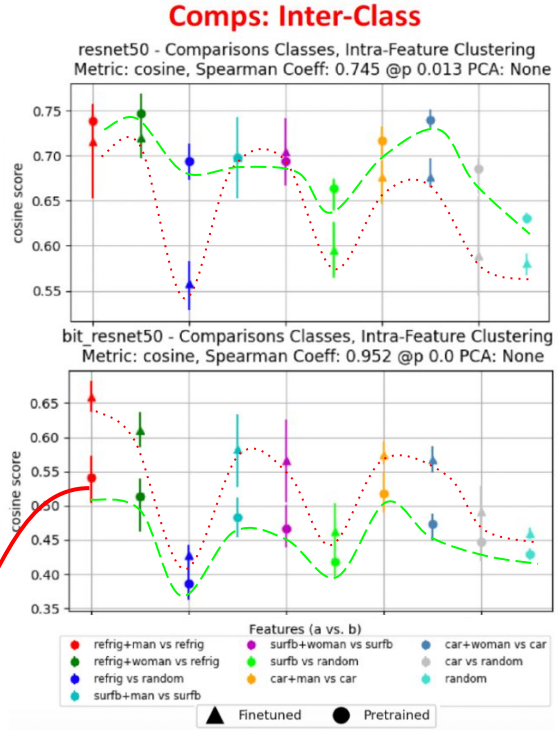
For intra-class similarity, the different pretraining datasets still resulted in similar biases in both the pretraining and finetuning stage. Biases not impacted by finetuning on COCO

Big Transfer ResNet50 and ResNet50 - Inter-Class

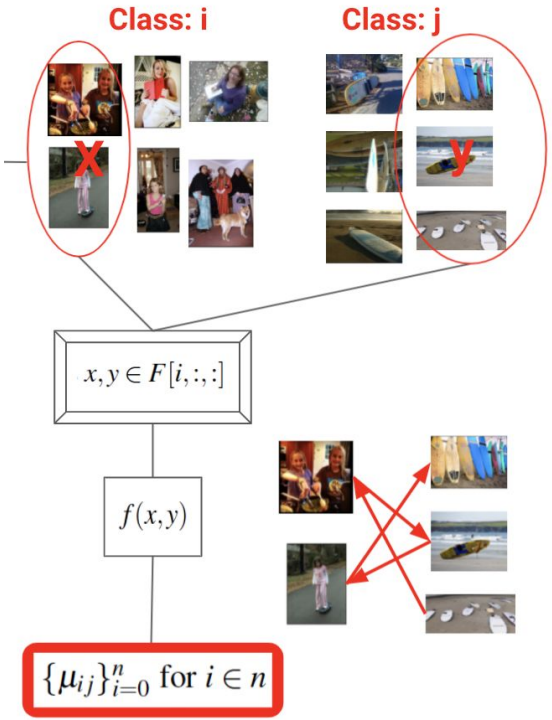


Resnet50: ImageNet 1K

Big Transfer Resnet50: ImageNet 21K

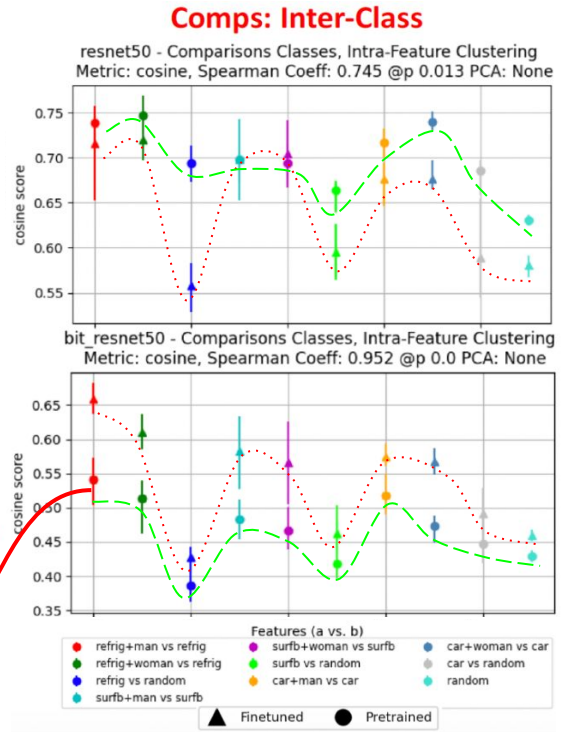


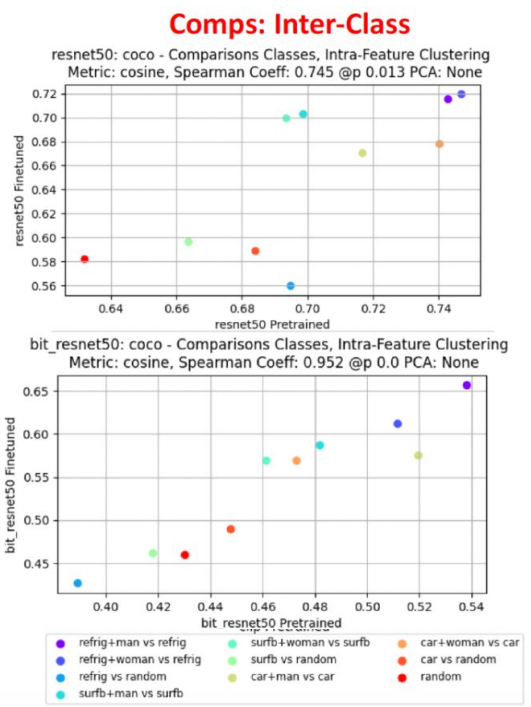
Big Transfer ResNet50 and ResNet50 - Inter-Class



