Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

CVPR'15

Abstract

- 近期工作，图片微不足道的改变可能会导致分类器的误分类
- 本文工作，生成人类完全无法识别的图像（噪声），但是分类器有很高的确信将其识别为特定的类别（噪音以99.99%的确信分类为狮子）
- Fooling images

Introduction

- DNN学习层次表示（hierarchical layers of representation）
- Human-competitive 接近人类的
- 本文使用人类完全无法识别的噪声欺骗DNN，使其以高可信度的将噪声识别为特定的类别
- 使用进化算法或者梯度下降在正确的训练集上训练DNN，然后产生对抗样本欺骗它
- 利用对抗样本对网络进行再训练以避免被fooling images欺骗并不容易，很容易再找到一组fooling images欺骗网络，即使网络被再训练了多轮

Methods

- Deep neural network models
  - 首先训练一个高性能的DNN网络
  - 使用Caffe提供的两个模型
    - AlexNet
    - LeNet

- Generating images with evolution
  - evolutionary algorithm，进化算法
  - 包含“有机体（organisms）”的总体，这些有机体交替面对选择（alliteratively face selection 保持最佳），然后随机扰动（random perturbations 变异和/或交叉）
  - 选择有机体的依据是适应度函数（fitness function）是DNN对属于一个类的图像的最高预测值
  - 传统EAs优化解决方案：以在单一目标或所有小目标上实现良好性能（例如，演变图像以匹配单个

ImageNet类）
  - 本文改进的方案：改为使用称为MAP-Elites的算法，能够同时演变包含在许多类别（例如，所有1000个ImageNet类别）上得分都很好的个体的群体。
  - MAP-Elites每个目标的当前最佳个体。 每次迭代时，从群体中随机选择一个有机体，随机变异，如果新个体对该目标具有更高的适应性，则替换当前目标的值。 这里，通过向DNN显示图像来确定适应度，如果图像为任一类别生成更高的预测分数，则新生成的个体将成为该类别档案中的冠军。
  - 用两种不同的编码测试EA，这意味着图像如何被表示为基因组。
    - 直接编码，对于MNIST，每个28×28像素有一个灰度整数，对于ImageNet，每个256×256像素有三个整数（H，S，V）。每个像素值在[0,255]范围内均匀随机噪声初始化。这些数字是独立变异的，首先确定哪些数字发生了突变，通过从0.1开始的速率（每个数字有10%的机会被选择为突变）并且每1000个代数下降一半。然后通过多项式变异算子改变选择要突变的数字，固定突变强度为15。
    - 间接编码，这更有可能产生规则图像，意味着图像包含可压缩模式（如对称和重复）。间接编码的图像往往是规则的，因为基因组中的元素可能影响图像的多个部分。具体而言，这里的间接编码是一个构图模式生成网络（CPPN），它可以演化复杂的，规则的图像，可以编排自然和人造物体。
    - 用CPNN进化的图像可以通过DNN识别，这提供了CPPN编码的EA可以产生人和DNN都可以识别的图像的存在证据，用户通过选择他们喜欢的图像作为进化算法中的适应度函数，成为下一代的父母。
    - CPPN类似于人工神经网络（ANN）。 CPPN将像素的（x，y）位置作为输入，并输出该像素的灰度值（MNIST）或HSV颜色值的元组（ImageNet）。 像神经网络一样，CPPN计算的函数取决于CPPN中神经元的数量，它们如何连接以及神经元之间的权重。 每个CPPN节点可以是一组激活函数（这里是正弦曲线，S形曲线，高斯曲线和线性曲线）中的一个，其可以为图像提供几何规律性。 例如，将x输入传递给高斯函数将提供左右对称，并且将y输入传递给正弦函数提供了上下重复。 进化决定了群体中每个CPPN网络的拓扑，权重和激活功能。
    - CPPN网络从没有隐藏节点开始，随着时间的推移添加节点，鼓励进化首先搜索简单，规则的图像，然后再增加复杂性
    - 代码

Results

- Evolving irregular images to match MNIST
  - 首先演变直接编码的图像，由LeNet高可信的声明为数字0到9（LeNet被训练识别来自MNIST数据集的数字）。
  - 多重独立的进化运行重复产生了MNIST DNN认为99.99%的置信度为数字的图像，但无法辨认。 在不到50代的时间里，每次进化的进程都会重复产生无法识别的MNIST DNN分类的每种数字类型的图像，并具有99.99%的置信度。 到200代，中位数信度为99.99%。 鉴于DNN的几乎确定性，人们可能会期望这些图像类似于手写数字。 相反，生成的图像与MNIST数据集中的手写数字完全不同。

- Evolving regular images to match MNIST
  - 由于CPPN编码可以演变出可识别的图像，测试了这种更有效的常规编码是否可能产生比直接编码

