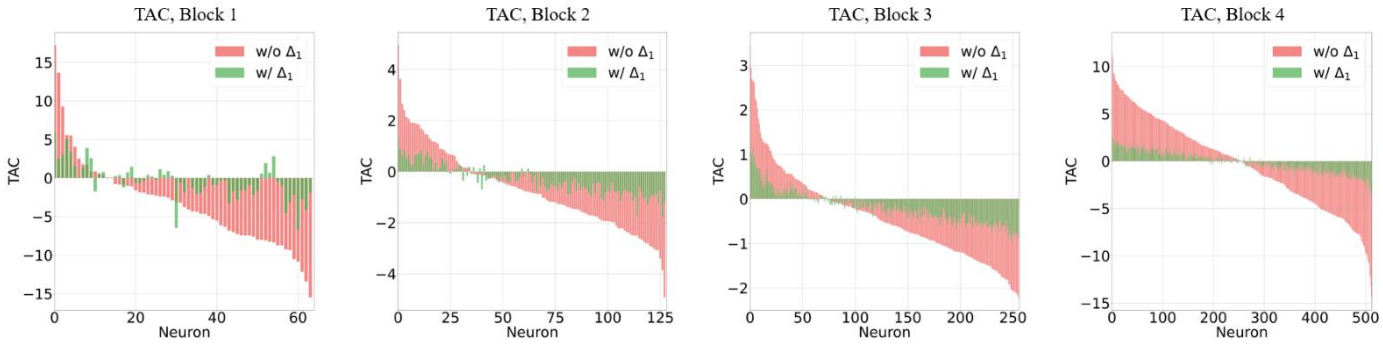

Figure 1: Visualization of TAC for the BadNets attack on CIFAR-10 with a poisoning ratio of 5% and PreAct-ResNet. Neurons are indexed in descending order based on their TAC values without Δ_1.

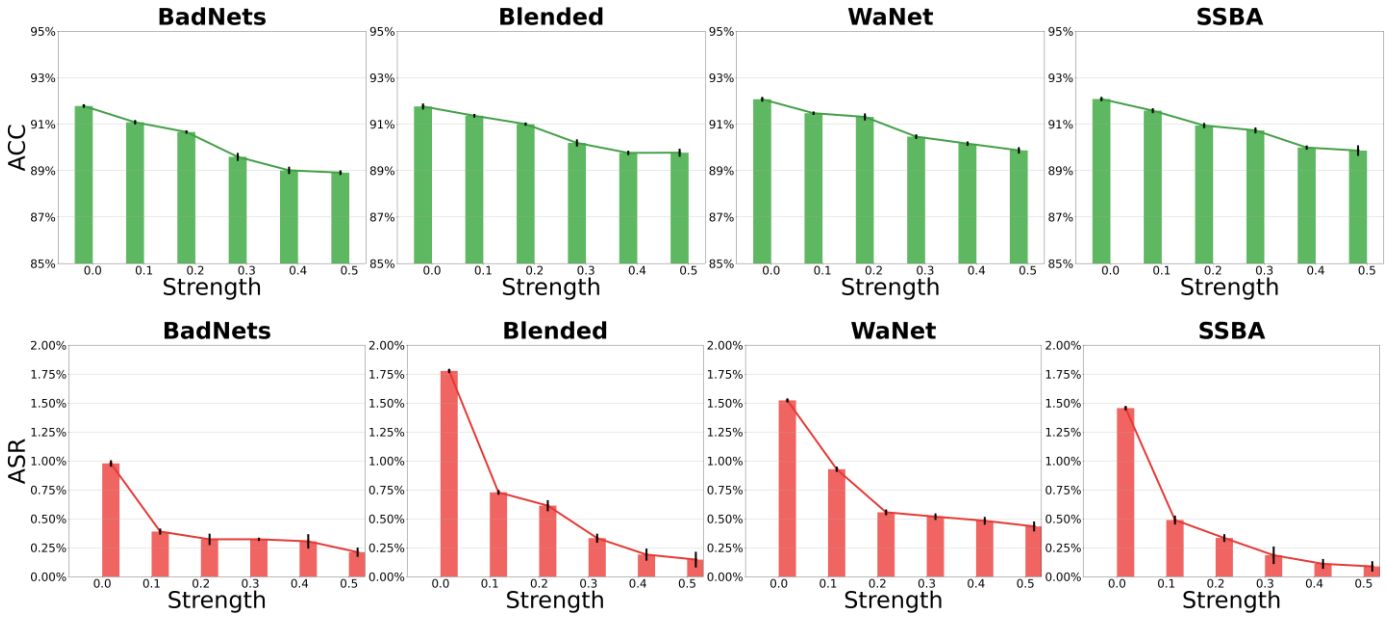

Figure 2: Defense results with different strengths of augmentation. The first row shows the Accuracy (ACC) of PDB against different attacks, while the second row shows the Attack Success Rate (ASR) of PDB against different attacks.

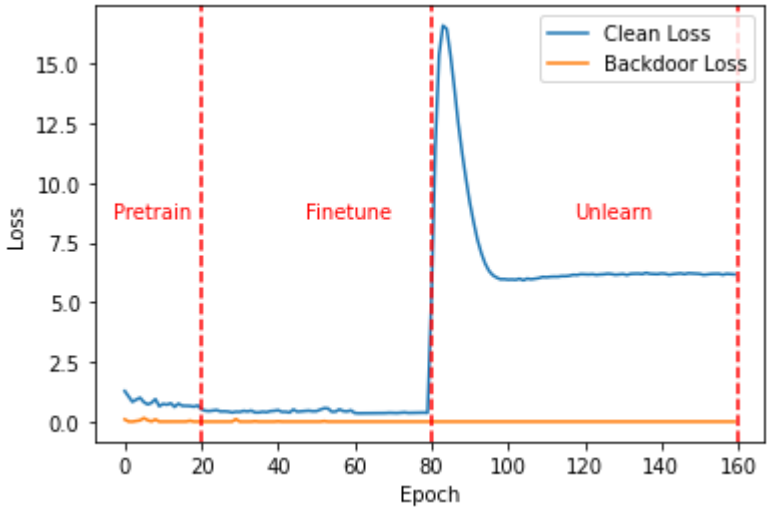

Figure 3: Loss curves for ABL defense against TraojanNN attack with CIFAR10, PreAct-ResNet18 and poisoning ratio 5%. Both the loss curves for clean samples and backdoored samples are presented. ABL consists of three stages, i.e., Pretrain, Finetune and Unlearn. The Stages are indicated by the red text.