Distributed Semi-Private Image Classification Based on Information-Bottleneck Principle

Shideh Rezaeifar, Maurits Diephuis, Behrooz Razeghi, Denis Ullmann, Olga Taran, Slava Voloshynovskiy
Stochastic Information Processing Group
Department of Computer Science, University of Geneva, Switzerland
{shideh.rezaeifar, maurits.miephuis, behrooz.razeghi, denis.ullmann, olga.taran, svolos}@unige.ch

***Abstract*—In this paper, we propose a framework for semi-privacy-preserving image classification. It allows each user to train a model on her/his own particular data class, after which the output features are shared centrally. The model parameters are never shared. Individual users each use an auto-encoder to empirically ascertain their private data distribution. The resulting features are sufficiently discriminative between the private datasets. A central server aggregates all labeled output features together with a subset of the private data into a final classifier over all classes from all users. The latter forms a trade-off between privacy and classification performance. We demonstrate the viability of this scheme empirically and showcase the privacy performance compromise.**

***Index Terms*—privacy, information bottleneck, image classification, semi-private model.**

I. Introduction

The last decade has seen tremendous progress in (deep) neural networks in a vast amount of application domains. Much progress has been made possible thanks to cheap computing power and the widely easily available amount of datasets needed for training. This has incentivised organizations to gather more (user) data in the hope that insight might be gained from it. However, most of the data captured from users are highly sensitive, and collecting or storing this data raises many privacy-related issues. The General Data Protection Regulation, a new EU initiative on data protection and privacy, clearly prohibits companies from storing user data for long periods. Moreover, sharing user data in many applications, especially medical or financial, is strictly regulated by law. Obviously, user data is a potentially very rich source of information, which will benefit many applications. Therefore, there is a huge need for privacy-preserving machine learning.

There is significant prior research on statistical and information theoretical privacy-preserving schemes. Abundant classical well-known statistical formulations, such as k-anonymity [1], ℓ-diversity [2], t-closeness [3], differential privacy [4], and Pufferfish [5], were proposed. Information theoretic (IT) privacy approaches [6]–[16], model and analyze privacy-utility trade-offs using the IT metrics to provide asymptomatic or non-asymptotic privacy-utility-guaranteed frameworks. Recent research [17]–[19] trends adopted the Information Bottleneck (IB) problem [20] and generative adversarial networks (GANs) [21] to address information-theoretic trade-offs and address potentially new data-driven frameworks for privacy-assuring data release mechanisms.

This research is supported by SNF project No 200021-182063.

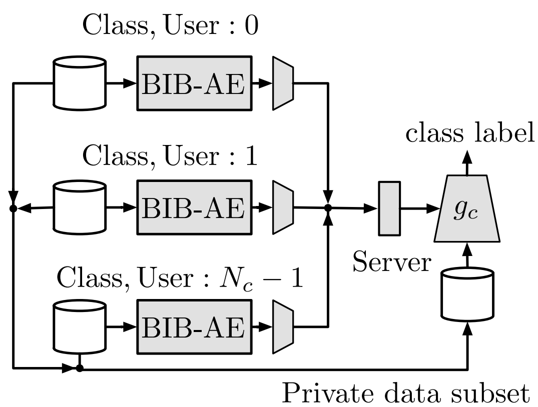

Fig. 1: Proposed client-server training architecture. BIB-AE stands for the bounded information bottleneck autoencoder proposed in [26].

There have been numerous studies on training deep neural networks models, while protecting the privacy of the original data. In this article, we investigate an image classification problem, in which there are two parties. The data owner/user/client, who holds a collection of images and the server, which would like to classify a query. In a classical approach, the model is simply trained on all available data, and significant performance loss incurs as the amount of said training data diminishes.

In the so called federated learning scheme [22], a shared model is trained individually by all users (owners) without sharing their private data. Each user trains his/her own model parameters based on his/her private data, but passes the model parameters to a central server, where all parameters are aggregated into a final model. Federated learning suffers from multiple issues. Firstly, a single model is trained end-to-end and is not easily expendable to new classes. In addition to that, since the learned parameters are shared, it is still vulnerable to adversarial attacks targeting to obtain sensitive information [23]–[25].