Resnet50: ImageNet 1K

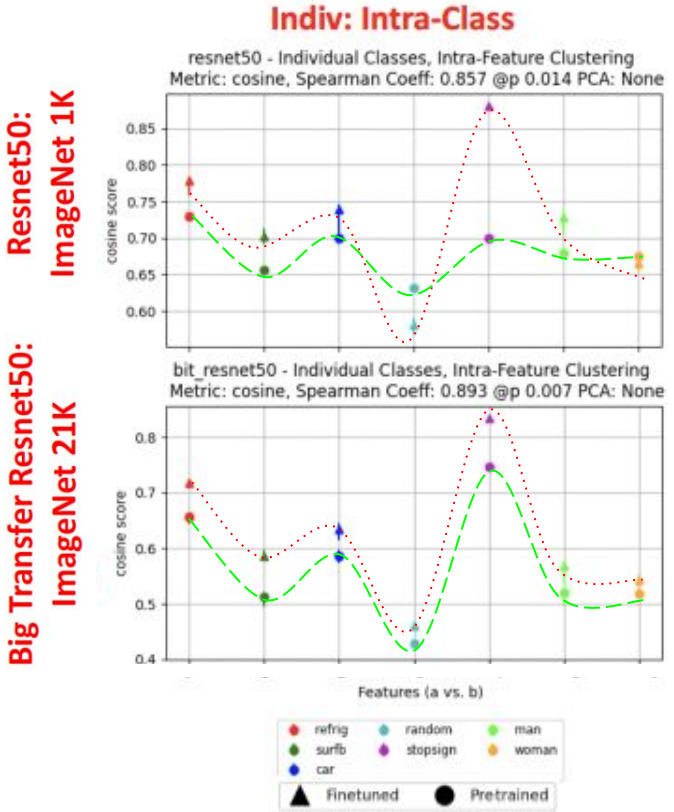
Big Transfer Resnet50: ImageNet 21K

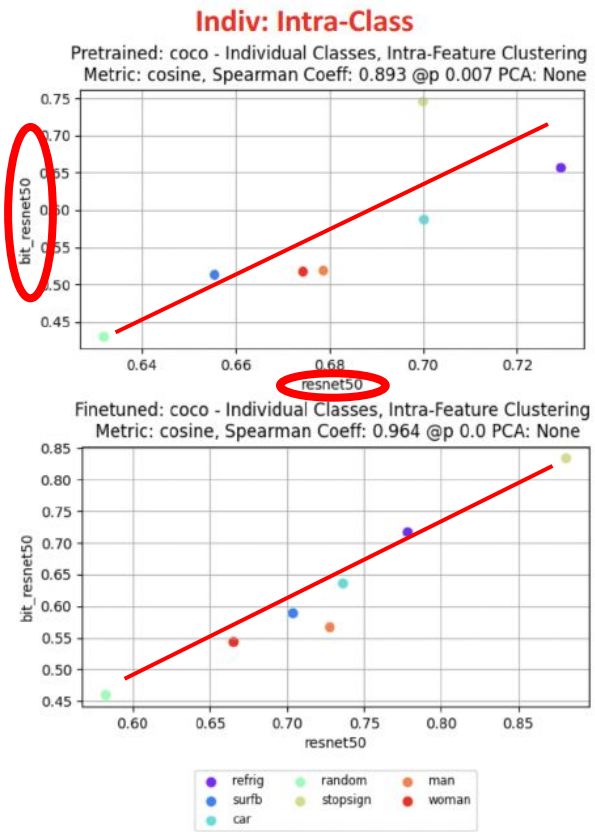




For inter-class similarity, the different pretraining datasets resulted in different biases across ResNet50 in the pretraining stage. Finetuning impacted biases for models trained on ImageNet1K more than ImageNet21K

