南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Causal Learning in Machine Learning

Tang K , et al. Long-tailed classification by keeping the good and removing the bad momentum causal effect[J]. Advances in Neural Information Processing Systems, 2020, 33: 1513-1524.

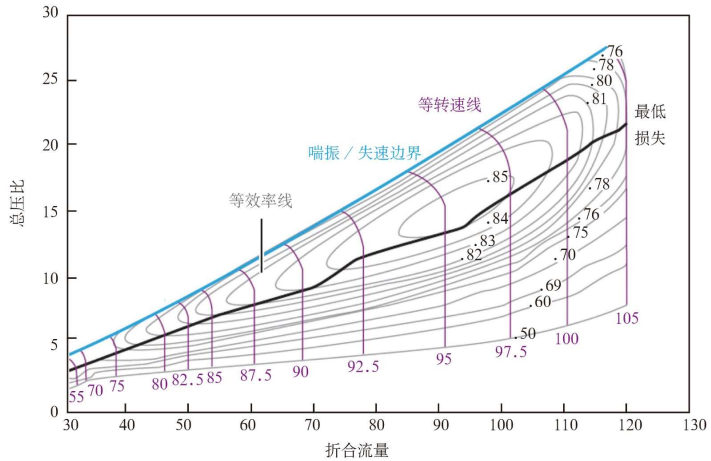
Hu X, et al. Distilling causal effect of data in class-incremental learning[C]//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. 2021: 3957-3966.

