Network and System Security Laboratory

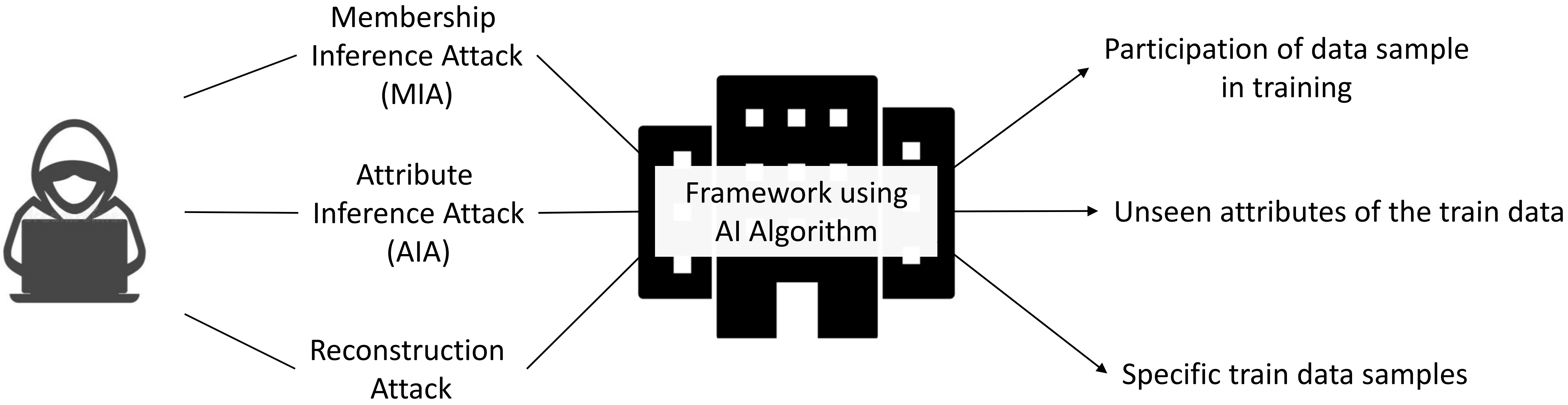
Closing the Loophole: Rethinking Reconstruction Attacks in Federated Learning from a Privacy Standpoint

SEUNG HO NA, HYEONG GWON HONG, JUNMO KIM, SEUNGWON SHIN

KAIST

Privacy Attacks in Federated Learning (FL)

- Privacy Attack
  - Attacks aiming at leaking private information



Privacy-preserving Technologies

- Differential privacy
  - Theoretical approach to quantifying information leakage
- Encryption methods
  - Key encryption schemes such as secure multi-party computation protocols
  - Incur heavy computation and communication costs

Original Image

Reconstruction Attack on Differential Privacy Model

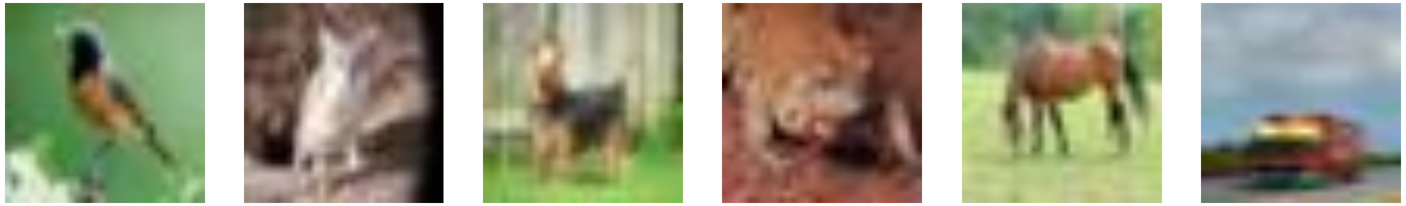
Privacy-preserving Technologies

- Differential privacy
  - Theoretical approach to quantifying information leakage
- Encryption methods
  - Key encryption sc… …on protocols
  - Incur heavy computation and communication costs

Research Question 1

What is this inconsistency between privacy attacks and privacy-preserving methods?

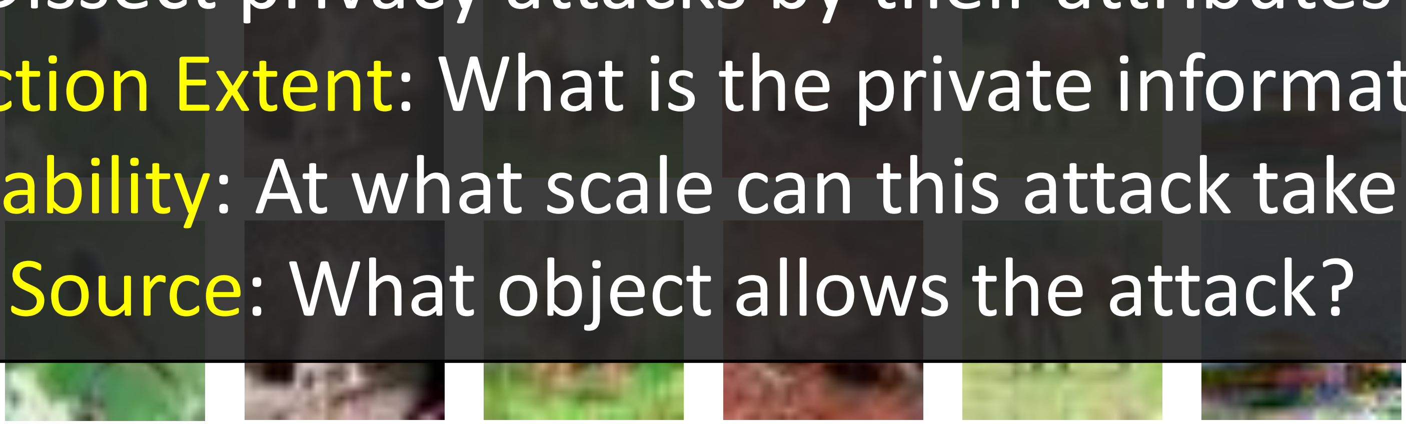
Original Image



Reconstruction Attack on Differential Privacy Model

Privacy-preserving Technologies

- Differential privacy
  - Theoretical approach to quantifying information leakage
- Encryption methods
  - Key encryption sc… on protocols
  - Incur heavy computation and communication costs



Research Question 1

What is this inconsistency between privacy attacks and privacy-preserving methods?

