基于 KNN 最小范化系数攻击在 CNN 上的利用

杨雨澄 李丹
四川大学锦城学院 四川 成都 610072

【摘 要】若要分析算法的鲁棒性，必须采用生成对抗样本的方式来评价算法鲁棒性。然而目前对于算法鲁棒性没有统一的评判标准，也就是说对于不同的算法，人们会针对性地采取不同的方法来生成对抗样本。在这些攻击算法中，Sitawarin 等人对于 KNN 及其衍生模型提出了基于 KNN 和及其衍生模型的攻击方法，其方法和其类似的 KNN 方法比较而言，能快速地找到优秀的对抗样本，在该攻击下对抗样本表现的也十分出色。那么该方法的迁移性能又如何呢，又是否能对其他的算法模型生成优秀的对抗样本呢？本文将利用 Sitawarin 等人的攻击模型对 CNN 算法模型进行攻击，并且进行对比分析造成该结果的原因。

【关键词】KNN；对抗样本；鲁棒性；对抗训练

1 背景

1.1 背景

对抗样本（adversarial examples）这一概念在 2014 年由 Szegedy 等人提出，对于进行处理的样本图片，Szegedy 等人给予其加入微量的噪音，即所谓的扰动后，导致训练后的算法模型对该样本产生错误的判断。

在如今机器学习，深度学习处于已经开始广泛应用的时代，由于对抗样本对算法模型本身具有攻击性和迷惑性，出于安全和实际角度这引起了人们的注意。

1.2 对抗样本意义

对抗样本由于其对算法模型具有迷惑性的特点，可以利用于对算法的攻击，这能造成算法模型错误的分类。出于安全角度考虑，这样的攻击将会造成许多可怕的后果，例如智能驾驶中，如果给计算机传入带有扰动的数据，会造成汽车做出错误的判断，从而酿成严重后果。

但另外一方面，这样的对抗样本也可以直接用来评价该算法鲁棒性，用于评价算法模型是否可以避免上述情况。因此近几年对于算法的鲁棒性的研究越来越多。

然而，目前并没有较通用的攻击模型算法，对抗样本的生成可以说都是基于算法模型而具有针对性的，其攻击都是对不同的算法模型和原理采取不同的攻击手段。在这些攻击中，Yang 等人提出了基于 KNN 的一种优秀的攻击算法，本文将不会针对 Yang 等人的算法进行过多讲述，具体请参考相关文献。Sitawarin 等人在 Yang 等人的基础上对基于 KNN 的攻击算法进行了更加深入的研究，提出了一种基于 KNN 及其衍生模型更加优秀的攻击算法。

本文就利用 Sitawarin 等人的攻击算法对 CNN 模型进行攻击，并分析其结果及原因。

2 Sitawarin 等人攻击简介

2.1 攻击简介

KNN 算法模型，对样本的判断是选取与其最近邻的 K 个样本，并分为较多的一类，这样的算法对其判断造成的影响，在于最近的几个样本之中，而不是所有的样本均有所影响，因此怎样选取可用于生成对抗样本的 K 个邻近样本，即如何选取指导样本，成为了难题。

Sitawarin 等人的基本思路是，选取距离样本最近的 K 个与目标样本有着不同标签值的样本作为指导样本。

Sitawarin 等人最先提出了基于梯度下降的方法找到优秀的对抗样本，引入 sigmoid 函数和惩罚干扰系数，并且通过二元搜索得到平衡常数 c。

2.2 攻击优化

Sitawarin 等人近期提出的基于 KNN 及其衍生模型的攻击是基于其之前提出的攻击算法的再次优化[1]，他的整体思想和之前的想法一致，但做出了一些细节上的优化，使得攻击更加优秀。

首先在指导样本上，Sitawarin 等人进行了优化，由于 KNN 算法实际分类是选取其距离最近的 K 个同类标签值最多的样本作为结果，因此直接选取 K 个不同类别的标签值其攻击性不会很强，Sitawarin 对其进行优化，改用选取距离样本最近的 K/2 个与却与目标样本有着不同标签值的样本作为指导样本。

在其提出的攻击算法基础上，Sitawarin 等人对具体实现方式也进行了优化，将 sigmod 函数用 ReLu 函数替代，引入 Δ 增加数值稳定性，并定期梯度优化阈值，公

式如下。

2.3 实验效果

我们通过 Sitawarin 等人优化前后的实验结果和 Yang 等人的采用思想所提出的基于 KNN 算法的攻击算法 [2] 进行对比实验，并且得到以下结果，见表 1：

表 1 sitawarin 等人对比实验结果

k	Attacks	Mean 2-Norm	time
1	Yang. 等人	2.4753	2h
	Sitawarin 等人	3.4337	2m
	Sitawarin 等人优化后	2.7475	5m
3	Yang. 等人	2.9857	11h
	Sitawarin 等人	3.9132	1m
	Sitawarin 等人优化后	2.9671	5m
5	Yang. 等人	3.2473	44h
	Sitawarin 等人	3.9757	1m
	Sitawarin 等人优化后	3.0913	5m

这里的 dist 指的是最小扰动的生成对抗样本到原样本的 l2 距离，即为扰动，该扰动越小，更能说明生成的对抗样本和原样本距离更近，也就意味着更小的扰动就能使得算法模型分类错误。因此 dist 可以用来直接评价攻击算法的好坏，dist 越小代表着该算法攻击效果越好。