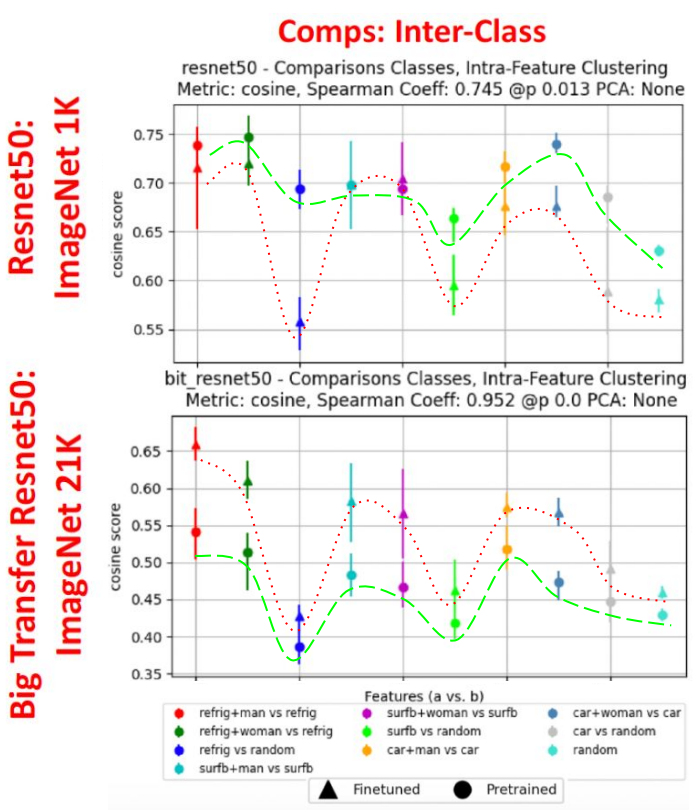
Big Transfer ResNet50 vs. ResNet50

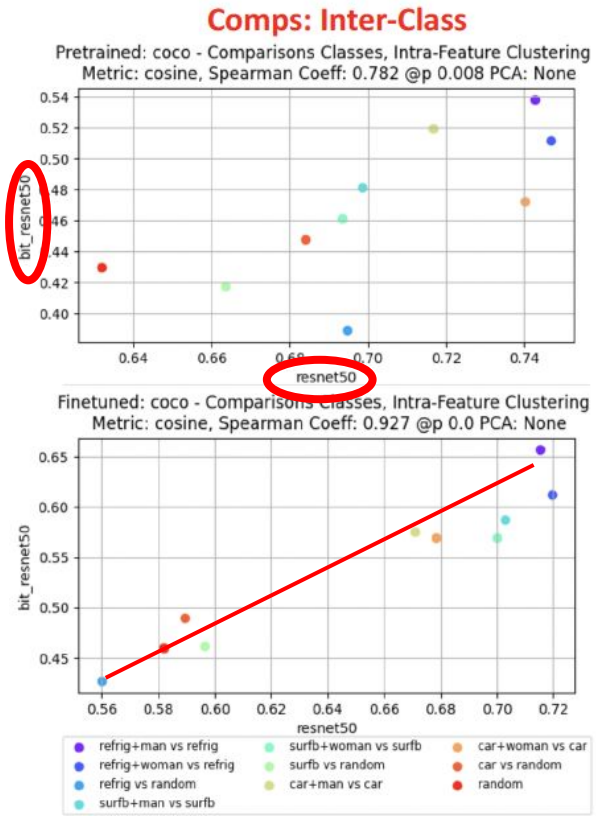




Comparing across Big Transfer and ResNet50, they encoded biases similarly in the pretraining and finetuning stage implying that the different pretraining datasets did not impact the way these biases are represented in the feature space

Big Transfer ResNet50 vs. ResNet50





Pretraining Stage

Finetuning Stage

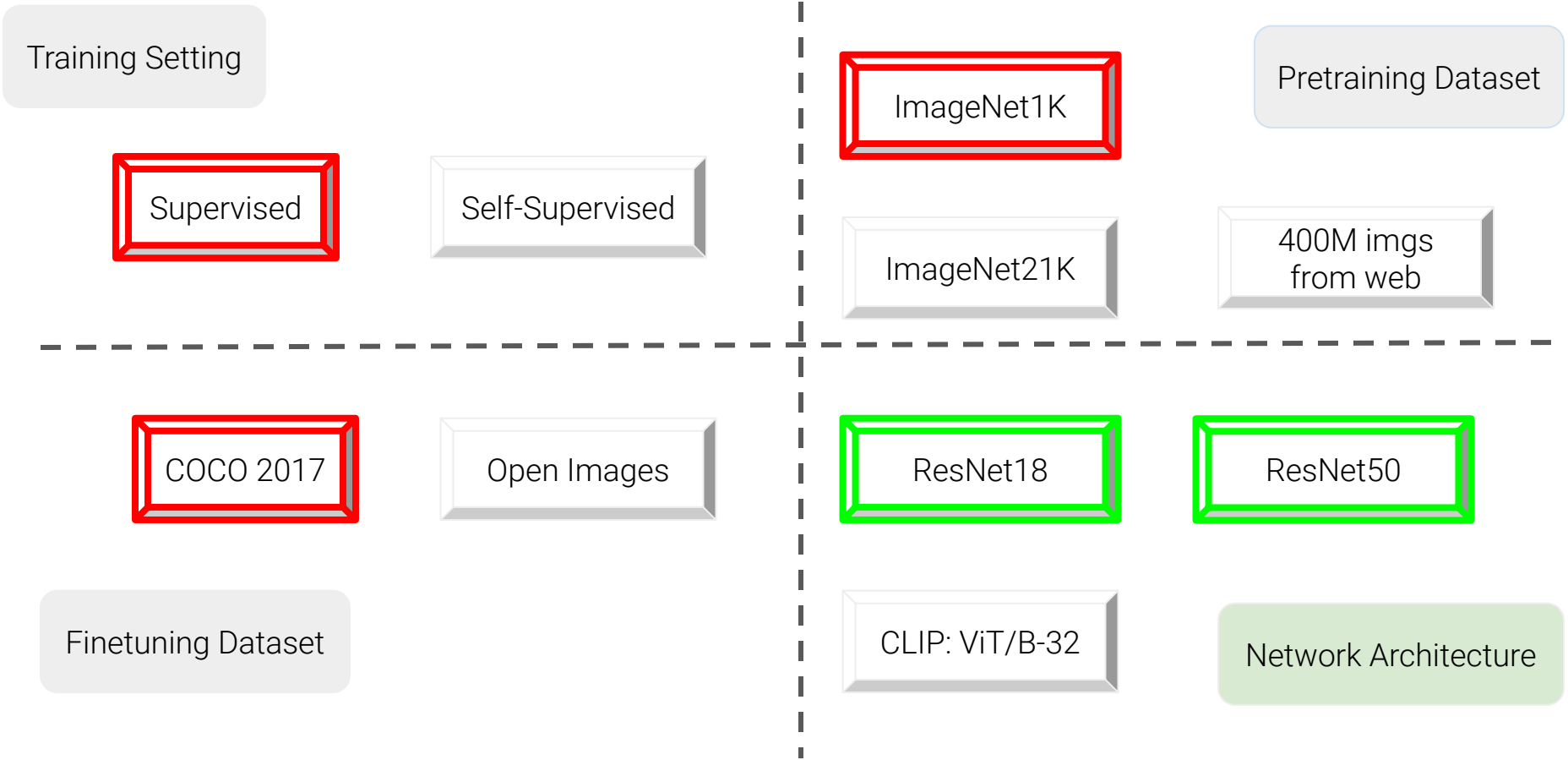
ResNet50 encoded the same biases in the pretraining and finetuning stage for intra-class similarity, but this was not the case for Inter-Class similarity where the biases in the pretraining stage were different across ResNets trained on different datasets

