REDUCING LEAKAGE IN DISTRIBUTED DEEP LEARNING FOR SENSITIVE HEALTH DATA

Praneeth Vepakomma, Otkrist Gupta, Abhimanyu Dubey, Ramesh Raskar
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{vepakom,otkrist,dubeya,raskar}@mit.edu

ABSTRACT

For distributed machine learning with health data we demonstrate how minimizing distance correlation between raw data and intermediary representations (smashed data) reduces leakage of sensitive raw data patterns during client communications while maintaining model accuracy. Leakage (measured using KL Divergence between input and intermediate representation) is the risk associated with the invertibility from intermediary representations, can prevent resource poor health organizations from using distributed deep learning services. We demonstrate that our method reduces leakage in terms of distance correlation between raw data and communication payloads from an order of 0.95 to 0.19 and from 0.92 to 0.33 during training with image datasets while maintaining a similar classification accuracy.

1 INTRODUCTION

Data sharing and computation with security, privacy and safety have been identified amongst most important current trends in healthcare (Stanford, 2018; Avancha et al., 2012; Halperin et al., 2008). Hosting of multi-modal data by multiple healthcare entities that do not trust each other due to sensitivity and privacy issues poses to be a barrier for distributed machine learning. This paper proposes a way to minimize reconstruction of raw data in distributed machine learning by minimizing distance correlation measure between raw data and any intermediary communication between entities while maintaining model accuracies. Our proposed solution makes it apt to empower resource and staff constrained local health centers to collaboratively train distributed deep learning models without any raw data sharing.

1.1 RELATED WORK

Distributed deep learning methods: Split learning (Gupta & Raskar, 2018; Vepakomma et al., 2018a) is a recently developed resource efficient method for distributed deep learning by sending intermediate representations (smashed data) of split layer to another entity which completes rest of the training. Other existing distributed deep learning methods include federated learning (Konečnỳ et al., 2016; McMahan et al., 2016) and large batch synchronous stochastic gradient decent (SGD) Chen et al. (2016). Our proposed method is a significant improvement of these methods in terms of reducing leakage of raw data patterns in any communications during the training of distributed deep learning models.

Distance Correlation methods: Our method is based on minimizing a statistical measure of dependence called distance correlation introduced in Székely et al. (2007) between raw data and all intermediary communications shared between entities partaking in distributed learning while still maintaining model accuracies in the context of split learning. Distance correlation has been used in recent non-deep learning applications including causal inference Liu & Chan (2016), sure-independence screening Li et al. (2012), hypothesis testing Sejdinovic et al. (2013), supervised dimensionality reduction Vepakomma et al. (2018b) and embeddings of distributions Muandet et al. (2017). Negative log of distance correlation has been used in the context of deep learning for supervised autoencoders Wang et al. (2018) where the goal is supervised dimensionality reduction and not for preventing reconstruction of raw data as in our case.

1.2 CONTRIBUTIONS

Our main contribution is a new technique that reduces invertibility of intermediate representations (leakage) using distance correlation and we demonstrate this in the context of split learning. We do

this by ensuring the communication payloads have a low distance correlation with raw input data while still maintaining their accuracy in predicting the output labels. We show how minimizing distance covariance minimizes product of KL divergences between intermediate representations and input.

2 METHOD

We now describe the key idea of split learning as part of the background for rest of this paper and we then describe our method that improves upon split learning for reducing the leakage of patterns in distributed deep learning. In the simplest of configurations of split learning each client forward propagates a partial deep network up to a specific layer known as the split layer. The outputs at the split layer are sent to another entity (server/another client) which completes the rest of training without looking at raw data from any client that holds the raw data. The gradients are now back propagated again from its last layer until the split layer in a similar fashion. The gradients at the split layer (and only these gradients) are sent back to clients. The rest of back propagation is now completed at clients. This process is continued until the distributed split learning network is trained without looking at each others raw data. The only communication payloads in split learning are transformed versions of raw data obtained at the intermediary deep learning layer known as the split layer as described above as against to federated learning where the entire model and weights are shared and updated by all entities.

