DKNN 结合不同模型的鲁棒性测试

刘应洴 李丹
四川大学锦城学院 四川 成都 610065

【摘 要】程序的鲁棒性是十分重要的，特别是在图像识别领域中。本文中对一种基于梯度的攻击 KNN 分类器的方法对 Deep KNN 的鲁棒性进行了测试。发现 Deep KNN 对鲁棒性的提高有较大的帮助。

【关键词】DKNN；模型；测试

近几年，随着深度学习的发展，也带动了一些相关的研究，例如在图像识别领域，随着图像识别的准确率的提升，人们不禁思考，如果模型在面对错误的输入时，能否给出正确的结果？

随着对抗样本的提出，很多研究都表明神经网络模型在其面前都十分脆弱，所以，如何增加神经网络模型的鲁棒性是当下比较前沿的研究。

1 背景介绍

对抗样本

对抗样本这一概念由 Szegedy 在 2014 年提出：对于输入样本故意添加一些人类无法感知的细微干扰，导致模型以高置信度得出一个错误的分类。

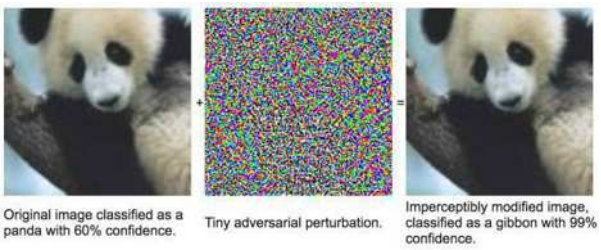

图 1

经典的例子就是一张置信度为 60% 的熊猫图片，再添加部分干扰后，在人眼中依旧是熊猫，但是深度学习模型却认为他是一张长臂猿的图片，并且给出了 99% 的置信度。

所以，以上这个例子可以转化为这样一个约束优化问题：

$$X_{adv} = x + \delta \quad \text{where} \quad \delta = \operatorname{argmax}(\text{Loss}(x + \delta, y))$$

其中代表一个足够小的干扰，Loss 代表某种损失函数，x 代表正确数据，y 代表真实标签。还应该具有某种约束使得它的 p 范数小于某个阈值。

2 基于 KNN 的防御

KNN 是一直十分受欢迎的分类器，它通过计算某种距离 (欧几里得距离) 来获得距离它最近的 k 个邻居，并且从这些邻居中投票来获得最终的标签。KNN 与鲁棒性的关系是十分值得研究的，近几年，很多论文都证明了 KNN 在对抗训练中的潜力。但是，KNN 在面对高纬度的数据时就会显得十分笨拙。因此，目前的一些研究将 KNN 与神经网络相结合，试图通过 KNN 来增加神经网络的鲁棒性。其中 Deep KNN 就是一种优秀的防御手段。从某种角度来说，Deep KNN 适合于任何模型，在每一层都进行 k 近邻搜索提供可解释性和鲁棒性。Deep KNN 利用置信度来对预测结果的正确的概率，对抗样本产生的置信度很低，能够被轻易的检测出来。在 Deep KNN 每一个监视层中，都会使用 KNN 分类器来对输入的样本进行检测。

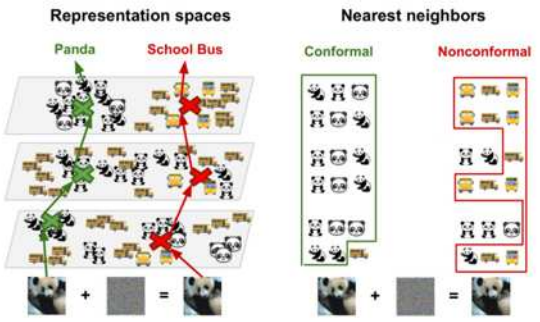

图 2

图 2 中的左图为 DKNN 中每层输出的表示，右图则为训练数据中每一层的最近邻，每一层的可解释性由最近邻提供，鲁棒性来自不同层输入数据的预测，当模型对一个输入进行错误分类时，它必定有一层输出了错误的表示，DKNN 通过识别 DNN 中低层和高层之间训练样本附近标签的变化来防止错误出现，从本质上来说，DKNN 消除了对抗样本能够利用大部分的漏洞，从而为

对抗样本攻击提供了鲁棒性。

3 鲁棒性

鲁棒性 (Robustness) 是指在系统在执行过程中处理异常，以及在遭遇异常输入等操作时正常运行的能力，在生物学中，鲁棒性是指一个生物系统在受到外部扰动或内部参数摄动等不确定性因素干扰时，系统仍保持其结构和功能稳定；在建筑领域，结构的鲁棒性是以避免结构垮塌为目标的整体结构安全性。鲁棒性并不等同与稳定性，稳定性一般来说是指物质的性质随着时间不变化的能力，鲁棒性则是被用来描述面对复杂系统的适应能力。在诸多对于鲁棒性的研究中，总结出鲁棒性的 3 个概念：一是模型具有较高的精度或有效性，这也是对于机器学习中所有学习模型的基本要求；二是对于模型假设出现的较小偏差，只能对算法性能产生较小的影响；三是对于模型假设出现的较大偏差，不可对算法性能产生“灾难性”的影响。

4 基于 k 近邻的攻击

在该论文中提出了基于 KNN 的攻击方法，首先选取出目标样本或者清洁样本作为 x，攻击对手将使用 x 作为生成对抗样本的起点。记处理后的 x 为 =x+, 训练集和测试集记为 (X,Y)，其中 X = {x1, x2, ..., xn} and Y = {y1, y2, ..., yn}.

