Software Engineering and Applications 软件工程与应用, 2022, 11(6), 1465-1472
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.116151

Hans汉斯

联邦学习的安全性问题分析

张建红，李　晶，崔广金

哈尔滨师范大学计算机科学与信息工程学院，黑龙江　哈尔滨

收稿日期：2022年11月30日；录用日期：2022年12月23日；发布日期：2022年12月31日

摘　要

随着大数据和人工智能的快速发展，人们对数据安全的需求也日益提高。为了解决目前面临的隐私保护和数据孤岛问题，联邦学习受到了各界学者的广泛关注与研究。虽然联邦学习是一种很有潜力的机器学习技术，使位于不同地理位置的多个用户可以在不共享数据的情况下协作训练机器学习模型，但是也存在一些安全性问题。因此本文对联邦学习的安全性问题进行了总结和分析，这对联邦学习的发展及应用具有重要的意义。本文首先对联邦学习的基本概念和分类进行了详细地阐述；接着，深入分析了联邦学习面临的安全性问题，包括投毒攻击、后门攻击和基于生成对抗网络(GAN)的攻击；然后，归纳总结了不同攻击的防御方法。最后，对联邦学习的应用前景以及未来的研究方向进行了总结与展望。

关键词

联邦学习，数据安全，防御机制

Analysis on the Security of Federated Learning

Jianhong Zhang, Jing Li, Guangjin Cui

College of Computer Science and Information Engineering, Harbin Normal University, Harbin Heilongjiang

Received: Nov. 30th, 2022; accepted: Dec. 23rd, 2022; published: Dec. 31st, 2022

Abstract

With the rapid development of big data and artificial intelligence, people's demand for data security is also increasing. In order to solve the problems of privacy protection and data islands, federated learning has been widely concerned and studied by scholars from all walks of life. Although federated learning is a promising machine learning technology that enables multiple users in different geographical locations to collaborate on training machine learning models without sharing data, there are also some security issues. Therefore, this paper summarizes and analyzes the secu-

文章引用: 张建红, 李晶, 崔广金. 联邦学习的安全性问题分析[J]. 软件工程与应用, 2022, 11(6): 1465-1472.
DOI: 10.12677/sea.2022.116151

rity issues of federated learning, which is of great significance to the development and application of federated learning. Firstly, this paper expounds the basic concepts and classification of federated learning in detail. Then, the security problems faced by federated learning are analyzed in depth, including poisoning attack, backdoor attack and Generative Adversarial Network (GAN)-based attack. Then, the defense methods of different attacks are summarized. Finally, the application prospect of federated learning and the future research direction are summarized and prospected.

Keywords

Federated Learning, Data Security, Defense Mechanism

1. 引言

随着人工智能的快速发展，机器学习(Machine Learning, ML)作为其核心技术在自动驾驶、智能家居、金融、医疗保健等领域得到了广泛的应用。机器学习的发展主要分为三个阶段，分别是集中式学习、分布式学习和联邦学习[1]。在集中式学习阶段，用户的数据都需要存储在中央服务器上进行集中训练，是一种“模型不动，数据动”的学习模式，这会造成在数据传输和集中训练数据模型的过程中存在严重的隐私泄露风险。随着个人和企业对隐私数据的重视，集中式学习已经无法满足社会发展需求，于是分布式学习被提出。在分布式学习阶段，可以利用多个计算节点执行机器学习任务，将最终的学习结果发送到中央服务器进行聚合，无需将敏感数据上传到中央服务器，有效解决了数据安全和隐私保护的问题。但是由于在分布式学习过程中各方难以实现数据共享，导致了严重的数据孤岛现象。

为了解决上述数据安全和数据孤岛的问题，谷歌在 2016 年提出了联邦学习(Federated Learning, FL)这一新的理论[2]，联邦学习是一种“数据不动，模型动”的学习模式。在联邦学习阶段，终端用户设备在本地进行数据存储和模型训练，并将训练得到的本地模型参数发送到中央服务器进行安全聚合，中央服务器在聚合后将全局模型参数反馈给用户，用户根据收到的全局模型参数进行本地模型的更新，从而有效保障了用户的数据安全，有利于各方之间的数据共享。

