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

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

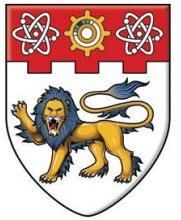
[1]Nanyang Technological University

[2]Damo Academy, Alibaba Group

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

Contents

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

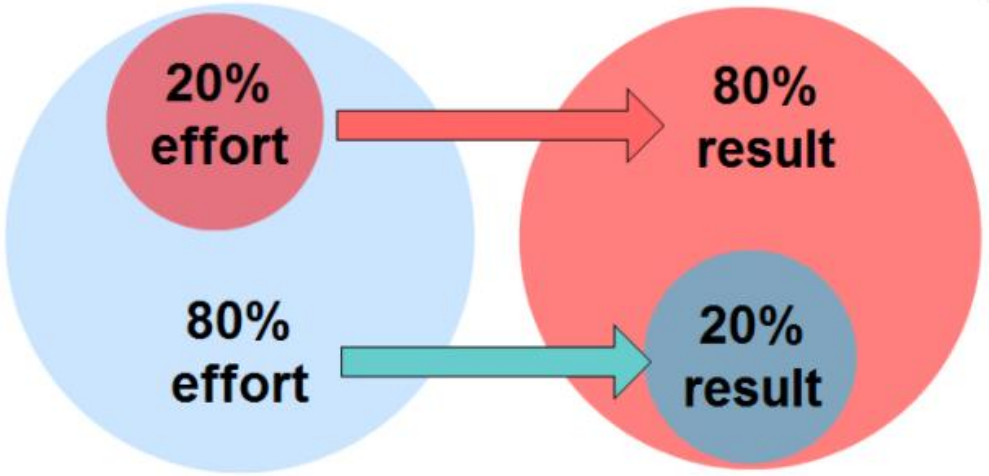


NANYANG TECHNOLOGICAL UNIVERSITY

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?

ImageNet (Image Classification)

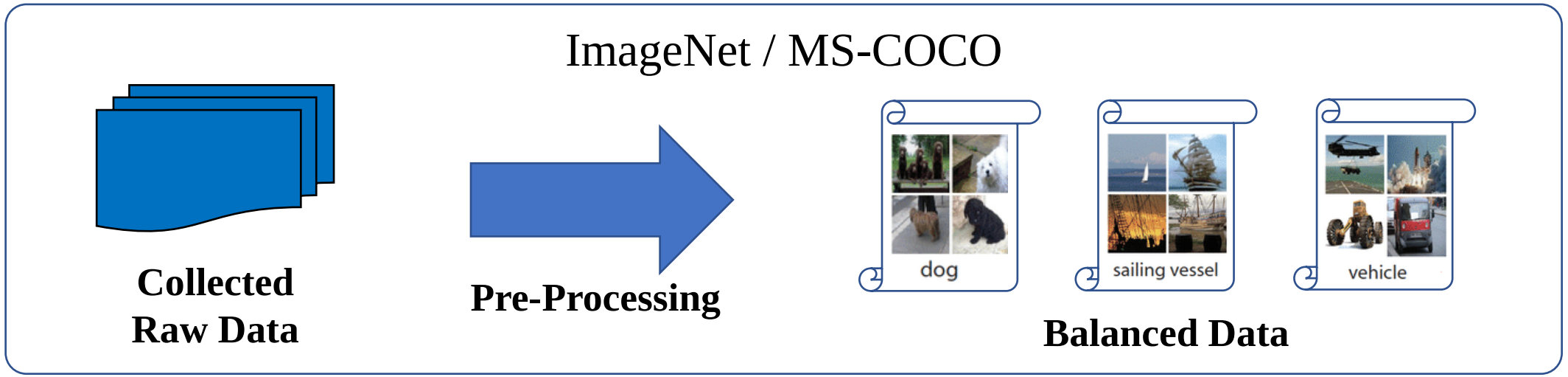
MS-COCO (Object Detection & Instance Segmentation)

[1] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE CVPR*. 2009.
[2] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *ECCV*. Springer, Cham, 2014.

Long-Tailed Distribution

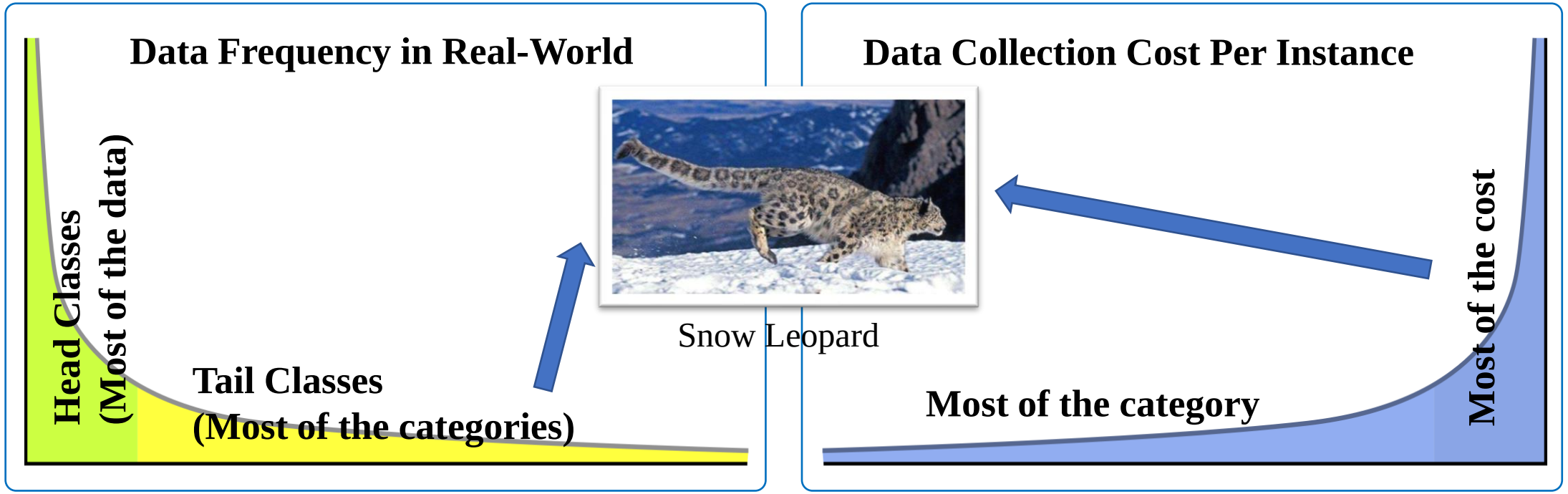
- Why we never heard about long tail problem in ML before?

