Learning to Retain while Acquiring: Combating Distribution-Shift in Adversarial Data-Free Knowledge-Distillation

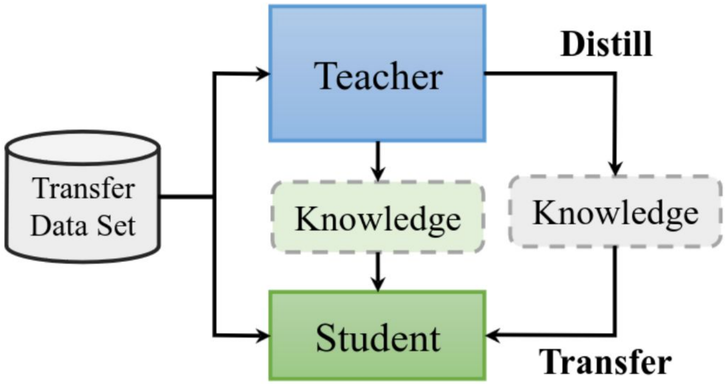
Gaurav Patel[†], Konda Reddy Mopuri[‡], and Qiang Qiu[†]
[†]Purdue University
[‡]Indian Institute of Technology Hyderabad

TUE-PM-350

Introduction

Knowledge Distillation

Knowledge distillation (KD) is a popular model compression technique that seeks to **transfer valuable information** from a **cumbersome teacher network** to **a similar-capacity or a compact student network**.



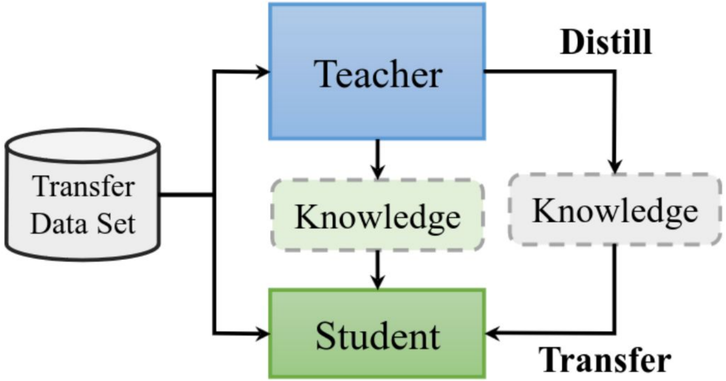
Liu et al., Data-Free Knowledge Transfer: A Survey, arXiv:2112.15278, 2021.

Introduction

PURDUE UNIVERSITY

Knowledge Distillation

However, one basic assumption that **KD** considers is the availability of a **Transfer Dataset (teacher's training data)**, used to query the **teacher** and the **student,** to conduct **KD**.

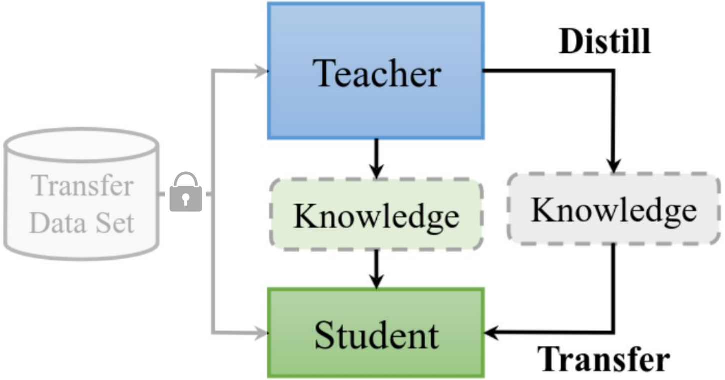


Liu et al., Data-Free Knowledge Transfer: A Survey, arXiv:2112.15278, 2021.