This article proposes a training setup where part of the work is done locally by users on their nodes without sharing their datasets or model parameters. The only output features of each node are centrally aggregated into the final global classifier.

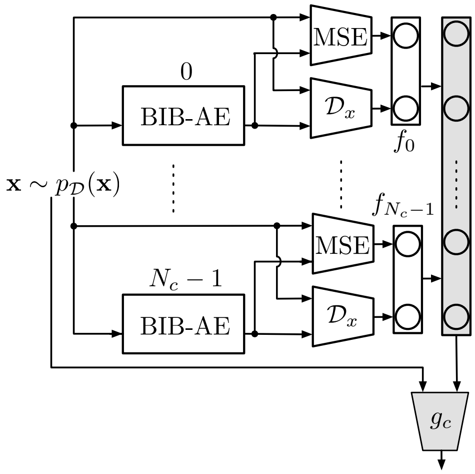

Fig. 2: Proposed classification architecture.

The proposed scheme differs from the federated learning setting in the following manners:

- Each user has the data for at most one class.
- There is no possibility of multi-session communication between the server and the users.
- No gradient or weight sharing occurs between the nodes.

Therefore, the proposed architecture has a number of advantages. As users only share the feature outputs from their models and not the parameters even less information is left exposed to possible attacks. Secondly, there are no issues stemming from gradient propagation and multi-session communication. Thirdly, it is relatively easy to add new classes/users.

Source code for the experiments will be made publicly available after review.

II. Proposed Architecture

As shown in Fig. 1 the proposed scheme consists of two entities, distributed data owners/clients and a single centralized server. Each client trains its own local model on an arbitrary assigned class and sends only the output features to the central hub. The server aggregates and concatenates all the received features and trains a final global classifier based upon those and a complementary subset of the original private data. The latter is a trade-off between privacy and classifier performance in the final model. Each user has access to the data of one single class only. All the users deploy an identical model architecture, but the model parameters are learned individually per user, per class.

Users deploy the so called Bounded Information Bottleneck Autoencoder (BIB-AE) [26] as their model to extract discriminative features. Normally, when training a classifier, one would need to see data from all classes in order to train a classifier. Contrarily, the BIB-AE allows us to approximate the distribution of data pertaining to a single class. Somewhat similar to anomaly detection systems, one can expect that autoencoders trained exclusively on a single class (inlier class), will exhibit both high reconstruction errors and low

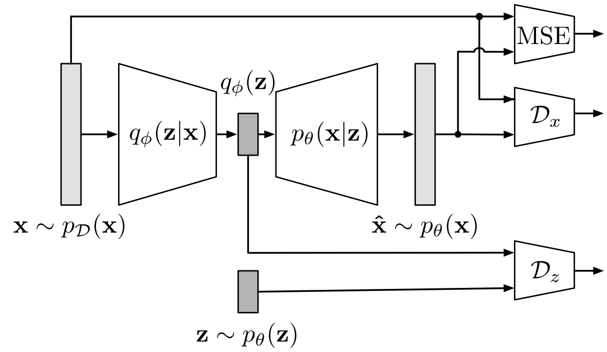

Fig. 3: Diagram of the Bounded Information Bottleneck Autoencoder (BIB-AE).

discriminative scores when tested on other unseen data classes (outlier classes). The BIB-AE is detailed in Section II-A.

The server classifier g_c is responsible for aggregating the final features (Section II-A) from each user. The complete architecture of the scheme is sketched in Fig. 2. The server trains the classifier g_c based on all the concatenated per-class features from all individual users, together with a subset of the original data. The latter is an obvious breach of privacy by design, as adding more side-information enhances performance. However, this is a controllable leak.