Figure 1 shows the architecture of our proposed method. The layers in the network are divided across the distributed entities based on the split layer as shown in the Figure. The loss function for the network is a combination of two losses of log distance correlation Székely et al. (2007) and categorical cross entropy used before and after split layer respectively. Distance correlation is a measure of non-linear (and linear) statistical dependence, and we reduce the log of distance correlation (DCOR) between the raw data and activations at the split layer during the training of the network. The categorical cross entropy (CCE) is optimized between predicted labels and ground-truth for classification. The total loss function for n samples of input data $\mathbf{X_n}$, estimated split layer activations $\hat{\mathbf{Z}}$, true labels $\mathbf{Y_n}$, predicted labels $\hat{\mathbf{Y}}$ and scalar weights α_1, α_2 is given by

$$\alpha_1 DCOR(\mathbf{X_n}, \hat{\mathbf{Z}}) + \alpha_2 CCE(\mathbf{Y_n}, \hat{\mathbf{Y}}) \tag{1}$$

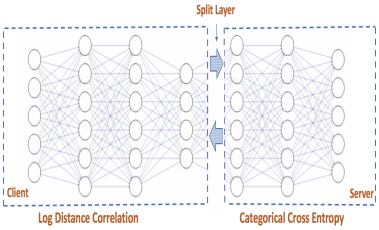

Figure 1: In our method log of distance correlation between raw input data and activations at split layer is minimized for privacy and categorical cross entropy loss between split activations and output labels is optimized for classification accuracy. The total loss is a weighted combination of these two losses.

3 CONNECTION: DISTANCE CORRELATION AND INVERTIBILITY

We use Kullback-Leibler divergence as a measure of invertibility of smashed data. In this section we derive a connection between distance covariance $DCOV(\mathbf{X}, \mathbf{Z})$ which is an unnormalized version of distance correlation and information-theoretic measures of Kullback-Leibler divergence D_{KL} and cross-entropy H. From Vepakomma et al. (2018b) we have that the sample statistic of distance covariance can be written in terms of covariance matrices $Cov(\mathbf{X})$, $Cov(\mathbf{Z})$.

$$\begin{aligned} DCOV(\mathbf{X}, \mathbf{Z}) &= Tr(\mathbf{X^T X Z^T Z}) \\ &= n^2 Tr\left(\mathbf{Cov}(\mathbf{X}).\mathbf{Cov}(\mathbf{Z})\right) \text{ (} \mathbf{X}, \mathbf{Z} \text{ are mean centered)} \end{aligned} \tag{2}$$

By arithmetic-geometric mean inequality we now have,

$$Tr\left(\mathbf{Cov}(\mathbf{X}).\mathbf{Cov}(\mathbf{Z})\right) \geq \det(\mathbf{Cov_{ZX}}) \det(\mathbf{Cov_{XZ}}) \tag{3}$$

where $\mathbf{Cov_{ZX}}$ is the cross-covariance matrix and $\det(\mathbf{Cov_{ZX}})$ is cross-entropy $H(\mathbf{X}, \mathbf{Z})$. But KL divergence is directly related to cross-entropy as

$$D_{KL}(\mathbf{X}||\mathbf{Z}) = H(\mathbf{X}, \mathbf{Z}) - H(\mathbf{X}) \tag{4}$$

Therefore combining the equations above we have the required result that minimizing distance covariance $DCOV(\mathbf{X}, \mathbf{Z})$ minimizes the product of KL divergences $D_{KL}(\mathbf{X}||\mathbf{Z})D_{KL}(\mathbf{Z}||\mathbf{X})$ with a deviation [1] of $\pm\sqrt{\frac{log(6/\delta)}{0.24N}} + \frac{C}{N}$ with probability at least $1-\delta$.

Regularizing distance covariance $DCOV(X, Z)$ with $\|\mathbf{X}-\mathbf{Z}\| + \|\mathbf{Z}\|$ gives us

$$DCOV(\mathbf{X}, \mathbf{Z}) = Tr(\mathbf{X}\mathbf{X}^\mathbf{T}\mathbf{Z}\mathbf{Z}^\mathbf{T}) + \|\mathbf{X}-\mathbf{Z}\| + \|\mathbf{Z}\| \tag{6}$$

We would like to bound the difference of KL divergences $D_{KL}(\mathbf{Z}||\mathbf{X}) - D_{KL}(\mathbf{X}||\mathbf{Z})$. Minimizing this difference can be interpreted as $\mathbf{X}$ being a good proxy dataset to construct $\mathbf{Z}$ but not as vice-versa interms of reconstructing $\mathbf{X}$ from $\mathbf{Z}$. This difference can be written in terms of cross-entropy and entropy terms as

$$D_{KL}(\mathbf{Z}||\mathbf{X}) - D_{KL}(\mathbf{X}||\mathbf{Z}) = H(\mathbf{Z}, \mathbf{X}) - H(\mathbf{Z}) - H(\mathbf{X}, \mathbf{Z}) + H(\mathbf{X}) \tag{7}$$

and this can be written in terms of determinants of covariances as

$$= \det(\mathbf{Z}^\mathbf{T}\mathbf{X}) - \det(\mathbf{Z}^\mathbf{T}\mathbf{Z}) - \det(\mathbf{X}^\mathbf{T}\mathbf{Z}) + \det(\mathbf{X}^\mathbf{T}\mathbf{X})$$

