THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學
Opening Minds • Shaping the Future • 啟迪思維 • 成就未來

大连理工大学 计算机科学与技术学院
DALIAN UNIVERSITY OF TECHNOLOGY SCHOOL OF COMPUTER SCIENCE AND TECHNOLOGY

THE WEB CONFERENCE ACM

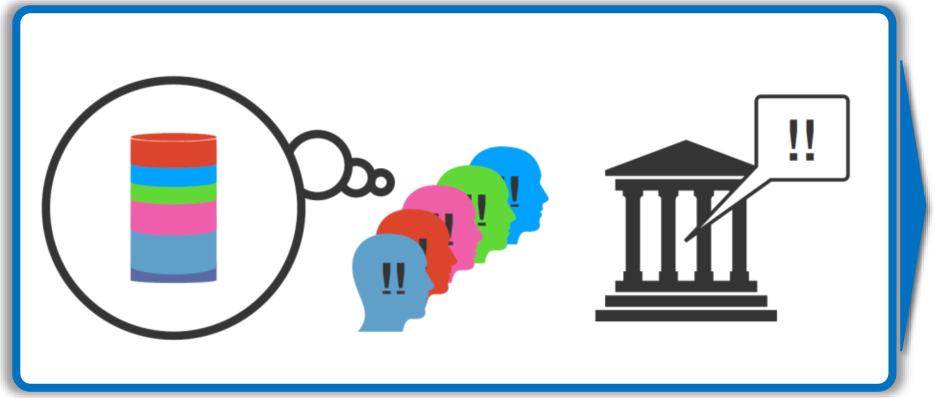
Federated Unlearning via Class-Discriminative Pruning

JUNXIAO WANG, SONG GUO, XIN XIE, HENG QI

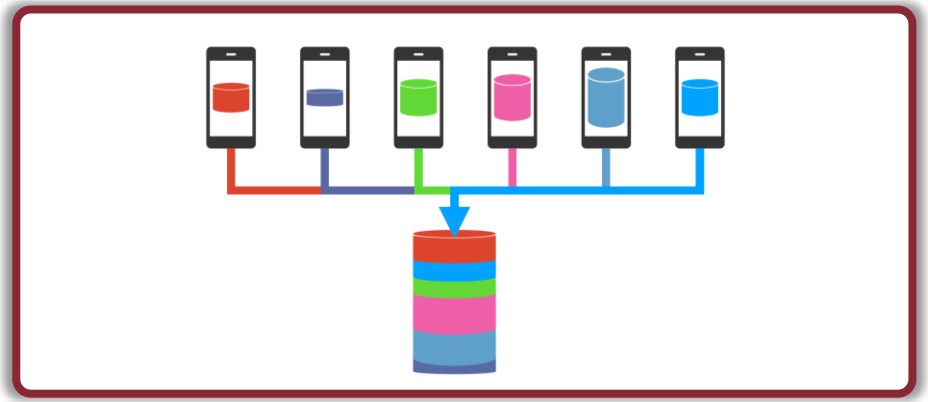
PolyU Edge Intelligence Lab

DEPARTMENT OF COMPUTING
電子計算學系

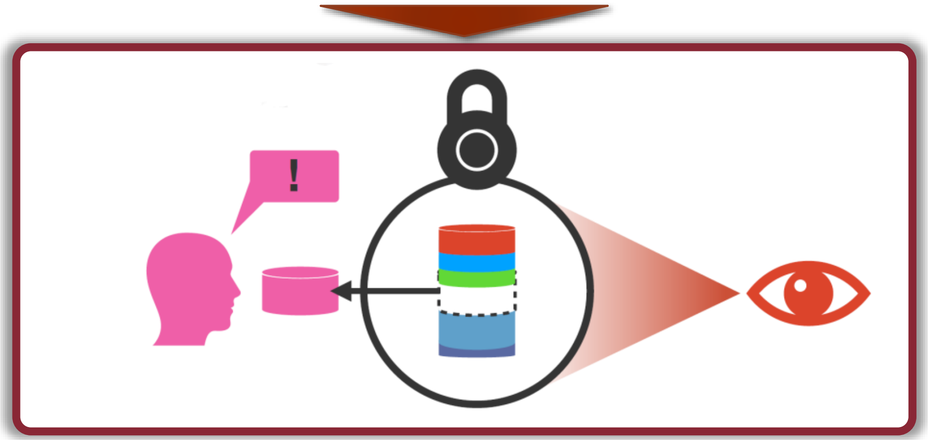
About Federated Unlearning



(a) Federated Unlearning



(b) Federated Learning



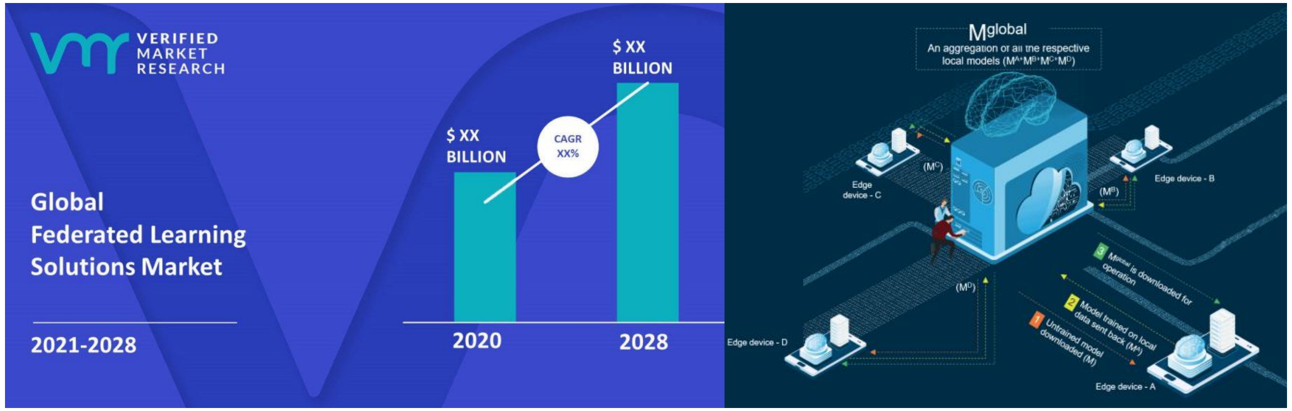
(c) Machine Unlearning

Topic: How to do <u>Machine Unlearning</u> in <u>Federated Settings</u>?

