Enhancing Privacy in Federated Learning via Early Exit

Invited Paper: Enhancing Privacy in Federated Learning via Early Exit

Yashuo Wu
Dept. of EECS, University of California, Irvine
Irvine, CA, USA
yashuow@uci.edu

Carla Fabiana Chiasserini
Politecnico di Torino
Torino, Italy
carla.chiasserini@polito.it

Francesco Malandrino
CNR-IEIIT
Torino, Italy
francesco.malandrino@cnr.it

Marco Levorato
Dept. of ICS, University of California, Irvine
Irvine, CA, USA
levorato@uci.edu

ABSTRACT

In this paper, we investigate the interplay between early exit mechanisms in deep neural networks and privacy preservation in the context of federated learning. Our primary objective is to assess how early exits impact privacy during the learning and inference phases. Through experiments, we demonstrate that models equipped with early exits perceivably boost privacy against membership inference attacks. Our findings suggest that the inclusion of early exits in neural models can serve as a valuable tool in mitigating privacy risks while, at the same time, retaining their original advantages of fast inference.

CCS CONCEPTS

• **Security and privacy → Distributed systems security**.

KEYWORDS

Federated learning, Early exit, Neural networks, Membership inference attacks, Deep learning

ACM Reference Format:
Yashuo Wu, Carla Fabiana Chiasserini, Francesco Malandrino, and Marco Levorato. 2023. Invited Paper: Enhancing Privacy in Federated Learning via Early Exit. In *Proceedings of Advanced tools, programming languages, and PLatforms for Implementing and Evaluating algorithms for Distributed systems (ApPLIED).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Deep neural networks (DNN) have achieved remarkable performance on various tasks. However, their training often necessitates large datasets and considerable computing and energy resources. In this context, federated learning (FL) [13] has emerged as a promising paradigm to enable the collaborative training of machine learning models while preserving user privacy. In FL, the models are trained locally on individual nodes and only the aggregated updates are shared with the server, fso the sharing of raw datasets is not required. Specifically, the models are trained collaboratively using the gradients from each device while keeping the training data decentralized [9]. While FL thus enhances privacy, it is still prone to attacks [7].

Membership inference attacks (MIA), first introduced by [15], are a class of privacy attacks that target general DNNs. In MIA, an adversary aims to determine if the record was in the model's training dataset, which may contain sensitive data such as human faces or medical records. Thus, it is critical to prevent such attacks. Existing studies have addressed MIA by considering black-box attacks, where the adversary only has access to the target model's outputs, and white-box attacks, where the adversary has full knowledge of the target model's architecture and parameters [5]. In this paper, we focus on black-box attacks that have direct applicability in real-life scenarios and pose a greater threat if successfully executed with limited knowledge, as described in [6], compared to white-box settings. In the black-box setting, the adversary can only receive the probabilistic output of the model after submitting a data sample to a target model. Many strategies have been proposed to perform MIAs, such as threshold attacks [15] and logistic regression attacks [14]. The former use a threshold, e.g., the confidence score produced by the target model to determine the membership of a specific data sample, whereas the latter ones train a binary classifier using the target model's confidence scores as input features. In this context, [15] trains shallow models to attack the membership of a given dataset, and [14] generalizes such an attack to frameworks in which the training of a shallow model is not needed.

In our work, we study the impact of Early Exit (EE) mechanisms in both centralized and FL settings and investigate their ability to improve the resilience of the training process and inference against MIA. Within the general context of dynamic neural networks, models equipped with early exits adapt the computing path of individual samples to be analyzed to their specific complexity [12]. EE mechanisms have been primarily applied in computer vision [10], natural language processing, and speech recognition tasks [22] [17], and, importantly, they have been shown to reduce energy consumption and latency [18]. Furthermore, since EE mechanisms share less information during the inference phase, they are potentially aligned with the goal of improving the privacy of FL. However, their impact on privacy in neural models trained under different settings has

ApPLIED, June 19, 2023, Orlando, FL

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

not been analyzed. To the best of our knowledge, this paper is the first to explore the potential of early exits in enhancing privacy protection in both centralized and FL settings, specifically against MIA. To do so, we leverage both threshold attacks and logistic regression attacks as relevant privacy assessment metrics to evaluate the impact of EE techniques on privacy preservation.

