Multi-Class Incremental Learning

Yaoyao Liu
liuyaoyao@tju.edu.cn

NUS National University of Singapore
天津大学 Tianjin University

Outline

- Background
- Methods
- Experiments
- Takeaways

Background

Motivation

Thousands of new users and items everyday

Update the model with incremental data

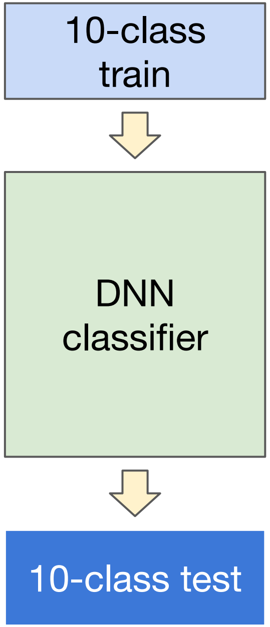
Limited memory

Taking too long time to retrain the model

(*Images from Internet*)

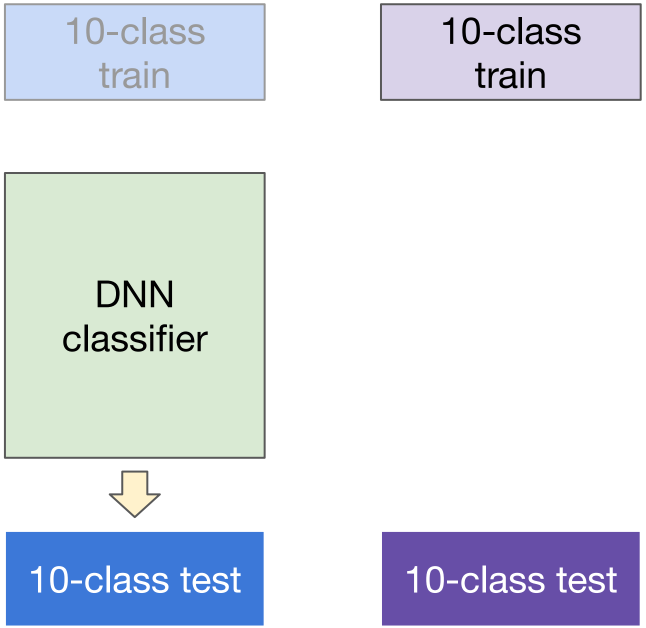
Problem Definitions

Incremental learning (also lifelong learning, continual learning)



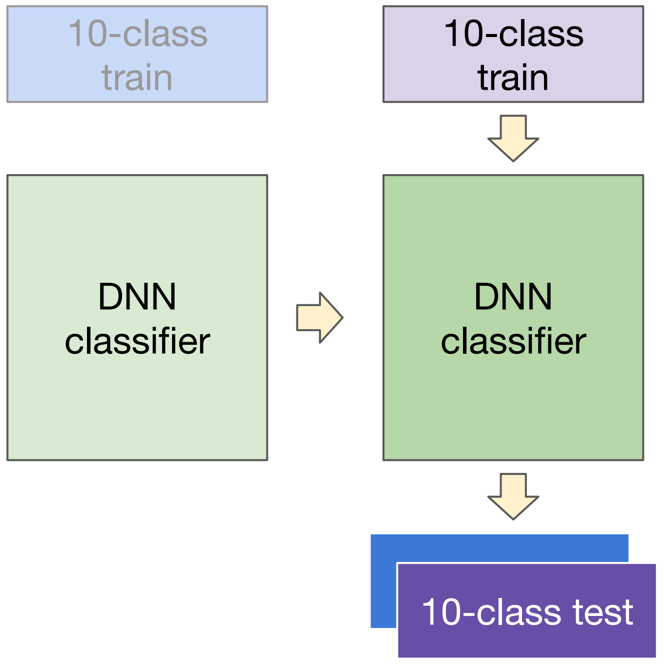
Problem Definitions

Incremental learning (also lifelong learning, continual learning)



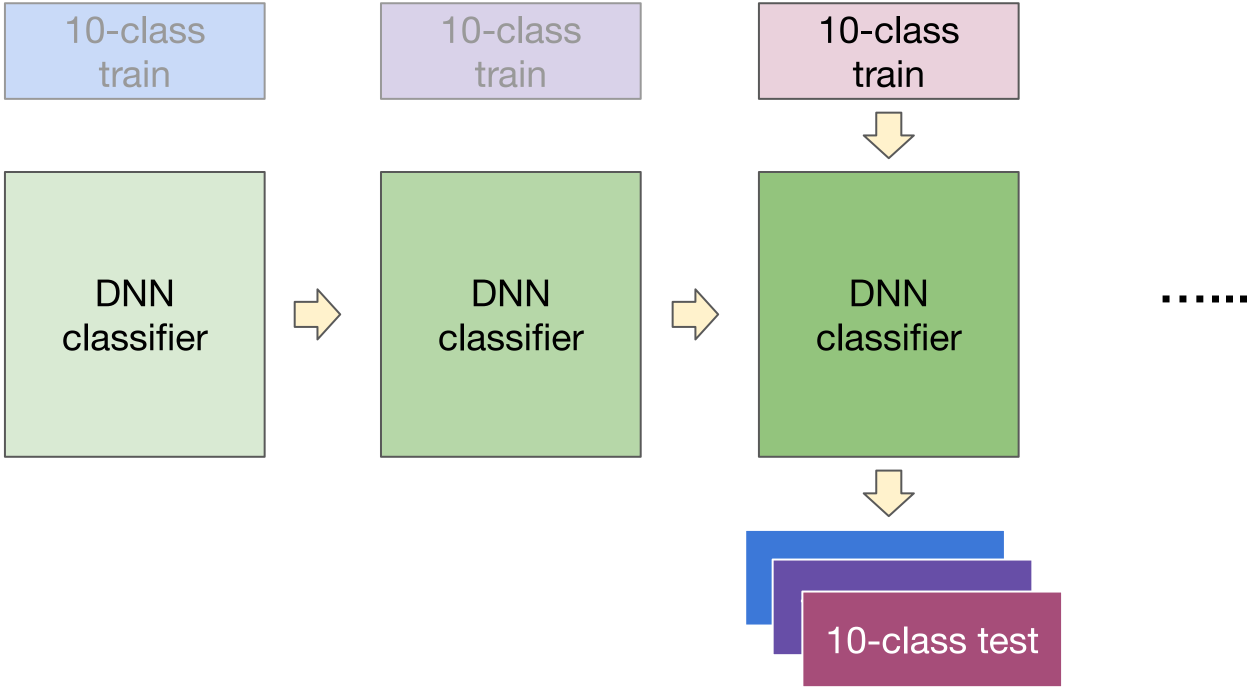
Problem Definitions

Incremental learning (also lifelong learning, continual learning)