A. *Bounded Information Bottleneck Auto-encoder*

We will consider a true data distribution $p_D(\mathbf{x})$ from which the training set $\{\mathbf{x}_1, \mathbf{x}_2, ..\mathbf{x}_m\}$ is sampled from. The Information Bottleneck (IBN) auto-encoder [26] can be considered as a "compression" of $\mathbf{x}$ to $\mathbf{z}$ via a parametrized mapping $q_\phi(\mathbf{z}|\mathbf{x})$ leading to a bottleneck representation $\mathbf{z}$ yet preserving a certain level of information $I_\mathbf{x}$ in $\mathbf{z}$ about $\mathbf{x}$. Accordingly, the unsupervised IBN problem can be formulated as:

$$\min_{\phi: I(\mathbf{z};\mathbf{x}) \geq I_\mathbf{x}} I_\phi(\mathbf{x};\mathbf{z}), \tag{1}$$

where $I(\cdot;\cdot)$ denotes the mutual information [27]. It can be also written in the Lagrangian formulation as a minimization of:

$$\mathcal{L}(\theta, \phi) = I_\phi(\mathbf{x};\mathbf{z}) - \beta I(\mathbf{z};\mathbf{x}), \tag{2}$$

with β to be a Lagrangian multiplier.

The first term $I_\phi(\mathbf{x};\mathbf{z})$ in (2) can be decomposed as:

$$\begin{aligned} I_\phi(\mathbf{x};\mathbf{z}) &= \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z},\mathbf{x})}{q_\phi(\mathbf{z}) p_D(\mathbf{x})}\right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z})} \frac{p_\theta(\mathbf{z})}{p_\theta(\mathbf{z})}\right] \\ &= \mathbb{E}_{p_D(\mathbf{x})}\left[KL(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}))\right] - \\ &\quad KL(q_\phi(\mathbf{z}) || p_\theta(\mathbf{z})). \end{aligned} \tag{3}$$

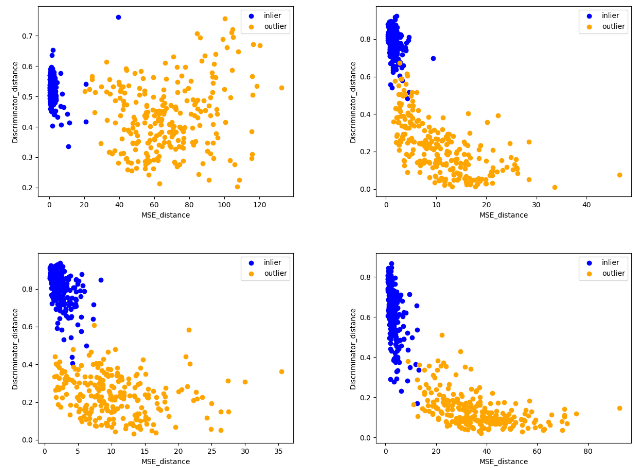

Fig. 4: Outlier (orange) versus inlier (blue) for different classes of MNIST, with a latent space dimensionality of 128.

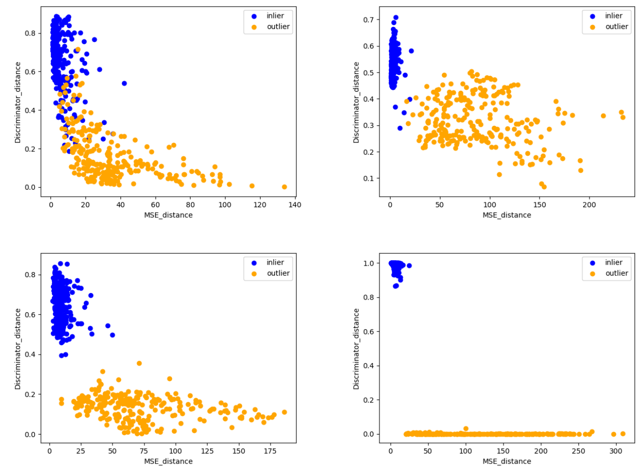

Fig. 5: Outlier (orange) versus inlier (blue) for different classes of Fashion-MNIST, with a latent space dimensionality of 128.