More specifically, by means of experiments, we evaluate the resiliency of different models and find that models without early exits are more vulnerable to MIAs. Specifically, we assess the vulnerability of the Convolutional Neural Network (CNN), AlexNet, and ResNet18 models to such attacks during the training process, and conduct experiments in both centralized and FL settings. Our results indicate the potential of modern dynamic architectures to boost privacy, thus encouraging privacy-specific study and design of this class of models.

2 RELATED WORK AND NOVEL CONTRIBUTION

The overarching objective of this paper is to bridge the gap between EE mechanisms and privacy preservation in the context of both centralized and FL settings.

In the recent literature, several approaches have been proposed to boost privacy in FL frameworks. For instance, [3] divides the model into two parts, with one part residing on the client side for personalization and the other part on the server side for generalization. However, this approach potentially suffers from privacy leakage, and a comprehensive analysis of the privacy risks associated with it is lacking.

Other studies, such as [8], [2], and [20], have focused on locking personalized layers on each edge device to guarantee privacy. These works have employed sensitive attribute inference and membership inference attacks to assess privacy risks. [19] introduces a noise layer to provide privacy guarantees, while [23] explores the privacy and communication efficiency trade-offs in split learning algorithms within federated settings, using attack resilience defined in the same paper as a privacy assessment metric.

Unlike previous work, our objective is to assess the versatility and effectiveness of EE mechanisms, especially in potential privacy benefits in an FL setting.

3 METHODOLOGY

In FL, MIA is a type of attack aimed at the global model that is trained through various rounds between federated server and client devices. Fig. 1 presents an overview of the scenario we address and use for our experiments. We consider a typical FL setting (depicted in the top image), which uses the FedAve algorithm [13]. Each of the N clients possesses its own dataset, ranging from $\mathcal{D}_1$ to $\mathcal{D}_N$. Throughout each communication round, every client trains its model locally with early exits and sends its weights to the federated server. The server then averages the weights and returns the updated model weights to the clients.

Each client has its own local model with branches whose architecture integrates multiple early exits. Fig. 1(bottom), as an example, depicts a CNN with two exit points, one is an early exit point we injected and the other is the regular output of the model. Each added branch consists of a subset of the model's layers and an exit point.

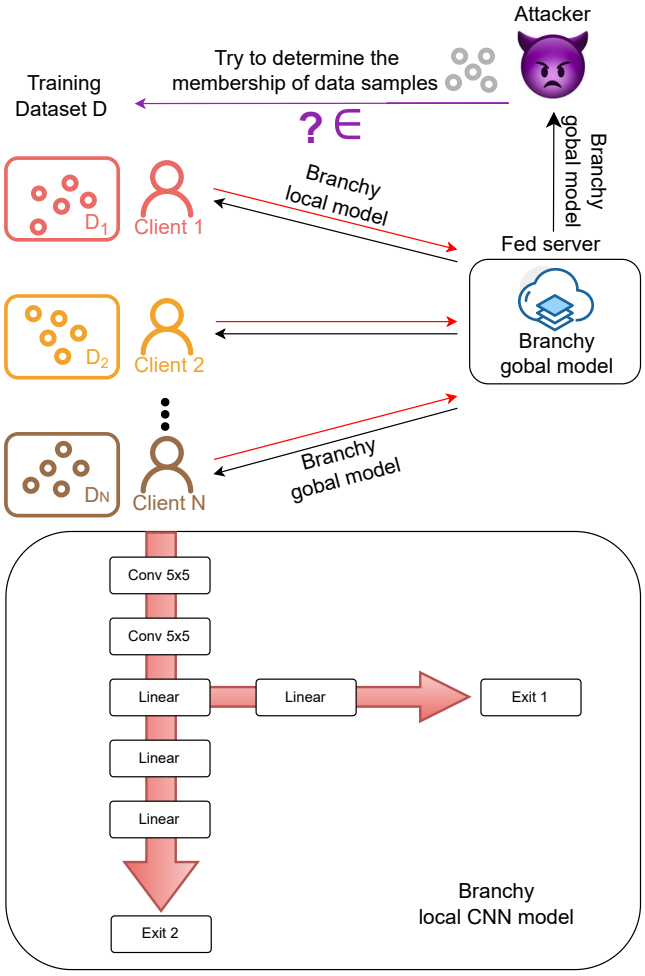

Figure 1: Overview of ours setting. Standard MIAs are adopted. A simple CNN model with an early exit is presented and noted as a branchy local CNN model that is trained locally at each client.

