DeepReDuce: ReLU Reduction for Fast Private Inference

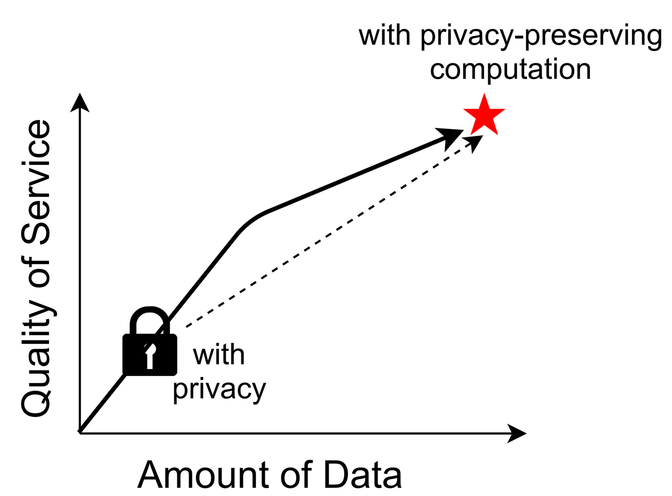
Nandan Kumar Jha, Zahra Ghodsi, Siddharth Garg, Brandon Reagen

CENTER FOR CYBER SECURITY

NYU

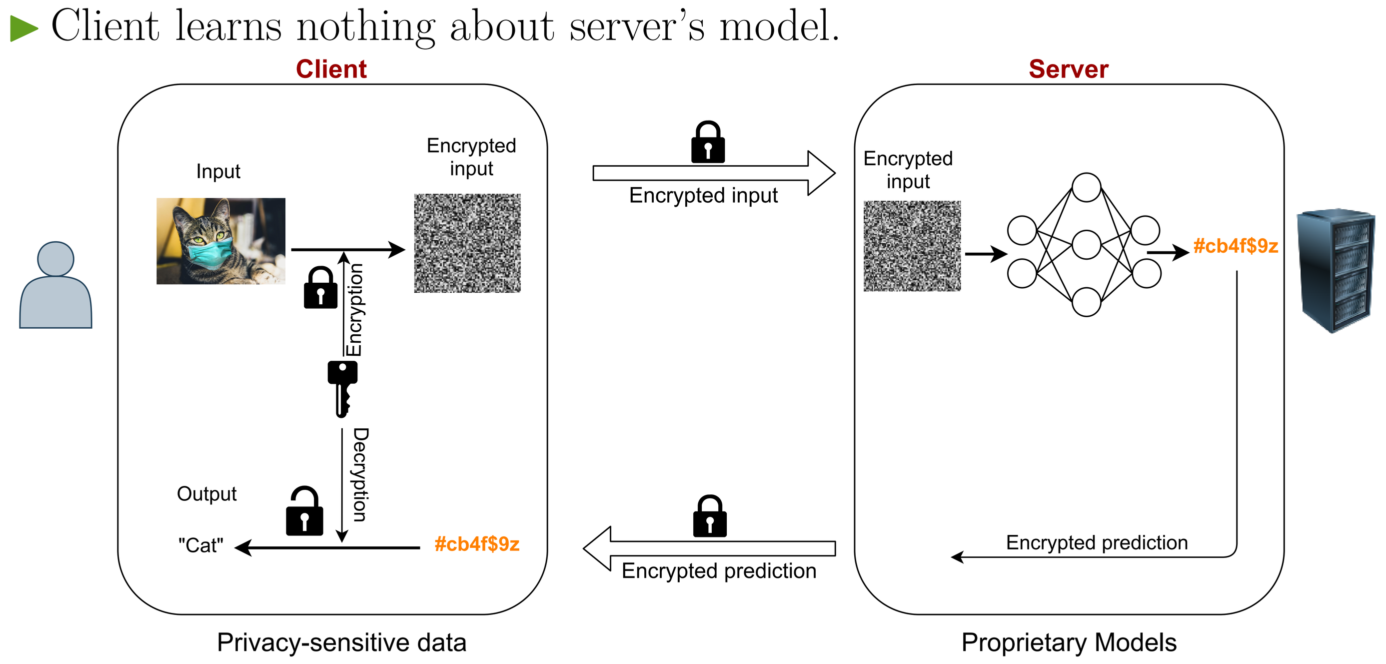
Motivation

- Deep learning as a service (DLaaS) gives rise to privacy concerns:
  - Client's input are privacy-sensitive and server's models are IP of service provider.
- There is an inherent tradeoff between privacy and QoS:
  - Users sacrifice the QoS for higher privacy guarantee.
- Privacy-preserving computation breaks the QoS-privacy tradeoff:
  - Users can get high QoS with higher privacy guarantee.