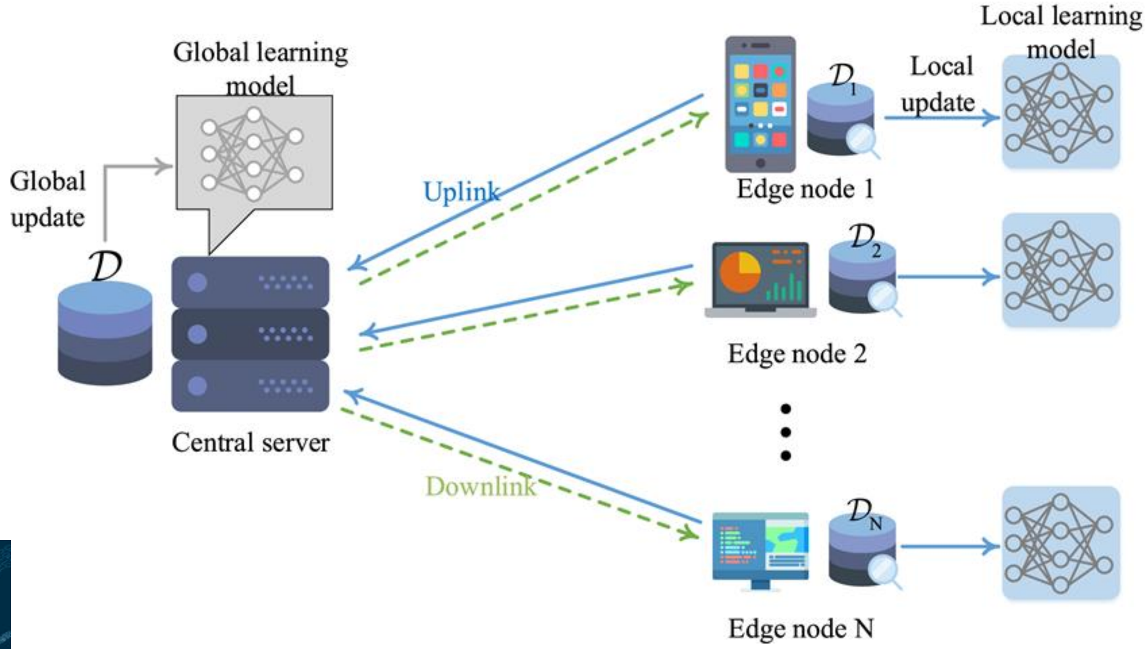
Federated Learning and Its Settings

tensorflow/
federated

(a) TensorFlow Federated (TFF): **a framework for implementing Federated Learning**



(b) Market Statistics and Application of FL

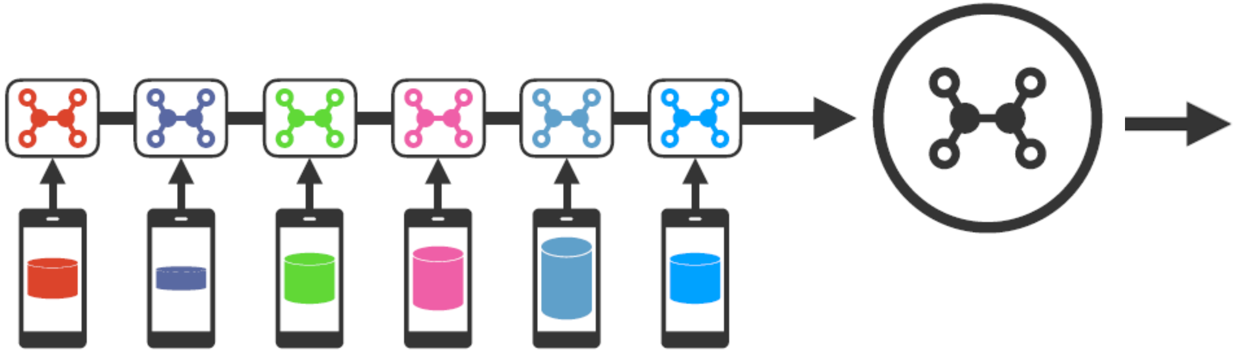


(c) FL workflow: How Federated Learning performs

[1]https://www.tensorflow.org/federated/
[2]https://www.everestgrp.com/
[3]https://www.verifiedmarketresearch.com/