The training methodology for the EE strategy applies joint training. According to this approach, an overall loss function is defined for each early exit classifier. The loss function for the generic j-th exit is defined as:

$$\mathcal{L}_{CE}(\hat{y}^j, y_{true}) = -\log\left(\frac{\exp(\hat{y}_{true})}{\sum_{c\in C}\exp(\hat{y}_c)}\right), \tag{1}$$

where $\hat{y}^j$ and y_{true} are, respectively, the output and the correct label given to a model input x, and $\hat{y}_c$ is the model returned probability for class $c \in C$, with C being the set of considered classes. The objective of the EE model is to minimize the weighted sum of $\mathcal{L}_{CE}$ with a weighted scalar λ, as represented by the following expression [12]:

$$\mathcal{L}_{joint}([\hat{y}^1, ..., \hat{y}^J], y_{true}) = \sum_{j=1}^{J} \lambda_j \mathcal{L}_{CE}(\hat{y}^j, y_{true}) \tag{2}$$

where $[\hat{y}^1, \ldots, \hat{y}^J]$ are the outputs from J exits, and the correct label y_{true} is shared to every branch in the model.

The inference process of EE models enables the classification of data samples at earlier stages, utilizing the early exit points injected during the training phase. Suppose a data sample achieves a predefined confidence score threshold at any exit. In that case, it can exit early from the larger model without being processed through the remaining layers, thus enabling faster inference.

An MIA involves constructing an unsupervised binary classifier $\mathcal{M}_{\text{attack}}$ to ascertain whether a data sample x_i is part of the training datasets $\mathcal{D}$ used to train a model $\mathcal{M}$. Each data sample x_i has its corresponding label y_i. The training dataset $\mathcal{D}$ is private and was used for training the model $\mathcal{M}$. We apply MIA in a black-box setting where the adversary is able to submit a data sample $x_{i,\text{attacker}}$ to the trained target model and receives the model's probabilistic output. More specifically, test labels and probabilistic test output would be provided to the $\mathcal{M}_{\text{attack}}$. The $\mathcal{M}_{\text{attack}}$ would perform MIA based on the received information. The output of $\mathcal{M}_{\text{attack}}$ is 1 (noted as a member of the training dataset) or 0 (noted as not a member of the training dataset).

As mentioned, two metrics can be used to perform MIA: a threshold attack metric [16], and a logistic regression attack metric [21]. Both metrics make the membership prediction based on the confidence of the target model's output $\hat{p}(y|x_{i,\text{attacker}})$. In threshold attacks, $\mathcal{M}_{\text{threshold attack}}$ extracts the highest posterior and tags the input data sample as a member of the posterior if such a quantity is above a predefined threshold η:

$$\mathcal{M}_{\text{threshold attack}}(\hat{p}(y|x)) = \mathbb{1}(\max \hat{p}(y|x) \geq \eta)\,. \tag{3}$$

The logistic regression attack is instead performed by training a logistic regression classifier with the input of training labels and the returned probability as features.

The effectiveness of these two attacks is quantified at the end of each FL round through two metrics, namely, the area under the curve (AUC) and the attacker advantage, shown later in Section 4. The AUC metric measures the ability of a classifier to distinguish between the positive and negative classes, with a value of 0.5 indicating random guesses and a value of 1 indicating perfect classification. A higher AUC value indicates that the adversary's model is more successful at distinguishing between training and non-training samples, which implies a higher privacy risk. The attacker advantage metric quantifies the difference between the true positive rate and the false positive rate of the attack. A value of 0 indicates that the attacker's performance is equivalent to random guessing. A positive value indicates that the adversary is performing better than random guessing, and a value of 1 means that the attacker has perfect accuracy in identifying whether a data point belongs to the training dataset $\mathcal{D}$.