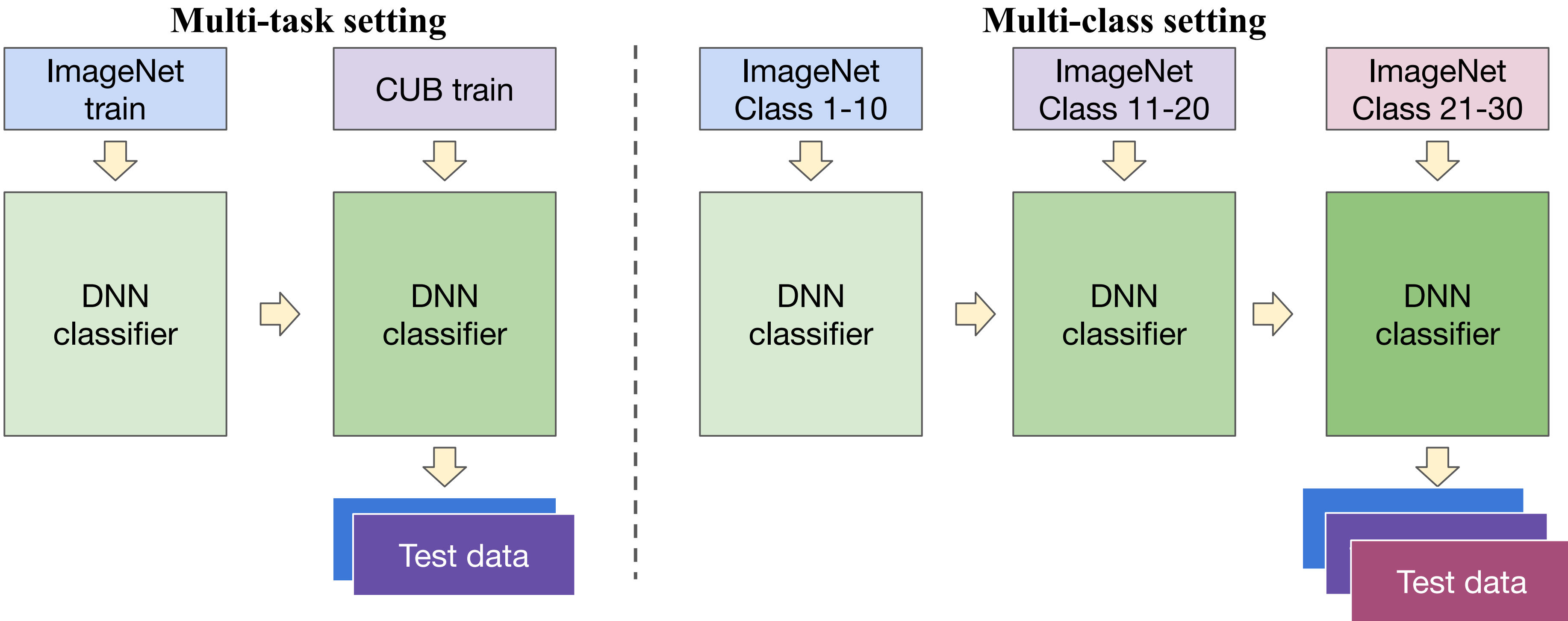
Problem Definitions

Incremental learning (also lifelong learning, continual learning)



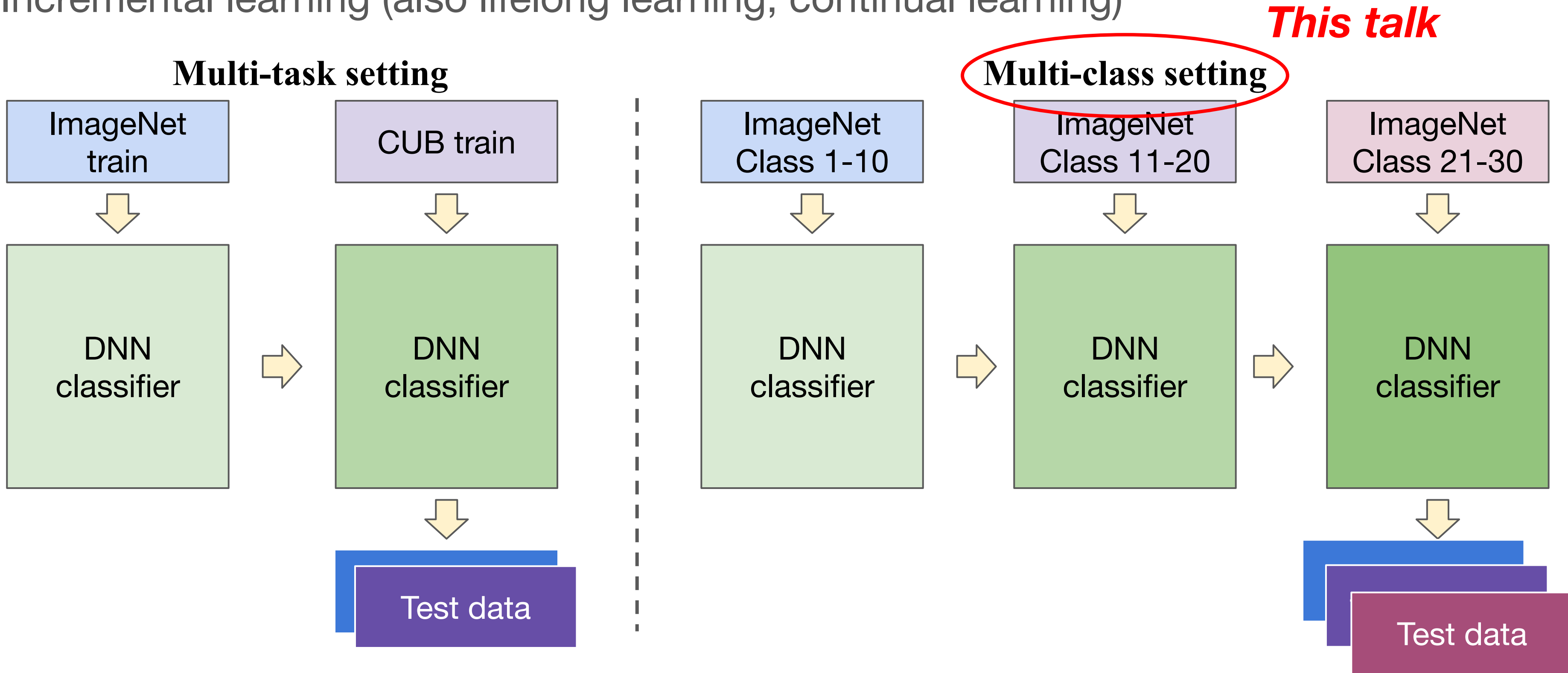
Problem Definitions

Incremental learning (also lifelong learning, continual learning)



