之前工作的总结

FL的抗投毒：相当于复现了该论文 Privacy-Enhanced Federated Learning Against Poisoning Adversaries（CCF-A）

遗留问题

- 抗投毒/抗后门攻击方面：当恶意work超过50%时，我们做的，以及Krum等防御方法不起作用。

  usenix 2020 Justinian's GAAvernor Robust Distributed Learning 结合了强化学习可以实现仅有一个良性work时也能抵御拜占庭攻击。
- 效率方面：使用了同态加密（防止泄露隐私），耗时特别长
- 假设了数据分布是IID的（联邦学习局部梯度成为全局梯度无偏估计的前提，使用Person相关系数检测出恶意梯度的前提）。猜很多防御方法都用了IID这个假设前提
- 假设了work能投毒，其他服务器是诚实好奇的

感兴趣的方面

Conferences and Journals Collection for Federated Learning from 2019 to 2021 **含incentive works， Poisoning， Personalization， Optimization Distribution， Communication

- 除了poisoning works了解一些，incentive works，Optimization_Distribution蛮想了解和研究的

Advances and Open Problems in Federated Learning：先做Privacy和Robust方面的，学习 **Effificiency and Effectiveness**，**Ensuring Fairness and Addressing Sources of Bias**，**Addressing System Challenges**

Threats to Federated Learning: A Survey

1. FL的数据重构攻击/梯度泄露攻击
   - NIPS 2019 Deep Leakage from Gradients 通过加大训练的 batchsize 可以规避那些攻击。条件过于苛刻，例如要求恢复的数据样本数量要远小于总类别数目。
   - NIPS 2020 Inverting Gradients - How easy is it to break privacy in federated learning? 在几个迭代或几个图像上平均梯度也不能保护用户的隐私，证明任何输入到全连接层的输入都可以被分析重建，与架构无关。
   - CAFE: Catastrophic Data Leakage in Vertical Federated Learning NIPS 2021 利用 VFL 框架下 data index alignment，通过逐层的还原，实现了 VFL 阶段过程中 大批量训练数据 的还原。
   - USENIX 2022 Exploring the Security Boundary of Data Reconstruction via Neuron Exclusivity Analysis 基于基本的梯度泄露攻击~~（原理太难看了..）~~
   - Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning 2019利用多任务GAN 训练，恢复 client 级别的多种隐私，包括 数据所属的类别、真假、来自哪一个 client，最后还能实现 某位client 训练数据的恢复。 缺陷是要求 batchsize=1时才能恢复。
   - A Novel Attribute Reconstruction Attack in Federated Learning 2021 ZJU Attribute Reconstruction Attack 利用了受害者模型每轮更新的梯度，构建了一批虚拟样本并计算其对应的更新梯度，通过梯度上升优化得到 虚拟样本梯度和受害者模型上传的真实梯度的 cos 相似度的最大值：

$$X_g^* = \underset{X_a}{\operatorname{argmax}} \operatorname{cosinesim}(\nabla w_t', \nabla w_t)$$

然后再更新

- Source Inference Attacks in Federated Learning 确定某一数据来源于哪一参与方，成员推理攻击不能
- GAN Enhanced Membership Inference: A Passive Local Attack in Federated Learning 在FL场景下，各 client 所持有的数据存在分布不一致的情况，而之前的 shadow model 的方法要求 攻击者拥有部分和 target model 训练集同分布的辅助数据集，才能进行攻击；所以之前的方法在 FL 场景下的效果会大打折扣，或者说需要辅助信息（和 target model 训练集同分布的数据集）。本文介绍了一种利用 GAN 获取 整体参与者们 训练数据的分布，从而进行了更加精准的成员推断攻击
  Efficient passive membership inference attack in federated learning 相较于之前基于更新量的成员推断攻击，本文提出一种黑盒的利用 连续次更新量（结合了 time series） 的 推断攻击，仅需 **极小的计算量**（虽然攻击准确率差不多）。

2. Privacy:

   Analyzing User-Level Privacy Attack Against Federated Learning CCF-A 首次尝试通过恶意服务器的攻击来探索用户级隐私泄露