4 PRIVACY EVALUATION RESULTS

In this section, we demonstrate the effectiveness of early exits attached to different models to mitigate privacy leakage, particularly in MIA. All experiments were conducted using the CIFAR-10 dataset. For the centralized setting, the entire dataset was utilized. In the Federated Learning (FL) scenario, the dataset was partitioned among 10 clients using an extreme label skew non-i.i.d. distribution. Each client was allocated data from two classes out of a total of 10 classes. Specifically, each client was given 10,000 samples in total, with 5,000 samples for each class. Our intention was to partition the data in a way that the local dataset of each client is not a representative sample of the entire dataset, unlike the data in the centralized setting. Tensorflow Privacy library [1] was used to generate the privacy metrics such as logistic attacks and threshold attacks.

We begin with a simple experiment to present the influence of EE techniques on privacy. We consider a CNN model with only one early exit attached to it without incorporating an FL setting. The CNN model consists of two convolutional layers and three fully connected layers. We then add a branch with one fully connected layer after the second convolutional layer. For AlexNet [11], two early exit points were added after the first convolutional layer and the second convolutional layer. Both early exits have a convolutional layer followed by a fully connected layer. For ResNet18 [4], one early exit point was added between the fourth and fifth ResNet blocks with one convolutional layer and a fully connected layer.

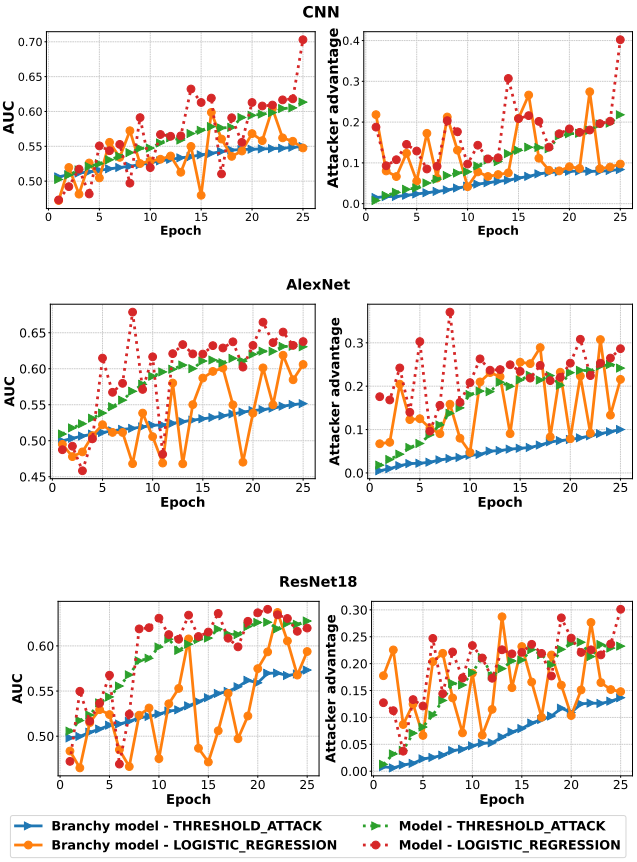

Figure 2: Privacy assessment obtained comparing a simple CNN with and without early exits, where the MIA is executed after each round during training.

In Fig. 2, CNN. AlexNet and ResNet18 show a striking separation between models with and without early exits attached, in the case of both logistic attacks and threshold attacks in a centralized setting.

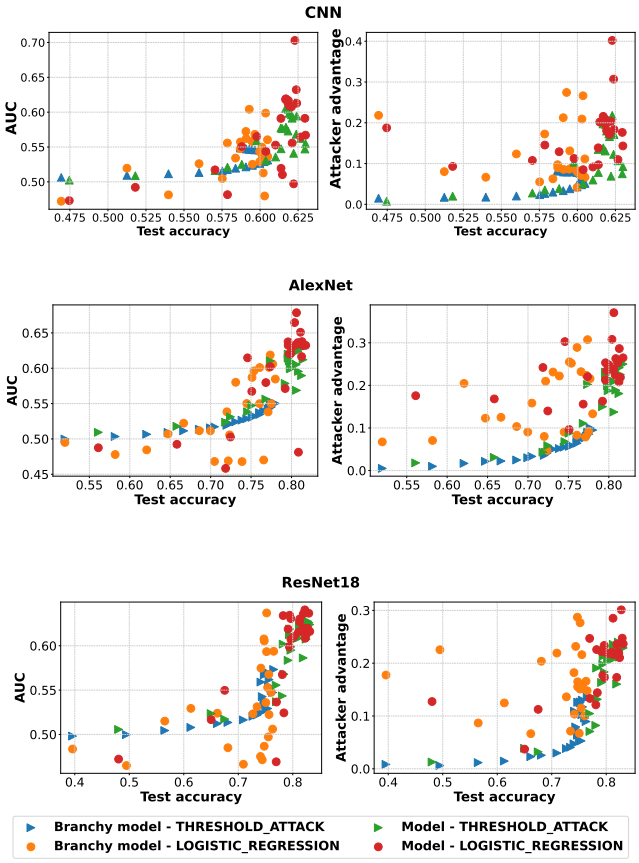

Figure 3: Privacy and accuracy trade-off with and without early exits in a centralized setting

