All Rivers Run to the Sea: Private Learning with Asymmetric Flows

Yue Niu[1], **Ramy E. Ali**[2], Saurav Prakash[3], Salman Avestimehr[1]

[1]University of Southern California (USC)

[2]Samsung (was with USC)

[3]University of Illinois Urbana-Champaign

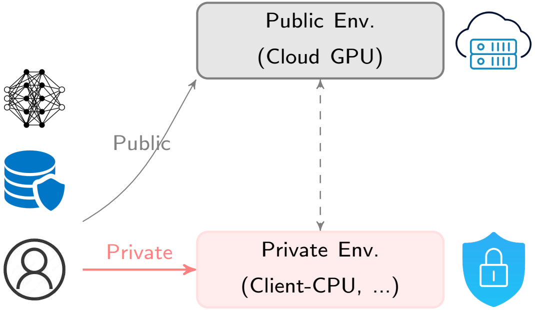
Feb. 22, 2024

Outline

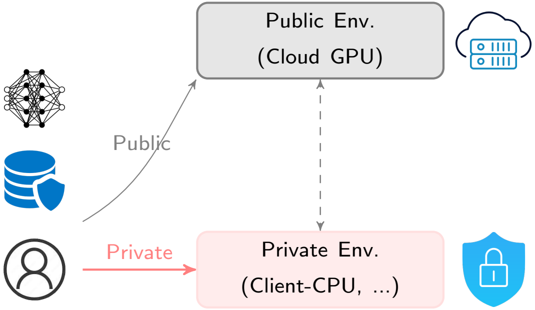
Outline

1. Background and Problem Setting
2. Delta: Private Learning with Asymmetric Flows
3. Empirical Evaluation: Utility, Privacy, Running Time
4. Discussion and Future Works

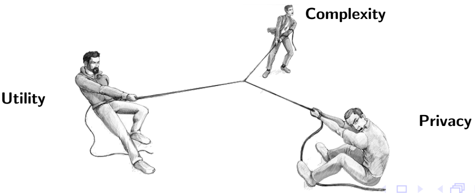
How to leverage cloud ML while ensuring privacy?



How to leverage cloud ML while ensuring privacy?

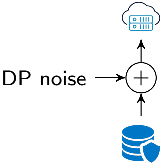


The Utility-Privacy-Complexity Trilemma



Privacy-Preserving ML Approaches

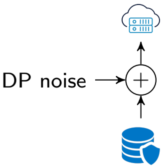
(Naive) DP-based ML



- Provable guarantee
- Severe accuracy drop

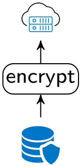
Privacy-Preserving ML Approaches

(Naive) DP-based ML



- Provable guarantee
- Severe accuracy drop

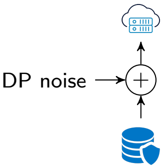
Crypto-based ML



- Strong protection
- High complexity

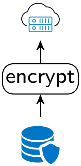
Privacy-Preserving ML Approaches

(Naive) DP-based ML



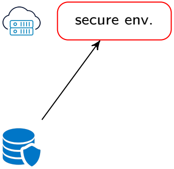
- Provable guarantee
- Severe accuracy drop

Crypto-based ML



- Strong protection
- High complexity

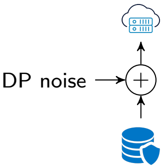
Secure Enclaves



- Hardware security
- Long running time

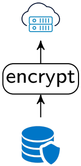
Privacy-Preserving ML Approaches

(Naive) DP-based ML



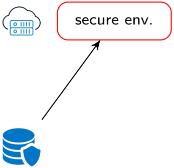
- Provable guarantee
- Severe accuracy drop

Crypto-based ML



- Strong protection
- High complexity

Secure Enclaves



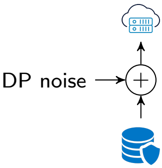
- Hardware security
- Long running time

Our work: Leverage both DP & Trusted hardware (local CPU, ...)