Similarly, the second term can be also formulated as:

$$\begin{aligned}
I(\mathbf{z};\mathbf{x}) &= \mathbb{E}_{p(\mathbf{z},\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{z})}{p_D(\mathbf{x})}\right] \\
&= -\mathbb{E}_{p_D(\mathbf{x})}\left[\log p(\mathbf{x})\right] - \mathbb{E}_{p_D(\mathbf{x})}\left[\log \frac{p_D(\mathbf{x})}{p_\theta(\mathbf{x})}\right] \\
&\quad + \mathbb{E}_{p_D(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] \\
&= H(p_D(\mathbf{x});p_\theta(\mathbf{x})) - KL(p_D(\mathbf{x})||p_\theta(\mathbf{x})) \\
&\quad + \mathbb{E}_{p_D(\mathbf{x})}\left[E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] \\
&\geq \mathbb{E}_{p_D(\mathbf{x})}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - KL(p_D(\mathbf{x})||p_\theta(\mathbf{x})) \\
&= I^U_{\theta,\phi}(\mathbf{z};\mathbf{x}). \qquad (4)
\end{aligned}$$

Given the principles of information bottleneck, a new auto-encoding framework was introduced in [26] as a bounded information bottleneck AE (BIB-AE). The BIB-AE Lagrangian is defined as:

$$\mathcal{L}_{\mathrm{BIB-AE}}(\theta,\phi) = I_\phi(\mathbf{x};\mathbf{z}) - \beta I^U_{\theta,\phi}(\mathbf{z};\mathbf{x}), \qquad (5)$$

where:

$$\begin{aligned}
I_\phi(\mathbf{x};\mathbf{z}) &= \underbrace{E_{p_D(\mathbf{x})}[KL(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))]}_{\text{A}} \\
&\quad - \underbrace{KL(q_\phi(\mathbf{z})||p_\theta(\mathbf{z}))}_{\text{B}}, \\
I^U_{\theta,\phi}(\mathbf{z};\mathbf{x}) &= \underbrace{E_{p_D(\mathbf{x})}[E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{C}} \\
&\quad - \underbrace{KL(p_D(\mathbf{x})||p_\theta(\mathbf{x}))}_{\text{D}}. \qquad (6)
\end{aligned}$$

The diagram explaining the BIB-AE setup is shown in Fig. 3. The reconstruction fidelity is ensured jointly by the terms (C) and (D), while the minimization of mutual information between $\mathbf{x}$ and $\mathbf{z}$ is guided by the targeted distribution of the latent space $p_\theta(\mathbf{z})$ according to the terms (A) and (B). After training the BIB-AE, each user would give the following features as the outputs:

- *MSE score*: Euclidean distance between the input probe image and the reconstructed one.
- *discriminator score*: output of the discriminator for the reconstructed image.

With a proper training of the model, we expect the images from the inlier class to have a low reconstruction error and high discriminator score in contrast to the outlier classes. To validate our hypothesis, examples of these scores for inlier and outlier classes are shown in Fig. 4 for MNIST database [28] and in Fig.5 for Fashion-MNIST database [29].

B. Global Classifier

The final server side classifier takes the aggregated features from all user models, next to a subset of original data. The original image data are passed through two convolutional layers outputting 8 channels. These intermediate features are stacked with the user features and fed through a final linear layer, followed by soft-max.

III. Experiments

In this section, we evaluate the proposed semi-private classification algorithm. The performance of the proposed model is assessed in terms of the classification accuracy over three public datasets, the MNIST and Fashion-MNIST. The description of the datasets are given below:

- The **MNIST** dataset contains 70000 28×28 grayscale images of handwritten digits from 0 to 9 [28].
- The **Fashion-MNIST** dataset contains 70000 28×28 gray scale images of fashion and clothing items that each sample associates with a label from 10 classes. It was created by Zalando as a compatible replacement for MNIST [29].

Each BIB-AE is trained using a subset of N_{AE} private data and the final classifier at the server is trained using only $N_{\mathrm{classifier}}$. labeled samples of the private data from all users.

The trade off between privacy and the classification accuracy comes with using different values of $N_{\text{classifier}}$.

We define privacy leakage as the percentage of images from the original dataset that is used in the training of the final global classifier.

$$L = \frac{N_{\text{classifier}}}{N_{\text{total}}}. \tag{7}$$

From another point of view, this scheme can be considered as a particular case of semi-supervised learning; nevertheless, one has access to all the dataset with a limited number of labels in semi-supervised learning. In our architecture, however, the total number of data together with their label is:

$$N_{\text{label}} = \min(N_{\text{classifier}}, N_c \times N_{\text{AE}}). \tag{8}$$

A. Discriminative Features

The BIB-AE blocks have two outputs, which are used as feature scores. The first is the reconstruction (MSE) loss, the second is the discriminator output. Before concatenation all are scaled between 0 and 1.

