PhD Defense

Striking the Balance:
Optimizing Privacy, Utility, and Complexity in Private Machine Learning

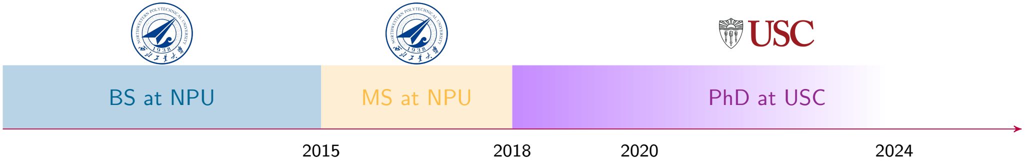
Yue Niu

Dept. of Electrical and Computer Engineering
University of Southern California

Committee: Salman Avestimehr (Chair), Murali Annavaram, Meisam Razaviyayn

USC Viterbi
School of Engineering
Ming Hsieh Department of Electrical and Computer Engineering

About Me



About Me

Research Areas:

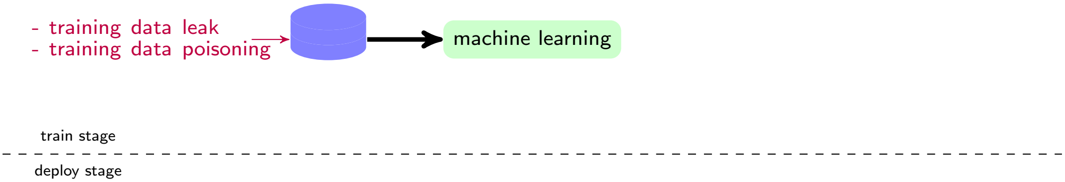
- ML/LLM compression and acceleration [CVPR'24, FPGA'20, HiPC'19, $\cdots$]
  - model pruning
  - low-rank compression
  - hardware architecture design
- **Efficient private ML** [CVPR'24, PETS'24, TMC'24, TMLR'23, NeurIPS-FL'23, PETS'22, $\cdots$]
  - differential privacy
  - federated learning
  - trusted execution environments
- LLM privacy, fairness and bias [NAACL'24, AAAI-ReLM'24]
- Stochastic optimization [TMLR'23, ICML'21 Workshop on Optimization]

About Me

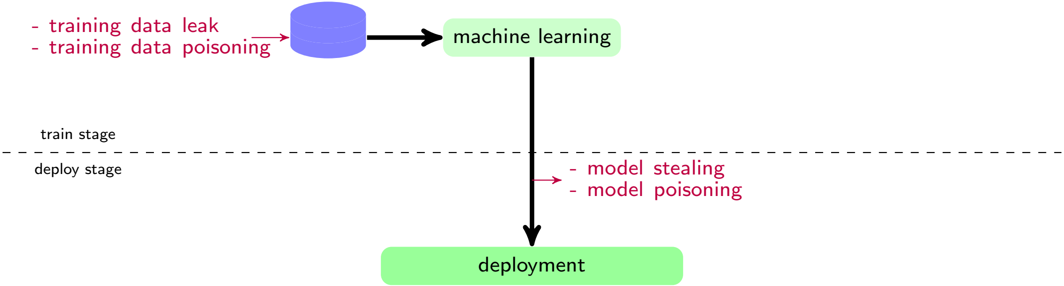
Research Areas:

- ML/LLM compression and acceleration [CVPR'24, FPGA'20, HiPC'19, · · ·]
  - model pruning
  - low-rank compression
  - hardware architecture design
- **Efficient private ML** [CVPR'24, PETS'24, TMC'24, TMLR'23, NeurIPS-FL'23, PETS'22, · · ·]
  - differential privacy
  - federated learning
  - trusted execution environments
- LLM privacy, fairness and bias [NAACL'24, AAAI-ReLM'24]
- Stochastic optimization [TMLR'23, ICML'21 Workshop on Optimization]

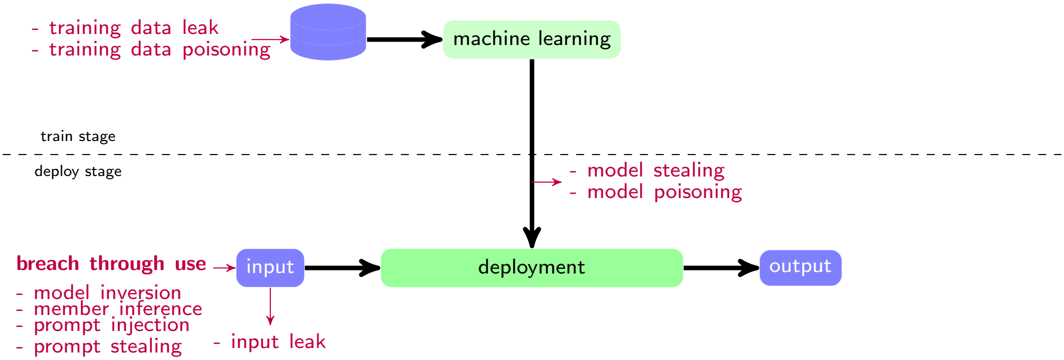
Privacy Breach in Machine Learning Pipeline



Privacy Breach in Machine Learning Pipeline



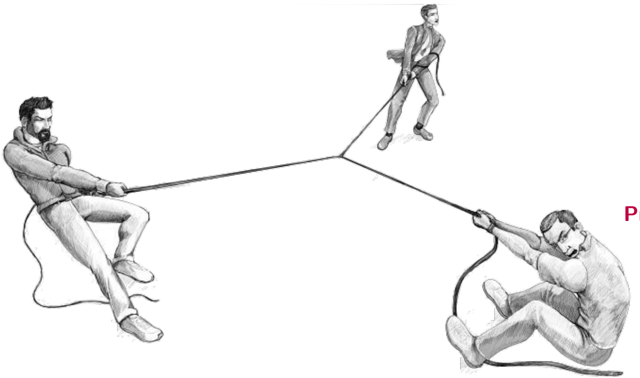
Privacy Breach in Machine Learning Pipeline



The Privacy-Utility-Complexity Trilemma

Complexity (MACs, time, ...)