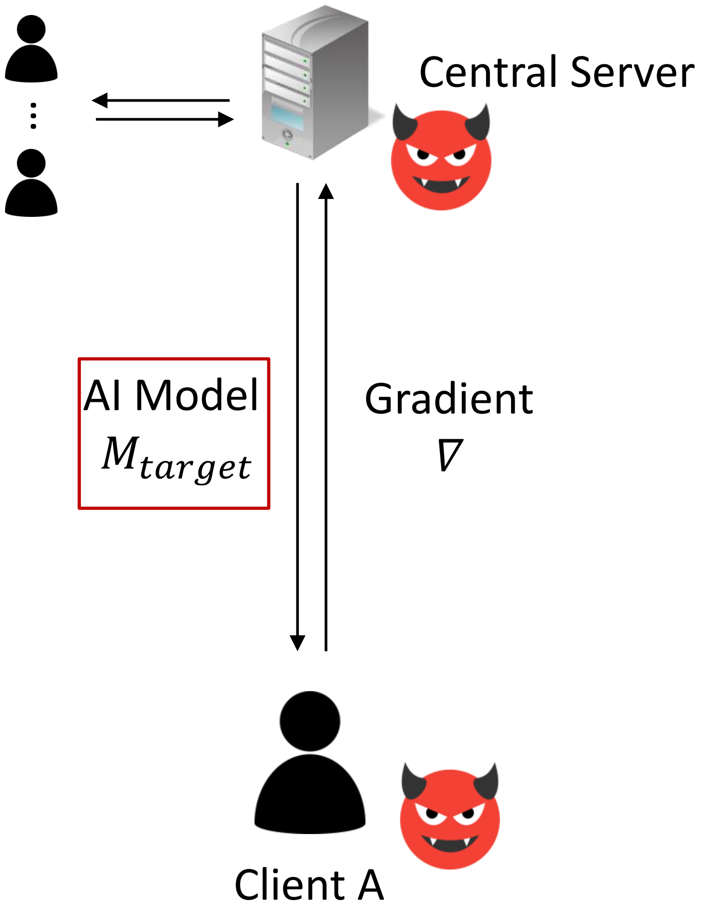
➡ Dissect privacy attacks by their attributes

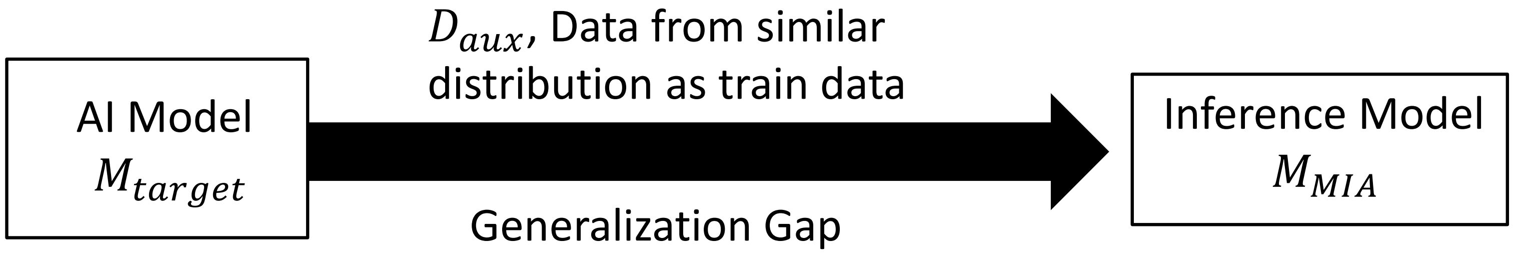
Extraction Extent: What is the private information?

Transferability: At what scale can this attack take place?

Source: What object allows the attack?

