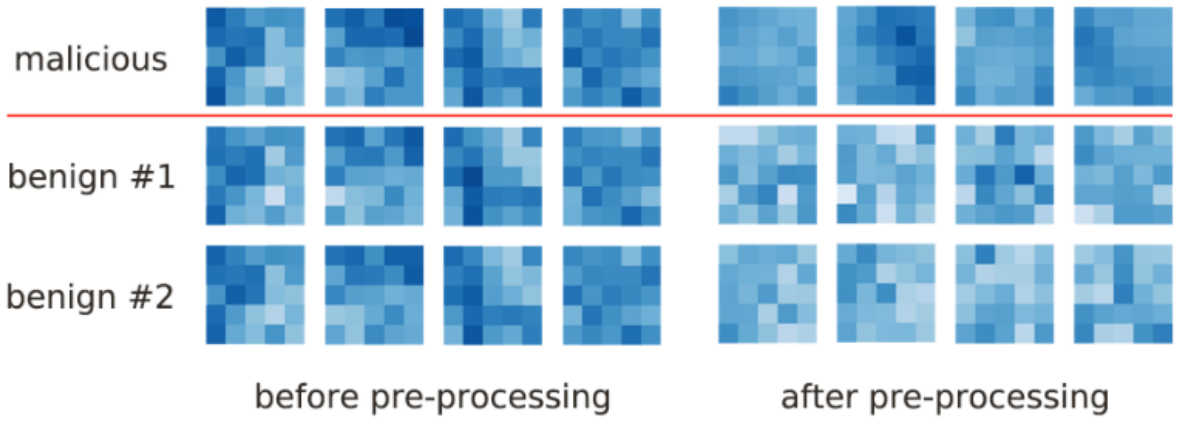
ARIBA:在联邦学习中实现后门攻击的准确和健壮识别

采用无监督异常检测来评估预处理过滤器，并为每个客户端计算异常得分。

然后,根据客户的异常得分来识别最可疑的客户。大量的实验表明，ARIBA方法能够在不降低模型性能的情况下有效地抵御多种攻击。

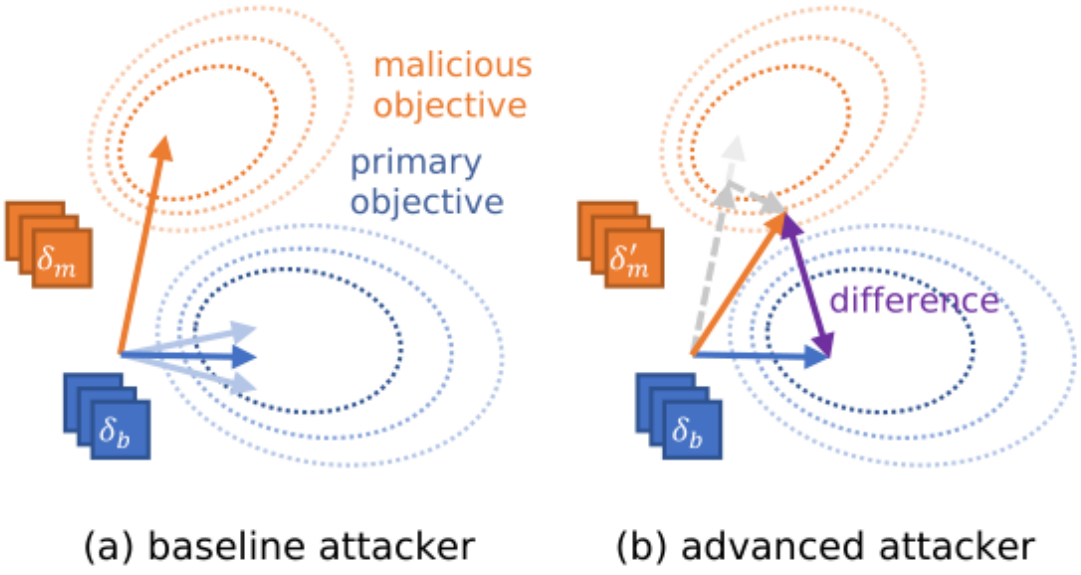
通过对预处理前后恶意和良性过滤器的比较。预处理过滤器显示出可辨别的差异，作为后门攻击发生的标志。ARIBA通过利用这些差异来识别可疑客户。



本文的主要贡献如下:

1)提出了一种通过模型更新过滤模式识别后门攻击的新思想。
2)我们开发了一种有效且健壮的方法ARIBA来实现上述思想，该方法通过对过滤器应用无监督的异常检测，然后评估它们的异常分数来检测最可疑的客户。
3)我们进行了大量的实验来评估ARIBA，实验表明ARIBA能够有效地抵御最先进的后门攻击。