This can be bounded using Hadamard's inequality as

$$\det(\mathbf{Z}^\mathbf{T}\mathbf{X}) - \det(\mathbf{Z}^\mathbf{T}\mathbf{Z}) + \det(\mathbf{X}^\mathbf{T}\mathbf{X}) - \det(\mathbf{X}^\mathbf{T}\mathbf{Z}) \leq \left\|\mathbf{Z}^\mathbf{T}\mathbf{X} - \mathbf{Z}^\mathbf{T}\mathbf{Z}\right\|_2 \frac{\left\|\mathbf{Z}^\mathbf{T}\mathbf{X}\right\|_2^n - \left\|\mathbf{Z}^\mathbf{T}\mathbf{Z}\right\|_2^n}{\left\|\mathbf{Z}^\mathbf{T}\mathbf{X}\right\|_2 - \left\|\mathbf{Z}^\mathbf{T}\mathbf{Z}\right\|_2} + \left\|\mathbf{X}^\mathbf{T}\mathbf{Z} - \mathbf{X}^\mathbf{T}\mathbf{X}\right\|_2 \frac{\left\|\mathbf{X}^\mathbf{T}\mathbf{Z}\right\|_2^n - \left\|\mathbf{X}^\mathbf{T}\mathbf{X}\right\|_2^n}{\left\|\mathbf{X}^\mathbf{T}\mathbf{Z}\right\|_2 - \left\|\mathbf{X}^\mathbf{T}\mathbf{X}\right\|_2}$$

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

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

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

Therefore the upper-bound on difference of KL-divergence can be minimized by minimizing $\|\mathbf{Z}\|$ and $\|\mathbf{X}-\mathbf{Z}\|$ to minimize terms $\left\|\mathbf{Z}^\mathbf{T}\mathbf{X} - \mathbf{Z}^\mathbf{T}\mathbf{Z}\right\|$,$\left\|\mathbf{X}^\mathbf{T}\mathbf{Z} - \mathbf{X}^\mathbf{T}\mathbf{X}\right\|$ in addition to minimizing $\left\|\mathbf{Z}^\mathbf{T}\mathbf{Z}\right\|$, $\left\|\mathbf{Z}^\mathbf{T}\mathbf{X}\right\|_2 = Tr(\mathbf{Z}^\mathbf{T}\mathbf{X}\mathbf{X}^\mathbf{T}\mathbf{Z}) = DCOV(\mathbf{X}, \mathbf{Z})$ to minimize terms $\frac{\left\|\mathbf{Z}^\mathbf{T}\mathbf{X}\right\|_2^n - \left\|\mathbf{Z}^\mathbf{T}\mathbf{X}\right\|_2^n}{\left\|\mathbf{Z}^\mathbf{T}\mathbf{X}\right\|_2 - \left\|\mathbf{Z}^\mathbf{T}\mathbf{Z}\right\|_2}$, $\frac{\left\|\mathbf{X}^\mathbf{T}\mathbf{Z}\right\|_2^n - \left\|\mathbf{X}^\mathbf{T}\mathbf{X}\right\|_2^n}{\left\|\mathbf{X}^\mathbf{T}\mathbf{Z}\right\|_2 - \left\|\mathbf{X}^\mathbf{T}\mathbf{X}\right\|_2}$.

3.1 Similarity and affine invariance of distance correlation

Distance correlation is also invariant Székely et al. (2007) to transformations of the form $\mathbf{X} \mapsto a_1 + b_1 C_1 \mathbf{X}$ and $\mathbf{Y} \mapsto a_2 + b_2 C_2 \mathbf{Y}$ where a_1, a_2 are arbitrary vectors, b_1, b_2 are arbitrary nonzero numbers and $\mathbf{C_1}$, $\mathbf{C_2}$ are arbitrary orthogonal matrices. and also alternate versions of distance correlation that achieve affine invariance exist Dueck et al. (2014). These are highly suitable properties for computer vision given that a leakage reduction measure should be able to find a representation beyond a simple orthogonal group transformation.