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



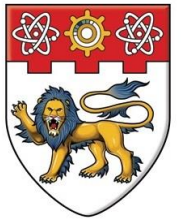
Limitations of Balanced Datasets

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



Contents

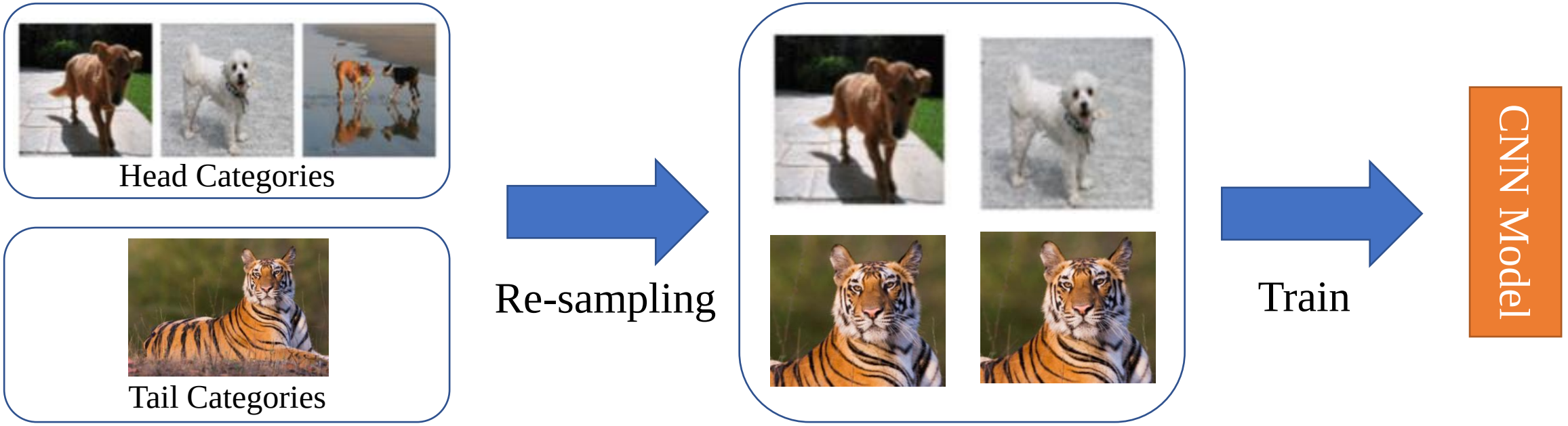
- Long-Tailed Classification
- Related Work
- The Proposed Causal Graph
- De-confound TDE
- Advantages
- Experiments



NANYANG TECHNOLOGICAL UNIVERSITY

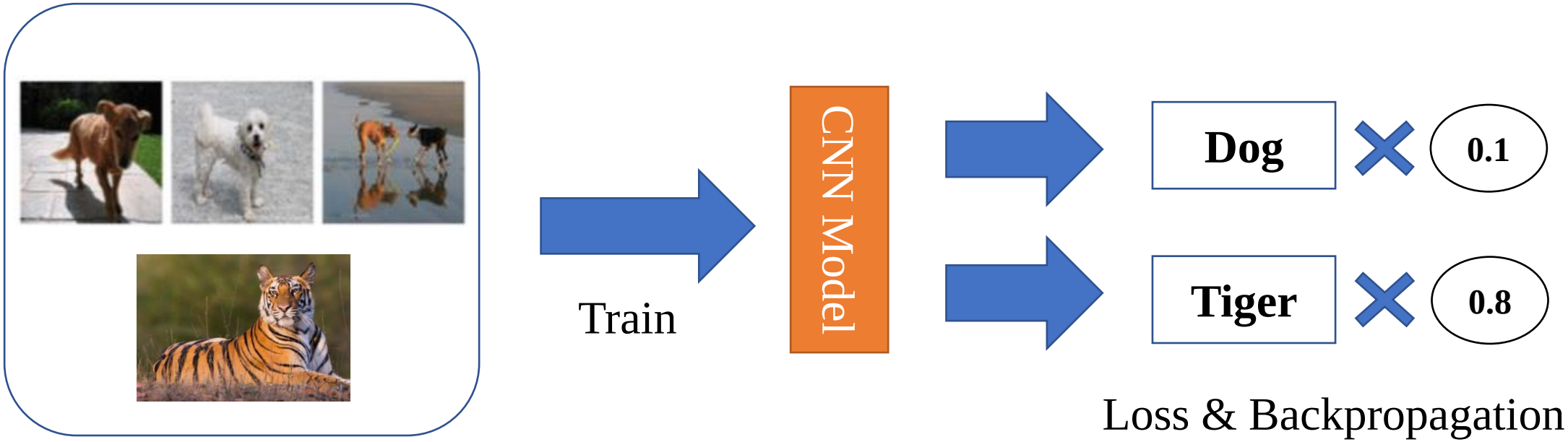
Re-balancing (Re-sampling/Re-weighting)

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



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

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

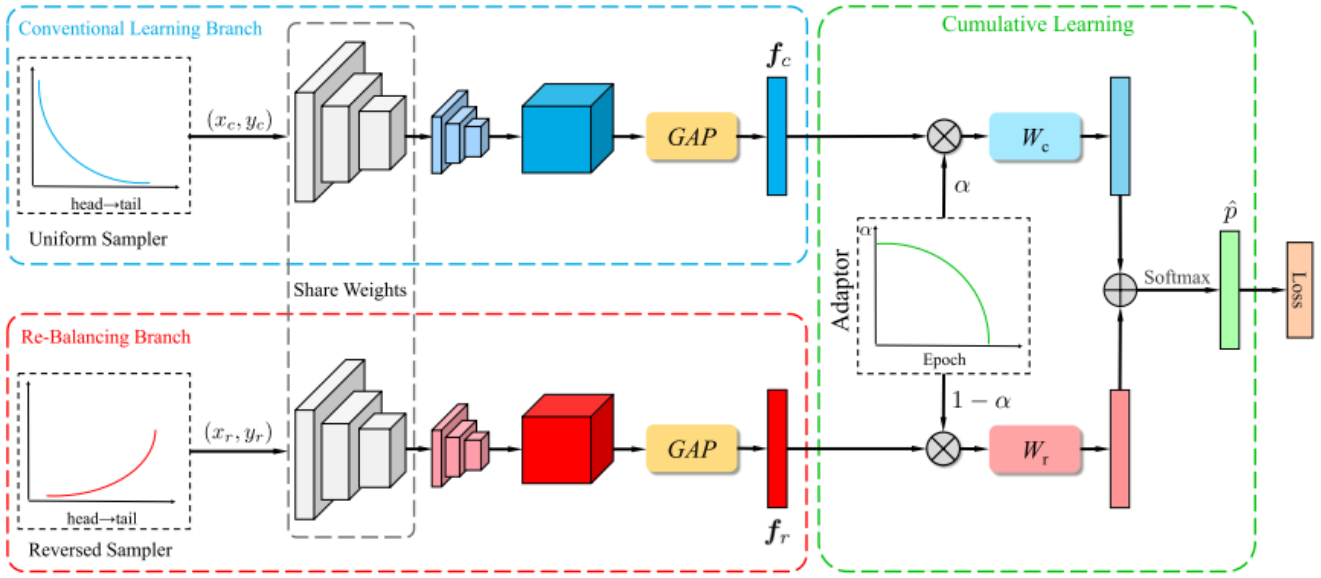


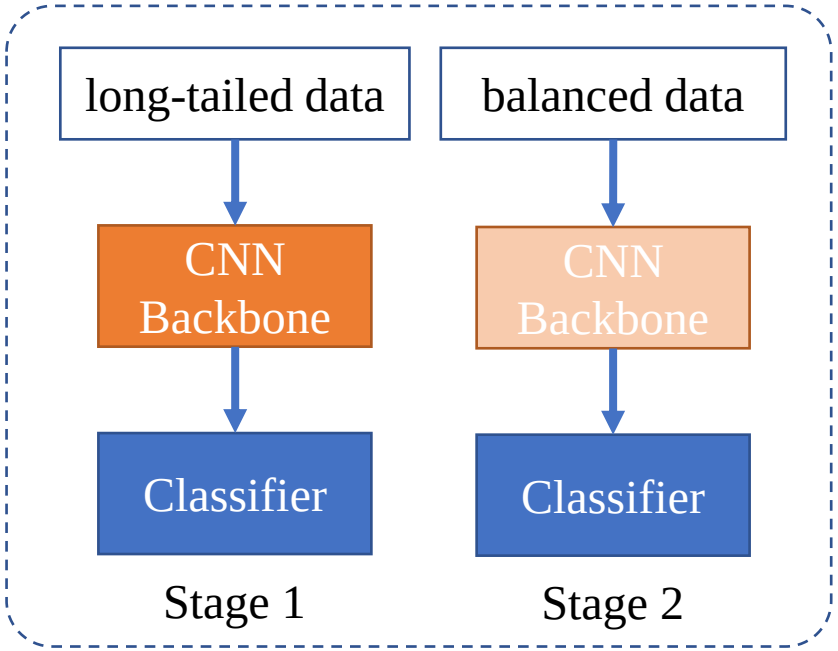
Two-Stage Re-balancing

- 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

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

Back To The Two-Stage SOTAs

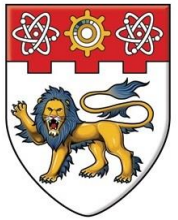
What's the problem of existing two-stage solutions?