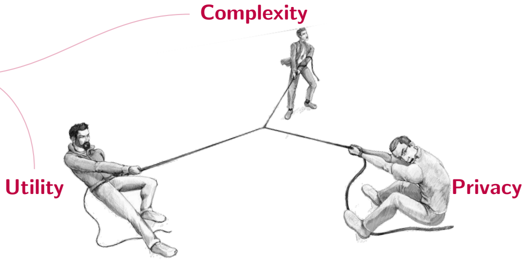
Utility (acc, ...)



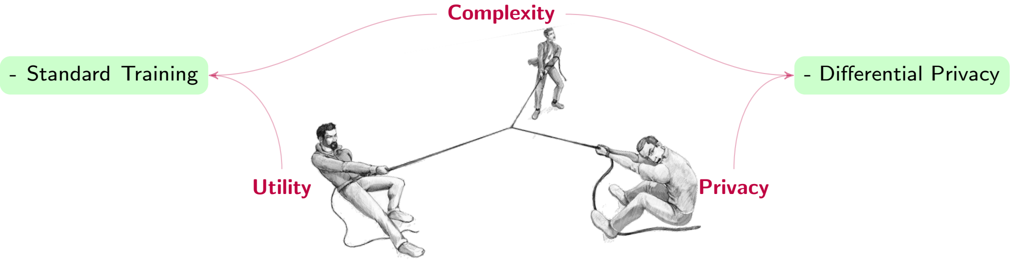
Privacy (attack, information, ...)

The Privacy-Utility-Complexity Trilemma

- Standard Training



The Privacy-Utility-Complexity Trilemma

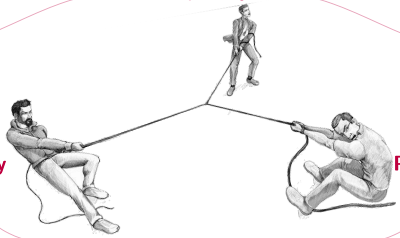


The Privacy-Utility-Complexity Trilemma

Complexity

- Standard Training

- Differential Privacy

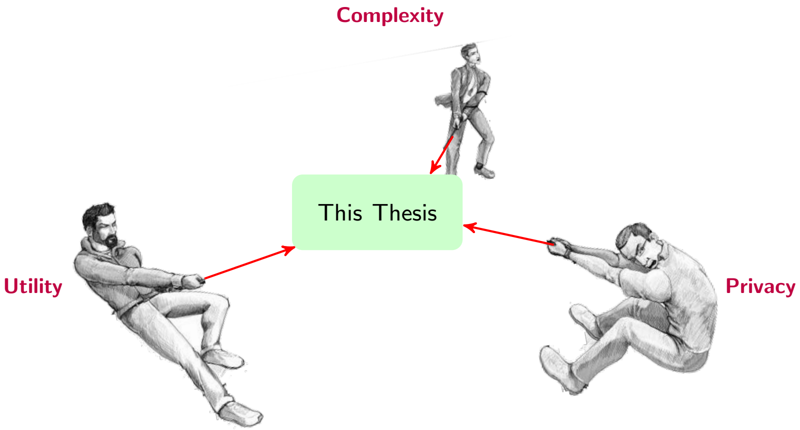


Utility

Privacy

- Homomorphic Encryption
- Trusted Execution

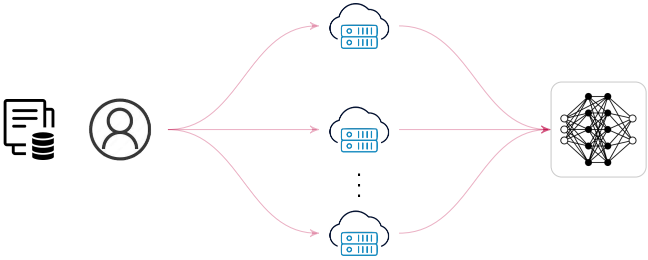
The Privacy-Utility-Complexity Trilemma



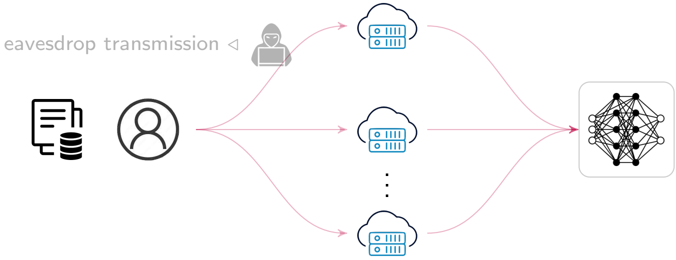
This Thesis

- Background: Data Privacy Breach in Machine Learning

Data Privacy Breach Overview



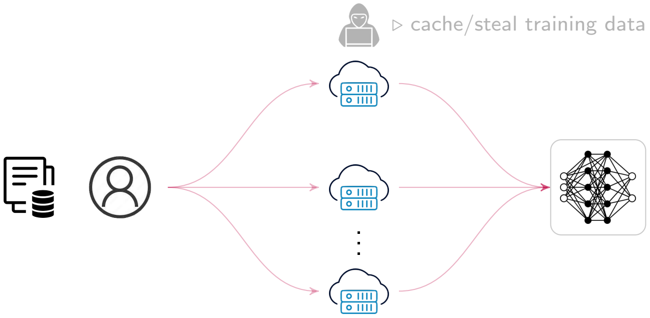
Data Privacy Breach Overview



▷ Case 1: Attackers obtain private data via unsafe transmission in

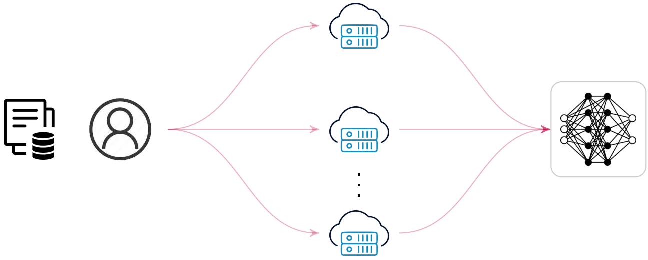
- user-cloud systems
- distributed systems

Data Privacy Breach Overview



▷ Case 2: Public cloud servers may cache or steal private data

Data Privacy Breach Overview



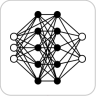
▷ attack through models

▷ Case 3: Private data can be leaked via models:

- model inversion
- membership inference
- . . .

