Intriguing properties of neural networks

ICLR'14

Abstract

- Highly expressive models: 高度表示性的模型
- State of the art: 最先进的
- 会产生一些反直觉的，不可预知的结果
  - Uninterpretable: 不可预知
  - Counter-intuitive: 反直觉

- 个体的高级单元和随意组合的高级单元没有差异，说明在神经网络的高层中，是空间而不是分立的单元包含了语义的信息
- 神经网络的输入和输出的映射是不连续的，可以通过一些微不足道的扰动，使网络对图像进行误分类，这是通过最大化预测误差实现的。这些扰动的具体性质不是随机的学习：相同的扰动会导致在数据集的不同子集上进行训练的不同网络，将相同的输入误分类。

Introduction

- 神经网络可以实现高性能，因为它通过包含数量适中的(a modest number of)大规模并行非线性步骤(massively parallel nonlinear steps)去表示任意计算。
- 最终的计算由反向传播自动发现，因此难以解释，并可能包含反直觉的特性。
- 第一个属性：关于单个神经元的语义信息
  - 前人：找到最大化激活给定单元的输出，隐性假设：最后一个特征层的神经元形成了最有区分度的信息
  - 本文的发现：是被激活的整个空间，而非单个神经元，包含了语义的信息。

- 第二个属性：关注神经网络在微小扰动上的稳定性
  - 应用微小的，非任意的扰动，可能会随机改变神经网络的预测结果。
  - 这些扰动通过优化输入来最大化预测误差实现—对抗样本
  - 对抗样本是相对稳健的，由包含不同层数、激活函数甚至是由不同的训练数据训练的神经网络共享（假如我们在一个神经网络上训练了对抗样本，那么它可以迁移到其他网络上去）

- 通过反向传播学习的深层神经网络具有非直观的特征(nonintuitive characteristics)和内在的盲点(intrinstic blind spots)，其结构以非明显的方式(non-obvious way)与数据分布相关联

Framework

- MNIST,正则化，分成两个不相交的数据集:
  - 全连接，单层，softmax分类器—“softmax”
  - 全连接，双层 — “FC”
  - 顶层是自编码器分类器 — “AE”
  - 标准CNN —“Cone”

- ImageNet:
  - AlexNet

- 10M image samples from Youtube:
  - QuocNet

Unit of phi(i)

- 传统CV系统依赖特征提取，检查特征空间的单个坐标，并将其链接回输入域中的有意义的变化
- 前人的工作，将隐藏单元的激活解释为一个有意义的特征。 他们寻找使这个单一特征的激活值最大化的输入图像

$$x' = \underset{x \in \text{image set}}{\arg\max} \langle \phi(x), e_i \rangle$$

- 本文的发现，在任意方向最大激活函数，也有相关的语义。

$$x' = \underset{x \in \text{image set}}{\arg\max} \langle \phi(x), v \rangle$$

- 每个单独结点的激活程度并不能唯一对应图像的某实际特征。图像特征是存在与整个空间内的，难以分开，不能将这些信息简单的分配给各个节点。

Blind spots in neural networks

- 输出层单元是其输入的高度非线性函数，当使用交叉验证训练，表示给定输入标签的条件分布。
- 一种说法，输出到单元可能为输入空间中不包含训练样本的区域分配小概率。
  - 隐含局部泛化(local generalization)的假设，在训练实例附近，按照预期工作。和在训练实例的极小半径的点，以高概率被分类。

- 本文发现，对深度网络，许多核方法的平滑假设不成立，通过简单的优化程序，可以找到对抗样本，使得分类错误。
- 本文描述一种高效的（通过优化）的方式遍历由网络表示的流形(manifold)，以在输入空间中找到对抗样本的方式。
- 对抗样本表示流形中低概率（高维）的”pockets”，难以通过对给定示例附近的输入进行简单采样而有效找到。
- 已经有多种最新的计算机视觉模型在训练期间采用输入变形来提高模型的鲁棒性和收敛速度。 然而，对