   A Framework for Evaluating Client Privacy Leakages in Federated Learning 分析本地训练的共享参数更新（来重建私有的本地训练数据。不同的超参数配置和攻击算法的不同设置对攻击有效性和攻击成本的影响。还评估了在使用通信效率高的FL协议时，不同梯度压缩率下的客户端隐私泄露攻击的有效性。

   PMLR 2021 Gradient Disaggregation: Breaking Privacy in Federated Learning by Reconstructing the User Participant Matrix inference attack

   Secure Aggregation is Insecure: Category Inference Attack on Federated Learning 类别隐私+Non-iid

   usenix 2022 Label inference attacks against vertical federated learning vertical federated learning安全研究相对较少

3. 防御各类隐私攻击

   Digestive neural networks: A novel defense strategy against inference attacks in federated learning COMPUTERS&Security2021 CCF-B
   本文在联邦学习场景下，提出了一种 Digestive neural networks （后称DNN，区别于传统的DNN），类似于输入数据的特征工程，用于“抽取原始数据的有效特征，并修改原始数据使之不同”，本地模型经过处理后的数据进行训练，从而大大降低了 FL 中各类推断攻击的成功率（假设server是攻击者）。

   MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers2021
   本文介绍了一种介于 client 和 server 中间的代理网络 MixNN proxy ，这样的代理网络有点类似同态加密（相当于利用神经网络进行加密），可以有效避免训练过程中的各类推断攻击。除此以外，本文还设计了一种利用 update 进行的 attribute 推断攻击，用于评估不同防御方法的防御效果。

   Efficient and Privacy-Enhanced Federated Learning for Industrial Artificial Intelligence2019
   本文提出了一种包含同态加密、差分隐私、多方安全计算的FL隐私保护框架。

4. 后门攻击：（和拜占庭攻击防御之间的关系？）

   https://github.com/THUYimingLi/backdoor-learning-resources 含Image and Video Classification attack-and-defense, Attack and Defense Towards Other Paradigms and Tasks(含FL), Evaluation, 是State-of-the-art一直在更新的

- usenix 2021 Blind Backdoors in Deep Learning Models 和BadNets差不多，~~但是这是顶会？？~~

  Main task和后门效果差的太多了吧，这都能实现：

Experiment	Main task	Synthesizer		T	Task accuracy ($\theta \rightarrow \theta^*$)	
		input μ	label ν		Main	Backdoor
ImageNet (full, SGD)	object recog	pixel pattern	label as 'hen'	2	65.3% → 65.3%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	pixel pattern	label as 'hen'	inf	69.1% → 69.1%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	single pixel	label as 'hen'	inf	69.1% → 68.9%	0% → 99%
ImageNet (fine-tune, Adam)	object recog	physical	label as 'hen'	inf	69.1% → 68.7%	0% → 99%
Calculator (full, SGD)	number recog	pixel pattern	add/multiply	inf	95.8% → 96.0%	1% → 95%
Identity (fine-tune, Adam)	count	single pixel	identify person	inf	87.3% → 86.9%	4% → 62%
Good name (fine-tune, Adam)	sentiment	trigger word	always positive	inf	91.4% → 91.3%	53% → 98%

- AISTATS 20 How To Backdoor Federated Learning
- dba distributed backdoor attacks against federated learning ICLR 20 分布式后门
- 联邦学习后门攻击总结（2019-2022）
- 《Backdoor Learning: A Survey》阅读笔记

Algorithm 1 Create a model that does not look anomalous and replaces the global model after averaging with the other participants' models.

