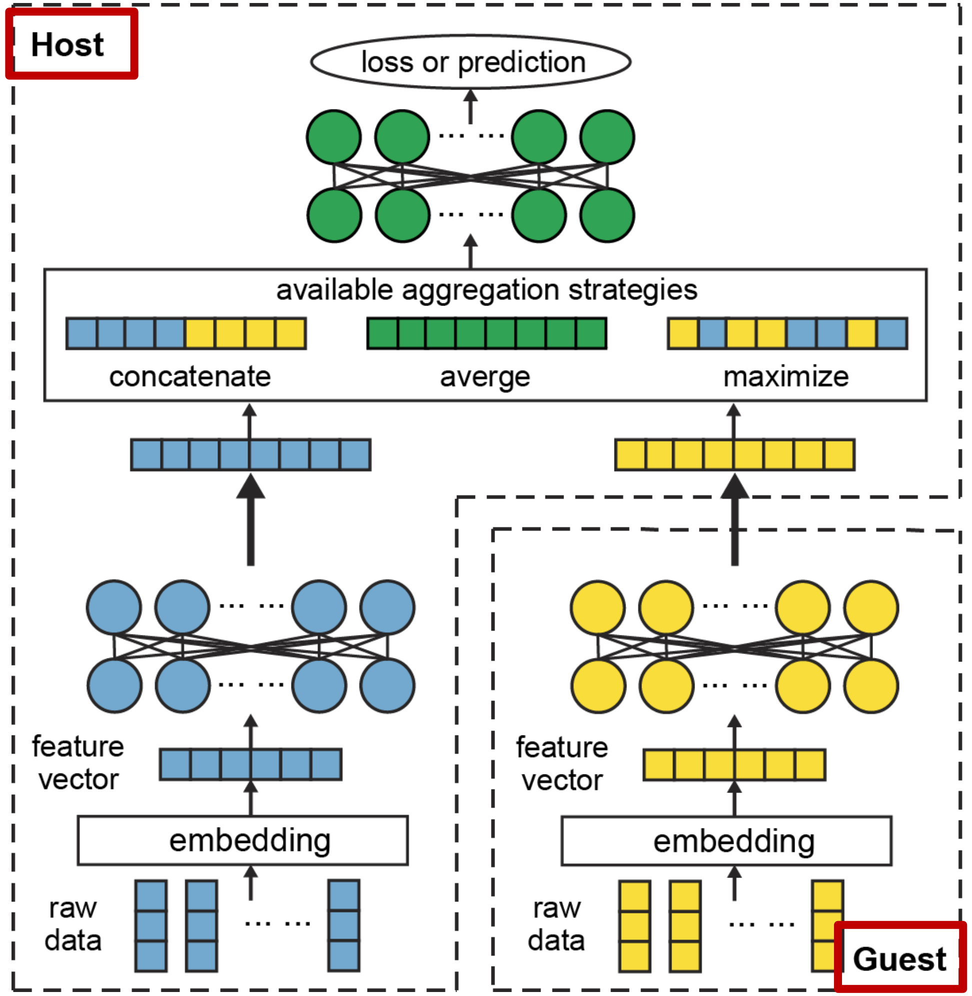
Secure Split Learning against Property Inference, Data Reconstruction, and Feature Space Hijacking Attacks