Membership Inference Attack in FL

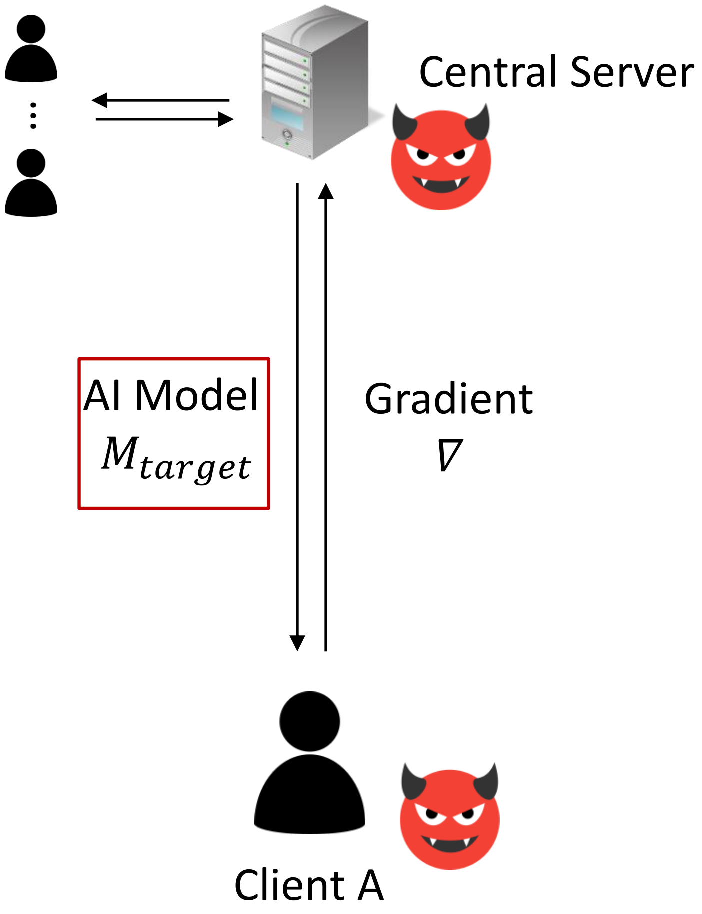




- Develop inference model to know data membership
- $M_{MIA}(x, M_{target}, D_{aux}) = \begin{cases} 1, x \in D_{train} \\ 0, x \in D_{test} \end{cases}$

Extraction Extent: private meta-information
Transferability: membership of all data X can be known
Source: M_{target}

Attribute Inference Attacks in FL



AI Model M_{target} → Inference Model M_{AIA}

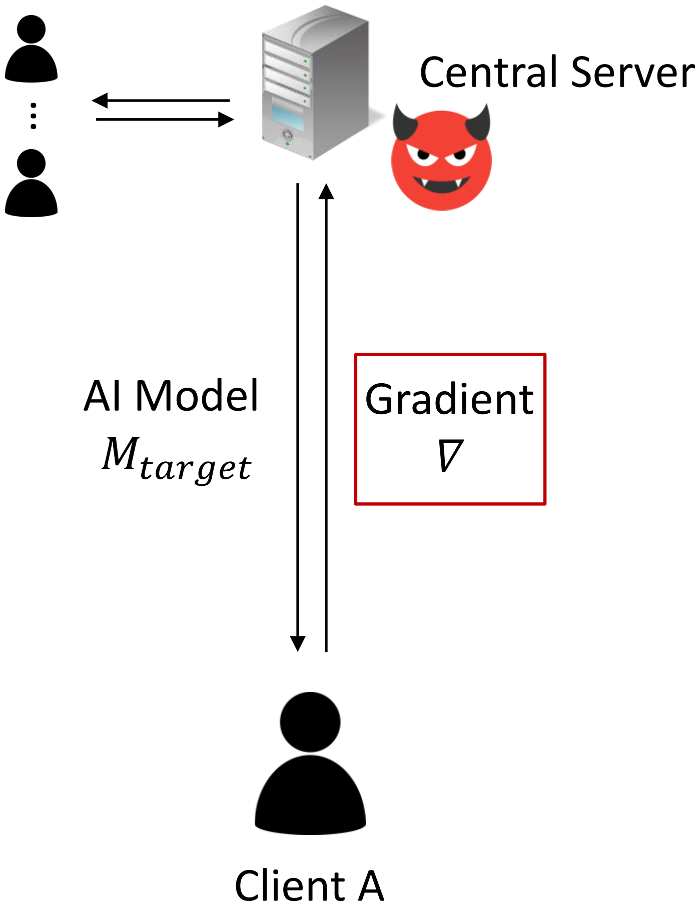
D_{aux}, Black box oracle mapping X→Z

Algorithms for Repurposing

- Develop inference model to exploit unseen characteristics of the data
- Given $M_{target}(x_i) = y_i$,
- $M_{AIA}(x_i, M_{target}, D_{aux}) = z_i$

Extraction Extent: private meta-information
Transferability: attributes of all data X can be known
Source: M_{target}