This observation suggests that attaching exits largely reduces the risks of privacy leakage in MIAs as training goes on. Also, notice that threshold attacks only used one threshold to determine the membership of data instead of training a classifier as in the logistic attacks, leading to less noisy curves.

The trade-off between privacy risks and accuracy is presented in Fig. 3 and 5. The model with branches in both centralized and federated settings achieves comparable accuracy compared to the model without branches. However, for the same level of accuracy, the model with branches exhibits lower privacy risks. In the federated setting, the performance of the branchy model is slightly better.

5 DISCUSSION AND CONCLUSION

It is important to highlight that the privacy risks associated with deep learning models are contingent upon various factors. As the depth of the model architecture grows, the model becomes less susceptible to MIA pointing to a trade-off between model complexity and privacy vulnerabilities. Early exit mechanisms result in 'reduced' confidence scores for the output, effectively providing less information for an attacker attempting to mimic the targeted model's output. This phenomenon indicates that early exit mechanisms may play a crucial role in enhancing privacy protection.

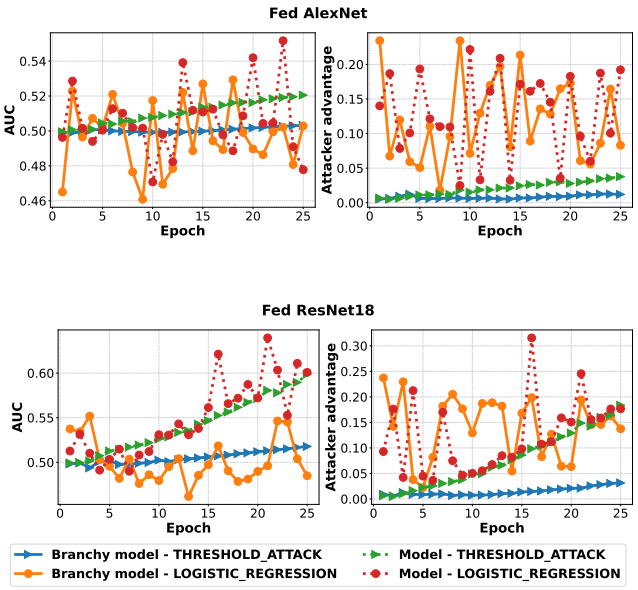

Figure 4: Privacy assessment with and without early exits in an FL setting.

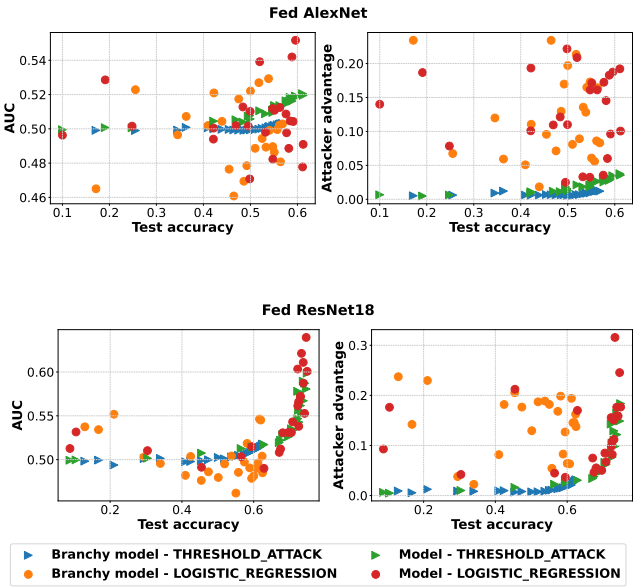

Figure 5: Privacy and accuracy trade-off with and without early exits in an FL setting.

