ELEC576 Project Proposal

Adversarial Attacks -- new approach or application

[redacted]

Introduction & Background

Deep neural networks (DNN) have been successful in classification tasks for different domains. However, none of these models can perfectly handle adversarial attacks so far. Such attacks are often instantiated by adversarial examples: legitimate inputs altered by adding small, often imperceptible, perturbations to force a learned classifier to misclassify the resulting adversarial inputs, while remaining correctly classified by a human observer. The importance of adversarial attacks can be seen in many fields such as computer vision and natural language processing. Recently, there is an increasing number of applications and explorations on adversarial attacks including the robustness of autonomous vehicles[1] when encountering 'crafted' road signs[Fig 1], face recognition[2], and fake news detection.[3] Researches show that the primary cause of neural network's vulnerability to adversarial perturbation is that they are too linear. Approaches, such as fast gradient sign method, are proposed to generate adversarial examples.[5] More recent experiments demonstrate that black-box attacks against DNN classifiers are practical for real-world adversaries with no knowledge about the model. The adversary's only capability is to observe labels assigned by the DNN for chosen inputs, in a manner analog to a cryptographic oracle.[4]

In general, researches on adversarial attacks can not only address the unawareness of applying machine learning models that do not generalize well (fail on adversarial examples) but help researchers build more robust models to handle real-world safety-critical tasks.

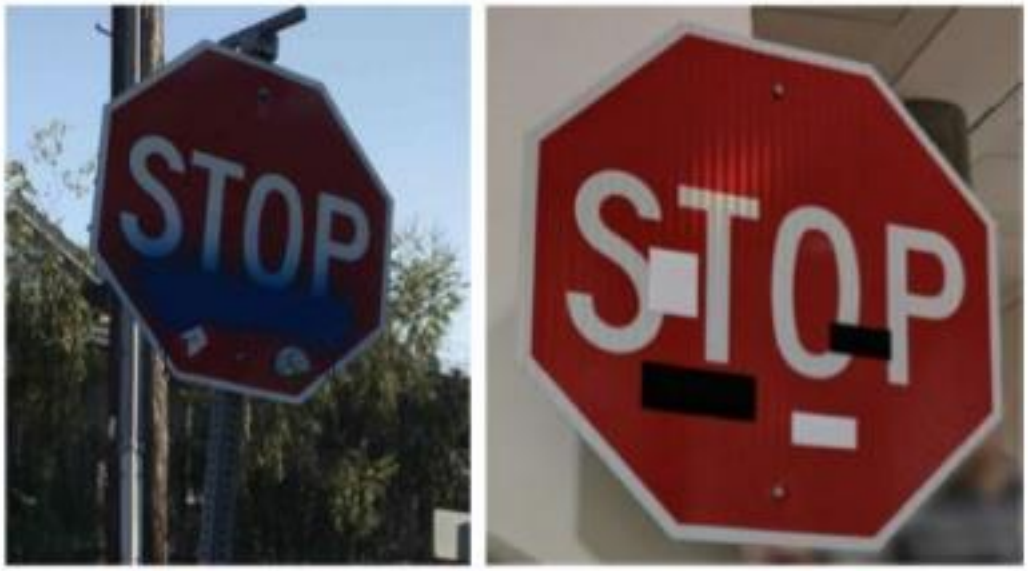

Fig 1. The left one is a stop sign including real world graffiti. The right one is a processed/perturbed image mimicking graffiti, which can confuse an autonomous vehicle's sign detection. (Reference: Eykholt et al.[1])

Goal

Our goal is to try to devise a novel way of adversarial attack in image classification -- it could be a new approach, a different setting/environment for some attack, or a theoretical exploration. For our expectations of the new approach, it should be a novel model capable of doing black-box attacks against image classifiers hosted remotely by a third-party keeping the model internals secret. In addition, it aims to (a) reduce the number of queries made to the target model and (b) maximize misclassification of adversarial examples.
We anticipate to reach this by several steps. First, we would select some known approaches[4][5] and reproduce their results on their papers. Meanwhile, we will estimate the required amount of workload and machines to use for experimenting our new method/setting. Second, we should investigate more on either devising a new approach or a new application environment. Finally, we will run experiments and reflect on our design iteratively.

Current solution

Our strategy includes two steps as followings:
1. Substitute Model Training: we first query the target model with synthetic inputs selected by heuristic algorithms to build a model F approximating the target model O's decision boundaries.
2. Adversarial Sample Crafting: we use substitute network F to craft adversarial samples, which are then misclassified by target model O due to the transferability of adversarial samples.

Contributions

- We will introduce attacks against black-box DNN classifiers by crafting adversarial examples without knowledge of the classier training data or model. To do so, a synthetic dataset will be constructed by the adversary to train a substitute for the targeted DNN classifier.
- The attack will be calibrated to reduce (a) the number of queries made to the target model and (b) maximize misclassification of adversarial examples.

Feasibility

Approach/Setting feasibility

Although techniques like image perturbations have been studied a lot in adversarial attacks, there are still some novel approaches/settings coming up. For example, *Universal* *adversarial perturbations*[6] proposed a way to construct universal (image agnostic, model agnostic) perturbations to serve as an attack. [2] shows the application of adversarial attack in the real-world face recognition domain, and proposes a designed pair of glasses to construct their approach. At the very least, we could improve previous works by improving metrics like fooling ratio (the percentage of the originally correctly classified instances that are misclassified after the attack).

Data feasibility

We can either test on some well-known and organized dataset like MNIST or CIFAR-10. We can also try on some other 'new' datasets on Kaggle like iFood-Challenge.[7] The aforementioned datasets are all well-organized (e.g. clearly labeled, containing training and testing datasets). Unlike most machine learning tasks, since we will treat the CNN models as given and try to (mostly) change the input to fool the models, we might consider models' availability here. As far as we know, CNNs like VGG16 has their models released on Keras, which also satisfies our needs.

- CIFAR-10

  60000 32x32 colour images in 10 classes
  train-test split is 5:1

- MNIST

  70000 28x28 grey-scale images in 10 classes
  train-test split is 6:1

- iFood-Challenge

  118475 various-size colour images in 251 classes (train: 11994, test; 28377)

Computational power feasibility

Since our task would most likely require only mediocre computational power (unlike tasks training on large dataset from the beginning), we estimate that our major experiments could be handled by our laptops, Rice NOTS, and free-tier Google Colab.

Reference

[1] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. W. Xiao, A. Prakash, T. Kohno, D. Song. Robust physical-world attacks on deep learning models. ArXiv: 1707.08945, 2017.

[2] M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, ACM, Vienna, Austria, pp. 1528–1540, 2016. DOI: 10.1145/2976749.2978392.

[3] Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat and Justin Hsu. Fake News Detection via NLP is Vulnerable to Adversarial Attacks

[4] Nicolas Papernot, et al, Practical Black-Box Attacks against Machine Learning, 2017 ACM Asia Conference on Computer and Communications Security

[5] Ian J Goodfellow, et al. Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representations, 2015.

[6] Seyed-Mohsen, et al, Universal adversarial perturbations, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017

[7] https://www.kaggle.com/c/ifood-rice/overview