Reconstruction Attack Attributes in FL

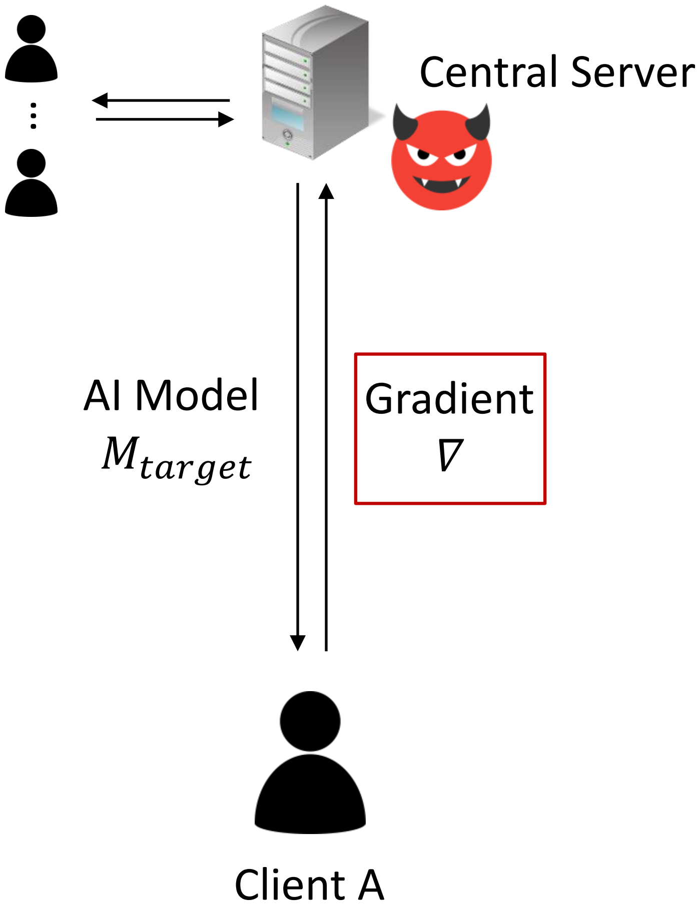


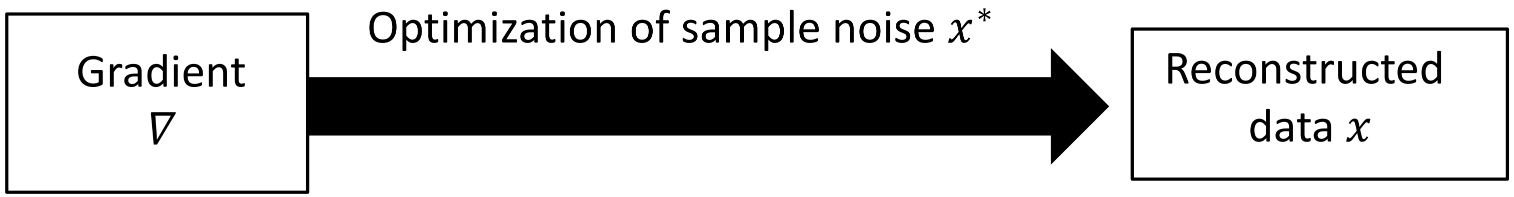
Gradient ∇ → Optimization of sample noise x^* → Reconstructed data x

- Recover original data from gradient information
- $\underset{x^*}{\operatorname{argmin}}\ 1 - \frac{<\nabla L_M(x,y), \nabla L_M(x^*,y)>}{|\nabla L_M(x,y)||\nabla L_M(x^*,y)|} + TV(x^*)$

Extraction Extent: private raw data
Transferability: attack specific to the input gradient
Source: ∇

Reconstruction Attack Attributes in FL





- Recover original data from gradient information
- $\underset{x^*}{\operatorname{argmin}}\, 1 - \frac{<\nabla L_M(x,y), \nabla L_M(x^*,y)>}{|\nabla L_M(x,y)||\nabla L_M(x^*,y)|} + TV(x^*)$

Extraction Extent: private raw data
Transferability: attack specific to the input gradient
Source: ∇

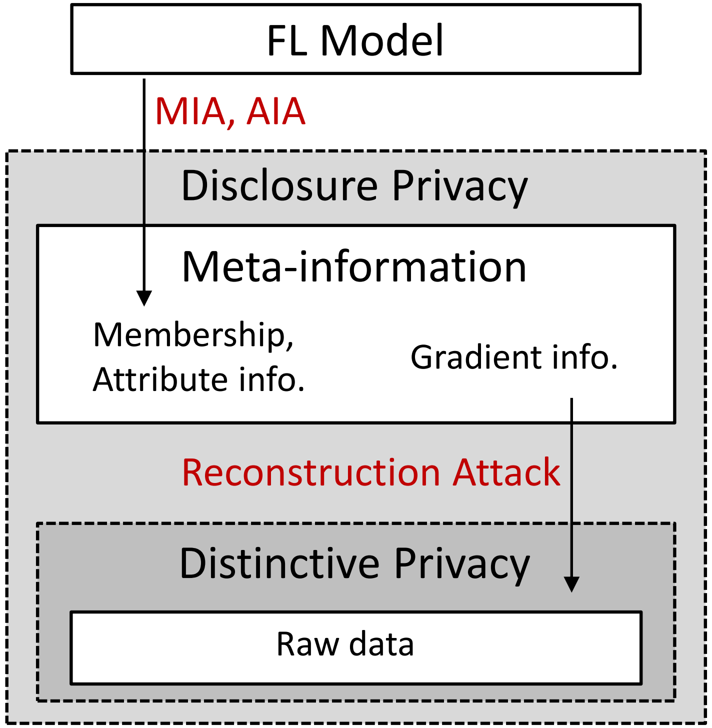
Breakdown of Privacy

Disclosure Privacy

“Privacy that ensures that any information cannot be inferred from the collaborative result”

Distinctive Privacy

“Privacy that ensures that the raw data will be secure and safe from exposure”



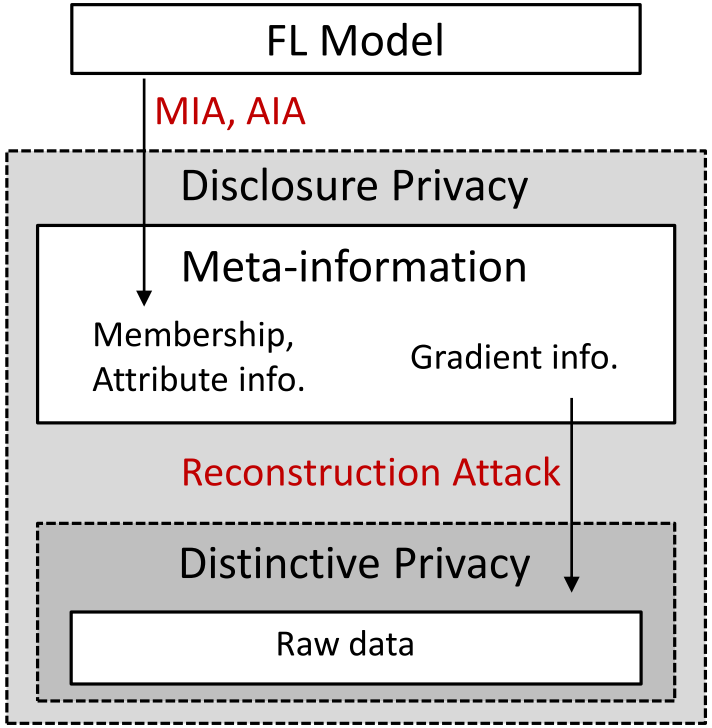
Breakdown of Privacy

Disclosure Privacy

In FL, the trained model should not leak any form of participant information (meta-info.).

Distinctive Privacy

In FL, the client data should be safe from reconstruction attempts.