Yunlong Mao*, Zexi Xin*, Zhenyu Li*†, Jue Hong#, Qingyou Yang#, and Sheng Zhong*

*Nanjing University
#ByteDance
†University of California, San Diego

Split Learning



- Forward pass

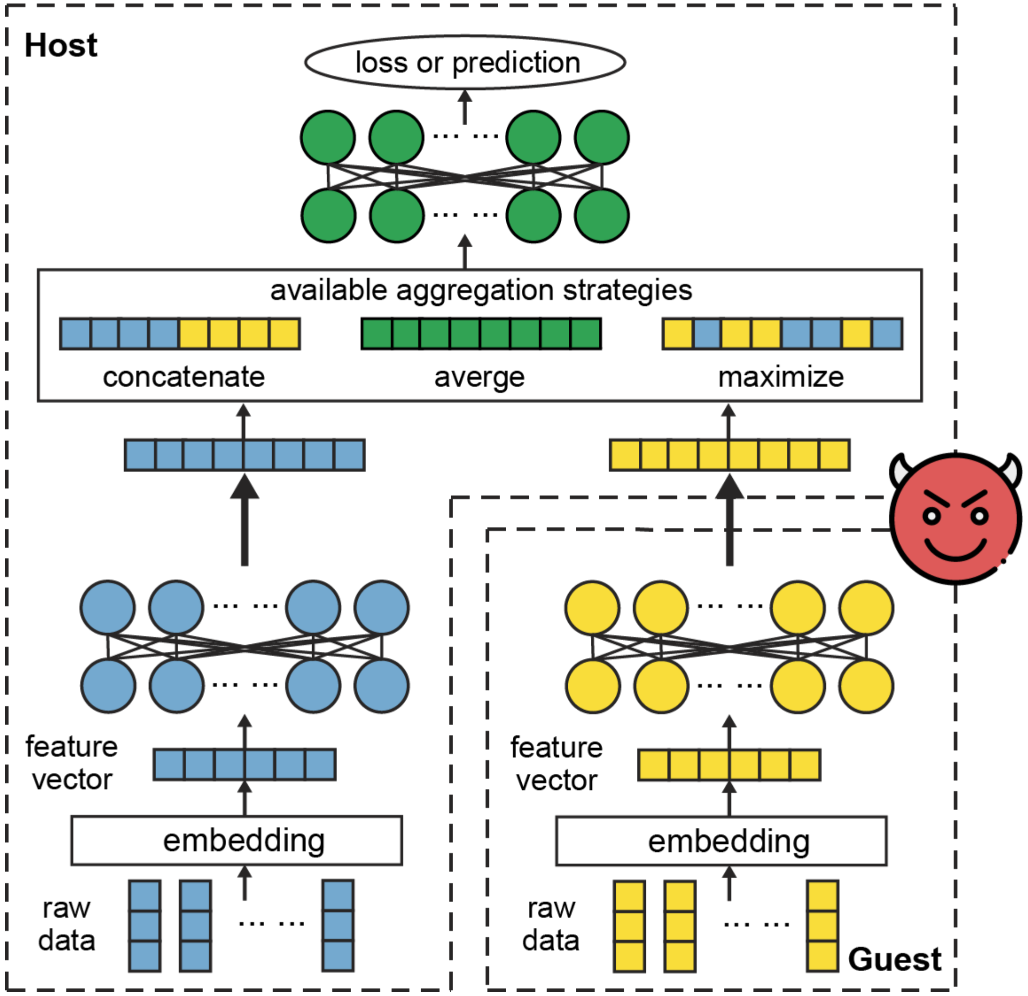
- Guest and Host calculate their forwarding results with their own raw data, respectively.
- Host aggregates the forwarding results.
- Host finishes the loss evaluation.

- Backward propagation

- Host calculates the gradients of her own model.
- Host propagates the partial loss of the guest model.
- Guest and calculates his gradients.
- Host and guest update their models separately.

- Raw data should not be disclosed.