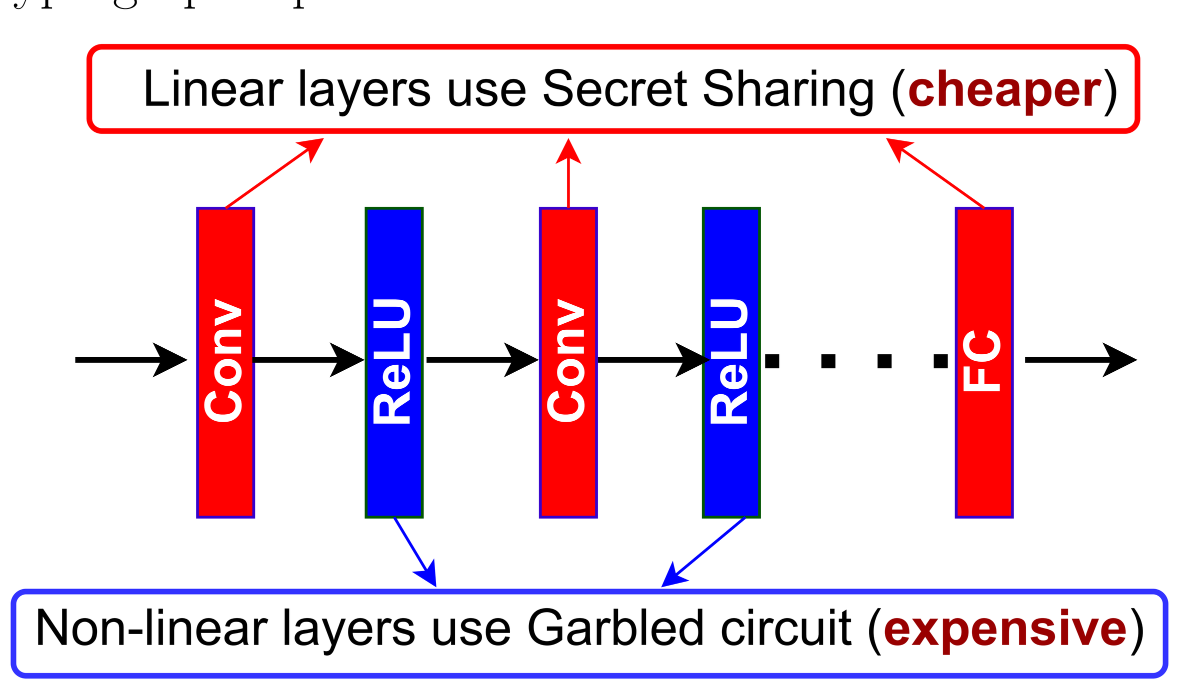
Private Inference

- In private inference, neural network computation is performed directly on encrypted data such that:
  - Server learns nothing about client's input.
  - Client learns nothing about server's model.



Source of slowdown in Private Inference

- In private inference, linear and nonlinear layers use different cryptographic protocols.



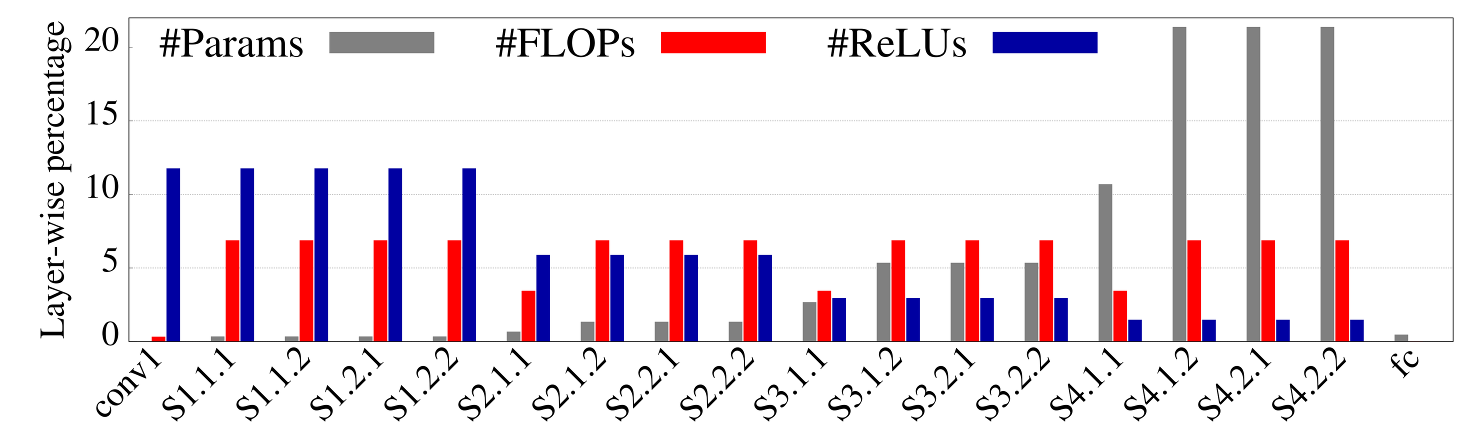
- Inverted operator cost in private inference:
  - ReLUs are 3 to 4 orders of magnitude slower than convolution [1].
  - ReLUs contribute ∼99% in total online latency [2].

