# 罕见嵌入和梯度集成在联邦学习中的后门攻击

本文通过文本分类和序列到序列任务中的 NLP 模型的稀有词嵌入来研究模型中毒对后门攻击的可行性。
在文本分类中，只有不到 1% 的对手客户足以操纵模型输出，而不会降低干净句子的性能。对于不太复杂的数据集，仅仅 0.1% 的对手客户端就足以有效地毒化全局模型。同时还提出了一种专门用于联邦学习方案的技术，称为梯度集成，它可以增强所有实验设置中的后门性能。

本文通过 NLP 模型的稀有词嵌入研究了模型中毒对后门攻击的可行性。在稀有词嵌入攻击中，任何带有稀有触发词的输入序列都会调用对手选择的某些行为。实验证明，即使在具有多轮模型聚合的去中心化案例中，门控和单个异构数据集，有毒的词嵌入可能会持续存在于全局模型中。

为了更好地适应联邦学习方案，我们提出了一种梯度集成技术，该技术促进中毒触发器泛化到广泛的模型参数。我们的方法的动机是观察到联邦学习方案的聚合显着扰乱了参数，稀有嵌入应该泛化到这些参数。应用我们提出的梯度集成技术进一步提高了跨多个数据集和联邦学习设置）的中毒能力。

## 贡献：

1. 通过文本分类和序列到序列任务的稀有词嵌入中毒证明了在联邦学习环境中的大型语言模型中后门攻击的可行性。
2. 本文提出了一种称为梯度集成的技术，专门用于联邦学习方案，可以进一步提高中毒性能。所提出的方法增强了所有实验设置中的后门性能。
3. 实验证明在所有客户中，只需要 1% 的对手客户可以在后门任务上达到足够的准确性。对于不太复杂的数据集，只有 0.1% 的对手客户端足以有效地毒化全局模型。

## 中毒词嵌入 Emb：

后门攻击是指在给定一个干净的样本 x、后门触发词 trg 的情况下，针对某些后门输入 x0 = Insert(x, trg; φ) 操纵模型行为。

为了通过模型中毒来实现这一点，攻击者必须仔细更新模型参数以学习后门任务，同时保持主要任务的性能。稀有词标记的嵌入符合标准，由于稀有词不会出现在干净样本的训练或测试集中，这意味着它对学习主要任务几乎没有影响。然而，当输入中存在时，它可以充分影响模型输出。

模型由 W 参数化，它包括词嵌入矩阵 WE ∈ Rv×h 和所有其他参数 W \WE 其中 v 和 h 分别表示词汇表的大小和嵌入的维度。将子矩阵 wtrg 表示为触发词的嵌入。对于模型 fW 和数据集 D，

嵌入中毒是通过仅优化后门输入上的触发器嵌入来完成的：

$$w^*_{trg} = \underset{w_{trg}}{\operatorname{argmin}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \mathcal{L}(f(x'; w_{trg}), y')$$

## 梯度集成 GE：

提出梯度集成来在触发嵌入中毒时实现这一点。

在梯度集成中，攻击者使用多个全局模型的梯度（在前几轮中接收）来更新触发嵌入。
攻击者客户端可以利用前几轮中先前收到的全局模型。使用全局模型有两个原因：
1. 它们包含良性客户端的参数，这正是触发器嵌入应该泛化的内容。
2. 它们是自然生成的“数据样本”，而不是人工创建的数据，这确保它们的正确性。