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



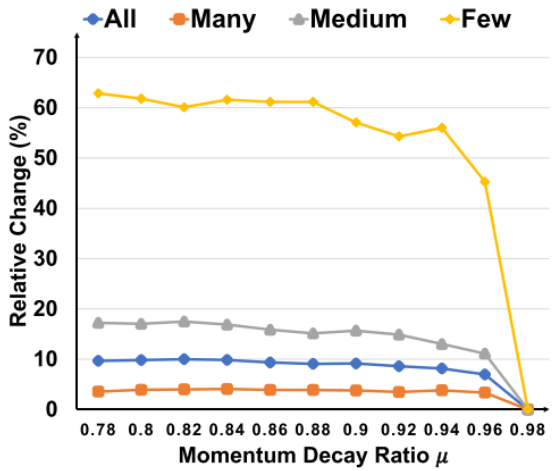
NANYANG TECHNOLOGICAL UNIVERSITY

Accumulative Momentum Effect

- The PyTorch implementation of SGD with momentum [8]:

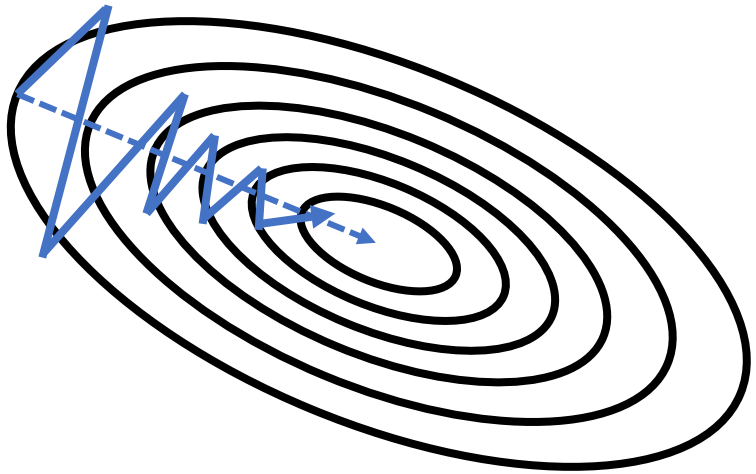
$$v_t = \underbrace{\mu \cdot v_{t-1}}_{momentum} + g_t, \qquad \theta_t = \theta_{t-1} - lr \cdot v_t,$$

- The moving average momentum will encode the data distribution, that creates a shortcut towards the head.

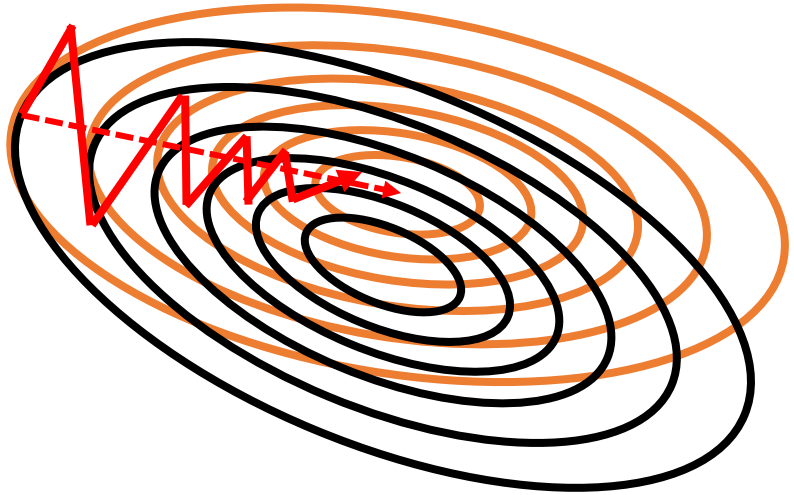


[8] https://pytorch.org/docs/stable/_modules/torch/optim/sgd.html

Accumulative Momentum Effect



SGD Momentum in
Balanced Dataset



SGD Momentum in
Long-Tailed Dataset

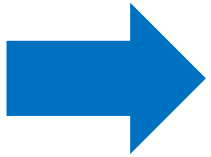
- Global Optima for All Categories
- Local Optima for Head Categories
- Momentum Direction in Balanced Data
- Momentum Direction in Long-Tailed Data

Causal Effect of Momentum

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

Remove Momentum:

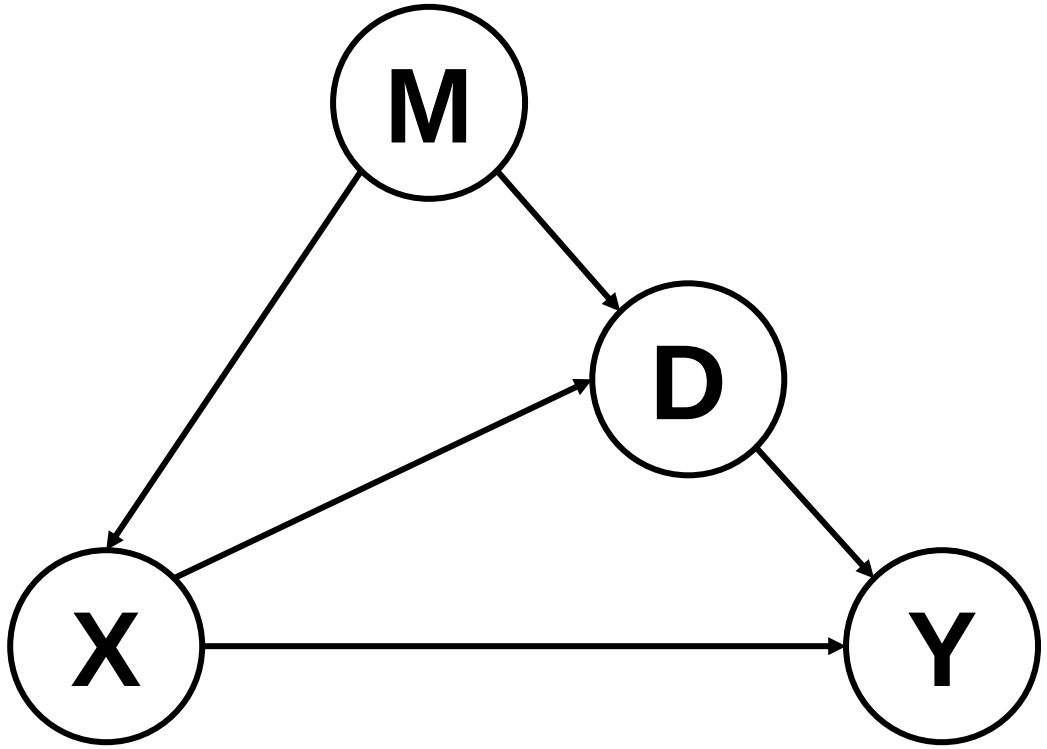
1. **Unstable Gradient**
2. **Local Optima**
3. **SGD Still Accumulates**