尽管联邦学习作为一种隐私保护的机器学习方法得到了大量关注与研究，但是仍然存在一定的安全性问题。本文主要对联邦学习的定义和分类进行了详细介绍，并对联邦学习中的安全性问题和现有的解决方法进行了总结和分析，最后对联邦学习未来的应用前景和发展方向进行了展望。

2. 联邦学习概述

2.1. 联邦学习的定义

联邦学习也可以被称为联合学习、联邦机器学习或联盟学习，是一种新型的隐私保护和去中心化学习范式，不需要收集各个用户所有的数据便能协作地训练一个模型的机器学习过程[3]。定义 N 个数据所有者 $\{F_1, F_2, \cdots, F_N\}$，他们各自的数据为 $\{D_1, D_2, \cdots, D_N\}$，为了训练机器学习模型需要将他们的数据进行合并。传统的方法是将所有数据放在一起并使用 $D = D_1 \cup D_2 \cdots \cup D_N$ 来训练模型 M_{SUM}。联邦学习是一个不需要集中数据训练的学习过程，其中数据所有者协作训练模型 M_{FED}，在此过程中，任何数据所有者 F_i 都不会将其数据 D_i 泄露给其他用户。此外，设 V_{FED} 和 V_{SUM} 分别表示 M_{FED}， M_{SUM} 的模型精度，V_{FED} 应

该非常接近于 V_{SUM} 的性能。形式上，设 δ 为一个非负实数，如果

$$\left|V_{FED}-V_{SUM}\right|<\delta \tag{1}$$

则联邦学习模型 M_{FED} 具有 δ 的精度损失。

2.2. 横向联邦学习

横向联邦学习(Horizontal Federated Learning, HFL)指的是用户之间数据的特征重叠较多，样本重叠较少的情况[4]。横向联邦学习也可以被称为按样本划分的联邦学习，其公式表示如下。

$$X_m=X_n,Y_m=Y_n,I_m\neq I_n,\forall D_m,D_n,m\neq n \tag{2}$$

其中，I_m 和 I_n 分别表示用户 m 和用户 n 的样本 ID 空间，D_m 和 D_n 分别表示用户之间拥有的数据集，(X_m,X_n) 和 (Y_m,Y_n) 分别表示用户之间的数据特征空间和标签空间对。

横向联邦学习适合同一行业或领域但是用户不同的机器学习任务，其架构如图 1 所示。

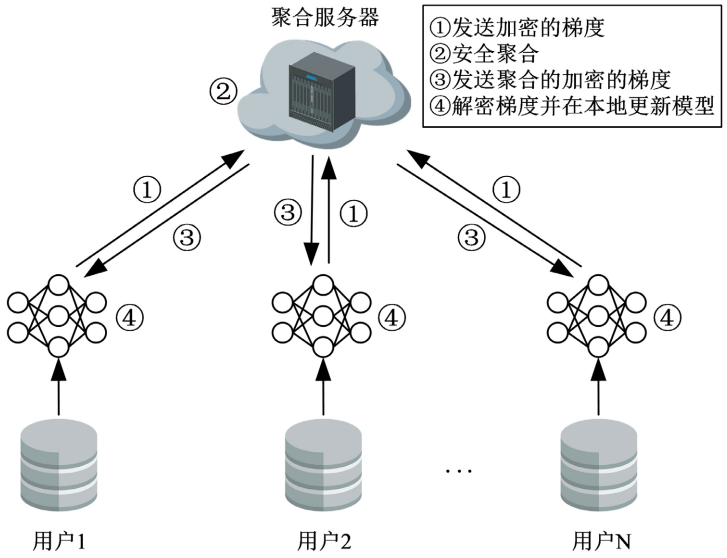

Figure 1. Diagram of the architecture of horizontal federated learning
图 1. 横向联邦学习架构图

2.3. 纵向联邦学习

纵向联邦学习(Vertical Federated Learning, VFL)指的是用户之间数据的特征重叠较少，样本重叠较多的情况[4]。纵向联邦学习也可以被称为按特征划分的联邦学习，其公式表示如下。

$$X_m\neq X_n,Y_m\neq Y_n,I_m=I_n,\forall D_m,D_n,m\neq n \tag{3}$$

其中，I_m 和 I_n 分别表示用户 m 和用户 n 的样本 ID 空间，D_m 和 D_n 分别表示用户之间拥有的数据集，(X_m,X_n) 和 (Y_m,Y_n) 分别表示用户之间的数据特征空间和标签空间对。

