Split Learning

A resource efficient distributed deep learning method without sensitive data sharing

Praneeth Vepakomma
vepakom@mit.edu
‘Invisible’ Health Image Data

‘Small Data’

‘Small Data’

‘Small Data’
a. Distributed Data
b. Patient privacy
c. Incentives
d. ML Expertise
e. Efficiency

ML for Health Images

- Low Bandwidth
- ‘Small’ Data
- Low Compute
Gupta, Raskar ‘Distributed training of deep neural network over several agents’, 2017
Intelligent Computing

Security, Privacy & Safety

… and predictive models can breach privacy too

Top Perceived Advantages of Using AI for Health Care

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health care would be easier and quicker for more people to access</td>
<td>34%</td>
</tr>
<tr>
<td>Faster and more accurate diagnoses</td>
<td>31%</td>
</tr>
<tr>
<td>Will make better treatment recommendations</td>
<td>27%</td>
</tr>
<tr>
<td>Like having your own health care specialist, available any time and on any device</td>
<td>27%</td>
</tr>
</tbody>
</table>


Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Marketers Can Glean Private Data on Facebook
Regulations

**GDPR:** General Data Protection Regulation

**HIPAA:** Health Insurance Portability and Accountability Act, 1996

**SOX:** Sarbanes-Oxley Act, 2002

**PCI:** Payment Card Industry Data Security Standard, 2004

**SHIELD:** Stop Hacks and Improve Electronic Data Security Act, Jan 1 2019
NOTABLE HEALTHCARE BREACHES

- AvMed, Inc. --- 1.2M victims
- North Bronx Healthcare Network --- 1.7M victims
- The Nemours Foundation --- 1M victims
- Advocate Medical Group --- 4M victims
- Anthem, Inc. --- 80M victims
- BlueCross BlueShield of Tennessee --- 1M victims
- TRICARE Management Activity --- 4.9M victims
- Health Net, Inc. --- 1.9M victims
- Community Health Systems --- 4.5M victims
Ms. Magrin’s location data shows other often-visited locations, including the gym and Weight Watchers.

In about four months’ of data reviewed by The Times, her location was recorded over 8,600 times — on average, once every 21 minutes.

Data reviewed by The Times includes duration of schools. Here a device is, most likely, a child’s, is tracked from a home to school.

More than 40 other devices appear in the school during the day. Many are traceable to nearby homes.
Challenges for Distributed Data + AI + Health

Distributed Data
Multi-Modal
Incomplete Data

Ledgering
Smart contracts
Maintenance

Regulations
Incentives
Cooperation
Ease

Resource-constraints
Memory, Compute, Bandwidth,
Convergence, Synchronization, Leakage
Automating ML : AI building AI

Teacher

Agent Samples Network Topology

Student

Conv

Pool

Softmax

Train Network

Agent Learns From Memory

Store in Replay Memory

Topologies:
- C(64,5,1)
- C(128,3,1)
- P(2,2)
- SM(10)

Performance: 93.3%

Sample Memory

Update Q-Values

Otkrist Gupta, Baker, Naik, Raskar, ICLR 2017
Training Deep Networks

No sharing of Raw Images

Server

Client

Invisible Data / Data Friction

AI: Bringing it all together
Ease | Incentive | Trust | Regulation

Blockchain

AI/ SplitNN

Overcoming Data Friction
Anonymize  |  Obfuscate  |  Encrypt

Protect Data
Anonymity is not enough …
Federated Learning
Nets trained at Clients
Merged at Server

Differential Privacy
Obfuscate with noise
Hide unique samples

Split Learning (MIT)
Nets split over network
Trained at both

Homomorphic Encryption
Basic Math over Encrypted Data (+, x)
Federated Learning

Server

How does it work?

\[ \sum_{k=1}^{K} \frac{n_k}{n} w_t^k \]

\[ \Delta w^1 \]

\[ \Delta w^2 \]

\[ \Delta w^3 \]

\[ \Delta w^K \]
<table>
<thead>
<tr>
<th>Distributed Training</th>
<th>Protect data</th>
<th>Partial Leakage</th>
<th>Differential Privacy</th>
<th>Homomorphic Encryption</th>
<th>Oblivious Transfer, Garbled Circuits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federated Learning</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Split Learning</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>⬜</td>
<td>⬜</td>
</tr>
</tbody>
</table>

Praneeth Vepakomma, Tristan Swedish, Otkrist Gupta, Abhi Dubey, Raskar 2018
When to use split learning?