Constrain-and-scale$(\mathcal{D}_{local}, D_{backdoor})$
Initialize attacker's model X *and loss function* l:
  $X \leftarrow G^t$
  $\ell \leftarrow \alpha \cdot \mathcal{L}_{class} + (1-\alpha) \cdot \mathcal{L}_{ano}$
for epoch $e \in E_{adv}$ **do**
  if $\mathcal{L}_{class}(X, D_{backdoor}) < \epsilon$ **then**
    // Early stop, if model converges
    break
  end if
  for batch $b \in \mathcal{D}_{local}$ **do**
    // inject c backdoors to the batch b
    $b \leftarrow \texttt{replace}(c, b, D_{backdoor})$
    $X \leftarrow X - lr_{adv} \cdot \nabla \ell(X, b)$
  end for
  if epoch $e \in step\_sched$ **then**
    // reduce learning rate
    $lr_{adv} \leftarrow lr_{adv} / step\_rate$
  end if
end for
// Scale up the model before submission.
$\widetilde{L}^{t+1} \leftarrow \gamma(X - G^t) + G^t$
return $\widetilde{L}^{t+1}$

其他后门攻击：（感觉很AI了）

2019 ACM SIGSAC Latent Backdoor Attacks on Deep Neural Networks 结合迁移学习，~~感觉比预会还难~~

见上面github总结

特定领域的后门攻击：

Backdoor Attack against Speaker Verification CCF-B

Backdoors in Federated Meta-Learning

5. FL_Robust (model poisoning/后门攻击/拜占庭攻击)的防御
    - 联邦学习中的后门攻击的防御手段
    - Robust Federated Training via Collaborative Machine Teaching using Trusted Instances AAAI 20
    - 神经网络中后门攻击的防御：加噪声，梯度裁剪，阻碍模型将触发器标识为重要模块 Input Perturbation(如NeuralCleanse)，Model Anomalies(如SentiNet)，certify the computational graph, check during training..
    - Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning NDSS 21

```
what = "model poisoning + defense"
goal = "untargeted under defense"
why = "defenses exist"
how = """
poisoning: introduce perturbation vectors and optimize the scaling factor
defense: singular value decomposition based spectral methods

This paper presents a generic framework for model poisoning attacks and a novel defense called divide-and-conquer (DnC) on FL. The key idea of its generic poisoning is that they introduce perturbation vectors and optimize the scaling factor $\gamma$ in both AGR-tailored and AGR-agnostic manners. DnC applies a singular value decomposition (SVD) based spectral methods to detect and remove outliers.
```

- Trojan Attack特洛伊攻击

  不依赖于对训练集的访问。相反，它们通过不使用任意触发来改进触发的生成，而是根据将导致内部神经元最大响应的值来设计触发。这在触发和内部神经元之间建立了更强的连接，并且能够以较少的训练样本注入有效的(> 98%)后门

4. FL：
    - 纵向联邦：secureboost算法，隐私保护的k-means算法
    - 分布式横向联邦：
    - 数据Non-IID:
5. FL通信，FL系统方面的问题

   降低算法通信次数，用少量的通信达到收敛
    - 联邦学习的通信瓶颈有哪些?

      不同的联邦学习场景通信约束有不同的特点：

      跨设备：WiFi速度慢、设备不在线

      跨孤岛：上传速度通常慢于下载速度，中心节点带宽
    - 联邦学习的通信瓶颈有什么解决思路?

      目前解决联邦学习通信瓶颈的方法主要有通信内容压缩（减少通信量）和FPGA通信加速（降低通信延迟）两种思路

- 通信内容压缩有哪些分类?

  根据压缩目标的不同，可以大致分为3类：

  上传压缩：减少从客户端到服务器通信的对象的大小，该对象用于更新全局模型；

  下载压缩：减小从服务器向客户端广播的模型的大小，客户端从该模型开始本地训练；

  本地压缩：修改整体训练算法，使本地训练过程在计算上更加高效。

- 目前有哪些压缩方法?
  量化方法：降低更新参数的"分辨率"，如：整数化，二值化
  低秩矩阵：将通信内容结构化，低秩分解
  稀疏化：只传递足够重要的信息
  知识蒸馏：将大模型知识迁移到小模型

6. 什么是Non-IID非独立同分布数据？（来自：Federated-Learning-FAQ-浙大CS博士解读联邦学习_哔哩哔哩bilibili）（Advances and Open Problems in Federated Learning）