DeepReDuce

- ReLUs in neural networks exhibit heterogeneity in terms of their impact on accuracy.

▶ ReLUs' Heterogeneity

- Layer-wise distribution of ReLU
  - Usually initial layers have higher #ReLUs and layer-wise ReLU count decreases in deeper layers.



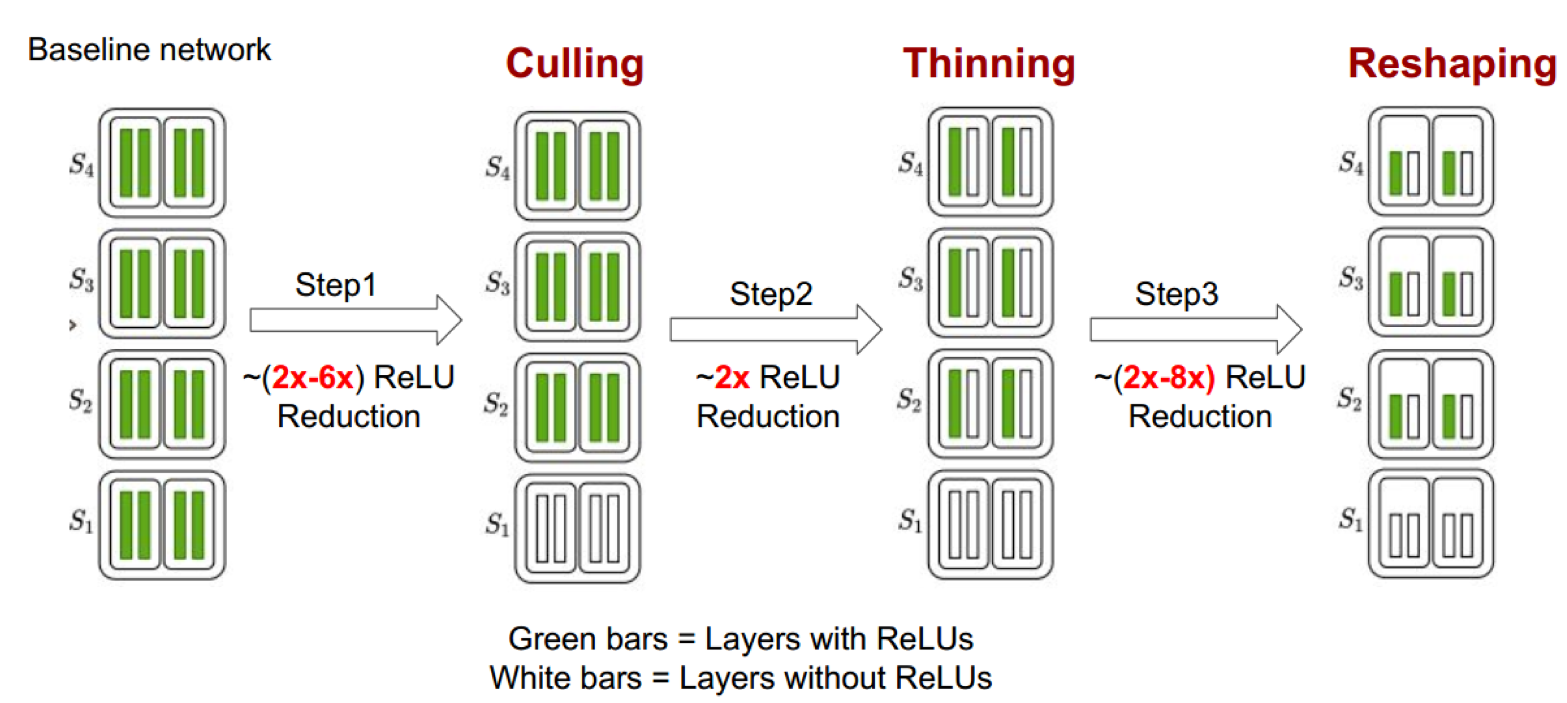
- ReLUs' criticality for network's accuracy.
  - ReLUs in middle layers are more critical than ReLUs in initial and last layers.