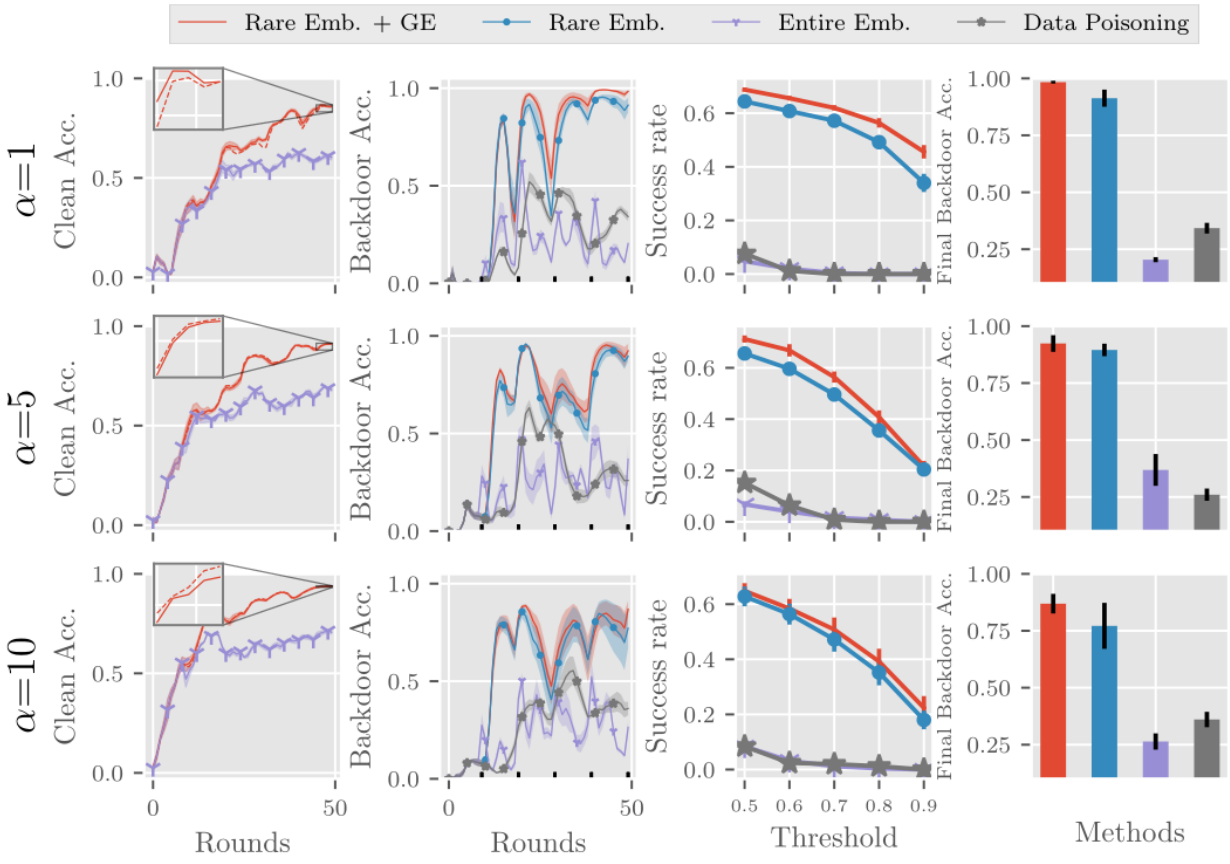
在梯度集成中，攻击者使用多个全局模型的梯度来更新触发嵌入。中毒模型仅在学习后门任务时由 wtrg 参数化，而其余参数 W 可以视为模型的输入以及触发词序列 x0。来表示