纵向联邦学习适合跨行业跨领域但是用户群体相同的机器学习任务，其架构如图 2 所示。

2.4. 联邦迁移学习

联邦迁移学习(Federated Transfer Learning, FTL)指的是用户之间数据的特征和样本都没有重叠或重叠较少的情况[4]。联邦迁移学习不需要服务器作为用户之间的协调者，其公式表示如下。

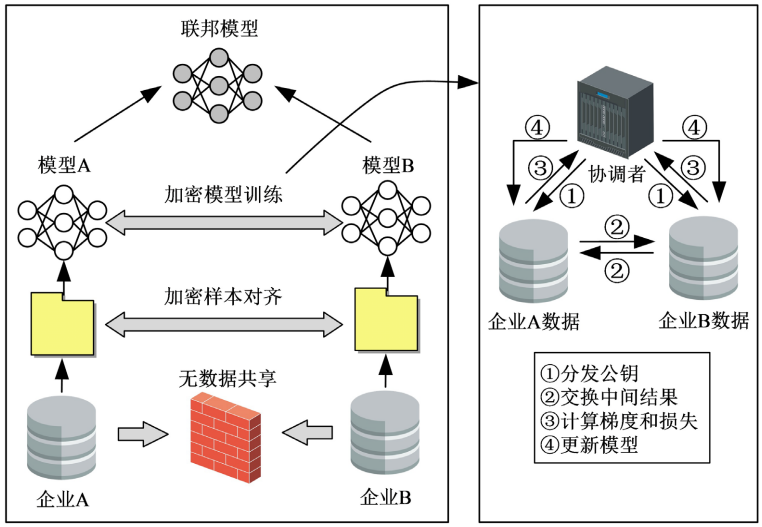

Figure 2. Diagram of the architecture of vertical federated learning
图 2. 纵向联邦学习架构图

$$X_m \neq X_n, Y_m \neq Y_n, I_m \neq I_n, \forall D_m, D_n, m \neq n \tag{4}$$

其中，I_m 和 I_n 分别表示用户 m 和用户 n 的样本 ID 空间，D_m 和 D_n 分别表示用户之间拥有的数据集，(X_m, X_n) 和 (Y_m, Y_n) 分别表示用户之间的数据特征空间和标签空间对。

联邦迁移学习适合跨行业跨领域并且用户不同的机器学习任务，其架构如图 3 所示。

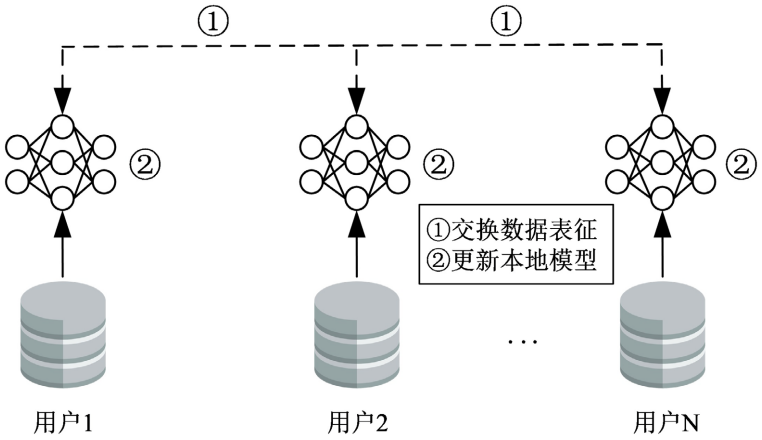

Figure 3. Diagram of the architecture of federated transfer learning
图 3. 联邦迁移学习架构图

3. 联邦学习中的安全性问题

3.1. 投毒攻击

投毒攻击(poisoning attack)是联邦学习中最常见的攻击方法[5]。这种类型的攻击可能具有不同的性质，但其原理基本相同，攻击者将发送伪造的数据以影响全局模型。这种攻击可以通过两种方式进行：

1) 数据投毒攻击(data poisoning attack)：为了在联邦学习过程中发起数据投毒攻击，攻击者对参与学习过程的智能终端设备中的训练数据进行投毒，从而危及全局模型的准确性。攻击者可以通过将中毒数

