Federated Learning:
• Hardly any single entity’s data is sufficient for batched linear operations

Data privacy in DNNs

Candidates:
• (Efficient) Homomorphic operation:
  – Independent & Competing Cloud Server Providers

For Goten: GPU + TEE for Private Training
• Federated Learning:
  – Each data contributor train DNN locally
  – They exchange the DNN’s weight frequently

• Problems:
  – Every data contributor can use the DNN
  – No rate-limiting, even for non-asynced uses
  – Data contributors may steal others’ data
  – Model Inversion Attack [Fredrikson et al.]
  – Noisy/Implicit data ⇒ Data privacy

Preliminary: TEE & GPU
TEE: Trusted Execution Environment (e.g., SGX)
• Protect the data’s privacy inside
  – Even the machine owner cannot read it
• Processes data efficiently as plaintext on GPU
  – Too slow for batched linear operations

GPU: Graphics Processing Unit
• GPU can speed up the linear layers in DNNs
  – The most time-consuming part in DNNs
• GPU does not have TEE
  – Lack of data privacy & model privacy

Goten: GPU + TEE for Private Training
• Contributors send their data to SGX’s TEE/enclosures
• Securely outsource linear-layer computation to GPUs
  – Resided in 3 non-colluding servers (U0, U1, U2)
  – Can reduce to 2 servers at (1/5 of the throughput)
• (Non-)linear layers in SGX

Why Federated Learning is not enough?
• Federated Learning:
  – Each data contributor train DNN locally
  – They exchange the DNN’s weight frequently

• Problems:
  – Every data contributor can use the DNN
  – No rate-limiting, even for non-asynced uses
  – Data contributors may steal others’ data
  – Model Inversion Attack [Fredrikson et al.]
  – Noisy/Implicit data ⇒ Data privacy

GPU-Outsourcing Protocol for Linear Layers (Overview)

Lightweight (Light-Weight) Crypto Tool: Additive Secret Shares (SS)
• x = <x0, x1, x2, ... > (mod q)
  – <x0, x1, x2, ... > is a pair of additive SSs for x
• Privacy (n) has no information about x
  – For each value of x, given <x0, x1, x2, ... > corresponding <x0, x1, x2, ... >
  – (Efficient) Homomorphic operation:
    – x0 + x1 = y0 + y1
  – For brevity, we will omit (mod q)

Gotten: GPU-Outsourcing Protocol for Linear Layers (Details)
• Goal: Compute y = W0 x (ϕ is the linear operation)
• Without leaking any (W, x, y) to (U0, U1, U2)
  1. U0 = a + Rand(r), a = Rand(r)
  2. U1 = a + e1, e1 = x + e0
     for e0 = 0 (in parallel)
  3. E, U2 : (W0, W1, W2, W3, w1, w2, w3, x1) → (W1, W2, W3, w1, w2, w3, x2)
     e = W0 − Rand(r), f = e − Rand(r)
  4. U2 = E, U2 = a + y0, y0 = (W0 f + e x0 + e1) + e0 + e1
  5. End

Non-Colluding Servers in Goten
• Each server holds a secret-share of the model/data
• Individual share by itself is totally meaningless
• Candidates:
  – Some of the Data Contributors
  – Government: Hospital/Monetary authority
  – Independent & Competing Cloud Server Providers

Training for Invasive Ductal Carcinoma (IDC) Detection
• Showcase application involving sensitive training data
• IDC: The most common type of breast cancer
• Dataset: Images of women’s breast tissue [Ozis-Ross et al.]

Accuracy

Speedup

Time (min)

Gotten vs. Falcon’s 5 weeks (accuracy not reported)

References


Lucien K. L. Ng, Sherman S. M. Chow, Anna P. Y. Woo, Donald P. H. Wong (CUHK) → NTU

Department of Information Engineering, Chinese University of Hong Kong (CUHK), Hong Kong

Gotten: GPU-Outsourcing Trusted Execution of Neural Network Training
Lucien K. L. Ng, Sherman S. M. Chow, Anna P. Y. Woo, Donald P. H. Wong (CUHK), and Yongjun Zhao (CUHK → NTU)

Privacy of "Big" Training Data
• Sensitive
  – Medical Image analysis, Child Experiment Imagery, etc.
• Privacy laws & Regulations, e.g., GDPR
• Massive
  – Hardly any single entity’s data is sufficient
• Private Training
  – No one learns anything about the model & other’s data

Gotten: NVIDIA V100 16GB
GPU (w/ SGX) Intel I7-7700K
• Network: Google Cloud (8GBs & ~5ms latency)

Conclusion
• Best of Both Worlds: TEE & GPU
• Our Techniques:
  – Lightweight Crypto for GPU-Outsourcing
  – Dynamic Quantization for Weight Flattening during Training
• Future Work: GPU-Friendly Pure-Crypto Solutions [Ng and Chow]
• Code: github.com/goten-team/Gotten

Performance on Training

CIFAR-10: Common Benchmark for Computer Vision
• Gotten attains >95% accuracy in 34 hours
  – vs. Falcon’s 5 weeks (accuracy not reported)
• 132x throughput speed up over Falcon

Falcon (Sameer Wagh et al.): State-of-the-Art Crypto Approach

Framework | GPU | TEE | DNN Arch. | Throughput | Speedup
Falcon | X | X | VGG-16 | 1482 | 132x
Caffe2Cone | X | X | VGG-11 | 28800 | 6.84x
Gotten | U | V | 196733 | -

[*] Our pure-TEE private training framework over Caffe & SCCONE (Secure Container Environment)

Accuracy

Speedup

Time (min)

85.53%
81%
82%
83%
84%
85%
86%
13.7x
13.1x
16.9x
132x
7.28x
6.33x
3.42x
31.2
4.68

GPU-Outsourcing Protocol for Linear Layers (Details)

Security

What each non-colluding server sees:
– U0: <W0>, x0, y0, f0
– U1: <W1>, x1, y1, f1
– U2: x2, y2
– They are all secret shares or random tokens:
  – <W0 f0, y0, f0> are secret shares (by definition)
  – e = (W0 x) + f0 and f = (x v) are secret shares
  – u and v are random tokens

Correctness

y = y0 + y1 + z
= <W0 f0 > + f1 = e + y0 y1 + <W0 f0 > + e + y0 y1
= W0 x V

References

Fredrikson, Somayeh Jha, Thomas Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. CCS 15.


(lucien@ie.cuhk.edu.hk)