Introduction

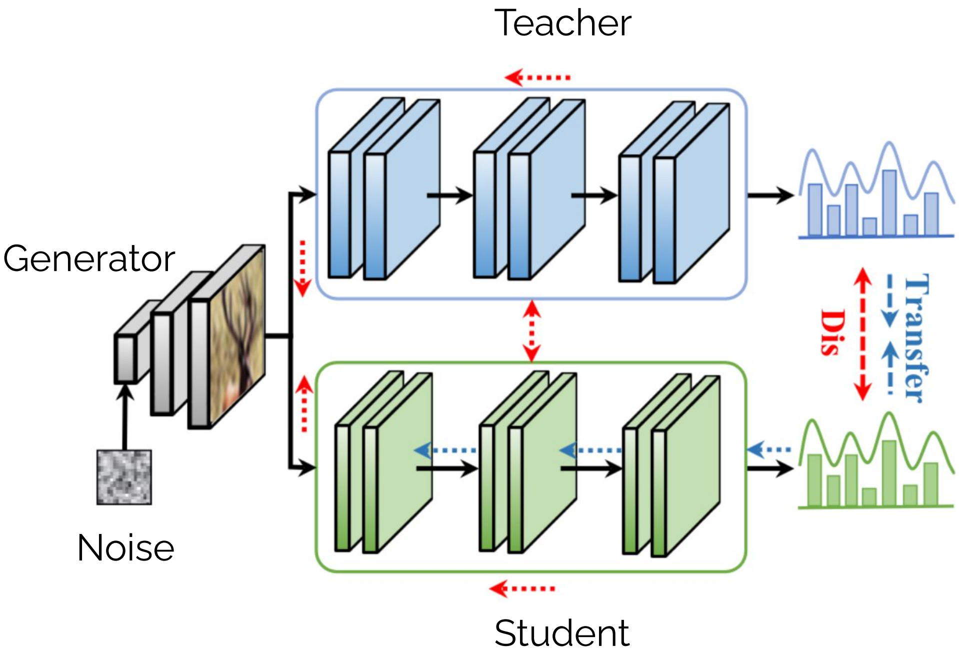
Data-Free Knowledge Distillation (**DFKD**)

Nonetheless, in real-world situations, the **transfer set** is not easily available once the teacher model is trained using them. Therefore, works have been explored in conducting **KD** in the absence of the training data.



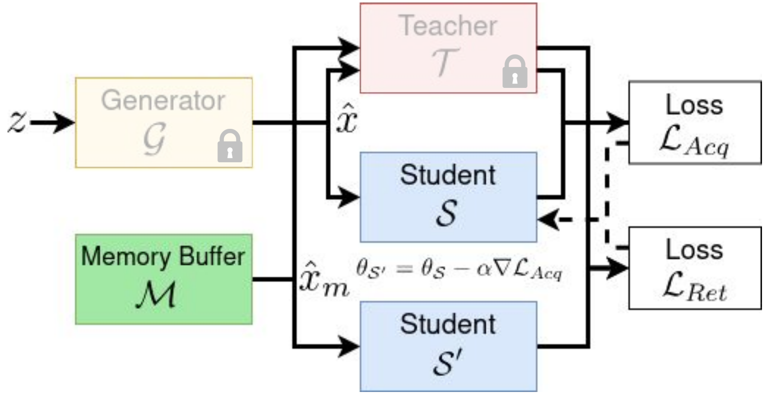
Liu et al., Data-Free Knowledge Transfer: A Survey, arXiv:2112.15278, 2021.

Adversarial DFKD



Liu et al., Data-Free Knowledge Transfer: A Survey, arXiv:2112.15278, 2021.

Learning to Retain while Acquiring



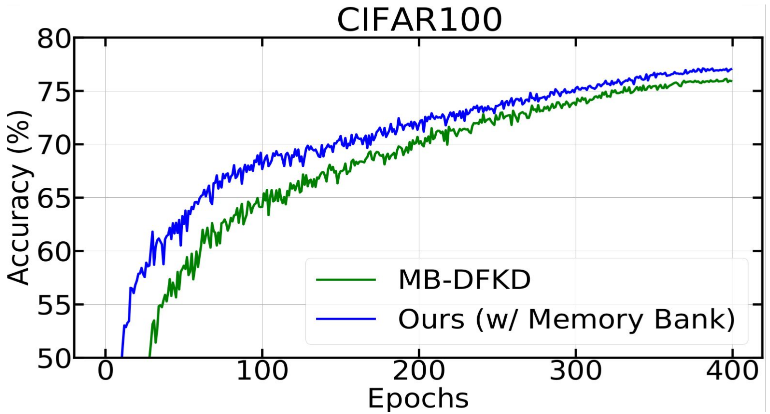
$\mathcal{L}_{Acq}$ Learning from current samples produced by the **generator**