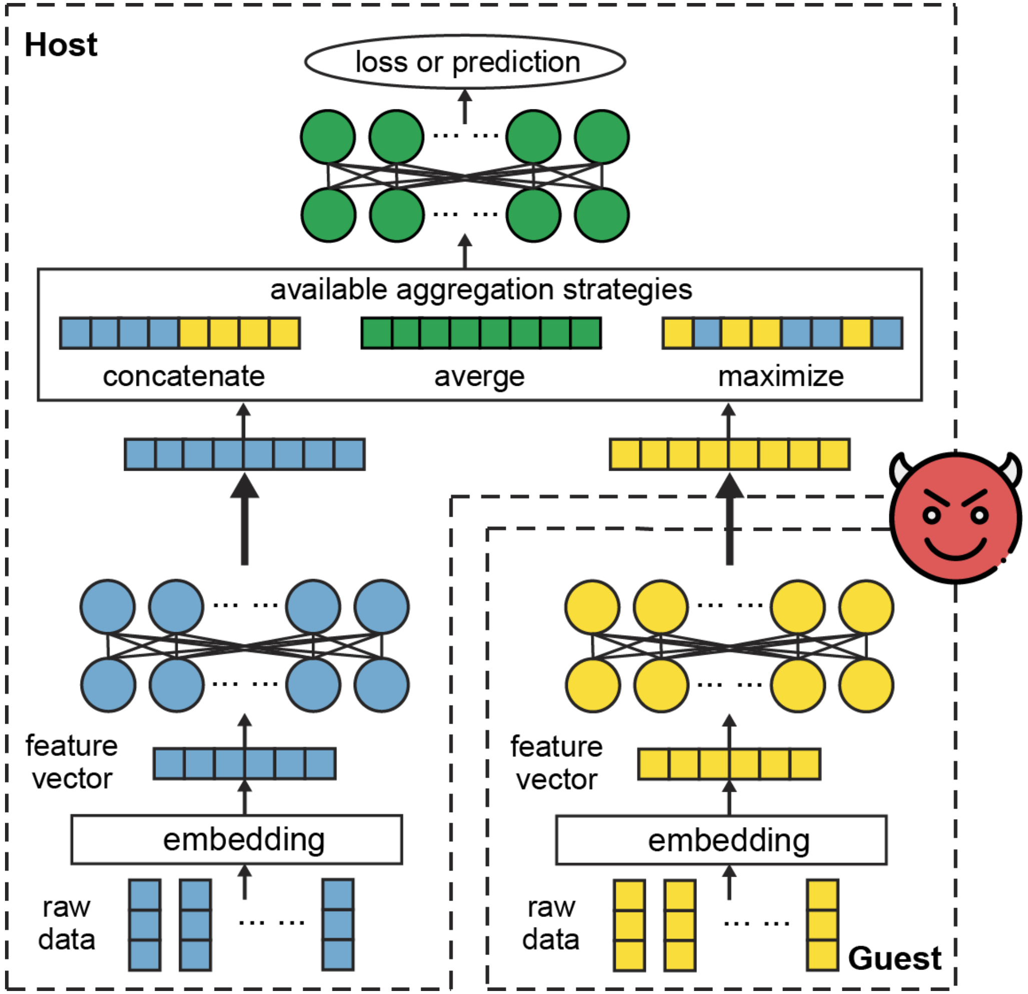
Threats in Split Learning



- Interactions leak privacy.

- **Assumptions**
  - **Honest** but **Curious**
  - Alllow additional computations
  - No out-of-band information exchange except for the interactions
  - Both parties can be adversarial

Threats in Split Learning



- Property inference attack

- Access to the output of the other side is a black-box query.
- Construct surrogate models
- Infer properties of data samples (such as gender or age)

- Data reconstruction attack

- GANs
- Construct a local generator
- Use the global model as a discriminator
- Reconstruct data samples

- Feature space hijacking attack (FSHA)

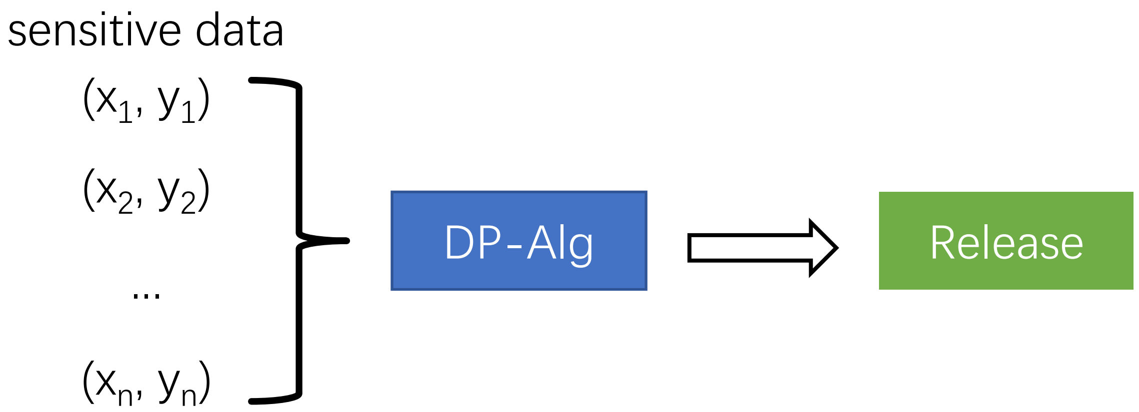
- malicious host
- Unleashing the tiger: Inference attacks on split learning

Defense

- Differential privacy

Definition 2.4 (Differential Privacy)**.** A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$



Defense

- **State-of-art in Deep Learning**
    - DP-SGD
- Advantage: generic
- Drawback: accuracy drop; not suitable for split learning

- Challenges in split learning:
  - ➢ Asymmetric parties (different models and different data)
  - ➢ Interactions happen in both forward and backward passes (only update information will be revealed in FL)
  - ➢ The host and the guest should be protected against each other (not required in FL)