$\rightarrow$ Overcome accuracy drop of naive-DP & long running time of TEEs

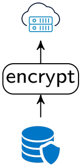
Privacy-Preserving ML Approaches

(Naive) DP-based ML



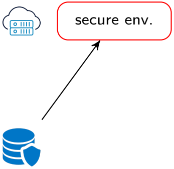
- Provable guarantee
- Severe accuracy drop

Crypto-based ML



- Strong protection
- High complexity

Secure Enclaves



- Hardware security
- Long running time

Related works leveraging TEEs

- Slalom'18: Inference only $\rightarrow$ This work: Inference and Training
- 3LegRace'21: Layerwise TEE-GPU communication $\rightarrow$ This work: No layer-wise communication

Outline

What does Delta do?

Decompose model & data into a low-dimensional part & a residual part

1. Lightweight model (client-side, TEEs, ...)
    - Fed with the low-dimensional information-sensitive part of the data
    - Confidential computing (no DP noise needed)

What does Delta do?

Decompose model & data into a low-dimensional part & a residual part

1. Lightweight model (client-side, TEEs, ...)
    - Fed with the low-dimensional information-sensitive part of the data
    - Confidential computing (no DP noise needed)
2. Large model (offloaded to cloud)
    - Fed with the quantized residual part of the data
    - The residual data is protected by a DP noise

What does Delta do?

Decompose model & data into a low-dimensional part & a residual part

1. Lightweight model (client-side, TEEs, ...)
   - Fed with the low-dimensional information-sensitive part of the data
   - Confidential computing (no DP noise needed)

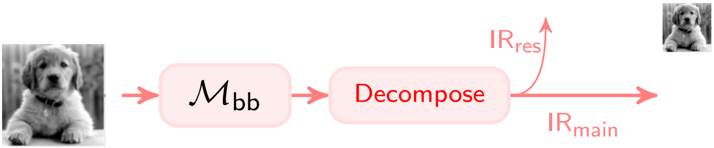
2. Large model (offloaded to cloud)
   - Fed with the quantized residual part of the data
   - The residual data is protected by a DP noise

$\Rightarrow$ Delta provides better utility-privacy trade-off than naive-DP methods

Delta Overview

Forward Propagation: Asymmetric Data Decomposition

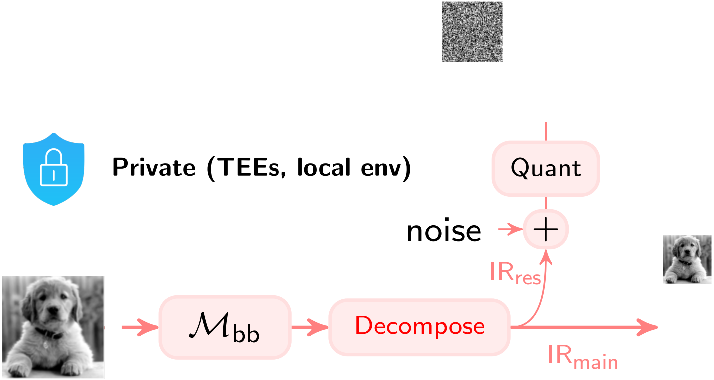
Private (TEEs, local env)



$\rightarrow$ To leverage the low-rank structure of the data

Delta Overview

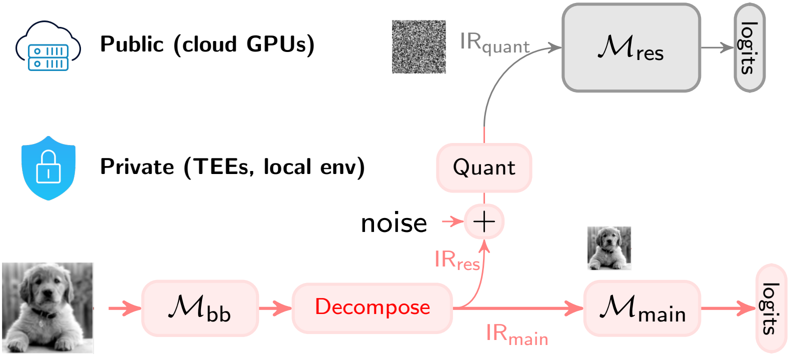
Forward Propagation: Perturbation & Binary Quantization