据直接注入目标设备或通过其他设备注入中毒数据来毒害用户的数据。为了不被发现，攻击者通常只是在每轮训练过程中稍微修改训练数据，但是在多次迭代训练后会对全局模型造成较大的影响。

2) 模型投毒攻击(model poisoning attack)：模型投毒攻击类似于数据投毒攻击，但是这类投毒方式比较直接，攻击者试图毒害本地模型而不是本地数据。模型投毒攻击背后的主要动机是在全局模型中引入错误。攻击者通过破坏智能终端设备并修改其本地模型参数来发起模型投毒攻击，从而影响全局模型的准确性。

3.2. 后门攻击

后门攻击(backdoor attack)可以归入模型投毒攻击的范畴。然而，相比之下，它们的透明度较低。在后门攻击的过程中，攻击者能够将隐藏的后门插入到全局模型中，在预测阶段，参与联邦学习的受损设备会触发后门攻击中的后门触发器，从而自动出现错误预测。使用后门攻击，攻击者可以在不影响全局模型准确性的情况下对某些任务进行错误标记。在将检测规避例程纳入攻击者的损失函数后，中毒模型更新的外观和行为类似于没有后门训练的模型。当后门模型与其他模型没有差异时，后门存在的检测变得复杂。后门攻击的影响可能是毁灭性的，因为后门能够被永久地隐藏在模型当中，当带有特定后门触发器的样本出现后将对其进行激活。后门攻击也被称为目标攻击，此类攻击的强度取决于受损设备的比例和联邦学习的模型容量。

3.3. 基于生成对抗网络(GAN)的攻击

投毒攻击已经演变，并提出了创建投毒模型的新方法，其中一种被称为 PoisonGAN [6]。该攻击使用生成对抗网络(Generative Adversarial Network, GAN)攻击联邦学习。攻击者主要依赖于迭代更新的全局模型参数生成由攻击者随意控制的真实数据集。然后利用所设计的中毒数据生成方法，有效地降低了攻击假设，使攻击在实际应用中变得可行。GAN 在给定的客户端上进行优化，依赖于模型更新，目的是操纵全局模型的参数。实际上，攻击者在进行攻击之前并不需要拥有数据集。这种攻击实际上是由联邦学习基础结构促成的，因为攻击者可以访问本地模型，而集中学习则更难。这种基于 GAN 的攻击是最难检测的，因为生成的数据是真实的。它的应用会给模型的准确性带来灾难性的后果，因为它引入了强而可控的偏差。

4. 联邦学习中的防御方法

4.1. 投毒攻击防御方法

针对投毒攻击的防御方法主要是连续监控客户端行为的反应性方法。Rodríguez-Barroso 等人[7]提出了一种筛选恶意客户端的方法，该方法通过使用人工智能来检测模型变化或不一致的数据分布。Cao 等人[8]中提出了一种遵循相同原理但适用于几个恶意用户联合攻击的过滤方法，这被称为 sniper。Levine 等人[9]提出了两种针对投毒攻击的新型可证明的防御方法，分别是深度分区聚合(Deep Partition Aggregation, DPA)和半监督 DPA (Semi-Supervised DPA, SS-DPA)。文献[10]中提出了另一种针对投毒攻击的防御措施，该技术包括在每次新模型更新时检查全局模型的性能。

4.2. 后门攻击防御方法

针对后门攻击的防御方法主要包括检测模型参数的异常进行删除以及最小化模型的大小，以降低其复杂性和容量，同时可能提高其准确性。Desai 等人[11]提出了一个基于区块链的混合型联邦学习框架，阻止后门攻击检测并惩罚攻击者。Liu 等人[12]提出了一种新的微剪枝技术，将剪枝和微调进行结合，能

够成功地削弱甚至消除后门攻击。由于生成的模型表现力较弱，因此后门攻击的执行更加复杂。这种方法也会带来一些有益的副作用，参数数量的减少确实降低了通信成本并降低了消息拦截概率。

4.3. 基于 GAN 的攻击防御方法

针对基于 GAN 的攻击防御需要专门的方法，因为这些特定的投毒攻击使用普通的检测方法较难解决。如可以利用先进的拜占庭演员检测算法[13]。此外，Li 等人[14]提出了一种基于模型蒸馏的异构联邦学习，通过使用迁移学习和知识蒸馏来开发一个通用的框架，在联邦学习过程中，每个用户不仅拥有自己的私有数据，而且具有唯一设计的模型，从而有效防御基于 GAN 的攻击。然而，目前来看，针对这类攻击的防御技术仍然缺乏开发和记录，需要进行进一步研究。