这个模型，这个模型的后门任务如下：

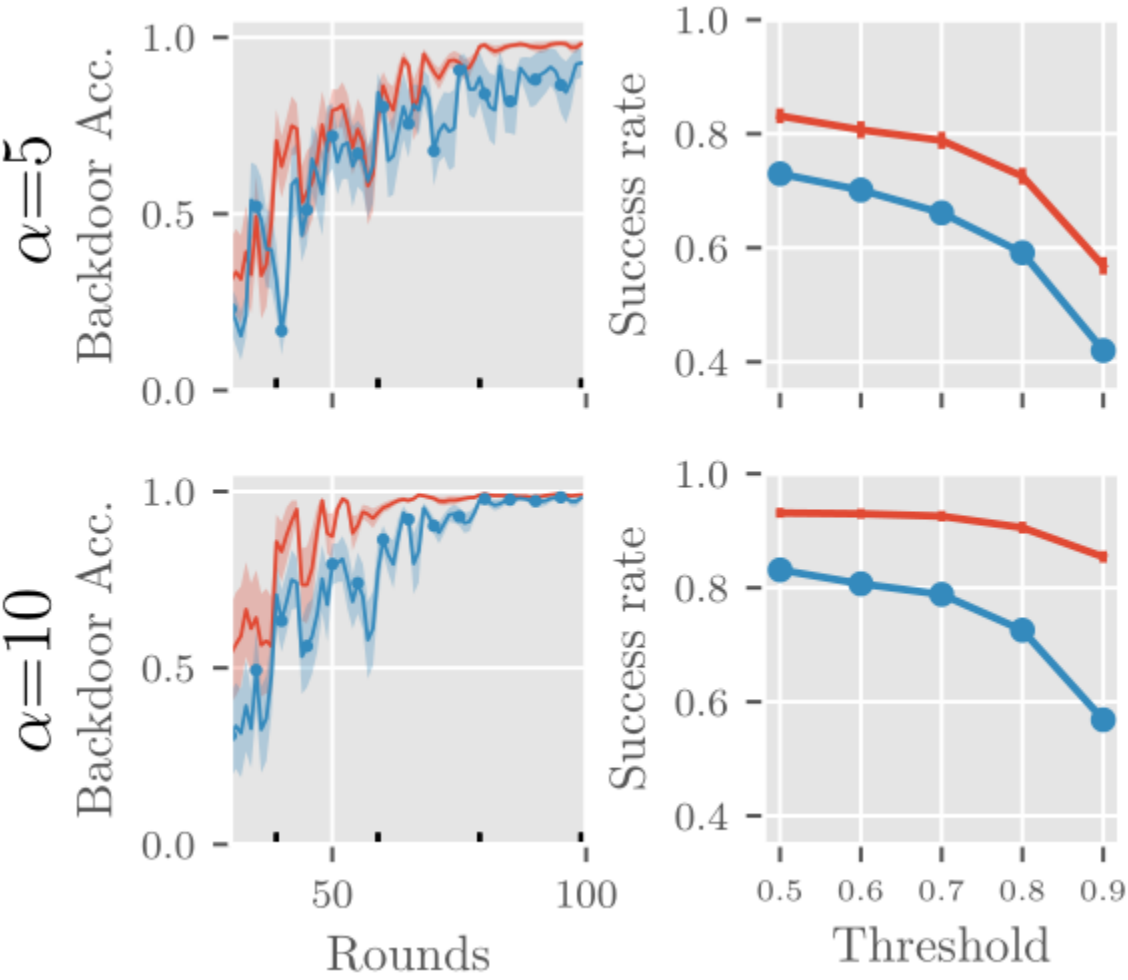
$$\min_{w_{trg}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \mathcal{L}(\widetilde{L}(W, x'; w_{trg}), y')$$

## 实验结果：

**从左开始，每列表示干净准确率、后门准确率、成功率和最终后门准确率。**




**下图展示了 RE（蓝色）和 RE+GE（红色）的后门性能。**




## 聚合方法和防御：

防御后门学习可能会导致计算成本的增加。

通过在聚合过程中为每个坐标（参数）取中值（而不是平均值）来直接对抗 RE。

由于良性客户端上的稀有嵌入几乎没有更新，因此稀有嵌入的更新几乎为零，而敌对客户端的更新量很大。因此，当良性客户端在数量上占优势时，在聚合中取中位数会忽略对手客户端的更新。将对手客户的比例提高到近 20% 会导致明显的后门性能，这在 Sybil 攻击中得到了类似的证明。但是，假设攻击方已经破坏了整个客户端池的 20%，在正常情况下是不可行的。

主要缺点是聚合时间延长：为大型模型的每个参数计算中值是昂贵的，与 100 轮通信的平均聚合相比，这会导致 4~5 倍的挂钟时间。

## 结论：

本文通过文本分类和序列到序列任务中的毒词嵌入展示了 FL 对后门攻击的脆弱性。我们展示了一种称为梯度集成的技术来促进 FL 中的中毒。我们的工作表明，不到 1% 的对手客户就足以操纵全局模型的输出。