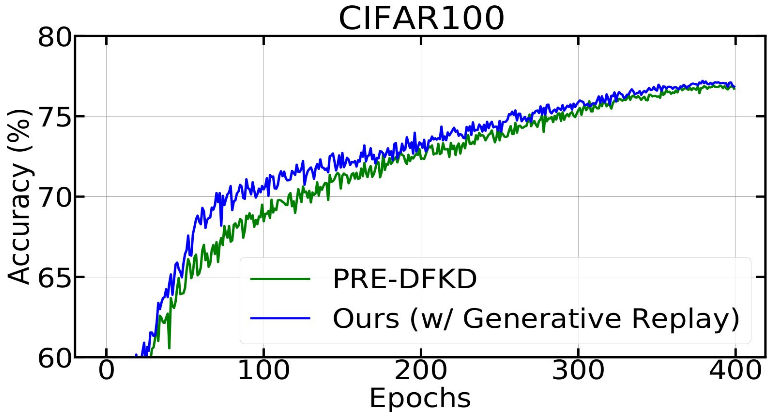
$\mathcal{L}_{Ret}$ Learning from previously generated samples stored in the **memory**

$$\min_{\theta_{\mathcal{S}}} \mathcal{L}_{Acq}(\theta_{\mathcal{S}}) + \mathcal{L}_{Ret}(\theta_{\mathcal{S}} - \alpha \nabla \mathcal{L}_{Acq}(\theta_{\mathcal{S}}))$$

Learning to Retain while Acquiring

Performance Improvements





Goal of Data-Free Knowledge Distillation

PURDUE UNIVERSITY

$$\min_{\theta_{\mathcal{S}}} P_{x\sim\mathcal{D}_{\mathcal{T}}}\left(\arg\max_i p_{\mathcal{S}}^i(x) \neq \arg\max_i p_{\mathcal{T}}^i(x)\right)$$

Goal of Data-Free Knowledge Distillation

$$\min_{\theta_{\mathcal{S}}} P_{x\sim\mathcal{D}_{\mathcal{T}}}\left(\arg\max_i p^i_{\mathcal{S}}(x) \neq \arg\max_i p^i_{\mathcal{T}}(x)\right)$$

$\mathcal{D}_{\mathcal{T}}$ ✗

Goal of Data-Free Knowledge Distillation (DFKD)

PURDUE UNIVERSITY

$$\min_{\theta_{\mathcal{S}}} \mathbb{E}_{\hat{x}\sim\mathcal{D}_{\mathcal{P}}}[\mathcal{L}(\mathcal{T}_{\theta_{\mathcal{T}}}(\hat{x}), \mathcal{S}_{\theta_{\mathcal{S}}}(\hat{x}))]$$

