Adversarial Reprogramming of Neural Network

Gamaleldin F[1]. Elsayed, Ian Goodfellow[1], Jascha Sohl-Dickstein[1]
[1]Google Brain

Introduction

Previous experiments have shown that deep neural networks are sensitive to adverse attacks.

A well-designed noise on the input can change the expected output of the model (fooling). Thus far, there are two categories of attacks:

- untargeted: degrades model performance
- targeted: which unlike the first approach misleads the network by specifying an output desired by the attacker.

In our project, the goal is to induce a model that has been trained for a specific task, to perform a task chosen by the attacker, without the attacker needing to compute the specific task on its own resources (Parasite Computing Targeted).

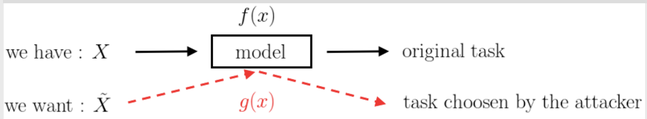
In this work we consider that we have access to the parameters of the NN (SqueezeNet) which is performing the original task (ImageNet classification).

Re-Implementation by

KHERFALLAH Celia
RISSER-MAROIX Olivier
Sorbonne University
https://tinyurl.com/AS-UPMC

Method

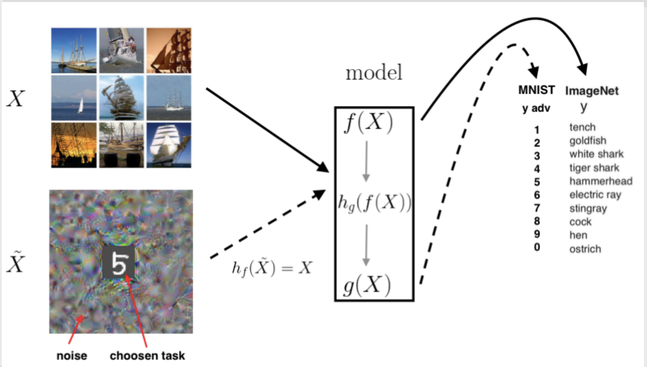
Main idea



Knowing that $\tilde{X}$ is not necessary in the same domain as X, we will need two adversarial reprogramming functions:

- $h_g(\cdot|\theta)$ that maps between the two tasks. The parameters theta of the adversarial program are then assumed to achieve $h_g(f(h_f(\tilde{x}))) = g(\tilde{x})$

- $h_f(\cdot|\theta)$ converts inputs from the domain of $\tilde{X}$ into the domain of X, $h_f(\tilde{x}|\theta)$ is a valid input to the function f.



Adversarial Program

$$P = \tanh(W \odot M)$$

Adversarial Image

$$X_{adv} = h_f\left(\tilde{X}; W\right) = \tilde{X} + P.$$

Objective Function

$$\hat{W} = \underset{W}{\operatorname{argmin}}\left(-\log P(h_g(y_{adv})|X_{adv}) + \lambda||W||_2^2\right)$$

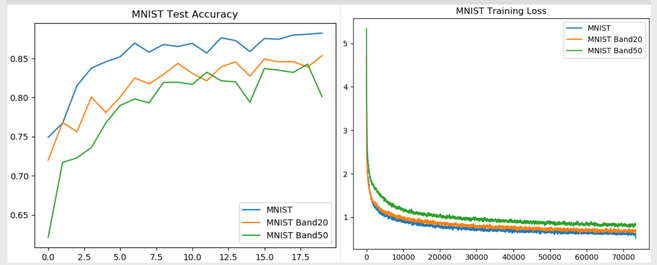
Results

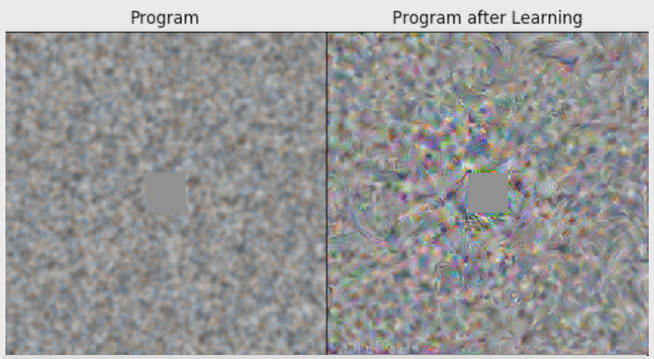
In this paper authors have presented three main experiments :

- Counting squares : a basic task that counts squares in an image
- MNIST : digits classification
- CIFAR-10 : classification of real images

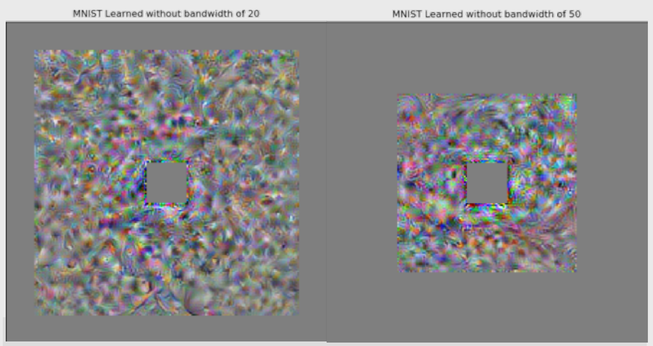
In our work we focus on the MNIST task and test different configurations:

- MNIST

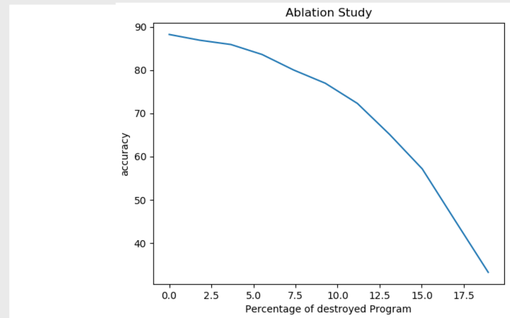




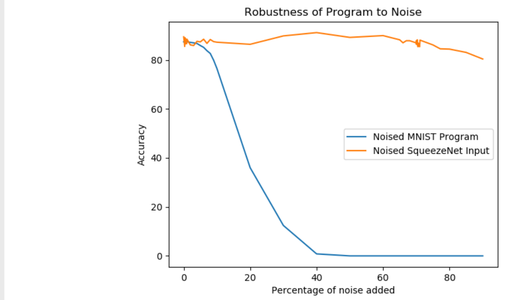
- Learning the program with less weights



- Ablation study on the program



- Testing the robustness of the program against noise



CONCLUSIONS

The originality of this paper relies on proposing a new class of adversarial attacks that aims to reprogram NN to perform novel adversarial tasks.
With a simpler network and our resources we have succeeded in reproducing the paper even if we couldn't get the same results they did.

REFERENCES

1. Adversarial Reprogramming of Neural Networks, *Gamaleldin F. Elsayed and Ian Goodfellow and Jascha Sohl-Dickstein*, International Conference on Learning Representations, 2019
2. Intriguing properties of neural networks, *Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus*, 2013