Attack through Models: Model Inversion

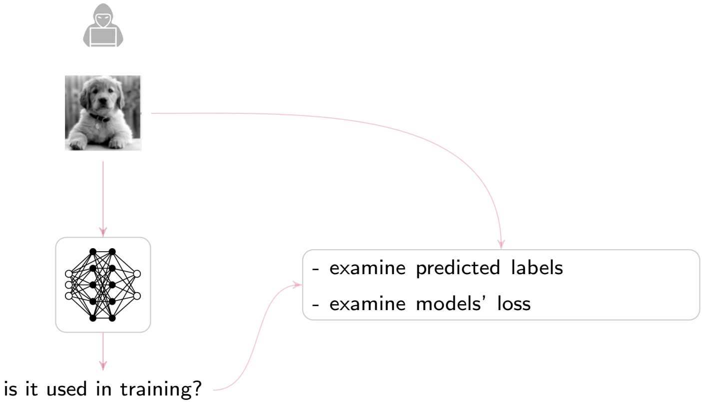
generative model



Optimize a generative model to

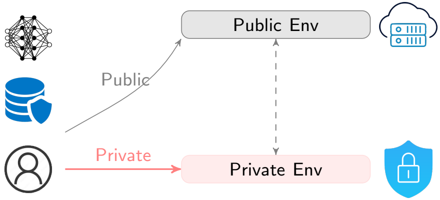
- mimic models' output loss
- mimic model's gradients
- mimic model's predicted labels

Attack through Models: Membership Inference



- Target Setup: Learning with Private and Public Environments

Target Setup

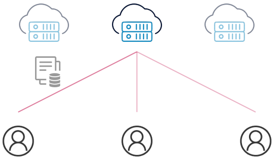


- Private Env: strong privacy guarantee; increasing complexity, less computation efficient
  - local clients
  - trusted execution
  - ...
- Public Env: no privacy guarantee; high computing performance
  - Cloud GPUs
  - ...

Target Setup

A Generic Setup Seen in Many Scenarios

Distributed ML



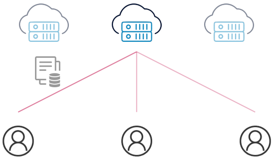
Distribute model and data in distributed systems
- distributed training
- federated learning
- data parallelism

- ...

Target Setup

A Generic Setup Seen in Many Scenarios

Distributed ML



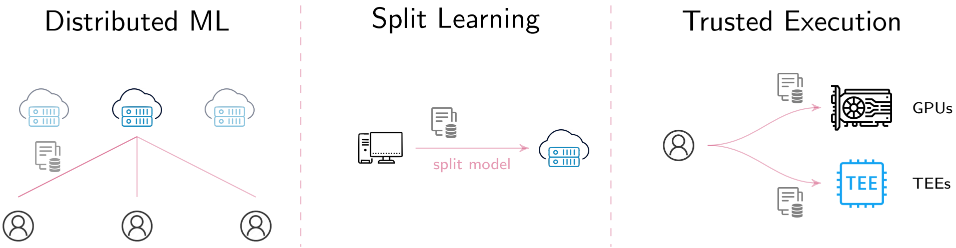
Split Learning

Split model and data onto multiple platforms

- model splitting
- model parallelism
- ...

Target Setup

A Generic Setup Seen in Many Scenarios



Target Setup

The Central Problem To Be Solved

How to leverage both private and public environments to achieve:

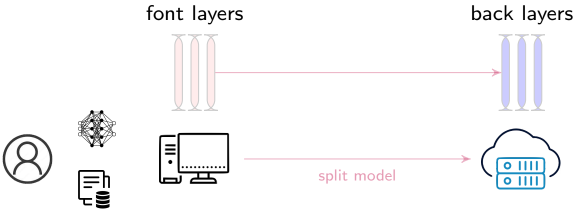
- private training & inference
- high model utility
- fast execution

This Thesis

- Review: Relevant Works

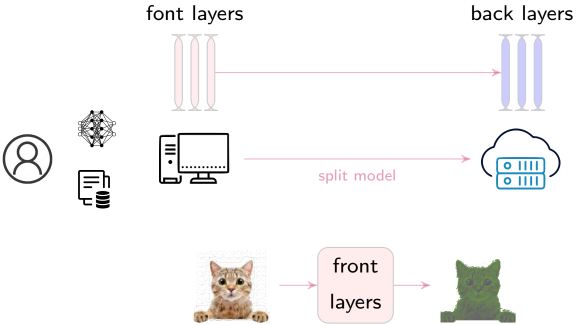
Relevant Works

▷ Split Learning



- protect raw data in local
- reduce computation from local

▷ Split Learning

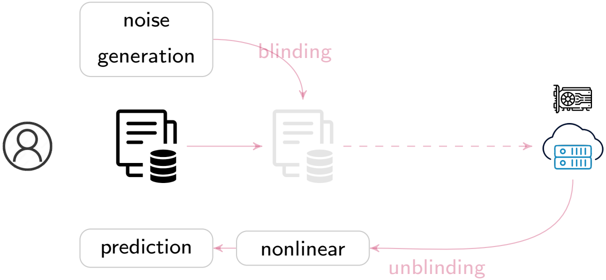


- protect raw data in local
- reduce computation from local

- not fully private
- not communication efficient
- fail against reconstruction attacks

