CSDS 452 Causality and Machine Learning

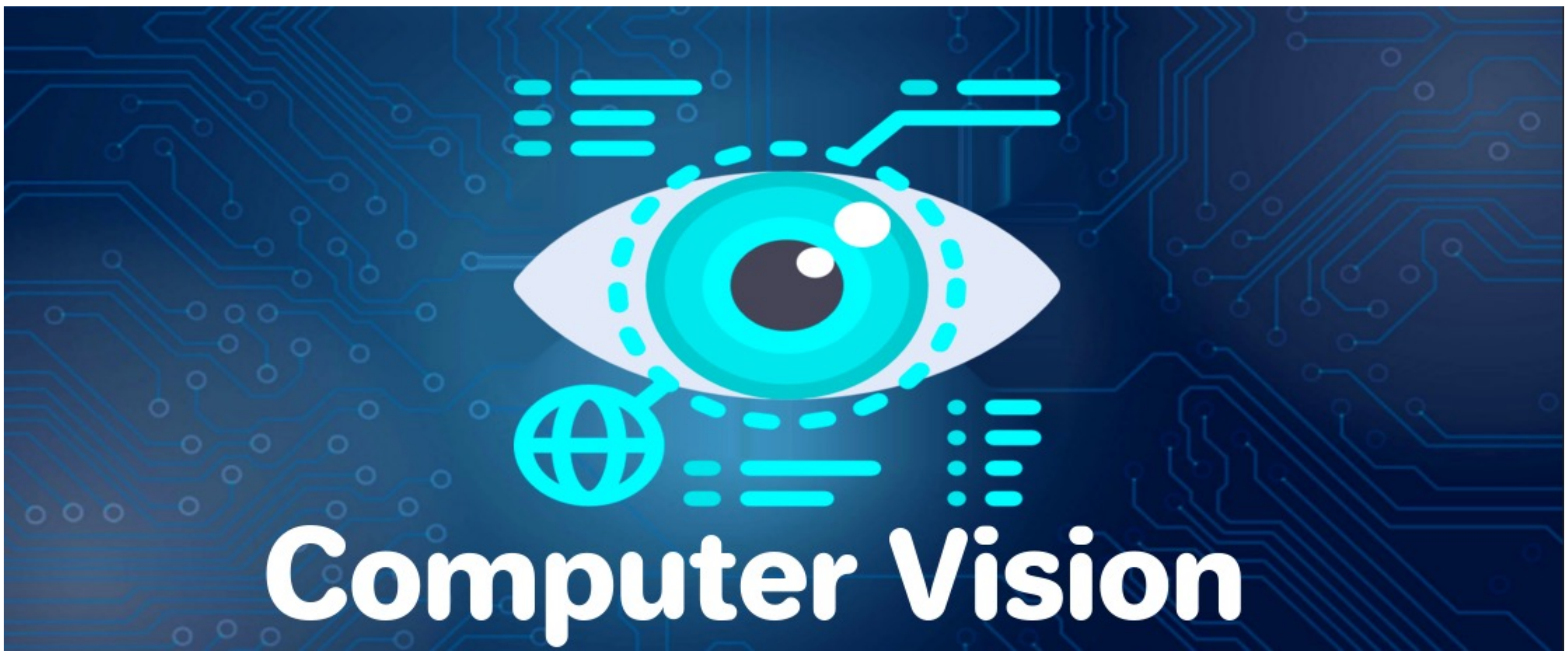
Lecture 20: Causal Computer Vision

Instructor: Jing Ma

Fall 2024, CDS@CWRU

Computer Vision

- “Vision is the act of knowing what is where by looking.” --Aristotle



General tasks in CV

- Image classification
- Object detection
- Pose estimation
- Image segmentation
- …

Challenges

- Biased data distribution
- Limited annotation
- …

Outline

- Momentum Causal Effect
- Interventional few-shot learning

Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect

Kaihua Tang[1], Jianqiang Huang[1,2], Hanwang Zhang[1]

NeurIPS 2020

[1]Nanyang Technological University

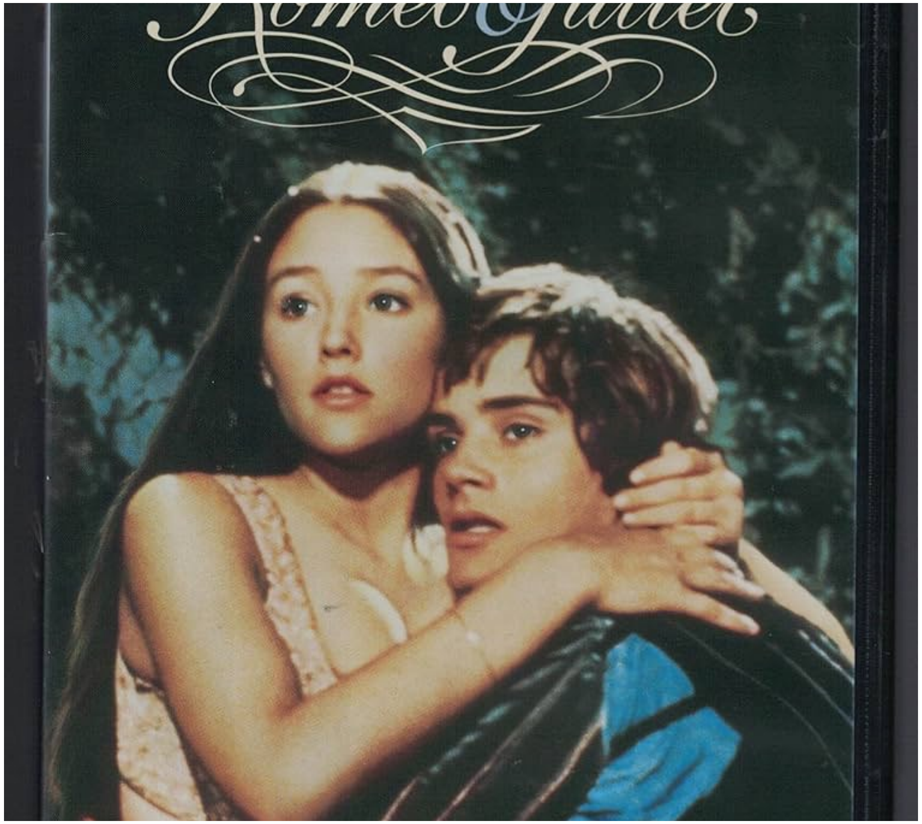
[2]Damo Academy, Alibaba Group

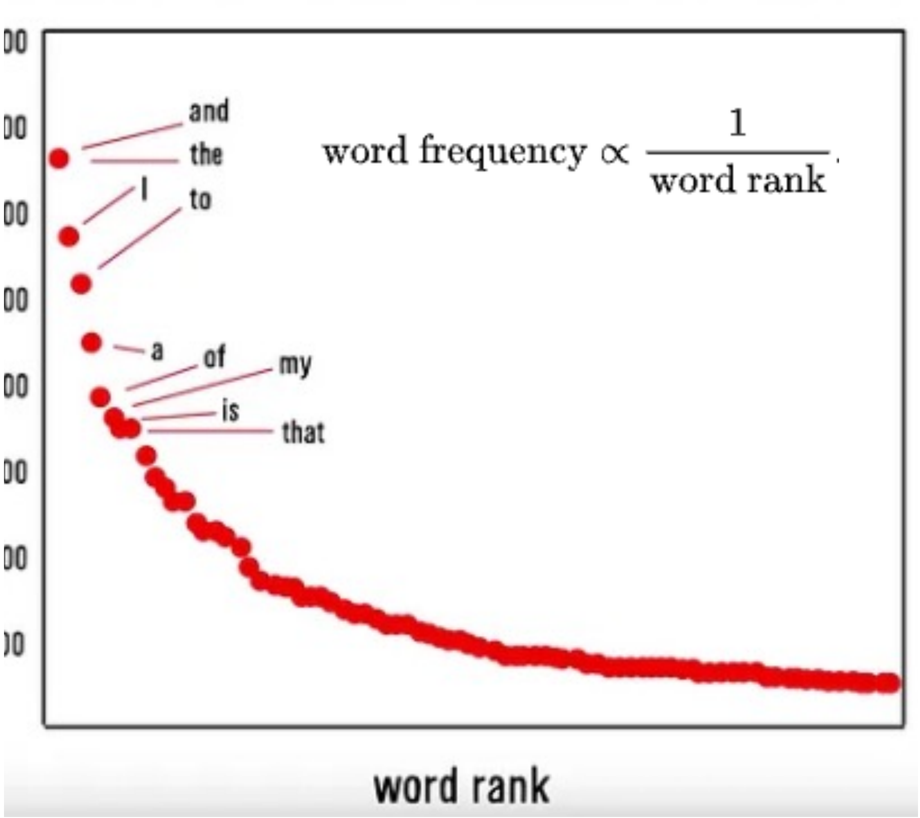
Github: https://github.com/KaihuaTang/Long-Tailed-Recognition.pytorch

Contents

Example: Zipf's law

Word frequency and rank in Romeo & Juliet.

