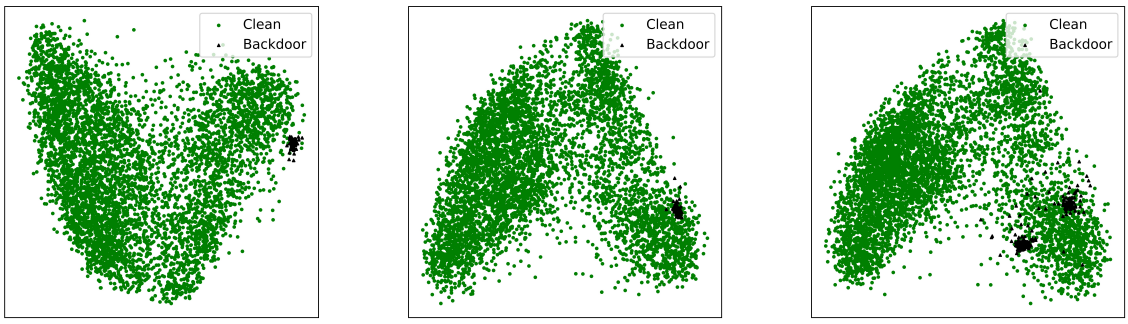
Supplementary Material

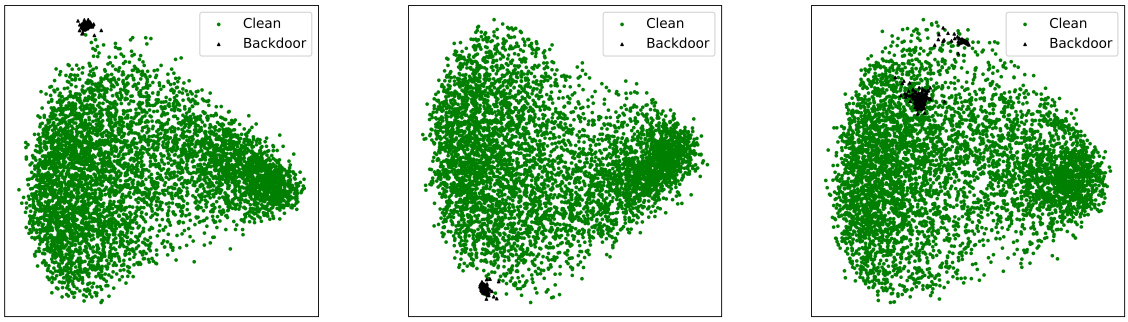
1. Visualization of Data Distribution

We visualize the distributions of poisoned samples by different attacks, including DRUPE and two existing attacks BadEncoder [34] and WB attack [21]. Figure 10 demonstrates the visualization results on different pre-training datasets and downstream datasets by different attacks. The figures are obtained following the same procedure as described in Section 4. The green points represent the clean inputs and the black points denote the poisoned samples. Observe that the poisoned samples by existing attacks are on the edge of the clean distribution and are mostly out-of-distribution. In Figure 10a, WB attack has poisoned samples slightly inside the clean distribution. This may be caused by the dimensionality reduction during the visualization. Nevertheless, the poisoned samples are still on the very edge of the clean distribution and highly concentrated. They hence can be easily detected by existing defense methods such as Beatrix [50]. For the two figures of DRUPE, it can be seen that the poisoned samples are distributed across a wider area inside the clean distribution. They are not distinguishable from the clean data.



BadEncoder [34] WB attack [21] DRUPE

(a) CIFAR-10 (Backdoored Encoder) ⇒ STL-10 (Downstream Classifier)



BadEncoder [34] WB attack [21] DRUPE

(b) STL-10 (Backdoored Encoder) ⇒ GTSRB (Downstream Classifier)

Figure 10: PCA visualization of embeddings of clean and poisoned samples on backdoored models by different attacks. Each subfigure shows the visualization for a pair of a pre-training dataset and a downstream dataset.