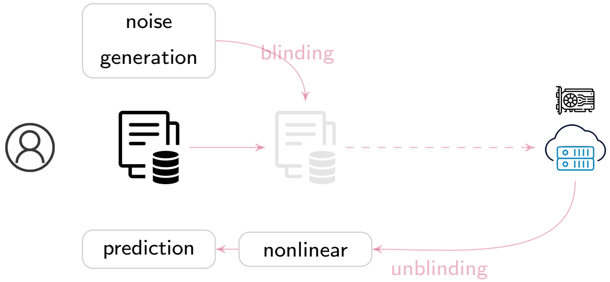
Relevant Works

▷ Data Blinding



- offload complex ops
- fully private in cloud

▷ Data Blinding

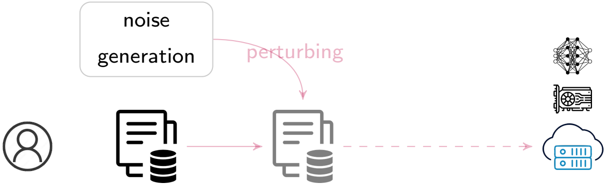


- offload complex ops
- fully private in cloud

- only for model inference
- heavy layerwise communication

Relevant Works

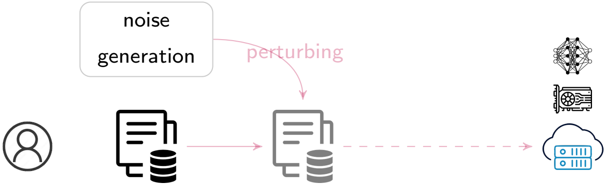
▷ Data Obfuscation



- completely offload computation
- no need for local computation

Relevant Works

▷ Data Obfuscation



- completely offload computation
- no need for local computation

- degraded model utility
- not fully private

The Key Argument in This Thesis:

Protecting data in private ML must be based on data,

and content-aware.

- This Thesis
  - Asymmetric Structure in Data (PETS'22)
  - 3LegRace: Layer-Wise Asymmetric Data Decomposition (PETS'22)
  - Theoretical Foundations (PETS'22)
  - Delta: ML with Fully Asymmetric Data Flow(CVPR'24)

Asymmetric Structure in Data

▷ Data in ML

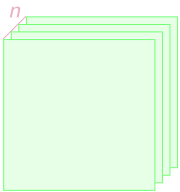
▷ Data in ML

Data Representations Are Redundant!!!

Asymmetric Structure in Data

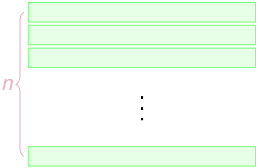
▷ Data Representation

image



Tensor: $n \times h \times w$

text



Tensor: $n \times l$

▷ Redundancy Analysis

For data $X \in \mathcal{R}^{n \times k}$, obtain singular values as

$$X \xrightarrow{SVD} U \cdot \mathtt{diag}(s) \cdot V^*$$

SVD-Entropy (PETS'22)

$$\mu_X = -\log\left(\sum_{j=1}^{n} \bar{s}_j^2\right)$$

$$\bar{s}_j = \frac{s_j}{\sum_{i=1}^{n} s_i}$$

Asymmetric Structure in Data

▷ Redundancy Analysis

Sufficiency (PETS'22)

$r = \lceil 2^{\mu_X} \rceil$ denote the number of components that *sufficiently* approximate X:

$$\frac{\sum_{j=1}^{r} s_j^2}{\sum_{j=1}^{n} s_j^2} \geq .97.$$

▷ Redundancy Analysis

$\xrightarrow{\text{SVD}}$ s: [0.94, 0.05, 0.007] $\rightarrow \mu = 0.17 \xrightarrow{r=2}$

3LegRace: Layer-Wise Asymmetric Data Decomposition

$$X \longrightarrow \boxed{\text{linear op}} \longrightarrow Y$$

3LegRace: Layer-Wise Asymmetric Data Decomposition

$X \longrightarrow \boxed{\text{linear op}} \longrightarrow Y$

$X \xrightarrow{\text{private}} \boxed{X_{\text{main}} = \sum_{j=1}^{r} U_j \cdot X'_j} \longrightarrow \boxed{\text{linear op}}$

$$
\begin{aligned}
Y_{\text{main}} &= \sum_{i=1}^{N} X_{\text{main,i}} \circledast W_i \\
&= \sum_{i=1}^{N} \sum_{j=1}^{r} U_{i,j} X'_j \circledast W_i \\
&= \sum_{j=1}^{r} X'_j \circledast W'_j
\end{aligned}
$$

3LegRace: Layer-Wise Asymmetric Data Decomposition

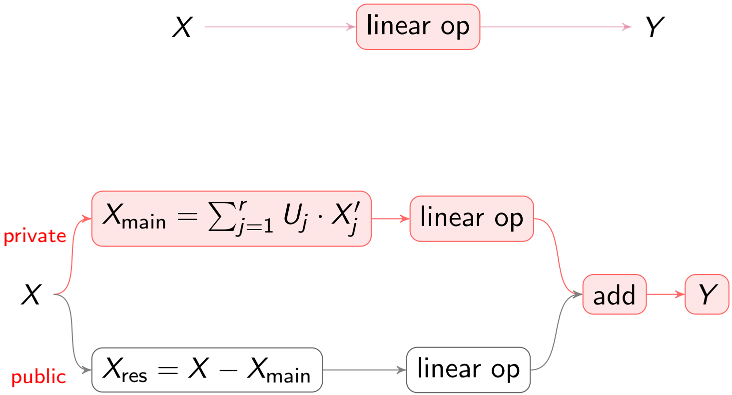
$$X \longrightarrow \boxed{\text{linear op}} \longrightarrow Y$$

$$X \xrightarrow{\text{private}} \boxed{X_{\text{main}} = \sum_{j=1}^{r} U_j \cdot X'_j} \longrightarrow \boxed{\text{linear op}}$$

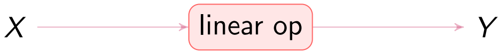
$$Y_{\text{main}} = \sum_{i=1}^{N} X_{\text{main,i}} \circledast W_i$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{r} U_{i,j} X'_j \circledast W_i$$
$$= \sum_{j=1}^{r} X'_j \circledast W'_j$$

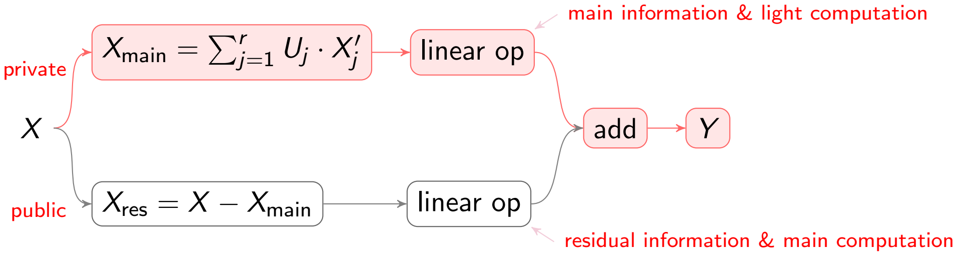
$$\nabla_{W_i} = X_i \circledast \nabla_Y \mathcal{L}$$
$$= \sum_{j=1}^{r} U_{i,j} \cdot X'_j \circledast \nabla_Y \mathcal{L}$$