在这里使用欧几里得距离作为判断距离的指标，KNN 会返回输入元素 x 附近的 k 个最近元素的顺序列表 $(x_{\pi_1(x)}, x_{\pi_2(x)}, \ldots, x_{\pi_k(x)})$，遵循一下规则

$$\left\| x - x_{\pi_{i(x)}} \right\|_2 < \left\| x - x_{\pi_{j(x)}} \right\|_2$$

其中 i<j,最后 KNN 的预测结果由投票函数 Maj 给出：

$KNN(x)=Maj(y_{\pi_1(x)}, y_{\pi_2(x)}, y_{\pi_k(x)})$

Chawin Sitawarin 提出了一种基于此改进的攻击，在之前对于 KNN 攻击的研究中，对于的求解被归纳为：

$$\delta_1 = \operatorname{argmin} \sum_{i=1}^{m} \omega_i \cdot \sigma(\|x_i - (x+\delta)\|_2^2 - \theta^2) + c\|\delta\|_2^2$$

其中σ是 sigmoid 函数，θ是阀值，c 则是通过二分搜索得到的惩罚系数。但是由于使用的是 sigmoid 函数，则会带来由指数引起的溢出问题。Chawin Sitawarin 提出了一种新的思路使用 ReLu 代替 sigmoid 函数，这样可以避免 sigmoid 所存在的问题，在经过一定的修改后也能够达到 sigmoid 相同的效果，改进后的公式如下：

$$\delta_1 = \operatorname{argmin} \sum_{i=1}^{m} \left\{\omega_i\left(\|x_i - (x+\delta)\|_2^2 - \theta^2\right) + \Delta, 0\right\} + c\|\delta\|_2^2$$

该公式与原始公式的区别就在于使用了 ReLu 代替了 sigmoid 函数，ReLu 函数可以轻松的将不符合要求的值区分出来；引入了Δ用来增加数值的稳定性，Chawin Sitawarin 不希望所有的指导样本距离都相同 (不管是否该样本是否被正确分类)。另外，还引入了额外的两个改进：

动态阀值：在之前的研究中使用的固定的阀值和初始化后就不会改变的初始指导样本，但是由于 xi 与 x 距离的增加，固定的阀值变得不再合适，xi 可能会与正确的训练样本靠近而不是靠近指导样本了，这些都会对 KNN 的决策造成干扰。为了解决这个问题，Chawin Sitawarin 提出了动态计算阀值和指导样本，并且每隔 p 步就使用梯度下降进行重新计算，理论上每一步都应该进行重新计算，但是代价过大。注意：p 的取值应该在 10-100 之间，直到 p 的影响降低到可以忽略或者程序的运行时间过长。

指导样本：Chawin Sitawarin 实验了不同的取值，最后发现从原始类别中取出一半的效果最好 (m/2 个)，剩下的 m/2 个从分类正确的样本中取出。参数也采用同样的方法从原始数据中选取，Chawin Sitawarin 建议 m 的值要尽可能的小，较小的 m 意味这需要优化的样本较少，这样能够有效的减少运行时间，因此，只能在攻击未成功时增加 m 的值。

5 实验步骤

使用 CIFAR10 作为数据模型，在对抗训练中，MNITS，CIFAR10 与 ImageNet 都是十分经典的数据集，CIFAR-10 数据集由 10 个类别组成，一共 60000 张 32×32 的彩色图片组成，每一个类都分别有 6000 个图像。通过 torchvision 将数据集读取，并且划分出测试集，验证集与训练集。我们使用 45000 张图片作为训练样本，5000 张作为验证集，15000 张图片作为测试集。

我们在 Chawin Sitawarin 的基础上进行了更多的测试，last-three Deep KNN 在神经网络模型的最后 3 层进行 k 最近邻搜索，Deep KNN 和前一种方式类似，是在每一层都进行 k 最近邻搜索，然后基于所有层的结果进行投票最后得到结果。

VGG 模型由 Oxford 的 Visual Geometry Group 提出，该网络证明了增加网络深度与模型的最终性能存在一定的关系。VGG16 由 13 个卷积层和 3 个全连接层叠加而成。使用对抗干扰的平均 L2 范数来评估攻击的表现和防御的鲁棒性，如果能够发现一个更小的干扰就能够欺骗到目标，这个攻击就更好，平均 L2 范数就越小。