5. 联邦学习的应用前景和研究方向

5.1. 联邦学习的应用前景

现阶段机器学习应用面临的挑战主要是如何在保障数据安全的前提下，实现不同行业之间的数据共享。联邦学习作为应对该挑战的新技术，在涉及大规模数据收集和模型训练等领域具有广泛的应用前景。

1) 智能医疗。在智能医疗领域中，各医疗机构为了保护患者的隐私信息一般不会进行数据共享，导致现有的智能医疗系统很难收集到能够全面描述患者症状的数据，训练的模型不够准确。因此可以运用联邦学习开发新型的智能医疗系统，将不同医疗机构的医疗图像数据、药物开发数据和专家知识数据等进行融合，在保护患者敏感数据的前提下打破医疗数据孤岛，从而训练性能更好的模型。

2) 智慧消费金融。在智慧消费金融领域中，为了给消费者提供更好的金融信息和消费服务，往往需要将消费者的购买开销、投资偏好和信用记录等数据进行收集并训练。但是这些数据通常由不同的部门进行收集，然而这些部门之间不能直接进行数据共享。同时由各方存储的数据通常是异构的，难以利用传统的机器学习处理。因此可以运用联邦学习解决这些问题，实现跨数据和跨领域的模型训练。

3) 智慧城市。在智慧城市领域中，城市计算能够对城市中不同信息源产生的大量异构数据进行整合和分析，从而解决当前城市中面临的空气污染、交通拥挤等问题。但是仍然存在城市管理中不同应用数据整合缺失的问题和一定的安全风险。为了更好地协同建设智慧城市，可以运用具有隐私保护性和协作性的联邦学习，在不损害数据安全和隐私的情况下，建立最优的模型，促进智慧城市的发展。

5.2. 联邦学习的研究方向

联邦学习由于其能够解决数据安全和数据孤岛的问题，在智能医疗、智慧金融服务和智慧城市等领域具有巨大的潜力。然而，它也面临着一些安全性问题需要进一步研究。

1) 联邦学习的可追溯性。联邦学习的一个主要挑战是在底层训练过程的整个生命周期中对全局模型进行跟踪。例如，当全局模型中的预测值发生变化的时候，需要通过联邦学习的反向跟踪能力对导致该变化的客户端聚合值进行识别，从而对恶意攻击者进行捕获。如果模型训练行为背后的逻辑是一个黑匣子，那么将无法对逻辑现实进行控制，而只能依赖于固定的人工智能技术。因此，实现模型异常的可追溯性需要进行进一步研究。

2) 联邦学习的标准化。联邦学习作为一种近年来刚提出的能够实现隐私保护的机器学习方法，在实际应用中需要对所有利弊进行详细分析，然后具体定义标准化的技术，以支持不同领域对联邦学习的相关需求。由于隐私保护是联邦学习中的一个关键因素，因此需要在研究联邦学习标准化的时候考虑到隐私性，重点是增强隐私的同时使每个需求的方法标准化，并定义实施此类增强方法的具体过程。

3) 联邦学习的权衡。在联邦学习的过程中，为了进一步抵抗模型参数受到攻击，提出了相关的防御

方案。但是现有的解决方案都是以牺牲联邦学习模型的性能为代价来保证数据安全和隐私，还给服务器带来了一定的计算压力，降低了联邦学习的效率[15]。因此，如何在进行联邦学习的过程中达到隐私性、可用性和高效性之间的权衡是未来值得研究的一个方向。

6. 总结

联邦学习作为一种新型的协作性机器学习方法，有效地解决了隐私保护和数据孤岛的问题，但是仍然存在一些潜在的安全性问题。本文对联邦学习的发展历程和基本概念进行了阐述，并详细介绍了横向联邦学习、纵向联邦学习和联邦迁移学习这三种典型的联邦学习框架。同时对联邦学习中的安全性问题进行了分析，主要分为投毒攻击、后门攻击和基于生成对抗网络(GAN)的攻击。在此基础上，对现有的攻击防御方法进行了总结和介绍。最后，对联邦学习的应用前景和未来研究方向进行了探讨。

基金项目

哈尔滨师范大学研究生培养质量提升工程“新工科背景下创新创业型研究生多维培养模式的研究——以网络安全方向研究生培养为例”和哈尔滨师范大学本科专业人才培养方案研究改革专项(XJGRYK2022012)。

参考文献

[1] 王帅, 李丹. 分布式机器学习系统网络性能优化研究进展[J]. 计算机学报, 2022, 45(7): 1384-1411.

