**Interpreting and Evaluating Neural Network Robustness**

**一、Abstract**

本文旨在通过损失可视化对攻击防御机制进行定性解释，并建立定量度量来评估神经网络模型的内在鲁棒性。

实验表明，本文的度量比传统的基于对抗测试精度的鲁棒性估计有几个优势:

（1）对不同结构、参数的模型，提供统一评估；

（2）精度更高，对不同test设置，评估更可靠；

（3）快速生成，省去了可观的测试成本。

**二、Introduction**

**Contributions：**

1. ·定义并可视化了一个新的损失面，叫做决策面，以此解释adversarial vulnerability and robustness。与基于交叉熵损失面相比，决策面包含隐式决策边界，可视化效果更好。
2. ·证明了对抗性欺骗是由神经网络邻域欠拟合引起的。可视化结果显示，对抗样本是在输入的附近自然存在的点，有时模型未能对其正确分类。
3. ·提出一种鲁棒性评估指标。结合了新的归一化方式，该方法可不受攻击与防御的影响，正确反映固有鲁棒性。
4. ·在防御蒸馏等特定情况下，常用PGD对抗性测试精度会给出不可靠鲁棒性估计，而本文度量则可以正确反映鲁棒性。

**三、Background and Related Work**

**1、Adversarial Attacks and Robustness**

目前，多采用来评估模型鲁棒性。但attack specififications对其影响很大，无法正确反映模型内在固有的鲁棒性。例如，实验表明前面提到的PGD攻击下的测试精度不可靠：具有高PGD测试精度的模型很容易被其他攻击破坏。

因此，本文提出的内在鲁棒性评估指标，目标在于可以不受模型、攻击、防御的改变的影响。

**2、Neural Network Loss Visualization**

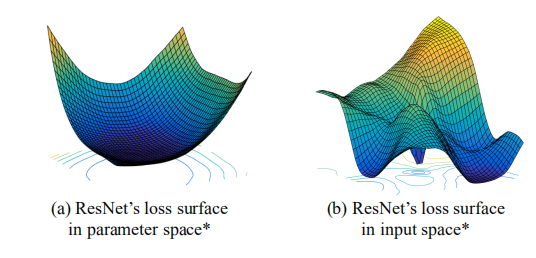
Goodfellow等人2015年提出了损失可视化，用于分析模型训练与收敛性。

本文采用了损失可视化的概念，分析神经网络在对抗扰动下的损失行为。同时也引入了归一化方法来解决reparameterization问题。

**3、Visualization Space Selection**

损失可视化主要是对神经网络的泛化能力进行评估，因此之前的工作主要是在parameter space上对模型的训练和收敛进行分析。本文将之前的方法扩展到了input space。

以ResNet模型为例，这两种空间可视化的区别在于：



（a）参数空间，损失面显示出平滑的极小值。

（b）输入空间，明显的非光滑，说明loss对输入扰动非常敏感。

因此，本文采用输入空间作为默认可视化设置。

**四、Adversarial Robustness Interpretation**

**1、Loss Visualization**

定义模型损失函数为F(θ, x)，（θ是参数设置包括weight和bias，x是输入）。

输入x常在高维空间中构造，所以无法直接可视化分析损失面，需要将高维损失面project到低维空间来可视化。选择两个向量*α、β*作为x-y超平面的基向量，对于输入o，可对其周围的点插值，相应损失值为：

clipboard.png

其中，F中的o表示原始图像，*α、β*是添加的单位扰动，（i,j）表示扰动强度。

损失可视化后，一个点的坐标可以理解为它在*α、β*方向上与原始图像o的发散程度。采集足够多的的点的损失值后，将有高维输入的F投影到选择的超平面上。

**决策面选什么：**

当F为交叉熵损失函数时，相应的损失面无法定性显示出一个输入的明确决策边界，因此，在损失可视化中提出一个决策面来取代损失面：

clipboard.png

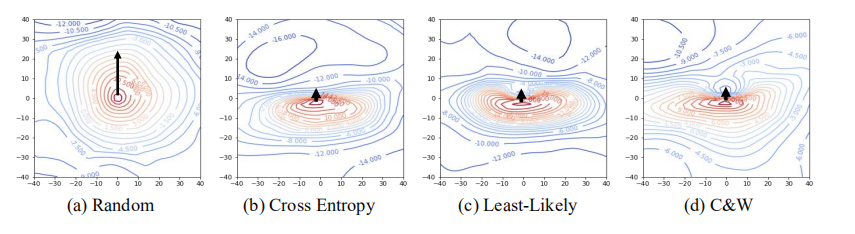
其中Z(x)是softmax层前的logit输出，t是输入x的真实标签。

决策函数S(x)评估预测的置信度，当S(x)>0时，预测正确，S(x)<0时，预测错误，S(x)=0时，表示决策边界。

这样定义的决策面有明确的决策边界，有助于对抗性分析。

**2 、Visualizing Adversarial Vulnerability**

这里做了一个小实验，基于损失可视化将loss投射到二维超平面上，比较模型在决策面上的四种不同损失行为，分别代表四种不同扰动，第一种是随机扰动，其余三种在图下方标出：