Takeaways

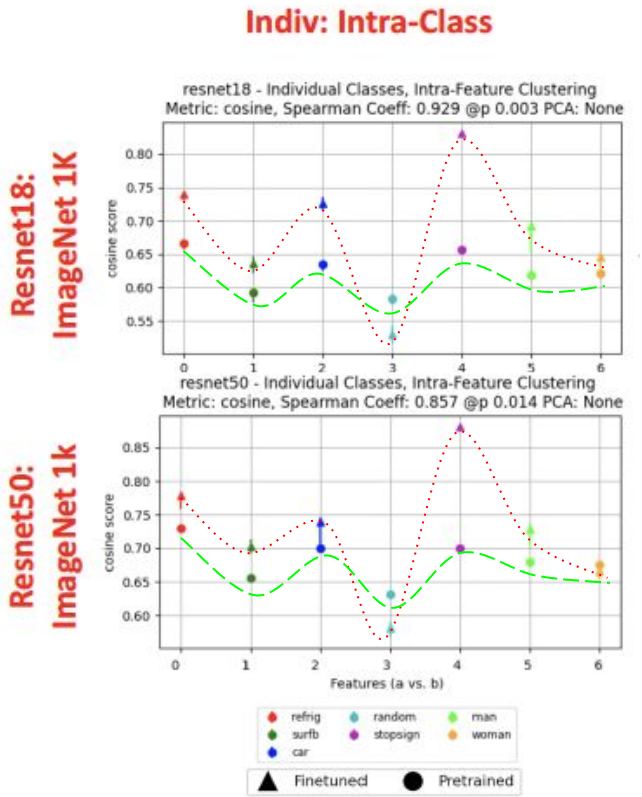
- For intra-class similarity, ResNet50s encode similar biases regardless of being trained on ImageNet1K or ImageNet21K
- For inter-class similarity, ResNet50s encode different biases in the pretraining stage but similar biases in the finetuning stage
    - Biases learned on the finetuning stage can be attributed to the dataset that the model is finetuned on
- Pretraining dataset (specifically ImageNet1K vs ImageNet21K) impacts biases in the pretraining stage for inter-class comparisons but after finetuning the model, the biases from the finetuning dataset impact the pretraining biases