使用 adam 优化器，学习率为 0.0001，训练后准确率达到了 84% 的 VGG 模型作为初始模型，在对未加入 DKNN 的 VGG 模型进行攻击耗时为 10 分钟，平均 L2 范数为 0.6846。

将 DKNN 加入到模型中，将 VGG 的最后三层全连

接层作为 DKNN 的监视层。最后将测试集中的前 200 个样本用于生成对抗样本。

ResNet 模型在图像识别领域已经是一个非常有名的模型，基于 VGG19 模型改进而来，希望通过残差学习处理深度神经网络的退化问题。在使用论文中对未加入 DKNN 的 VGG 模型进行攻击耗时为 8 分钟，平均 L2 范数为 0.18。

在 ResNet 模型上使用同样的数据与设置：

表 1 DKNN 防御对比

k	模型	Mean l2-Norm	运行时间
	ResNet	1.69	85mins
1	VGG16	0.94	76mins
	ResNet	1.55	89mins
3	VGG16	1.52	70mins
	ResNet	1.53	78mins
5	VGG16	2.01	68mins
	ResNet	1.43	139mins
1	VGG16	1.21	73mins

6 实验结果

这里的 L2 范数越大，表示图形的鲁棒性越强。通过上述实验不难发现，Deep KNN 能够对模型提供一定的抵抗对抗样本能力，当 k 值在 3，5 的时候，训练时间相对较短且防御效果更佳。

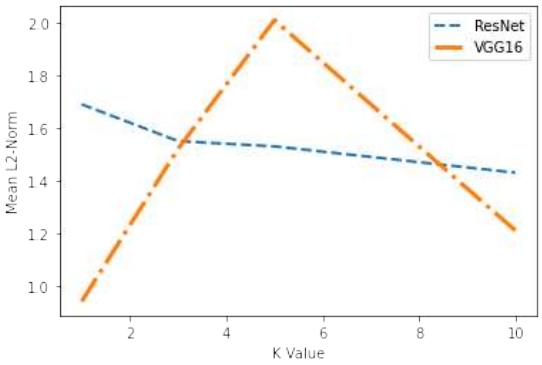

图 3 K 值与 L2 距离关系图

7 总结

Deep KNN 的特点是既能产生分类，又能产生可信度评分。可信性得分能够直接影响它预测的准确率。在对 DKNN 鲁棒性进行测试的时候，我们很难在短时间内找到合适的对抗样本，我们生成的对抗样本只有在被大部分近邻分类错误的时候才进行保存。在原始攻击代码中，原始论文中为了节约设备性能，使用了固定的阀值与固定的指导样本，这让迭代次数提高后的函数出现了欠拟合。Chawin Sitawarin 在对原始算法的优化中调整了终止条件，并且对参数的计算进行了改进，达到更好的效果。

从测试中，可以看出当 DKNN 取值在 3 和 5 的时候，能够使模型达到更强的鲁棒性，训练时间也能够得到相应的减少。使用曼哈顿距离可能能够增加不同预测结果的置信度。

对抗样本的出现是机器学习路上的一大绊脚石，由于有对抗样本的存在，机器学习模型的可靠性大大降低。如果对传入神经网络模型的特征进行微小的干扰，就能使其做出错误的分类判断，这其中存在巨大的安全隐患。如在广泛使用的人脸识别技术中，这让不法分子通过简单的伪装就能够盗取个人信息，所以，一个轻量化的防御框架就十分必要，例如 Deep KNN 框架，能够有效的使得神经网络的鲁棒性得到了提高。

不止在图像识别领域，在文字识别领域，也存在这对抗样本的问题，当前的很多处理方法都不能很好地描述出文字语言的特征，使用对抗样本的攻击和防御不失为一种很好的研究方式。

8 结语

由于文本数据和图像数据的不同，微小的扰动是不会对语义造成影响，所以，在文本处理领域，只需要对可识别的特征提供鲁棒性支持。总之，在图像识别领域使用对抗样本是为了解决恶意的梯度攻击，在文本处理中是为了提高模型的泛化能力。

【参考文献】

[1] 张静文 , 刘耕涛 . 基于鲁棒性目标的关键链项目调度优化 [J]. 系统工程学报 ,2015(01).

[2] 何正文 , 刘人境 , 胡信布 . 基于合同双方交互作用的项目调度优化 [J]. 管理科学学报 ,2014(08).

[3] 田文迪 , 胡慕海 , 崔南方 . 不确定性环境下鲁棒性项目调度研究综述 [J]. 系统工程学报 ,2014(01).

[4] 张静文 , 李若楠 . 关键链项目调度方法研究评述综述与评论 [J]. 控制与决策 ,2013(09).