于一个给定的例子，这些变形在统计上是低效的：它们高度相关，并且在整个模型训练过程中从相同的分布中得出。

- 和hard negative mining类似。hard negative mining 包括识别模型给出的低概率的训练集（但是期望给出高概率），然后改变训练集分布（从负样本中选出有代表性的样本），进一步进行模型训练
  - 在bootstrapping方法中,我们先用初始的正负样本(一般是正样本+与正样本同规模的负样本的一个子集)训练分类器,然后再用训练出的分类器对样本进行分类,把其中错误分类的那些样本(hard negative)放入负样本集合,再继续训练分类器,如此反复,直到达到停止条件(比如分类器性能不再提升).

- 公式表示（box-constrained L-BFGS）
  - Minimize $\|r\|_2$ subject to:
    1. $f(x+r) = l$
    2. $x + r \in [0,1]^m$
  - Minimize $c|r| + \text{loss}_f(x+r, l)$ subject to $x + r \in [0,1]^m$

- 利用对抗样本反馈训练，可以得到更高的准确率。利用对抗样本，不断加入训练集，可以得到错误率1.2%，只使用正则化1.6%，正则化+dropout 1.3%
- 对抗样本对每一层输出生成，并用来训练所有层，高层的对抗样本更有效。
- 实验部分
  - 三个实验：单一模型试验，跨模型超参数实验，跨模型跨训练集实验
  - Table 1: mnist 结果
  - L-BFGS训练
  - 前三个：线性模型
    - 其中一个lamda=1
    - 另两个，2隐含层，1分类器
    - 最后一个，单层稀疏自编码器，400节点softmax分类器
    - 最后一列表示使训练集达到0精度所需的最小像素失真（扰动）

  - Table 2:
    - 表示当每个模型采用Table各行所示的，达到训练集0准确率的扰动时，得到的错误率（对角线为100%，显然）
    - 最后两行表示高斯随机干扰的结果，即使干扰的标准差大于大多数对抗样本，但是干扰程度不及。
    - 结论：对抗样本对于使用不同超参数训练的模型有泛化性能，虽然自编码器最有弹性(resilient)，但是也不能完全免疫(immune)

  - 第三个实验：
    - MNIST数据集分成两部分，3个非卷积神经网络，前两个（1，2）训练第一部分，后一个（3）训练第二部分，目的是想实验同时改变超参数以及训练集的影响，1和3超参数一样
    - 对测试数据集生成对抗样本

    - Table 3: 三个神经网络训练的原始数据，以及使训练集达到0准确率所需的最小失真
    - Table 4:
        - 前半部分，使用Table 3中对应行的最小失真时，每个网络的错误率
        - 后半部分，提高失真到0.1时，每个网络的错误率

    - 结论：对抗样本对于跨超参数、跨训练样本的网络仍有迁移性，尽管效果会降低

Conclusion

(参考自知乎专栏)

本文晦涩难懂，共发现了神经网络的两个反常特性：

- 采样层内部，每个单独节点的激活程度并不能唯一的对应图像的某种实际特征。也就是说，在他们之前就有人发现，某一些靠后的采样层中的节点对于图像的某种特征，比如说白色的花，有比较强的感应，即会输出比较大的值。前人就认为了，肯定是这一个节点单独学会了辨认这种特征啊。但Szegedy等人取出某一层的所有输出，乘上一个随机的权重向量，这样得到的就是一个随机的未经训练的节点。结果发现这个节点依然会感应一些图片的特征，比如白色的汪星人。所以他们就认为这些图像特征是存在于整个空间内的，难以分开，不能将这些信息简单的分配给各个节点。

- 他们能够通过在图像上添加极微小的但经过计算的干扰，使分类器得出完全不同的结果。这样添加干扰后的样本叫做对抗样本(adversarial samples)。他们进一步发现这些对抗样本也能骗过一些用不同样本训练的其他模型，证明了这些对抗样本的健壮性。产生这种样本的损失函数基本取决于原模型犯错误的损失函数(即犯错的坚定程度= =) 与这个干扰的大小之和，通过L-BFGS使损失函数最小化。