的不规则白噪声静态更多的可识别图像。
    - 结果包含更多的笔画和其他规则，但仍然导致MNIST DNN在仅仅几次生成后将无法识别的图像标记为99.99%置信度的数字， 到200代，中位数置信度为99.99%。
    - 某些图形在一些数字类别中反复演变，表现为该数字的指示。 被分类为1的图像倾向于具有竖条，而分类为2的图像倾向于在图像的下半部分具有水平条。 在其他50次运行中也观察到了类似的鉴别特征（补充材料）。 这一结果表明，EA利用了与MNIST DNN学习的手写数字相对应的具体区分特征。

- Evolving irregular images to match ImageNet
    - 为了验证是否是过拟合的影响，对图像直接编码，在ImageNet 2012数据集上训练的卷积DNN上进行分类。
    - 在这种情况下，直接编码的EA在生成高可信度图像方面不太成功。 即使在2万代后，进化也未能产生许多类别的高可信度图像。
    - 然而，进化确实能够产生45类图像，以99%信度归类为自然图像。

- Evolving regular images to match ImageNet
    - CPPN，产生许多人类无法识别，但DNN置信度为99.99%的图像，5000代后，中位置信度为88.11%，高信心的图像可以在大多数类别中找到。
    - 生成的图像通常包含目标类别的某些功能。 例如，海星图像包含水的蓝色和海星的橙色，棒球在白色背景上红色缝合，遥控器具有按钮网格等。这是因为进化只需要产生对一个类是独特的或区分的特征，而不是产生包含一个类的所有典型特征的图像。
    - 图像产生惊人数量的多样性
        - 图像的不可察觉的变化可以改变DNN的类别标签，所以可能出现这样的情况，演变会产生非常相似的，对于所有类别都高可信度的图像
        - 许多图像在系统发育上彼此相关，这导致进化为密切相关的类型产生相似的图像。例如，一种图像类型对于三种类型的蜥蜴获得高置信度分数，另一种图像类型针对三种类型的小型蓬松狗获得高置信度分数。
        - 然而，不同的进化过程会为这些相关的类别产生不同的图像类型，揭示出进化所利用的每个类有不同的区别性特征。这表明有很多不同的方法可以欺骗DNN
        - 多样性表明图像是非随机的，相反，演化会产生每个目标类别的区分特征。

    - 许多CPPN图像具有重复多次的图案。为了测试这种重复是否提高了置信度得分，或者重复是否仅由CPPN倾向于产生规则图像的事实，消除（即移除）了一些重复元素以查看该图像的DNN置信度分数是否下降。在许多图像中，消除重复元素确实会导致性能下降，这意味着重复元素使DNN更加确信图像属于目标类。
    - 这个结果与先前的一篇文献[26]一致，这些结果表明，DNN倾向于学习低层和中层特征，而不是全局物体结构。如果DNN正确地学习全局结构，如果图像包含很少出现在自然图像中的对象子组件的重复，例如许多狐狸耳朵或无尽遥控按钮，那么图像应该获得较低的DNN置信度分数。
    - 表现不佳的类别是狗和猫，这些类别在ImageNet数据库中的比例过高。
        - 可能的解释：网络被调整为识别许多特定类型的狗和猫。因此，它最终会有更多的单位专门用于这种图像类型。换句话说，它所训练的猫狗数据集的规模大于其他类别，这意味着它有写过拟合，因此更难以愚弄。如果这是真的，那么这个解释意味着更大的数据集可以改善DNN容易被愚弄的问题。

        - 一个替代的，但不是相互排斥的解释是，因为有更多的猫类和狗类，EA很难在特定的狗类（例如日本的小狗）中找到高分的图像，但是在其他相关类别（例如布伦海姆猎犬），考虑到最终的DNN层是softmax，这对于产生高置信度是必需的。这个解释表明具有更多类的数据集可以帮助改善愚蠢。

- Images that fool one DNN generalize to others
    - DNN学习和进化所利用的一类图像具有区别性特征。 一个问题是不同的DNN是否为每个类学习相同的特征，或每个训练的DNN是否学习不同的判别特征。
    - 解决这个问题的一个方法就是看看愚弄一个DNN的图像是否也欺骗了另一个。 用一个DNN（DNNA）演变了CPPN编码的图像，然后将这些图像输入到另一个DNN（DNNB）。
    - 测试了两种情况
        - DNNA和DNNB具有相同的架构和训练，仅在随机初始化方面有所不同
        - DNNA和DNNB具有不同的DNN体系结构，但是在相同的数据集上进行了训练。
        - 我们对MNIST和ImageNet DNN执行了此测试。
    - 进化得到的图像在DNNA和DNNB 都获得了99.99%的置信度得分。 因此，DNN的一些一般属性被CPPN编码的EA所利用。 然而，也有一些图像经过精心调整以在DNNA上获得高分，但在DNNB上没有。