Models	Metrics	No ReLUs	Conv1	S_1	S_2	S_3	S_4
ResNet18	#ReLUs	0	66K	262K	131K	66K	33K
	W/o KD (%)	18.49	46.22	61.93	67.63	67.41	58.90
	W/ KD (%)	18.34	45.07	59.85	68.79	69.92	63.16
ResNet34	#ReLUs	0	66K	393K	262K	197K	49K
	W/o KD(%)	18.16	45.42	60.77	69.47	70.04	57.44
	W/ KD(%)	18.07	45.13	62.88	70.93	72.61	64.23

DeepReDuce achieves ReLU saving with minimal impact on accuracy by dropping the less critical while preserving most critical ReLUs.

▶ ReLU optimization steps in DeepReDuce

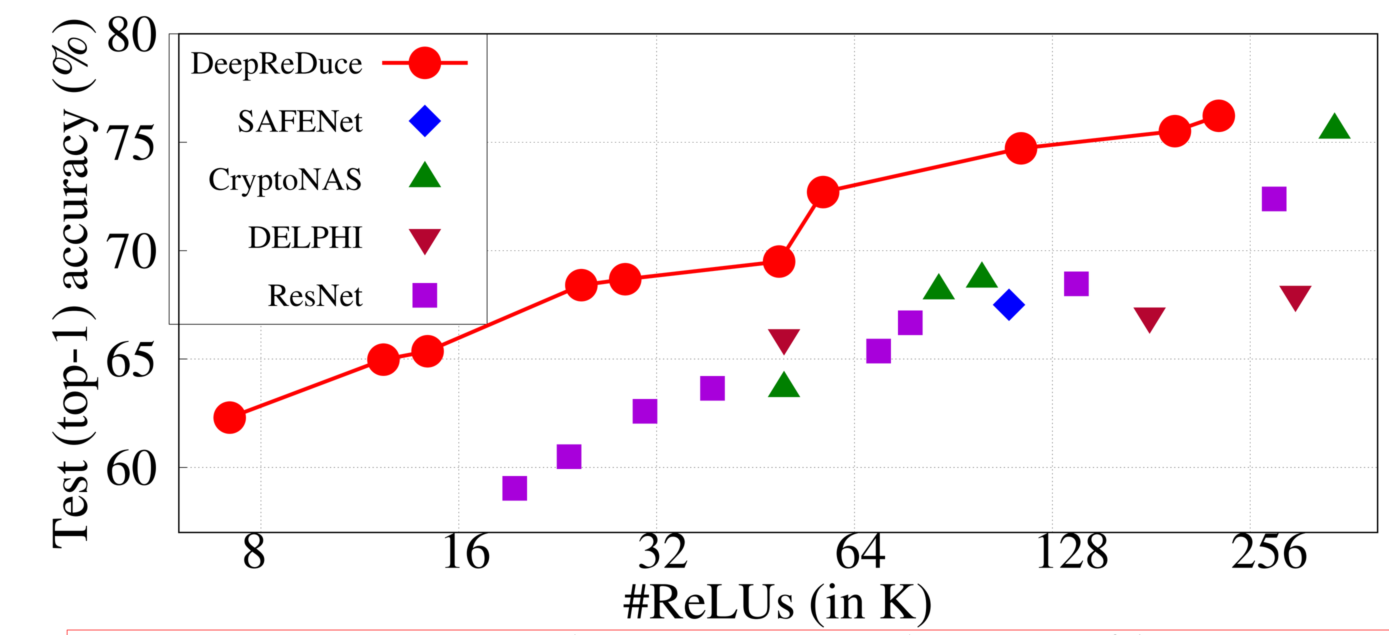
- ReLU Culling
  - Given a baseline full ReLU network, it first drops/removes ReLUs from least critical stage.
- ReLU Thinning
  - Drops ReLUs from the alternate layers in the remaining non-Culled stages.
- ReLU Reshaping
  - Employ conventional channel and/or feature map resolution scaling in all the layers of network to achieve very low ReLU count.



DeepReDuce outputs a Pareto-frontier of ReLU optimized networks with different ReLU counts and accuracy.