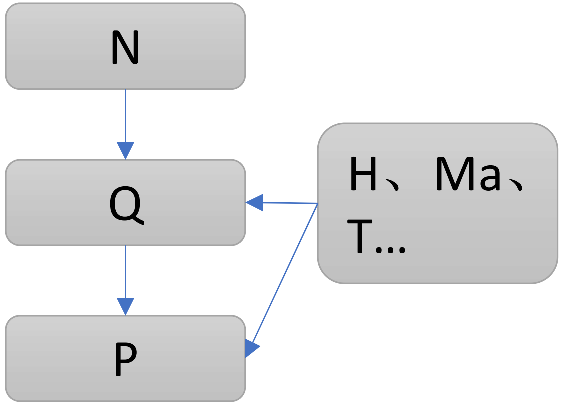
Motivation

As for **physical model**

- **Explainability**

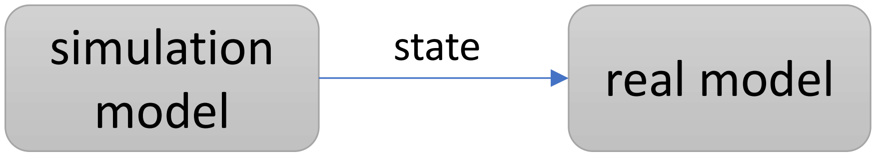
  We cannot understand AI model





- **Stability**

  We cannot trust AI model



Motivation

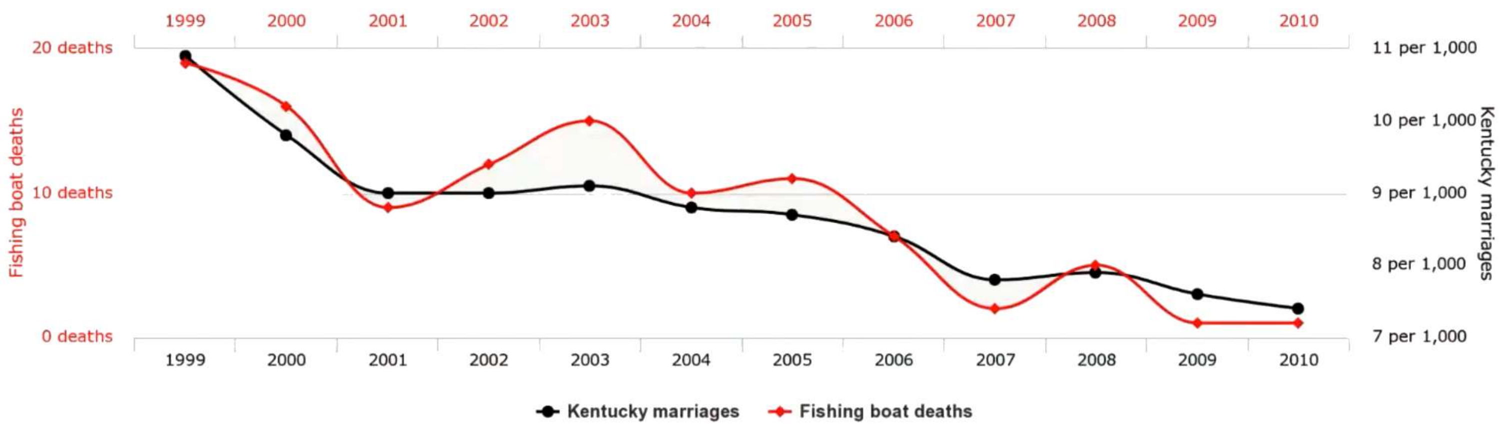
As for **physical model**

- **Explainability**

Correlation

↓

Causality

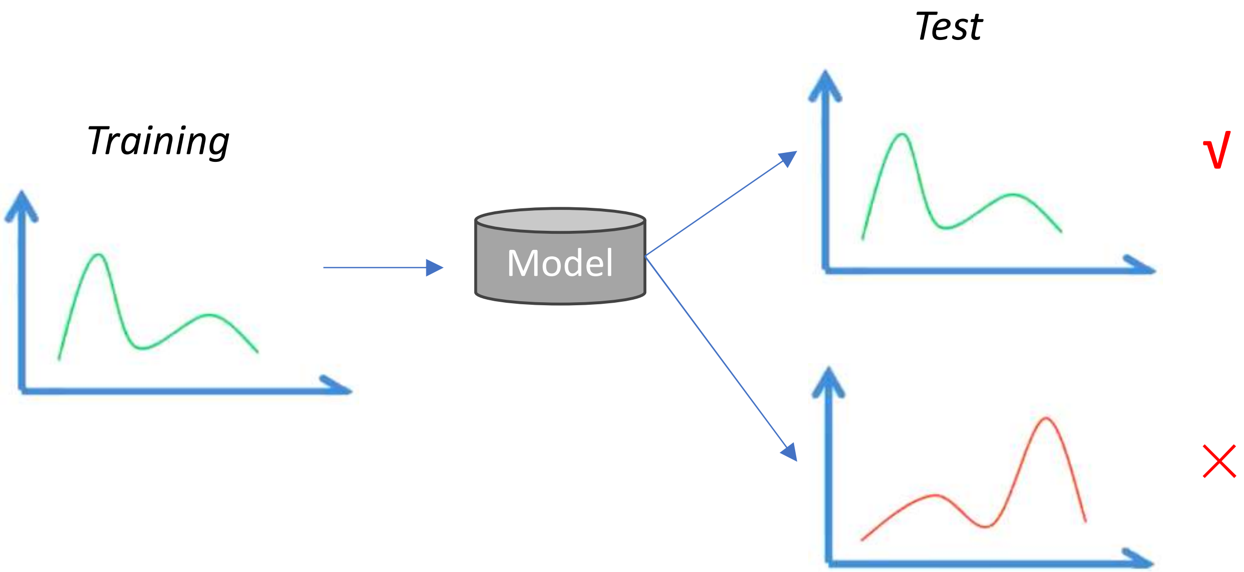


- **Stability**

I.I.D hypothesis

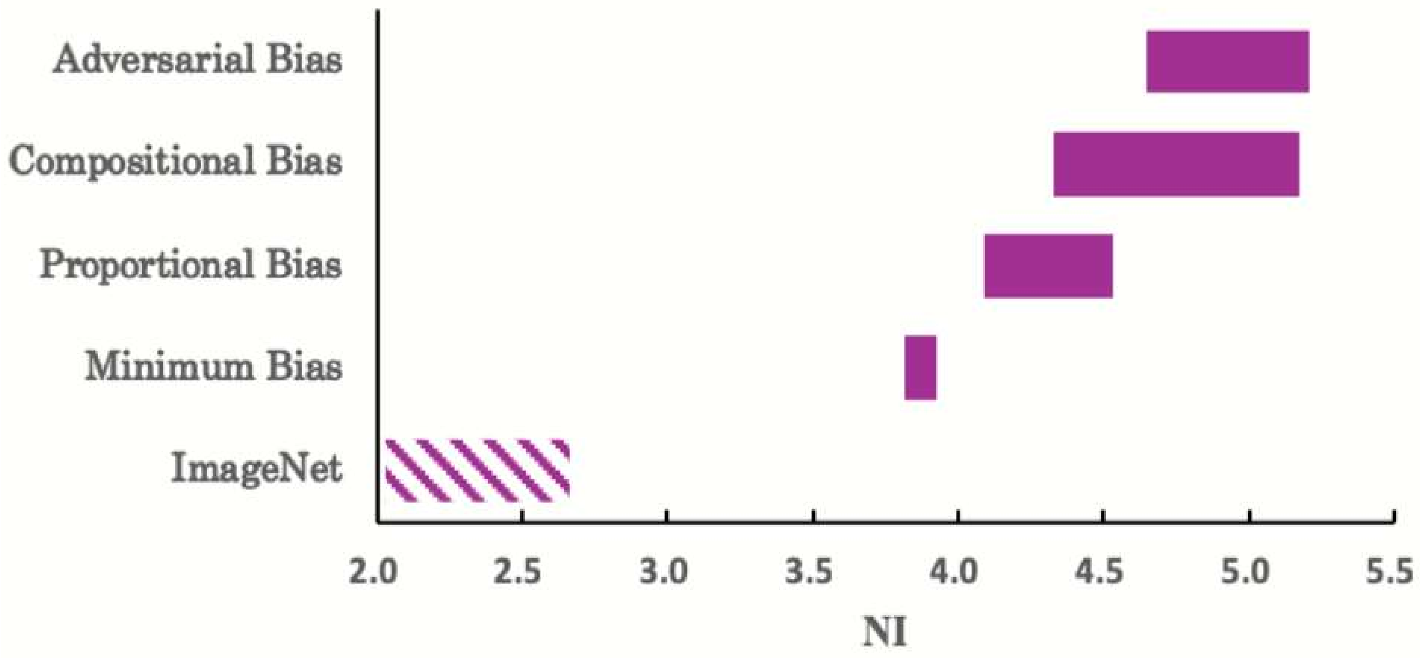
↓

Distribution Shift



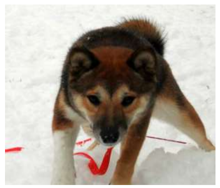
Motivation

As for **visual field**



NICO

- **spurious correlation：** grass --- dog label



- **mediation effect：** dog hair color --- dog label ✗ dog nose/ears --- dog label √

He Y , et al. Towards non-iid image classification: A dataset and baselines. Pattern Recognition, 2021.

Background

Causal Regularizer

Re-weighting to decorrelate features

Set feature *j* as treatment variable

$$\sum_{j=1}^{p}\left\|\frac{X_{-j}^{T}\cdot(W\odot I_j)}{W^T\cdot I_j}-\frac{X_{-j}^{T}\cdot(W\odot(1-I_j))}{W^T\cdot(1-I_j)}\right\|_2^2,$$

- X_{-j}^{T}: All features excluding treatment *j*
- W^T: Sample Weights
- I_j: Indicator of treatment status

Potential weakness

- destruction of characteristic interaction structure
- no specific causality involved

e.g. logistic regression

$$\begin{aligned}
\min\quad & \textstyle\sum_{i=1}^{n} W_i\cdot\log(1+\exp((1-2Y_i)\cdot(x_i\beta))),\\
s.t.\quad & \textstyle\sum_{j=1}^{p}\left\|\frac{X_{-j}^{T}\cdot(W\odot I_j)}{W^T\cdot I_j}-\frac{X_{-j}^{T}\cdot(W\odot(1-I_j))}{W^T\cdot(1-I_j)}\right\|_2^2\le\gamma_1,\\
& W\ge 0,\ \ \|W\|_2^2\le\gamma_2,\ \ \|\beta\|_2^2\le\gamma_3,\ \ \|\beta\|_1\le\gamma_4,\\
& (\textstyle\sum_{k=1}^{n} W_k-1)^2\le\gamma_5,
\end{aligned}$$

continuous variables case

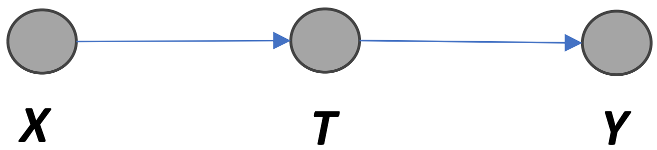
$$\sum_{j=1}^{p}\left\|\mathbf{X}_{,j}^{T}\Sigma_W\mathbf{X}_{,-j}/n-\mathbf{X}_{,j}^{T}W/n\cdot\mathbf{X}_{,-j}^{T}W/n\right\|_2^2$$

Zheyan Shen, et al. Causally Regularized Learning on Data with Agnostic Bias. ACM MM, 2018.
Kun Kuang, et al. Stable Prediction with Model Misspecification and Agnostie Distribution Shift. AAAI, 2020.