Adversarial DFKD

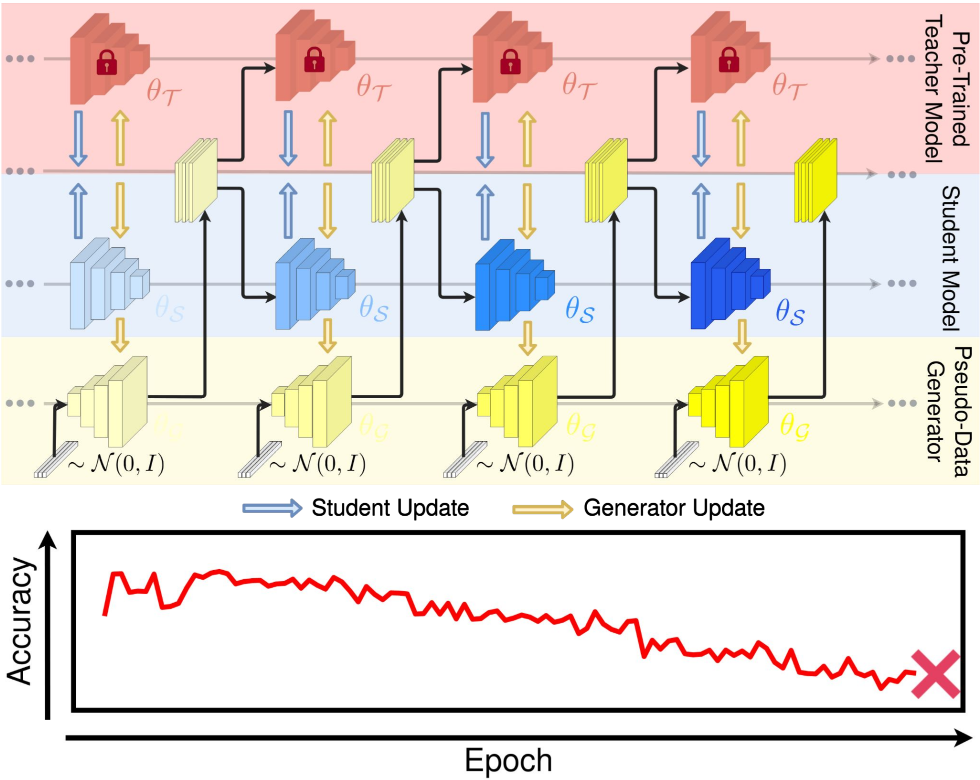
Knowledge-Acquisition

In the typical **Adversarial DFKD** setup, the student update objective with the generated pseudo samples, i.e., the ***Knowledge-Acquisition*** task, is formulated as:

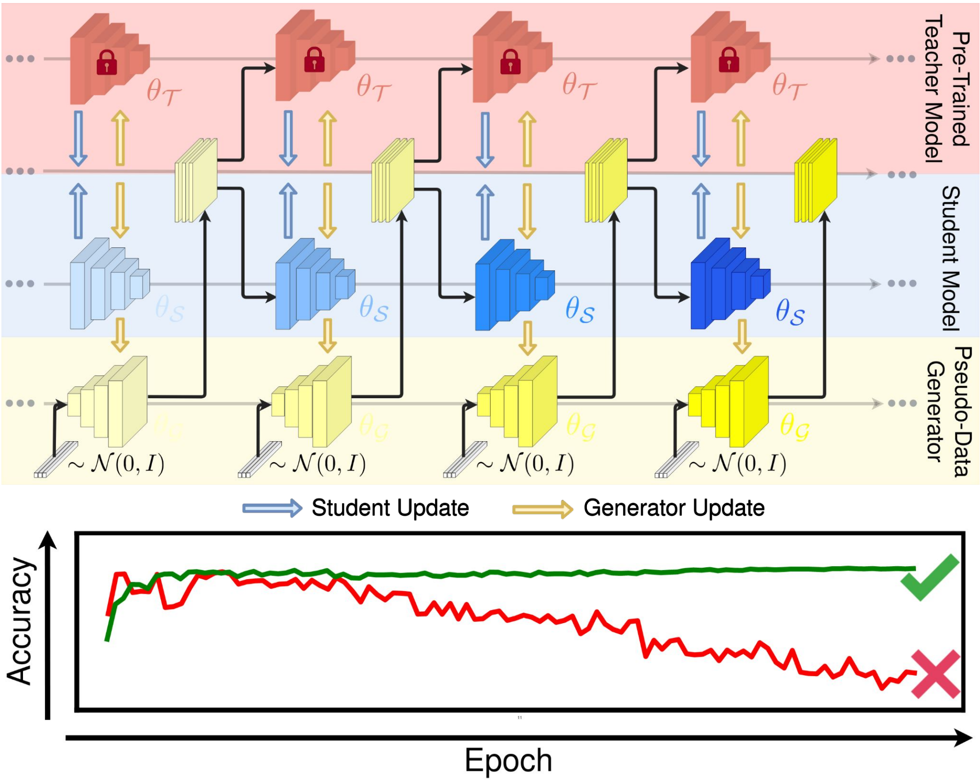
$$\min_{\theta_\mathcal{S}} \mathcal{L}_{Acq}(\theta_\mathcal{S}) = \min_{\theta_\mathcal{S}} \mathbb{E}_{\hat{x}}[\mathcal{L}(\mathcal{T}_{\theta_\mathcal{T}}(\hat{x}), \mathcal{S}_{\theta_\mathcal{S}}(\hat{x}))]$$