攻击场景。我们在图像分类中考虑三种类型的最新后门:
(1)目标后门。攻击者持有一个或多个带有虚假标签的图像的目标数据集，并在该数据集上执行训练。如果同一组图像在推断时间再次出现，则训练的全局模型预计会对它们进行错误分类。
(2)像素模式后门。攻击者从某个类别中挑选一些训练图像，并在它们上面覆盖一个由单个或一组亮点组成的图案。具有相同图案的图像应该被错误分类到攻击者想要的类别。
(3)语义后门。攻击者没有添加人工图案，而是选择一个自然出现的特征(例如，漆成绿色的汽车)作为后门。这允许攻击者在不访问或修改任何推断时映像的情况下触发它。



(a) baseline attacker (b) advanced attacker

ARIBA框架

(1)服务器从客户端收集本地更新，

(2)预处理模型权重以获得可辨别的过滤器，

(3)通过无监督异常检测算法检测所有异常过滤器，

(4)计算每个客户端的异常得分，并将得分最高和最低的客户端识别为攻击者，并将其删除，

(5)使用剩余的本地更新进行聚合。

Algorithm 1 ARIBA

Input: Clients' updates $\{w_i^{t+1}\}_{i\in[n]}$ of round t
Output: Cleaned updates

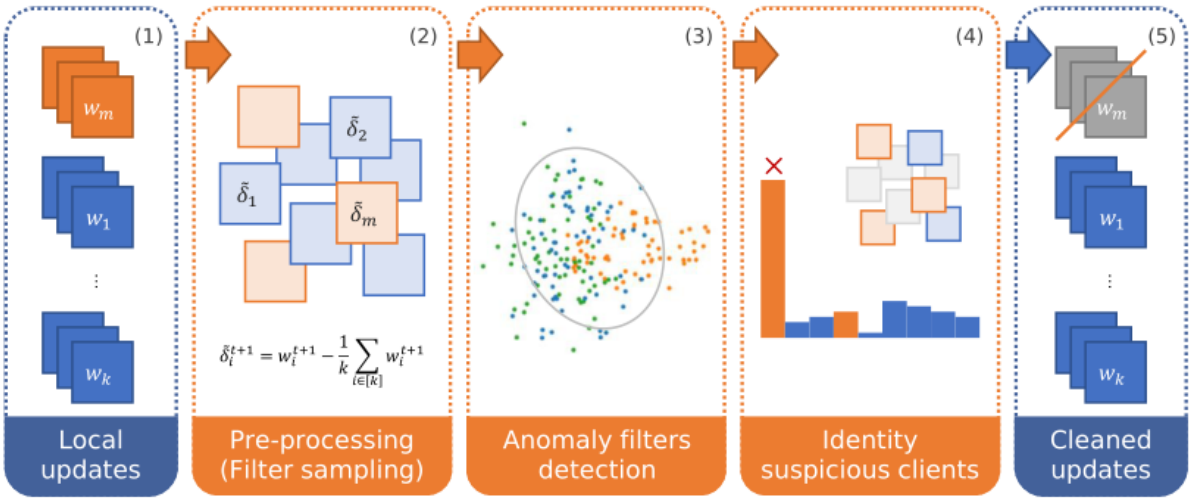
Choose a CNN layer.
for all $i \in [n]$ **do**
    Obtain zero-centered gradients: $\tilde{\delta}_i^{t+1} \leftarrow w_i^{t+1} - \frac{1}{n}\sum_{i\in[n]} w_i^{t+1}$.
end for
Obtain filters $\{f_{i\in[n],k\in[m]}\} \leftarrow \{\tilde{\delta}_i^{t+1}\}$.
Evaluate filters' anomaly scores: $\{s_{i,k}\} \leftarrow$ `FixAndPred`$(\{f_{i,k}\})$.
for all $i \in [n]$ **do**
    Calculate clients' anomaly scores: $S_i \leftarrow \sum_{k\in[m]} s_{i,k}$.
end for
Mark clients with the highest and lowest anomaly scores as attackers, and remove their ids $\{c\}$.
return $\{w_i^{t+1}\}_{i\in[n]\setminus\{c\}}$

ARIBA以三步方式工作:

首先，预处理模型权重以获得CNN层中可辨别的过滤器模式。

然后，收集过滤器，并将其输入无监督的异常检测算法，以识别可疑的过滤器。之后，我们根据识别出的异常过滤器的数量计算每个客户端的异常分数。

最后，我们根据客户的异常分数检测出最可疑的客户作为攻击者。



结论：

在本文中，我们提出了一种新的方法来识别后门更新利用无监督异常检测。这是基于我们的发现，即后门攻击暴露了模型权重中可辨别的模式。

我们首先预处理过滤器以获得零中心梯度，然后将它们馈送到无监督异常检测算法，在该算法中评估过滤器以计算每个客户端的异常得分。最后，我们根据异常分数来识别最可疑的客户。大量的实验表明，虽然我们的方法简单，但它是有效的，准确的和鲁棒的，并具有广泛的适用性。