What's Machine Unlearning

Privacy Legislation – Selective Forgetting from Trained Models



Not to Learn, Unlearn, Forget

Particular Data Samples

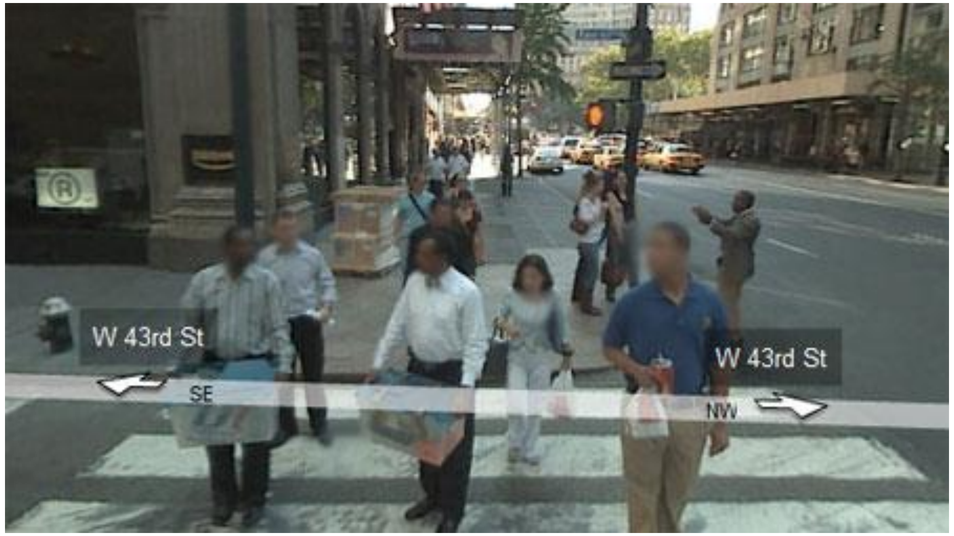
Machine **Learning**

Machine **Unlearning**

Class-wise Machine Unlearning

We focus on Class-wise Machine Unlearning in Federated Learning Settings.

A specific class of data needs to be removed from Trained FL Model.



Scenario
Task: Image Classification
Model: CNNs

Automotive Domain: Street View Images with Facial

Class-wise Machine Unlearning

General Way
Retrain from Scratch

Pros
Convincing in terms of Forgetting.

Cons
Expensive Overhead in Computation.

Class-wise Machine Unlearning

Approximate Way - Directly Update Model

Most of them (Centralized Methods) can be categorized into one of three groups:

1) Fisher unlearning method. **2) I**nfluence unlearning method.

3) Gradient unlearning method.

Pros

Speedup unlearning process with lower computation.

Cons

Heavily rely on the global data access.

Class-wise Federated Unlearning

*Practical Constrains in FL Settings

1) Lack of direct data access.

2) Communication cost.

3) Non-IID data distribution.

With incomplete and severely biased local training data …

Addressing these challenges to unlearn class is a key contribution of our work.

Class-wise Federated Unlearning

*Novel Unlearning Paradigm in FL is Required!

Practical Constrains in FL:
Data used for training are impossible to access globally

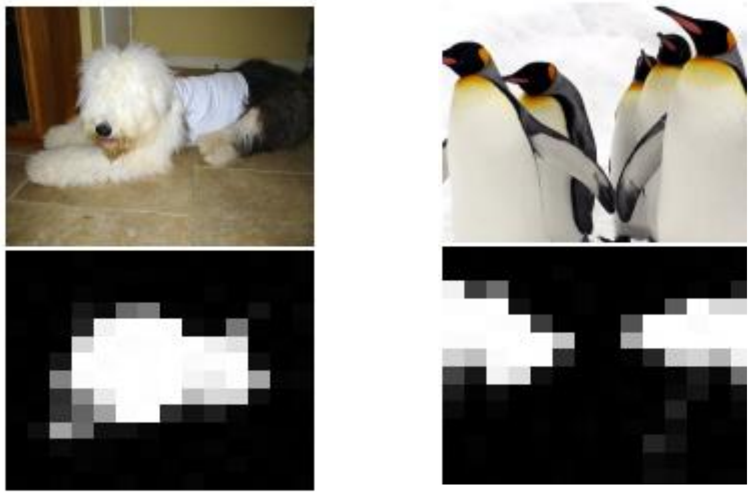
Existing Approximate Unlearning Methods ✗

We need to revisit the class discrimination of model …

↓

1) Find the most discriminative channels of the target class,
2) then prune those channels.

Visualization of Channels' Class Discrimination



(a) The channels highlight head information

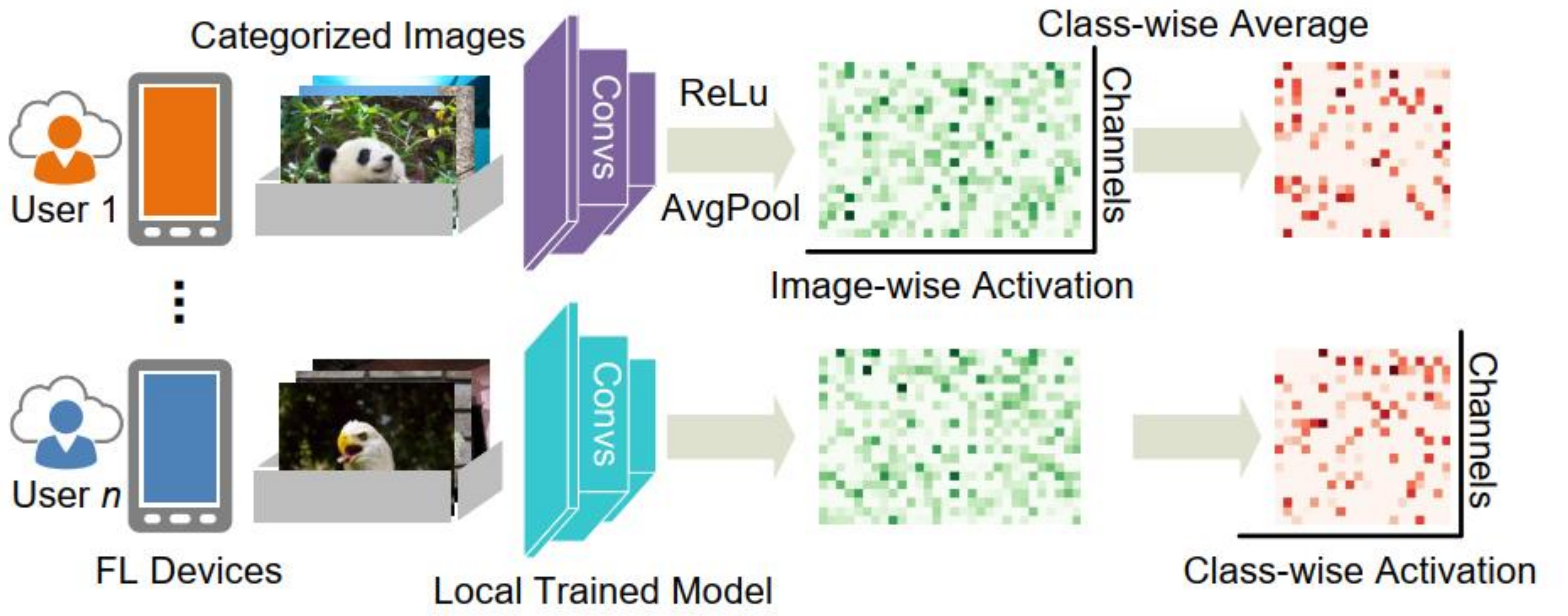
(b) The channels highlight text information