Problem Definitions

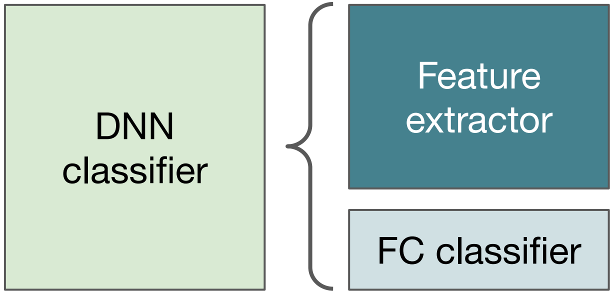
Incremental learning (also lifelong learning, continual learning)



Related Learning Methods

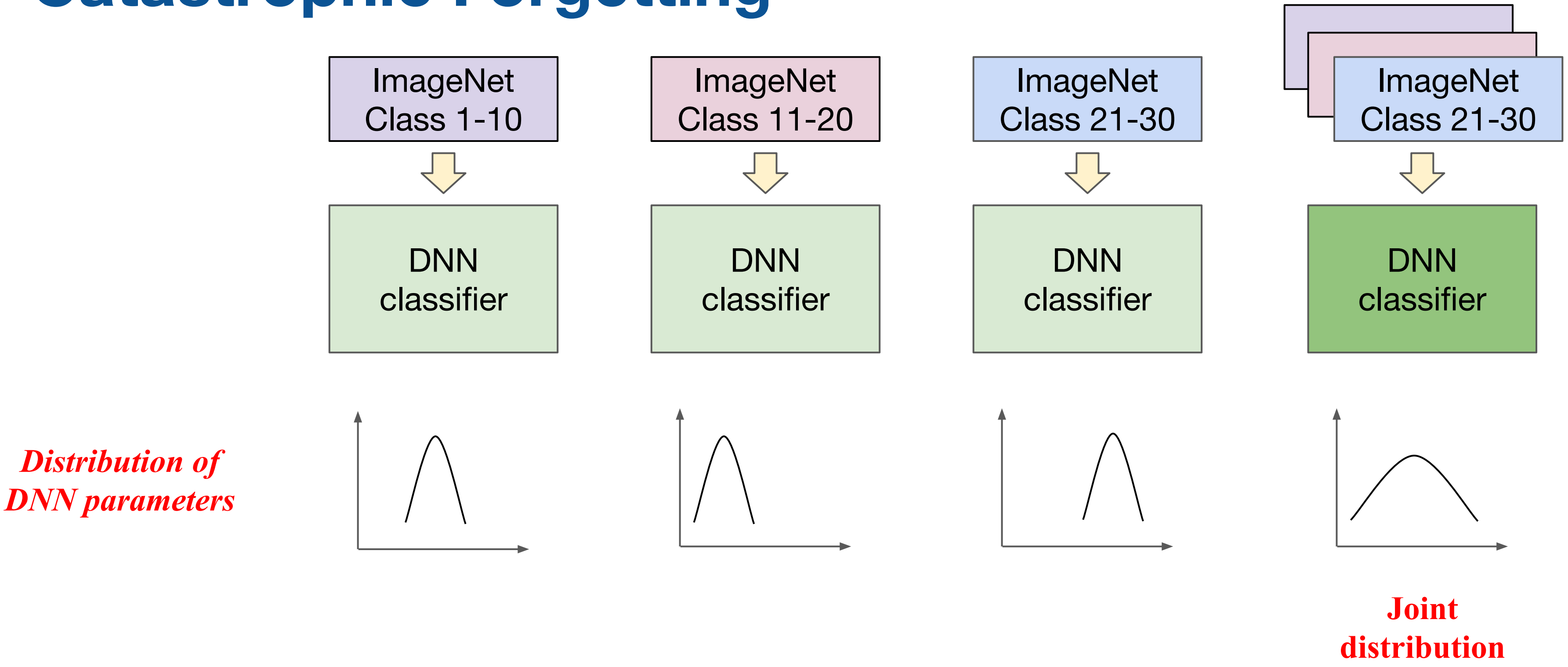
	Transfer learning	Multi-task learning	Multi-task incremental learning	Multi-class incremental learning
Target task(s)	Single	Multiple	Multiple	Single
Source task(s)	Multiple	Multiple	Multiple	Single
Data arrival	Constantly / Once	Once	Constantly	Constantly

Challenges

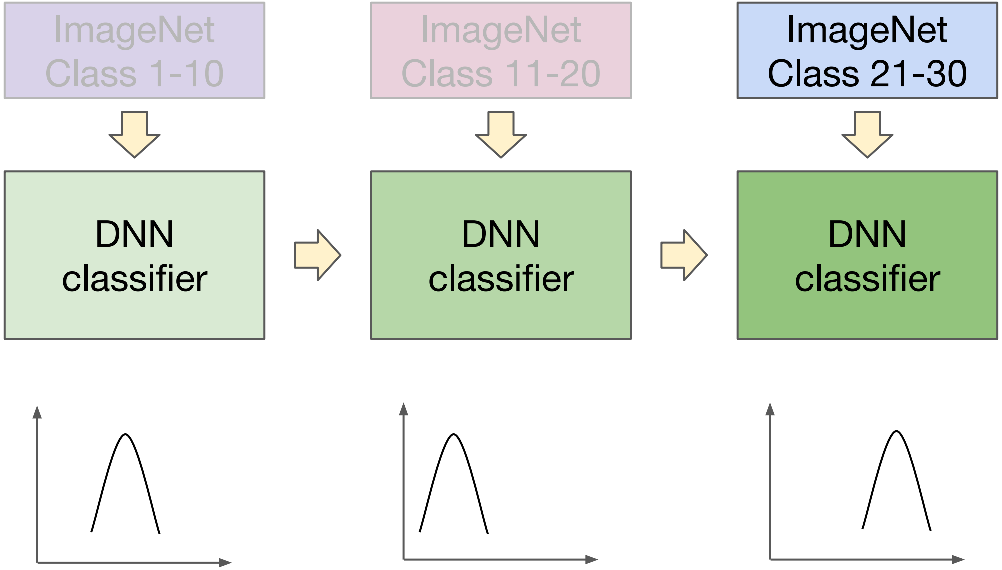


1. ***Catastrophic Forgetting***
   Model bias on the latest class group
2. ***FC classifier is not extendable***
   10 classes -> 20 classes
3. ***Memory resources may be limited***
   Not able to retain all previous samples