Large number of clients: Split learning shows positive results

Project Page and Papers: https://splitlearning.github.io/
## Quantitative Results

<table>
<thead>
<tr>
<th>Method</th>
<th>100 Clients</th>
<th>500 Clients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large Scale SGD</td>
<td>29.4 TFlops</td>
<td>5.89 TFlops</td>
</tr>
<tr>
<td>Federated Learning</td>
<td>29.4 TFlops</td>
<td>5.89 TFlops</td>
</tr>
<tr>
<td>Our Method (SplitNN)</td>
<td>0.1548 TFlops</td>
<td>0.03 TFlops</td>
</tr>
</tbody>
</table>

Table 1. Computation resources consumed per client when training CIFAR 10 over VGG (in teraflops)

<table>
<thead>
<tr>
<th>Method</th>
<th>100 Clients</th>
<th>500 Clients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large Scale SGD</td>
<td>13 GB</td>
<td>14 GB</td>
</tr>
<tr>
<td>Federated Learning</td>
<td>3 GB</td>
<td>2.4 GB</td>
</tr>
<tr>
<td>Our Method (SplitNN)</td>
<td>6 GB</td>
<td>1.2 GB</td>
</tr>
</tbody>
</table>

Table 2. Communication Bandwidth consumed per client when training CIFAR 100 and Resnet 50 (in gigabytes)
Label Sharing

No Label Sharing
# Distribution of parameters in AlexNet

<table>
<thead>
<tr>
<th>Layer Name</th>
<th>Tensor Size</th>
<th>Weights</th>
<th>Biases</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Image</td>
<td>227x227x3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Conv-1</td>
<td>55x55x96</td>
<td>34,848</td>
<td>96</td>
<td>34,944</td>
</tr>
<tr>
<td>MaxPool-1</td>
<td>27x27x96</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Conv-2</td>
<td>27x27x256</td>
<td>614,400</td>
<td>256</td>
<td>614,656</td>
</tr>
<tr>
<td>MaxPool-2</td>
<td>13x13x256</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Conv-3</td>
<td>13x13x384</td>
<td>884,736</td>
<td>384</td>
<td>885,120</td>
</tr>
<tr>
<td>Conv-4</td>
<td>13x13x384</td>
<td>1,327,104</td>
<td>384</td>
<td>1,327,488</td>
</tr>
<tr>
<td>Conv-5</td>
<td>13x13x256</td>
<td>884,736</td>
<td>256</td>
<td>884,992</td>
</tr>
<tr>
<td>MaxPool-3</td>
<td>6x6x256</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FC-1</td>
<td>4096x1</td>
<td>37,748,736</td>
<td>4,096</td>
<td>37,752,832</td>
</tr>
<tr>
<td>FC-2</td>
<td>4096x1</td>
<td>16,777,216</td>
<td>4,096</td>
<td>16,781,312</td>
</tr>
<tr>
<td>FC-3</td>
<td>1000x1</td>
<td>4,096,000</td>
<td>1,000</td>
<td>4,097,000</td>
</tr>
<tr>
<td>Output</td>
<td>1000x1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>62,378,344</strong></td>
</tr>
</tbody>
</table>
Versatile Configurations of Split Learning

Split learning for health: Distributed deep learning without sharing raw patient data, Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, Ramesh Raskar, (2019)
NoPeek SplitNN: Reducing Leakage in Distributed Deep Learning

\[ \alpha_1 DCOR(X_n, \hat{Z}) + \alpha_2 CCE(\hat{Y}, Y_n) \]

Reducing leakage in distributed deep learning for sensitive health data, Praneeth Vepakomma, Otkrist Gupta, Abhimanyu Dubey, Ramesh Raskar (2019)
No peak deep learning with conditioning variable

Setup:

- **Supervised**: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \subset X \times Y$
- **Output**: $y \in \mathbb{R}$
- **Goal**: To find a projection $S_{Y|X}$ such that, $Y \independent X|Z$.