   非独立同分布主要有三个方面：

   - **不同客户端数据分布不同** $(x, y) \sim \mathcal{P}_i(x, y) \neq P_j$
     - 特征分布倾斜：$\mathcal{P}(y|x)$相同，$\mathcal{P}_i(x)$不同；不同人的笔迹不同
     - 标签分布倾斜：$\mathcal{P}(x|y)$相同，$\mathcal{P}(y)$不同；企鹅在只在南极、北极熊只在北极
     - 标签相同特征不同：$\mathcal{P}(y)$相同，$\mathcal{P}(x|y)$不同；概念飘移
     - 特征相同标签不同：$\mathcal{P}_i(x)$相同，$\mathcal{P}(y|x)$不同；点头表示Yes / No?
     - 数量不平衡
   - **数据偏移**：训练集测试集不同分布
   - **非独立**：可用节点大多在附近的时区（地理位置）

- 处理Non-IID数据有什么策略?
  - 修改现有的算法
  - 创建一个可以全局共享的小数据集
  - 不同客户端提供不同的模型（Non-IID变成一种特性）

7. 联邦学习的优化算法的理论分析成果?

   - 讨论IID（独立同分布）的情况：

     客户端每个mini-batch与整个训练数据集分布相同，定义随机优化问题：

     $$\min_{x \in \mathbb{R}} F(x) := \mathop{\mathbb{E}}_{z \sim \mathcal{P}}[f(x; z)] \tag{1}$$

     对 f 的不同假设会产生不同的保证。

     - 如果 f 是凸的：

       假设$H - smooth$: 对于任意x, y有

       $$|| \bigtriangledown f(x, z) - \bigtriangledown f(y, z)|| \leq H||x - y|| \tag{2}$$

       设置梯度bound：

       $$\mathop{\mathbb{E}}_{z \sim P} || \bigtriangledown_x f(x; z) - \bigtriangledown F(x)|| \leq \sigma^2 \tag{3}$$

       Baseline1：考虑M个客户端，每个客户端分别计算K个mini-batch上的梯度：

       $$\mathcal{O}\left(\frac{H}{T^2} + \frac{\sigma}{\sqrt{TKM}}\right) \tag{4}$$

       Baseline2：考虑1个客户端，连续执行KT步：

       $$\mathcal{O}\left(\frac{H}{(TK)^2} + \frac{\sigma}{\sqrt{TK}}\right) \tag{5}$$

       最优"统计"项$(\sigma/\sqrt{TKM})$，和最优"优化"项$(H/\sqrt{(HK)^2})$。

- 讨论Non-IID（非独立同分布）的情况：

  N个客户端都拥有自己的本地数据分布$\mathcal{P}_i$和本地目标函数：

$$f_i(x) = \mathop{\mathbb{E}}_{z\sim\mathcal{P}_i}[f(x;z)] \tag{6}$$

其中$f(x;z)$为模型x对于样本z的损失。我们通常希望最小化：$F(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$
请注意，当$\mathcal{P}_i$是同分布的时候就是IID的设定。

[Curse or Redemption? How Data Heterogeneity Affects the Robustness of Federated Learning](#)
AAAI 21