$$\hat{x} = \mathcal{G}(z), z \sim \mathcal{N}(0, I)$$

Adversarial DFKD



Adversarial DFKD



Adversarial DFKD

Knowledge-Retention

Moreover, to alleviate the **distribution drift** during **KD** in the adversarial setting, a **memory buffer** of previously encountered sampes is maintained, and samples are **replayed** to help the student recall the knowledge. Therefore, performing ***Knowledge-Retention*** as follows:

$$\min_{\theta_\mathcal{S}} \mathcal{L}_{Ret}(\theta_\mathcal{S}) = \min_{\theta_\mathcal{S}} \mathbb{E}_{\hat{x}_m}[\mathcal{L}(\mathcal{T}_{\theta_\mathcal{T}}(\hat{x}_m), \mathcal{S}_{\theta_\mathcal{S}}(\hat{x}_m))]$$

$$\hat{x}_m \sim \mathcal{M}$$

Adversarial DFKD

Student update objective:

$$\min_{\theta_{\mathcal{S}}} \mathcal{L}_{Acq}(\theta_{\mathcal{S}}) + \mathcal{L}_{Ret}(\theta_{\mathcal{S}})$$

However, the objective above, attempts to simultaneously optimizes ***Knowledge-Retention*** and ***Knowledge-Acquisition***, but does not seek to align the objectives, which leaves them to potentially interfere with one another.

Learning to Retain while Acquiring

PURDUE UNIVERSITY

Proposed Method

- The proposed meta-learning inspired approach, seeks to align the two tasks.
- We take cues from **Model-Agnostic Meta-learning (MAML)**.
- Typically, **MAML-like** methods are framed as a **bi-level optimization problem**, where the objective is defined as:

$$\min_{\omega} \mathcal{L}_{\text{outer}}(\arg\min_{\omega} \mathcal{L}_{\text{inner}}(\omega, \mathcal{D}_{\text{train}}), \mathcal{D}_{\text{test}})$$

Learning to Retain while Acquiring

PURDUE UNIVERSITY

Proposed Method

- Likewise, we pose ***Knowledge-Acquisition*** and ***Knowledge-Retention*** as **meta-train** and **meta-test**, respectively.

- We perform a **single gradient descent step** on the ***Knowledge-Acquisition*** objective, using samples from current distribution, and then optimize the student parameters on the ***Knowledge-Retention*** objective, using the samples in the memory.

- Hence, the overall student learning objective is defined as follows:

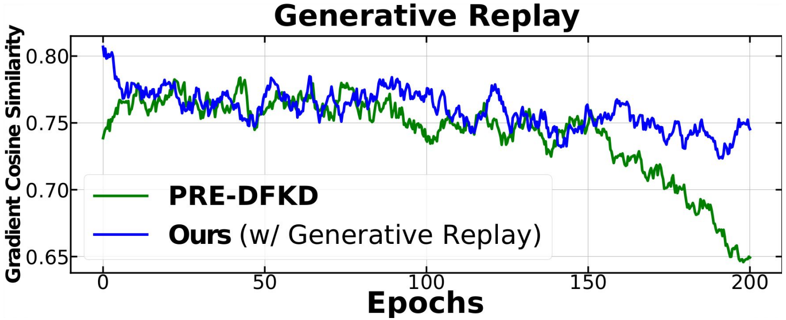
$$\min_{\theta_{\mathcal{S}}} \mathcal{L}_{Acq}(\theta_{\mathcal{S}}) + \mathcal{L}_{Ret}(\theta'_{\mathcal{S}}) = \min_{\theta_{\mathcal{S}}} \mathcal{L}_{Acq}(\theta_{\mathcal{S}}) + \mathcal{L}_{Ret}(\theta_{\mathcal{S}} - \alpha \nabla \mathcal{L}_{Acq}(\theta_{\mathcal{S}})).$$