**Ideal Goal**: To find such a conditioning variable $Z$ within the framework of deep learning such that the following directions are approximately satisfied:

1. $Y \independent X \mid Z$ (Utility property as $X$ can be thrown away given $Z$ to obtain prediction $E(Y|Z)$)
2. $X \independent Z$ (One-way property preventing proper reconstruction of raw data $X$ from $Z$)

*Note: \independent denotes statistical independence*
Possible measures of non-linear dependence

- COCO: Constrained Covariance
- HSIC: Hilbert-Schmidt Independence Criterion
- DCOR: Distance Correlation
- MMD: Maximum Mean Discrepancy
- KTA: Kernel Target Alignment
- MIC: Maximal Information Coefficient
- TIC: Total Information Coefficient
Why is it called distance correlation?

Definition 3.1. Sample Distance Covariance [3]: Given i.i.d samples \( \mathcal{X} \times \mathcal{Y} = \{(x_k, y_k) | k = 1, 2, 3, \ldots, n\} \) and corresponding double centered Euclidean distance matrices \( \widehat{E}_X \) and \( \widehat{E}_Y \), then the squared sample distance correlation is defined as,

\[
\hat{r}^2(X, Y) = \frac{1}{n^2} \sum_{k, l=1}^{n} [\widehat{E}_X]_{k,l} [\widehat{E}_Y]_{k,l},
\]
Distance Covariance (Székely, G. (2007))

\[
\nu^2(X, Y; w) = \int_{\mathbb{R}^{h+m}} |f_{X,Y}(t, s) - f_X(t)f_Y(s)|^2 w(t, s) dt ds
\]

where \( f_X, f_Y, f_{X,Y} \) are the characteristic functions of \( X, Y, X \times Y \) and \( w(t, s) \) is a suitably chosen weight function.

Sample Distance Covariance (2nd order)

\[
\hat{\nu}^2(X, Y) = \frac{1}{n^2} \text{Tr} \left( L_X^T L_Y \right)
\]

where \( L_X = D_X - HE_x H \) and \( L_Y = D_Y - HE_y H \).

Lemma 3.1. Given matrices of squared Euclidean distances \( E_X \) and \( E_Y \) and Laplacians \( L_X \) and \( L_Y \) formed over adjacency matrices \( \hat{E}_X \) and \( \hat{E}_Y \), the square of sample distance correlation \( \hat{\nu}^2(X, Y) \) is given by

\[
\hat{\nu}^2(X, Y) = \frac{\text{Tr} \left( X^T L_Y X \right)}{\sqrt{\text{Tr} \left( Y^T L_Y Y \right) \text{Tr} \left( X^T L_X X \right)}}.
\]
Colorectal histology image dataset (Public data)
Reduced leakage during training over colorectal histology image data from 0.96 in traditional CNN to 0.19 in NoPeek SplitNN

Reducing leakage in distributed deep learning for sensitive health data, Praneeth Vepakomma, Otkrist Gupta, Abhimanyu Dubey, Ramesh Raskar (2019)
Similar validation performance
Effect of leakage reduction on convergence

Figure 9: $\alpha_1 = 0.05$

Figure 10: $\alpha_1 = 0.15$
Robustness to reconstruction

Figure 7: $\alpha_1 = 0.1$

Figure 8: $\alpha_1 = 0.9$
Proof of one-Way Property:

\[
DCOV(X, Z) = \text{Tr}(XX^TZZ^T) + \|X - Z\| + \|Z\|
\]

\[
D_{KL}(Z|X) - D_{KL}(X|Z) = H(Z, X) - H(Z) - H(X, Z) + H(X)
\]