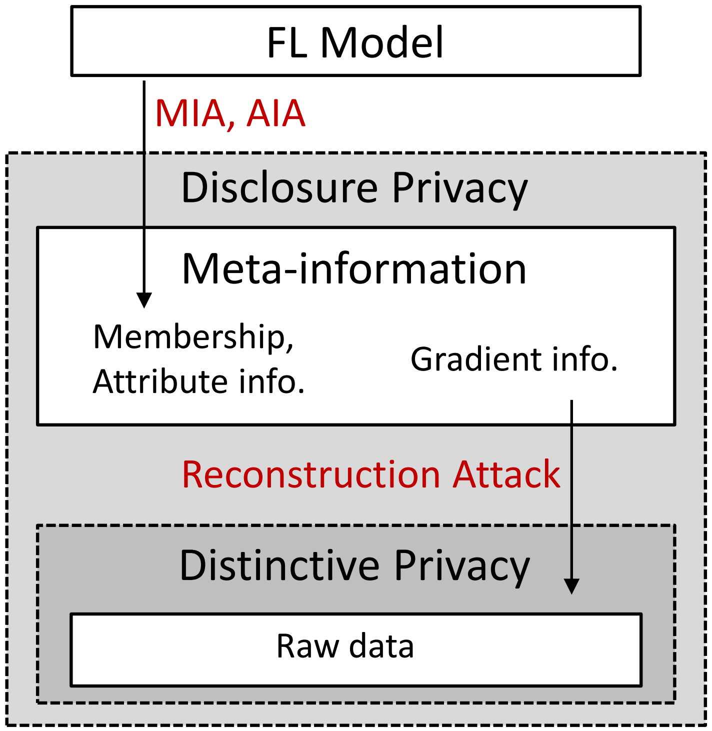
Breakdown of Privacy

Disclosure Privacy

By definition, differential privacy preserves disclosure privacy by training a safer model.

Distinctive Privacy

To conceal gradient information, encryption protocols accompanied by *computation and communication overhead* are used.



Breakdown of Privacy

Disclosure Privacy

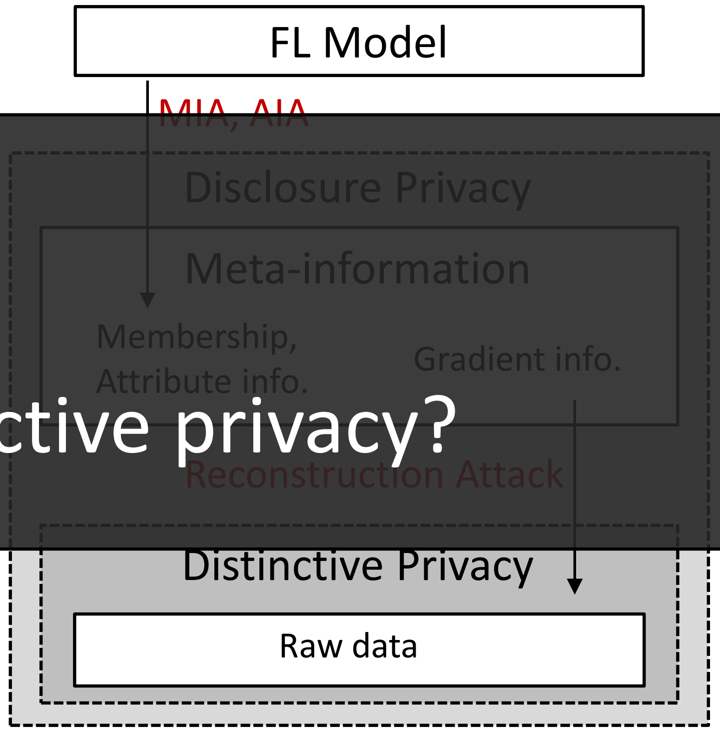
By definition, differential privacy preserves disclosure privacy by training a safer model.

<u>Research Question 2</u>

Is there a light method of ensuring distinctive privacy?

Distinctive Privacy

To conceal gradient information, encryption protocols accompanied by *computation and communication overhead* are used.



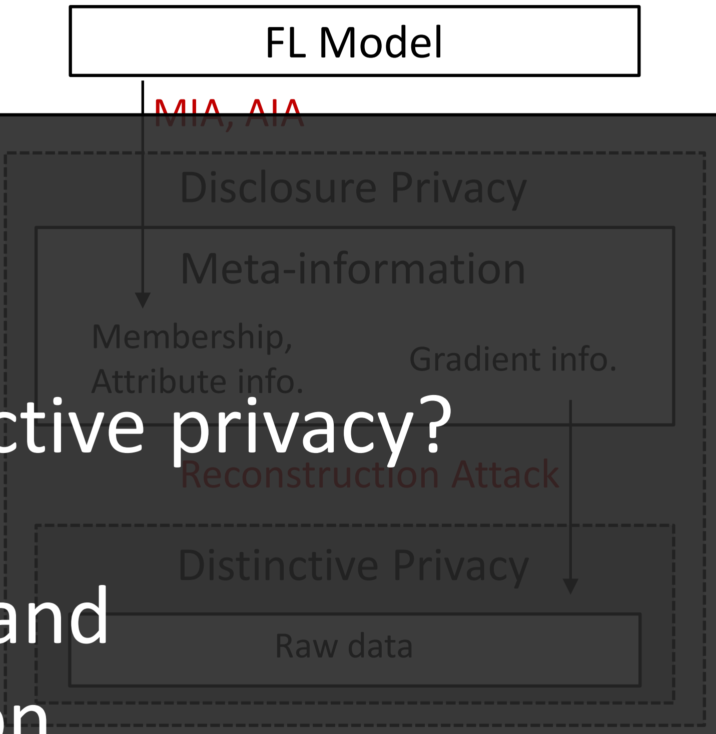
Breakdown of Privacy

Disclosure Privacy

By definition, differential privacy preserves disclosure privacy by training a safer model.