Catastrophic Forgetting



Catastrophic Forgetting



Distribution of DNN parameters

Overfit to the latest class group

Literature Review

- **To improve the feature extractor**

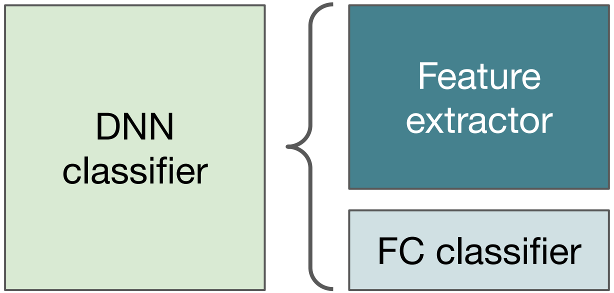
  LwF[1], ...

- **To improve the classifier:**

  iCaRL[4], BiC[3], ...

- **To improve both:**

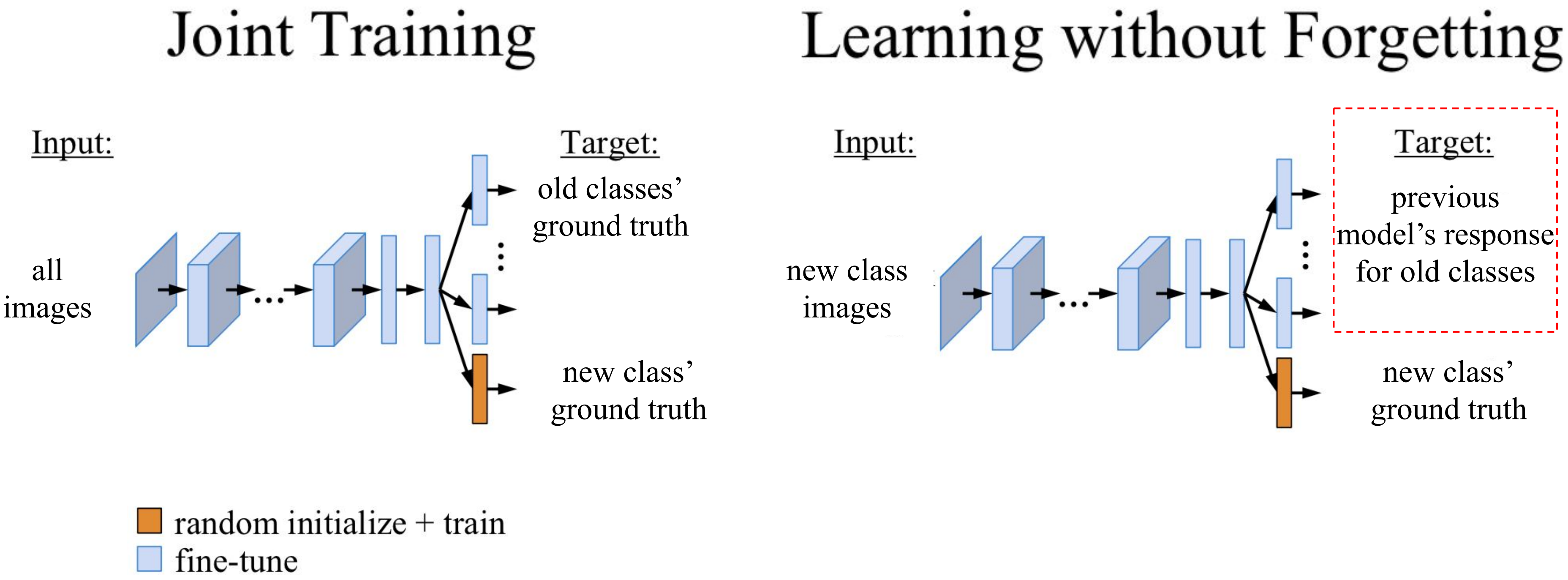
  Hou et al.[5], EEIL[2]......



Reference
[1] Li, Zhizhong, and Derek Hoiem. "Learning without forgetting." T-PAMI 2017;
[2] Castro, Francisco M., et al. "End-to-end incremental learning." ECCV 2018;
[3] Wu, Yue, et al. "Large scale incremental learning." CVPR 2019;
[4] Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning." CVPR 2017;
[5] Hou, Saihui, et al. "Learning a Unified Classifier Incrementally via Rebalancing." CVPR 2019.

Methods

Learning without Forgetting (LwF)



Reference

[1] Li, Zhizhong, and Derek Hoiem. "Learning without forgetting." T-PAMI 2017.

Learning without Forgetting (LwF)

Idea: discouraging the old classes output to change **[1]**

Knowledge distillation

Proposed by Hilton et al.[2] for ensemble modeling

Classification $$\mathcal{L}_{\text{new}}(\mathbf{y}_n, \hat{\mathbf{y}}_n) = -\mathbf{y}_n \log \hat{\mathbf{y}}_n$$

Distillation $$\mathcal{L}_{\text{old}}(\mathbf{y}_o, \hat{\mathbf{y}}_o) = -\mathbf{y}_o \log \hat{\mathbf{y}}_o$$

Full objective $$\mathcal{L} = \mathcal{L}_{\text{old}} + \mathcal{L}_{\text{new}}$$

In which,

$\mathbf{y}_o = \Phi_{\text{old}}(x)$

$\mathbf{y}_n$: ground truth

$[\hat{\mathbf{y}}_o, \hat{\mathbf{y}}_n] = \Phi_{\text{current}}(x)$

Reference

[1] Li, Zhizhong, and Derek Hoiem. "Learning without forgetting." T-PAMI 2017;

[2] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv 2015.