[1] Sejdinovic et al. (2013) shows an equivalence between distance correlation and another popular measure of statistical dependence called Hilbert Schmidt independence criterion (HSIC) by just a constant. 2 is based on empirical estimate of DCOV which comes with an Hoeffding bound around its true estimate, Gretton et al. (2005) as

$$|DCOV(p_{xy}, \mathcal{F}, \mathcal{G}) - DCOV(Z, \mathcal{F}, \mathcal{G})| \lessapprox \sqrt{\frac{log(6/\delta)}{0.24n}} + \frac{C}{n} \tag{5}$$

with probability at least $1-\delta$.

4 EXPERIMENTS

In this section we share our experimental results with our proposed method of NoPeekNN which is an improvement over the Vanila SplitNN (split learning) method in terms of leakage reduced via distance correlation. We run experiments with a dataset of colorectal histology images without any data augmentation. In Figure 2 we share some example images from this dataset along with corresponding class labels. In Figure 3, we show that our technique NoPeekNN converges to a similar validation accuracy of 0.69 as the VanillaSplitNN 4. In Figure 5 we show the reduction in distance correlation between smashed data and raw data with respect to increasing epochs by NoPeekNN over the colorectal histology image dataset. We show that this leakage distance correlation is drastically reduced from 0.92 in Vanilla SplitNN to 0.33 in NoPeekNN. We also show that even in the first few epochs the leakage distance correlation has been minimized to less than 0.46 thereby not allowing any leakage during the training. In Figure 6 we perform the same experiment over the MNIST handwritten recoginition dataset and again show a drastic reduction from a distance correlation of 0.95 in traditional convolutional networks (CNN) and Vanilla SplitNN to about 0.19 in NoPeekNN. We also observe that even in the first few epochs the leakage distance correlation is contained below 0.34. In Figures 7, 8 we show via a NoPeekNN architecture for autoencoders that the privatized split layer prevents reconstruction of raw images by decoder layers. The layer in between the encoder and decoder layers is made the split layer. This shows that the proposed NoPeekNN is able to block the flow of critical information required for reconstriction of raw data in this experiment with increasing levels of α_1 in our loss function as desired. As a baseline, this same architecture upon removing the split layer was in comparison reasonably able to reconstruct the images. In Figures 10, 11 we show the convergence plots of validation accuracy for NoPeekNN on MNIST data with increasing levels of α_1. As expected in the convergence plots for colorectal image dataset as well as MNIST dataset we observe that we drastically reduce leakage while maintaining high accuracy levels at the cost of requiring to run for more epochs.

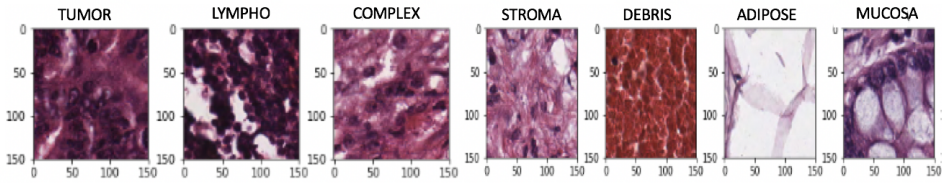

Figure 2: Some example classes from images of colorectal histology dataset. This dataset was used in our experiments to measure i) reduction in distance correlation between raw data and smashed data and ii) preservation of model accuracy

5 CONCLUSIONS

In this paper we show how to minimize the distance correlation between the smashed data and raw input while reducing classificational cross entropy. We experimentaly demonstrate how our tech-

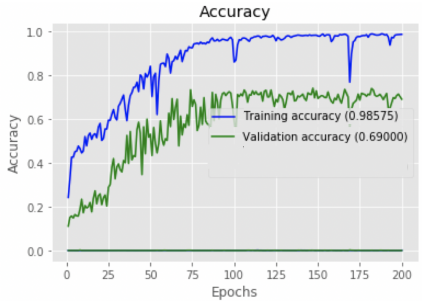

Figure 3: Convergence of validation accuracy in NoPeekNN over colorectal histology image data

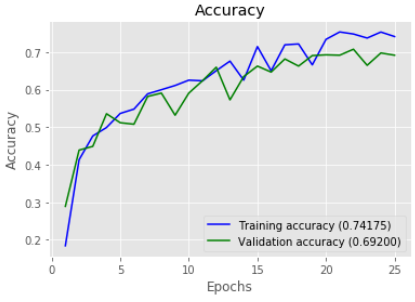

Figure 4: Convergence of validation accuracy in Vanilla SplitNN of colorectal histology image data

