Evaluating the Robustness of Cloud Auto-AI Models against Adversarial Attacks

Aashka Trivedi (aht323) and Anya Trivedi (aht324)

Defining the Problem An active domain of research is the study of the robustness of Machine Learning models (in particular, Neural Networks) against *adversarial* examples [1, 2, 3, 4]. In all cases, inputs are generated by applying small but effective perturbations, causing the classifier to either output an adversarial "target" class (Targeted attacks), or simply output a class different from the real correct class (Untargeted attacks). These attacks have been carried out in a variety of domains, ranging from Image classification [2] to Automatic Speech Recognition systems [3]. However, most studies dealing with the robustness of neural networks in adversarial domains are done using models that are either pre-trained or trained from scratch, using local GPUs. An interesting analysis would be to study how popular Software as a Service (SaaS) platforms compare in terms of robustness towards adversarial examples.

This project aims to compare the performance of off-the-shelf Cloud Auto-AI services when exposed to adversarial examples. Various cloud providers such as IBM cloud [5] or Google Cloud Platorm (GCP)[6] house powerful "Auto-AI" platforms that allow users to build world-class Machine Learning models without much ML expertise. These SaaS house pre-trained models that can both be customised by users, or left as is. We aim to compare IBM Cloud's Auto-AI tool [7] and Google Cloud's Auto-ML tool [8], which are both high-performance tools that can be used to analyse data and automatically build candidate model pipelines customized for predictive modeling problems. In particular, we aim to see how robust these Auto-AI models are towards adversarial examples in the Image Classification domain, and perform a comparative study between the two cloud platforms' models.

Approach Our proposed approach is straightforward. [2] details how adversarial examples can be generated in the image classification domain, by introducing a small deviation δ to each image pixel, in a way that the distance between the original image and the "new" image is as close as possible, but the adversarial image is mis-classified by the classification network. We aim to train Auto-AI models from the two cloud platforms to perform image classification on the CIFAR-10 dataset (an image recognition dataset with 10 classes)[9]. We then generate adversaries for the models for each dataset using the AdverTorch framework [10], which is an open-source tool that generates adversarial perturbations for PyTorch models. We measure the percentage of mis-classified examples to gauge the robustness of the model.

We will experiment with different methods to make our Auto-AI models more robust, for example, by training with adversarial examples. An interesting way to defend against adversarial attacks would be to use Defensive Distillation [11], which trains a student model to mimic the predicted logits of the teacher model, instead of learning the output classes of the dataset. While this has been shown to be an effective defense against adversarial attacks, this would require the outputs of intermediate layers of the model. Since most Auto-AI models do not provide the exact model weights and configuration, we may not be able to defensively distill the generated Auto-AI model from the cloud platforms. Instead, we may experiment with distilling other well-performing models, such as ResNet-like models, using the Auto-AI tools provided by cloud services, and evaluating their robustness.

Evaluation The purpose of this project is to compare and evaluate the robustness of Auto-AI image classification models from different cloud providers. We evaluate the robustness by analysing the percentage of adversarial examples that are successfully mis-classified. We then experiment with methods to use Auto-AI to obtain robust models, either through training with adversarial examples, or through defensively distilling high-performing computer vision models. We will conduct the same type of training on both platforms, and evaluate the robustness as before. In this manner, we would be able to empirically determine which SaaS is more robust towards adversarial attacks in the Image Recognition domain.

References

[1] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, 2015. *Explaining and Harnessing Adversarial Examples*

[2] Nicholas Carlini and David Wagner, 2017. *Towards Evaluating the Robustness of Neural Networks*

[3] Lea Schonherr et al, 2018. *Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding*

[4] Nicholas Carlini et al, 2019. *On Evaluating Adversarial Robustness*

[5] *IBM Cloud*, https://www.ibm.com/cloud

[6] *Google Cloud Platform*, https://cloud.google.com/

[7] *Watson Studio's AutoAI*, https://www.ibm.com/cloud/watson-studio/autoai

[8] *Cloud AutoML*, https://cloud.google.com/automl

[9] Alex Krizhevsky, and George Hinton, 2009. *Learning multiple layers of features from tiny images*

[10] Ding, Gavin Weiguang and Wang, Luyu and Jin, Xiaomeng. *AdverTorch v0.1: An Adversarial Robustness Toolbox based on PyTorch*, https://github.com/BorealisAI/advertorch.

[11] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, Ananthram Swami. *Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks.*