Background

Data relationships

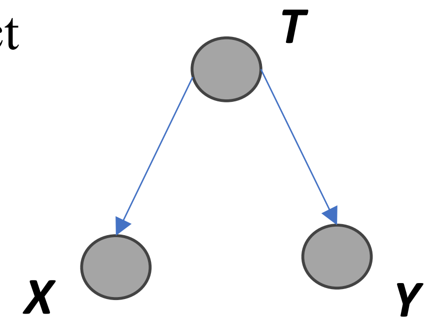
- **Chain**



T : **mediator**

Transmit causal effect

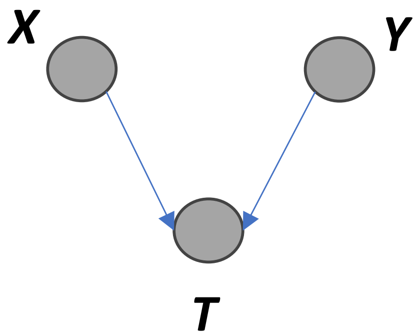
- **Fork**



T : **confounder**

Transmit causal effect

- **Collision**



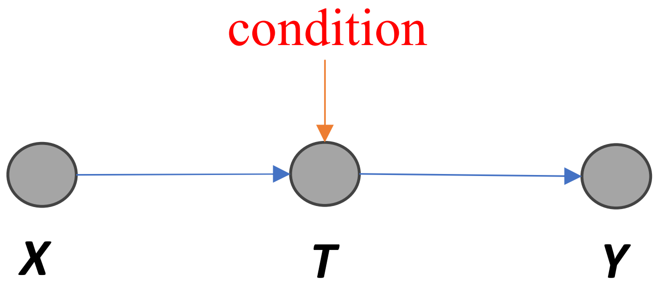
T : **collider**

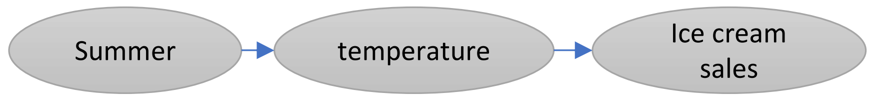
Block causal effect

Background

Data relationships

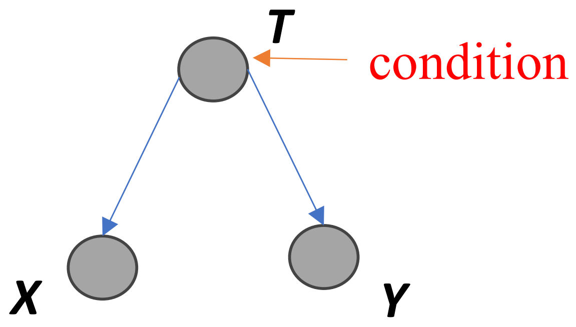
- **Chain**

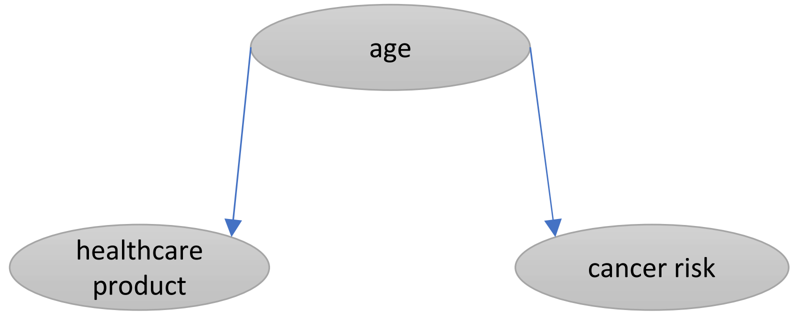




Block causal effect

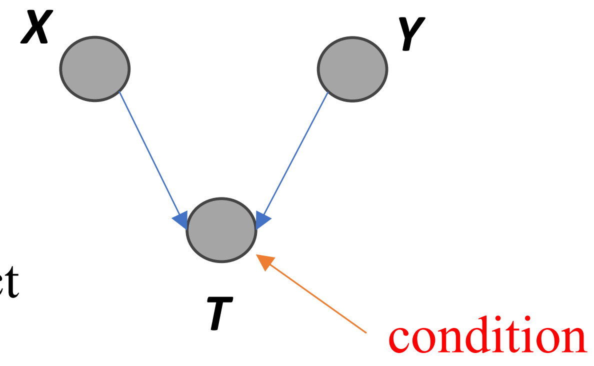
- **Fork**

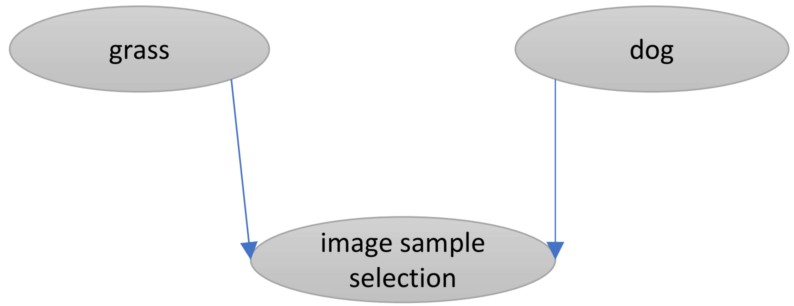




Block causal effect

- **Collision**



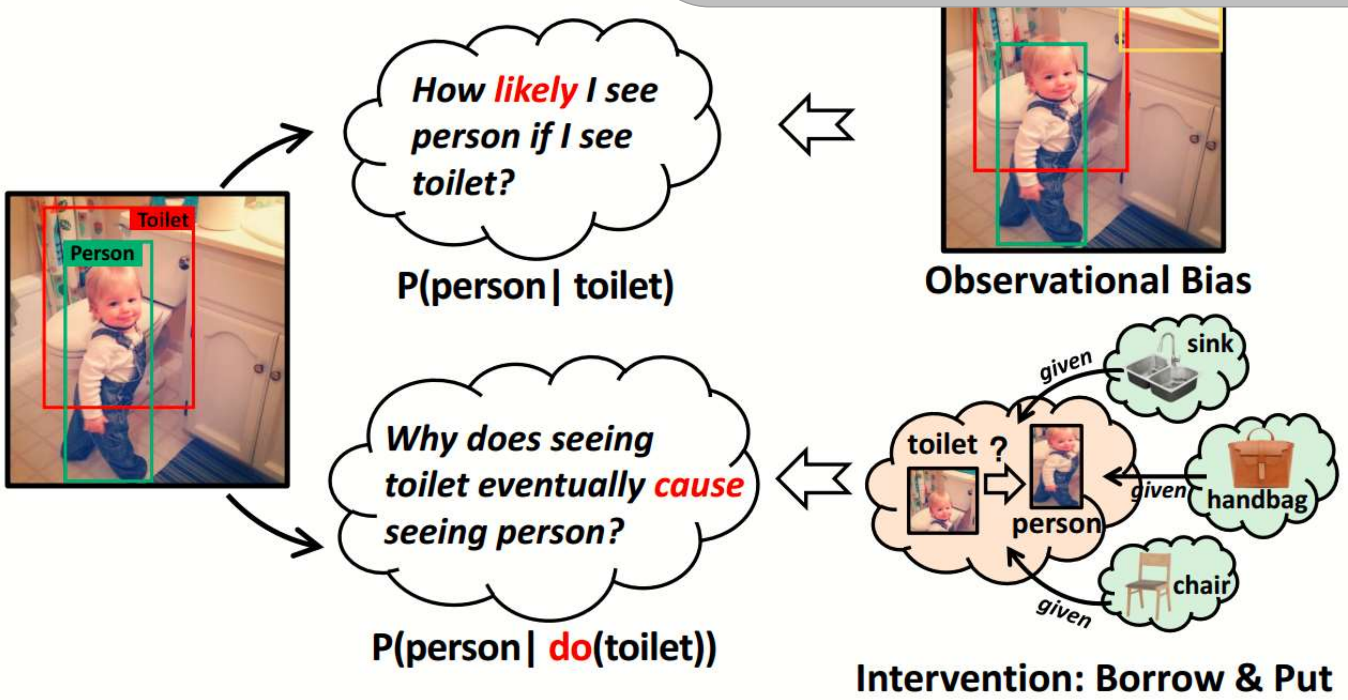


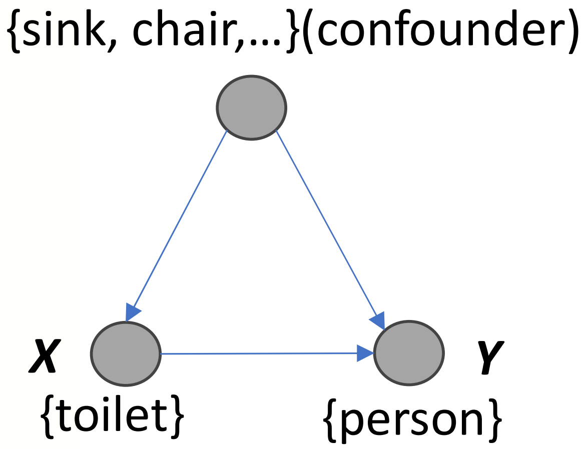
Transmit causal effect

Background

Pearl's theory of causal graph

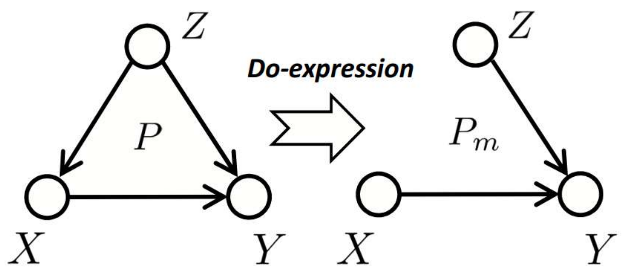
$$
\begin{aligned}
P(Y=y|do(X=x)) &= P_m(Y=y|X=x) \\
&= \sum_z P_m(Y=y|X=x,Z=z)P_m(Z=z|X=x) \\
&= \sum_z P_m(Y=y|X=x,Z=z)P_m(Z=z) \\
&= \sum_z P(Y=y|X=x,Z=z)P(Z=z)
\end{aligned}
$$





$$P(Y|X) = \sum_z P(Y|X,z)P(z|X) = \frac{P(Y,X)}{P(X)}$$

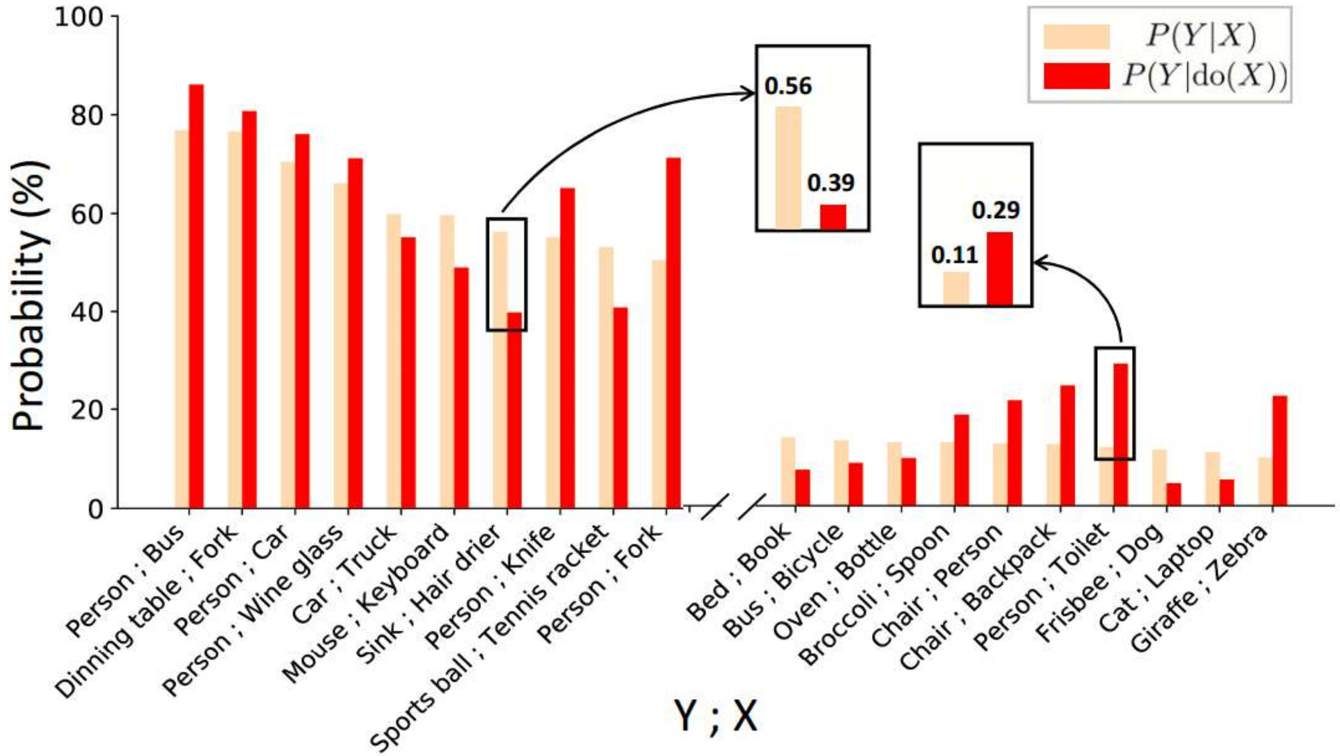
$$P(Y|\mathrm{do}(X)) = \sum_z P(Y|X,z)P(z) = \sum_z \frac{P(Y,X,z)P(z)}{P(X,z)}$$



Wang et al. Visual Commonsense R-CNN. CVPR, 2020

Background

Pearl's theory of causal graph



Simpson's Paradox: (from Wikipedia)

a phenomenon in probability and statistics in which a trend appears in several groups of data but **disappears or reverses** when the groups are combined.