→ To ensure privacy and reduce communication cost

Delta Overview

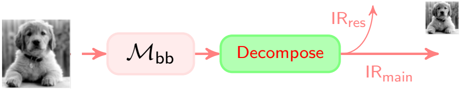
Forward Propagation: Model Decomposition



$\rightarrow$ To ensure low complexity in the private environment

Delta: Detailed Procedure

Asymmetric Data Decomposition



$\mathbf{IR} \xrightarrow{\text{SVD}} \begin{cases} \mathbf{IR}_{\text{SVD main}} \xrightarrow{\text{DCT}} \begin{cases} \mathbf{IR}_{\text{main}} \\ \mathbf{IR}_{\text{DCT res}} \end{cases} \\ \mathbf{IR}_{\text{SVD res}} \end{cases}$

$\mathbf{IR}_{\text{DCT res}} + \mathbf{IR}_{\text{SVD res}} \rightarrow \mathbf{IR}_{\text{res}}$

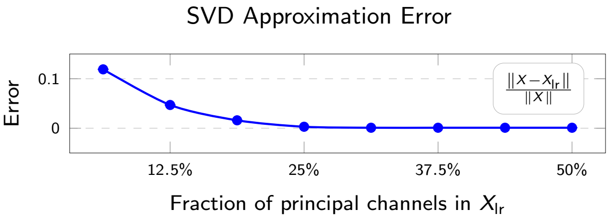
$\mathbf{IR}_{\text{main}}$: few channels & smaller feature size

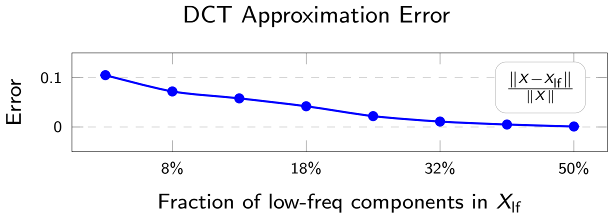
$\mathbf{IR}_{\text{res}}$: few info & large dimension

- SVD $\rightarrow$ asymmetric decomposition along channel dimension
- DCT $\rightarrow$ asymmetric decomposition along spatial dimension

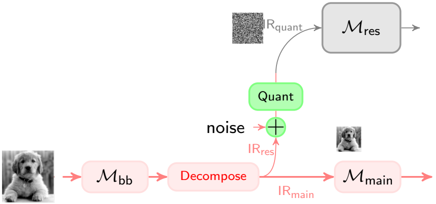
Delta: Detailed Procedure

Why asymmetric decomposition?





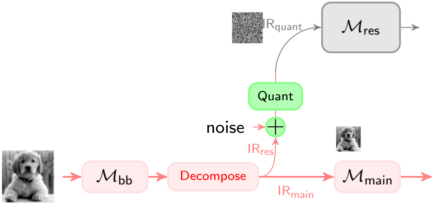
Random Binary Quantization



$$\text{IR}_{\text{quant}}(\cdot) = \mathrm{BinQuant}(\text{IR}_{\text{noisy}}(\cdot)) = \begin{cases} 0 & \text{IR}_{\text{noisy}}(\cdot) < 0 \\ 1 & \text{IR}_{\text{noisy}}(\cdot) \geq 0 \end{cases}$$