Keep Momentum in Training

Remove Bad Causal Effect

The Proposed Causal Graph



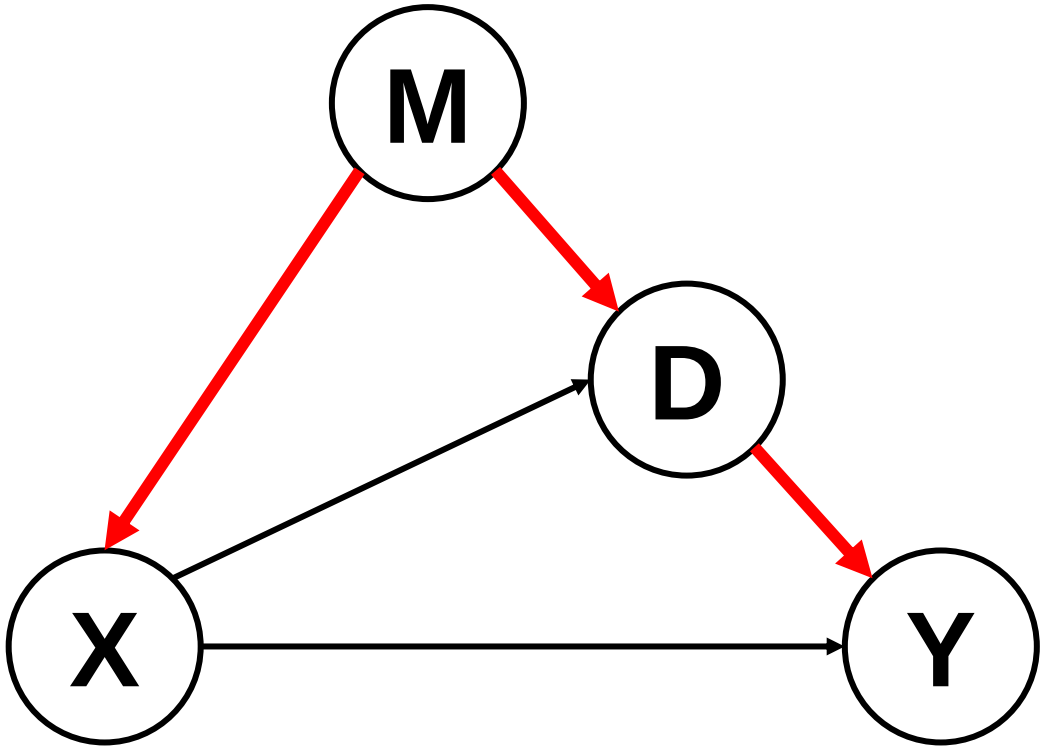
X : Feature

Y : Prediction

M: Momentum

D : Projection on Head

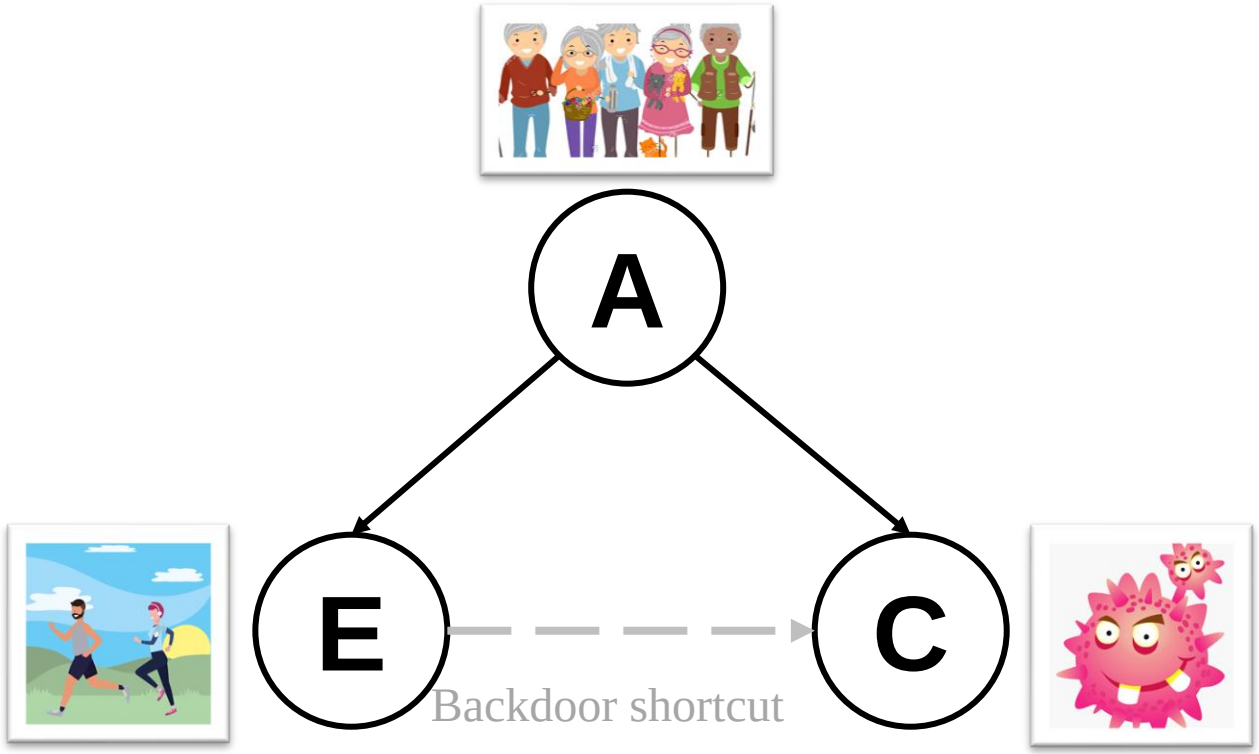
Two Undesired Causal Effects of Momentum



Two Undesired Causal Effects of Momentum:

1. Backdoor shortcut
2. Indirect Mediator Effect

Confounder and backdoor shortcut

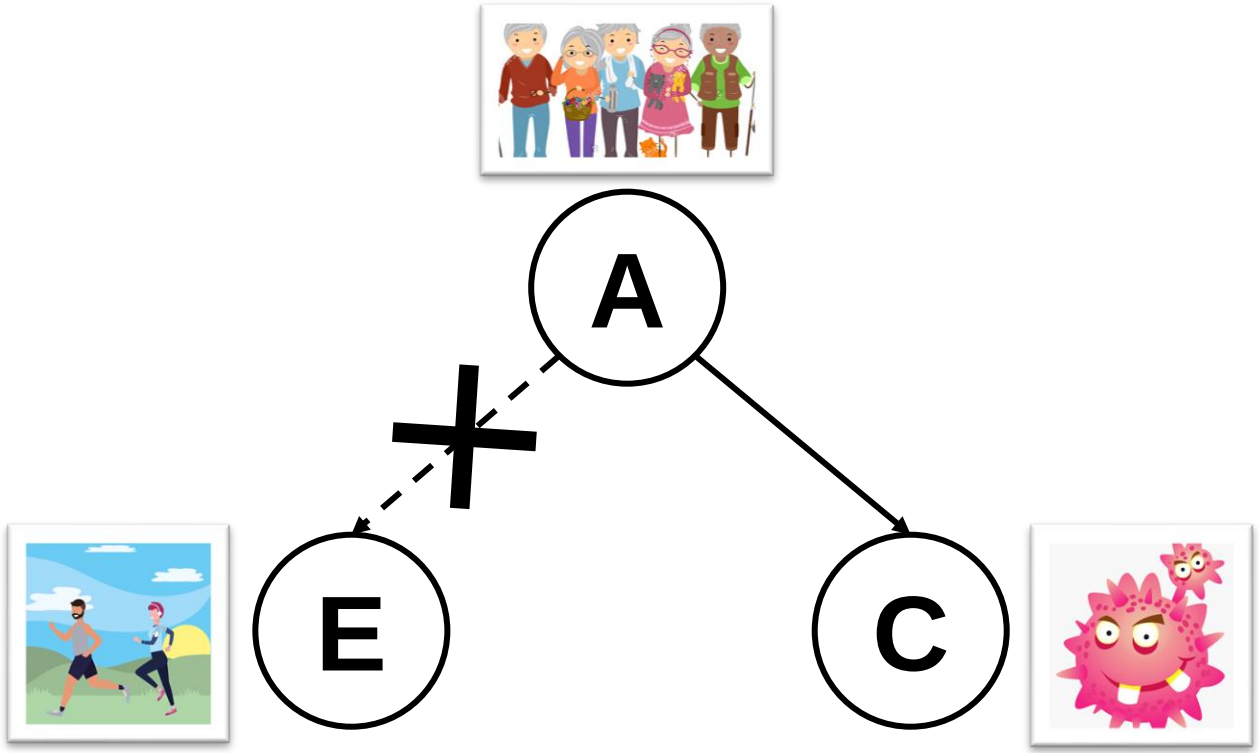


A: age **E: exercise** **C: cancer**

Backdoor shortcut:

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

backdoor Adjustment



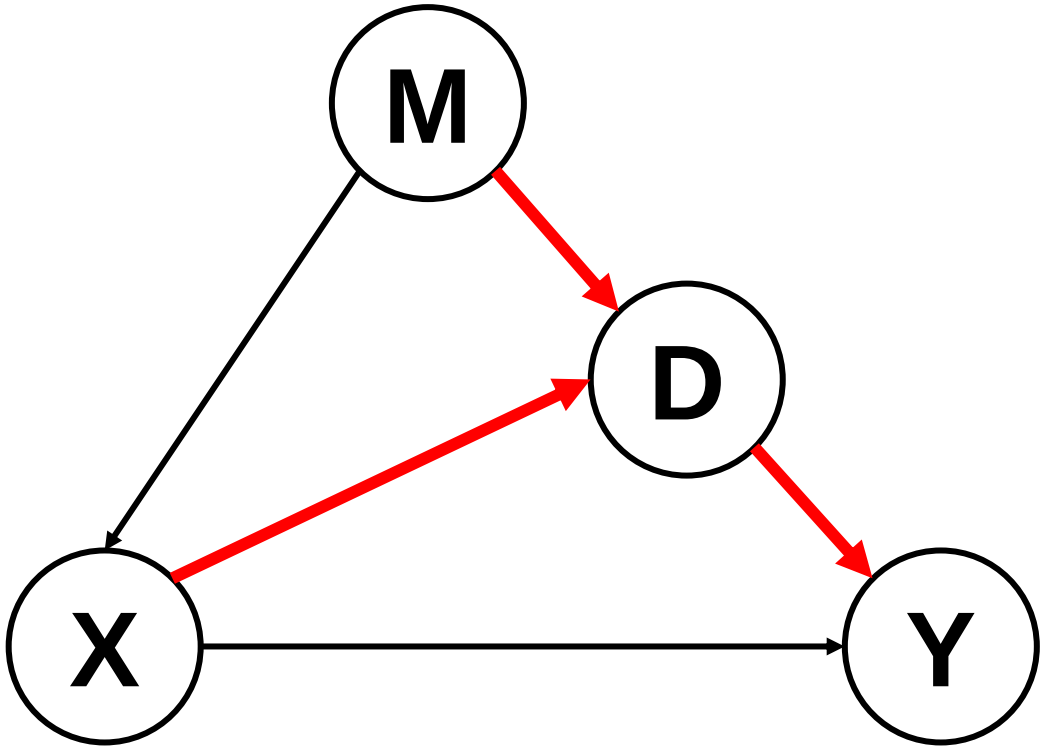
A: age E: exercise C: cancer

Backdoor Adjustment

Intervention on E:

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

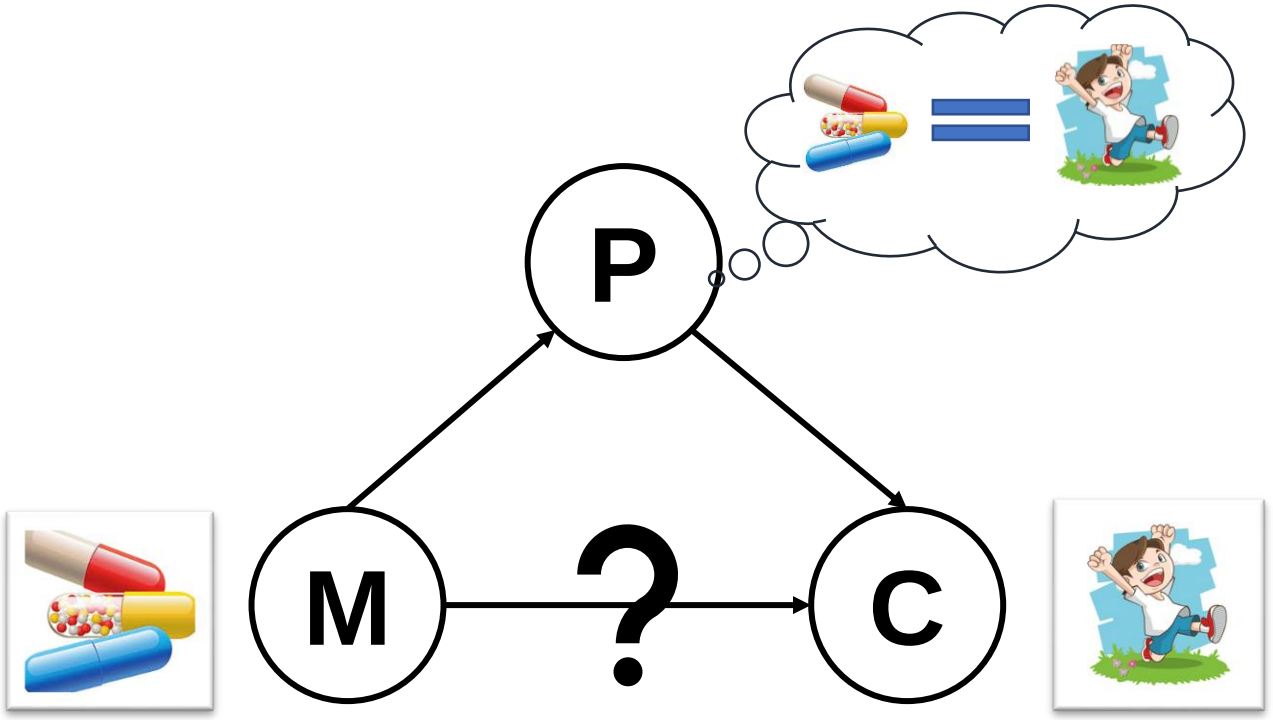
Two Undesired Causal Effects of Momentum



Two Undesired Causal Effects of Momentum:

1. Backdoor shortcut
2. Indirect Mediator Effect

Mediator and indirect effect

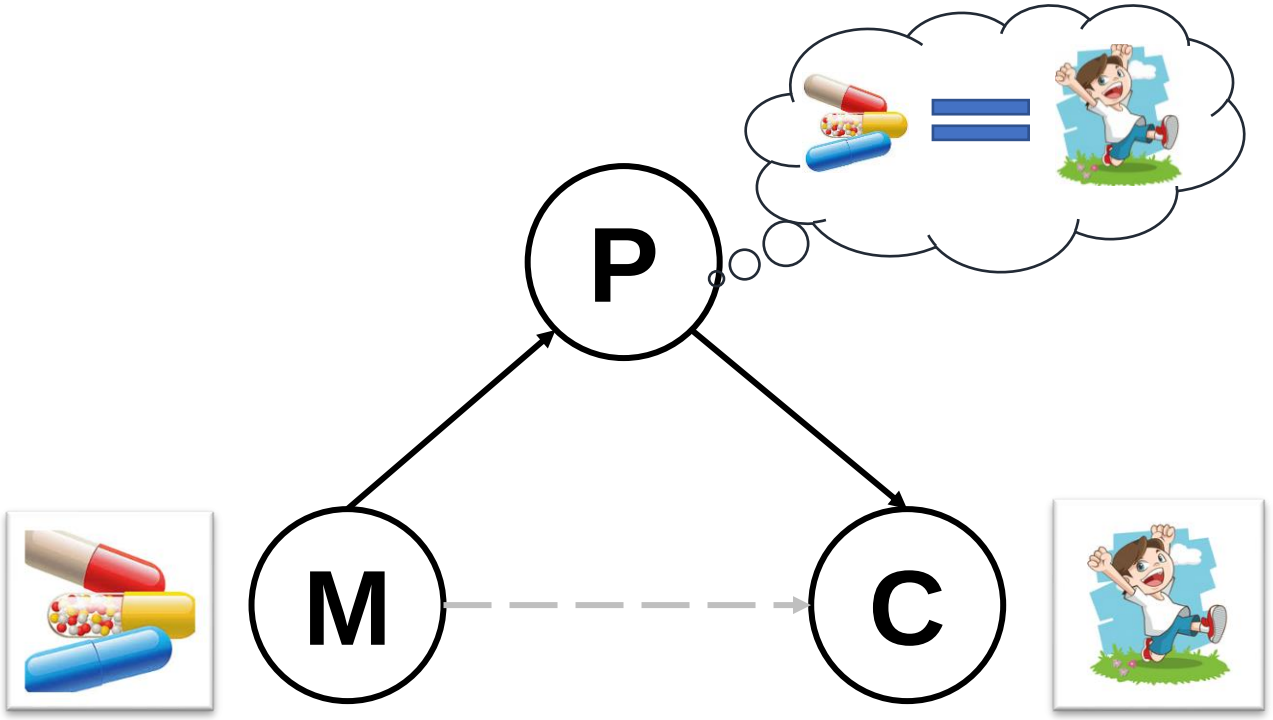


M: medicine P: placebo C: cure

Indirect effect:

1. $M \Rightarrow P$
2. $P \Rightarrow C$
3. $M \Rightarrow ? C$

Removing Placebo effect



Placebo effect:

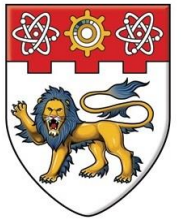
$$C(M = m_0, P = p)$$

Control Group

M: medicine P: placebo C: cure

Contents

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

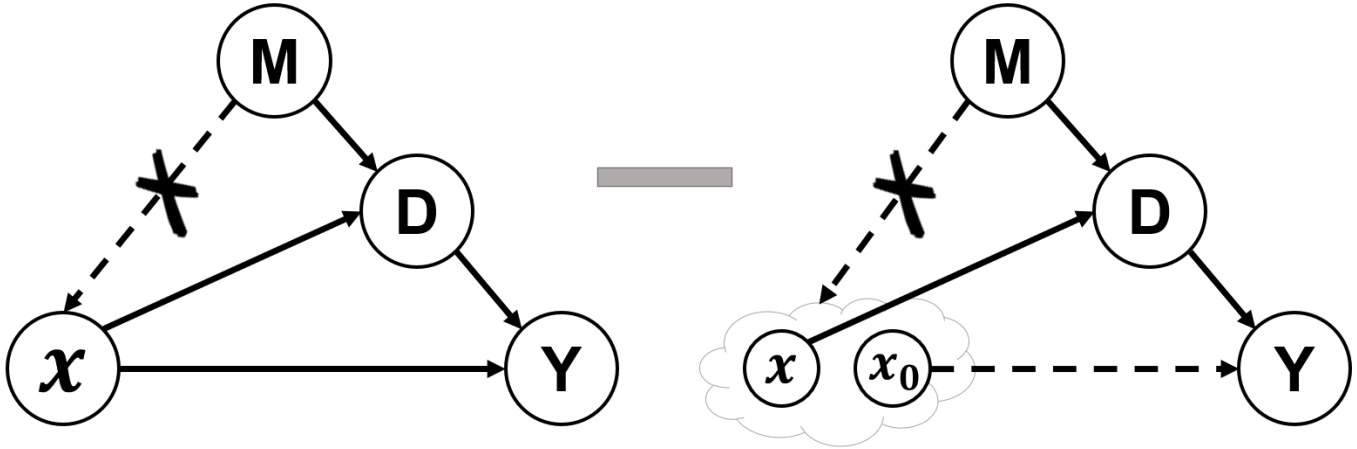


NANYANG TECHNOLOGICAL UNIVERSITY

De-confound TDE Classifier

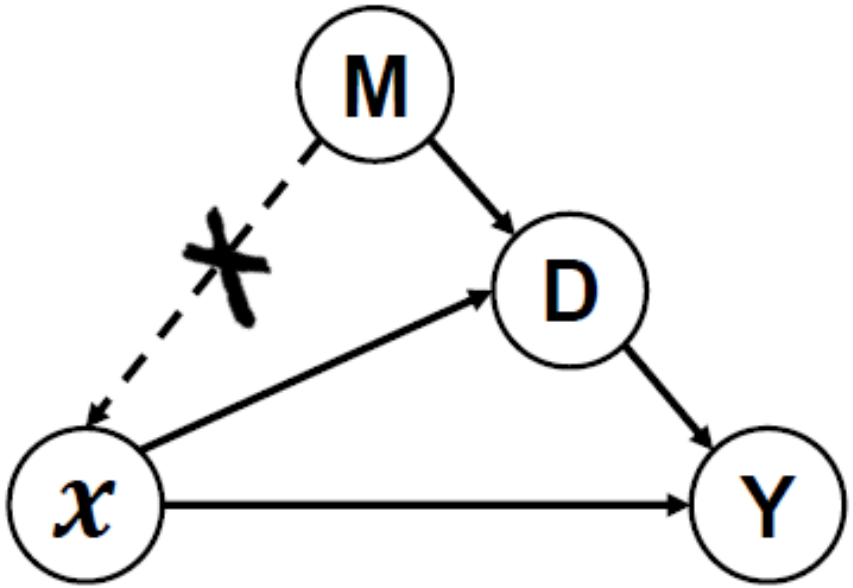
The definition of Total Direct Effect (TDE):

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



The proposed classifier = De-confounded Training + TDE Inference

De-confounded training (An Open Question)



- Approximation of the backdoor adjustment:

$$P(Y = i|do(X = x)) = \sum_{m} P(Y = i|X = x, M = m)P(M = m)$$

Inverse Probability Weighting [8]:

1. Skip the prohibitive P(M)
2. Applying the propensity score to reduce the confounding effects [9]

[8] Judea Pearl, et.al., Causal inference in statistics: A primer. 2016.
[9] Austin, Peter C. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." *Multivariate behavioral research*.

Inverse Probability Weighting

1. Skip the prohibitive P(M):
   - When there are infinite possible values of the confounder M, if we can only observe one (i, x, d) given one certain m, we can assume the number of possible m values is equal to the number of (i, x, d) samples, i.e., **the values of variables $(\mathbf{i}, \boldsymbol{x}, \boldsymbol{d})$ and $\boldsymbol{m}$ are one-to-one mapping**, we can skip the P(M) when we condition on X and D.

   - **Explanation of one-to-one mapping between (X, D) and M**:
     - Fix the random seed of model initialization.
     - Fix the hyper-parameters like learning rate, weight decay, etc.
     - Momentum M is the accumulated gradient of all past learning samples, which contains all the information of dataset and sampling strategies, i.e., M has one-to-one mapping with backbone parameters that used to generate feature X (which contains head deviation D).

   - We also adopt multi-head K on feature X, which means K times sampling on (X, D) for better approximation to M.
   - $\sum_m P(Y = i|X = x, M = m)P(M = m) \approx \frac{1}{K}\sum_{k=1}^{K} \tilde{P}(Y = i, X = x^k, D = d^k)$

Inverse Probability Weighting

2. Applying the propensity score to reduce the confounding effects:

- Although skipping the prohibitive P(M) is important, it barely changes anything in our model.
- Definition of the effect $(X \to Y)$: the prediction logits $f\left(x^k, d^k; w_i^k\right)$.
- Definition of the propensity score: a normalizing term of the effect $g\left(x^k, d^k; w_i^k\right)$.