Learning to Retain while Acquiring

Implicit Gradient Matching

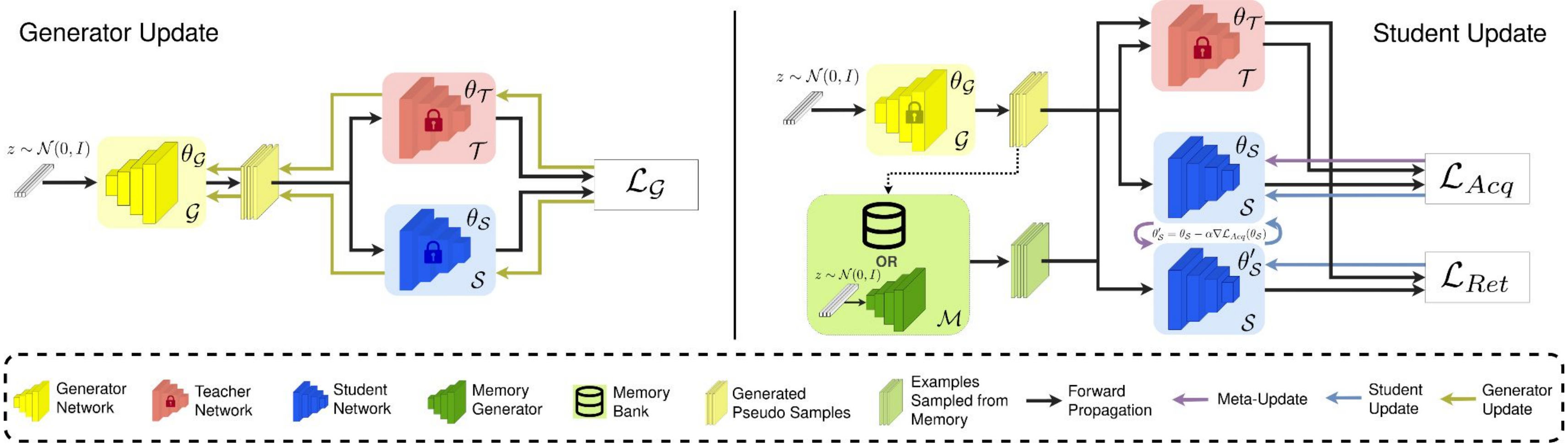
Analyzing the objective using Taylor's expansion.

$$\frac{\partial \mathcal{L}_{Ret}(\theta'_{\mathcal{S}})}{\partial \theta_{\mathcal{S}}} = \frac{\partial \mathcal{L}_{Ret}(\theta_{\mathcal{S}} - \alpha \nabla \mathcal{L}_{Acq}(\theta_{\mathcal{S}}))}{\partial \theta_{\mathcal{S}}} = \nabla \mathcal{L}_{Ret}(\theta_{\mathcal{S}}) - \alpha \nabla \underbrace{\left(\nabla \mathcal{L}_{Ret}(\theta_{\mathcal{S}}) . \nabla \mathcal{L}_{Acq}(\theta_{\mathcal{S}}) \right)}_{Gradient\ \ Alignment} + \mathcal{O}(\alpha^2)$$



