CryptGPU: Fast Privacy-Preserving Machine Learning on the GPU

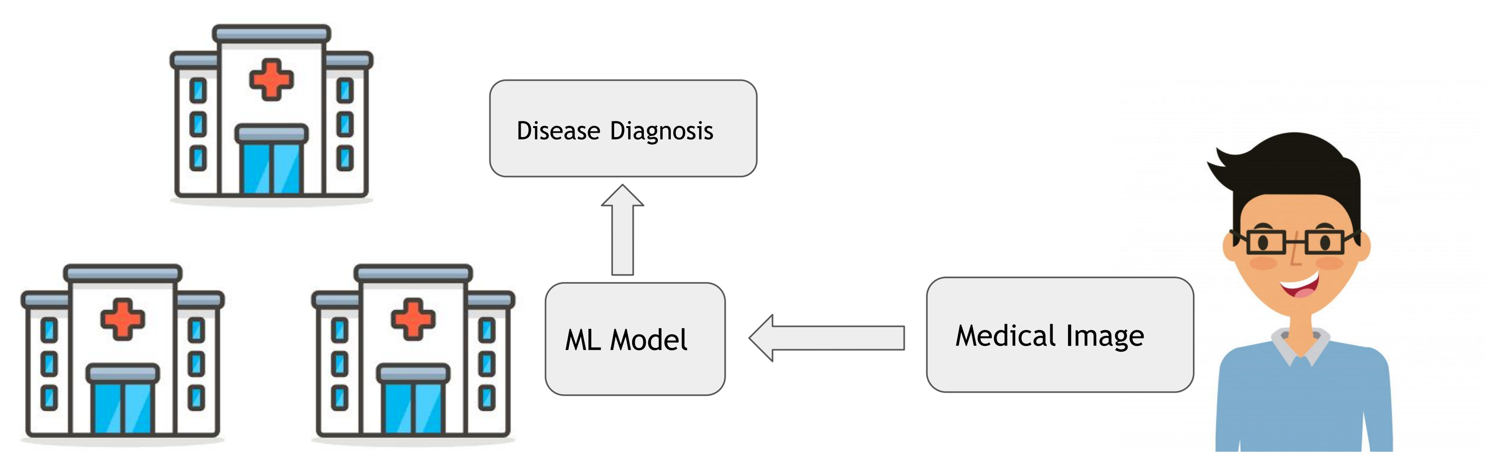
Sijun Tan, University of Virginia, st8eu@virginia.edu