3LegRace: Layer-Wise Asymmetric Data Decomposition



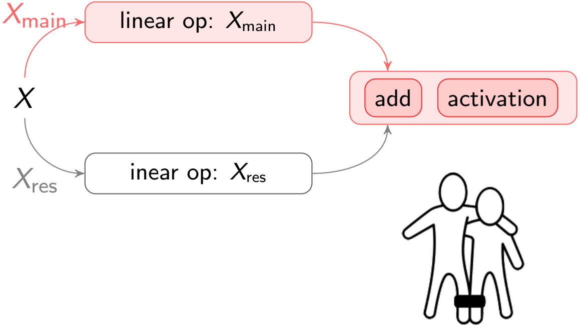
3LegRace: Layer-Wise Asymmetric Data Decomposition





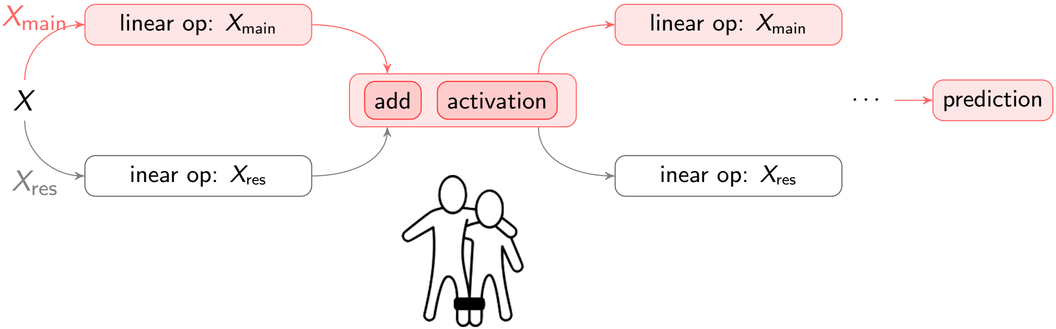
3LegRace: Layer-Wise Asymmetric Data Decomposition

▷ Complete Flow

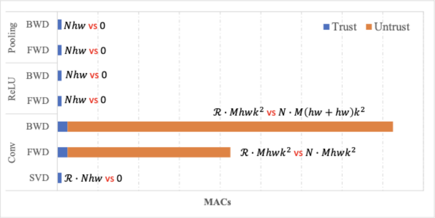


3LegRace: Layer-Wise Asymmetric Data Decomposition

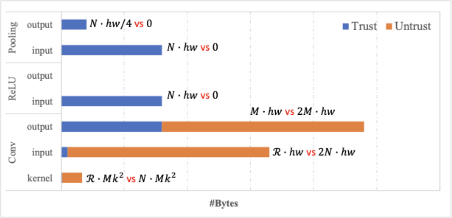
▷ Complete Flow



3LegRace: Layer-Wise Asymmetric Data Decomposition



Computation Complexity



Memory Complexity

Low-Rank Structure Is Preserved in Models

Low-Rank Structure in a 1×1 Conv Layer (PETS'22)

Given input $X \in \mathcal{R}^{n \times h \times w}$ with SVD-entropy μ_X, and kernel $W \in \mathcal{R}^{m \times n \times 1 \times 1}$, the SVD-entropy of the output is upper-bounded by:

$$\mu_Y \leq \log\left(\lceil 2^{\mu_X} \rceil\right).$$

Low-Rank Structure Is Preserved in Models

Low-Rank Structure in a $k \times k$ Conv Layer (PETS'22)

Given input $X \in \mathcal{R}^{n \times h \times w}$ with SVD-entropy μ_X, and kernel $W \in \mathcal{R}^{m \times n \times k \times k}$, the SVD-entropy of the output is upper-bounded by:

$$\mu_Y \leq \log \left(\sum_{j=1}^{r} \lceil 2^{\mu_j} \rceil \right) \cong \mu_X + c(k).$$

Low-Rank Structure Is Preserved in Models

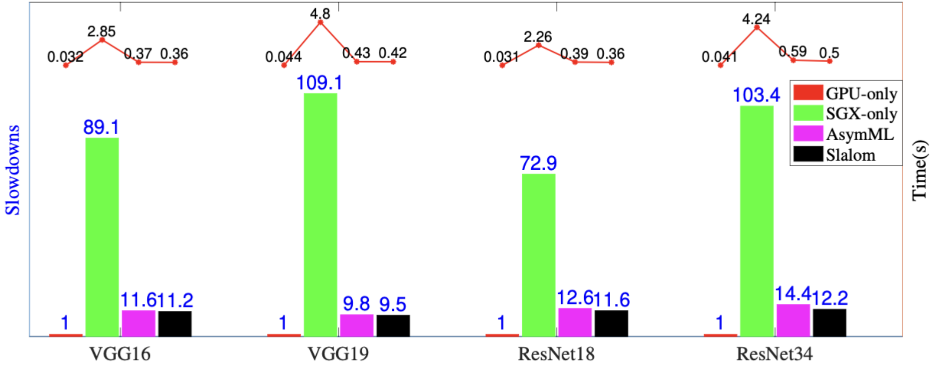
Low-Rank Structure in a Batch Norm Layer (PETS'22)