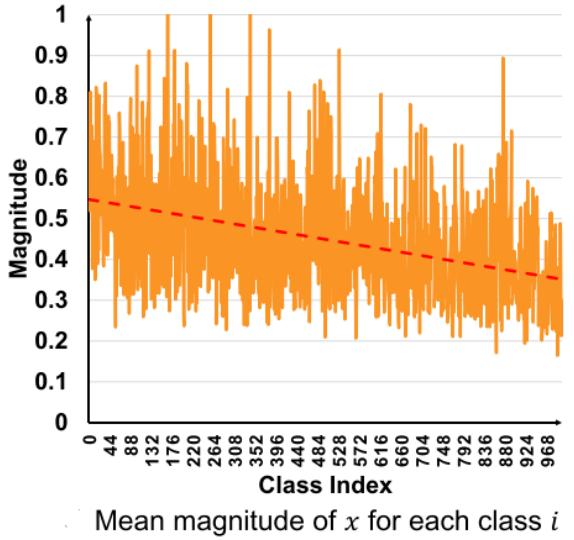
$$\tilde{P}\left(Y = i, X = x^k, D = d^k\right) \propto \tau \frac{f\left(x^k, d^k; w_i^k\right)}{g\left(x^k, d^k; w_i^k\right)}$$

De-confounded Training

- Logit of $P(Y = i|do(X = x))$ in the training phase is:

$$[Y = i|do(X = x)] = \frac{\tau}{K}\sum_{k=1}^{K}\frac{\left(w_i^k\right)^T(\dot{x}^k + d^k)}{\|x^k\|\cdot\|w_i^k\| + \gamma\|x^k\|}$$

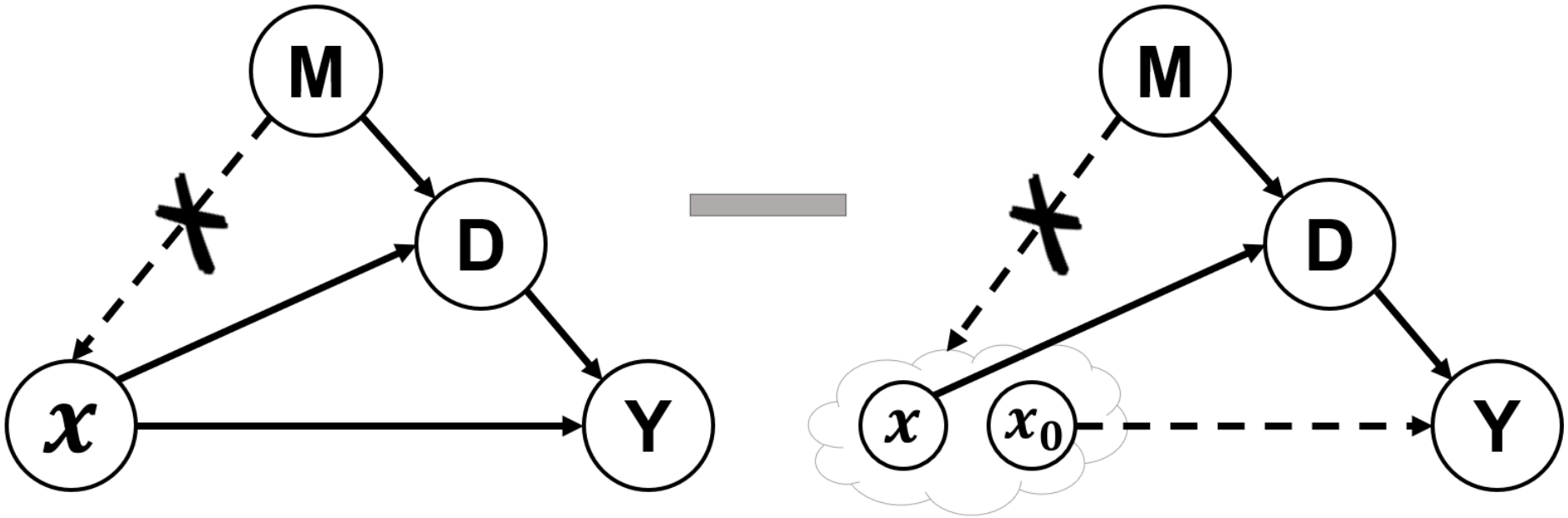
$$= \frac{\tau}{K}\sum_{k=1}^{K}\frac{\left(w_i^k\right)^T x^k}{(\|w_i^k\| + \gamma)\|x^k\|}$$



Mean magnitude of x for each class i

- Why not use cosine classifier (i.e., $g\left(x^k, d^k; w_i^k\right) = \|w_i^k\| \cdot \|x^k\|$) ?
- The proposed normalization term $g\left(x^k, d^k; w_i^k\right) = \|w_i^k\| \cdot \|x^k\| + \gamma\|x^k\|$

De-confound TDE Inference

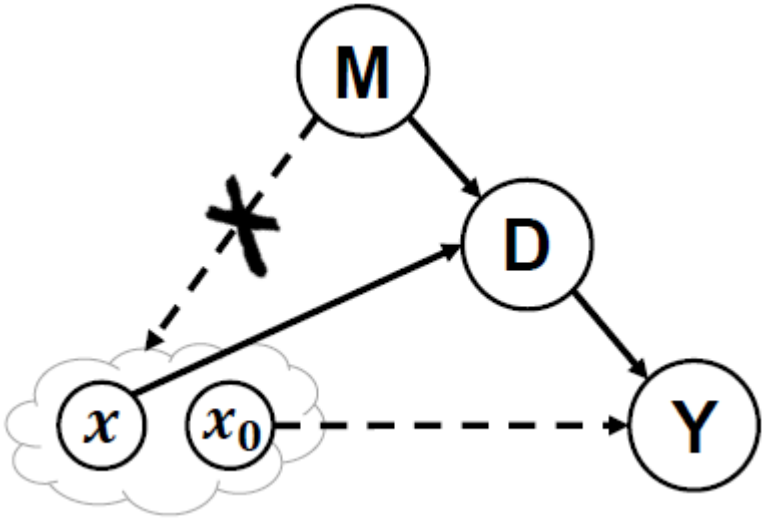


$$TDE(Y_i) = \frac{\tau}{K}\sum_{k=1}^{K}\left(\frac{\left(w_i^k\right)^T x^k}{\left(\left\|w_i^k\right\| + \gamma\right)\|x^k\|} - \alpha \cdot \frac{\cos\left(x^k, \hat{d}^k\right) \cdot \left(w_i^k\right)^T \hat{d}^k}{\left(\left\|w_i^k\right\| + \gamma\right)}\right)$$

The Counterfactual Bias (Placebo Effect)

$$[Y = i|do(X = x)] = \frac{\tau}{K}\sum_{k=1}^{K}\frac{\left(w_i^k\right)^T(x^k+d^k)}{\|x^k\|\cdot\|w_i^k\|+\gamma\|x^k\|}$$