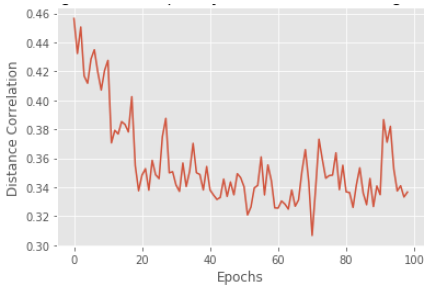

Figure 5: Reduced leakage during training over colorectal histology image data from 0.92 in traditional CNN and Vanilla SplitNN to 0.33 in NoPeekNN

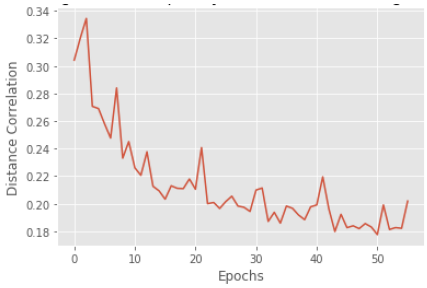

Figure 6: Reduced leakage during training over MNIST handwritten digit image data from 0.95 in traditional CNN and Vanilla SplitNN to 0.19 in NoPeekNN

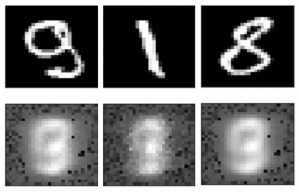

Figure 7: $\alpha_1 = 0.1$

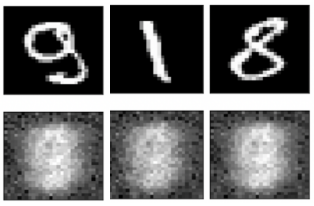

Figure 8: $\alpha_1 = 0.9$

Figure 9: *
In Figures 7,8 reconstruction results from NoPeekNN experiment for autoencoder shows that the privatized split layer prevents reconstruction of raw MNIST data with increasing levels of α_1.

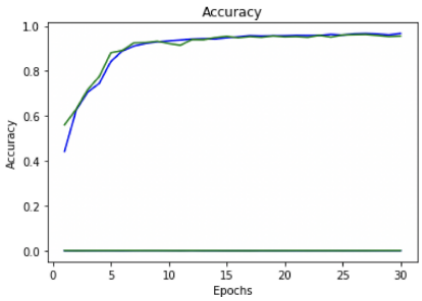

Figure 10: $\alpha_1 = 0.05$

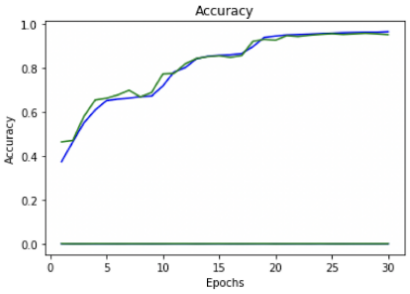

Figure 11: $\alpha_1 = 0.15$

Figure 12: *
Epochs Vs. accuracy plots on MNIST show that it takes a larger number of epochs for larger values of α_1 to reach a higher accuracy as expected in NoPeekNN.

nique can both reduce the leakage (distance correlations) and achieve accuracy when we implement it in the context of split learning applied over health datasets. We hope our method can pave way for remote communities to pool together health data during emerging threats like epidemics or slow moving threats like obesity or diabetes.

REFERENCES

Sasikanth Avancha, Amit Baxi, and David Kotz. Privacy in mobile technology for personal healthcare. *ACM Computing Surveys (CSUR)*, 45(1):3, 2012.

Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. Revisiting distributed synchronous sgd. *arXiv preprint arXiv:1604.00981*, 2016.

Johannes Dueck, Dominic Edelmann, Tilmann Gneiting, Donald Richards, et al. The affinely invariant distance correlation. *Bernoulli*, 20(4):2305–2330, 2014.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications*, 116:1–8, 2018.

Daniel Halperin, Thomas S Heydt-Benjamin, Kevin Fu, Tadayoshi Kohno, and William H Maisel. Security and privacy for implantable medical devices. *IEEE pervasive computing*, 7(1):30–39, 2008.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Runze Li, Wei Zhong, and Liping Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012.

Furui Liu and Laiwan Chan. Causal inference on discrete data via estimating distance correlations. *Neural computation*, 28(5):801–814, 2016.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, et al. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013.

Stanford. The democratization of health care. In *Stanford Medicine 2018 Health Trends Report*, 2018.

Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564*, 2018a.

Praneeth Vepakomma, Chetan Tonde, Ahmed Elgammal, et al. Supervised dimensionality reduction via distance correlation maximization. *Electronic Journal of Statistics*, 12(1):960–984, 2018b.

Rick Wang, Amir-Hossein Karimi, and Ali Ghodsi. Distance correlation autoencoder. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2018.