The outlier versus inlier test was setup as follows. For all individual classes, a single one is selected as inlier and tested against all others as outliers. Fig. 4 and Fig. 5 show the results from a couple of single classes tested against the other classes(outlier classes), for MNIST and Fashion-MINST, respectively. As expected some inlier classes overlap more than others. To quantify this we use two metrics: the Class Scatter Measure (CSM) [30] and the Davis Bouldin (DB) metric [31].

The Class Scatter Measure (CSM) index is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters (where dispersion is defined as the sum of distances squared) and is higher when clusters are dense and well separated. For a dataset of size n_E which has been clustered into k clusters, the CSM index is defined as the ratio of the between-clusters dispersion mean and the within-cluster dispersion:

$$s = \frac{\mathrm{trace}(B_k)}{\mathrm{trace}(W_k)} \times \frac{n_E - k}{k-1}, \tag{9}$$

where $\mathrm{trace}(B_k)$ is the trace of the between group dispersion matrix and $\mathrm{trace}(W_k)$is the trace of the within-cluster dispersion matrix defined by:

$$W_k = \sum_{q=1}^{k} \sum_{\mathbf{x} \in C_q} (\mathbf{x} - \boldsymbol{\mu}_q)(\mathbf{x} - \boldsymbol{\mu}_q)^T, \tag{10}$$

$$B_k = \sum_{q=1}^{k} n_q (\boldsymbol{\mu}_q - \boldsymbol{\mu})(\boldsymbol{\mu}_q - \boldsymbol{\mu})^T, \tag{11}$$

with C_q the set of points in cluster q, $\boldsymbol{\mu}_q$ the center of cluster q , $\boldsymbol{\mu}$ the global center, and n_q the number of points in cluster q .

The Davis Bouldin (DB) score is the average similarity score of each cluster with its most similar cluster. The similarity is defined as the ratio of within-cluster distances to between-cluster distances. The lower DB score corresponds to better separability [31].

class	0	1	2	3	4	5	6	7	8	9
CSM	283	246	202	153	226	157	180	**328**	152	159
DB	14.3	51.2	25.4	9.6	**9.1**	22.6	11.0	39.3	15.9	13.3

TABLE I: CSM and Davis Bouldin separability scores for inlier vs outlier classes for MNIST.

class	0	1	2	3	4	5	6	7	8	9
CSM	346	**1063**	689	664	451	851	345	782	607	827
DB	0.8	**0.4**	0.6	0.5	0.7	0.5	0.8	0.5	0.6	0.4

TABLE II: CSM and Davis Bouldin separability scores for inlier vs outlier classes for Fashion-MNIST.

B. Classification Accuracy

The classification accuracy of the proposed architecture is evaluated for different values of N_{AE} and $N_{\text{classifier}}$ in Table III and IV. Supplying the central classifier with an additional 100 images per class, ensures competitive performance.

Note that in a semi-supervised settings it is custom to use all images and the limited number of labels. In contrast, our scheme uses a very limited subset of labeled images.

As mentioned earlier, the setting is entirely different from the federated learning and thus can not be compared with.

		$N_{\text{classifier}}$		
		1000	3000	6000
N_{AE}	100	98.13	98.78	99.28
	1000	98.31	98.9	99.33

TABLE III: Classification accuracy for the MNIST dataset.

		$N_{\text{classifier}}$		
		1000	3000	6000
N_{AE}	100	85.7	88.97	90.66
	1000	87.7	89	90.7

TABLE IV: Classification accuracy for the Fashion-MNIST dataset.

		Labels	
		1000	3000
Methods	VAE (M1+M2) [32]	97.6	97.82
	ladder Network [33]	99.1	-
	SS_{CNN} [34]	94.5	-
	Adversarial autoencoder [35]	98.4	-
	VAT [36]	98.6	**98.75**
	Proposed model	98.13	98.79

TABLE V: MNIST semi-supervised classification accuracy for SOTA methods.