Learning without Forgetting (LwF)

Summary

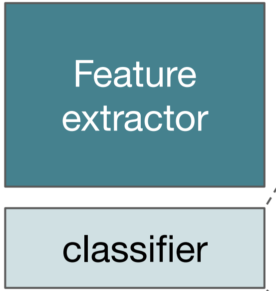
+ Distillation loss -> improve the learning of feature extractor
+ Don’t need to retain data for old classes

- Using a simple way to deal with the FC classifier without solving the bias problem

* This method is proposed for multi-task setting. However, it is usually used as a baseline of multi-class incremental learning papers

Reference
[1] Li, Zhizhong, and Derek Hoiem. "Learning without forgetting." T-PAMI 2017.

iCaRL: Incremental Classifier and Representation Learning



Algorithm 1 iCaRL CLASSIFY

input x // image to be classified
require $\mathcal{P} = (P_1, \dots, P_t)$ // class exemplar sets
require $\varphi : \mathcal{X} \to \mathbb{R}^d$ // feature map
  for $y = 1, \dots, t$ **do**
    $\mu_y \leftarrow \frac{1}{|P_y|} \sum_{p \in P_y} \varphi(p)$ // mean-of-exemplars
  end for
  $y^* \leftarrow \operatorname*{argmin}_{y=1,\dots,t} \|\varphi(x) - \mu_y\|$ // nearest prototype
output class label y^*

Idea: FC classifier -> nearest-mean-of-exemplars (NME) *NME is used only in test phase

Reference

[1] Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning." CVPR 2017.

iCaRL: Incremental Classifier and Representation Learning

Algorithm 4 iCaRL ConstructExemplarSet

input image set $X = \{x_1, \dots, x_n\}$ of class y
input m target number of exemplars
require current feature function $\varphi : \mathcal{X} \to \mathbb{R}^d$

$\mu \leftarrow \frac{1}{n} \sum_{x \in X} \varphi(x)$ // current class mean

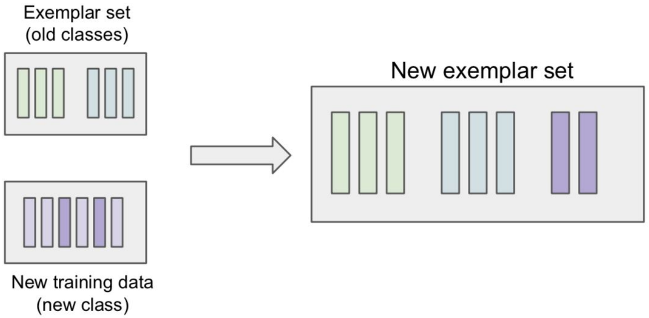
for $k = 1, \dots, m$ **do**

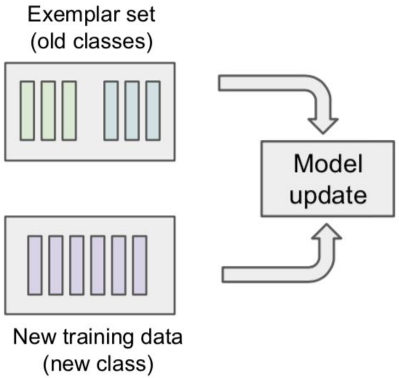
$$p_k \leftarrow \underset{x \in X}{\operatorname{argmin}} \left\| \mu - \frac{1}{k}\left[\varphi(x) + \sum_{j=1}^{k-1} \varphi(p_j)\right] \right\|$$

end for

$P \leftarrow (p_1, \dots, p_m)$

output exemplar set P





Reference
[1] Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning." CVPR 2017.

(*Images from Ramon Morros*)

iCaRL: Incremental Classifier and Representation Learning

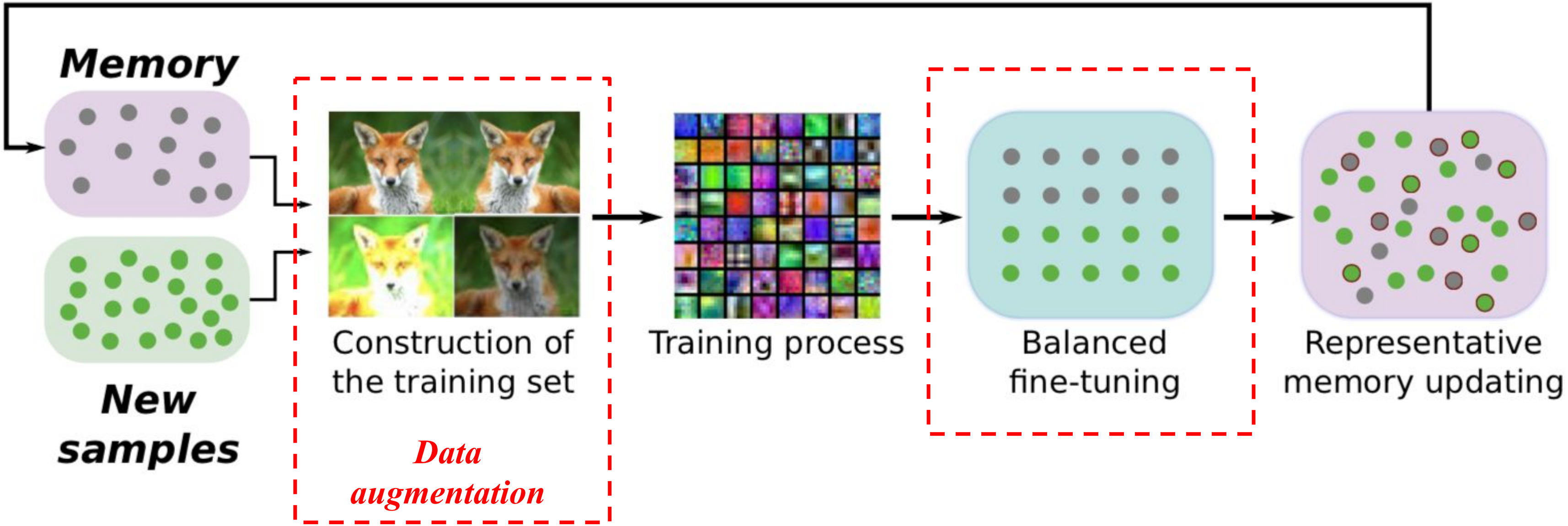
Summary

+ Solving the bias problem for the classifier

- Need to retain parts of old data
- Non-parametric classifier may fail in some novel similar classes
- Training and testing using different types of classifier (train: fc, test: NME)

Reference
[1] Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning." CVPR 2017.

End-to-end Incremental Learning (EEIL)



Reference

[1] Castro, Francisco M., et al. "End-to-end incremental learning." ECCV 2018.