Brian Knott, Facebook AI Research, brianknott@fb.com

Yuan Tian, University of Virginia, yuant@virginia.edu

David J. Wu, University of Virginia, dwu4@virginia.edu

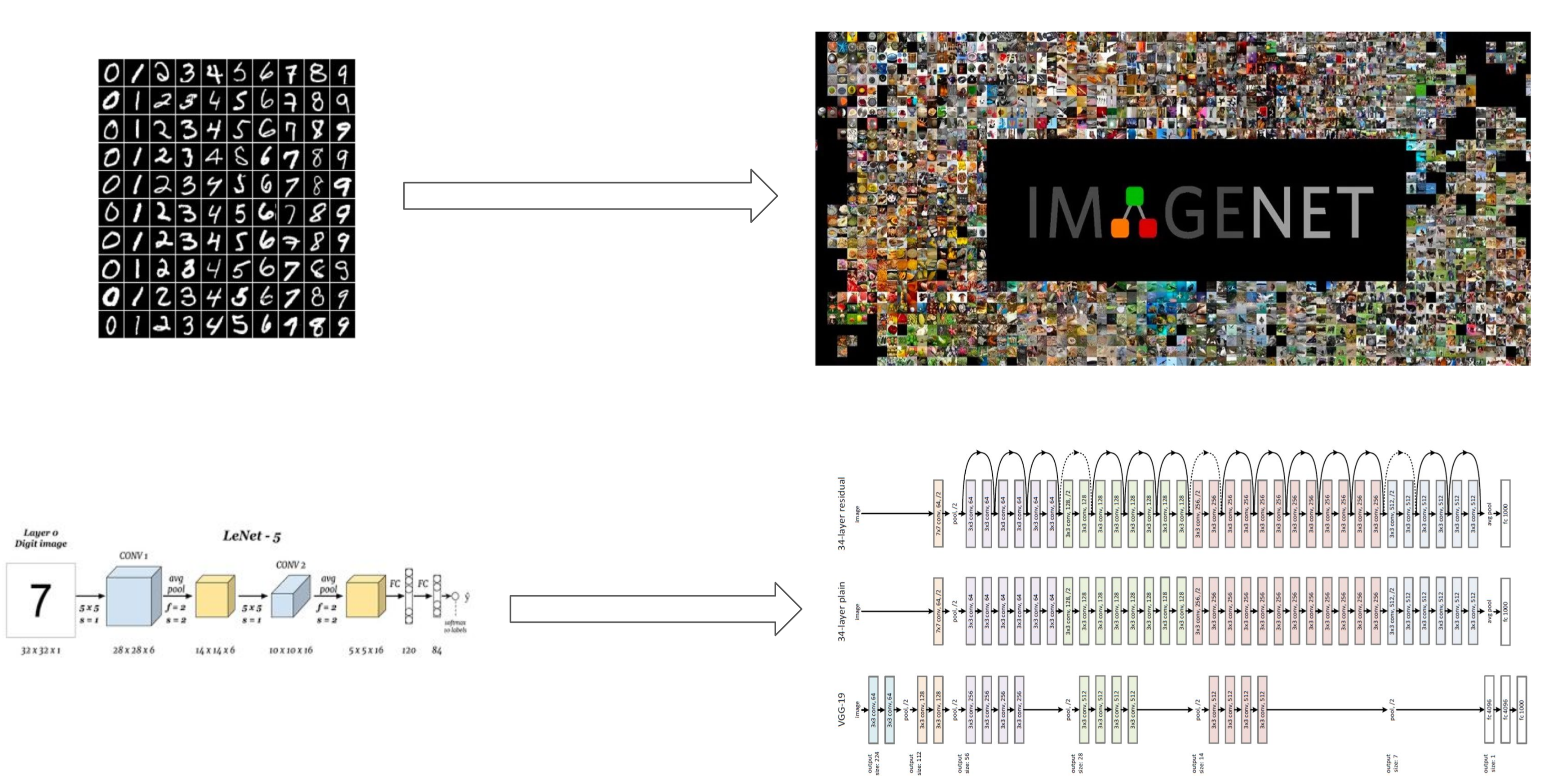
Privacy-Preserving ML



- Hospitals should not learn patient's medical data
- Patient should not learn the weights of the ML model

Can be achieved with **secure multiparty computation**

Scalability Challenge in PPML



- There is a significant performance gap between plaintext and private ML (2300x in private inference, 42000x in private training)
- Linear layers are the major performance bottleneck
- GPU acceleration is necessary for scalability

Our System and Benchmarks

A system that supports end-to-end private training/inference on GPU

- Supports private inference/training in the **3PC semi-honest setting**
- Keep all computations on the GPU
- Significantly improve performance of private inference/training

Embedding fixed-point arithmetic into floating-point CUDA kernels

$$(A_1 + A_2) \cdot (B_1 + B_2) = A_1 B_1 + A_1 B_2 + A_2 B_1 + A_2 B_2$$

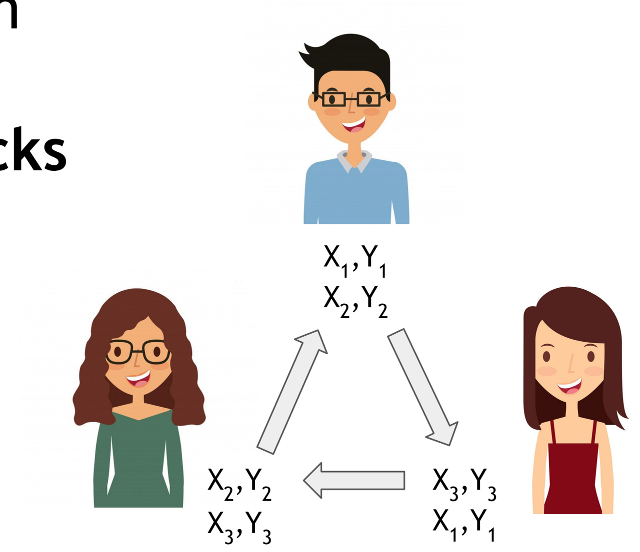
- Convert product of 64-bit integers into sums of product of 16-bit integers
- Use CUDA kernels to compute product of 16-bit integers in floating-point