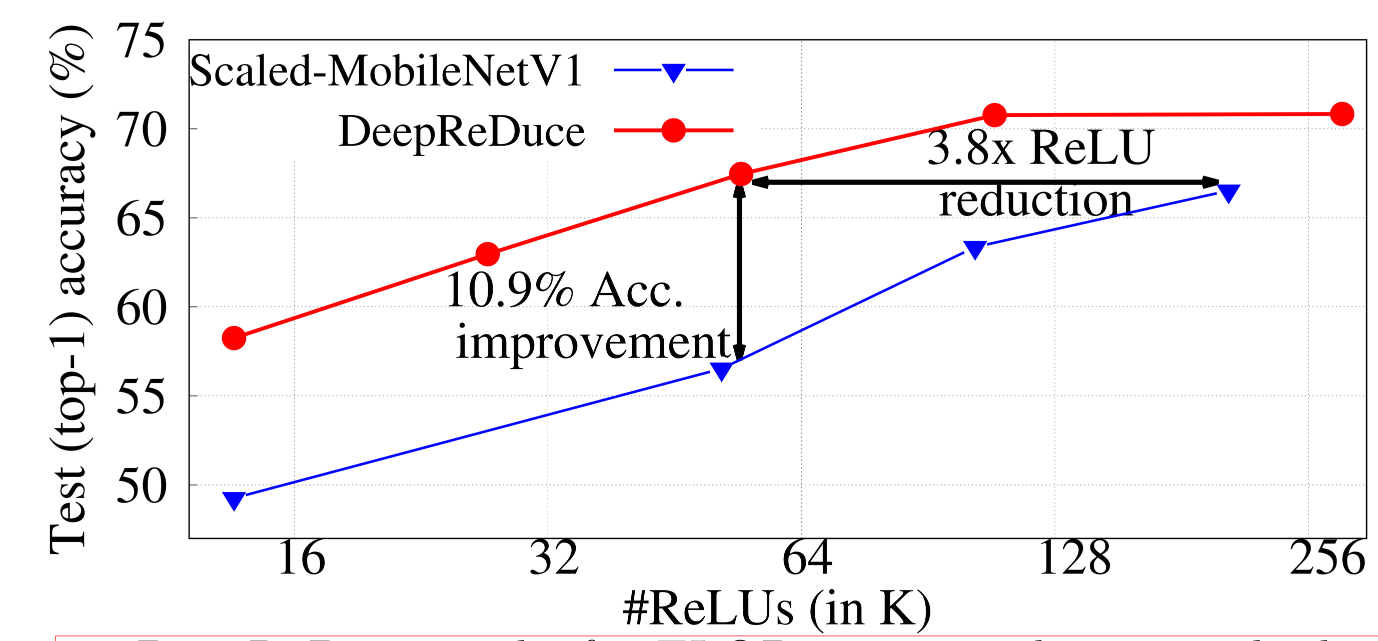
DeepReDuce Optimizations Evaluation

- Comparison with state-of-the-art in private inference



3.5× ReLU saving (at iso-accuracy) and 3.5% accuracy improvement (at iso-ReLU count) on CIFAR-100

- Generality case study with MobileNetV1 on CIFAR-100



DeepReDuce works for FLOPs-optimized non-residual network. Hence, DeepReDuce generalize beyond ResNet

- Comparison with state-of-the-art channel pruning method

	Method	Baseline Acc.(%)	Pruned Acc.(%)	Acc. ↓(%)	FLOPs	ReLUs
C10	Ch. pruning [3]	93.59	93.34	-0.25	59.1M	311.7K
	DeepReDuce	93.48	94.07	+0.59	87.7M	221.2K
			93.16	-0.32	66.5M	147.5K
C100	Ch. pruning [3]	71.41	70.83	-0.58	60.8M	311.7K
	DeepReDuce	70.93	73.66	+2.57	87.7M	221.2K
			71.68	+0.59	66.5M	147.5K

2× more ReLU saving with similar FLOPs and accuracy on CIFAR-10 (C10) and CIFAR-100 (C100).

References

[1] Z. Ghodsi *et al.*, "CryptoNAS: Private inference on a relu budget," *Neural Information Processing Systems*, 2020.

[2] Q. Lou *et al.*, "SAFENet: A secure, accurate and fast neural network inference," in *International Conference on Learning Representations*, 2021.

[3] Y. He *et al.*, "Learning filter pruning criteria for deep convolutional neural networks acceleration," in *CVPR*, 2020, pp. 2009–2018.

Contact

nj2049 [at] nyu [dot] edu