End-to-end Incremental Learning (EEIL)

Summary

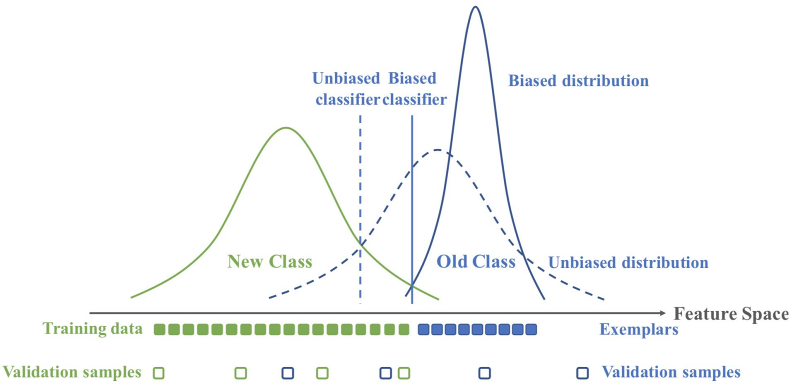
+ End-to-end, improvement on both feature extractor and classifier
+ A series of data augmentation techniques

- Improvements may come from tricks

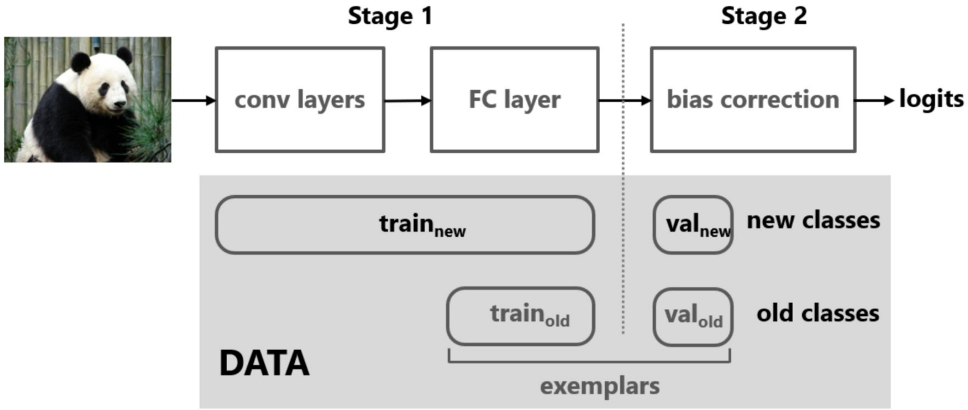
Reference
[1] Castro, Francisco M., et al. "End-to-end incremental learning." ECCV 2018.

Large Scale Incremental Learning (BiC)

Problem: bias on novel classer



BiC: Bias Correction



$$q_k = \begin{cases} o_k & 1 \le k \le n \\ \alpha o_k + \beta & n+1 \le k \le n+m \end{cases}$$

$$L_b = -\sum_{k=1}^{n+m} \delta_{y=k} \log[softmax(q_k)]$$

Reference

[1] Wu, Yue, et al. "Large scale incremental learning." CVPR 2019.

Large Scale Incremental Learning (BiC)

Summary

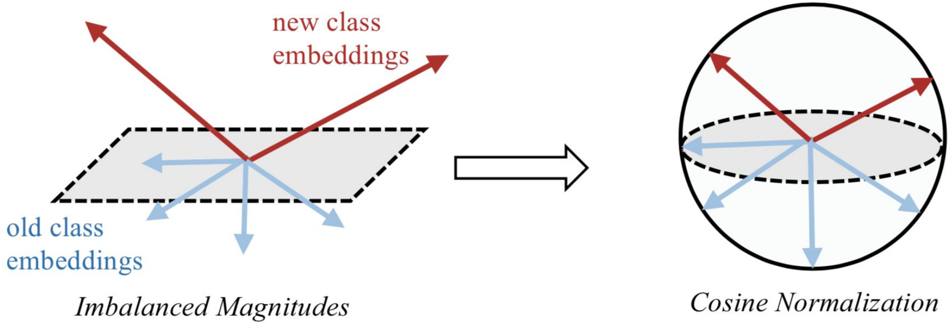
+ Solve the bias problem on classifier

- The correction function works only on large scale datasets

Reference
[1] Wu, Yue, et al. "Large scale incremental learning." CVPR 2019.

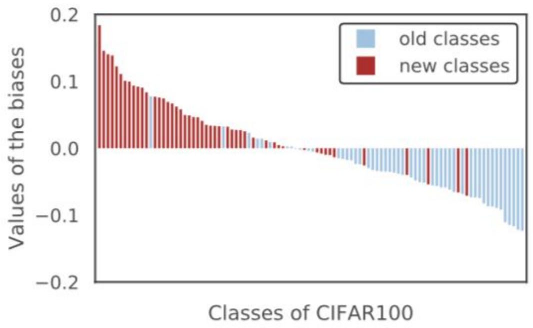
Learning a Unified Classifier Incrementally via Rebalancing (Hou et al.)

Cosine normalization



FC classifier

$$p_i(x) = \frac{\exp(\theta_i^{\mathrm{T}} f(x) + b_i)}{\sum_j \exp(\theta_j^{\mathrm{T}} f(x) + b_j)}$$



Cosine distance

$$p_i(x) = \frac{\exp(\eta\langle\bar{\theta}_i, \bar{f}(x)\rangle)}{\sum_j \exp(\eta\langle\bar{\theta}_j, \bar{f}(x)\rangle)}$$

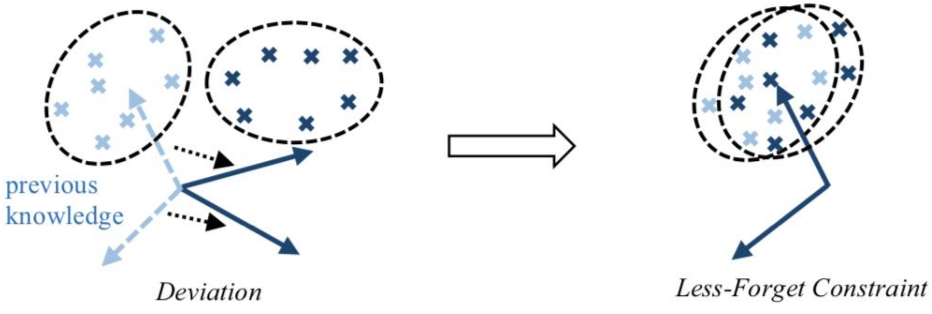
Improve classifier

Reference

[1] Hou, Saihui, et al. "Learning a Unified Classifier Incrementally via Rebalancing." CVPR 2019.

Learning a Unified Classifier Incrementally via Rebalancing (Hou et al.)

