WeeklyNote

2019.10.20

張慕琪

ILLUSTRATION BY EDGAR BĄK

DEEP TROUBLE FOR DEEP LEARNING

BY DOUGLAS HEAVEN

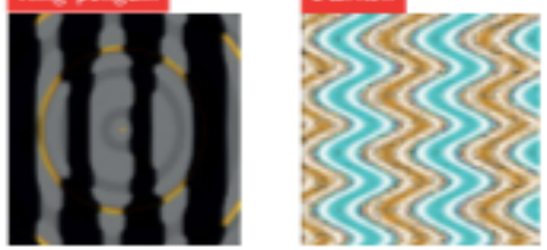
Introduction

FOOLING THE AI

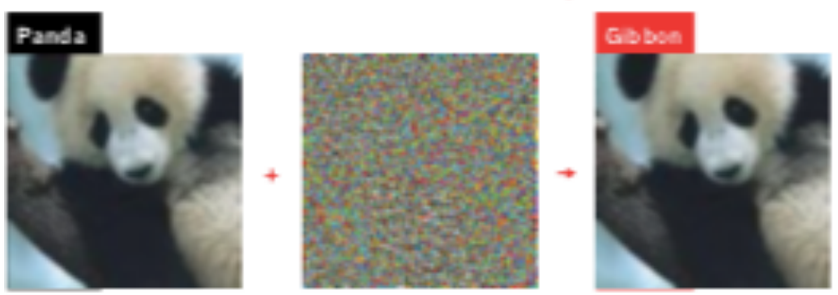
Deep neural networks (DNNs) are brilliant at image recognition — but they can be easily hacked.

These stickers made an artificial-intelligence system read this stop sign as 'speed limit 45'.

Scientists have evolved images that look like abstract patterns — but which DNNs see as familiar objects.



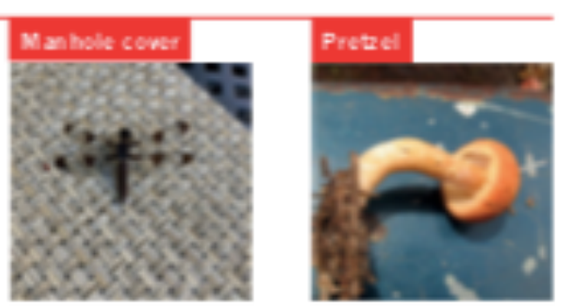
Adding carefully crafted noise to a picture can create a new image that people would see as identical, but which a DNN sees as utterly different.



In this way, any starting image can be tweaked so a DNN misclassifies it as any target image a researcher chooses.

Rotating objects in an image confuses DNNs, probably because they are too different from the types of image used to train the network.

Even natural images can fool a DNN, because it might focus on the picture's colour, texture or background rather than picking out the salient features a human would recognize.



There are no fixes for the fundamental brittleness of deep neural networks.

DNNs do not actually understand the world.

Introduction

2013年，谷歌研究员Christian Szegedy首次提出一个新概念——“对抗样本”。在印本“神经网络的有趣的可能”中，他提出他们小组发现DNN可以成功认出狮子的图像，但通过更改个别像素数据，DNN则认为自己在观看完全不同的一幅图像，例如一个图书馆。Clune也通过实验得到了类似的结论，他认为这种错误在人类大脑中是完全不可能想象的。

这也意味着黑客们有各种不同的方法来攻击一个系统，而当攻击开始后工作者往往很难解决这类问题。

Why Great Power Comes Great Gragility?

目前，大家公认的解决办法是向AI输入更多的数据，然而这样的训练往往是一个漫长的过程；训练仅仅一个模型所需要的数据有可能需要花费几你那的时间来完成，并且数据并不完全是可靠的，而传感器的校正可能会随着时间变化，硬件设施的性能也会随着时间降低。

针对使用较少数据进行学习的方法，人们提出了“转移学习”对训练方法。即使用几个甚至一个例子即可训练出一个新的网络。这个想法建立在已有一个提前训练好的DNN的前提之上，例如，有一个DNN已经见过犯罪数据中几百万的面部图片并习得了一些有用的信息，现在向其提供一张新的图片，DNN可以快速找到数据集中与之最相近的一个图像。

Learning From Less Data

然而，即使是目前最成功的AI系统例如AlphaZero也只能在很狭窄的领域中获得成功，AlphaZero的算法仅针对国际象棋和GO，但两种竞技游戏并不是同时训练的，同时训练将由于各自的干扰而降低胜率。然而从人类的角度出发，会觉得这是件很荒诞的事情，因为人类是不会轻易忘记曾经学过的知识并且人类是可以学以致用。

DNNs don’t have a good model of how to pick out what matters.

科学家希望理想的DNNs输出应该是不会收到图像细节改变的干扰的， 而这件事目前没有人可以优化。

Learning From Less Data

AlphaZero的成功不仅取决于有效的强化学习，还有一种新算法的帮助（Monte Carlo tree search）。换句话说，AI可以通过指导了解如何从环境中最好的进行学习。Chollet认为，AI的下一步，将是赋予DNN自行编写代码的能力，而非使用人类提供的代码。

现在而言，即使科学家们早已意识到DNN对于数据大小的依赖性和DNN的易攻击性，目前尚未有人提出过真正可以完善这两个问题的方法。

谢谢