Different channels have a varying contribution to different class in image classification …

1) Find the most discriminative channels of the target class,
2) then prune those channels.

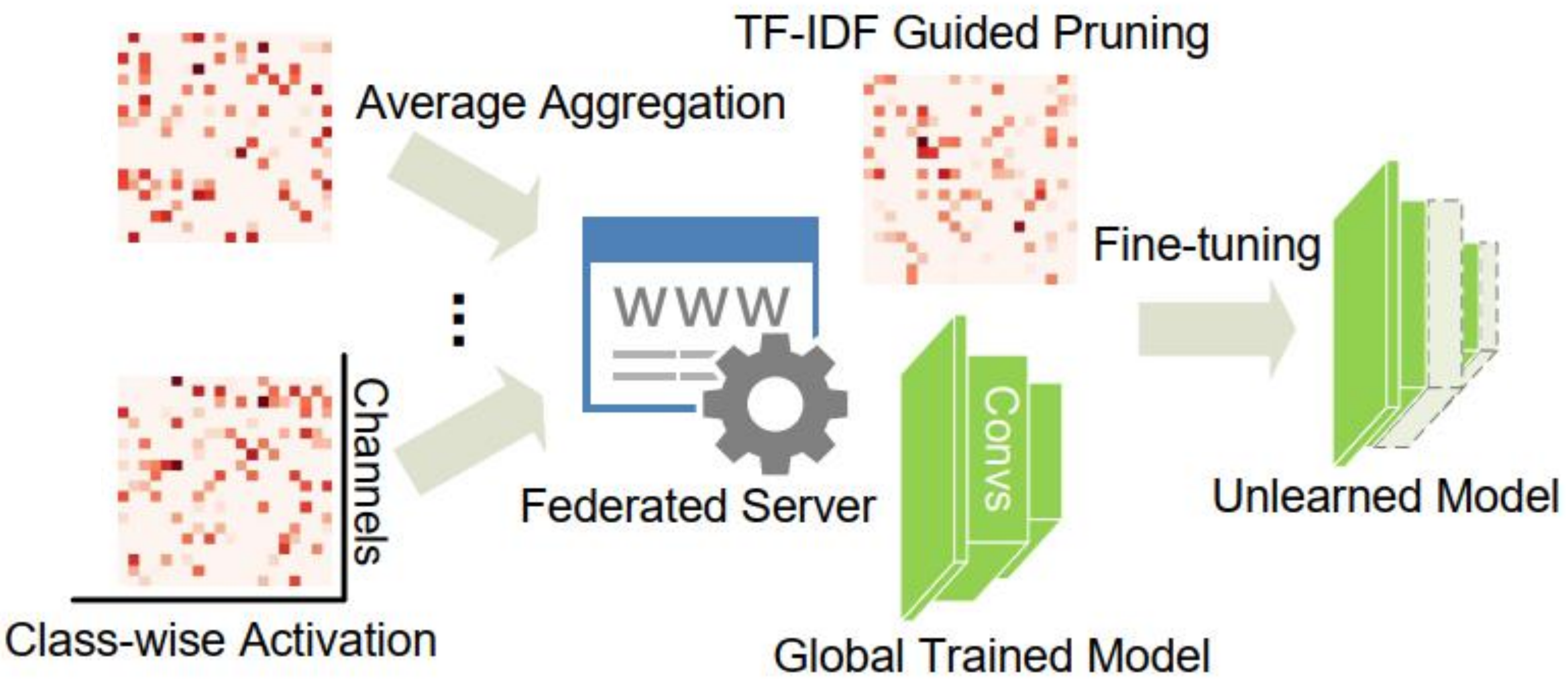
Federated Unlearning Framework

Local Channel Scoring on their Class Discrimination



Federated Unlearning Framework

Global Pruning on their Most Discriminative Channels

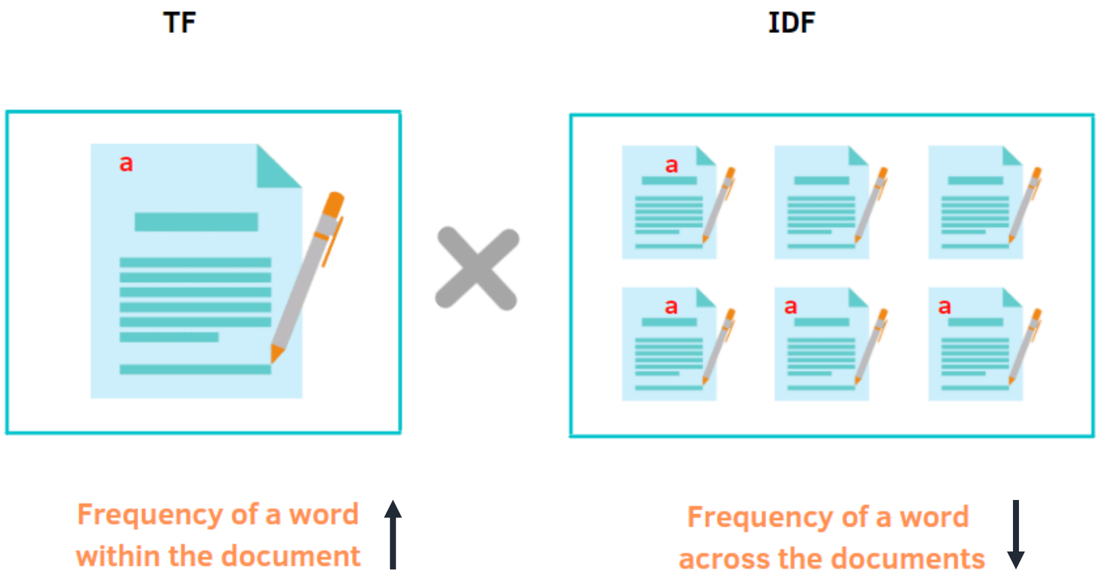


Federated Unlearning Framework

$\boxed{\text{Channels} \longrightarrow \text{Class Discrimination}}$ $\longrightarrow$ Channel Pruning $\longrightarrow$ Class Unlearning

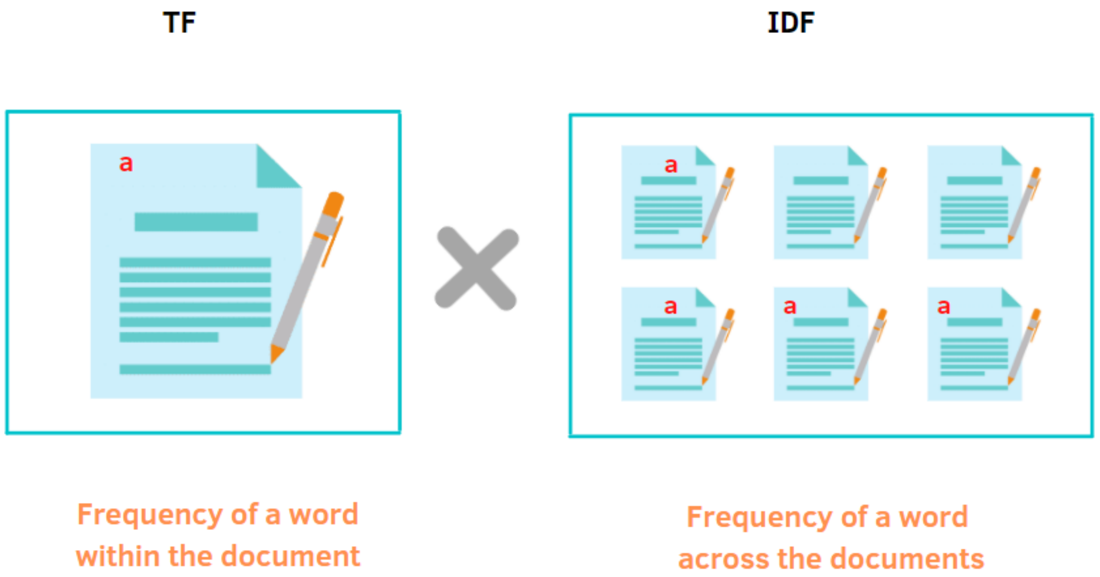
Term Frequency Inverse Document Frequency (TF-IDF)

TF-IDF 1) a statistical measure that evaluates how relevant a word is to a document in a set of documents,
2) very useful for scoring words in machine learning algorithms for Natural Language Processing.