Distinctive Privacy

To conceal gradient information, encryption protocols accompanied by *computation and communication overhead* are used.



<u>Research Question 2</u>

Is there a light method of ensuring distinctive privacy?

➡ Select safe layers for exposure and mask the gradient information

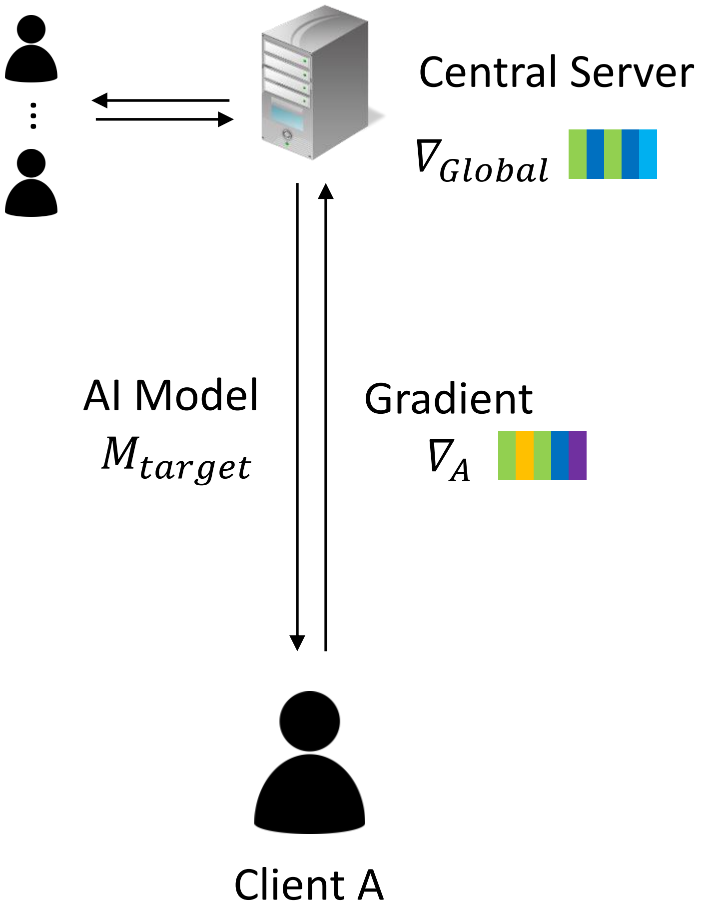
Obscuring Client Gradients Problem

To ensure distinctive privacy and prevent reconstruction, *mask the gradient information*

Problem: Finding obscuring function f that obscures the gradient ∇ such that:

Robustness	Light	Trade-off
$X\left(Recon(f(\nabla))\right) > X(f(\nabla))$ for defense capability X (e.g., MSE, PSNR)	$Cost(f(\nabla)) \leq Cost(\nabla)$ in terms of communication cost	Allows adjustment in trade-off of model performance and defense capability.

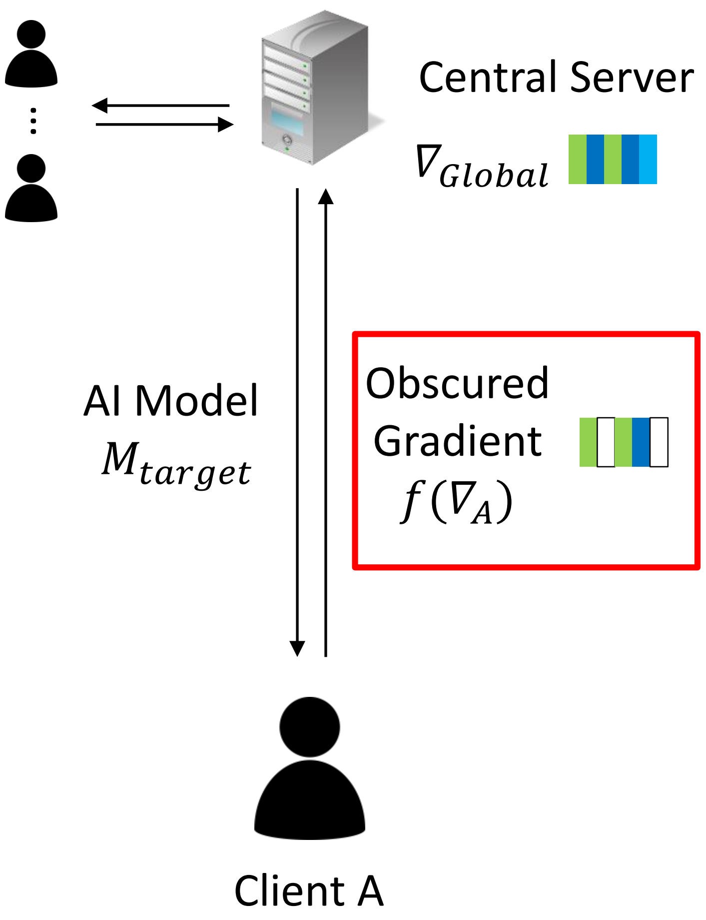
Intuition: Global Gradient



Aggregated gradient $\nabla_{global} = \frac{1}{N}\sum_{i=1}^{N}\nabla_i$

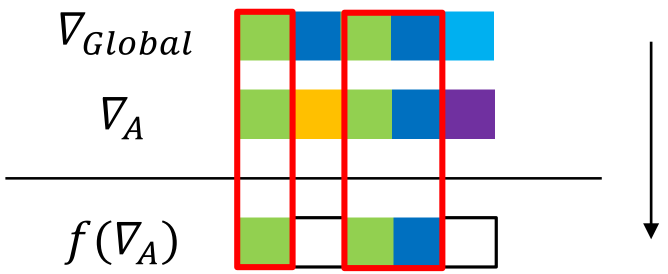
By being closer to ∇_{global}, the more generalized the gradient is.