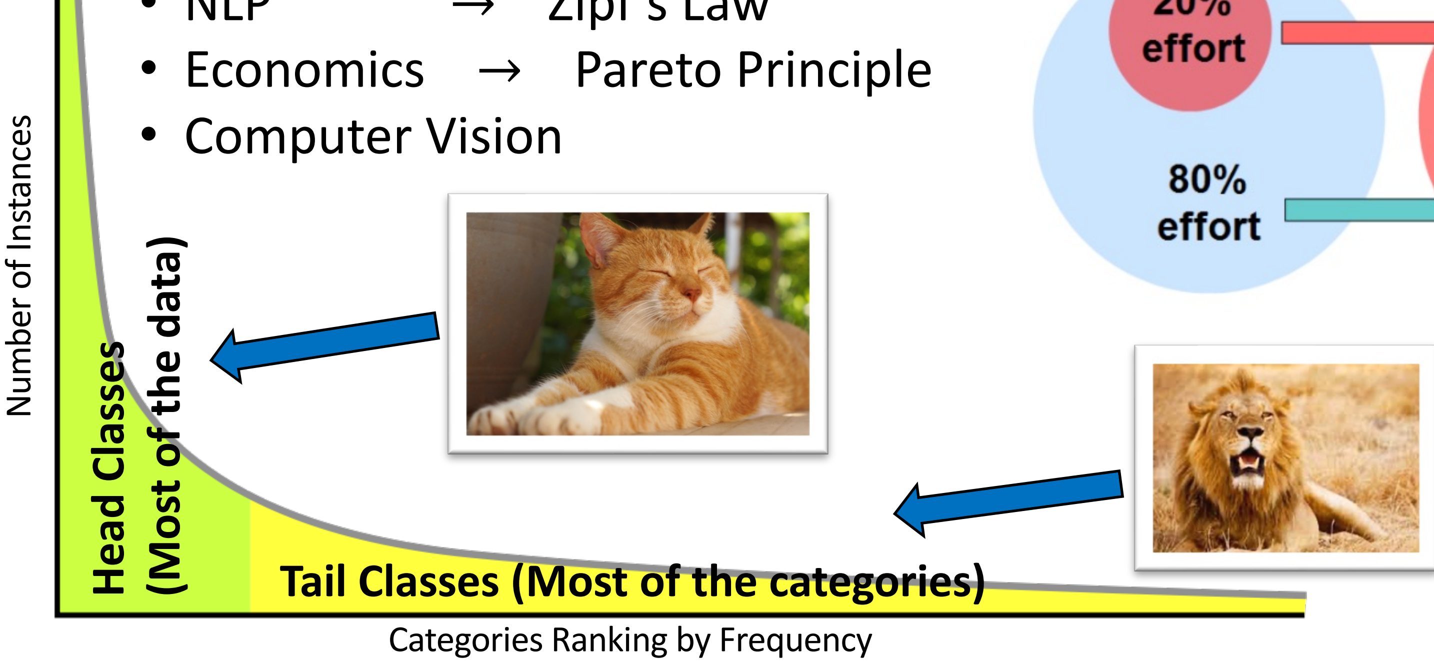


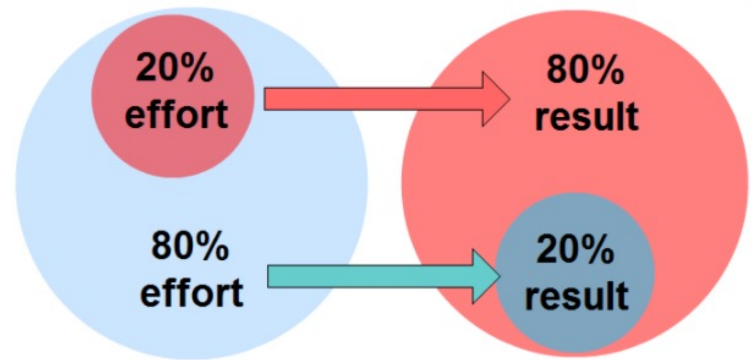


Long-Tailed Distribution

What is long-tailed distribution?

- NLP → Zipf's Law
- Economics → Pareto Principle
- Computer Vision

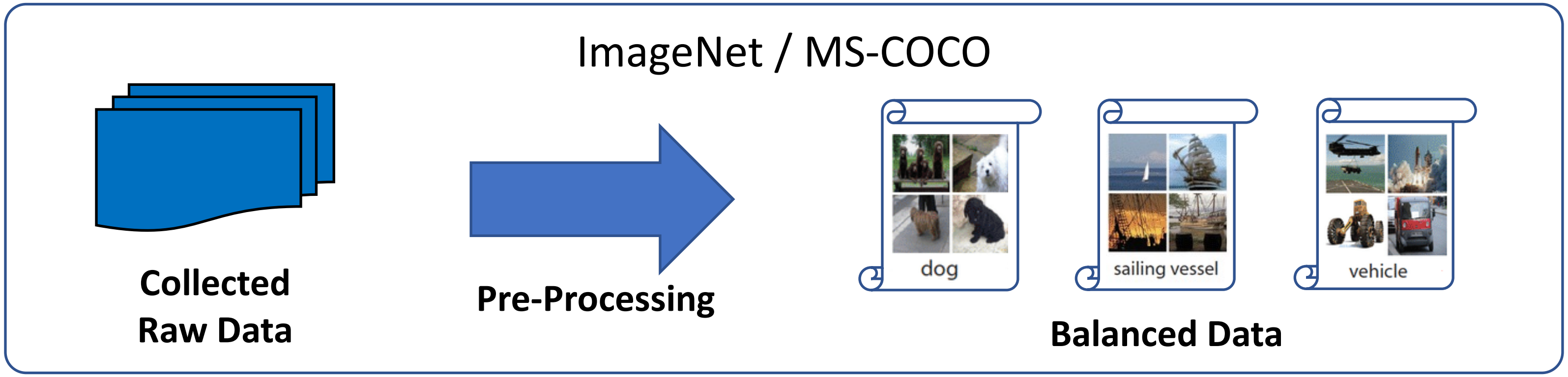




Long-Tailed Distribution

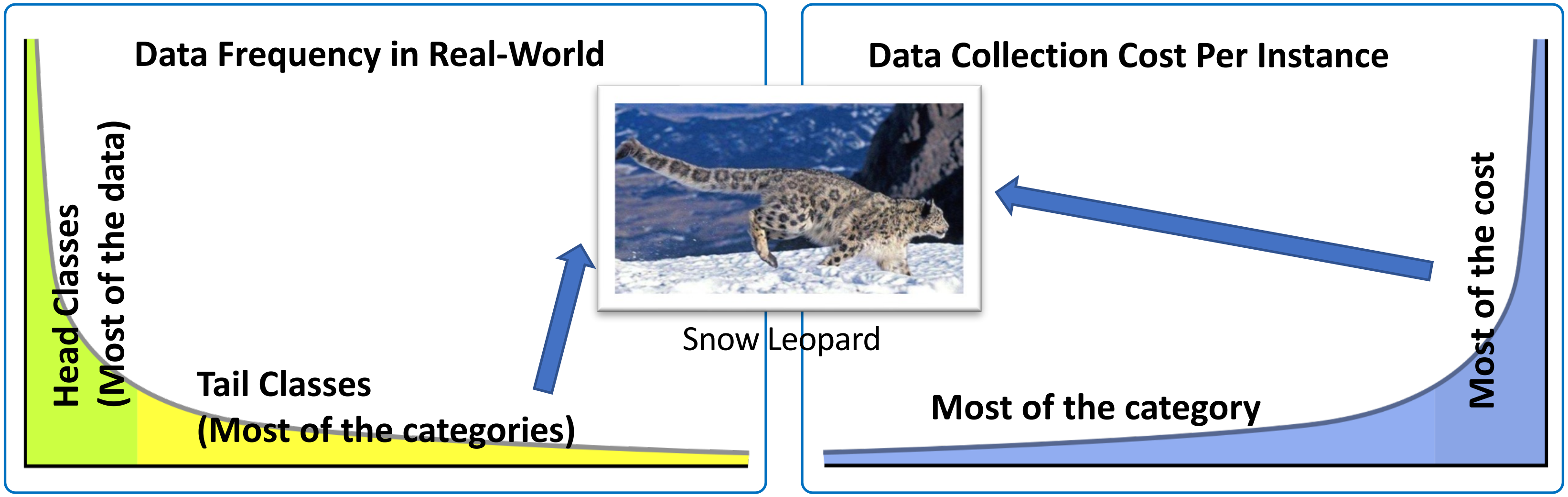
- Why we never heard about long tail problem in ML before?

It's because the dataset we saw has already been balanced by the pre-processing in the data collection stage.



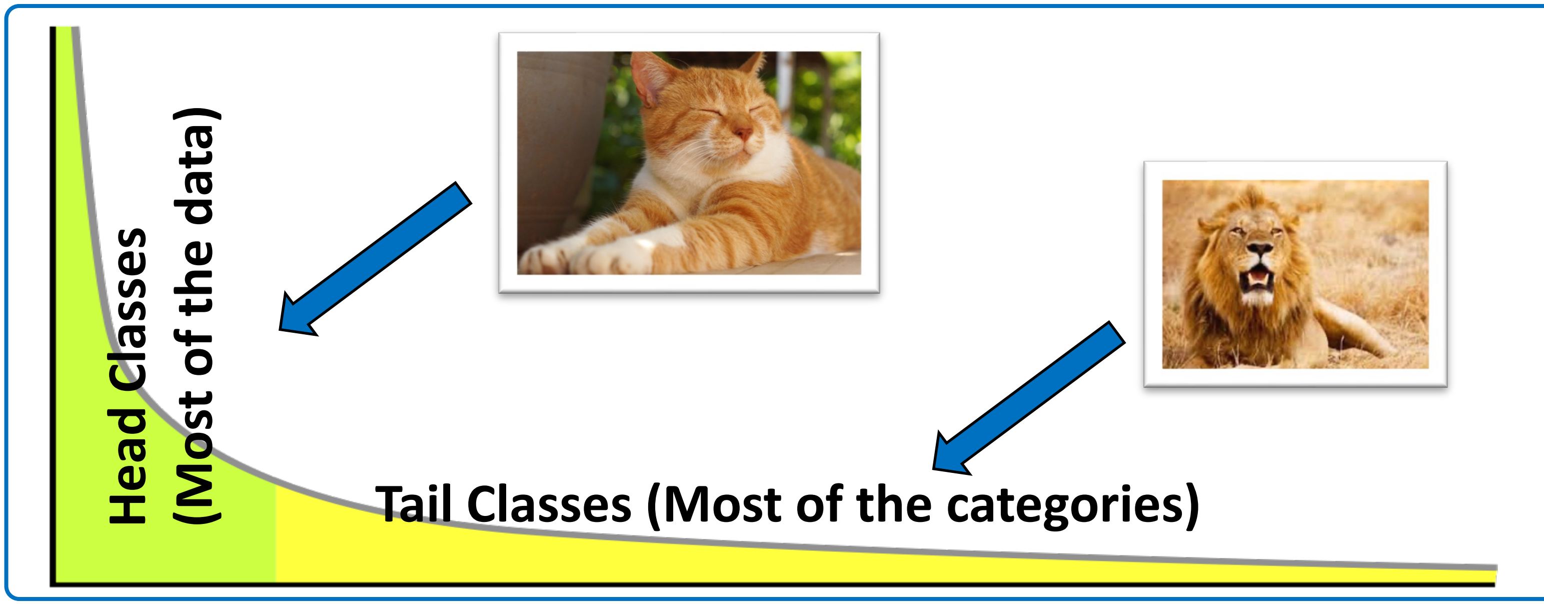
Limitations of Balanced Datasets

- Question 1: What's the problem of balancing all the dataset?