Less-forget constraint



Distillation loss

$$L_{\text{dis}}^{\text{C}}(x) = -\sum_{i=1}^{|\mathcal{C}_o|} \| \langle \bar{\theta}_i, \bar{f}(x) \rangle - \langle \bar{\theta}_i^*, \bar{f}^*(x) \rangle \|$$

Cosine distance of feature

$$L_{\text{dis}}^{\text{G}}(x) = 1 - \langle \bar{f}^*(x), \bar{f}(x) \rangle$$

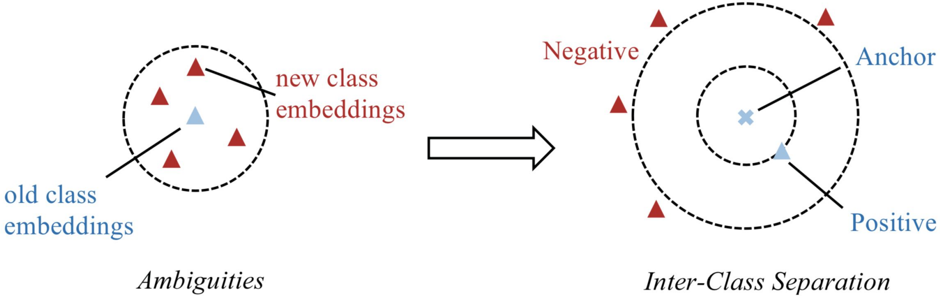
Improve feature extractor

Reference

[1] Hou, Saihui, et al. "Learning a Unified Classifier Incrementally via Rebalancing." CVPR 2019.

Learning a Unified Classifier Incrementally via Rebalancing (Hou et al.)

Inter-class separation



Add margin threshold to top-K classes

$$L_{\mathrm{mr}}(x) = \sum_{k=1}^{K} \max(m - \langle \bar{\theta}(x), \bar{f}(x) \rangle + \langle \bar{\theta}^k, \bar{f}(x) \rangle,\ 0)$$

Improve classifier

Reference

[1] Hou, Saihui, et al. "Learning a Unified Classifier Incrementally via Rebalancing." CVPR 2019.

Learning a Unified Classifier Incrementally via Rebalancing (Hou et al.)

Summary

+ Improvement on both classifier (cosine distance, inter-class separation) and feature extractor (less-forgot constraint)

- The first group requires more classes than other groups
  (require a good initialization for the CONV networks)
- Extremely slow with inter-class separation strategy

Reference
[1] Hou, Saihui, et al. "Learning a Unified Classifier Incrementally via Rebalancing." CVPR 2019.

Comparison

	LwF[1]	iCaRL[2]	EEIF[3]	BiC[4]	Hou et al.[5]
Feature Extractor	**Distillation Loss**	Distillation Loss	Distillation Loss	Distillation Loss	**Less-Forget Constraint**
		Exemplar	Exemplar, **Balanced Fine-tuning**	Exemplar	Exemplar,
Classifier	FC	**NME**	FC	FC, **Bias Correction**	FC, **Cosine Normalization**, **Inter-Class Separation**

Reference
[1] Li, Zhizhong, and Derek Hoiem. "Learning without forgetting." T-PAMI 2017;
[2] Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning." CVPR 2017;
[3] Castro, Francisco M., et al. "End-to-end incremental learning." ECCV 2018;
[4] Wu, Yue, et al. "Large scale incremental learning." CVPR 2019;
[5] Hou, Saihui, et al. "Learning a Unified Classifier Incrementally via Rebalancing." CVPR 2019.

Experiments

Datasets and Benchmark

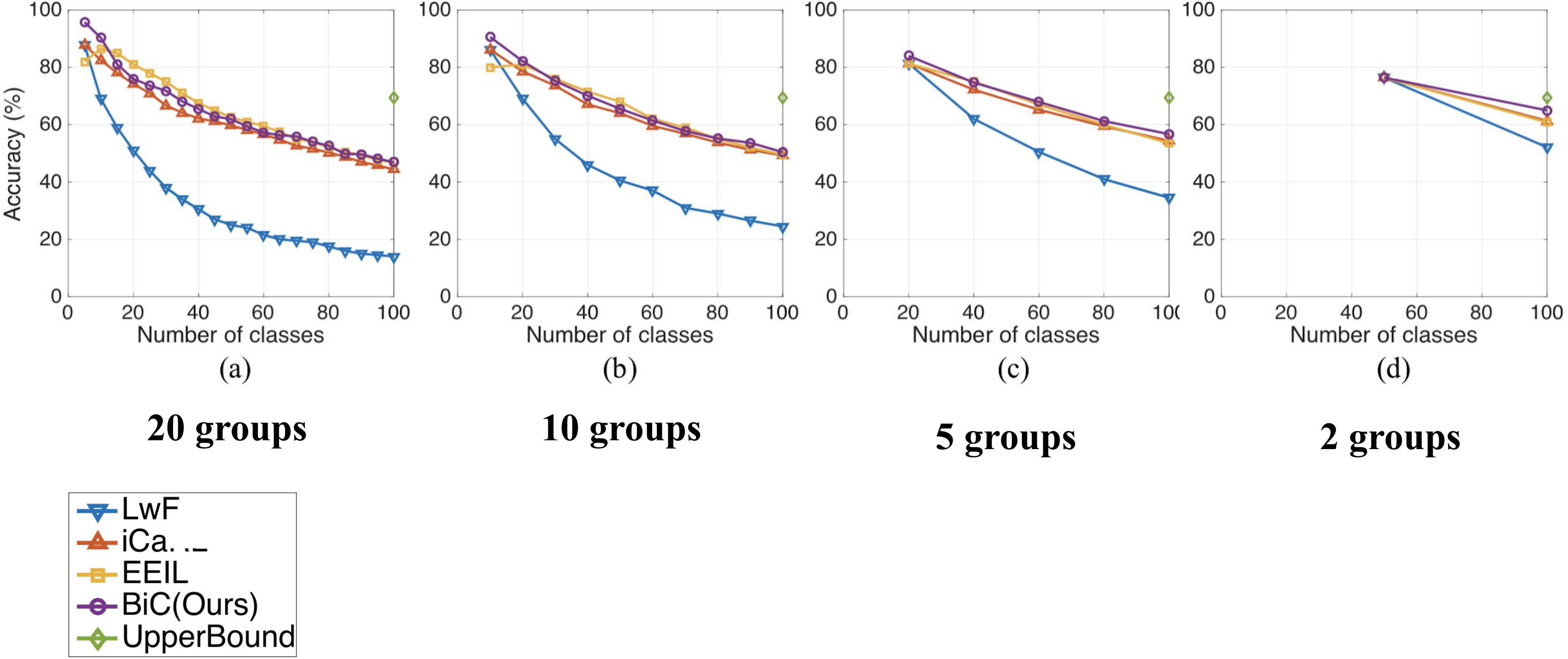
Datasets: CIFAR100, ImageNet-Sub (100 classes subset), ImageNet