R^3eLU

- Pure randomized response
  - advantage: good at statistical analysis
  - drawback: hard to deal with learning
- Pure Laplace mechanism
  - advantage: generic recipe for continuous variables
  - drawback: sensitive
- **R^3eLU (randomized-response ReLU)**
  - activations as item sets
  - add noise on values

$$\mathrm{R}^3\mathrm{eLU}(v) = \begin{cases} \max(0, v+z), & \text{with probability } p, \\ 0, & \text{with probability } (1-p). \end{cases}$$

R^3eLU

- Pure randomized response
    - advantage: good at statistical analysis
    - drawback: hard to deal with learning
- Pure Laplace mechanism
    - advantage: generic recipe for continuous variables
    - drawback: sensitive
- **R^3eLU (randomized-response ReLU)**
    - activations as item sets
    - add noise on values

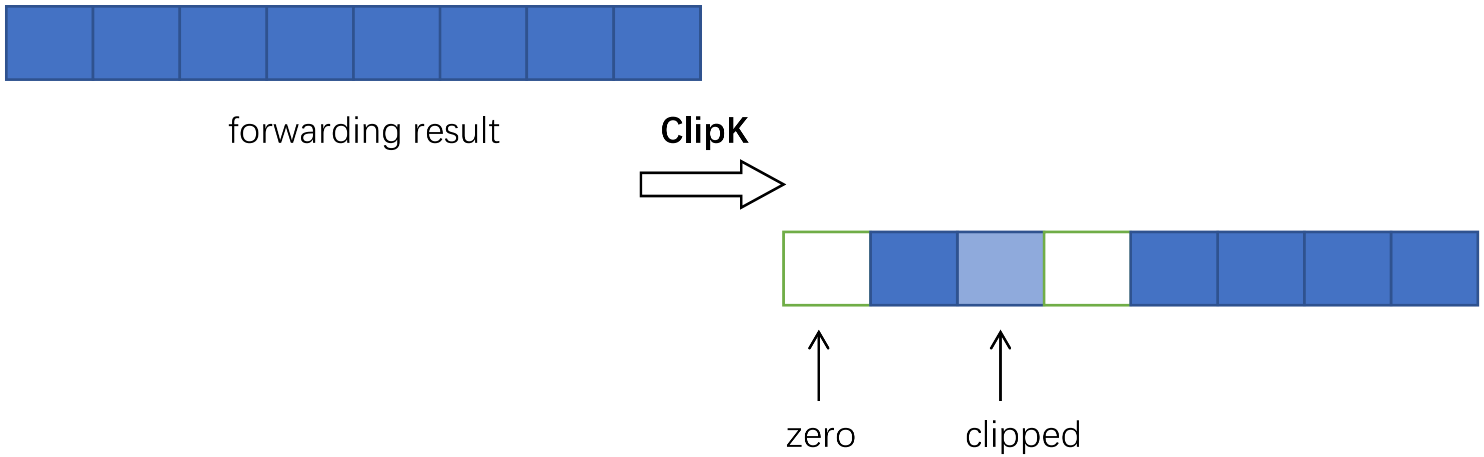
Key idea: activation states should be protected as well

$$\mathrm{R^3eLU}(v) = \begin{cases} \max(0, v+z), & \text{with probability } p, \\ 0, & \text{with probability } (1-p). \end{cases}$$

Forward Propagation with R^3eLU

- Replace ReLU with R^3eLU
    - protect the guest from an adversarial host

- Pre-process: ClipK
    - select the top **K** largest elements of forwarding result
    - clip the value by **C**
    - constrain the sensitivity to 2KC