Given input $X \in \mathcal{R}^{n \times h \times w}$ with SVD-entropy μ_X, the SVD-entropy of the output is upper-bounded by:

$$\mu_Y \leq \log\left(\lceil 2^{\mu_X} \rceil + 1\right).$$

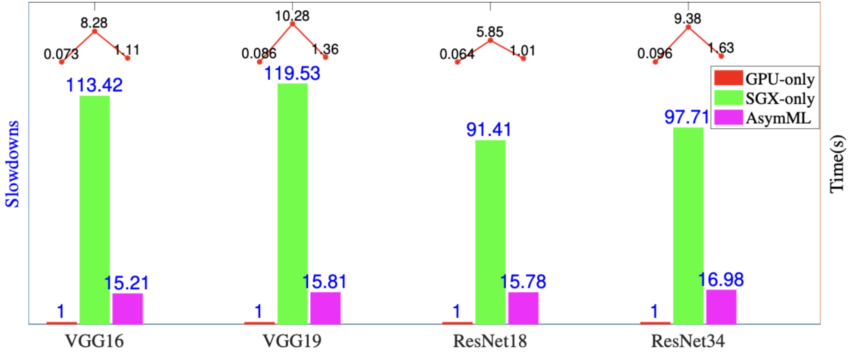
Performance

▷ Performance



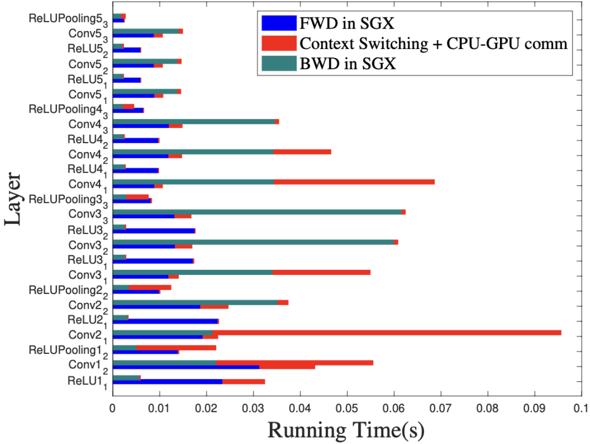
Inference Time (on ImageNet)

▷ Performance



Train Time (on ImageNet)

▷ Performance

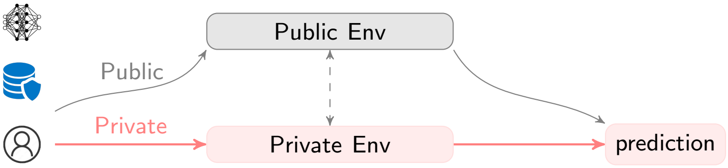


Time Breakdowns
(on VGG16)

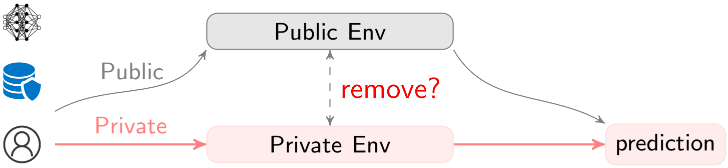
Still Not Good Enough:

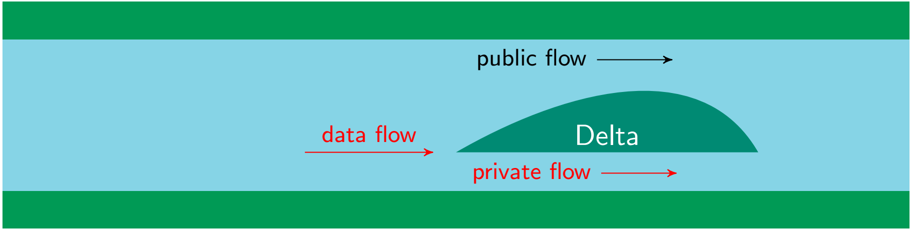
- Heavy layer-wise communication
- Formal privacy guarantee in public environments

Delta: ML with Fully Asymmetric Data Flow

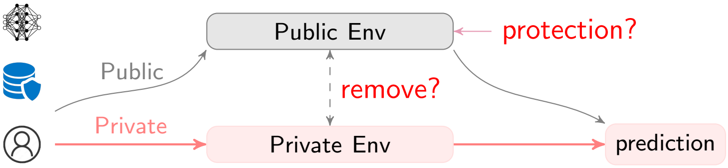


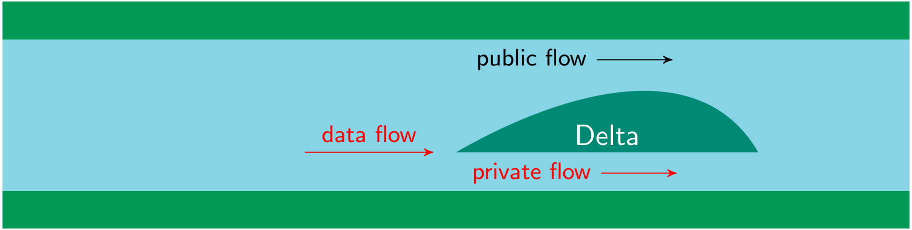
Delta: ML with Fully Asymmetric Data Flow





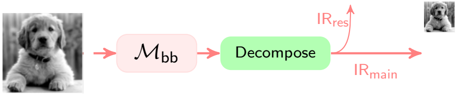
Delta: ML with Fully Asymmetric Data Flow





Delta: ML with Fully Asymmetric Data Flow

Private (TEEs, local env)

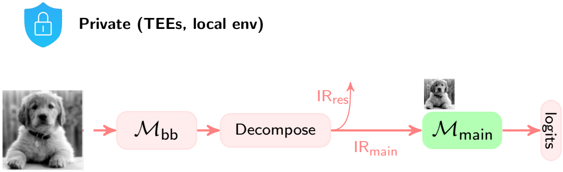


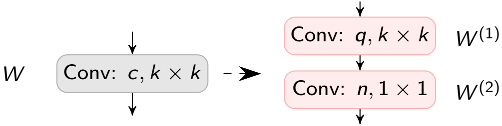
$$\textbf{IR} \xrightarrow{\text{SVD}} \begin{cases} \textbf{IR}_{\text{SVD main}} \xrightarrow{\text{DCT}} \begin{cases} \textbf{IR}_{\text{main}} \\ \textbf{IR}_{\text{DCT res}} \end{cases} \\ \textbf{IR}_{\text{SVD res}} \end{cases}$$

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

$\textbf{IR}_{\text{DCT res}} + \textbf{IR}_{\text{SVD res}} \rightarrow \textbf{IR}_{\text{res}}$ few info & large dimension

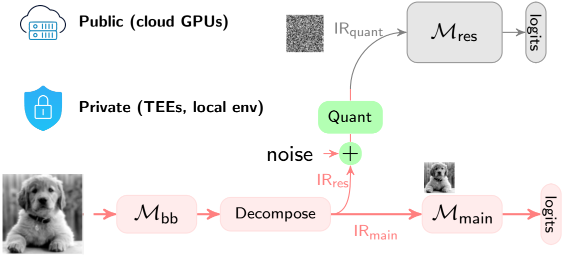
Delta: ML with Fully Asymmetric Data Flow