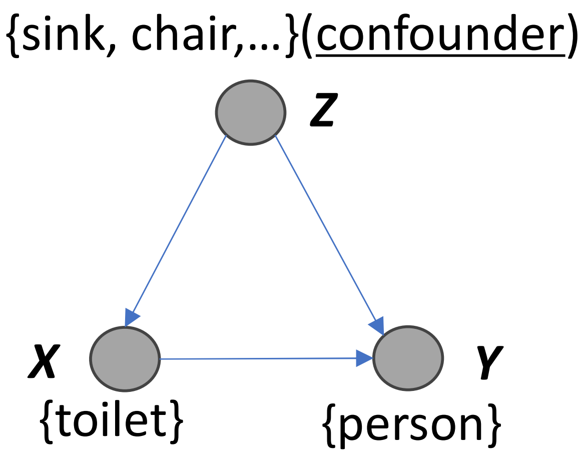
Wang et al. Visual Commonsense R-CNN. CVPR, 2020

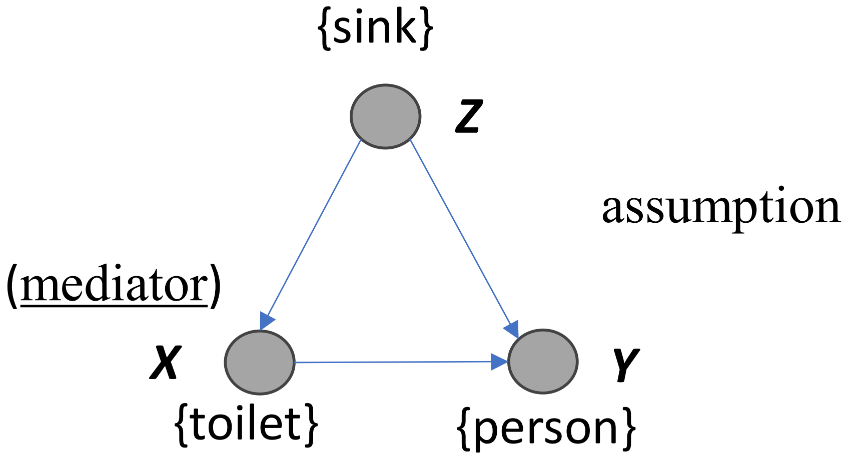
Some thoughts

From (Segmenting of confounds by clustering) -- decrease in effective samples/Pearson

Zheyean Shen ,et al.. Stable Learning via Differentiated Variable Decorrelation. KDD, 2020.

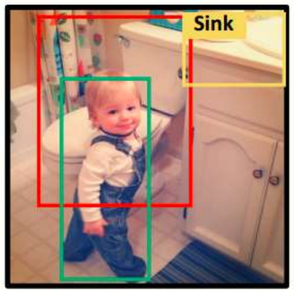
Confounder or Mediator is absolute truth?





Context (objects and backgrounds)

- purely bad (backgrounds) –de-confounder
- containing good and bad –TDE
- purely good (component)



How to define/or use relative effect to revise model

Case 1 Long-tailed visual recognition

Solid points:

- Shifting the Perspective of Intervention to **Momentum**
  - Constructing the causal graph --- **separating out the mediation**
  - **Potential contribution** on the long tail of features
- Compliance with the end-to-end learning framework
  - **De-confounder** training stage
  - Total Direct Effect (**TDE**) **inference** stage
- **Unifying** the long-tail learning methods at the time under the causal framework

Tang K , et al. Long-tailed classification by keeping the good and removing the bad momentum causal effect. NeurIPS, 2020

Case 1 Long-tailed visual recognition

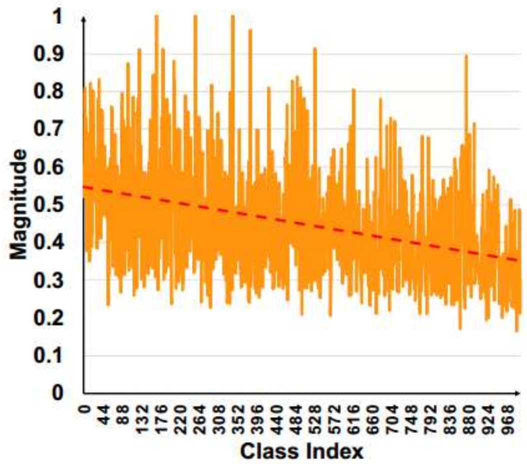
Constructing the causal graph :

Variables ： {X、M、D、Y}

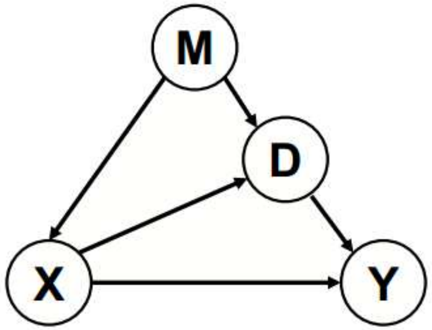
1) M→X：

The **backbone parameters** used to **generate** feature vectors X, are **trained under** the effect of M.

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



(b) Mean magnitude of x for each class i



Case 1 Long-tailed visual recognition

Constructing the causal graph :

Variables： {X、M、D、Y}

2) (M, X) → D：

$\ddot{\boldsymbol{x}}$ discriminative feature

$\boldsymbol{d}$ projection on head direction $\boldsymbol{d} = \hat{\boldsymbol{d}} cos(\boldsymbol{x}, \hat{\boldsymbol{d}}) \|\boldsymbol{x}\|$

the **unit vector** of
exponential moving average features

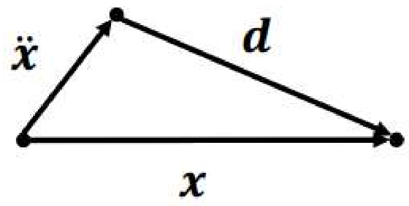
3) X → D → Y & X → Y：

Effect of X can be **disentangled** into
an indirect (mediator) and a direct effect.

By the way, Orthogonality is equivalent to disentanglement?

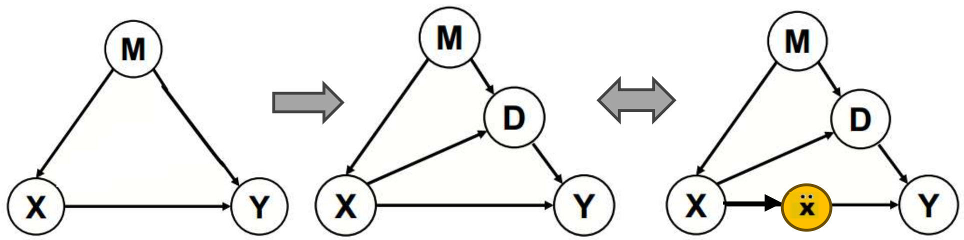
$$v_t = \underbrace{\mu \cdot v_{t-1}}_{momentum} + g_t, \quad \theta_t = \theta_{t-1} - lr \cdot v_t,$$

Orthogonal decomposition $\boldsymbol{x} = \ddot{\boldsymbol{x}} + \boldsymbol{d}$



$$\hat{\boldsymbol{d}} = \overline{\boldsymbol{x}}_T / \|\overline{\boldsymbol{x}}_T\|, \ where \ \overline{\boldsymbol{x}}_t = \mu \cdot \overline{\boldsymbol{x}}_{t-1} + \boldsymbol{x}_t$$

linear approximation

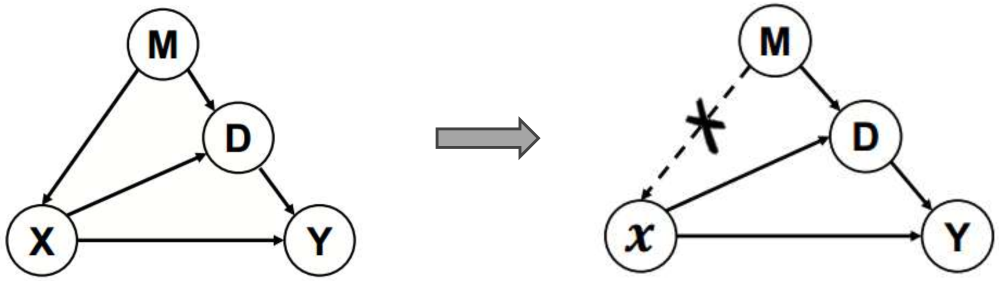


Case 1 Long-tailed visual recognition

How to realize the target:

a) de-confounder

- the Backdoor adjustment formula



- inverse probability weighting (IPW)

$$P(Y=y|do(X=x)) = \sum_z \frac{P(Y=y, X=x, Z=z)}{P(X=x|Z=z)}$$

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

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

- muti-head strategy splits the feature space

$$P(Y=i|do(X=\boldsymbol{x})) \approx \frac{1}{K}\sum_{k=1}^{K} \widetilde{P}(Y=i, X=\boldsymbol{x}^k|M=\boldsymbol{m}),$$