Fragmented Federated Learning (FFL)



Aggregated gradient $\nabla_{global} = \sum_{i=0}^{N} \nabla_i$

By being closer to ∇_{global}, the more generalized the gradient is.



Selecting similar layers with selection ratio 0.6

Selecting the similar layers by cosine distance to the global gradient allows sending the layers of the private gradient that is most like the general distribution i.e. less private and more safe to send.

Fragmented Federated Learning (FFL)

Obscuring function f needs to be light in terms of 1. *communication* and 2. *computation* cost

Light Communication

Because the global gradient is used to update the model, estimate by

$$\nabla_{global} \approx M_{current} - M_{prev}$$

Light Computation

To decrease the computation in selecting the safe portion of a gradient, we use *layer-wise* selection

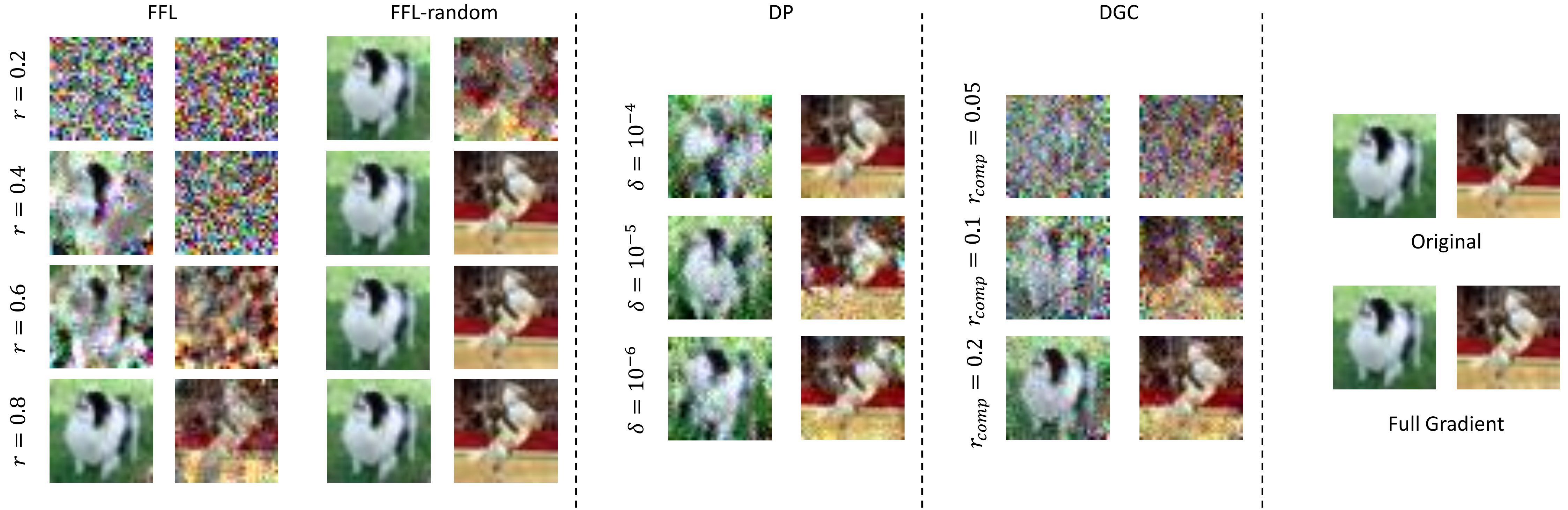
Experiment Setup

We evaluate FFL by attacking its gradients and attempting reconstruction by inverting gradients[1]

Comparison	Description	Variations
FFL-random	instead of selecting similar layers, random layer selection	Selection ratio of $r = 0.2,\ 0.4,\ 0.6,\ 0.8$
DP[2]	Differential privacy work applied to federated learning by Geyer et al.	privacy budget threshold of $\varepsilon = 8$ when $\delta = 10^{-4}, 10^{-5}, 10^{-6}$
DGC[3]	Gradient compression algorithm for efficient communication in federated learning	Compression ratios of $r_{comp} = 0.05, 0.1, 0.2$

