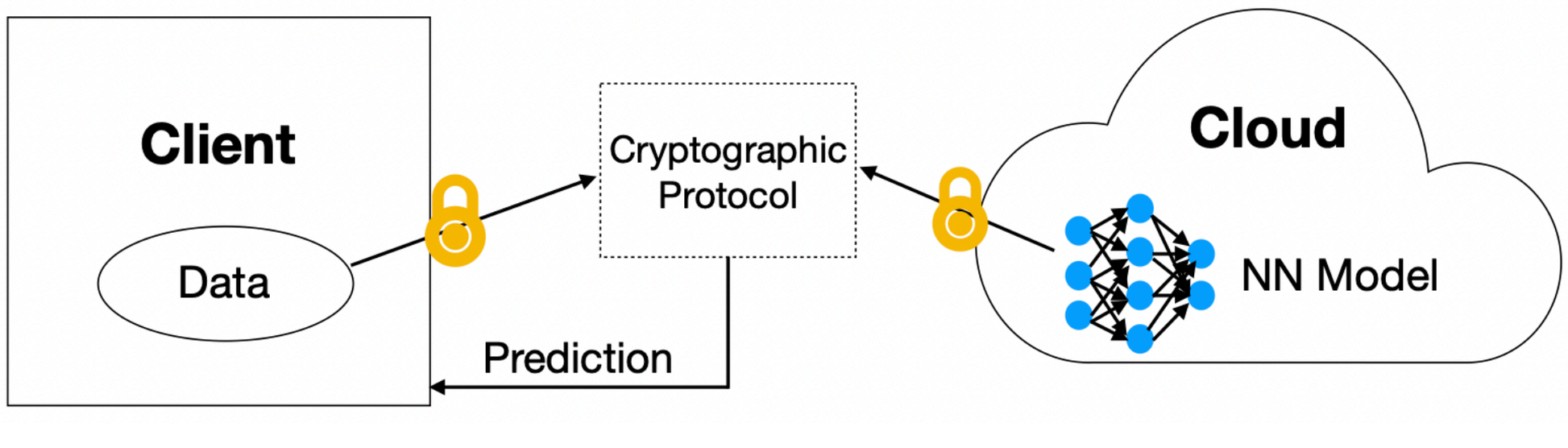
Selective Network Linearization for Efficient Private Inference

Minsu Cho, Ameya Joshi, Siddharth Garg, Brandon Reagen, Chinmay Hegde

New York University

International Conference on Machine Learning, July 2022

Motivations



ReLU are main computational bottleneck
in Private Inference

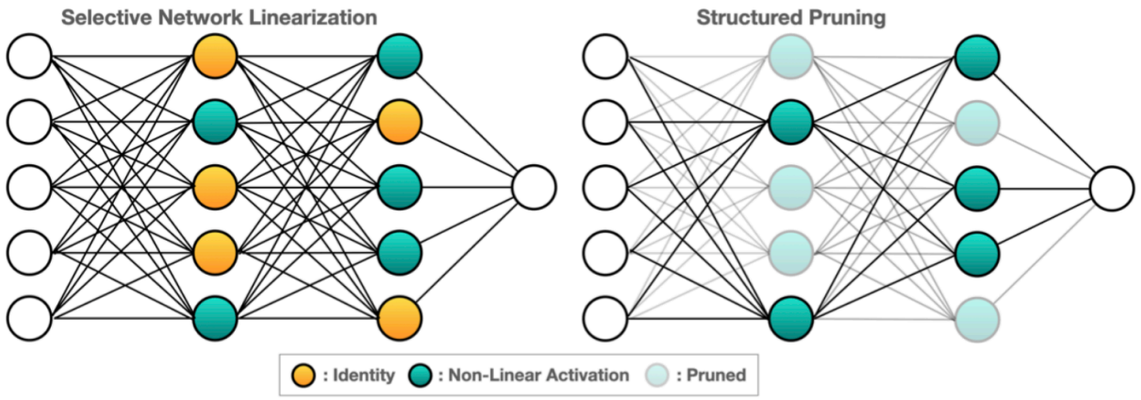
Mishra et al. DELPHI: A Cryptographic Inference Service for Neural Networks, USENIX Security 20

Prior Works on Secure Inference

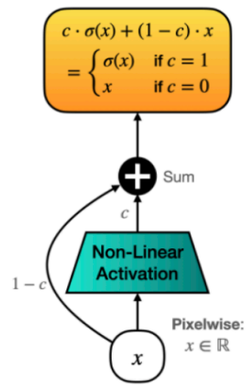
Approach	Methods	Reduce ReLUs	Units that are removed
CryptoNAS (Ghodsi et al., 2020)	NAS	Yes	layers
Sphynx (Cho et al., 2021a)	NAS	Yes	layers
DELPHI (Mishra et al., 2020)	NAS + polynomial approx.	Yes	layers
SAFENet (Lou et al., 2021)	NAS + polynomial approx.	Yes	channels
Unstructured Pruning	N/A	No	not exist
Structured Pruning	N/A	Yes	channels, layers
DeepReDuce(Jha et al., 2021)	manual	Yes	layers
SNL (ours)	gradient-based	Yes	pixels, channels

Table 1. Comparison of various techniques that reduce ReLU operations in deep networks. NAS stands for neural architecture search. "Pruning" techniques eliminate entire neurons. SNL, our proposed gradient-based network linearization method, achieves the accuracy-latency Pareto frontier in private inference.