Results: Network Architecture - COCO

Network Architecture

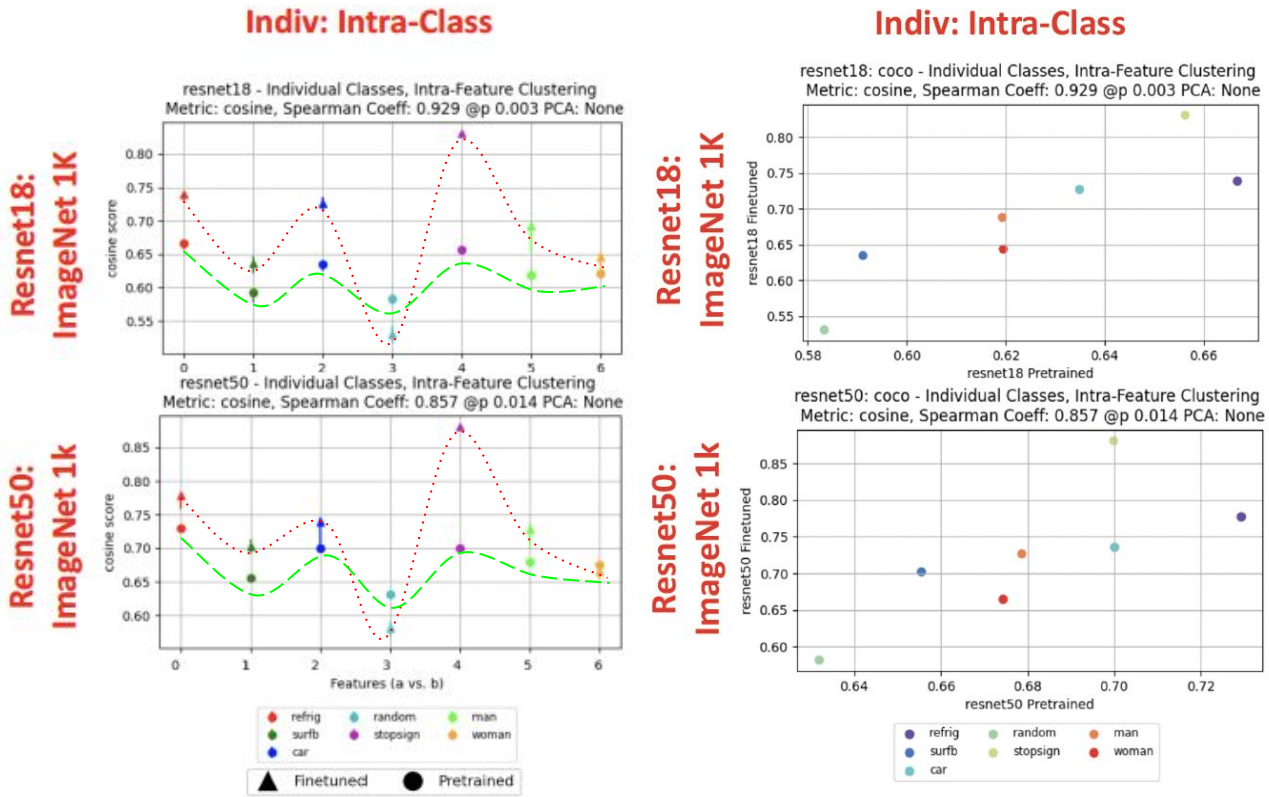


ResNet18 and Resnet50 - Intra Class



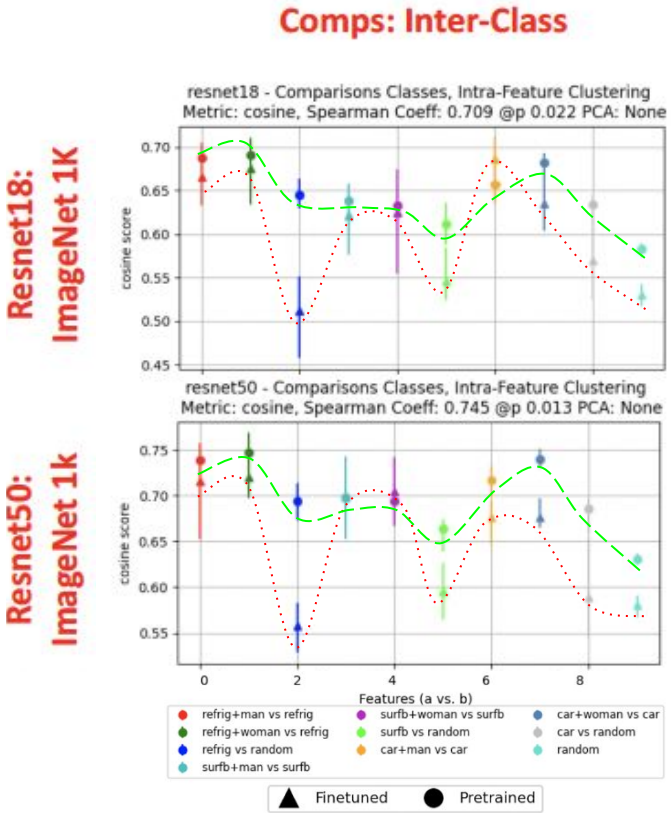
ResNet18 and ResNet50 encoded similar biases in the pretraining and finetuning space for intra-class comparisons, but this was not the case for inter-class comparisons where finetuning introduced new biases

ResNet18 and Resnet50 - Intra Class



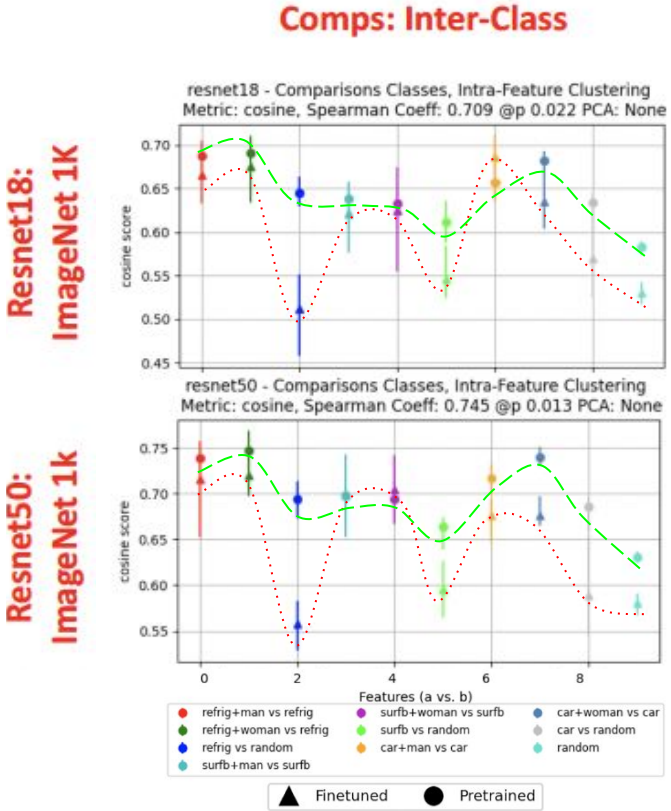
ResNet18 and ResNet50 encoded similar biases in the pretraining and finetuning space for intra-class comparisons, but this was not the case for inter-class comparisons where finetuning introduced new biases

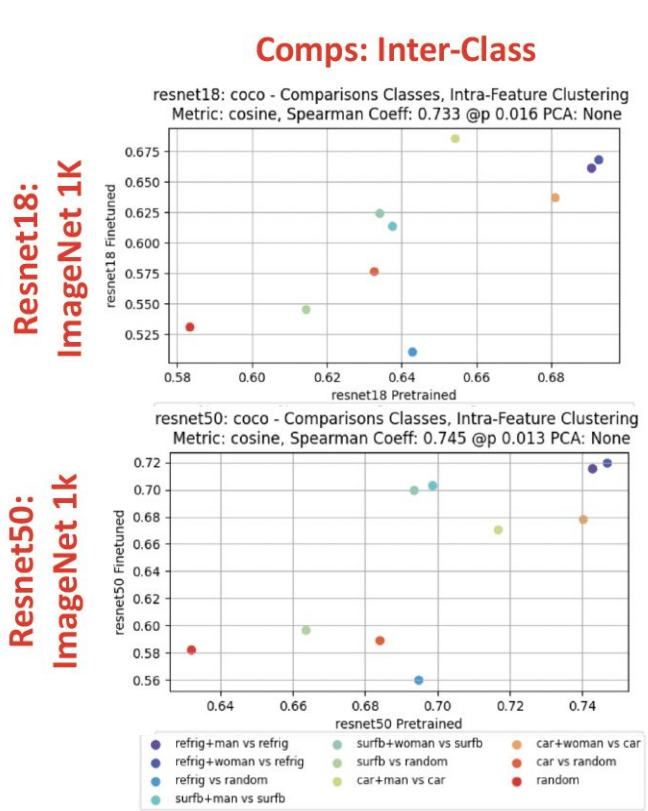
ResNet18 and Resnet50 - Inter-Class



ResNet18 and ResNet50 encoded similar biases in the pretraining and finetuning space for intra-class comparisons, but this was not the case for inter-class comparisons where finetuning introduced new biases