Delta: Detailed Procedure

Random Binary Quantization

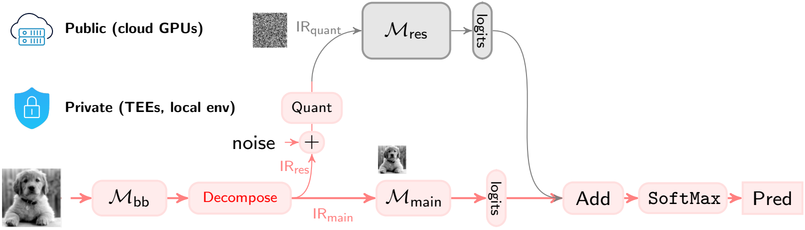


$$\mathsf{IR}_{\mathsf{quant}}(\cdot) = \mathrm{BinQuant}(\mathsf{IR}_{\mathsf{noisy}}(\cdot)) = \begin{cases} 0 & \mathsf{IR}_{\mathsf{noisy}}(\cdot) < 0 \\ 1 & \mathsf{IR}_{\mathsf{noisy}}(\cdot) \geq 0 \end{cases}$$

Theorem: `Delta` ensures that any operation in the public environment satisfy (ϵ, δ)-DP given noise $\mathcal{N}(0, p\Delta/\epsilon \cdot \sqrt{2\log(1.25/\delta)})$ and mini-batch size b, where $p = b/N$ is the sampling probability.

Delta: Detailed Procedure

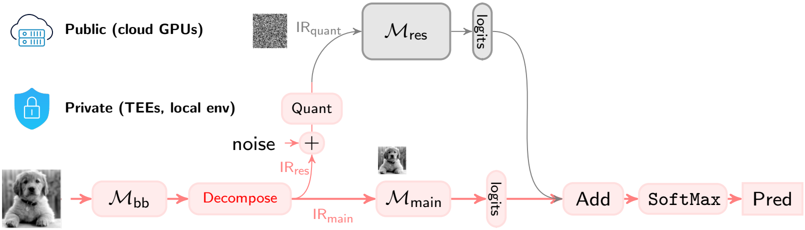
Private Backpropagation



$$\mathcal{M}_{\text{main}} : \boldsymbol{o}_{\text{tot}}(i) = \frac{e^{\boldsymbol{z}_{\text{main}}(i)+\boldsymbol{z}_{\text{res}}(i)}}{\sum_{j=1} e^{\boldsymbol{z}_{\text{main}}(j)+\boldsymbol{z}_{\text{res}}(j)}} \quad \text{for} \quad i = 1, \cdots, L$$

$$\mathcal{M}_{\text{res}} : \boldsymbol{o}_{\text{res}}(i) = \frac{e^{\boldsymbol{z}_{\text{res}}(i)}}{\sum_{j=1} e^{\boldsymbol{z}_{\text{res}}(j)}} \quad \text{for} \quad i = 1, \cdots, L,$$

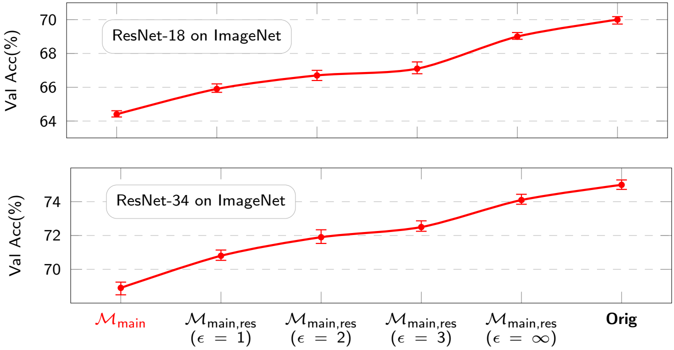
Delta: Full Picture



- Asymmetric data decomposition
- Efficient model design
- Random binary quantization
- Private backpropagation

Outline

Experiment Highlights: Model Utility



$\rightarrow$ Lightweight model achieves good accuracy, but still residuals are useful