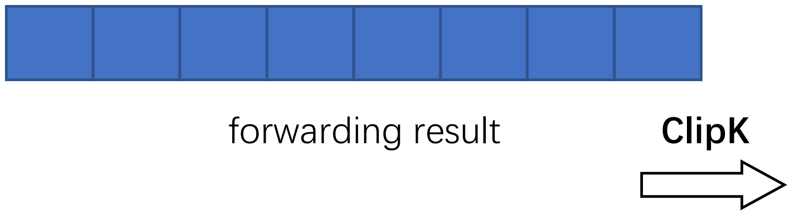
Forward Propagation with R³eLU

- R³eLU
  - randomly deactivate each activation with probabiliy 1-p_i

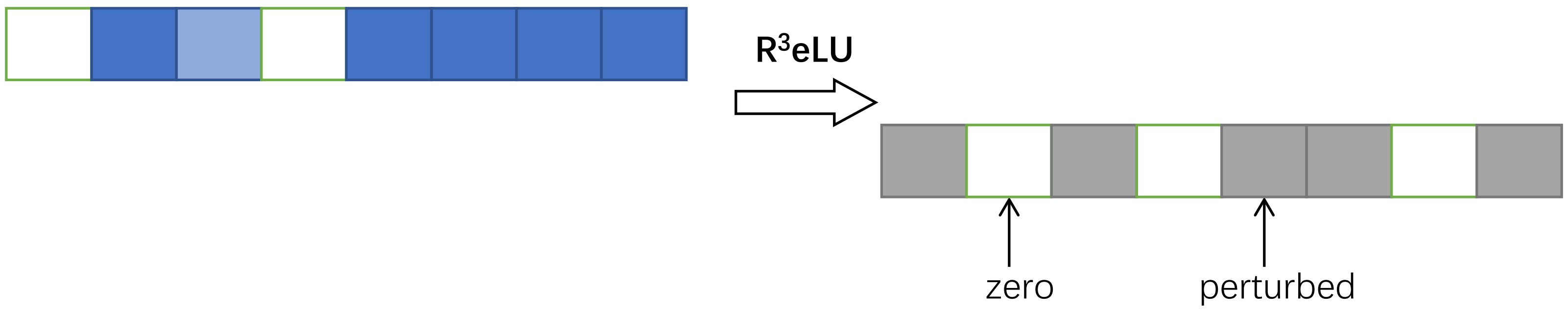
$$p_i = \frac{1}{2} + \boxed{\frac{\hat{v}_i}{\|\hat{\boldsymbol{v}}\|_\infty}} \cdot \left(\frac{e^{\frac{\epsilon_p}{K}}}{1+e^{\frac{\epsilon_p}{K}}} - \frac{1}{2}\right),$$

  - add Laplacian noise to the remaining value

$$\mathtt{Lap}(0, \frac{2KC}{\epsilon_l})$$



Corollary 1. *Given privacy budgets ϵ_p and ϵ_l for randomized response and Laplace mechanism respectively, the output of R^3eLU-forward procedure is $(\epsilon_p + \epsilon_l)$-DP.*

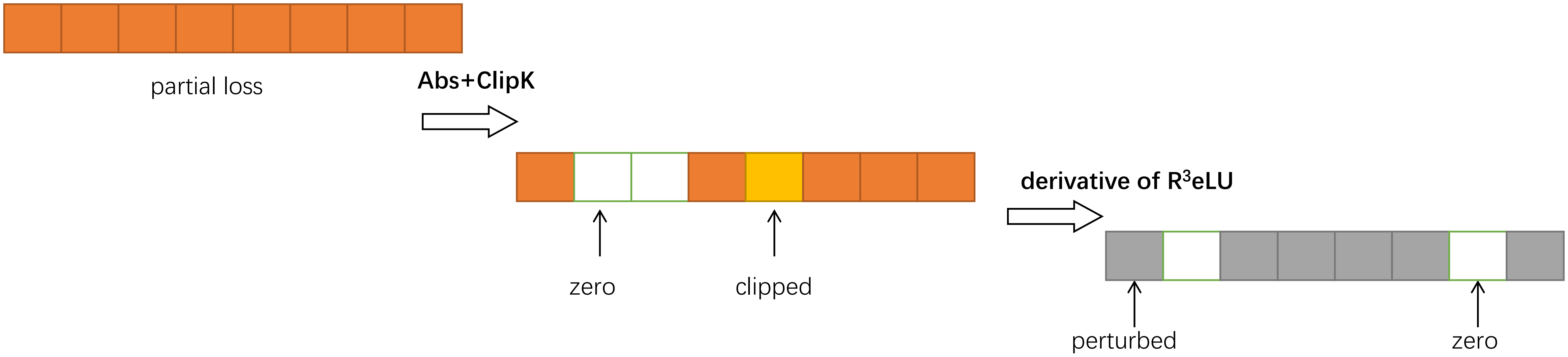


Private Backward Propagation

- construct a privacy-preserving tunnel
    - protect the host from an adversarial guest

- derivative of R³eLU

$$\nabla \mathrm{R}^3\mathrm{eLU}(\boldsymbol{\delta}^g, \tilde{\boldsymbol{a}}^g, \boldsymbol{v}^g) = \begin{cases} \boldsymbol{\delta}^g + \boldsymbol{z}, & \text{with probability } p, \\ 0, & \text{with probability } (1-p), \end{cases}$$