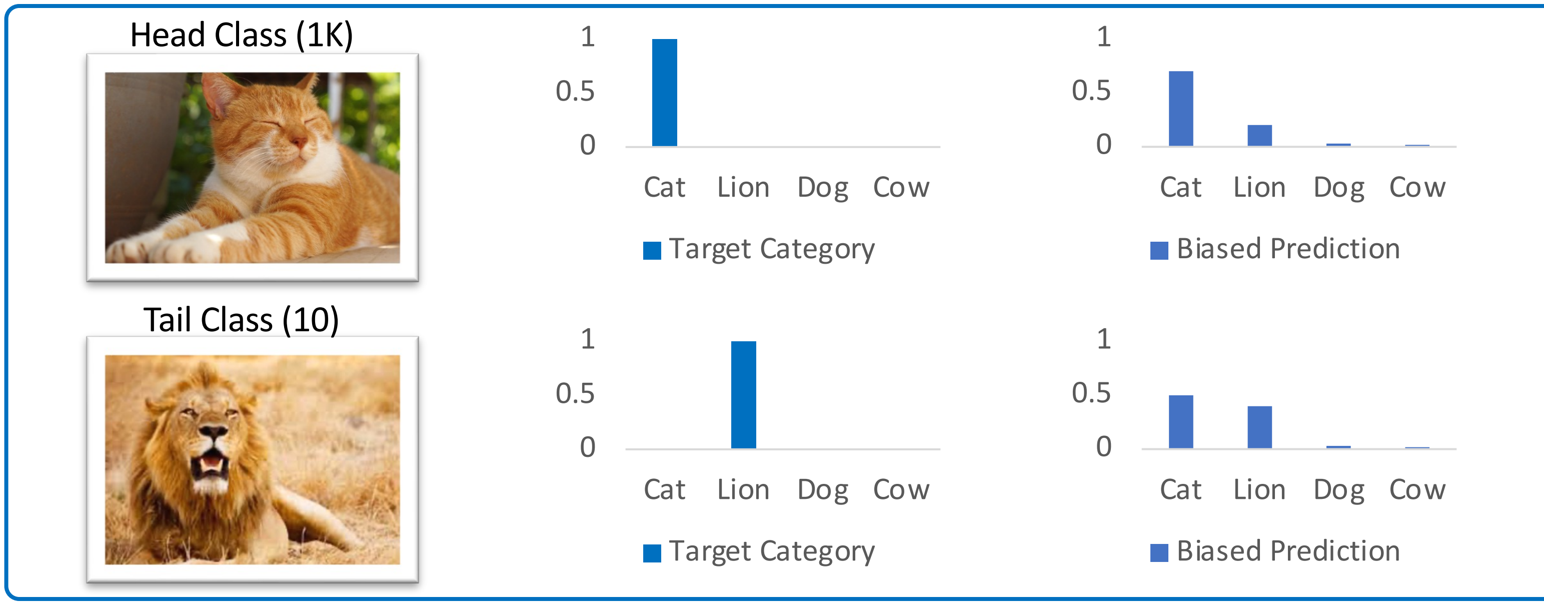
Problems of Long-Tailed Datasets

- Question 2: Why not directly use the long-tailed dataset?



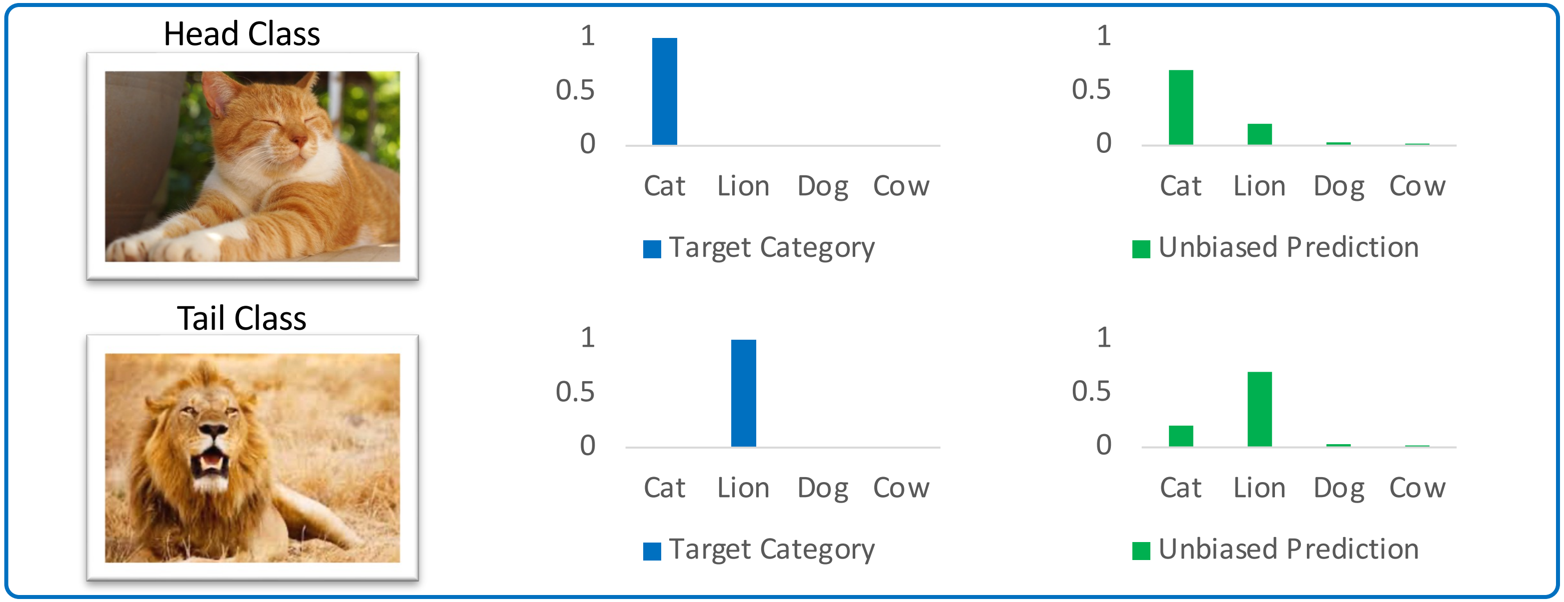
Long-Tailed Classification

- The Problem of Long-Tailed Datasets



Long-Tailed Classification

- The Target of Long-Tailed Classification

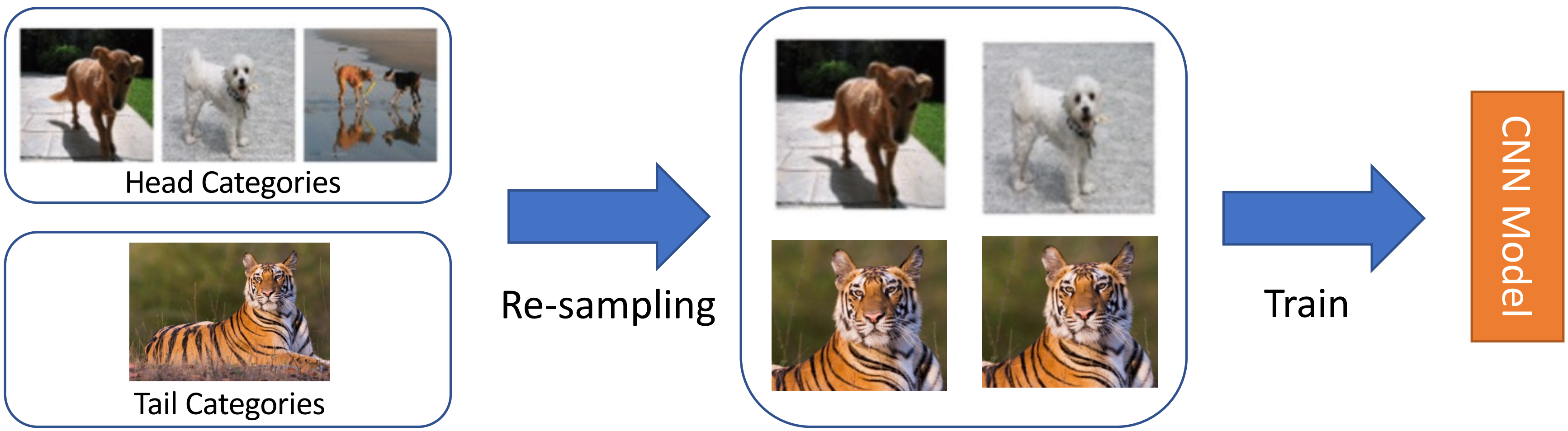


Contents

- Long-Tailed Classification
- Related Work
- The Proposed Causal Graph
- De-confound TDE
- Advantages
- Experiments

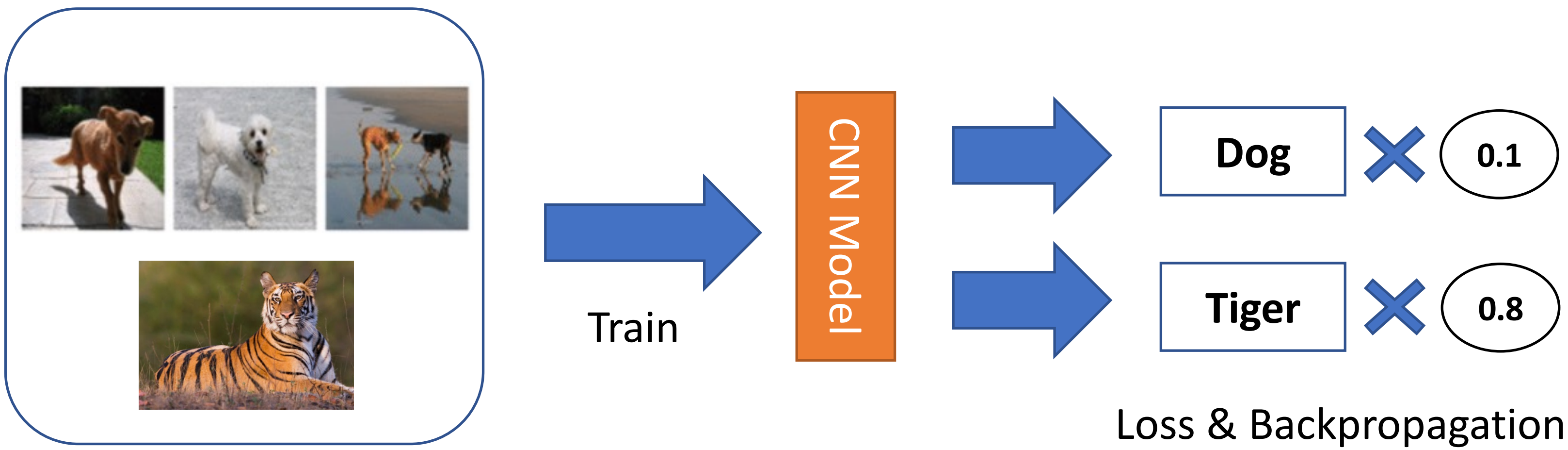
Re-balancing (Re-sampling/Re-weighting)