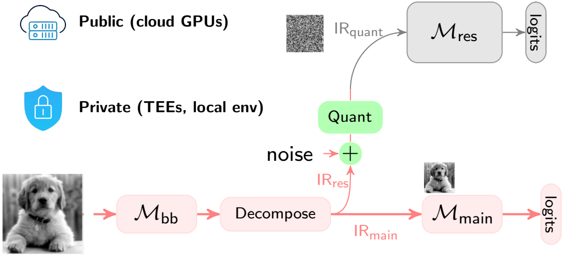
Theorem: By optimizing W^1, W^2, then $\min_{W^1, W^2} \|\mathrm{Op}(W, X) - \mathrm{Op}(W^1, W^2, X)\| = 0$

Delta: ML with Fully Asymmetric Data Flow



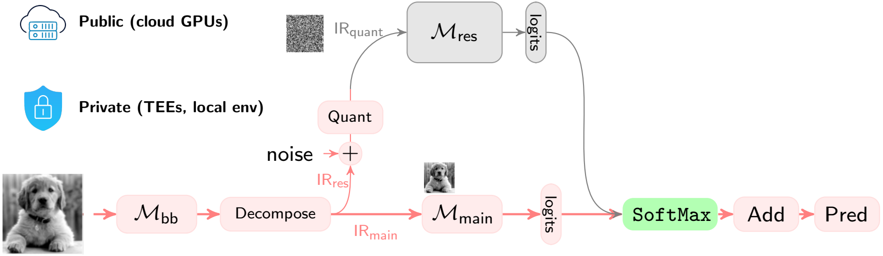
$$\text{IR}_{\text{quant}}(\cdot) = \mathrm{BinQuant}(\text{IR}_{\text{noisy}}(\cdot)) = \begin{cases} 0 & \text{IR}_{\text{noisy}}(\cdot) < 0 \\ 1 & \text{IR}_{\text{noisy}}(\cdot) \geq 0 \end{cases}$$

Delta: ML with Fully Asymmetric Data Flow



Theorem: Delta ensures that the perturbed residuals and operations in the public environment satisfy (ϵ, δ)-DP given noise $\mathcal{N}(0, 2C^2 \cdot \log(2/\delta')/\epsilon')$ given sampling probability p, and $\epsilon = \log(1 + p(e^{\epsilon'} - 1)), \delta = p\delta'$.

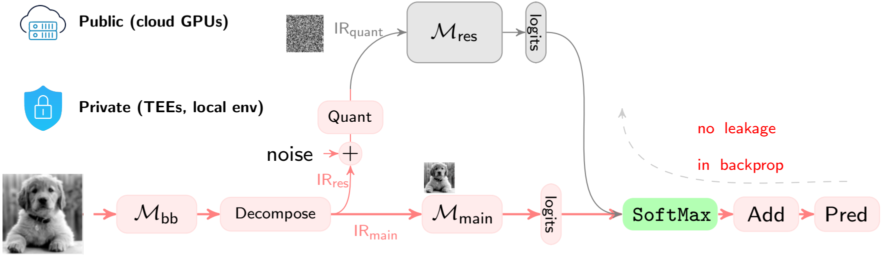
Delta: ML with Fully Asymmetric Data Flow



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

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

Delta: ML with Fully Asymmetric Data Flow

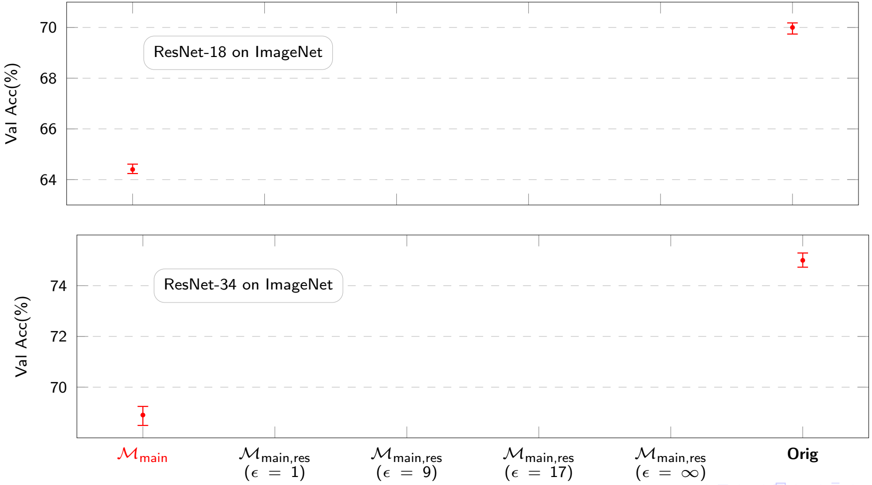


$$\mathcal{M}_{\text{main}} : \boldsymbol{o}_{\text{tot}}(i) = \frac{e^{z_{\text{main}}(i)+z_{\text{res}}(i)}}{\sum_{j=1} e^{z_{\text{main}}(j)+z_{\text{res}}(j)}} \quad \text{for} \quad i = 1, \cdots, L$$

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

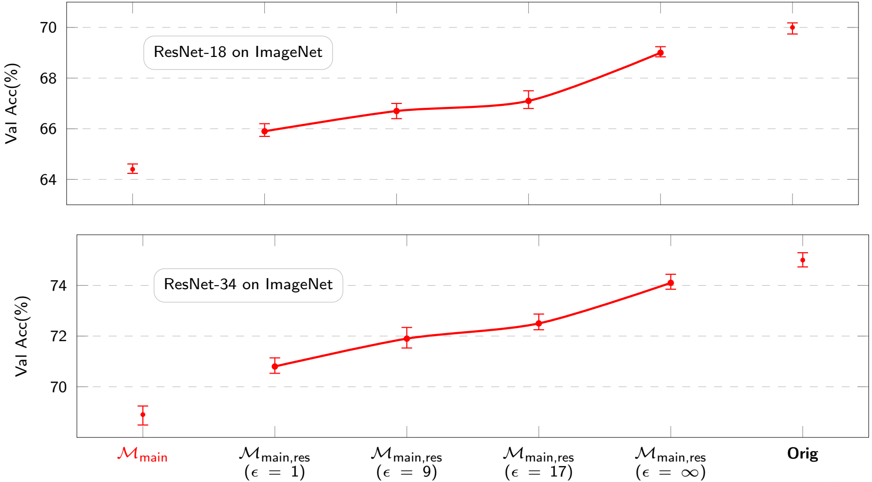
Delta: ML with Fully Asymmetric Data Flow