Sample

$$\widetilde{P}(Y=i, X=\boldsymbol{x}^k) \propto E(i, \boldsymbol{x}^k; \boldsymbol{w}_i^k) = \tau\frac{f(i, \boldsymbol{x}^k; \boldsymbol{w}_i^k)}{g(i, \boldsymbol{x}^k; \boldsymbol{w}_i^k)},$$

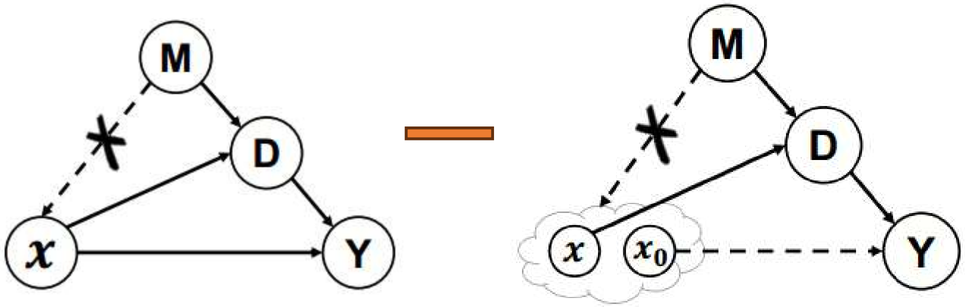
$$\|\boldsymbol{x}^k\| \cdot \|\boldsymbol{w}_i^k\| + \gamma\|\boldsymbol{x}^k\|$$

Case 1 Long-tailed visual recognition

How to realize the target:

b) TDE inference

not remove placebo effect totally



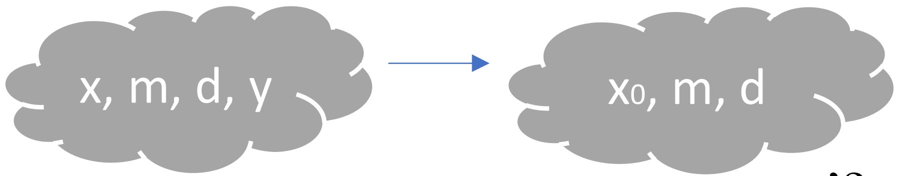
$$\arg\max_{i\in C} \; TDE(Y_i) = [Y_{\boldsymbol{d}} = i|do(X=\boldsymbol{x})] - [Y_{\boldsymbol{d}} = i|do(X=\boldsymbol{x}_0)],$$

$$\boldsymbol{x} = \ddot{\boldsymbol{x}} + \boldsymbol{d}$$

null

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

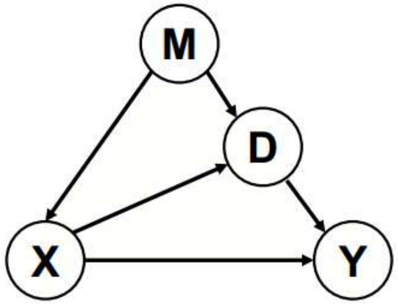
world1 world2



y'?

Case 1 Long-tailed visual recognition

Unifying the long-tail learning methods :



CDE: remove d $[Y \mid X = x, do(D = d)]$

NDE: get a fair d $[Y \mid X = x, D = d_0] - [Y \mid X = x_0, D = d_0]$

TDE: get d with good part $[Y_{D=d} \mid X = x)] - [Y_{D=d} \mid X = x_0]$

狮鹫（尾部类）=狮子（头部类）+鹰（头部类）

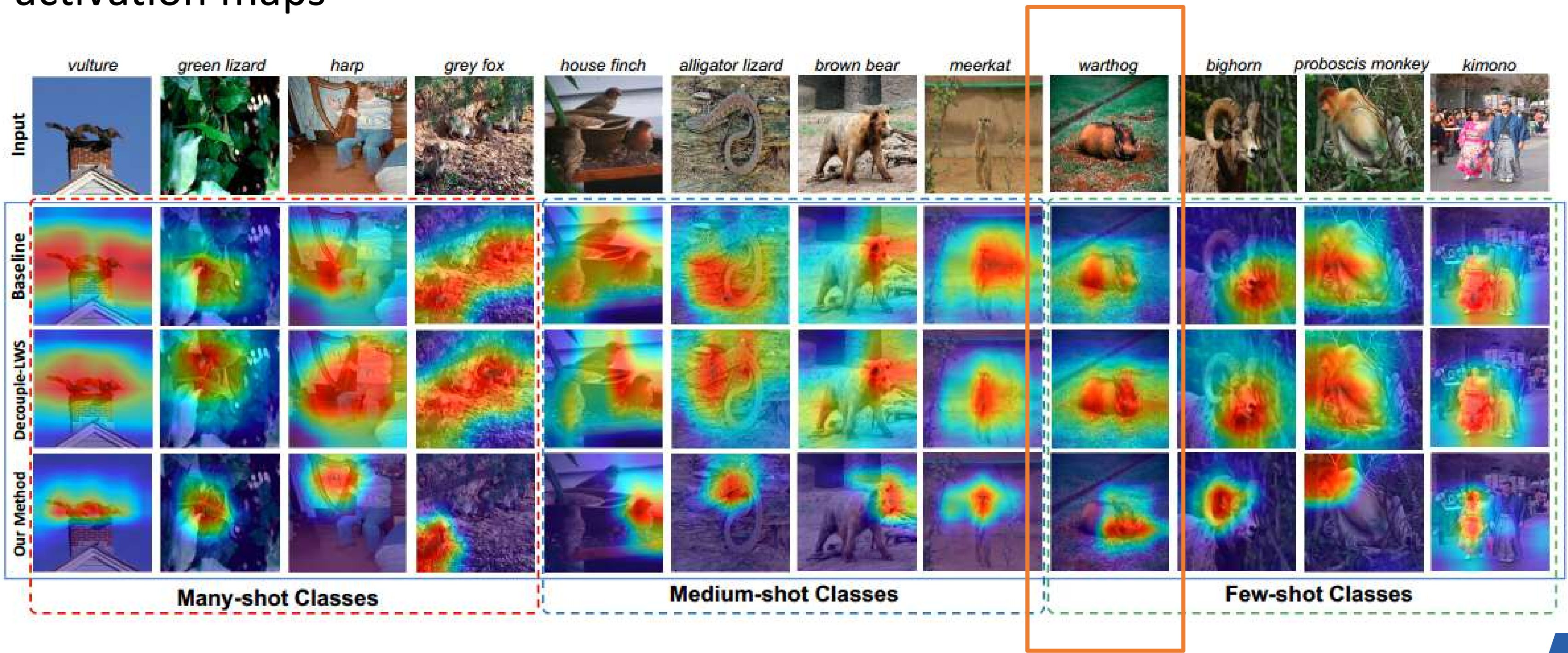
Case 1 Long-tailed visual recognition

Experiments:

- Top-1 accuracy

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**

- Visualized activation maps

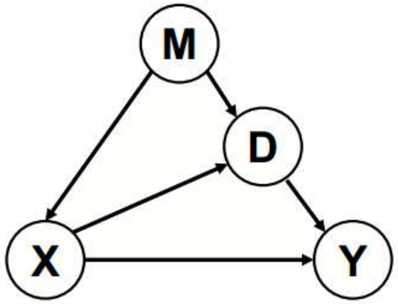


Case 1 Long-tailed visual recognition

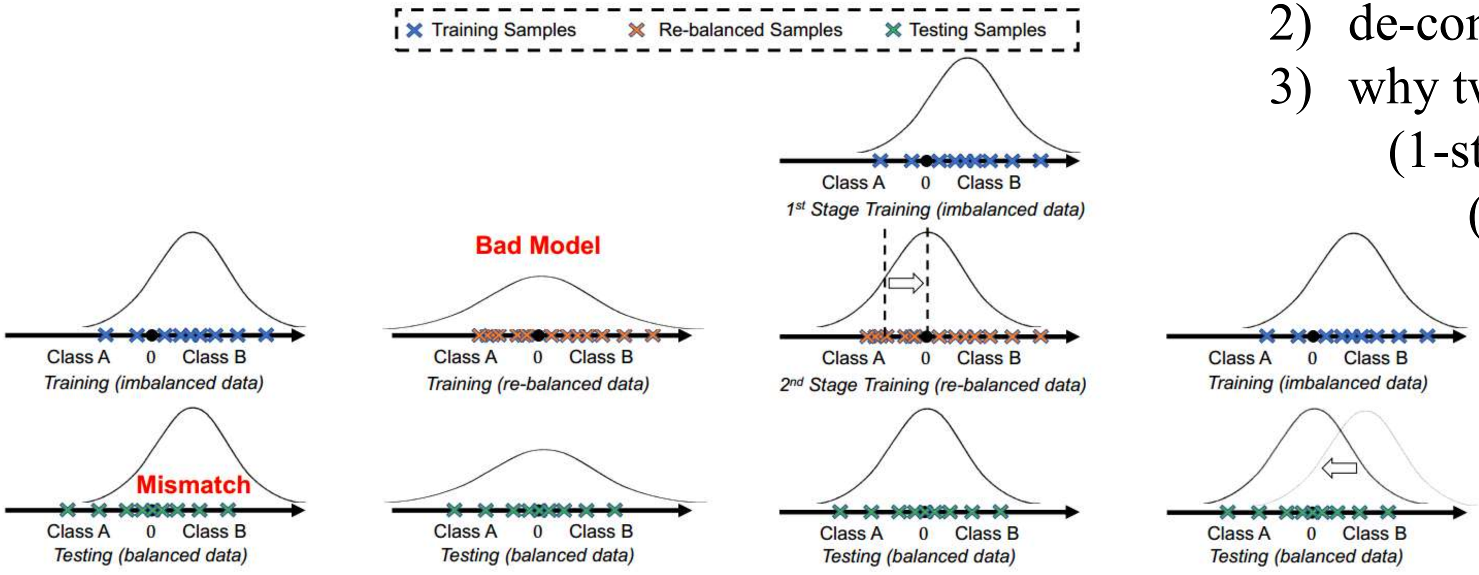
Unifying the long-tail learning methods :

Methods	Two-stage	Re-balancing ($do(D)$)	De-confound ($do(X)$)	Direct Effect
Cosine [50, 51]	-	-	✔	-
LDAM [12]	-	✔	✔	CDE
OLTR [9]	✔	✔	-	NDE
BBN [10]	✔	✔	-	NDE
Decouple [11]	✔	✔	-	NDE
EQL [17]	-	✔	-	-
Our method	-	-	✔	TDE

Table 1: Revisiting the previous state-of-the-arts in our causal graph. CDE: Controlled Direct Effect. NDE: Natural Direct Effect. TDE: Total Direct Effect.



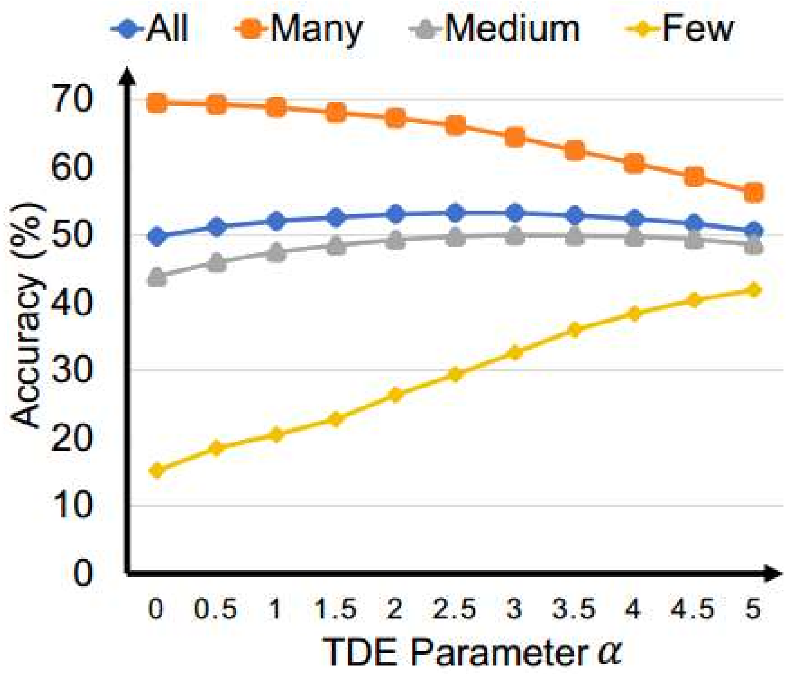
1) one-stage vs. two-stage
2) de-confounder + OLTR… vs. Ours
3) why two-stage work
   (1-step: keep D but no de-con)
   (2-step: de-con but hurt D)



(a) Baseline (b) One-stage Re-balancing (c) Two-stage Re-balancing (d) The Proposed TDE

Case 1 Long-tailed visual recognition

Experiments: **(ImageNet-LT)**



Accuracy for different TDE parameter α

Methods	Many-shot	Medium-shot	Few-shot	Overall
Focal Loss† [28]	64.3	37.1	8.2	43.7
OLTR† [9]	51.0	40.8	20.8	41.9
Decouple-OLTR† [9, 11]	59.9	45.8	27.6	48.7
Decouple-Joint [11]	65.9	37.5	7.7	44.4
Decouple-NCM [11]	56.6	45.3	28.1	47.3
Decouple-cRT [11]	61.8	46.2	27.4	49.6
Decouple-τ-norm [11]	59.1	46.9	30.7	49.4
Decouple-LWS [11]	60.2	47.2	30.3	49.9
Baseline	66.1	38.4	8.9	45.0
Cosine† [50, 51]	67.3	41.3	14.0	47.6
Capsule† [9, 54]	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**

Top-1 accuracy

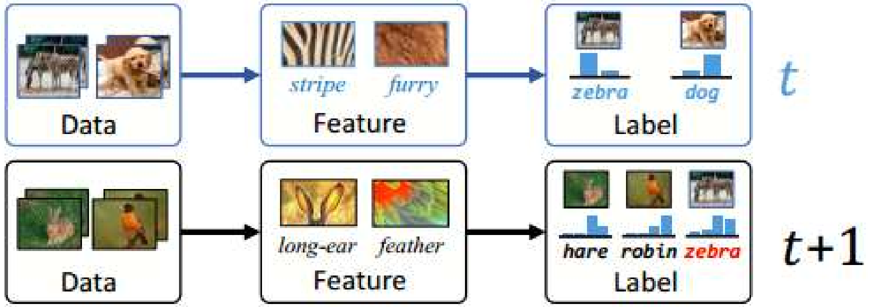
1) muti-head ——play little role
2) α —— nonlinear

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

Nonlinear systems cannot get the exact TDE

Case 2 Class-Incremental Learning

Solid points:



- Active usage of interventions to **create relevant**

Catastrophic forgetting problem

- Data replay effects **without consuming memory**

- Removal of momentum effects **in incremental processes**

Class imbalance problem

Hu et al. Distilling Causal Effect of Data in Class-Incremental Learning. CVPR'21

Case 2 Class-Incremental Learning

Under the causal framework:

- Forgetting Problem

$$Effect_D = P(Y=y \mid do(D=d)) - P(Y=y \mid do(D=0))$$
$$= P(Y=y \mid D=d) - P(Y=y \mid D=0),$$
$$= P(Y=y) - P(Y=y) = 0.$$

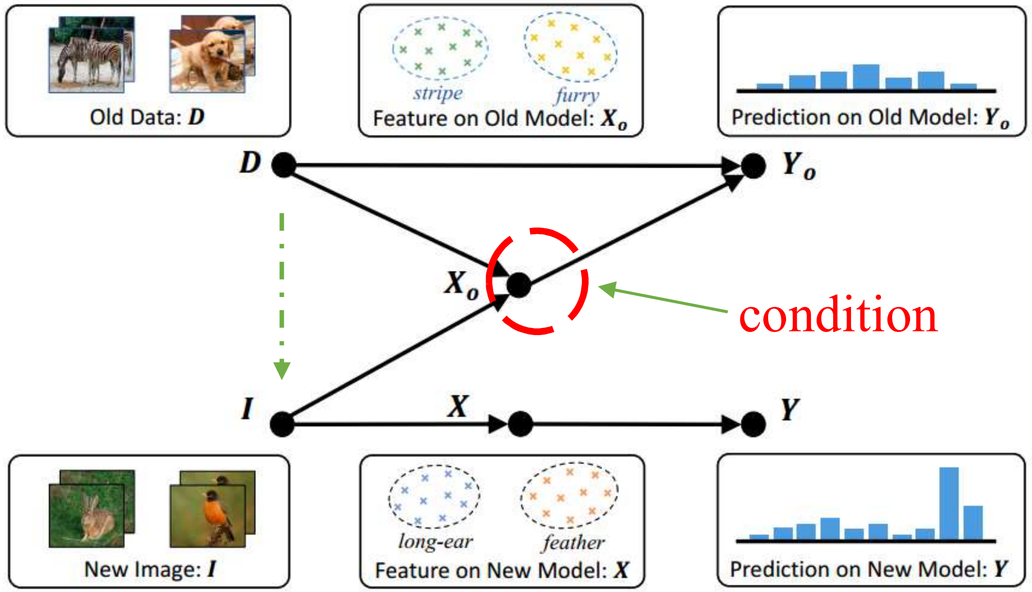
TE

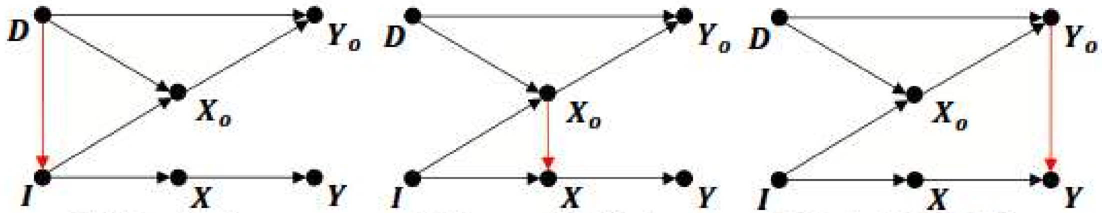
- Data Replay

$$Effect_D = \sum_I P(Y \mid I, D=d)\, P(I \mid D=d)$$
$$-\sum_I P(Y \mid I, D=0)\, P(I \mid D=0)$$
$$= \sum_I P(Y|I)\ (P(I|D=d) - P(I|D=0)) \neq 0,$$

