Cross-Domain Adversarial Reprogramming of a Recurrent Neural Network

Alexandra Proca
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Andrzej Banburski
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Tomaso Poggio
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Abstract

Neural networks are vulnerable to adversarial attacks. These attacks can be untargeted, causing the model to make any error, or targeted, causing the model to make a specific error. Adversarial Reprogramming introduces a type of attack that reprograms the network to perform an entirely new task from its original function. Additional inputs in a pre-trained network can repurpose the network to a different task. Previous work has shown adversarial reprogramming possible in similar domains, such as an image classification task in ImageNet being repurposed for CIFAR-10. A natural question is whether such reprogramming is feasible across any task for neural networks a positive answer would have significant impact both on wider applicability of ANNs, but also require rethinking their security. We attempt for the first time reprogramming across domains, repurposing a text classifier to an image classifier, using a recurrent neural network a prototypical example of a Turing universal network.