We show: Minimizing regularized distance covariance minimizes the difference of Kullback-Leibler divergences
\[ \det(Z^T X) - \det(Z^T Z) - \det(X^T X) + \det(X^T X) \]

This can be bounded using Hadamard’s inequality as

\[ \det(Z^T X) - \det(Z^T Z) + \det(X^T X) - \det(X^T Z) \leq \left\| Z^T X - Z^T Z \right\|_2 \left\| \frac{Z^T X}{Z^T Z} - \frac{Z^T Z}{Z^T X} \right\|_2 \]

\[ + \left\| X^T Z - X^T X \right\|_2 \left\| \frac{X^T Z}{X^T X} - \frac{X^T X}{X^T Z} \right\|_2 \]

The fractional terms \( \frac{\left\| Z^T X \right\|_2^n - \left\| Z^T X \right\|_2^n}{\left\| Z^T X \right\|_2 - \left\| Z^T Z \right\|_2} \) and \( \frac{\left\| X^T Z \right\|_2^n - \left\| X^T X \right\|_2^n}{\left\| X^T Z \right\|_2 - \left\| X^T X \right\|_2} \) can be written as a sum of geometric-series, with factors of change of \( \frac{\left\| Z^T X \right\|_2}{\left\| Z^T Z \right\|_2} \) and \( \frac{\left\| X^T Z \right\|_2}{\left\| X^T X \right\|_2} \) respectively because

\[ \frac{\left\| Z^T X \right\|_2^n - \left\| Z^T Z \right\|_2^n}{\left\| Z^T X \right\|_2 - \left\| Z^T Z \right\|_2} = \frac{1 - \left( \frac{\left\| Z^T X \right\|_2}{\left\| Z^T Z \right\|_2} \right)^n}{1 - \frac{\left\| Z^T X \right\|_2}{\left\| Z^T Z \right\|_2}} = \sum_{p=0}^{n-1} \left( \frac{\left\| Z^T X \right\|_2}{\left\| Z^T Z \right\|_2} \right)^p \left\| Z^T Z \right\|_2^{p-1} \]

Therefore these fractional terms can be minimized by minimizing \( \left\| Z^T X \right\|_2 \) and \( \left\| Z^T Z \right\|_2 \) as the sums of products of decreasing functions of norms are also decreasing. By Cauchy-Schwarz inequality \( \left\| Z^T (X - Z) \right\| \leq \left\| Z \right\| \left\| X - Z \right\| \).

Therefore the upper-bound on difference of KL-divergence can be minimized by minimizing \( \left\| Z \right\| \) and \( \left\| X - Z \right\| \) to minimize terms \( \left\| Z^T X - Z^T Z \right\|, \left\| X^T Z - X^T X \right\| \) in addition to minimizing \( \left\| Z^T Z \right\|, \left\| Z^T X \right\|_2 = Tr(Z^T X X^T Z) = \text{DCOV}(X, Z) \) to minimize terms \( \frac{\left\| Z^T X \right\|_2^n - \left\| Z^T X \right\|_2^n}{\left\| Z^T Z \right\|_2 - \left\| Z^T X \right\|_2} \), \( \frac{\left\| X^T Z \right\|_2^n - \left\| X^T X \right\|_2^n}{\left\| X^T Z \right\|_2 - \left\| X^T X \right\|_2} \).
Distributed Private Machine Learning for Computer Vision: Federated Learning, Split Learning and Beyond

Brendan McMahan, Jakub Konečný, Ramesh Raskar, Otkrist Gupta, Hassan Takabi, Praneeth Vepakomma
Thanks and acknowledgements to: Otkrist Gupta (MIT/LendBuzz), Ramesh Raskar (MIT), Jayashree Kalpathy-Cramer (Martinos/Harvard), Rajiv Gupta (MGH), Brendan McMahan (Google), Jakub Konečný (Google), Abhimanyu Dubey (MIT), Tristan Swedish (MIT), Sai Sri Sathya (S20.ai), Vitor Pamplona (MIT/EyeNetra), Rodmy Paredes Alfaro (MIT), Kevin Pho (MIT), Elsa Itambo (MIT)
THANK YOU