Federated Unlearning Framework

Channels → Class Discrimination → Channel Pruning → Class Unlearning



TF-IDF in Federated Unlearning

Word -> Activations of a channel
Document -> Feature map of a Class

Federated Unlearning Framework

$\boxed{\text{Channels} \longrightarrow \text{Class Discrimination}}$ $\longrightarrow$ Channel Pruning $\longrightarrow$ Class Unlearning

Find the Most Discriminative Channels for the Target Class

It doesn't matter if it's through the TF-IDF idea or something else …

TF-IDF is straightforward to project the relationship between channels and classes.

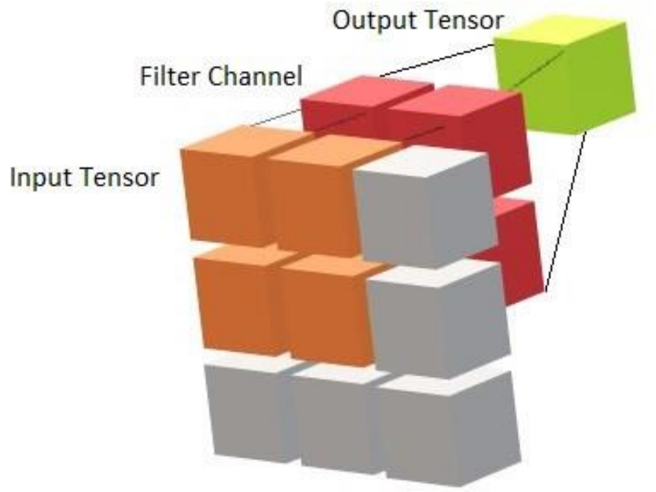
Federated Unlearning Framework

Channels → Class Discrimination → Channel Pruning → Class Unlearning

Channel Pruning 1) structured model update,
2) well supported by general-purpose hardware,
3) well adapted to Basic Linear Algebra Subprograms (BLAS) libraries.