Selective Network Linearization



(a) Visual comparison on SNL and Structured Pruning



(b) Visual representation of Equation 2

Figure 2. Visualization of SNL, structured pruning, and Equation 2. (a) Both SNL and structured pruning have two non-linear activations. While all 55 parameters are on in SNL, the network from structured pruning has only 18 parameters. We note that number of non-linear activations (especially ReLU) is what matters in PI. (b) Visual representation of the convex combination between x and $\sigma(x)$. If non-linear activation σ is ReLU and $c \in \mathbb{R}$, then this convex combination is equivalent to PReLU.

Algorithm 1 SNL: Selective Network Linearization

Inputs: $f_{\mathbf{W}}$: pre-trained network, λ: Lasso coefficient, κ: scheduling factor, B: ReLU budget, ϵ: threshold.
Set $\mathbf{C} = 1$: same dimensions to all feature maps.
$\overline{\mathbf{W}} \leftarrow (\mathbf{W}, \mathbf{C})$
while ReLU Count $> B$ **do**
  Update $\overline{\mathbf{W}}$ via ADAM for one epoch.
  ReLU Count $\leftarrow \|\mathbb{1}(\mathbf{C} > \epsilon)\|_0$
  if ReLU count not decreased **then**
    Increment Lasso coefficient $\lambda \leftarrow \kappa \cdot \lambda$.
  end if
end while
$\mathbf{C} \leftarrow \mathbb{1}(\mathbf{C} > \epsilon)$
Freeze $\mathbf{C}$ and finetune $f_{\mathbf{W}}$.

Pareto Analysis on Test Acc. vs ReLU

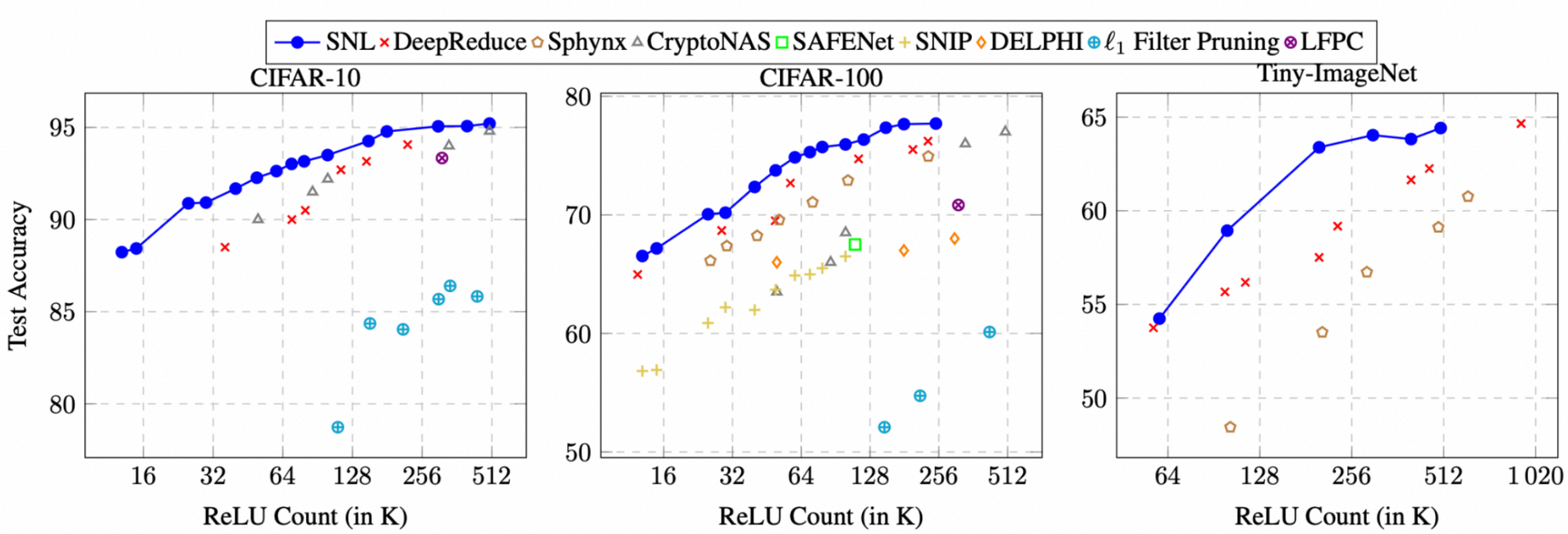


Figure 3. SNL achieves Pareto frontiers of ReLU counts versus test accuracy on CIFAR-10, CIFAR-100, and Tiny-ImageNet. SNL outperforms the state-of-the-art methods (DeepReDuce, SAFENet, and CryptoNAS) in all range of ReLU counts on all three dataset. ℓ_1 Filter Pruning (Li et al., 2016) and LFPC (He et al., 2020b) are structured pruning techniques.