ResNet18 and Resnet50 - Inter-Class





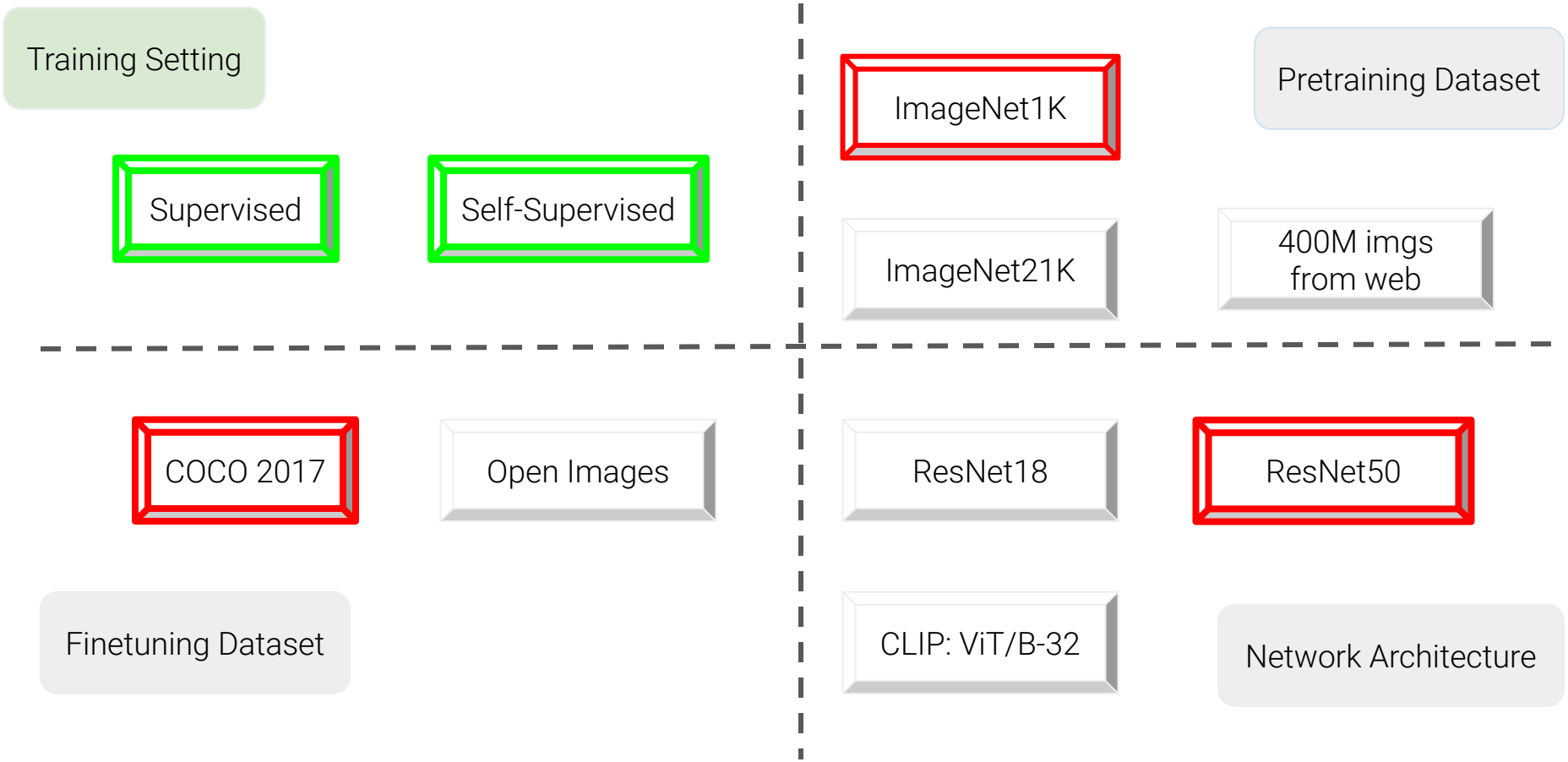
ResNet18 and ResNet50 encoded similar biases in the pretraining and finetuning space for intra-class comparisons, but this was not the case for inter-class comparisons where finetuning introduced new biases

Takeaways

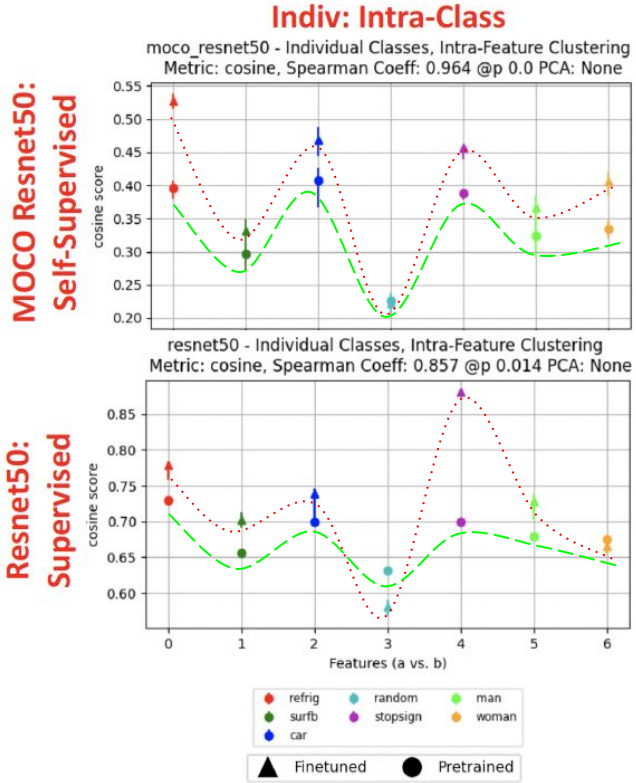
- The network architecture most influenced biases for inter-class comparisons
- Finetuning introduced new biases for inter-class comparisons
  - These new biases could be attributed to the dataset that the model was finetuned on
- The biases after finetuning for inter-class comparisons were different across ResNet18 and ResNet50 → ResNet18 is more susceptible to change in biases in the finetuning stage
- ResNet50 and ResNet18 encoded similar biases for intra-class comparisons implying that the change in network architecture did not impact biases for the ResNet models
  - Instead, the dataset the models were pretrained on, or the training setting might have a greater impact on the biases