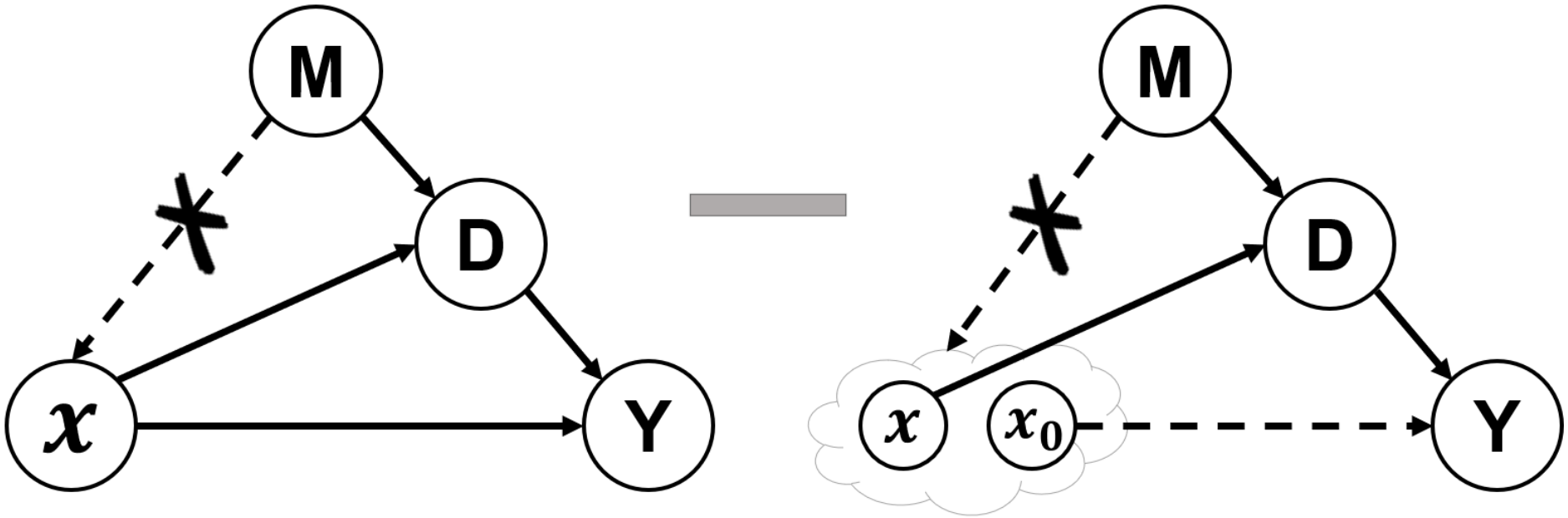
$\ddot{x}^k \Rightarrow x_0$ (zero vector)



$$[Y_d = i|do(X = x_0)] = \frac{\tau}{K}\sum_{k=1}^{K}\frac{\left(w_i^k\right)^T d^k}{(\|w_i^k\|+\gamma)\|x^k\|}$$, where $d = \|d\|\cdot\hat{d} = \cos(x,\hat{d})\cdot\|x\|\cdot\hat{d}$

$$= \frac{\tau}{K}\sum_{k=1}^{K}\frac{\cos(x^k,\hat{d}^k)\cdot\left(w_i^k\right)^T\hat{d}^k}{(\|w_i^k\|+\gamma)}$$

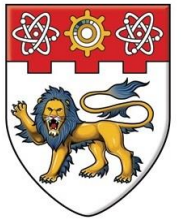
De-confound TDE Inference



$$TDE(Y_i) = \frac{\tau}{K}\sum_{k=1}^{K}\left(\frac{\left(w_i^k\right)^T x^k}{\left(\left\|w_i^k\right\| + \gamma\right)\|x^k\|} - \alpha \cdot \frac{\cos\left(x^k, \hat{d}^k\right) \cdot \left(w_i^k\right)^T \hat{d}^k}{\left(\left\|w_i^k\right\| + \gamma\right)}\right)$$

Contents

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



NANYANG TECHNOLOGICAL UNIVERSITY

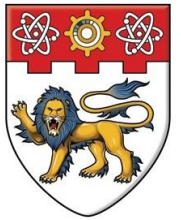
Advantages

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



NANYANG TECHNOLOGICAL UNIVERSITY

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**

Cosine Classifier:

$$g\left(x^k, d^k; w_i^k\right) = \left\|x^k\right\| \cdot \left\|w_i^k\right\|$$

Capsule Classifier:

$$g\left(x^k, d^k; w_i^k\right) = \left\|x^k\right\| \cdot \left\|w_i^k\right\| + \left\|w_i^k\right\|$$

Image Classification: ImageNet-LT

- Does the improvement come from multi-head trick?

Methods	#heads K	Many-shot	Medium-shot	Few-shot	Overall
Cosine† [5, 6]	1	67.3	41.3	14.0	47.6
Cosine† [5, 6]	2	67.5	42.1	14.1	48.1
Capsule† [8, 10]	1	67.1	40.0	11.2	46.5
Capsule† [8, 10]	2	67.7	41.3	12.6	47.6
(Ours) De-confound	1	67.3	41.8	15.0	47.9
(Ours) De-confound	2	**67.9**	42.7	14.7	48.6
(Ours) Cosine-TDE	1	61.8	47.1	30.4	50.5
(Ours) Cosine-TDE	2	63.0	47.3	31.0	51.1
(Ours) Capsule-TDE	1	62.3	46.9	30.6	50.6
(Ours) Capsule-TDE	2	62.4	47.9	31.5	51.2
(Ours) De-confound-TDE	1	62.5	47.8	**32.8**	51.4
(Ours) De-confound-TDE	2	62.7	**48.8**	31.6	**51.8**

Image Classification: Long-Tailed CIFAR

- Will the improvement be consistent across different imbalance ratio?

Dataset	Long-tailed CIFAR-100			Long-tailed CIFAR-10		
Imbalance ratio	100	50	10	100	50	10
Focal Loss [28]	38.4	44.3	55.8	70.4	76.7	86.7
Mixup [56]	39.5	45.0	58.0	73.1	77.8	87.1
Class-balanced Loss [13]	39.6	45.2	58.0	74.6	79.3	87.1
LDAM [12]	42.0	46.6	58.7	77.0	81.0	88.2
BBN [10]	42.6	47.0	59.1	79.8	82.2	88.3
(Ours) De-confound	40.5	46.2	58.9	71.7	77.8	86.8
(Ours) De-confound-TDE	**44.1**	**50.3**	**59.6**	**80.6**	**83.6**	**88.5**

Detection & Instance Segmentation

- Experiment Results on LVIS V0.5/V1.0 Val

Methods	LVIS Version	AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f	AP_{bbox}
Focal Loss† [28]	V0.5	21.1	32.1	22.6	3.2	21.1	28.3	22.6
(2019 Winner) EQL [17]	V0.5	24.9	37.9	26.7	10.3	27.3	27.8	27.9
Baseline	V0.5	22.6	33.5	24.4	2.5	23.0	30.2	24.3
Cosine† [50, 51]	V0.5	25.0	37.7	27.0	9.3	25.5	30.8	27.1
Capsule† [9, 54]	V0.5	25.4	37.8	27.4	8.5	26.4	**31.0**	27.1
(Ours) De-confound	V0.5	25.7	38.5	27.8	11.4	26.1	30.9	27.7
(Ours) Cosine-TDE	V0.5	28.1	42.6	30.2	20.8	28.7	30.3	30.6
(Ours) Capsule-TDE	V0.5	**28.4**	42.1	**30.8**	21.1	**29.7**	29.6	30.4
(Ours) De-confound-TDE	V0.5	**28.4**	**43.0**	30.6	**22.1**	29.0	30.3	**31.0**
Baseline	V1.0	21.8	32.7	23.2	1.1	20.9	31.9	23.9
(Ours) De-confound	V1.0	23.5	34.8	25.0	5.2	22.7	**32.3**	25.8
(Ours) De-confound-TDE	V1.0	**27.1**	**40.1**	**28.7**	**16.0**	**26.9**	32.1	**30.0**

Background-Exempted Inference

- Background-Exempted Inference (**when there are good heads**):

$$\arg\max_{i \in C} \begin{cases} (1 - p_0) \cdot \frac{q_i}{1 - q_0} & i \neq 0 \\ p_0 & i = 0 \end{cases},$$

where $i = 0$ is the background category, $p_i = P(Y = i|do(X = x))$, q_i is the probability of original TDE.

Methods	BG-Exempted	AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f	AP_{bbox}
De-confound	✗	25.7	38.5	27.8	11.4	26.1	**30.9**	27.7
De-confound-TDE	False	23.4	35.7	24.9	13.1	23.6	27.1	24.8
De-confound-TDE	True	**28.4**	**43.0**	**30.6**	**22.1**	**29.0**	30.3	**31.0**

Grad-cam Visualization on ImageNet-LT

What does our model see from images?

NANYANG TECHNOLOGICAL UNIVERSITY

DAMO
ALIBABA DAMO ACADEMY

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You

Code Link

Paper Link