一、文献信息

论文作者：Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson

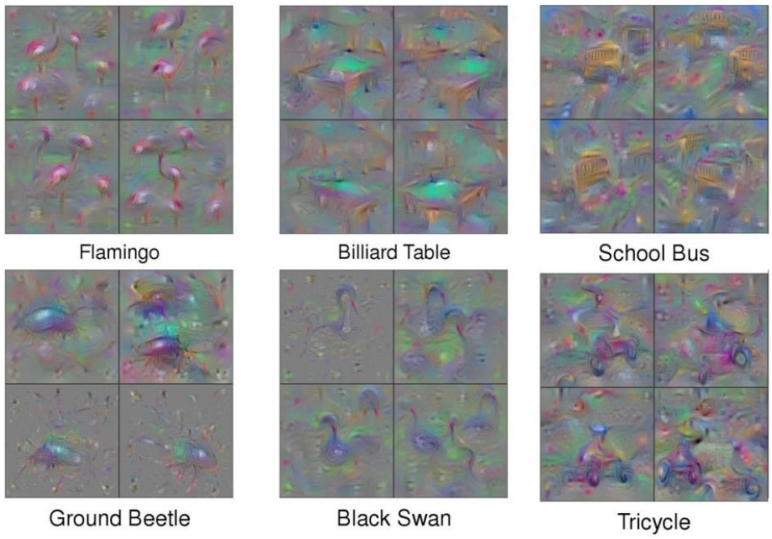
论文题目：《Understanding Neural Networks Through Deep Visualization》

发表途径：Digital Fountain, Inc.

发表时间：2015 年 6 月 26 日

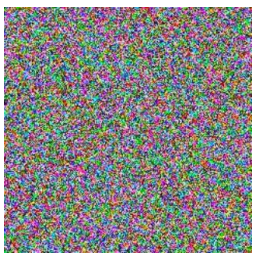
二、问题意义

从历史上看，深度神经网络被认为是"黑匣子"，这意味着他们的内在运作是神秘且难以理解的。最近，作者和其他人一起研究这些黑匣子，以便更好地了解每个神经元是如何学习的，以及它执行的计算。如下图，这些图像是合成的，以最大方式激活深度神经网络 （DNN） 中的单个神经元。它们显示出每个神经元"想要看到什么"，从而显示每个神经元所学的寻找的内容。



三、思路方法

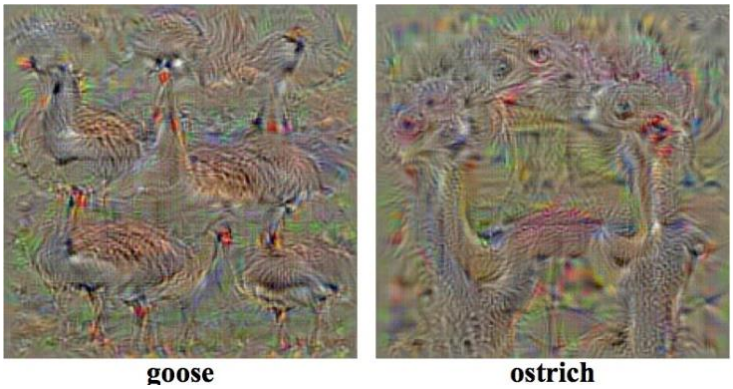
为了使神经网络中特定单元的功能可视化，我们合成能导致该单元具有高激活的输入。首先我们从随机图像开始，这意味着我们为每个像素随机选择一种颜色。图像最初将看起来像静态的彩色电视：



接下来，我们使用此图像执行正向传递，x 作为网络的输入来计算激活 $a_i(x)$，从而引

起 x 出现在一些神经元 i 网络的某处。然后，我们做一个向后传递计算出 $a_i(x)$与网络的早期激活有关。在向后通道的末尾，我们只剩下渐变$\partial a_i(x)/\partial x$，或如何更改每个像素的颜色，以增加神经元的激活 i.我们通过增加一小部分来做到这一点：$x \leftarrow x + \alpha \cdot \partial a_i(x)/\partial x$。一直重复这样做，直到找到一个图像 x*能导致相关神经元的高度激活。