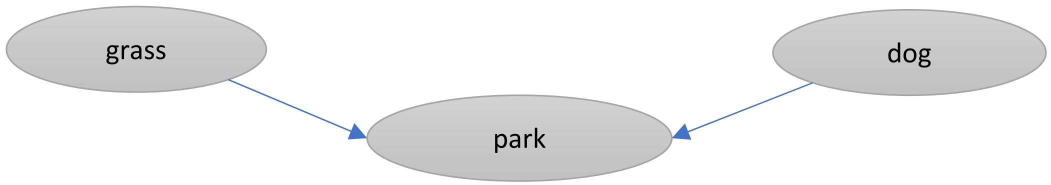
- Feature & Label Distillation

$$Effect_D = \sum_X P(Y|X)\ (P(X|D=d) - P(X|D=0)) \neq 0,$$





(b) Data Replay (c) Feature Distillation (d) Label Distillation



Case 2 Class-Incremental Learning

How to realize the target:

past $Effect_1 = \sum_I P(Y|I)\ (P(I|D=d) - P(I|D=0))$

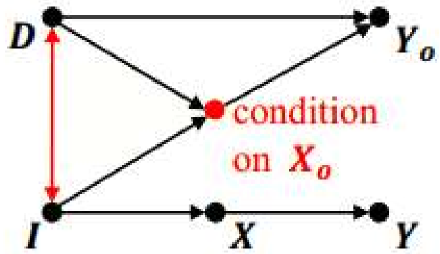
now $Effect_D = \sum_I P(Y|I, X_o)\ \underline{(P(I|X_o,D=d) - P(I|X_o,D=0))}$

$= \sum_I P(Y \mid I)\ \boxed{W(I, X_o, D),}$ → similarity metric

$$Effect_D = \sum_{i \in \{N_1, N_2, \ldots, N_K\}} P(Y \mid I = i)\ W_i,$$

where $\{W_{N_1}, \ldots, W_{N_K}\}$ subject to

$$\begin{cases} W_{N_1} \geq W_{N_2} \geq \cdots \geq W_{N_K} \\ W_{N_1} + W_{N_2} + \cdots + W_{N_K} = 1\,. \end{cases}$$



Algorithm 1 One CIL step with Colliding Effect Distillation

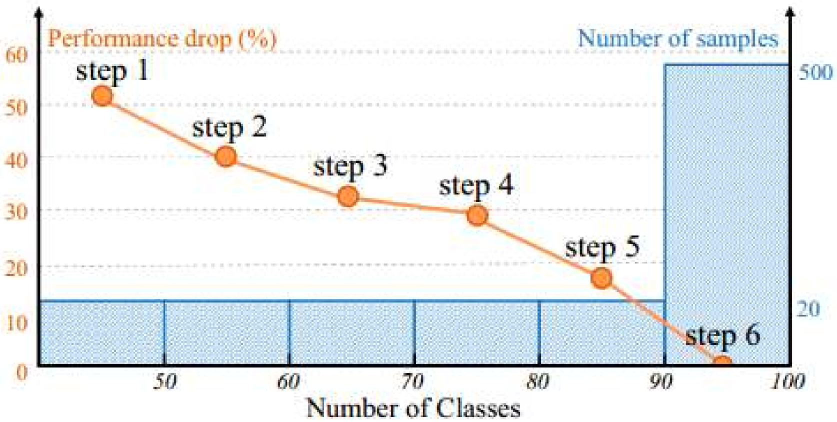
Input : $\mathcal{I}$, Ω_o ▷ new training data, old model
Output : Ω ▷ new model
$\Omega \leftarrow \Omega_o$ ▷ initialize new model
$\mathcal{X}_o \leftarrow \Omega_o(\mathcal{I})$ ▷ represent new images in old features
repeat for $I \in \mathcal{I}$
  $\mathcal{N}_K \leftarrow$ K-NEAREST-NEIGHBOR$(\Omega_o(I), \mathcal{X}_o)$
  $P(Y|N_1), \ldots, P(Y|N_K) \leftarrow \Omega(\mathcal{N}_K)$
  $W_1, \ldots, W_K \leftarrow$ WEIGHTASSIGN(K) ▷ Eq. (7)
  $Effect \leftarrow \sum_{j=1}^{K} W_j\ P(Y|N_j;\ \Omega)$
  $\Omega \leftarrow \arg\min_{\Omega}(-\log(Effect))$
until converge

Some experiments on CIFAR-100 with 5-step:

R	Baseline	Top1	Top5	Top10	Rand	Bottom	Variant1	Variant2
5	50.76	55.92	58.12	58.53	54.88	41.70	58.22	58.41
10	61.68	63.51	63.66	64.04	53.80	51.41	63.54	63.97
20	63.57	64.76	64.93	65.18	64.16	57.07	64.87	65.22

Case 2 Class-Incremental Learning

not the same as regular long-tail problems



CIFAR-100 with 5-step-20-relay

How to realize the target:

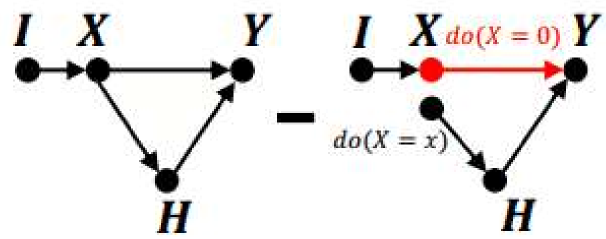
Static $\hat{d} = \overline{\boldsymbol{x}}_T / \|\overline{\boldsymbol{x}}_T\|, \text{ where } \overline{\boldsymbol{x}}_t = \mu \cdot \overline{\boldsymbol{x}}_{t-1} + \boldsymbol{x}_t$

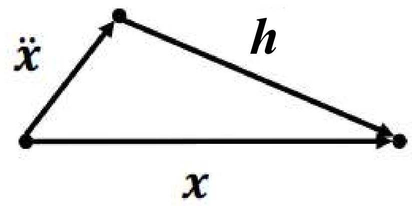
Increment $h = (1-\beta)\, h_{t-1} + \beta\, h_t$ ← Step

For each image in I:

$$[Y \,|\, I=i] = [Y \,|\, do(X=x)] - \alpha \cdot [Y \,|\, do(X=0, H=h)]$$
$$= [Y \,|\, X=x] - \alpha \cdot [Y \,|\, X=x^h],$$

Tips: α and β **were trained** in the finetuning stage
using a re-sampled subset containing balanced old and new data





Algorithm 2 CIL with Incremental Momentum Effect Removal

Input : $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_T$ ▷ training data of T steps
Input : I ▷ a testing image
For $t \in 1, 2, \ldots, T$ ▷ each CIL step
  $\alpha, \beta, h_t \leftarrow$ MovingAverage($\mathcal{I}_t$; Ω_t) ▷ training
  $h \leftarrow (1-\beta)\, h_{t-1} + \beta\, h_t$ ▷ head direction
  $X \leftarrow$ FeatureExtractor(I) ▷ inference
  $Y \leftarrow$ Classifier($X - x^h$) ▷ Eq. (8)

Case 2 Class-Incremental Learning

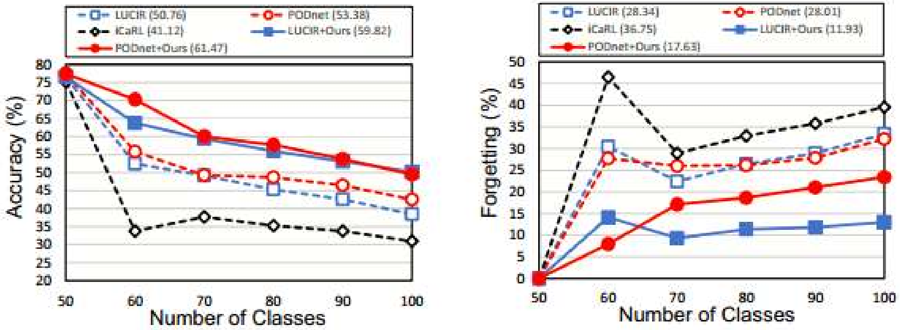
Experiments:

Average Incremental Accuracies

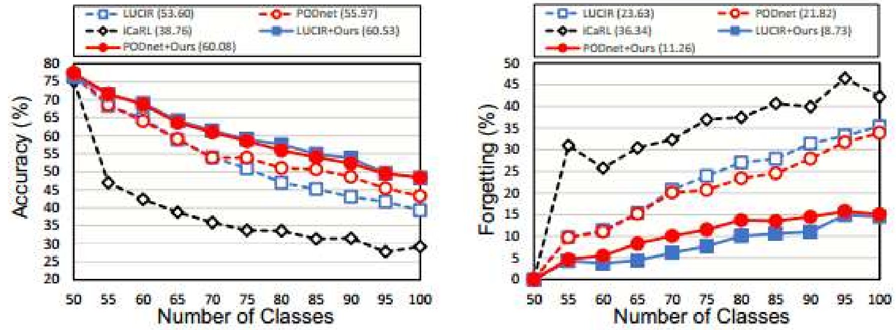
	Methods	CIFAR-100		ImageNet-Sub		ImageNet-Full	
		T=5	*10*	*5*	*10*	*5*	*10*
(1) R=5	LUCIR†	50.76	53.60	66.44	60.04	62.61	58.01
	+ DDE (ours)	59.82+9.06	**60.53**+6.93	70.86+4.42	68.30+8.26	**66.18**+3.57	**62.89**+4.88
	PODNet†	53.38	55.97	71.43	64.90	61.01	55.36
	+ DDE (ours)	**61.47**+8.09	60.08+4.11	**74.59**+3.16	**69.26**+4.36	63.15+2.14	58.34+2.98
(2) R=10	LUCIR†	61.68	58.30	68.13	64.04	65.21	61.60
	+ DDE (ours)	**64.41**+2.73	**62.00**+3.70	71.20+3.07	69.05+5.01	**67.04**+1.83	**64.98**+3.38
	PODNet†	61.40	58.92	74.50	70.40	62.88	59.56
	+ DDE (ours)	63.40+2.00	60.52+1.60	**75.76**+1.26	**73.00**+2.60	64.41+1.53	62.09+2.53
(3) R=20	LUCIR†	63.57	60.95	70.71	67.60	66.84	64.17
	+ DDE (ours)	65.27+1.70	62.36+1.41	72.34+1.63	70.20+2.60	**67.51**+0.67	**65.77**+1.60
	PODNet†	64.70	62.72	75.58	73.48	65.59	63.27
	+ DDE (ours)	**65.42**+0.72	**64.12**+1.40	**76.71**+1.13	**75.41**+1.93	66.42+0.83	64.71+1.44
	iCaRL† [35]	57.17	52.27	65.04	59.53	51.36	46.72
	BiC [47]	59.36	54.20	70.07	64.96	62.65	58.72
	LUCIR [13]	63.17	60.14	70.84	68.32	64.45	61.57
	Mnemonics [25]	63.34	62.28	72.58	71.37	64.54	63.01
	PODNet [9]	64.83	63.19	75.54	74.33	66.95	64.13
	TPCIL [42]	65.34	63.58	76.27	74.81	64.89	62.88

Case 2 Class-Incremental Learning

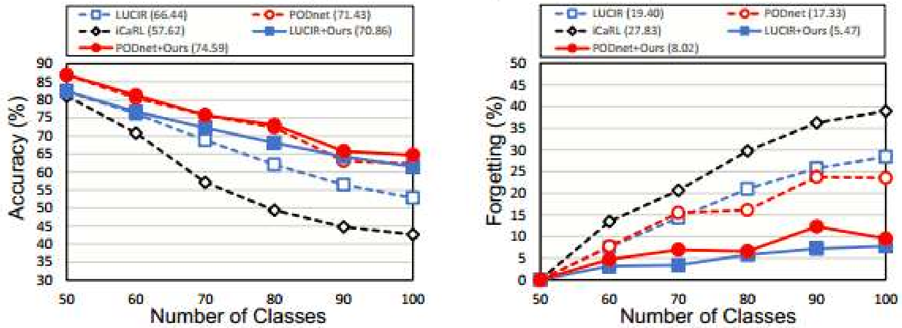
Experiments:



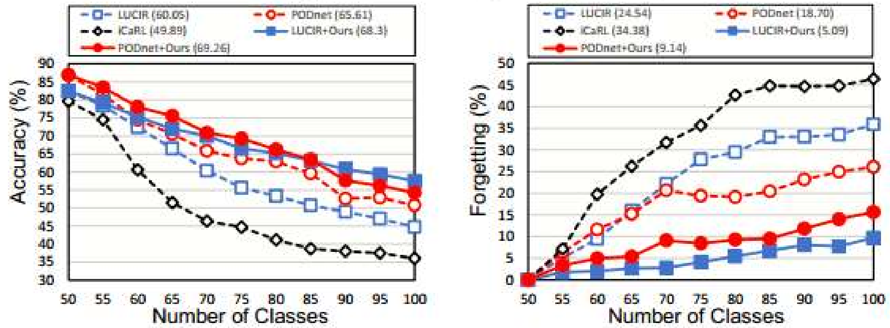
(a) CIFAR-100 ($T = 5$, $R = 5$)



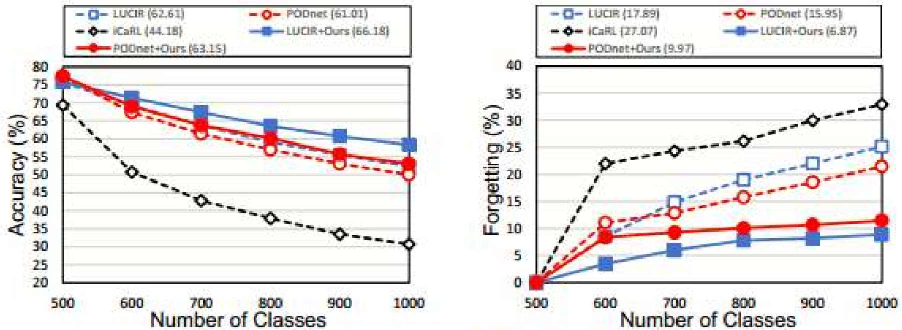
(b) CIFAR-100 ($T = 10$, $R = 5$)



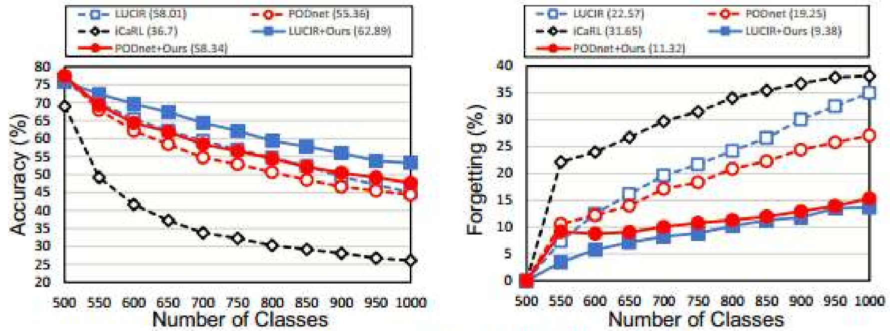
(c) ImageNet-Sub ($T = 5$, $R = 5$)



(d) ImageNet-Sub ($T = 10$, $R = 5$)



(e) ImageNet-Full ($T = 5$, $R = 5$)



(f) ImageNet-Full ($T = 10$, $R = 5$)

Case 2 Class-Incremental Learning

Ablation:

Distillation of Colliding Effect (DCE)
incremental Momentum Effect Removal (MER)

	Methods	R = 5	10	20
LUCIR →	Baseline [13]	50.76	61.68	63.57
Accuracy (%)	+All	59.82+9.06	64.41+2.73	65.27+1.70
	+DCE	58.53+7.77	64.04+2.36	65.18+1.61
	+MER	53.55+2.79	63.40+1.72	64.43+0.86
Forgetting (%)	Baseline [13]	28.34	17.51	14.08
	+All	11.93-16.41	6.23-11.28	7.11-6.97
	+DCE	16.82-11.52	10.16-7.35	8.41-5.67
	+MER	17.45-10.89	12.15-5.36	10.01-4.07

Under no replay data

Methods	CIFAR-100		ImageNet-Sub	
	T = 5	10	5	10
Baseline	45.57	32.72	58.55	45.06
Ours	**59.11**+13.54	**55.31**+22.59	**69.22**+10.67	**65.51**+20.45

THANKS