Results: Training Setting - COCO

Training Setting

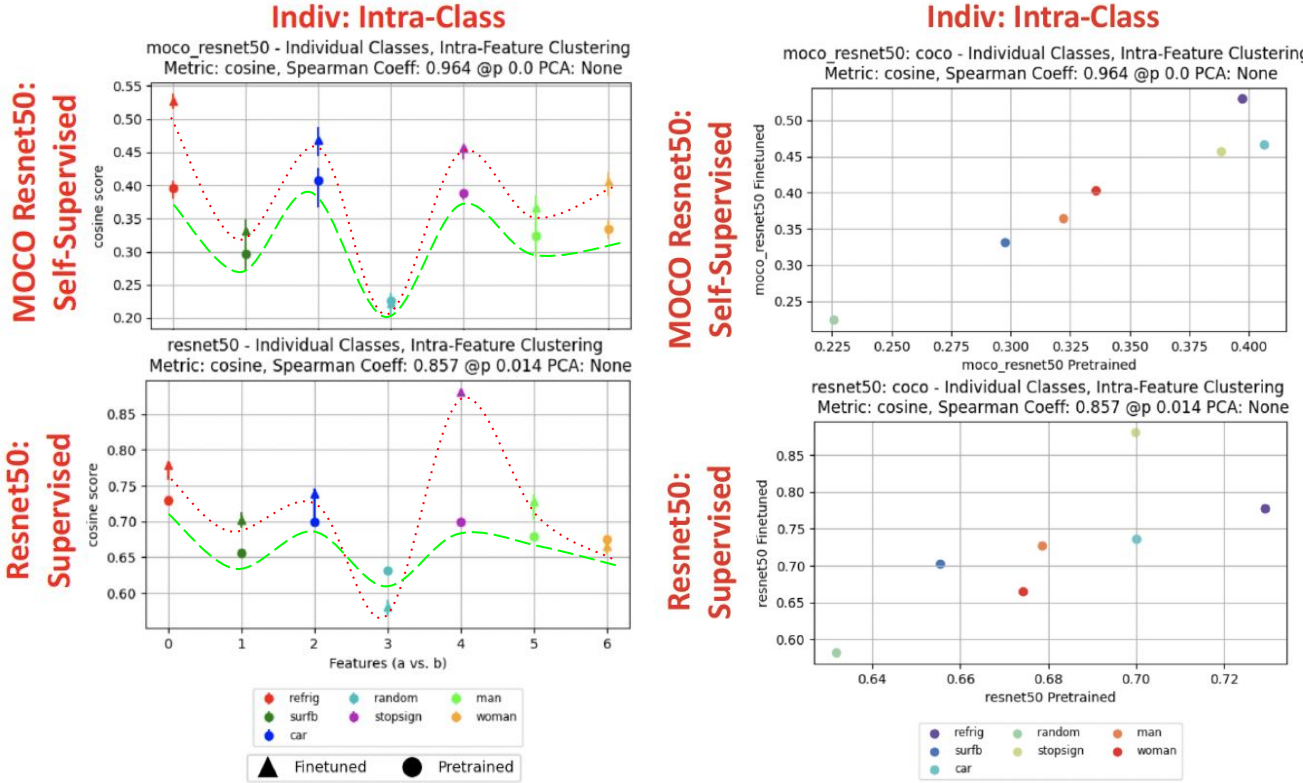


MoCo ResNet50 and ResNet50 - Intra-Class



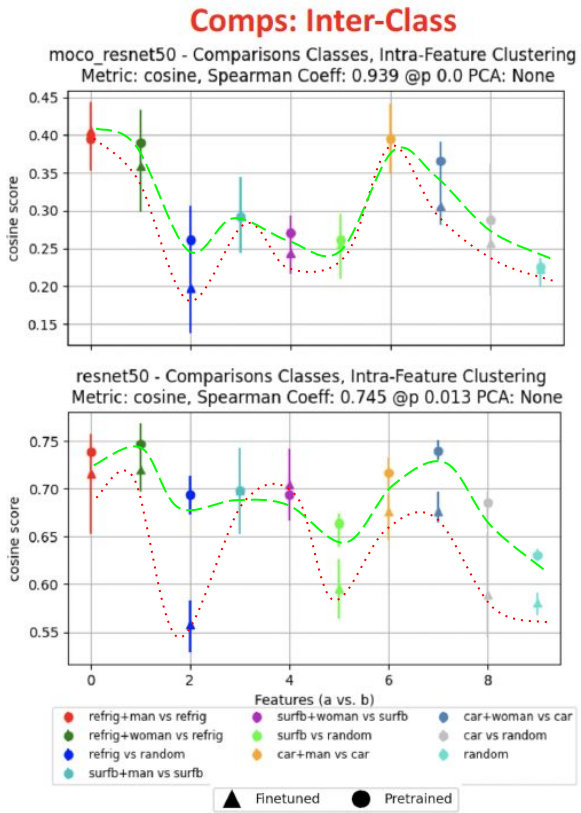
MoCo better preserved it's biases before and after finetuning whereas ResNet50 was more susceptible to changes in its biases after finetuning