Original FL Model →

Channels to be Pruned →



→ Pruned FL Model

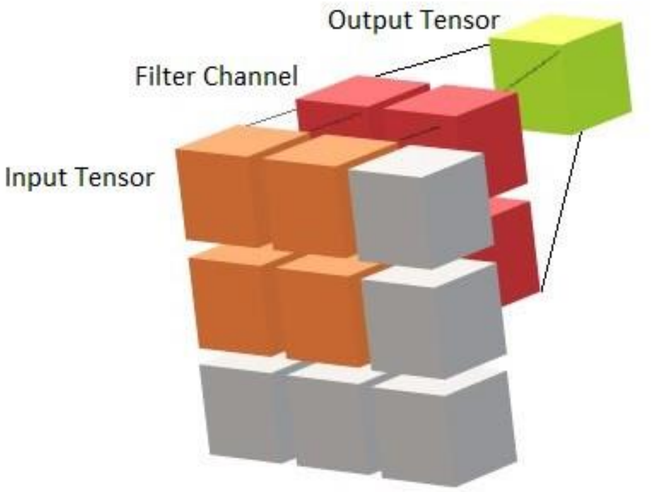
Federated Unlearning Framework

Channels → Class Discrimination → Channel Pruning → Class Unlearning

One-shot Channel Pruning with Pruning ratio (Hyper-parameter),
Specific weights of the discriminative channels are zeroed from models.

Original
FL Model →

Channels
to be Pruned →



→ Pruned
FL Model

Federated Unlearning Framework

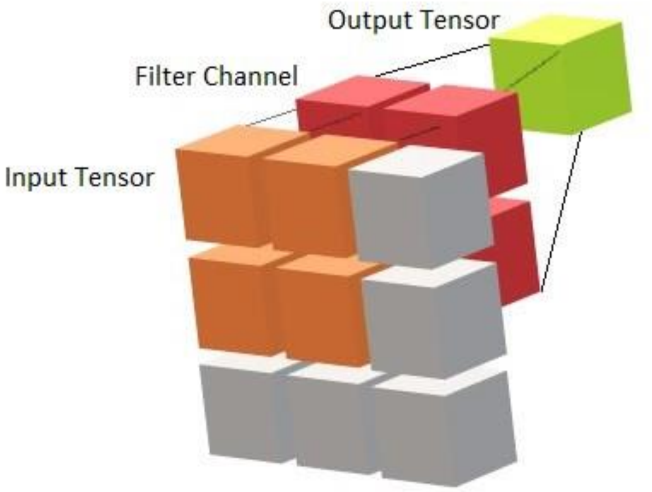
Channels → Class Discrimination → Channel Pruning → Class Unlearning

Unlearning multiple classes

Pruning process is executed multiple times, removing one class each time.

Original FL Model →

Channels to be Pruned →



× N → Pruned FL Model

Federated Unlearning Framework

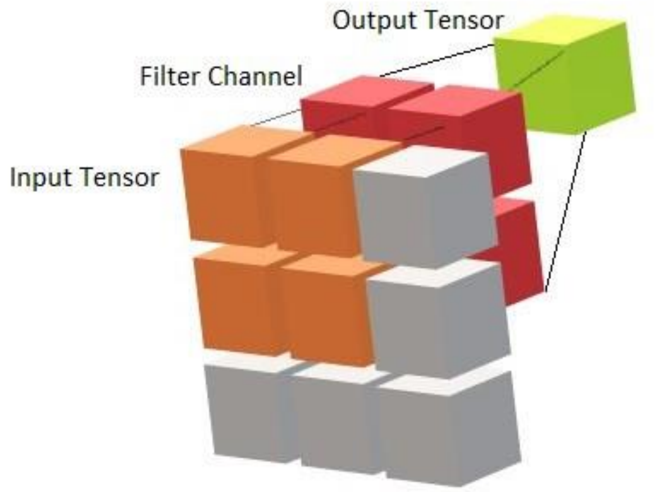
Channels → Class Discrimination → Channel Pruning → Class Unlearning

Channel Pruning is followed by the Fine-tuning process

1) Same as the normal training procedure of federated learning,
2) Compensate accuracy degradation of the pruned model,
3) Prune once and retrain to fine-tune.

Original FL Model →

Channels to be Pruned →



→ Pruned FL Model $\xrightarrow{\text{Fine-tuning}}$

Federated Unlearning Framework

Discussion

Can this federated unlearning framework be applied to centralized scenarios?

Of course, it can …

1) Measure of class discriminative channels can be easily obtained with global access to the data,
2) Data privacy protection and communication overhead optimization is no longer required.

Far greater diversity of class-unlearning designs should be there …

Yet it's specific to federated settings.

1) **L**ack of direct data access,
2) **N**on-IID data distribution,
3) **C**ommunication cost.

Experimental Settings

- Datasets
  - CIFAR10, CIFAR100. → Federated Settings
    1) Incomplete participant data,
    2) Biased participant data towards certain classes.
- Model
  - [1] ResNet20, ResNet32, ResNet44, ResNet56.
  - [2] VGG11, VGG13, VGG16, VGG19.
- Baseline
  - [1] Gold Standard – Retraining from scratch with the remaining data.
  - [2] Centralized Approximate Unlearning – Fisher unlearning method.
- Cared Metrics
  - [1] Unlearning speedup ratio.
  - [2] Information erasure effect.