		Labels	
		1000	3000
Methods	GMVAE + auxiliary tasks [37]	84.15	-
	SS_{CNN} [34]	83.6	-
	Proposed model	**85.7**	**88.97**

TABLE VI: Fashion-MNIST semi-supervised classification accuracy for SOTA methods.

IV. Conclusion

We have proposed a distributed semi-private image classification scheme. In particular it allows each user to train an auto-encoder model on his/her private data subsets and only share the resulting output features at the classification stage. A server aggregates all these intermediate results into a final classifier. Our experiments showed that for (Fashion) MNIST the auto-encoder features per class are discriminative enough to build a central classifier on top. The latter reaches competitive performance compared to SOTA when using only 100 additional original images per class. Future work will focus on eliminating the need for additional images at server-side completely using synthetically generated images. Moreover, the experiments should be extended to other datasets.

References

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. IEEE, 2006, pp. 24–24.

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 106–115.

[4] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.

[5] D. Kifer and A. Machanavajjhala, "A rigorous and customizable framework for privacy," in *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*. ACM, 2012, pp. 77–88.

[6] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2012, pp. 1401–1408.

[7] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[8] F. P. Calmon, M. Varia, M. Médard, M. M. Christiansen, K. R. Duffy, and S. Tessaro, "Bounds on inference," in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2013, pp. 567–574.

[9] A. Makhdoumi and N. Fawaz, "Privacy-utility tradeoff under statistical uncertainty," in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2013, pp. 1627–1634.

[10] S. Asoodeh, F. Alajaji, and T. Linder, "Notes on information-theoretic privacy," in *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, pp. 1272–1278.

[11] F. P. Calmon, A. Makhdoumi, and M. Médard, "Fundamental limits of perfect privacy," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1796–1800.

[12] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "Managing your private and public data: Bringing down inference attacks against your privacy," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1240–1255, 2015.

[13] Y. O. Basciftci, Y. Wang, and P. Ishwar, "On privacy-utility tradeoffs for constrained data release mechanisms," in *2016 Information Theory and Applications Workshop (ITA)*. IEEE, pp. 1–6.

[14] K. Kalantari, L. Sankar, and O. Kosut, "On information-theoretic privacy with general distortion cost functions," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2865–2869.

[15] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1512–1534, 2018.

[16] H. Hsu, S. Asoodeh, F. du Pin Calmon, and N. Fawaz, "Information-theoretic privacy watchdogs," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019.

[17] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Context-aware generative adversarial privacy," *Entropy*, vol. 19, no. 12, p. 656, 2017.

[18] A. Tripathy, Y. Wang, and P. Ishwar, "Privacy-preserving adversarial networks," *arXiv preprint arXiv:1712.07008*, 2017.

[19] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Generative adversarial privacy," *arXiv preprint arXiv:1807.05306*, 2018.

[20] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[22] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016. [Online]. Available: https://arxiv.org/abs/1610.05492

[23] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *NDSS*, 2018.

[24] B. Hitaj, G. Ateniese, and F. Pérez-Cruz, "Deep models under the gan: Information leakage from collaborative deep learning," *ArXiv*, vol. abs/1702.07464, 2017.

[25] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *ArXiv*, vol. abs/1704.01155, 2018.

[26] S. Voloshynovskiy, M. Kondah, S. Rezaeifar, O. Taran, T. Hotolyak, and D. J. Rezende, "Information bottleneck through variational glasses," in *NeurIPS Workshop on Bayesian Deep Learning*, Vancouver, Canada, December 2019.

[27] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[28] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," 2010.

[29] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[30] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[31] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[32] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[33] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems*, 2015, pp. 3546–3554.

[34] F. Berkhahn, R. Keys, W. Ouertani, N. Shetty, and D. GeiÃler, "One model to rule them all," 08 2019.

[35] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[36] T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii, "Distributional smoothing with virtual adversarial training," *arXiv preprint arXiv:1507.00677*, 2015.

[37] J. A. Figueroa, "Semi-supervised learning using deep generative models and auxiliary tasks," in *NeurIPS Workshop on Bayesian Deep Learning*, December 2019.