Experiment Highlights: Model Utility

Setting: ResNet-18 with $\epsilon = 1$

	`Delta`: perturb IR_{res}	naive-DP: perturb IR
CIFAR-10	92.4%	69.6% ($\downarrow -22.8$)
CIFAR-100	71.4%	48.3% ($\downarrow -23.1$)
ImageNet	65.9%	34.4% ($\downarrow -31.5$)

$\rightarrow$ Delta improves accuracy by up to 31.5%

Experiment Highlights: Model Complexity

MACs of the modules in `Delta`

	$\mathcal{M}_{\text{bb}}+\mathcal{M}_{\text{main}}$	SVD	DCT	$\mathcal{M}_{\text{res}}$
ResNet-18	48.3 M	0.52 M	0.26 M	547M
ResNet-34	437 M	1.6 M	0.7 M	3.5G

- Small model $\mathcal{M}_{\text{main}}$ only costs 10% complexity of $\mathcal{M}_{\text{res}}$
- Costs of SVD and DCT are marginal

Experiment Highlights: Speedup

Running time with one single input

	Priv-only	3LegRace	Slalom	Delta
Train (ms/speedup)	1372	237 (6×)	-	62 (22×)
Inference (ms/speedup)	510	95 (5×)	84 (6×)	20 (25×)

3LegRace [Niu, et al, PETs 2022]: layer-wise feature decomposition on linear layers
Slalom [Tramer, et al, ICLR 2019]: layer-wise computation distribution on linear layers

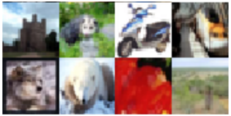
- Significant speedup compared to solely using private envs
- Faster compared to baselines due to reduced communication

Experiment Highlights: Protection Against Attacks

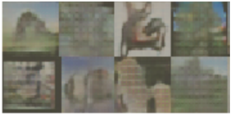
Procedure: Train a GAN with the quantized residuals

Setting: ResNet-18, CIFAR-100

Against model inversion attack [`SecretRevealer`, CVPR'20]



Original samples



Reconstruction (no noise)

Reconstruction ($\epsilon = 1$)

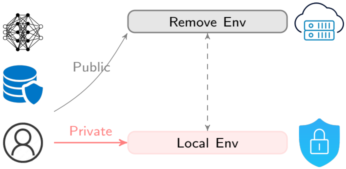
- Attack can succeed on certain samples (e.g., row 1, col 3)
- Random quantization provide further protection

Outline

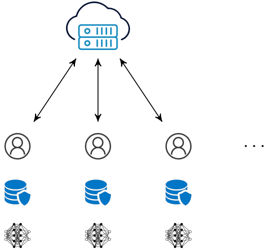
Future Works

Extend to More General Settings

User-Server Setting



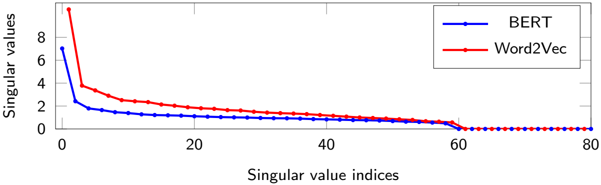
Federated Setting



Future Works

Extend to LMs

LMs' embedding also exhibits a low-ranks structure



Original text (top) and approximated (bottom) text with 1/5 principal vectors.

Large Language Models are foundational machine learning models that use deep learning algorithms to process and understand natural language. These models are trained on massive amounts of text data to learn patterns and entity relationships in the language.

Large Language Models can perform many types of language tasks, such as translating languages, analyzing sentiments, chatbot conversations, and more. They can understand complex textual data, identify entities and relationships between them, and generate new text that **is** coherent and grammatically accurate.

Large Language Models are foundational machine learning models that use deep learning algorithms to process and understand natural language. These models are trained on massive amounts of text data to learn patterns and entity relationships in the language.

Large Language Models can perform many types of language tasks, such and translating languages, analyzing sentiments, chatbot conversations, and more. They can understand complex textual data, identify entities and relationships between them, and generate new text that **are** coherent and grammatically accurate.

Questions?