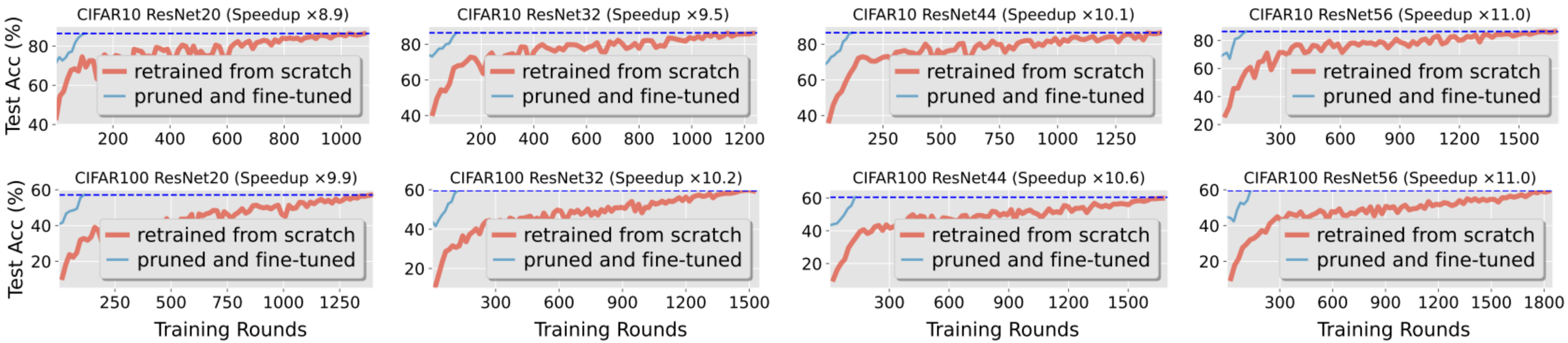
Experimental Settings

- Datasets
  - CIFAR10, CIFAR100. → Federated Settings
    1) Incomplete participant data,
    2) Biased participant data towards certain classes.
- Model
  - [1] ResNet20, ResNet32, ResNet44, ResNet56.
  - [2] VGG11, VGG13, VGG16, VGG19.
- Baseline
  - [1] Gold Standard – Retraining from scratch with the remaining data.
  - [2] Centralized Approximate Unlearning – Fisher unlearning method.
- Cared Metrics
  - [1] Unlearning speedup ratio.
  - [2] Information erasure effect.

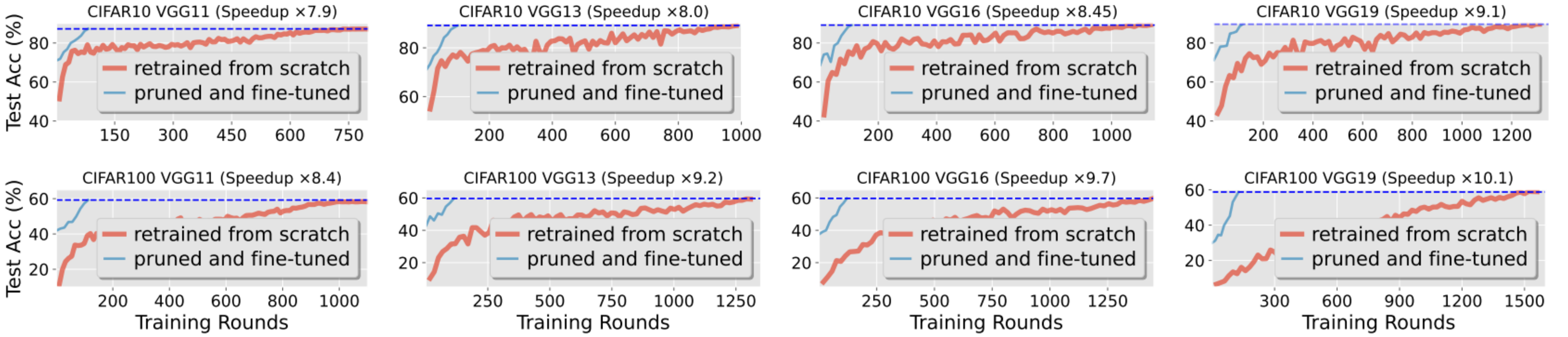
  → Efficiency and Efficacy
    1) Unlearning process time ↓
    2) Gap with full retraining ↓

