什么是对抗样本

深度神经网络（DNN）很强大，经过大量数据训练之后，可以实现非常复杂的功能，在语音识别、图像识别、自然语言理解等任务上取得了极大成功。

然而，Google 研究者在 2013 年发现 DNN 可能非常脆弱，一个人类无法察觉的噪声就可能让机器产生错判。如图 1 所示，最左边的图是一个熊猫，DNN 也确实将其识别成了熊猫，且对这一识别有 57%的信心，一切正常。之后，研究者将中间的噪音加到熊猫上，生成一幅带噪声的熊猫图，如右图所示。对于人来说，加噪后的熊猫看起来和原来没什么区别，但对 DNN 来说，就认成了长臂猿，而且这个识别结果的信任度达到 99%。这些人类无法察觉但对机器产生严重影响的样本称为“对抗样本”。

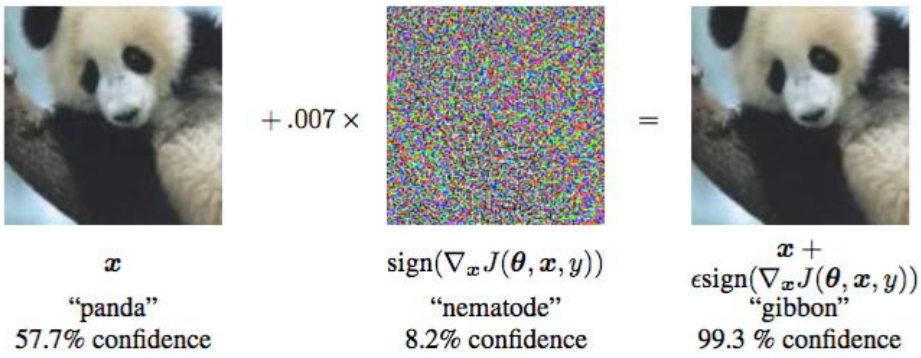

图 1：对抗样本例子。通过给熊猫图片加入少许噪声，可使 DNN 将其识别成长臂猿[2]。

后来，人们发现对抗样本不仅可以通过加入噪声得到，也可以用其它方式生成。如图 2 所示，将一图幅稍微旋转一个角度，DNN 马上把手枪识别成了捕鼠器（左），把雄鹰识别成了猩猩（中），把船识别成了狗（右）。

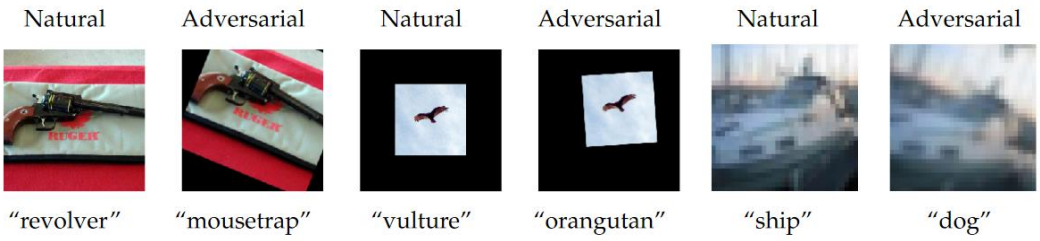

图 2：对抗样本例子。通过对图片进行轻微旋转，使 DNN 识别出错 [3]。

是什么使得强大的 DNN 模型错的如此离谱？研究者提出了很多解释，如高度非线性带来的分类面扭曲、高维空间不同维度上的误差积累、人不可见而机器能观察到的特殊模式等[1,2]。

不论是什么原因，对抗样本的存在具有深刻的意义。一方面，对抗样本意味着深度学习系统的脆弱性，虽然在绝大多数场景下可能表现得很强大，但对特别设计的攻击可能无法抵抗；另一方面，它也说明深度学习这一方法本身的缺陷，我们在利用 DNN 灵活性进行学习的时候，可扩展性问题还没有很好地解决。最后，对抗样本的存在还意味着基于当前的深度学习方法，机器还没有学会人对事物的理解方式。对抗样本本质上是人和机器的理解出显著分歧的样本，这些样本对于理解和改进机器的行为具有重要价值。

图 3：戴上一副对抗眼镜，可以轻松骗过人脸识别系统[4]

[1]Szegedy et al., Intriguing properties of neural networks, https://arxiv.org/pdf/1312.6199.pdf
[2]Goodfellow et al., EXPLAINING ANDHARNESSINGADVERSARIALEXAMPLES, https://arxiv.org/pdf/1412.6572.pdf
[3]Engstrom L, Tran B, Tsipras D, et al. Exploring the landscape of spatial robustness[C]//International Conference on Machine Learning. PMLR, 2019: 1802-1811.
[4] https://m.sfccn.com/article/20210205/herald/NzU4LTM2MTIONA==.html