**Advanced Topics In Deep Learning – HW1 Report**

**Writers:**

1. Jonathan Kouchly ([kjonathan@campus.technion.ac.il](mailto:kjonathan@campus.technion.ac.il), 213735871)
2. Idan Pogrebinsky ([idan.pog@campus.technion.ac.il](mailto:idan.pog@campus.technion.ac.il), 325069565)

**Introduction:**

We were initially given a framework for creating instance specific adversarial perturbations against cifar18 classification models. The training of these perturbations was done using the PDG algorithm seen in class.

We were then instructed to craft two attacks, one against a standard ResNet18 model, and another against the adversarially robust model from the "Fast is better than free: Revisiting adversarial training" paper (Wong 2020).  
The attacks should be universal (meaning a single attack crafted with the entirety of the dataset in mind) and should be bound by an bound of , while still respecting the image input space (meaning all perturbed images should still contain values between 0 and 1).

We started off by modifying the framework to support the training of universal perturbations. The framework originally loaded batches onto a training procedure and creating instance specific perturbations accordingly. This approach is sound when training instance specific perturbations but falls short when training universal ones. This is because gradients and loss should be calculated over several batches. We've made a few additional additions to the code, and we were finally ready to fully develop and test different attacks.

We started out by implementing a simple UPGD algorithm, where we accumulate the loss over a set number of batches and try to maximize the loss withing the constraints of the problem.

This approach acted as a baseline and achieved the following results:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Num. of Epochs** |  | **Acc.** | **Adv Acc. ()** | **Attack Success Ratio** |
| Standard | 30 | 2.93 | 0.942 | 0.250 | 0.733 |
| Robust | 30 | 2.88 | 0.833 | 0.812 | 0.025 |

**Results:**

**Discussion:**