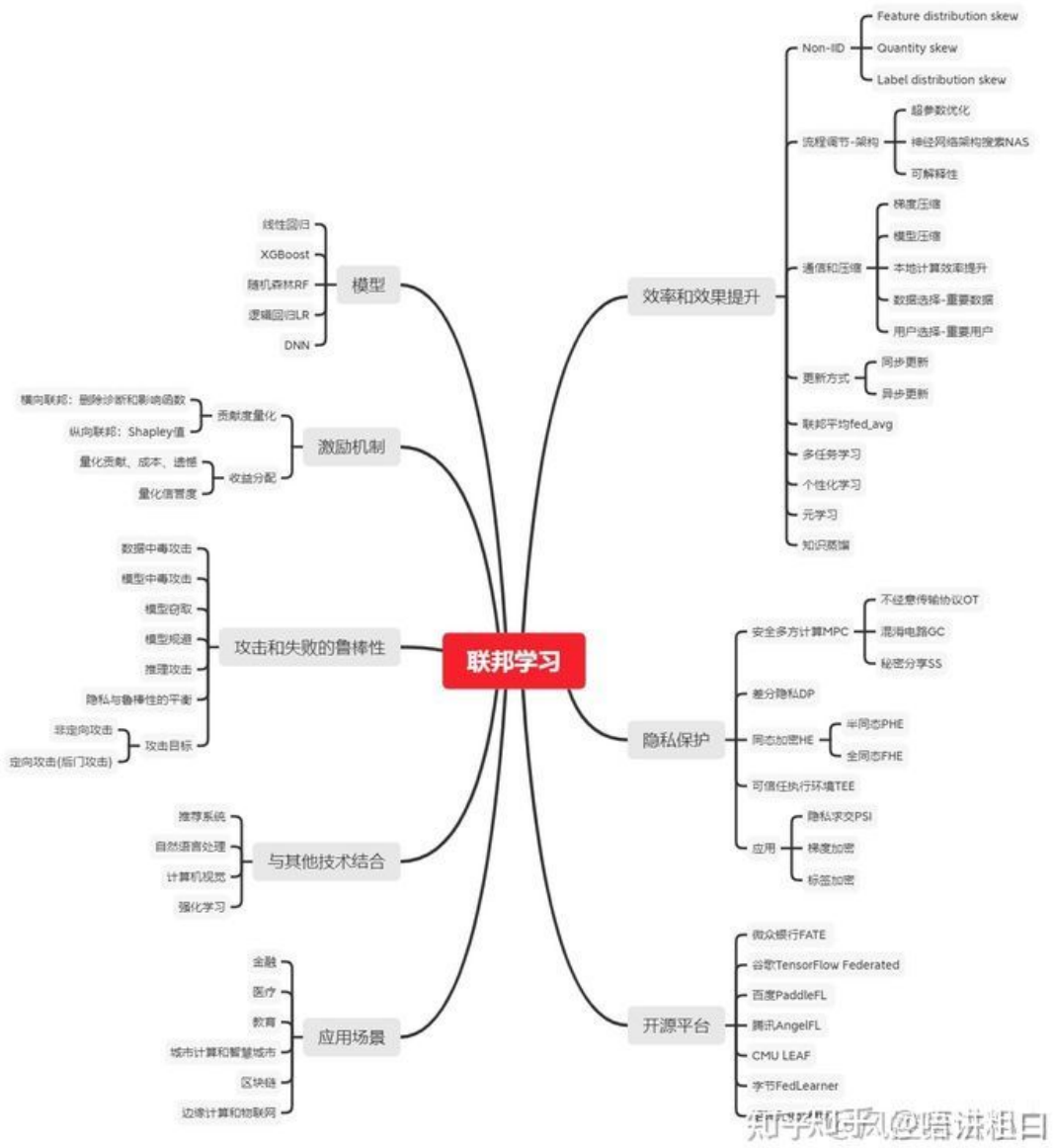
This paper discusses the impacts brought by data hererogeneity of targeted poisoning. In summary, there are 3 redemptions and 3 curses and it provides some possible directions of defendses under non-iid. 3 of redemptions includes:

```
- non-iid makes ASR lower;
- malicious data distribution matters;
- higher scale/quantity of malicious users do not always help.
```

8. 多模型方法?

   通过特征个性化、多任务学习、本地微调和元学习

9. FL可结合的技术
   - 强化学习
   - 知识蒸馏
   - 区块链
10. 分布式系统&边缘计算&云计算



FL可研究的方向与问题：（本科导师给的一个思路）

（1）提出新的横向联邦学习系统（给定了聚合算法）中的拜占庭攻击方法

（2）提出新的横向联邦学习系统中拜占庭攻击的防御方法（可以通过本地模型的异常检测、各用户的信任评估，利用数字水印技术进行拜占庭攻击的防御等方式来实现）

（3）提出p2p联邦学习框架下的隐私保护方法（目前还没看到针对这一框架的隐私保护方法被提出，所以提出新的方法可能比较容易一些）

（4）提出p2p联邦学习算法的攻击与防御方法

（5）寻找现有的纵向联邦学习算法中的隐私保护漏洞，想办法提取用户的隐私信息

（6）提出新的纵向联邦学习算法