▷ Experiment Highlights: Utility



Delta: ML with Fully Asymmetric Data Flow

▷ Experiment Highlights: Utility



▷ Experiment Highlights: Utility

	`Delta`: perturb $\mathrm{IR}_{\mathrm{res}}$	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)

Delta: ML with Fully Asymmetric Data Flow

▷ Experiment Highlights: 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

Delta: ML with Fully Asymmetric Data Flow

▷ Experiment Highlights: 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

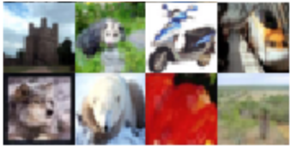
Running time with one single input

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

Delta: ML with Fully Asymmetric Data Flow

▷ Experiment Highlights: Privacy Protection

Against model inversion attack with ResNet-18 [SecretRevealer, CVPR'20]



Original samples

Reconstruction (no noise)

Reconstruction ($\epsilon = 1$)

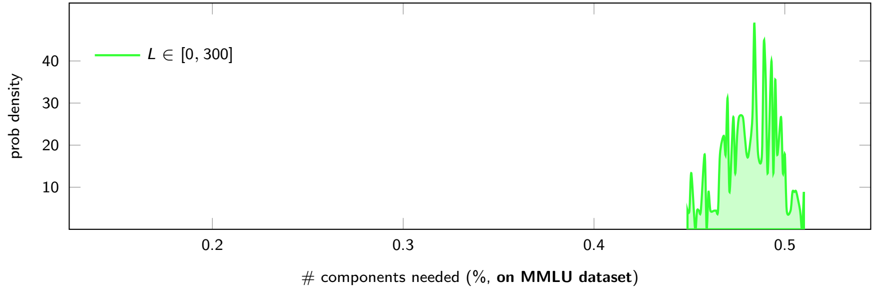
This Thesis

- Discussion of Future Works

Potential in Language Models

Internal activations exhibit highly low-rank structure [arXiv'24]

$$\text{low-rank approximation: } X \xrightarrow{SVD} X_{lr} = U(:, 1:r) \cdot S(1:r, 1:r) \cdot V(1:r, :)$$

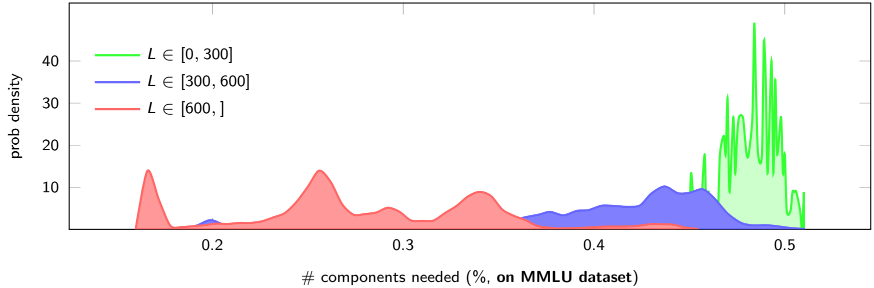


- input sequences can be approximated w. a few principal components

Potential in Language Models

Internal activations exhibit highly low-rank structure [arXiv'24]

$$\text{low-rank approximation: } X \xrightarrow{SVD} X_{lr} = U(:, 1:r) \cdot S(1:r, 1:r) \cdot V(1:r, :)$$



- input sequences can be approximated w. a few principal components
- long sequences exhibit more low-rank structure

Potential in Language Models

An example ···

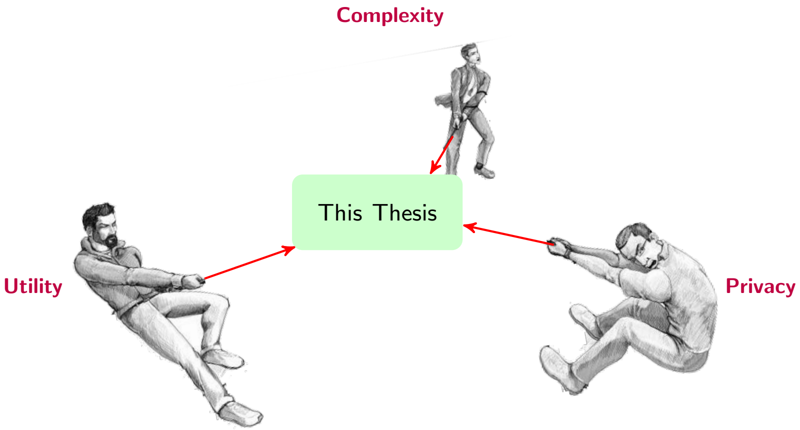
Large Language Models are foundational machine learning models that use deep learning algorithms to process and understand natural language. These models are trained on massive amounts of text data to learn patterns and entity relationships in the language. Large Language Models can perform many types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

(a) Original text

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 **are** coherent and grammatically accurate.

(b) approximated text with 20% principal vectors from Word2Vec.

Conclude: The Privacy-Utility-Complexity Trilemma



Thank You All

Prof. Salman Avestimehr (advisor)

Prof. Murali Annavaram

Prof. Meisam Razaviyayn

Also Advised By: Prof. Mahdi Soltanolkotabi, Prof. Viktor Prasanna

Collaborators: Ramy E. Ali (was a PostDoc at USC), Saurav Prakash (graduated from USC), Sunwoo Lee (was a PostDoc at USC), Lei Gao, Sara Babakniya, Tingting Tang, Tuo Zhang, Zalan Fabian

Labmates: Amir Ziashahabi, Asal Mehradfar, Duygu Nur Yaldiz, Emir Ceyani, Erum Mushtaq, Roushdy Elkordy, Yavuz Faruk Bakman, Chaoyi Jiang, James Flemings, Jonghyun Lee, Rachit Rajat, Tara Renduchintalam