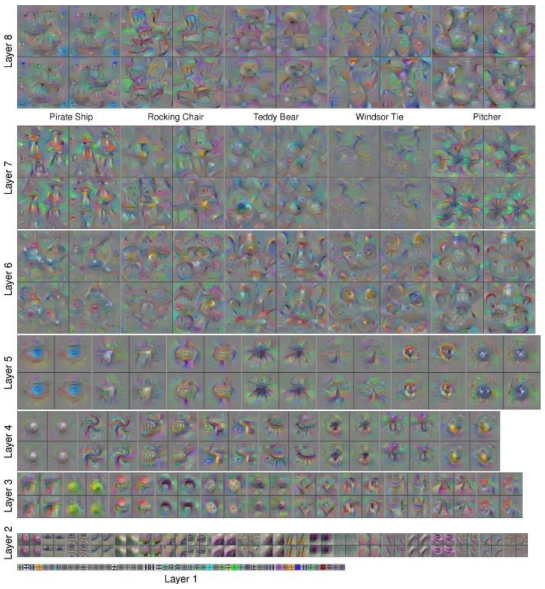
为了生成可识别程度更高的图像，研究人员尝试优化图像：(1)最大地激活神经元，(2)具有与自然图像相似的样式（例如，像素没有极值）。这些图像看起来像正常图像称为"自然图像优先"或"正规化"。通过在优化过程中添加弱正则化来能达到此要求，在生成的图像中，可以开始识别某些类：



goose **ostrich**

添加正则化会有所帮助，但它仍然倾向于生成不自然、难以识别的图像。它们主要由导致高激活的"hacks"组成，而不是由清晰可识别的对象组成：极端像素值、结构化高频模式以及没有全局结构的局部图案的副本。

下面是来自网络各层的示例图像，这些图像可以产生于 DNN 中的任何神经元，包括隐藏的神经元。这样做可以揭示每个层都学到的功能，这有助于我们了解当前的 DNN 是如何工作的，并有助于激发如何改进它们的直觉。



上图展示的是照亮网络上所有八个图层的示例要素的图像，类似于 AlexNet。图像反映不同图层要素的真实大小。对于每个图层中的每个要素，我们显示 4 个随机梯度下降运行的可视化效果。人们可以识别不同尺度上的重要特征，如边缘、角、车轮、眼睛、肩膀、面部、手柄、瓶子等。随着较高层要素合并较低层的简单要素，复杂性会增加。模式的变化也在较高层中增加。特别是从第 5 层（最后一个卷积层）跳转到第 6 层（第一个完全连接的图层）图像会极大的改变。

四、实验意义

与 DNN 交互可以教我们一些关于它们如何工作的事情。这些互动可以帮助建立我们的直观感受，这反过来又可以帮助我们设计更好的模型。到目前为止，可视化效果和工具箱已经教会了我们一些东西：

神经网络学到了人脸识别和文本识别等重要功能，尽管我们并没有专门要求它学习这些东西。它学习它们，是因为它们有助于帮它完成其他任务（例如，识别通常与面孔配对的弓形和书柜，这些书柜通常装满标有文本的书籍）。

有些人认为DNN表示是高维度分布的,因此任何单个神经元或维度都是无法解释的。我们的可视化显示，许多神经元以更局部的方式（例如人脸、车轮、文本等）表示抽象特征，从而表示可解释的。

人们认为，受监督的 DNN 忽略了物体的许多方面（例如海星的轮廓和它有五条腿的事实），而只针对一个或几个独特的事物，通过这些事物就能识别该物体（例如海星皮肤的粗糙、橙色纹理）。相反，这些新的综合图像表明，经过片面性性训练的神经网络实际上比我们想象的要学到更多，包括关于对象的全局结构。