GPU friendly protocol design

- Component-wise operations (e.g multiplication) are fast on GPUs
- Conditional statements are slow on GPUs
- Design protocols that better utilize parallelism

Replicated secret-sharing as basic building blocks

- A type of additive secret-sharing scheme
- Each party holds 2-out-of-3 secret shares
- Communication efficient in the 3PC setting



	ResNet-50 (ImageNet)		ResNet-101 (ImageNet)		ResNet-152 (ImageNet)	
	Time	Comm. (GB)	Time	Comm. (GB)	Time	Comm. (GB)
CrypTFlow	25.9	6.9	40*	10.5*	60*	14.5*
CryptGPU	9.31	3.08	17.62	4.64	25.77	6.56
Plaintext	0.011	—	0.021	—	0.031	—

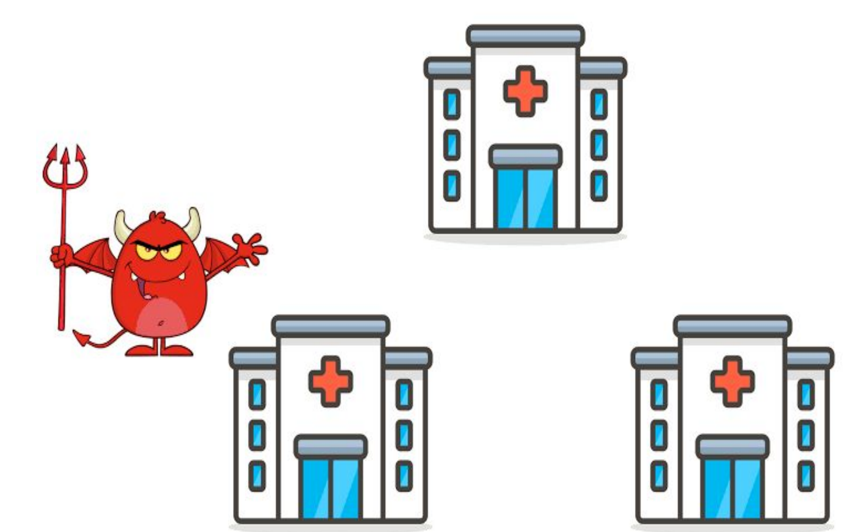
A 2.5x improvement over CrypTFlow on private inference

	LeNet (MNIST)		AlexNet (CIFAR-10)		VGG-16 (CIFAR-10)		AlexNet (TI)		VGG-16 (TI)	
	Time	Comm.	Time	Comm.	Time	Comm.	Time	Comm.	Time	Comm.
Falcon*	14.90	0.346	62.37	0.621	360.83†	1.78†	415.67	2.35	359.60‡	1.78‡
CryptGPU	2.21	1.14	2.91	1.37	12.14†	7.55†	11.30	6.98	13.89‡	7.59‡
Plaintext	0.0025	—	0.0049	—	0.0089	—	0.0099	—	0.0086	—

A 7x-36x improvement over Falcon on private training

Threat Model

3PC semi-honest security with honest-majority



- **Honest-majority:** Allowing a single semi-honest party for corruption
- **Semi-honest:** Corrupt parties follow the protocol, but try to gather information out of the protocol

Summary and Future Work

Summary

- We present the first PPML system that keep all computations on the GPU
- We demonstrate that GPU can significantly accelerate bottleneck in linear layers
- Training AlexNet on TinyImageNet previously takes **over a year**, and now it takes roughly **over a week** (~10 days)

Future Work

- Support multiple GPUs
- Design more efficient MPC protocols that leverages GPU parallelism