- The most common solutions:
  - Re-sampling
  - Re-weighting



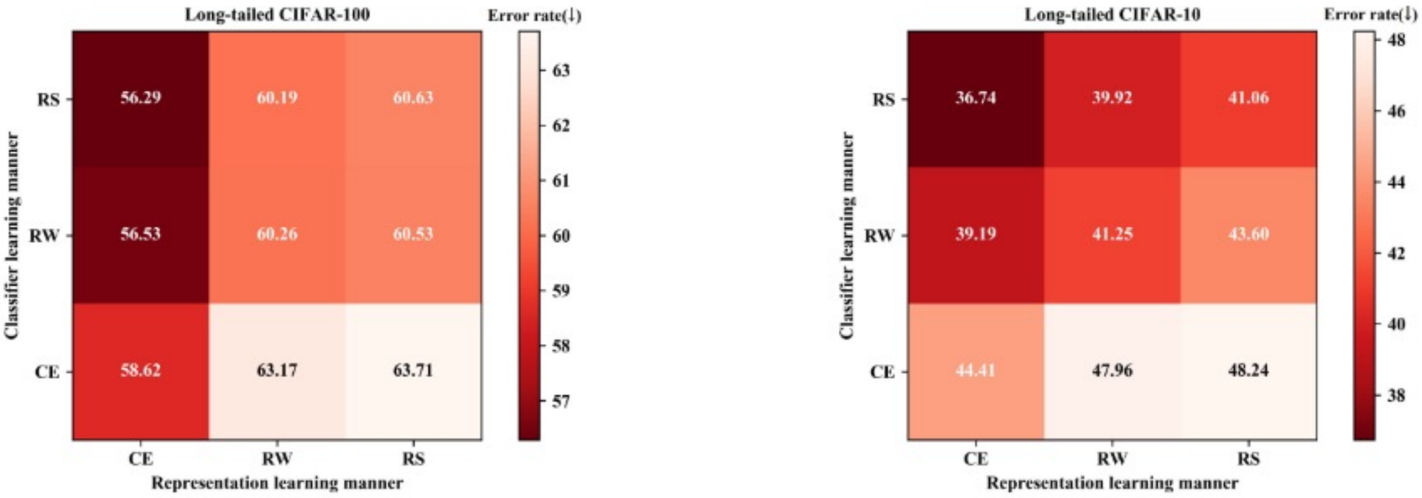
Re-balancing (Re-sampling/Re-weighting)

- The most common solutions:
  - Re-sampling
  - Re-weighting



Two-Stage Re-balancing

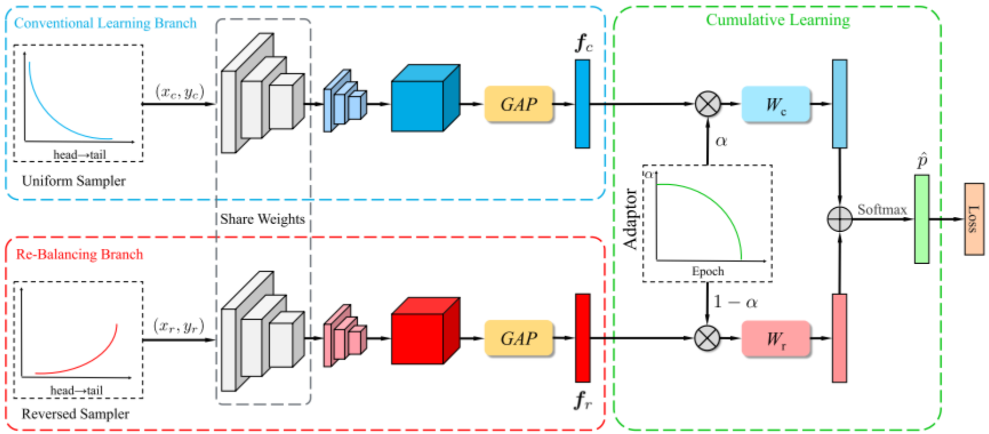
- Drawbacks of conventional re-balancing:
  - Foreknowledge towards the data: knowing the future data distribution before learning
  - Under-fitting to the head
  - Over-fitting to the tail

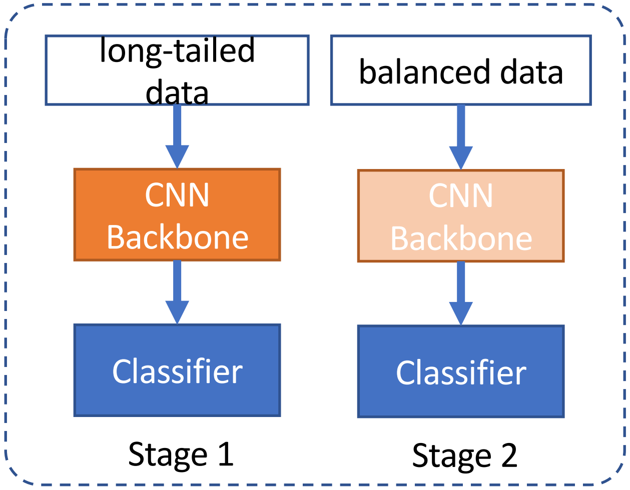


BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition, CVPR 2020

Two-Stage Re-balancing

- The two-stage solutions for the above drawbacks:
  - Smoothly adapted bilateral-branch training [3]
  - Decoupled two-stage training [4]





[3] BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition, CVPR 2020
[4] Decoupling Representation and Classifier for Long-Tailed Recognition, ICLR 2020

What’s the problem of existing two-stage solutions?

They fail to explain the whys and wherefores of their solutions:

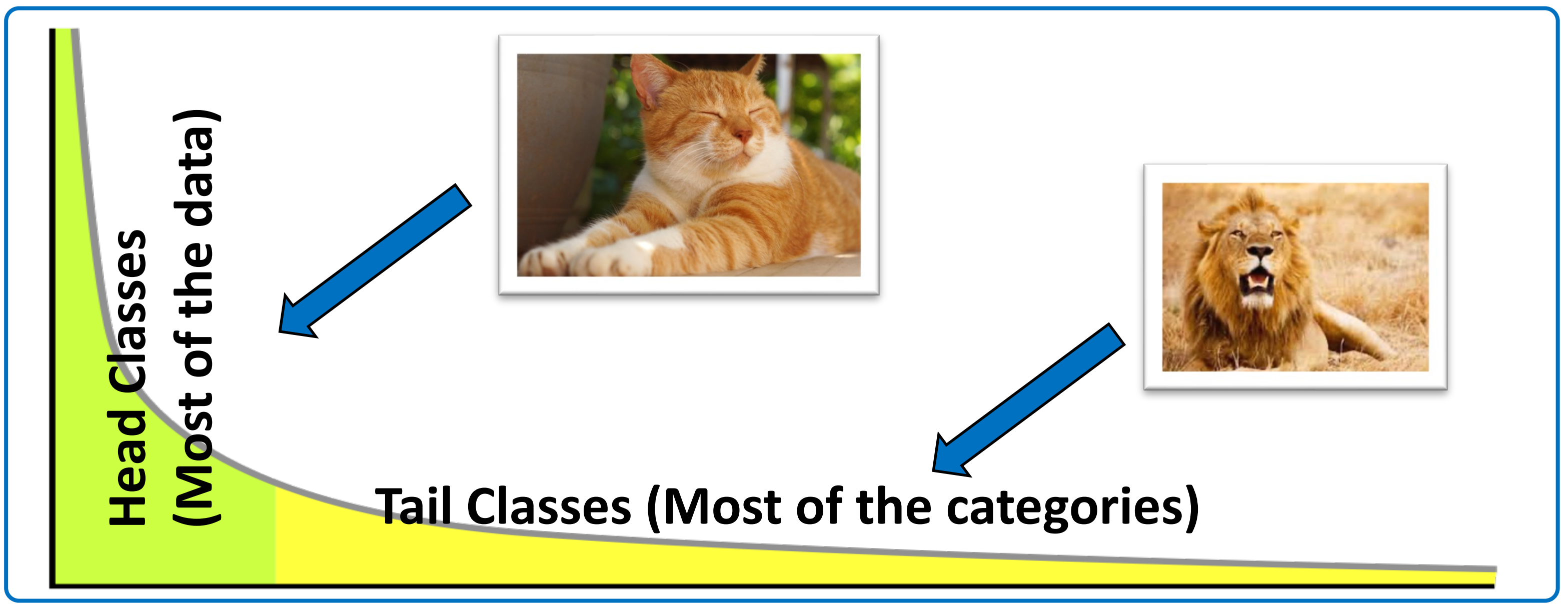
- why is the re-balanced classifier good but the re-balanced feature learning bad?
- why does the two-stage training significantly outperform the end-to-end one in long-tailed classification?