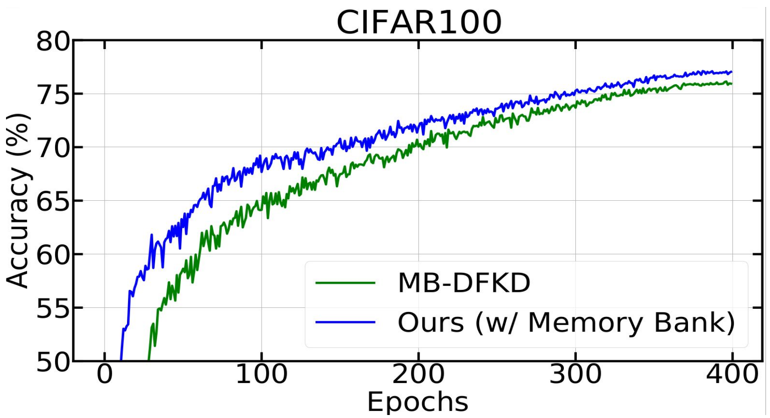
Learning to Retain while Acquiring

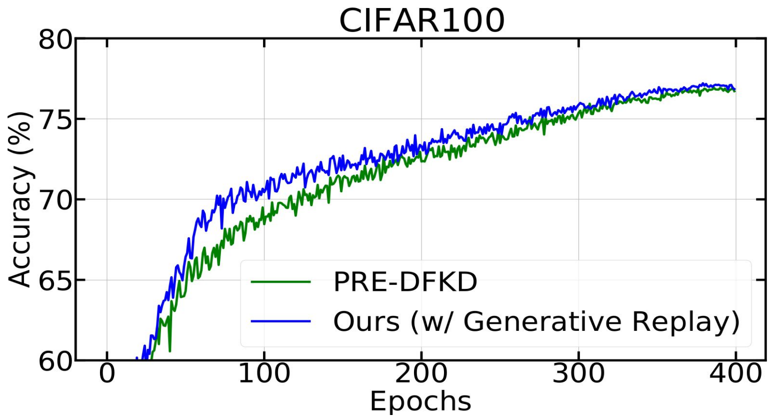
Overview of the proposed method



Learning to Retain while Acquiring

Learning evolution improvements





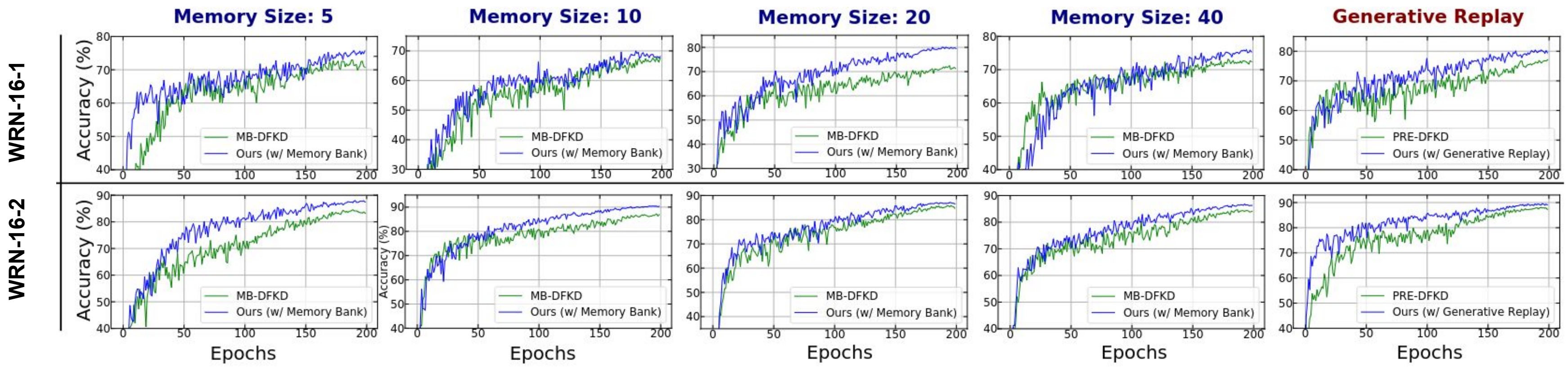
Learning to Retain while Acquiring

Learning evolution improvements

	$\mu \uparrow$	$\sigma^2 \downarrow$	$\mathrm{Acc}_{\mathrm{max}} \uparrow$
MB-DFKD	66.05	207.29	76.14
PRE-DFKD	70.23	86.63	76.93
Ours (w/ Memory Bank)	**69.87**	**75.67**	**77.11**
Ours (w/ Generative Replay)	**71.49**	**60.17**	**77.21**

Learning to Retain while Acquiring

Improvements across replay schemes



Thank You!