- Training networks to recognize fooling images
    - 看似一个网络可以被重新训练，并被告知先前欺骗它的图像不应该被认为是任何原始类的成员，而应该被认为是一个新的“愚蠢的图像”类。
    - MNIST和ImageNet DNN上的CPPN编码图像测试了该假设。
        - 在数据集（例如ImageNet）上训练DNN1，然后演变CPPN图像，这些图像对于数据集中的n个类别产生DNN1的高置信度分数
        - 然后将这些图像添加到数据集中新类别n + 1;然后我们在这个放大的“+1”数据集上训练DNN2
        - (可选)，重复这个过程，但是将DNN2演变的图像放在n + 1类别中（不需要+2类别，因为任何欺骗DNN的图像都是“愚蠢的图像”，因此可能会在n + 1类）。具体来说，在每一轮迭代，为了表示不同类型的图像，我们在这个n + 1类中添加m个图像，从第一代和最后一代多次进化中随机采样。在MNIST或ImageNet上进行的每次演变分别生成20张和2000张图像，其中一半从第一代开始，另一半从最后一次开始。
        - 训练过的DNNi的误差率与DNN1相似。

- Training MNIST DNNs with fooling images
    - 为了使n + 1类具有与其他MNIST类相同数量的图像，第一次迭代时，将6000图像添加到训练集（取自300次进化运行）。
    - 对于每个额外的迭代，添加1000个新图像到训练集。
    - 鲁棒性并没有因为fooling images的再训练而加强。 进化仍然会为DNN2生成许多无法识别的图像，信心分数为99.99%。

- Training ImageNet DNNs with fooling images
    - 最初的ILSVRC 2012培训数据集增加第1001个类别，添加9000个图片，这些图片愚弄了DNN1。
    - 每个ImageNet类有1300个图像，愚弄类图像的数目增加7倍是为了强调训练中的愚蠢图像。

    - 与前一部分的结果相反，对于ImaNet模型，Evolution对于DNN2的高置信度图像不如DNN1高。 中位信度得分从DNN1的88.1%显着下降到DNN2的11.7%。
    - 为了观察这个DNN2是否学习了专用于伪造DNN1的CPPN图像的特征，或者DNN2是否学习了所有CPPN图像的特征，甚至可以识别的特征，将可识别的CPPN图像从Picbreeder.org输入到DNN2。DNN2正确标记了45个70（64%，前1个预测）的PicBreeder图像作为CPPN图像。
    - 这样学习的再培训模型对CPPN图像具有通用性，有助于解释为什么生成愚弄DNN2的新图像更困难。

- Producing fooling images via gradient ascent
    - 产生高可信度，但大部分不可识别的图像的另一种方式是通过在像素空间中使用梯度上升。
    - 计算特定类的后验概率的梯度 - 这里是DNN的softmax输出单元 - 相对于使用back-prop的输入图像。
    - 然后按照梯度增加选定单元的激活。这种技术遵循[26]。
    - 通过使用L2-正则化，生成的图像具有一些可识别的特征类（例如狗脸，狐狸耳朵和杯子手柄）。
    - 图像也可以DNN以99.99%的信心进行分类，尽管它们大部分无法辨认。
    - 这些优化的图像揭示了第三种愚弄DNN的方法，这些方法产生的定性不同于本文中的两种演化方法。
    - 通过梯度上升最大化softmax输出来获得图像。 优化从图像平均值开始（加上小高斯噪声以破坏对称性）并持续到目标类别的DNN置信度达到99.99%。

Discussion

- 进化产生了巨大的图像多样性。一个不同的预测可能是，由于局部最优，进化将不能产生高置信度分数。也可能出现这样的情况：无法辨认的图像在所有类别中的信任度往往较低，而不是对一个类别的信度很高。事实上，这些预测都没有结果。相反，演变产生了高信心，但无法辨认的图像。
- 分类模型 - 或者学习标签向量y和输入样例X的p（y | X）的模型，可以创建将数据划分到分类区域的决策边界。在高维输入空间中，区分模型分配给类别的面积可能远大于该类别的训练样例占用的面积。
- 远离决策边界并深入到分类地区的合成图像可能会产生高可信度的预测，即使它们远离类别中的自然图像。由于其局部线性性质和高维输入空间的组合，在某些判别模型中显示出高置信度的大区域。
- 代表完整联合密度p（y，X）的生成模型不仅可以计算p（y | X），还可以计算p（X）。 这样的模型可能更难以愚弄，因为愚蠢的图像可以通过它们的低边际概率p（X）来识别，并且当p（X）低时，DNN在这样的图像的标签预测中的置信度可以被忽略。
- 不幸的是，当前的生成模型不能很好地扩展到像ImageNet这样的数据集的高维度，因此测试它们可能被愚弄的程度必须等待生成模型的提升
- 一旦类标签已知，一些生成的图像就可以作为其目标类的成员识别。
- CPPN EA也可以被认为是一种新颖的技术，可以将DNN学到的特征可视化。 同一类在不同运行中生成的模式的多样性表明了该类学习的特征的多样性。

Conclusion

- DNN模型很容易被愚弄，因为它们将许多不可识别的图像以接近确定（near certainly）的形式归类为可

识别的类的成员。

- 两种不同的编码进化算法的方式产生两种不同类型的不可辨认的“愚蠢图像”，梯度上升是第三种。
- DNN将这些物体看作是可识别图像的近乎完美的例子，它揭示了DNN和人类识别物体的方式之间的差异，提出了有关DNN的真正泛化能力。