Contents

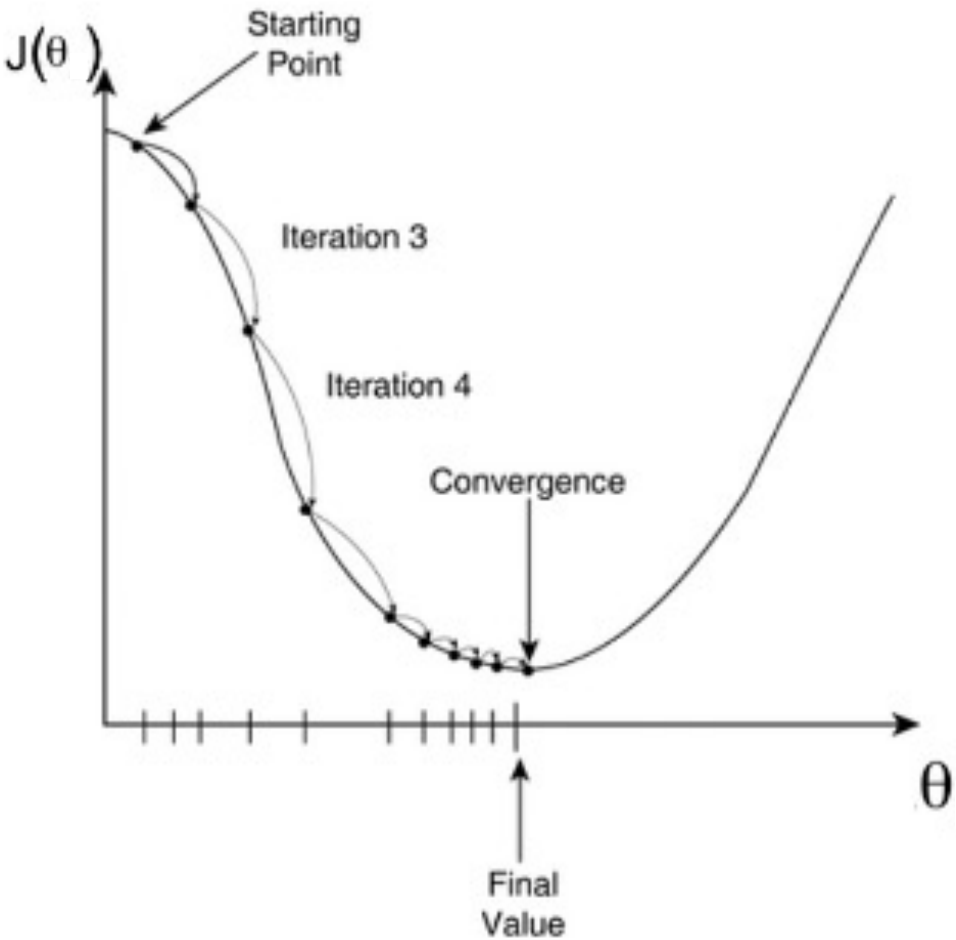
- Long-Tailed Classification
- Related Work
- The Proposed Causal Graph
- De-confound TDE
- Advantages
- Experiments

We should not blame the dataset

- We, human beings, also live in a long-tailed world.
- The problem must reside in the learning framework.



Gradient descent



gradient

$$\theta_t = \theta_{t-1} - lr \cdot v_t$$

model parameters

learning rate

Accumulative Momentum Effect

- The PyTorch implementation of SGD with **momentum** [1]:

Momentum Decay Ratio

$$v_t = \underbrace{\mu \cdot v_{t-1}}_{momentum} + g_t, \quad \theta_t = \theta_{t-1} - lr \cdot v_t,$$

velocity gradient model parameters learning rate

- Modern SGD variants (e.g., Adam, SGD-M, etc.) often involve momentum (or acceleration), which accumulates historical gradients to speed up convergence, akin to a heavy ball rolling down the loss function landscape.
- The moving average momentum will encode the data distribution, that creates a <u>shortcut</u> towards the <u>head</u>.

[1] https://pytorch.org/docs/stable/_modules/torch/optim/sgd.html

Accumulative Momentum Effect

SGD Momentum in *Balanced* Dataset

SGD Momentum in *Long-Tailed* Dataset

- Global Optima for All Categories
- Local Optima for Head Categories
- Momentum Direction in Balanced Data
- Momentum Direction in Long-Tailed Data

Causal Effect of Momentum

Why not remove the momentum when training the long-tailed dataset?

Remove Momentum:

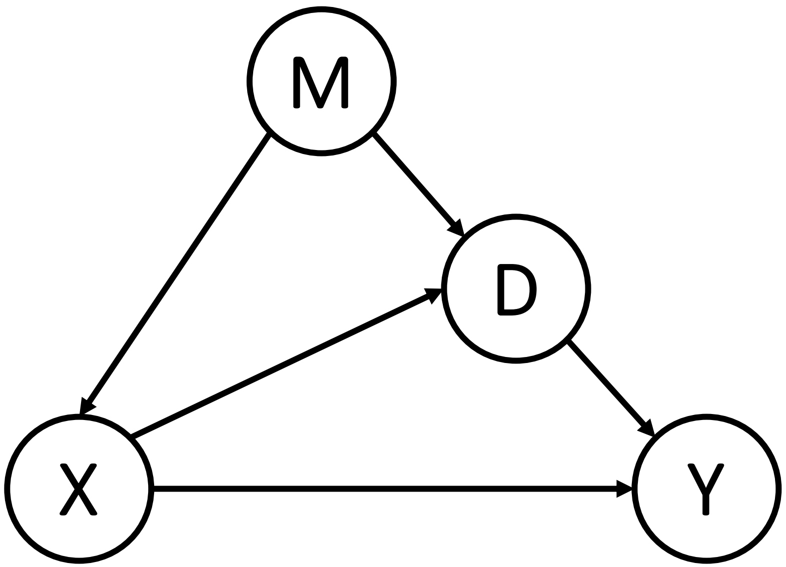
1. **Unstable Gradient**
2. **Local Optima**
3. **SGD Still Accumulates**

Keep Momentum in Training

+

Remove Bad Causal Effect

The Proposed Causal Graph



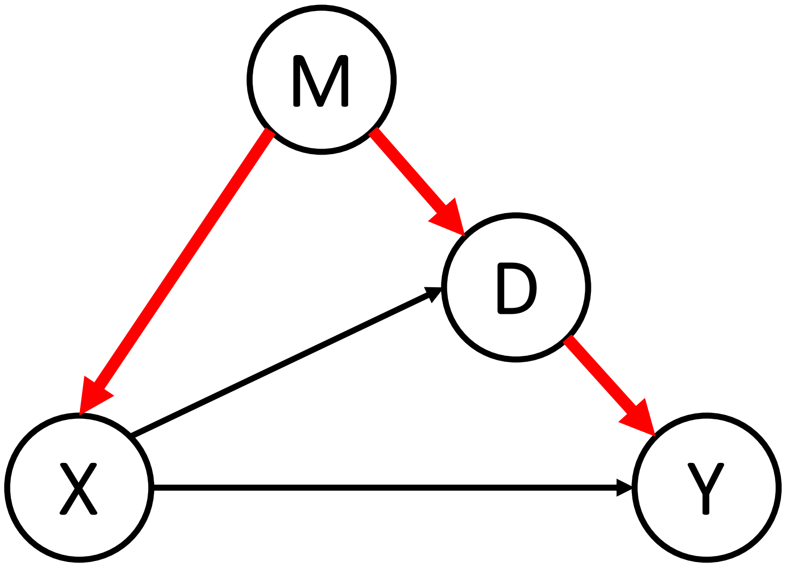
X : Feature

Y : Prediction

M: Momentum

D : Projection on Head

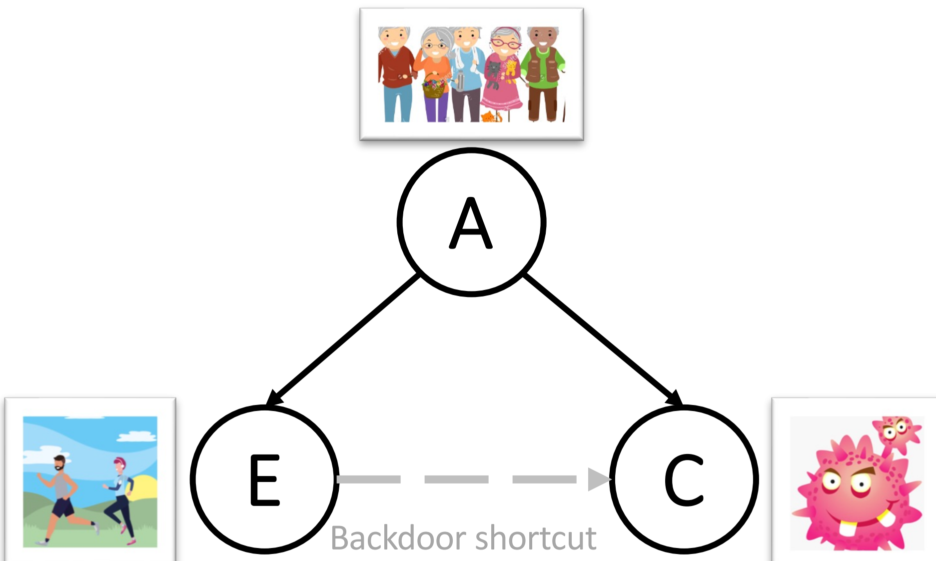
Two Undesired Causal Effects of Momentum



Two Undesired Causal Effects of Momentum:

1. Backdoor shortcut
2. Indirect Mediator Effect

Confounder and backdoor shortcut

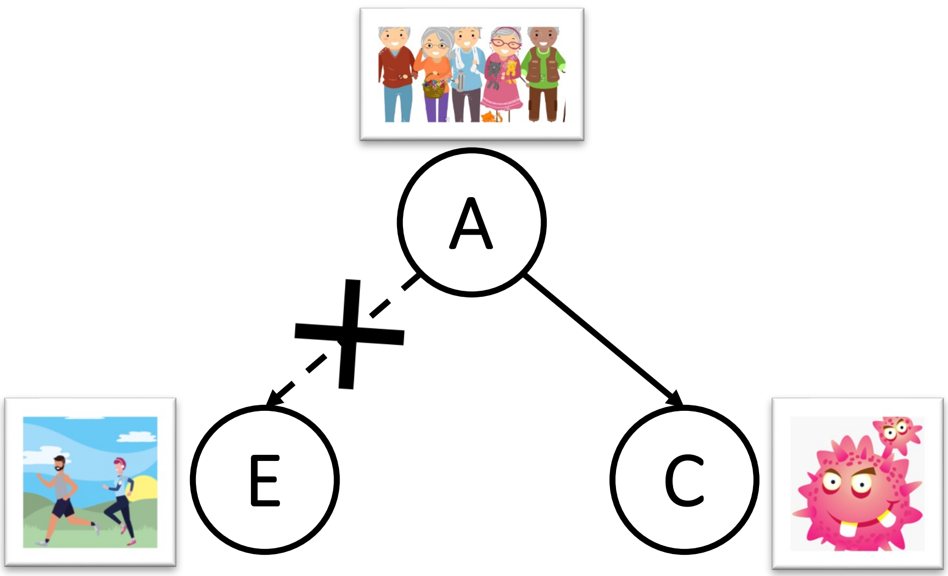


A: age **E: exercise** **C: cancer**

Backdoor shortcut:

1. $A \uparrow \;\Rightarrow\; E \uparrow$
2. $A \uparrow \;\Rightarrow\; C \uparrow$
3. $E \uparrow \;\Rightarrow ?\; C \uparrow$

backdoor Adjustment



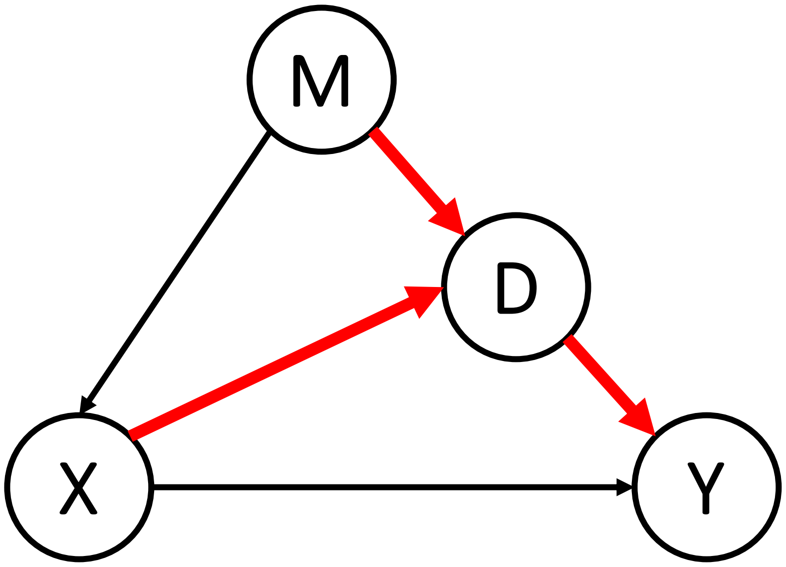
A: age   E: exercise   C: cancer

Backdoor Adjustment

Intervention on E:

$$P(C|do(E)) = \sum_{a} P(C|E, A = a)P(A = a)$$

Two Undesired Causal Effects of Momentum



Two Undesired Causal Effects of Momentum：

1. Backdoor shortcut
2. Indirect Mediator Effect

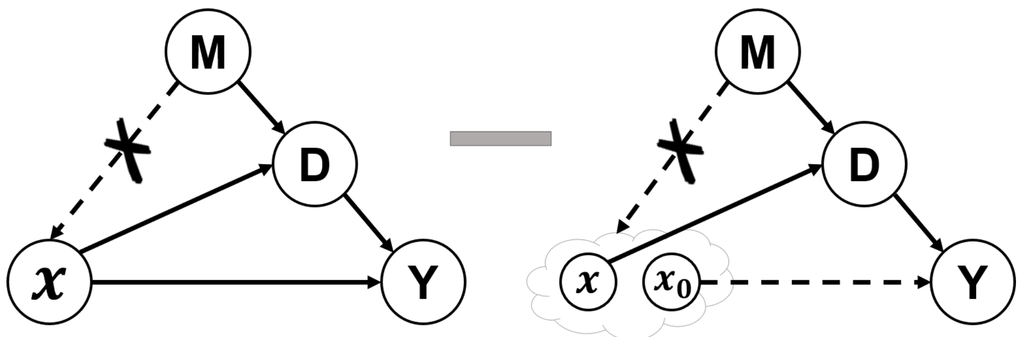
Contents

- Long-Tailed Classification
- Related Work
- The Proposed Causal Graph
- De-confound TDE
- Advantages
- Experiments

De-confound TDE Classifier

The definition of Total Direct Effect (TDE):

$$argmax_{i \in C}\, TDE(Y_i) = [Y_d = i | do(X = x)] - [Y_d = i | do(X = x_0)]$$



The proposed classifier = De-confounded Training + TDE Inference

Contents

Advantages

The proposed de-confound TDE **simple**, **adaptive**, and **agnostic** to the prior statistics of the class distribution:

1. It doesn’t introduce any additional stages or modules.
2. It can be applied to a variety of tasks, including but not limited to image classification, object detection, instance segmentation.
3. It doesn’t rely on the accessibility of data distribution.

Contents

- Long-Tailed Classification
- Related Work
- The Proposed Causal Graph
- De-confound TDE
- Advantages
- Experiments

Image Classification: ImageNet-LT

- Experiments on ImageNet-LT

Methods	Many-shot	Medium-shot	Few-shot	Overall
Focal Loss† [24]	64.3	37.1	8.2	43.7
OLTR† [8]	51.0	40.8	20.8	41.9
Decouple-OLTR† [8, 10]	59.9	45.8	27.6	48.7
Decouple-Joint [10]	65.9	37.5	7.7	44.4
Decouple-NCM [10]	56.6	45.3	28.1	47.3
Decouple-cRT [10]	61.8	46.2	27.4	49.6
Decouple-τ-norm [10]	59.1	46.9	30.7	49.4
Decouple-LWS [10]	60.2	47.2	30.3	49.9
Baseline	66.1	38.4	8.9	45.0
Cosine† [38, 39]	67.3	41.3	14.0	47.6
Capsule† [8, 42]	67.1	40.0	11.2	46.5
(Ours) De-confound	**67.9**	42.7	14.7	48.6
(Ours) Cosine-TDE	61.8	47.1	30.4	50.5
(Ours) Capsule-TDE	62.3	46.9	30.6	50.6
(Ours) De-confound-TDE	62.7	**48.8**	**31.6**	**51.8**

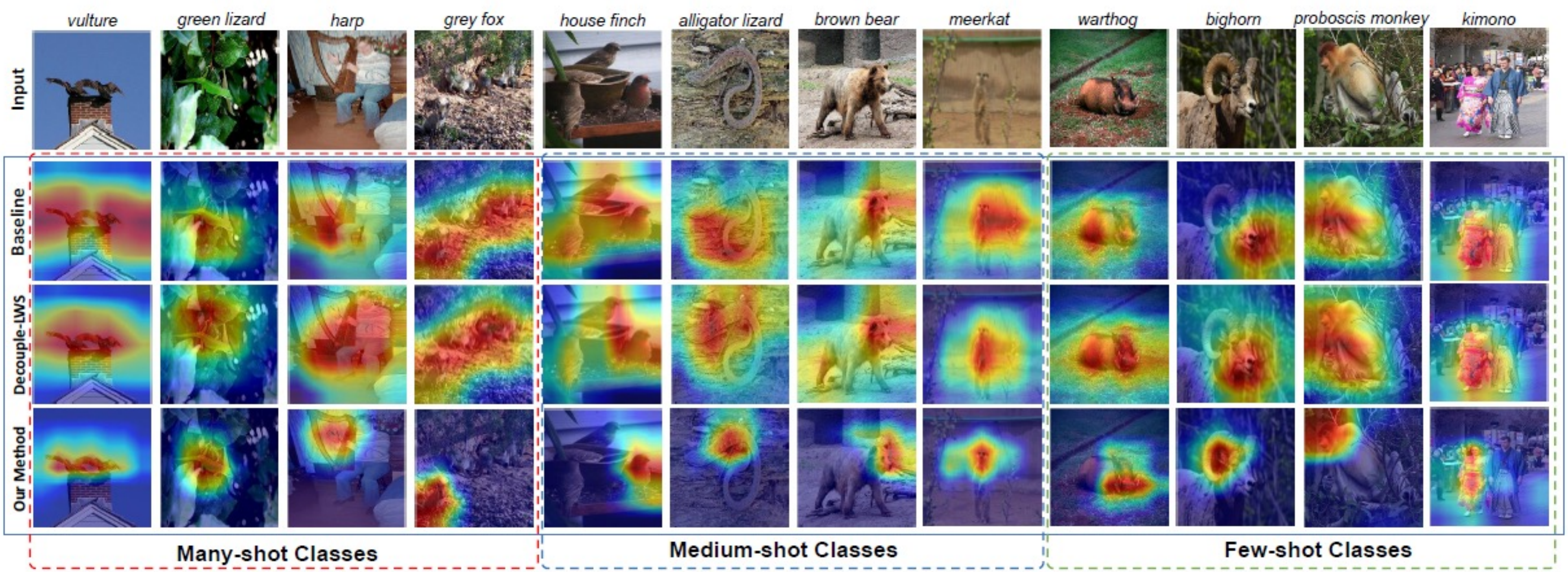
Image Classification: ImageNet-LT

- Will the improvement be consistent across different backbones?

Methods	Backbone	Many-shot	Medium-shot	Few-shot	Overall
Baseline	ResNeXt-50	66.1	38.4	8.9	45.0
De-confound	ResNeXt-50	67.9	42.7	14.7	48.6
De-confound-TDE	ResNeXt-50	62.7	48.8	31.6	51.8
Baseline	ResNeXt-101	68.7	42.5	11.8	48.4
De-confound	ResNeXt-101	**68.9**	44.3	16.5	50.0
De-confound-TDE	ResNeXt-101	64.7	**50.0**	**33.0**	**53.3**

Grad-cam Visualization on ImageNet-LT

What does our model see from images?