Dynamic Privacy Budget Allocation

- Importance estimation
  - Key: give important neuron higher privacy budget
  - parameter's importance: $\hat{I}_j = (\nabla_{\theta_j} \mathcal{L}(\boldsymbol{\theta}, x) \cdot \theta_j)^2.$
  - neuron's importance: joint importance of relevant parameters $U_j = \sum_{\theta_k \in \boldsymbol{\theta}_{U_j}} \hat{I}_k,$
  - dynamic estimation: $$U_j^q = \frac{\sum_{\theta_k \in \tilde{\boldsymbol{\theta}}_j} \hat{I}_k + U_j^{q-1} \times (q \times \lfloor T/n_t \rfloor + (t \bmod n_t) - 1)}{q \times \lfloor T/n_t \rfloor + (t \bmod n_t)},$$

The importance may change during the training. The importance of a neuron will be accumulated as the training epoch increases. The additional cost is only $O(N_u)$.

- application:

budget allocation: $\epsilon_j = \epsilon \times U_j^q,$

probability adjustment: $$p_i = \frac{1}{2} + \frac{U_i^q}{\|\boldsymbol{U}\|_\infty} \cdot \left(\frac{e^{\frac{\epsilon_p}{K}}}{1 + e^{\frac{\epsilon_p}{K}}} - \frac{1}{2}\right).$$

Dynamic Privacy Budget Allocation

- Iteration budget allocation
    - earlier iterations have higher budget for utility, later iterations have lower budget for the privacy concern.
    - a recommendation for iteration budget allocation:

$$\epsilon_i = \frac{\epsilon_T}{2^i}$$

Experiments

- **Setup**

  - Datasets: MovieLens and BookCrossing for recommendation
  MNIST and CIFAR100 for image classfication

  - Model Architecture: MLP for the recommendation
  ResNet for the image classifation

  $\epsilon_p = \epsilon_l = \frac{\epsilon}{2}$

  - Hyperparameters: batch size 32, learning rate 0.01, Adam optimizer
  , K=half of the features of the cutlayer, C=10

  - Baselines: split learning without any protection
  DPSGD
  Laplacian mechanism

Evaluation on dynamic importance estimation

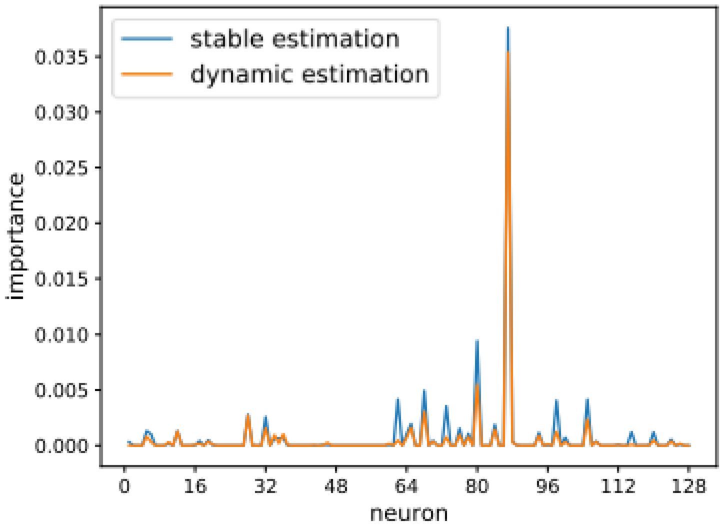


Fig. 1. Estimation results of neuron importance.

Correctness of dynmaic importance estimation

Existence of unbalanced feature importance

Evaluation on model usability

- metric: model accuracy

	MovieLens	BookCrossing	MNIST	CIFAR100
Baseline	56.62%	61.70%	98.00%	76.20%

Table 2. Model usability results while preserving the privacy of the guest.

ϵ	MovieLens			BookCrossing			MNIST			CIFAR100		
	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**
0.1	30.84%	32.29%	**34.03%**	57.02%	55.89%	**58.18%**	17.43%	30.21%	**32.41%**	20.25%	34.21%	**37.87%**
0.5	41.25%	43.69%	**43.87%**	57.67%	56.14%	**58.54%**	27.33%	58.43%	**60.38%**	39.74%	51.36%	**58.22%**
1.0	48.16%	49.09%	**50.56%**	58.02%	56.56%	**58.42%**	31.05%	75.58%	**76.60%**	46.19%	60.48%	**65.32%**
2.0	49.32%	50.38%	**50.49%**	58.74%	56.91%	**59.24%**	38.92%	92.90%	**93.53%**	56.30%	69.72%	**73.35%**
4.0	49.26%	**50.86%**	50.73%	59.01%	57.16%	**59.26%**	95.37%	**95.87%**	94.12%	57.04%	70.86%	**74.41%**

Table 3. Model usability results while preserving the privacy of the host.

