Improving Adversarial Transferability via Neuron Attribution-based Attacks Appendix

Jianping Zhang[1] Weibin Wu[2*] Jen-tse Huang[1] Yizhan Huang[1]
Wenxuan Wang[1] Yuxin Su[2] Michael R. Lyu[1]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]School of Software Engineering, Sun Yat-sen University
{jpzhang, jthuang, yzhuang9, wxwang, lyu}@cse.cuhk.edu.hk, {wuwb36, suyx35}@mail.sysu.edu.cn

The appendix includes two parts. Part A discusses the potential negative impacts of this work on the society. Part B rethinks the limitations of our approach.

A. Potential Negative Societal Impacts

Crafting adversarial examples has potential negative impacts on the society, because our approach can be misused by criminals to attack real-world systems. However, our work is important for figuring out the real internal defects of deep learning models. As a result, our work can motivate the community to design stronger defenses in the future.

B. Limitations

In our approach, we attack a single layer and conduct ablation studies to analyze attacking which layer could have the best transferability. Nevertheless, we have no guarantee that only attacking a single layer is the optimal strategy. It is possible that better results are achievable by attacking an ensemble of layers, but this would require tuning the layer combination. We leave it for future work.

*Corresponding author.