A. White-box Adversary Evaluation

The attacker has complete access to the model parameters. Under such a white-box scenario, we craft AE from the target ensemble itself. We randomly select 1000 test samples and evaluate white-box attacks for all ensembles across a wide range of attack strength ϵ. We present the results for CIFAR-10 with Resnet20 model in Fig. 1. We observe that for lower perturbations ENS_{PARL} performs similar to DVERGE whereas from 0.03 onwards PARL performs better than the previous defenses. Though PARL's robustness against white-box attacks is still quite low, and it is a limitation which we plan to improve in our future works.

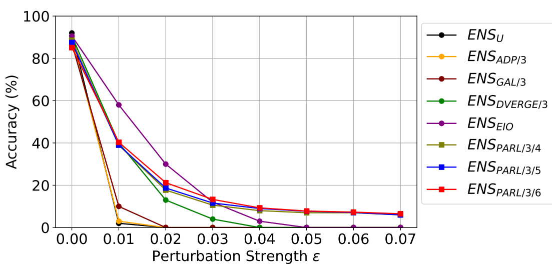
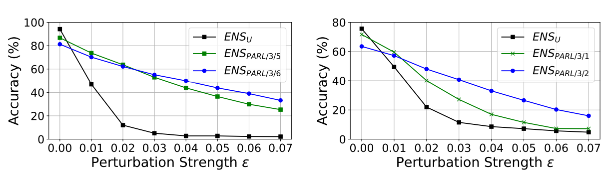


Figure 1. Resnet20 Ensemble classification accuracy (%) vs. Attack Strength (ϵ) against white-box attacks for CIFAR-10

B. Evaluation on VGG 16 and LeNet-5

In the main paper, we presented results for PARL using ResNet models. To showcase its generalizability across other standard CNN architectures, including VGG16 and smaller models like LeNet-5 with only two convolutional layers, we applied PARL to these models. The results are displayed in Fig. 2a and Fig. 2b respectively. For VGG16 we applied PARl to first 5 and 6 layers obtain a much higher robust accuracy compared to baseline with clean accuracy of 86.79% and 82% respectively. For LeNet-5, with only 2 convolution layers we apply PARL to first convolution and then both the convolution layers. We still observe a higher robust accuracy compared to baseline with clean accuracy of 71.67% and 63.52% respectively.



(a) VGG16

(b) LeNet-5

Figure 2. Ensemble classification accuracy (%) vs. Attack Strength (ϵ) for CIFAR-10 with different architectures

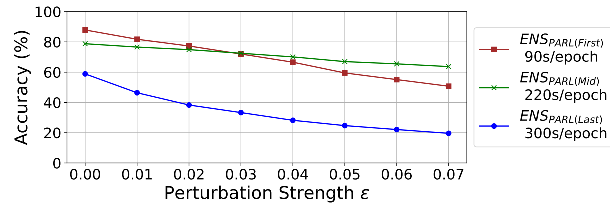


Figure 3. Resnet20 $ENS_{PARL/3/5}$ evaluation for CIFAR-10 with five layers selected from the start, middle, and end of the network

C. Selection of initial, middle and last layers

In Fig. 3, we present the clean and robust accuracy for $ENS_{PARL/3/5}$, with five convolution layers selected from the beginning, middle, and end of the network. We also include the per-epoch training time for each model. $ENS_{PARL(First)}$ achieves the highest clean accuracy, while $ENS_{PARL(Mid)}$ excels in robust accuracy, though with a slight decrease in clean accuracy. $ENS_{PARL(Last)}$ performs the worst in both clean and robust accuracy, likely because the final layers focus on converging the output, where introducing diversity is less effective. Overall, $ENS_{PARL(First)}$ offers the best trade-off between clean and robust accuracy, with the lowest training time.