Further investigation is necessary to understand the interplay between EE mechanisms and privacy preservation. Future research involves examining the impact of varying the number and position of early exits connected to the models, adjusting the scalar of the loss function for joint training, and exploring alternative attack strategies. Additionally, comparisons can be made between the ability of the EE mechanisms to enhance privacy and the performance of existing privacy metrics, such as differential privacy.

REFERENCES

[1] 2019. TensorFlow Privacy. https://github.com/tensorflow/privacy.

[2] Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. 2021. When the Curious Abandon Honesty: Federated Learning Is Not Private. https://doi.org/10.48550/ARXIV.2112.02918

[3] Dong-Jun Han, Do-Yeon Kim, Minseok Choi, Christopher G. Brinton, and Jaekyun Moon. 2022. SplitGP: Achieving Both Generalization and Personalization in Federated Learning. https://doi.org/10.48550/ARXIV.2212.08343

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[5] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2021. Membership Inference Attacks on Machine Learning: A Survey. https://doi.org/10.48550/ARXIV.2103.07853

[6] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. arXiv:2103.07853 [cs.LG]

[7] M. S. Jere, T. Farnan, and F. Koushanfar. 2021. A Taxonomy of Attacks on Federated Learning. *IEEE Security amp; Privacy* 19, 02 (mar 2021), 20–28. https://doi.org/10.1109/MSEC.2020.3039941

[8] Théo Jourdan, Antoine Boutet, and Carole Frindel. 2021. Privacy Assessment of Federated Learning using Private Personalized Layers. (2021). https://doi.org/10.48550/ARXIV.2106.08060

[9] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2019. Advances and Open Problems in Federated Learning. https://doi.org/10.48550/ARXIV.1912.04977

[10] Alexandros Kouris, Stylianos I. Venieris, Stefanos Laskaridis, and Nicholas D. Lane. 2021. Multi-Exit Semantic Segmentation Networks. https://doi.org/10.48550/ARXIV.2106.03527

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (Eds.), Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[12] Yoshitomo Matsubara, Marco Levorato, and Francesco Restuccia. 2022. Split Computing and Early Exiting for Deep Learning Applications: Survey and Research Challenges. *ACM Comput. Surv.* 55, 5, Article 90 (dec 2022), 30 pages. https://doi.org/10.1145/3527155

[13] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2016. Communication-Efficient Learning of Deep Networks from Decentralized Data. (2016). https://doi.org/10.48550/ARXIV.1602.05629

[14] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. https://doi.org/10.48550/ARXIV.1806.01246

[15] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2016. Membership Inference Attacks against Machine Learning Models. https://doi.org/10.48550/ARXIV.1610.05820

[16] Liwei Song and Prateek Mittal. 2020. Systematic Evaluation of Privacy Risks of Machine Learning Models. arXiv:2003.10595 [cs.CR]

[17] Raphael Tang, Karun Kumar, Ji Xin, Piyush Vyas, Wenyan Li, Gefei Yang, Yajie Mao, Craig Murray, and Jimmy Lin. 2022. Temporal Early Exiting for Streaming Speech Commands Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7567–7571. https://doi.org/10.1109/ICASSP43922.2022.9746863

[18] Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. 2017. BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks. https://doi.org/10.48550/ARXIV.1709.01686

[19] Chandra Thapa, M. A. P. Chamikara, Seyit Camtepe, and Lichao Sun. 2020. SplitFed: When Federated Learning Meets Split Learning. https://doi.org/10.48550/ARXIV.2004.12088

[20] Yuxin Wen, Jonas A. Geiping, Liam Fowl, Micah Goldblum, and Tom Goldstein. 2022. Fishing for User Data in Large-Batch Federated Learning via Gradient Magnification. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 23668–23684. https://proceedings.mlr.press/v162/wen22a.html

[21] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. arXiv:1709.01604 [cs.CR]

[22] Ji Won Yoon, Beom Jun Woo, and Nam Soo Kim. 2022. HuBERT-EE: Early Exiting HuBERT for Efficient Speech Recognition. https://doi.org/10.48550/ARXIV.2204.06328

[23] Zongshun Zhang, Andrea Pinto, Valeria Turina, Flavio Esposito, and Ibrahim Matta. 2023. Privacy and Efficiency of Communications in Federated Split Learning. https://doi.org/10.48550/ARXIV.2301.01824