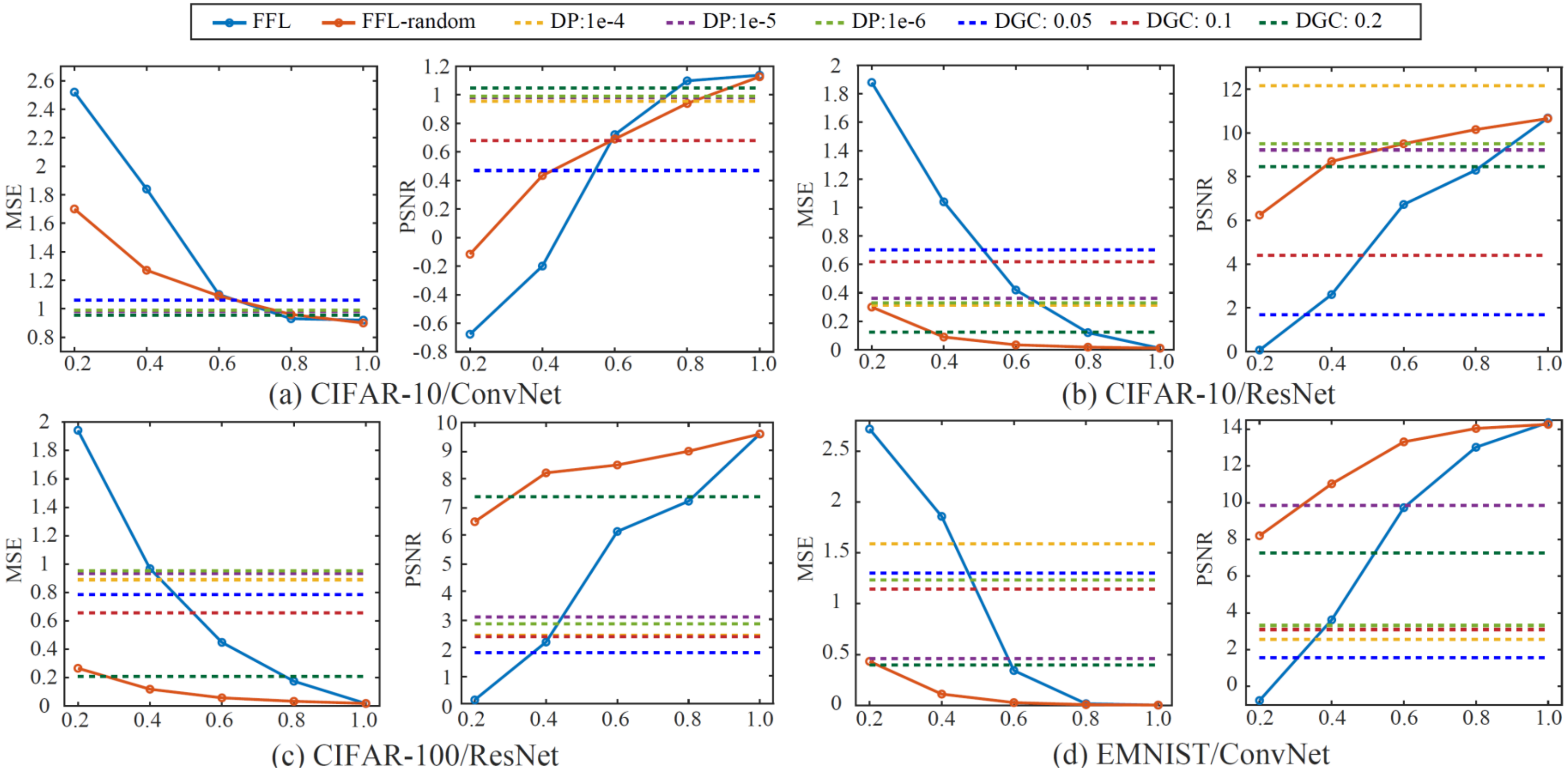
[1] Geiping, Jonas, et al. "Inverting gradients-how easy is it to break privacy in federated learning?." *Advances in Neural Information Processing Systems* 33 (2020): 16937-16947.
[2] Geyer, Robin C., Tassilo Klein, and Moin Nabi. "Differentially private federated learning: A client level perspective." *arXiv preprint arXiv:1712.07557* (2017).
[3] Lin, Yujun, et al. "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training." *International Conference on Learning Representations*. 2018.

Qualitative Evaluation



- As selection ratio r is decreased, there is a larger degree of failure.
- While the comparisons seem to reconstruct a noisy image, FFL reconstructions are patched, possibly due to the fact that full layers are dropped.

Quantitative Evaluation



- At lower ratios, FFL shows to be the most effective in preventing reconstruction and therefore ensuring distinctive privacy.
- Although different for each dataset/architecture pair, $r = 0.4$ shows to be the threshold for dominance in defense capability.

Communication Cost Evaluation

Arch.	r	Param. #	Size
ConvNet	0.2	134K	533KB
	0.4	563K	2.24MB
	0.6	2.44M	9.72MB
	0.8	3.48M	13.8MB
	1.0	3.49M	13.9MB
ResNet	0.2	1.67M	6.69MB
	0.4	3.08M	12.3MB
	0.6	15.1M	60.5MB
	0.8	25.9M	104MB
	1.0	44.7M	179MB

- Transmission bits in FFL
- As the layer ratio decreases, the number of parameters in bits decrease
- Because layers may contain different number of parameters, layer ratio r does not show a linear relationship with the number of parameters

Dataset Architecture	FFL	
	Train(sec)	Selection (sec)
CIFAR10/ConvNet	0.94	0.03 (3.19%)
CIFAR10/ResNet	1.27	0.09 (7.09%)
CIFAR100/ResNet	1.34	0.10 (7.46%)
EMNIST/ConvNet	1.79	0.03 (1.68%)

- Computation time consumed in one round of FFL
- For all dataset/architecture pairs, layer selection introduces a marginal computation overhead compared to training.
- ResNet has more layers than ConvNet, hence the increased time in selection

Accuracy Evaluation

Dataset/ Architecture	Methods	Accuracy (%)				
		Layer Selection Ratio r				
		0.2	0.4	0.6	0.8	1.0
CIFAR10/ConvNet	FFL	84.42	84.28	84.31	85.05	85.08
	FFL-random	83.35	84.02	84.02	84.67	
CIFAR10/ResNet	FFL	78.19	80.77	86.45	89.75	89.53
	FFL-random	83.18	86.31	87.78	88.21	
CIFAR100/ResNet	FFL	63.48	64.02	68.94	72.29	72.99
	FFL-random	67.3	69.77	71.00	71.89	
EMNIST/ConvNet	FFL	94.79	94.59	94.95	94.99	94.95
	FFL-random	94.73	94.91	94.94	95.00	

- FFL shows sharper decrease in accuracy than FFL-random, meaning that the 'safe' layers are beneficial in terms of model performance
- $r = 0.6$ seems to be the most appropriate with an average of (-1.96%) in terms of model performance degradation for FFL

Conclusion

- We conducted a holistic study of privacy attacks in FL and suggest two different forms of privacy breach: disclosure privacy and distinctive privacy
- We propose FFL as a framework that provides distinctive privacy while being light
- FFL is a practical solution in that it introduces near negligible overhead, and shows to be the most effective in terms of defense capability
- We hope that our decomposition of privacy in FL can be used to better understand and promote privacy-preserving methods

Thank you for listening!

Contact: harry.na@kaist.ac.kr