Interventional Few-Shot Learning

Yue, Zhongqi, et al

NeurIPS 2020

Few-shot Learning

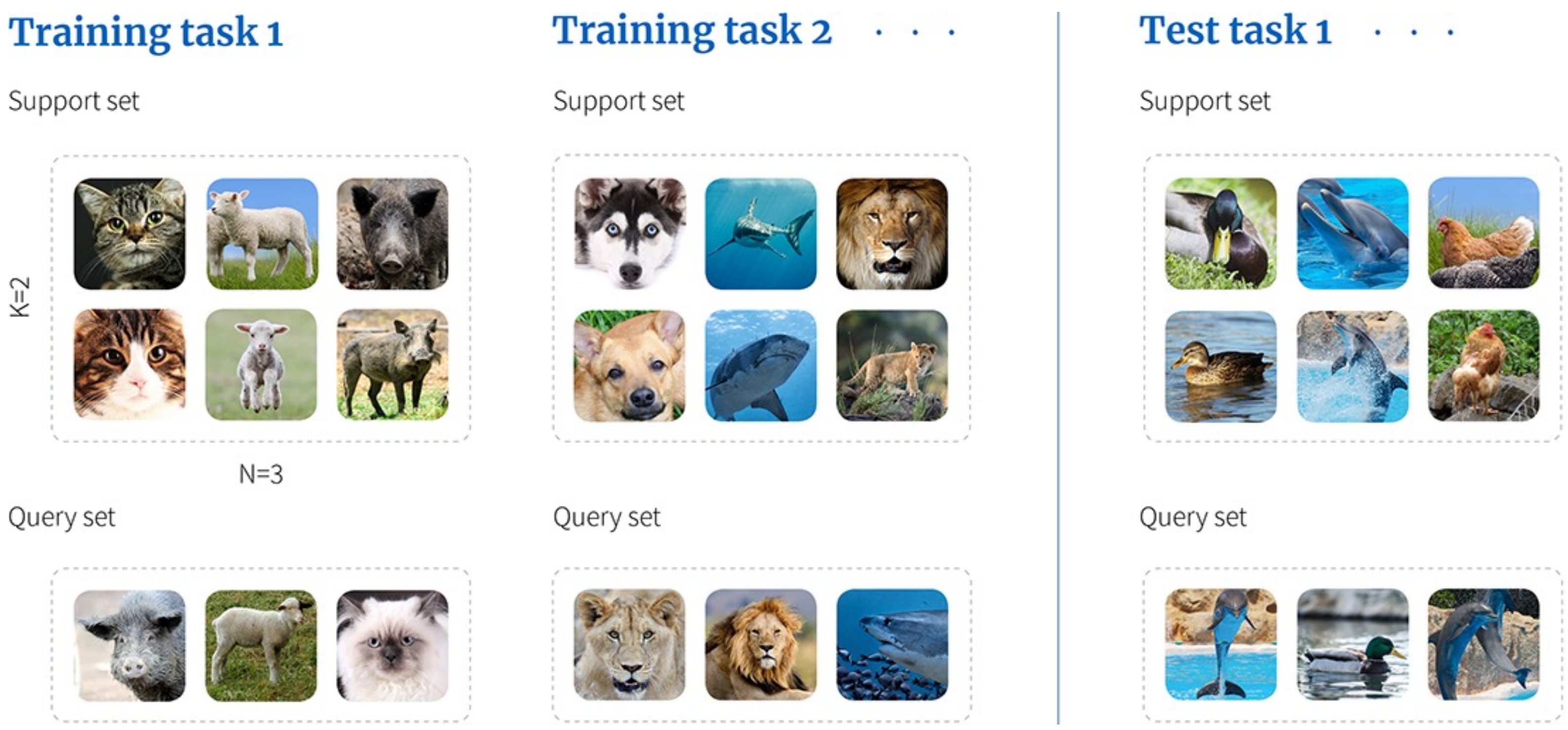
- Few-shot learning is usually studied using *N-way-K-shot classification*.

  N-way-K-shot classification

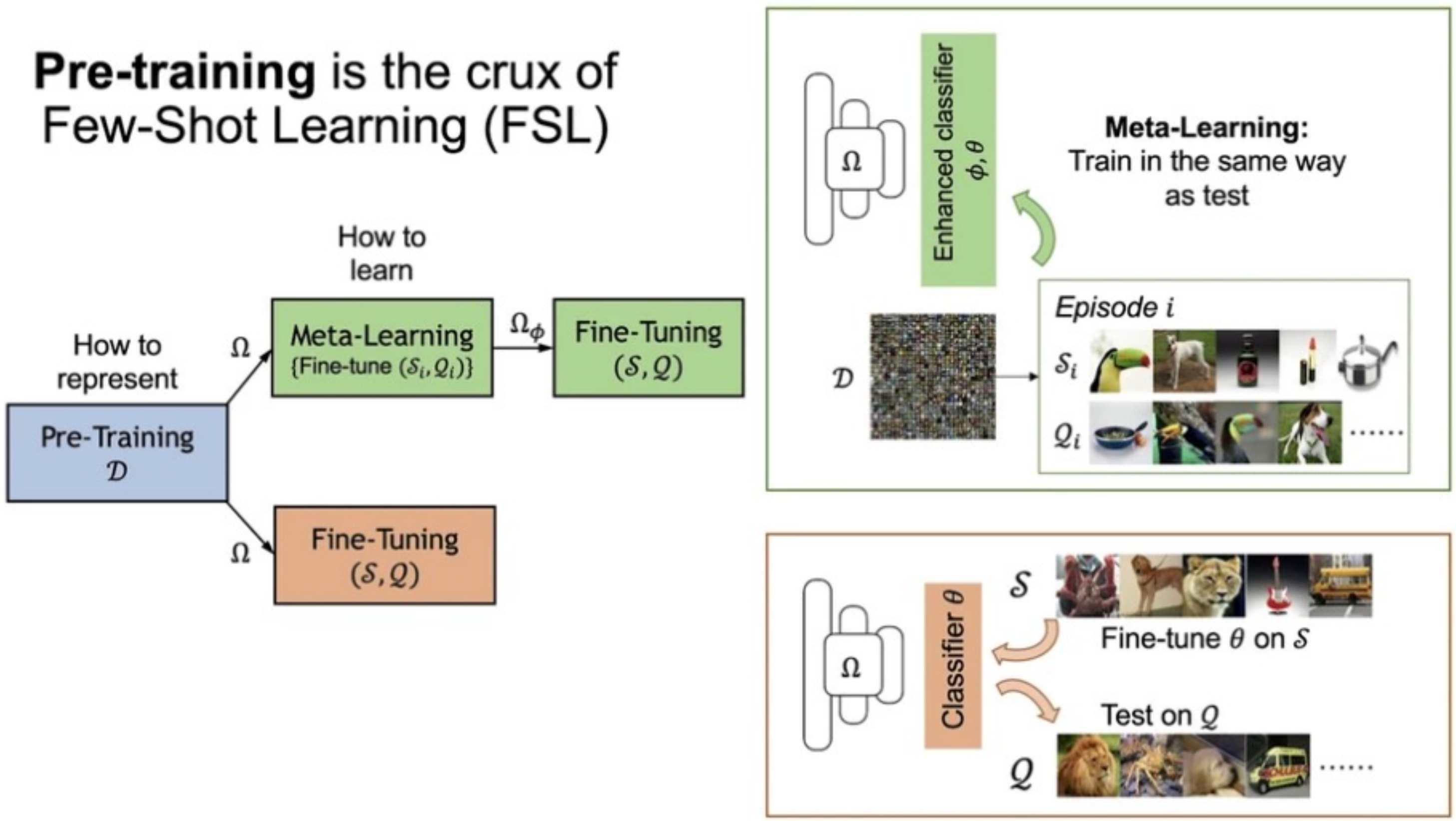
  - N classes with K samples of each.
  - E.g., classify N = 10 classes with only K = 5 samples from each to train from.

Meta-learning

- “learn to learn”: learn a meta-model that quickly adapts to different few-shot datasets.