MoCo ResNet50 and ResNet50 - Intra-Class



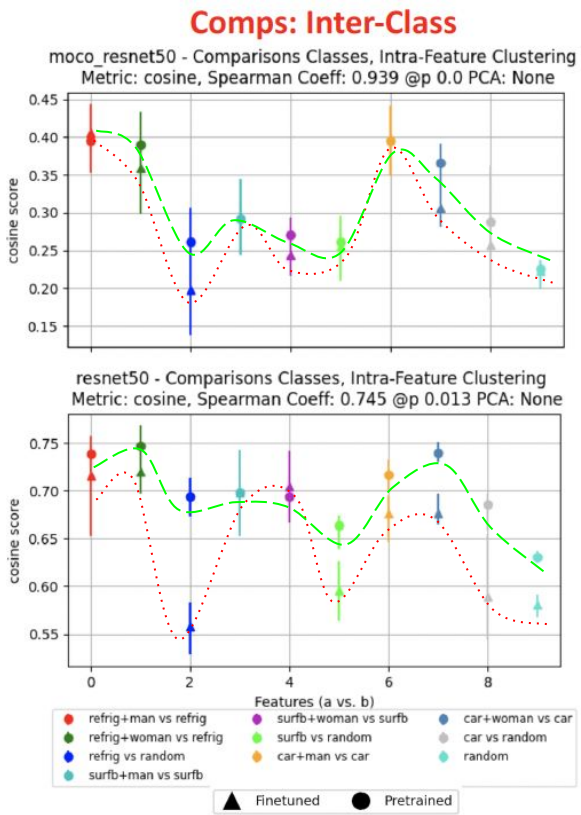
MoCo better preserved it's biases before and after finetuning whereas ResNet50 was more susceptible to changes in its biases after finetuning

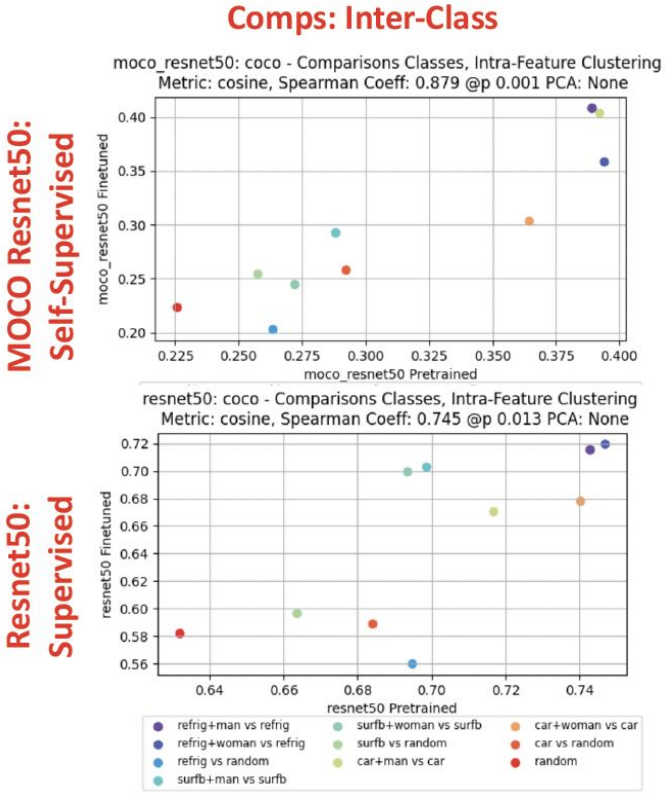
MoCo ResNet50 and ResNet50 - Inter-Class



MoCo better preserved it's biases before and after finetuning whereas ResNet50 was more susceptible to changes in its biases after finetuning

MoCo ResNet50 and ResNet50 - Inter-Class





MoCo better preserved it's biases before and after finetuning whereas ResNet50 was more susceptible to changes in its biases after finetuning

Takeaways

- ResNet50 was more susceptible to changes in its biases after finetuning than MoCo ResNet50 implying that the self-supervised setting more strongly preserves its biases even after finetuning
- ResNet50 and MoCo ResNet50 have similar biases in the pretraining stage implying that the training setting does not have an impact on the biases in this stage
- Training setting has an impact on the biases after the model has been finetuned since the biases across ResNet50 and MoCo ResNet50 were different after finetuning

Conclusions

- Analyze gender bias at the class level using implicit feature representations of models
- Analysis sets containing images with objects that co-occur with gender
- Identify factors that contribute to gender bias
  - Pretraining dataset
  - Network architecture
  - Training setting
  - Finetuning dataset
- Qualitative comparison across models for biases

Future Directions

- Larger, more targeted analysis sets with less noise that are scalable
- Experiment with multi-dimensional metrics such as mahalanobis distance and distance correlation to better capture non-linearity of latent space
- Transformations such as PCA to better capture and represent the variation in our dataset
- Expand experimentation with more models, and more controlled variables
- Formal verification pipeline
- “Who” vs “How” genders are being represented in the latent space
  - Qualitative analysis of analysis sets
- Explore beyond two genders

Acknowledgements

Thank you to my advisors, professors and fellow lab mates

Prof. Vicente Ordóñez

Prof. Baishakhi Ray

Tianlu Wang

Thank You