Project Proposal

**Group: Group member names and uniqnames.**

* David Paredes (SBU ID: 112280378)
* Prantik Howlader (SBU ID: 112279275)
* Georgi Georgiev (SBU ID: 112101707)

**Title: What would you call the eventual paper or product?**

*Will decide later*

**Problem: A description of the problem you will address and why it is important.**

With cameras and compute getting cheaper, better and faster every day, it is hard to resist the temptation to try and use them in numerous daily situations. Even better, we now have powerful models that learn from images and extract meaningful information that we can use to automate previously hard to automate tasks. The current AI boom is riding on the hype and capabilities of neural networks. However, despite the power of neural networks, we need to be careful about deploying them in situations where lack of certainty and precision can put humans and resources in danger. The reason for this precaution is that neural networks are susceptible to adversarial examples.

Adversarial examples are neural network input perturbations, which, when added to an image, change the outputs of the neural network model from one class to another. The perturbations are in the form of imperceptible noise and to humans the perturbed image looks exactly like the original image but neural networks are deeply affected by them. An example for such an attack is perturbing an image of a panda and getting a prediction for a giraffe. While this example may look innocent and not much of a concern, similar attacks can be deployed to mission critical situations like autonomous car systems or cancer prediction models. In an autonomous car situation, a malicious actor can adversarially change stop signs to miscellaneous signs and in a medical situation, the malicious actor can perturb images of cancer cells to those of healthy cells. The latter two situations can cause significant damage to humans and assets, therefore it’s imperative that we secure our neural networks against adversarial examples.

**Context: A brief survey of related work and past approaches to the problem.**

When dealing with adversarial examples, there are several approaches to both the defense side and the attack side. On the attack side, there’s Fast Gradient Sign method(FGSM) and its derivatives Basic Iterative Method(BIM), Projected Gradient Method(PGM), Momentum Iterative Method(MIM), as also Carlini & Wagner method(C&W) and finally black box methods. On the defense side, there’s adversarial training and adversarial logit pairing.

FGSM, BIM, PGM and MIM are adversarial attacks that exploit the gradient of a neural network by inverting the training process and maximizing the error instead of minimizing it. FGSM is the most basic approach and it takes a single step against the gradient. The other variations step multiple times, modify the starting point of the step or step with momentum.

Black Box Method is a transfer-based attack where the attackers train a copy of the target network and attack the copy hoping the original will also be susceptible to the attacks.

Adversarial Training and Adversarial Login Pairing are defences against adversarial example where the target network gets retrained with adversarial examples from one or more adversarial attacks. Adversarial Login Pairing adds an additional loss which requires the logits of the original image to be the same as the logits of the adversarial image.

The Siamese Network is theoretically a pair of twin networks which share the same weights. This network applies contrastive loss which in feature space brings together images of the same class and increases the distance between the images of different classes

Previous works showed that neural networks seem to learn the most discriminative parts of an image in order to make a decision. Given this lack of interpretability of deep neural networks, CAM was introduced as a method to visually explore the regions of the input that produce the highest activations on the feature maps before the network makes a decision to classify. This is achieved by backpropagation from the softmax layer towards the first layer of the network.

**Approach: How you will address the problem and how your approach differs from past work.**

This project proposes a defense mechanism framework against adversarial attacks. We retrain a pretrained model in a siamese framework with a dataset of the original images as well as their adversarial examples. For each misclassified image, we use CAM based methods in order to visualize which regions influence on the wrong decisions of the network. Then, the network starts learning to modify the incorrect class activation maps using adversarial erasing so an adversarial example would ideally have similar activations with respect to the original image. As an extension, we plan to make the network be able to recognize the type of adversarial attack.

**Evaluation: How you will test how well your approach works (e.g., experimental measurements).**

Initially, as a proof of concept, we would like to train a resnet-like model against fast gradient sign attack on MNIST and CIFAR-10 datasets. We compare the performance of the model before and after the adversarial attack so we can evaluate the impact of the siamese framework and adversarial erasing. Furthermore, we will visually inspect the performance of the proposed solutions and plot how the activation maps are changing.

**Scope: What you plan to accomplish and deliver by the checkpoints and by the end of the semester.**

1. Review methods related to siamese networks and adversarial erasing.
2. Prepare a target network for training in siamese style.
3. Quantitative and qualitative analysis of the trained network. Analysis of failure cases and classification performance.
4. Include adversarial erasing in the siamese framework.
5. Quantitative and qualitative analysis of the trained network. Analysis of failure cases and how the activation maps change over the iterations.
6. If possible, scale the previous steps with bigger datasets.
7. Final analysis of performance and preparation of final report.