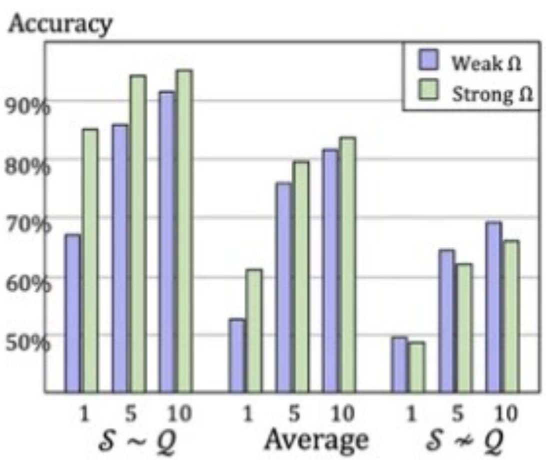


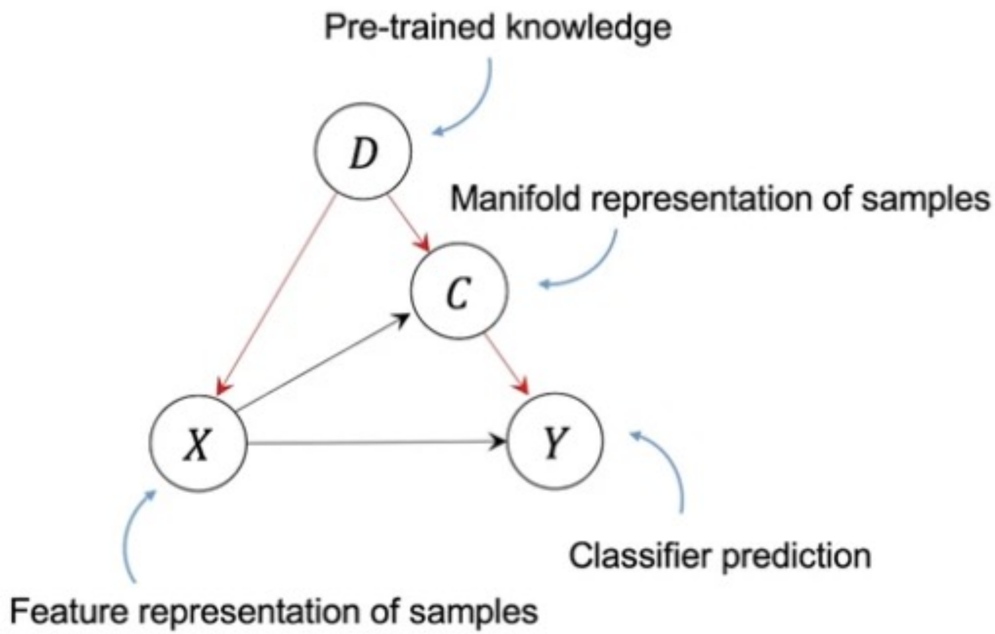
Pre-training



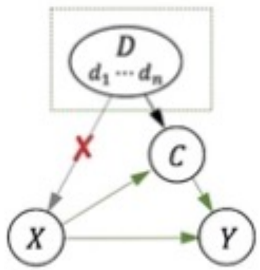
Pre-training confounds FSL

An example of $\mathcal{S} \nsim \mathcal{Q}$





Intervention removes bias from confounding

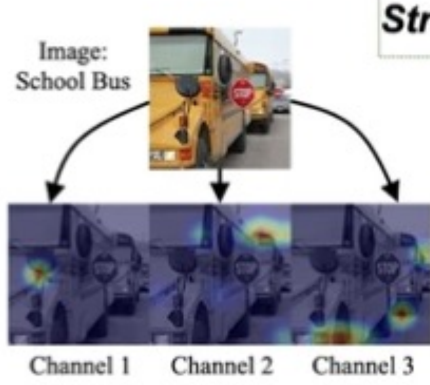


$P(Y|do(X))$

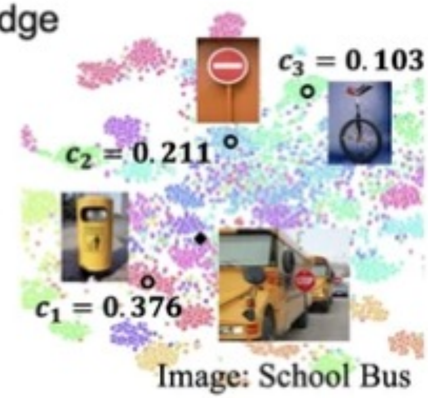
Use *Backdoor Adjustment* to achieve $P(Y|do(X))$

$$P(Y|do(X=\boldsymbol{x})) = \sum_{i=1}^{n} P(Y|X=\boldsymbol{x}, D=d_i, C=g(\boldsymbol{x}, d_i))\boxed{P(D=d_i)}$$

Stratify the pre-training knowledge

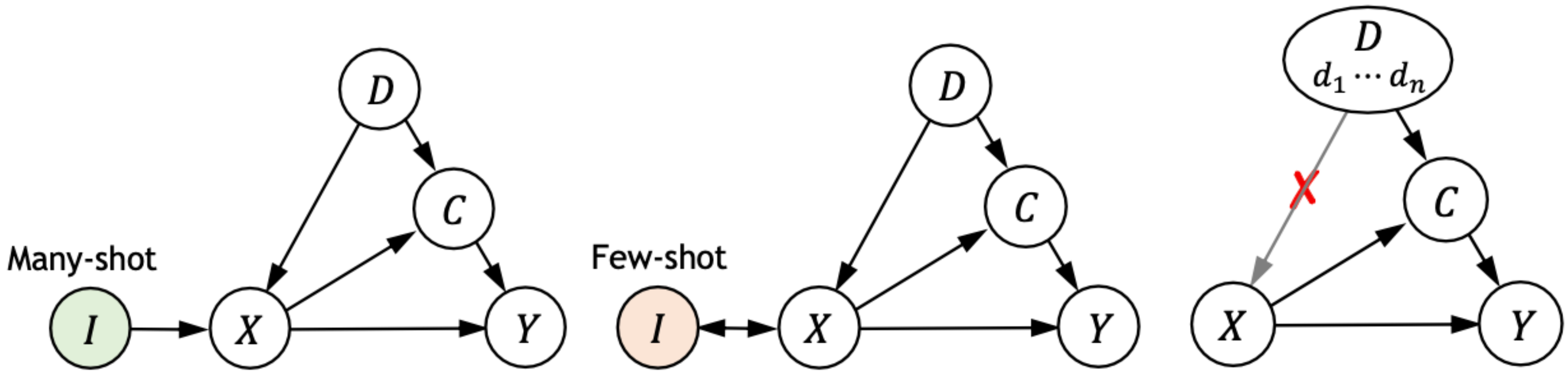


Feature-wise



Class-wise

Many-shot, Few-shot, and interventional few-shot



(a) MSL, where $p(y \mid \boldsymbol{x}) \approx p(y \mid \mathrm{do}(\boldsymbol{x}))$

(b) FSL, where $p(y \mid \boldsymbol{x}) \not\approx p(y \mid \mathrm{do}(\boldsymbol{x})$

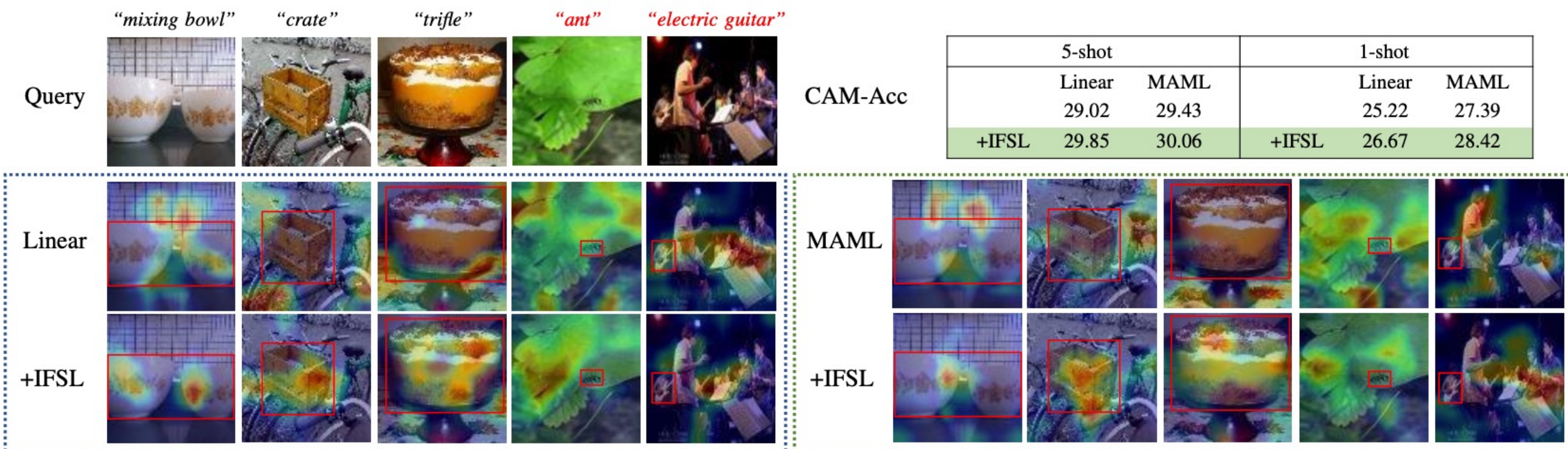
(c) IFSL, where $p(y \mid \mathrm{do}(\boldsymbol{x}))$ through back-door adjustment.

Consistent Improvement after Intervention

	Method		ResNet-10				WRN-28-10			
			*mini*ImageNet		*tiered*ImageNet		*mini*ImageNet		*tiered*ImageNet	
			5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot
Fine-Tuning	Linear		76.38	56.26	81.01	61.39	79.79	60.69	85.37	67.27
		+IFSL+2.19	77.97+1.59	60.13+3.87	82.08+1.07	64.29+2.9	80.97+1.18	64.12+3.43	86.19+0.82	69.96+2.69
	Cosine		76.68	56.40	81.13	62.08	79.72	60.83	85.41	67.30
		+IFSL+1.77	77.63+0.95	59.84+3.44	81.75+0.62	64.47+2.39	80.74+1.02	63.76+2.93	86.13+0.72	69.36+2.06
	k-NN		76.63	55.92	80.85	61.16	79.60	60.34	84.67	67.25
		+IFSL+3.13	78.42+1.79	62.31+6.36	81.98+1.13	65.71+4.55	81.08+1.48	64.98+4.64	86.06+1.39	70.94+3.69
Meta-Learning	MAML [18]		70.85	56.59	74.02	59.17	73.92	58.02	77.20	61.40
		+IFSL+5.55	76.37+5.52	59.36+2.77	81.04+7.02	63.88+4.71	79.25+5.33	62.84+4.82	85.10+7.90	67.70+6.30
	LEO [53]		74.49	58.48	80.25	65.25	75.86	59.77	82.15	68.90
		+IFSL+1.94	76.91+2.42	61.09+2.61	81.43+1.18	66.03+0.78	77.72+1.86	62.19+2.42	85.04+2.89	70.28+1.38
	MTL [56]		75.65	58.49	81.14	64.29	77.30	62.99	83.23	70.08
		+IFSL+2.02	78.03+2.38	61.17+2.68	82.35+1.21	65.72+1.43	80.20+2.9	64.40+1.41	86.02+2.79	71.45+1.37
	MN [61]		75.21	61.05	79.92	66.01	77.15	63.45	82.43	70.38
		+IFSL+1.34	76.73+1.52	62.64+1.59	80.79+0.87	67.30+1.29	78.55+1.40	64.89+1.44	84.03+1.60	71.41+1.03
	SIB [29]		78.88	67.10	85.09	77.64	81.73	71.31	88.19	81.97
	(transductive)	+IFSL+1.15	80.32+1.44	68.85+1.75	85.43+0.34	78.03+0.39	83.21+1.48	73.51+2.20	88.69+0.50	83.07+1.10
	SIB [29]		75.64	57.20	81.69	65.51	78.17	60.12	84.96	69.20
	(inductive)	+IFSL+2.05	77.68+2.04	60.33+3.13	82.75+1.06	67.34+1.83	80.05+1.88	63.14+3.02	86.14+1.18	71.45+2.25

Visualization

- IFSL achieves similar or better results in all settings.
  - Focus more on objects



CAM-Acc

	5-shot			1-shot	
	Linear	MAML		Linear	MAML
	29.02	29.43		25.22	27.39
+IFSL	29.85	30.06	+IFSL	26.67	28.42

References & Reading Lists

- Tang, Kaihua, Jianqiang Huang, and Hanwang Zhang. “Long-tailed classification by keeping the good and removing the bad momentum causal effect.” *NeurIPS 2020*.
- Yue, Zhongqi, et al. "Interventional few-shot learning." *NeurIPS* 2020.
- [Awesome Causality in Computer Vision https://github.com/Wangt-CN/Awesome-Causality-in-CV](https://github.com/Wangt-CN/Awesome-Causality-in-CV)

Thank you!
Questions?