箭头标出了跨越决策边界的最短距离。

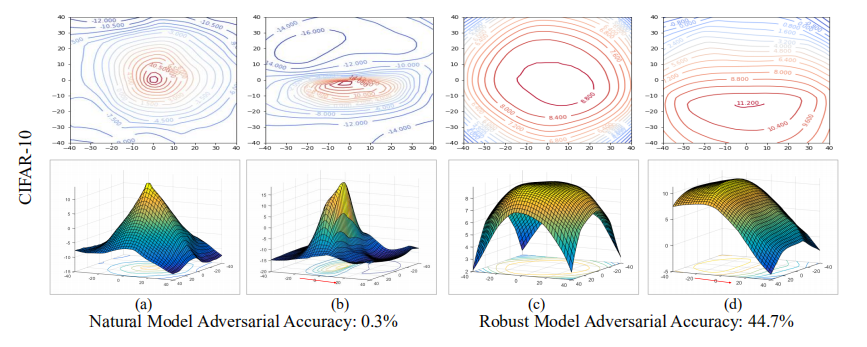
观察：(a)所需距离比(b)~(d)明显长很多。

结论：带有小的随机噪声的自然图像不会明显降低准确性，而加入了无法察觉的小扰动也会误导决策，产生对抗样本。

由上可知：对抗样本不是由敌手“generated”的，而是“naturally existed”的，并使模型无法正确地学习。为了处理神经网络这种固有的脆弱性，本质的、最终的鲁棒性增强应该集中在解决“邻域欠拟合”问题上。

**3 、Interpreting Adversarial Robustness**

分别在MNIST和CIFAR10上训练了natural、robust模型，两种模型结构一样，但natural模型采用natural training，robust模型采用MinMax training。



观察：natural模型的决策面有尖锐的峰值，斜坡很抖，决策的可信度可能会迅速下降到负值区域(错误的分类区域)；robust决策面上其原始输入附近所有邻域点都L(x)>0（正确分类区域），邻域内曲面也很平坦。

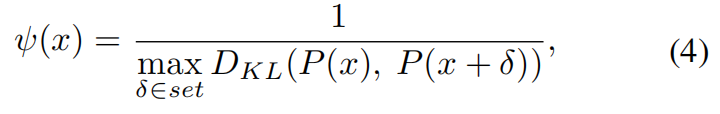
上面的图像验证了上一条结论：增强鲁棒性需要解决邻域欠拟合问题。**在决策曲面上，原始点周围有个平坦、宽阔的平台，是robust模型最理想的特征之一。**

**五、Adversarial Robustness Evaluation**

**1、定义鲁棒性评估**

给定一个可行的扰动集，可以用原预测与最坏情况对抗预测之间的最大散度来表示模型的vulnerability程度。

先计算原始输入与对抗输入之间的divergence（这里使用KL散度：DKL），DKL越低，鲁棒性越强，因此则鲁棒性可以用下式来评估：



其中P(·)表示模型输出的预测值，得到DKL最大值的方式是直接用梯度上升算法。

**2、Invariant Normalization against Model Reparameterization**

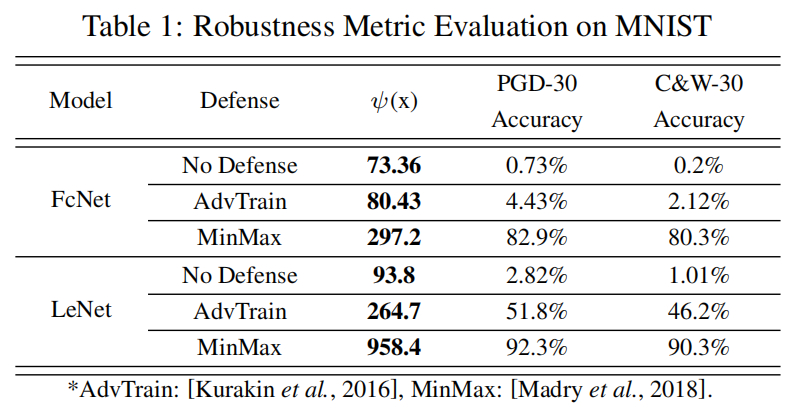
以往的鲁棒性评估存在model reparameterization问题：当weight和bias同时增大相同系数，预测结果和鲁棒性不变。 而本文定义中的KL散度变化会很大。

解决思想：在logit层输出后，加一个scale-invariant normalization层。

具体方法：得到logit层的输出向量，可能包含正值或负值；将其除以最大绝对值，标准化到（-1,1），再集中到正范围（0,2）中；最后，用简单的求和归一化将向量转化为有效的概率分布。

**六、Experiments**

1、在MNIST上，用本文方式进行鲁棒性评估，结果如下：

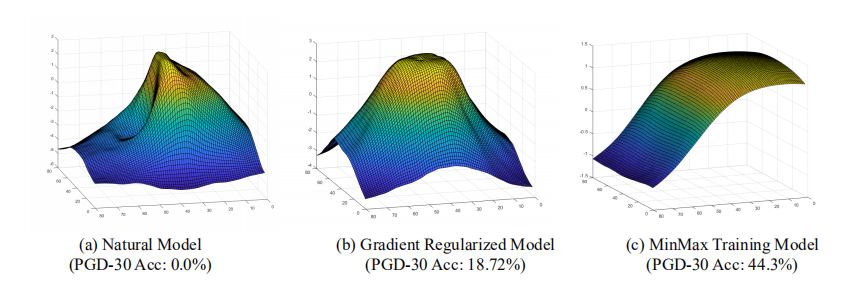


结论：

（1）该方法较好地反映了同一模型的不同鲁棒性（在不同防御下其鲁棒性不同，且与PGD、C&W攻击下的精度对应）。

（2）表明了该方法在这两种模型上的通用性。

2、损失可视化：



模型鲁棒性越高，几何形状更平滑、稳定。