Feature Squeezing:

Detecting Adversarial Examples in Deep Neural Networks

作者：

Weilin Xu, David Evans, Yanjun Qi

University of Virginia

evadeML.org

会议：

In Network and Distributed Systems Security Symposium (NDSS) 2018CCS’17, October 30-November 3, 2017, Dallas, TX, USA

## 中文标题：

特征挤压：

在深度神经网络中检测敌对案例

## Questions:

### What is the problem?

针对敌对样本在分类器中的攻击问题，提出了一个新的检测策略：Features Queezing

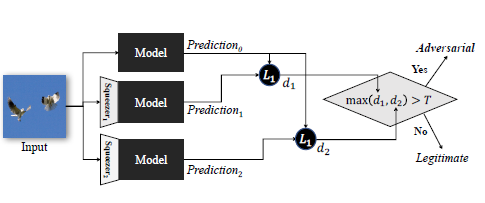
### Why is the problem important?

L

### What is the old technique?

机

### What is the new technique?



### Why is the problem in the futuer?

## 摘要：

Abstract—Although deep neural networks (DNNs) have achieved great success in many tasks, they can often be fooled by adversarial examples that are generated by adding small but purposeful distortions to natural examples. Previous studies to defend against adversarial examples mostly focused on refining the DNN models, but have either shown limited success or required expensive computation. We propose a new strategy, feature squeezing, that can be used to harden DNN models by detecting adversarial examples. Feature squeezing reduces the search space available to an adversary by coalescing samples that correspond to many different feature vectors in the original space into a single sample. By comparing a DNN model’s prediction on the original input with that on squeezed inputs, feature squeezing detects adversarial examples with high accuracy and few false positives. This paper explores two feature squeezing methods: reducing the color bit depth of each pixel and spatial smoothing. These simple strategies are inexpensive and complementary to other defenses, and can be combined in a joint detection framework to achieve high detection rates against state-of-the-art attacks.

摘要 - 虽然深度神经网络（DNNs）在许多任务中取得了巨大的成功，但它们往往会被对立的例子所迷惑，这些例子是通过对自然例子添加小但有目的的扭曲而产生的。以前的研究为抵御对抗性的例子主要集中在提炼DNN模型，但是要么显示有限的成功，要么需要昂贵的计算。我们提出了一种新的策略，特征压缩，可以通过检测敌对的例子来加强DNN模型。特征压缩通过将对应于原始空间中的许多不同特征向量的样本合并为单个样本来减少对手可用的搜索空间。通过将DNN模型对原始输入的预测与对压缩输入的预测进行比较，特征压缩以高准确度和少量误报检测敌对示例。本文探讨了两种特征压缩方法：减小每个像素的颜色位深度和空间平滑。这些简单的策略价格低廉，并且与其他防御措施相辅相成，并且可以结合在一个联合检测框架中，以实现针对最先进攻击的高检测率。

## 亮点解读

## 论文总结

The effectiveness of feature squeezing seems surprising since it is so simple and inexpensive compared to other proposed defenses. Developing a theory of adversarial examples remains an illusive goal, but our intuition is that the effectiveness of squeezing stems from how it reduces the search space of possible perturbations available to an adversary.

Although we have so far only experimented with image classification models, the feature-squeezing approach could be used in many domains where deep learning is used. For example, Carlini et al. demonstrated that lowering the sampling rate helps to defend against the adversarial voice commands [5]. Hosseini et al. proposed that correcting the spelling on inputs before they are provided to a character-based toxic text detection system can defend against adversarial examples [15].

As discussed in Section 5-D, feature squeezing is not immune to adversarial adaptation, but it substantially changes the challenge an adversary f aces. Our general detection framework opens a new research direction in defending against adversarial examples and understanding the limits of deep neural networks in adversarial contexts.

特征压缩的有效性似乎令人惊讶，因为与其他提议的防御相比，特征压缩如此简单和轻省。发展对抗性例子的理论仍然是一个虚幻的目标，但我们的直觉是，压缩的有效性源于它如何减少对手可用的扰动的搜索空间。

虽然迄今为止我们只是尝试使用图像分类模型，但是特征压缩方法可以用于许多使用深度学习的领域。例如，Carlini等人表明降低采样率有助于防御对抗的语音命令[5]。 Hosseini等人。提出在提供给基于字符的有毒文本检测系统之前纠正输入的拼写可以抵御敌对的例子[15]。

正如在第5节中讨论的那样，特征压缩并不是免疫对抗性适应，但它大大改变了对手所面临的挑战。我们的通用检测框架开创了一个新的研究方向，可以防御敌对的案例，并了解对抗情况下深度神经网络的局限性。