E.g. CIFAR-100

Number of class groups

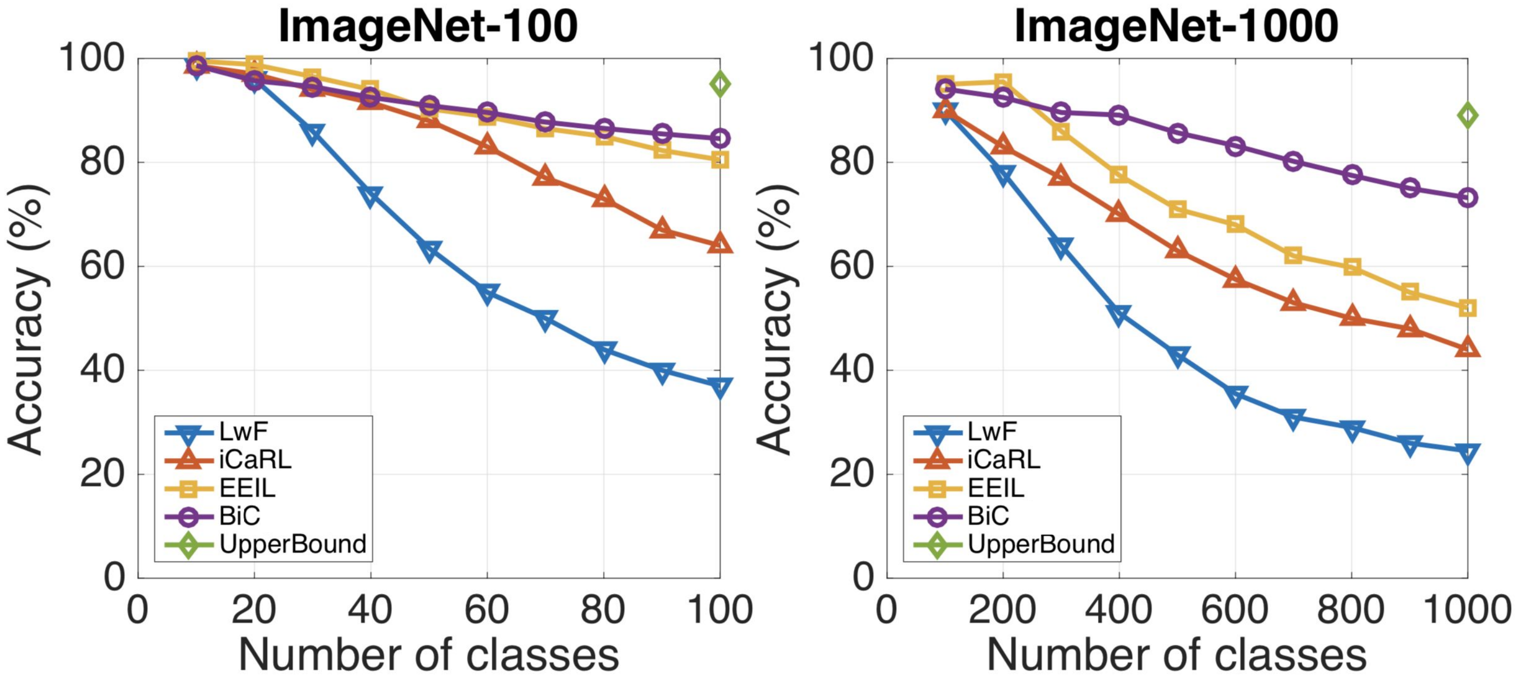
2 groups	50 classes					50 classes				
10 groups	10	10	10	10	10	10	10	10	10	10

Experiments on CIFAR-100



20 groups **10 groups** **5 groups** **2 groups**

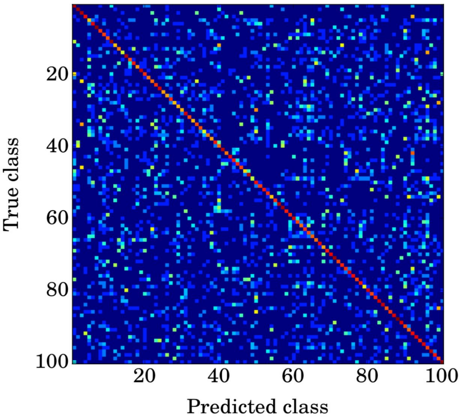
Experiments on ImageNet



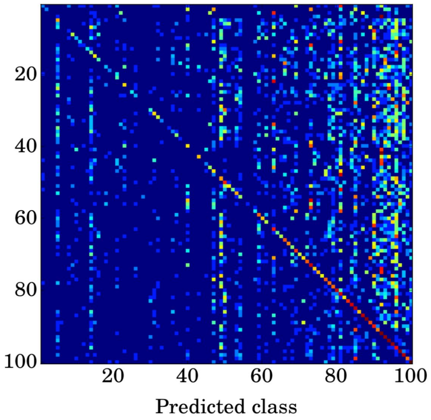
(a) (b)

10 groups

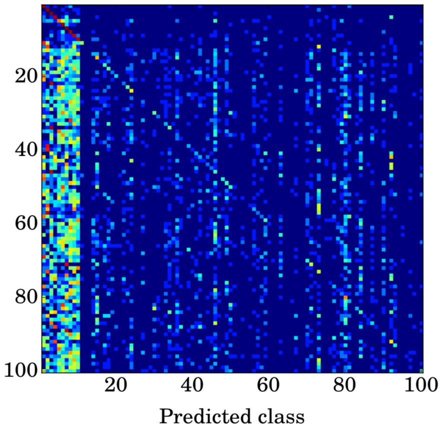
Confusion Matrices



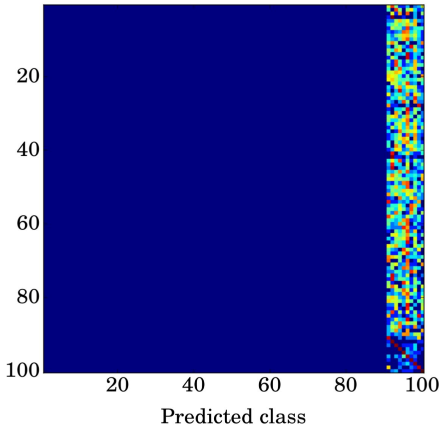
iCaRL



LwF



fixed representation



fine-tuning

Takeaways

Important techniques:

1. Distillation Loss -> retain knowledge for old classes
2. Nearest-Mean-of-Exemplar Classifier -> no-bias classifier

Future work:

1. Other strategy for retaining knowledge for old classes
2. Shareable parametric classifier -> meta-learning?

Thanks!
Any questions?

liuyaoyao@tju.edu.cn