ϵ	MovieLens			BookCrossing			MNIST			CIFAR100		
	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**
0.1	31.47%	30.68%	**33.98%**	57.37%	57.46%	**58.26%**	27.64%	**33.45%**	32.36%	17.04%	34.22%	**38.96%**
0.5	41.75%	42.31%	**42.67%**	58.62%	58.24%	**58.59%**	55.38%	65.28%	**67.83%**	25.82%	43.96%	**53.72%**
1.0	47.43%	48.29%	**50.39%**	59.49%	58.44%	**59.77%**	71.95%	**89.74%**	88.14%	37.69%	55.28%	**61.48%**
2.0	49.86%	50.43%	**51.47%**	59.34%	59.97%	**60.27%**	89.15%	**92.66%**	92.52%	51.87%	65.67%	**69.60%**
4.0	49.57%	50.09%	**51.62%**	59.55%	**60.75%**	60.66%	94.61%	**95.37%**	95.01%	51.87%	66.89%	**70.70%**

Accuracy improves

Evaluation on privacy preservation

- **Defense against property inference attack**

  - metric: attack accuracy

	MovieLens	BookCrossing	MNIST	CIFAR100
Adversarial host	80%	79%	94%	87%
Adversarial guest	80%	78%	57%	53%

Table 4. Results of defending the guest against property inference attack.

ϵ	MovieLens			BookCrossing			MNIST			CIFAR100		
	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**
0.1	66.99%	77.71%	**60.99%**	**54.76%**	73.29%	55.78%	**43.27%**	53.95%	44.33%	**50.76%**	79.14%	53.97%
0.5	66.16%	74.23%	**64.16%**	**54.97%**	74.52%	56.33%	46.92%	54.23%	**45.59%**	**51.47%**	79.26%	55.13%
1.0	**67.19%**	78.65%	68.18%	**55.03%**	74.96%	58.65%	47.58%	54.26%	**47.51%**	**50.40%**	79.37%	55.72%
2.0	68.65%	73.06%	**68.56%**	**54.85%**	74.26%	58.14%	**48.06%**	54.65%	52.87%	60.76%	79.37%	**58.81%**
4.0	**69.14%**	76.18%	71.91%	**54.92%**	74.33%	60.76%	**48.47%**	54.57%	55.73%	60.81%	79.35%	**58.03%**

Table 5. Results of defending the host against property inference attack.

ϵ	MovieLens			BookCrossing			MNIST			CIFAR100		
	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**
0.1	53.46%	78.59%	**51.86%**	**54.55%**	74.35%	59.42%	60.34%	80.29%	**48.74%**	50.42%	51.46%	**41.89%**
0.5	53.46%	75.64%	**51.89%**	**54.62%**	74.36%	59.42%	59.82%	81.92%	**49.71%**	50.38%	52.23%	**44.26%**
1.0	53.46%	73.54%	**52.75%**	**54.95%**	74.39%	59.52%	59.74%	82.80%	**50.48%**	49.95%	51.95%	**44.78%**
2.0	**53.47%**	75.05%	59.77%	**54.40%**	74.39%	58.13%	60.38%	88.88%	**50.57%**	50.77%	51.67%	**50.16%**
4.0	**53.48%**	79.28%	56.52%	**54.95%**	74.39%	62.04%	60.62%	89.73%	**50.47%**	51.52%	51.76%	**51.27%**

Evaluation on privacy preservation

- **Defense against data reconstruction attack**

  - metric: MSE

	MovieLens	BookCrossing	MNIST	CIFAR100
Adversarial host	0.2412	0.2629	0.9612	2.6335
Adversarial guest	0.2369	0.2402	1.6998	5.7534

Table 6. Results of defending the guest against data reconstruction attack.