Unlearning Speedup

- ResNet



- VGG



Information Erasure

- Baseline – Full retraining from scratch

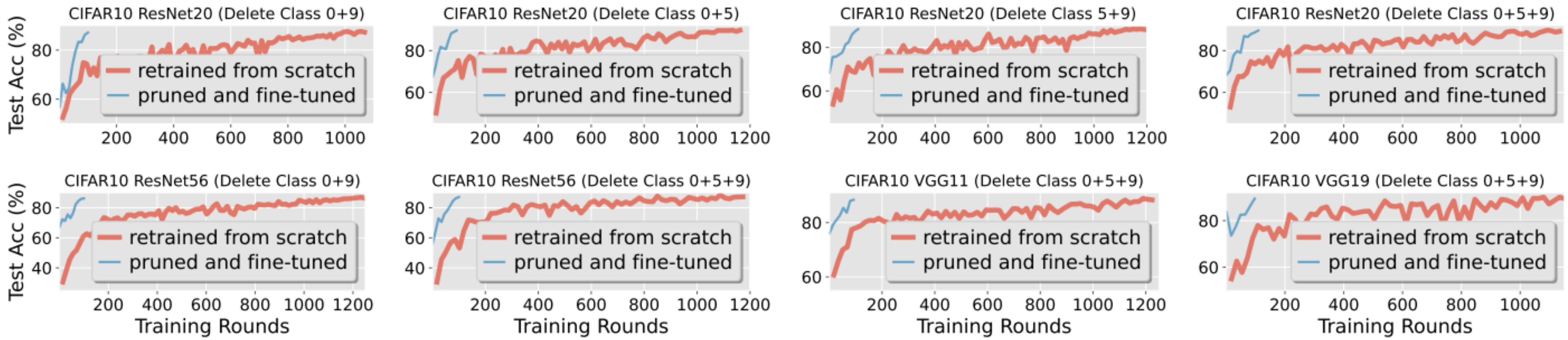
	CIFAR10								CIFAR100							
	Raw model		After-pruned		Fine-tuned		Re-trained		Raw model		After-pruned		Fine-tuned		Re-trained	
Accuracy	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**
ResNet20	91.50%	83.33%	00.00%	20.79%	00.00%	86.40%	00.00%	86.33%	54.00%	50.01%	00.00%	05.38%	00.00%	57.17%	00.00%	57.11%
ResNet32	94.20%	83.71%	00.00%	11.58%	00.00%	86.40%	00.00%	86.14%	52.00%	51.67%	00.00%	01.06%	00.00%	59.62%	00.00%	59.42%
ResNet44	89.90%	83.94%	00.00%	22.19%	00.00%	86.48%	00.00%	86.34%	48.00%	53.25%	00.00%	01.22%	00.00%	60.41%	00.00%	59.85%
ResNet56	93.10%	84.02%	00.00%	11.11%	00.00%	86.42%	00.00%	86.38%	44.00%	52.91%	00.00%	01.32%	00.00%	59.60%	00.00%	59.28%
VGG11	88.20%	84.72%	00.00%	18.29%	00.00%	87.24%	00.00%	87.13%	50.00%	53.55%	00.00%	01.28%	00.00%	59.25%	00.00%	58.20%
VGG13	91.50%	84.19%	00.00%	15.17%	00.00%	89.18%	00.00%	89.09%	60.00%	51.88%	00.00%	03.82%	00.00%	59.65%	00.00%	59.27%
VGG16	91.60%	84.38%	00.00%	17.79%	00.00%	89.20%	00.00%	89.30%	44.00%	50.34%	00.00%	01.46%	00.00%	59.72%	00.00%	59.57%
VGG19	88.80%	83.53%	00.00%	11.11%	00.00%	89.72%	00.00%	89.62%	52.00%	52.15%	00.00%	01.02%	00.00%	58.78%	00.00%	58.96%

- Baseline – Fisher unlearning method

	CIFAR10						CIFAR100					
	Rounds of training			Test accuracy on U/R-set			Rounds of training			Test accuracy on U/R-set		
Bias probability	**0.10**	**0.45**	**1.00**	**0.10**	**0.45**	**1.00**	**0.01**	**0.35**	**1.00**	**0.01**	**0.35**	**1.00**
Our method	113	135	181	00.00/80.13%	00.00/74.45%	00.00/66.87%	110	163	235	00.00/50.34%	00.00/46.99%	00.00/39.45%
Fisher method	610	750	1110	22.47/80.00%	28.54/73.79%	19.10/66.04%	700	820	1190	15.33/49.86%	14.71/45.30%	17.09/38.32%

Multi Class Removal

- Unlearning speedup



- Information erasure

ResNet20 CIFAR10	Raw model		First class pruned		Last class pruned		Fine-tuned		Re-trained		
Model Accuracy	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**Speedup**
Delete class 0+9 from [0-9]	88.70%	83.01%	02.10%	24.57%	00.00%	32.41%	00.00%	87.12%	00.00%	87.29%	×8.71
Delete class 0+5 from [0-9]	81.90%	84.71%	00.25%	25.04%	00.00%	26.74%	00.00%	89.62%	00.00%	89.75%	×10.62
Delete class 5+9 from [0-9]	84.70%	84.01%	02.10%	31.74%	00.00%	37.71%	00.00%	88.37%	00.00%	88.21%	×8.92
Delete class 0+5+9 from [0-9]	85.10%	83.74%	01.57%	28.01%	00.00%	30.00%	00.00%	89.62%	00.00%	89.23%	×8.45

ResNet56 CIFAR10	Raw model		First class pruned		Last class pruned		Fine-tuned		Re-trained		
Model Accuracy	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**Speedup**
Delete class 0+9 from [0-9]	91.75%	83.23%	01.20%	12.50%	00.00%	19.38%	00.00%	87.22%	00.00%	86.38%	×10.33
Delete class 0+5+9 from [0-9]	85.57%	84.66%	03.82%	14.29%	00.00%	33.33%	00.00%	87.10%	00.00%	87.23%	×9.66

VGG11 CIFAR10	Raw model		First class pruned		Last class pruned		Fine-tuned		Re-trained		
Model Accuracy	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**U-set**	**R-set**	**Speedup**
Delete class 0+5+9 from [0-9]	83.60%	85.70%	00.63%	17.77%	00.00%	20.61%	00.00%	88.40%	00.00%	88.34%	×10.77

Take Home Message

- Class discrimination of channels is the key for class unlearning,

  especially under the federated settings.

1) Find the most discriminative channels of the target class,
2) then remove those discriminative channels.

1) **L**ack of direct data access,
2) **N**on-IID data distribution,
3) **C**ommunication cost.

Take Home Message

- Sample-wise unlearning is a more strict problem due to its challenges, especially under the federated settings.

1) Remove specific data samples from the trained model,
2) Still maintaining output knowledge of that class.

1) **Requires a more elaborate design**,
2) **data point contributions to the model are difficult to evaluate without access to the raw data**.

Thank you!

THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學