[2] McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B.A. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, 20-22 April 2017, 1273-1282.

[3] 李少波, 杨磊, 李传江, 张安思, 罗瑞士. 联邦学习概述: 技术、应用及未来[J]. 计算机集成制造系统, 2022, 28(7): 2119-2138. https://doi.org/10.13196/j.cims.2022.07.018

[4] 周传鑫, 孙奕, 汪德刚, 葛桦玮. 联邦学习研究综述[J]. 网络与信息安全学报, 2021, 7(5): 77-92.

[5] 汤凌韬, 陈左宁, 张鲁飞, 吴东. 联邦学习中的隐私问题研究进展[J/OL]. 软件学报. https://doi.org/10.13328/j.cnki.jos.006411 1-33, 2022-11-15.

[6] Zhang, J., Chen, B., Cheng, X., Binh, H.T.T. and Yu, S. (2020) PoisonGAN: Generative Poisoning Attacks against Federated Learning in Edge Computing Systems. *IEEE Internet of Things Journal*, **8**, 3310-3322. https://doi.org/10.1109/JIOT.2020.3023126

[7] Rodríguez-Barroso, N., Martínez-Cámara, E., Luzón, M., *et al*. (2007) Dynamic Federated Learning Model for Identifying Adversarial Clients. ArXiv Preprint ArXiv: 2007.15030.

[8] Cao, D., Chang, S., Lin, Z., Liu, G. and Sun, D. (2019) Understanding Distributed Poisoning Attack in Federated Learning. 2019 *IEEE* 25*th International Conference on Parallel and Distributed Systems* (*ICPADS*), Tianjin, 4-6 December 2019, 233-239. https://doi.org/10.1109/ICPADS47876.2019.00042

[9] Levine, A. and Feizi, S. (2020) Deep Partition Aggregation: Provable Defense against General Poisoning Attacks. ArXiv Preprint ArXiv: 2006.14768.

[10] Bhagoji, A.N., Chakraborty, S., Mittal, P. and Calo, S. (2019) Analyzing Federated Learning through an Adversarial Lens. *Proceedings of the* 36*th International Conference on Machine Learning*, Long Beach, 9-15 June 2019, 634-643.

[11] Desai, H.B., Ozdayi, M.S. and Kantarcioglu, M. (2021) BlockFLA: Accountable Federated Learning via Hybrid Blockchain Architecture. *Proceedings of the* 11*th ACM Conference on Data and Application Security and Privacy*, Virtual Event USA, 26-28 April 2021, 101-112. https://doi.org/10.1145/3422337.3447837

[12] Liu, K., Dolan-Gavitt, B. and Garg, S. (2018) Fine-Pruning: Defending against Backdooring Attacks on Deep Neural Networks. In: Bailey, M., Holz, T., Stamatogiannakis, M. and Ioannidis, S., Eds., *Research in Attacks, Intrusions, and Defenses*. *Lecture Notes in Computer Science*, Vol. 11050, Springer, Cham, 273-294. https://doi.org/10.1007/978-3-030-00470-5_13

[13] Hayes, J. and Ohrimenko, O. (2018) Contamination Attacks and Mitigation in Multi-Party Machine Learning. 32*nd Conference on Neural Information Processing Systems* (*NeurIPS* 2018), Montréal, 3-8 December 2018.

[14] Li, D. and Wang, J. (2019) Fedmd: Heterogenous Federated Learning via Model Distillation. ArXiv Preprint ArXiv:

1910.03581.

[15] Isaksson, M. and Norrman, K. (2020) Secure Federated Learning in 5G Mobile Networks. *GLOBECOM* 2020-2020 *IEEE Global Communications Conference*, Taipei, 7-11 December 2020, 1-6. https://doi.org/10.1109/GLOBECOM42002.2020.9322479