ϵ	MovieLens			BookCrossing			MNIST			CIFAR100		
	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**
0.1	0.2459	0.2455	**0.3223**	0.3216	0.2907	**0.3329**	1.8849	1.8885	**2.0181**	**12.5145**	2.8622	3.6983
0.5	0.2453	0.2451	**0.3222**	0.3202	0.2902	**0.3329**	1.8024	1.8137	**1.9875**	**12.5262**	2.8537	3.6891
1.0	0.2453	0.2451	**0.3222**	0.3202	0.2902	**0.3221**	1.7857	1.7509	**1.9533**	3.6419	2.8351	**3.6624**
2.0	0.2452	0.2451	**0.3222**	0.3202	0.2902	**0.3221**	1.7336	1.7469	**1.9391**	2.9453	2.7998	**3.6383**
4.0	0.2452	0.2451	**0.3222**	0.3202	0.2902	**0.3221**	1.7014	1.7440	**1.9206**	2.9502	2.7743	**3.6365**

Table 7. Results of defending the host against data reconstruction attack.

ϵ	MovieLens			BookCrossing			MNIST			CIFAR100		
	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**	Laplace	DPSGD	**Ours**
0.1	0.4032	0.2417	**0.5486**	0.4237	0.2758	**0.5066**	1.2887	1.0875	**1.8257**	**13.2849**	6.0999	6.5283
0.5	0.4024	0.2419	**0.5357**	0.4222	0.2756	**0.5149**	1.2778	1.0685	**1.7758**	**12.8057**	5.9302	6.3719
1.0	0.4008	0.2422	**0.5285**	0.4217	0.2743	**0.5235**	1.2602	1.0422	**1.7528**	**12.7936**	5.9283	6.3531
2.0	0.3982	0.2421	**0.5083**	0.4214	0.2697	**0.5150**	1.2613	1.0333	**1.7334**	6.0397	5.9256	**6.3453**
4.0	0.3960	0.2422	**0.4819**	0.4194	0.2683	**0.5046**	1.2549	0.9996	**1.7262**	6.0143	5.9247	**6.3396**

Evaluation on privacy preservation

- **Defense against feature space hijacking attack (FSHA)**

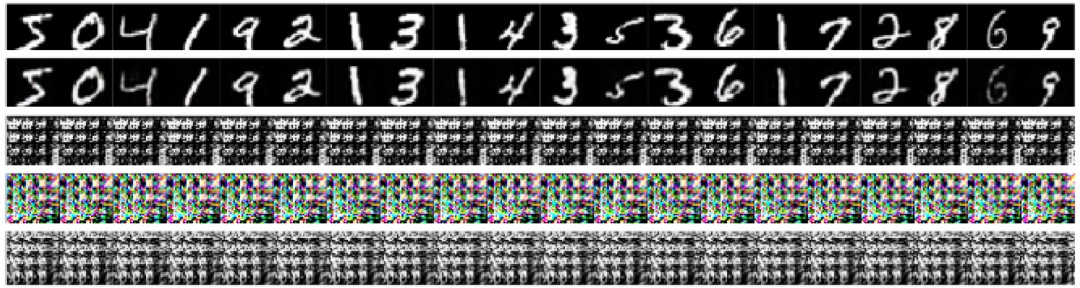


Fig. 2. Reconstruction results of FSHA against the guest's data in the first row. The following rows are attack results against the original SplitNN and our solution ($\epsilon = 0.1, 1.0, 4.0$), respectively.

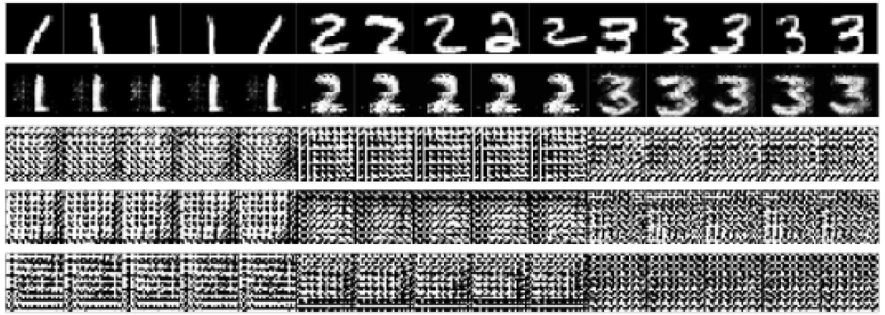


Fig. 3. Reconstruction results of FSHA against the host's data in the first row. The following rows are attack results against the original SplitNN and our solution ($\epsilon = 0.1, 1.0, 4.0$), respectively.

Thanks!
Q & A