通过上表，我们不难发现，Sitawarin 等人的算法在时间效率上明显优于 Yang. 等人，在 K>1 时其优化前的攻击算法的 dist 值略微大于 Yang 等人的攻击算法。但其优化后的攻击算法，虽然多花费了一些时间成本，但其结果显示 l2 范化系数更小，其攻击也更为优秀。

3 实验设计

3.1 实验目的

将 Sitawarin 等人的攻击方法 [1] 对 CNN 模型进行攻击，并通过分析比较该基于 KNN 及其衍生模型的攻击方法对 KNN 衍生模型的产生的攻击效果和对 CNN 模型的产生的攻击效果之间的差异。评价该算法的迁移性，以及分析造成差异的原因。

3.2 CNN 算法简介

CNN 又称为卷积神经网络，通过卷积提取图片中的特征值，结合神经网络的思想，通过各层输入值与权重进行运算，得到最终结果。

由于该实验是利用的 MNIST 数据集和 CIFAR10 数据集，这两个数据集的特点都是以图片作为数据，因此采用 CNN 算法进行分类是十分合适的。

3.3 CNN 与 KNN 模型的差异

CNN 又称为卷积神经网络，其思想在于卷积和神经网络，卷积的核心目的在于提取图片的特征值，比且依赖神经网络作为了运算的主要思想，神经网络是通过所有数据的输入训练建立的模型，他的每个权重是由数据集内所有的数据决定的，而 KNN 算法模型则是对其距离较为近的数据集影响决定的，因此可能是导致结果差异的重要原因。

3.4 实验数据

根据 Sitawarin 等人之前的实验，本文采用相同的数据集进行实现。手写 MNIST 数据集和 CIFAR10 数据集。

3.5 实验步骤

（1）我们采用 Sitawarin 等人基于 KNN 及其衍生的攻击模型，且在 MNIST 数据集和 CIFAR10 数据集上攻击 KNN 算法模型。

（2）我们采用 Sitawarin 等人基于 KNN 及其衍生的攻击模型，且在 MNIST 数据集和 CIFAR10 数据集上攻击 CNN 算法模型。

（3）对比这四种攻击的最小 l2 范化系数（即 dist）的差异，并且分析原因。

4 结果及分析

4.1 实验结果

dist 数值结果，见表 2：

表 2 该攻击在 KNN 和 CNN 算法模型上的结果

数据集	算法模型	Dist	ACC
MNIST	KNN	3.109	0.965
MNIST	CNN	1.740	0.961
CIFAR10	KNN	1.104	0.846
CIFAR10	CNN	0.843	0.643

实验结果过如表 1，我们可以看到，在 MNIST 和 CIFAR10 数据集中，Sitawarin 等人的攻击方法 [1] 都对 CNN 算法模型攻击的更加优秀。这似乎可以直接说明该算法的也适用于对 CNN 攻击，但是其 dist 值与 CNN 和 KNN 这两种算法是否有直接原因？提出了以下猜想。

4.2 结果猜想

猜想一：卷积神经网络，其思想在于卷积和神经网络，卷积目的是提取图片的特征值，神经网络作为了运算的主要思想，神经网络是通过所有数据的输入训练建立的模型，他的每个权重是由数据集内所有的数据决定的，而 KNN 算法模型则是对其距离较为近的样本影响决定的，因此可能是导致结果差异的重要原因。

猜想二：不同的算法模型分类得到的分界线有所区别，其分界线的不同，若分界线距离指导样本越近，就越容易产生攻击效果更优秀的对抗样本。导致攻击的 dist 受到影响。

4.3 猜想验证

基于猜想一，可根据作者的攻击算法思想来分析 K 值对攻击的影响。Sitawarin 等人的攻击方法。Sitawarin 等人的 1 攻击算法是选取 K/2 个最邻近的且分类不同的

样本作为生成对抗样本的指导样本。如图 1 a) 所示

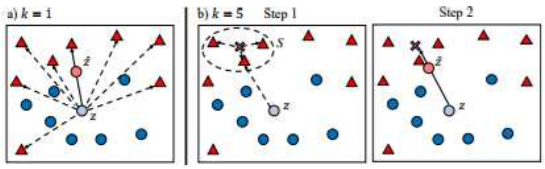

图 1 指导样本示意图

当 K 选取为 1 的时候，原样本选取距离最近的不同类别的样本，梯度下降，直到找到最近的对抗样本位置，并计算出原样本到对抗样本之间的 L2 距离，即 dist。

当 K>1 时，如图 1（b）所示，则是在原样本的基础上选取最近的 K/2 个不同类别的样本，再取其均值进行梯度下降。

若 K 值过大，会导致选取许多距离过于远的样本作为指导样本，这样会整体使得 dist 偏大，得到的攻击效果就会变弱。

本文选取 MNIST 数据集作为实验，分析样本和的关系：

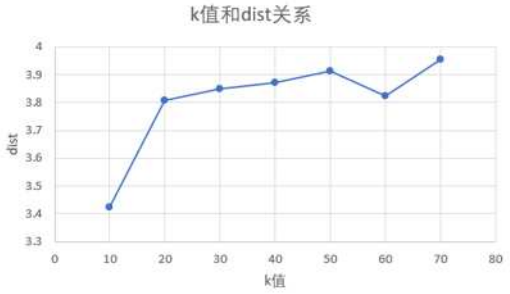

图 2 MNIST 数据集下 K 值和 dist 的关系图

又由图 2 可以看出，K 值和 dist 几乎成正向关，K 值越大，dist 越大，可能是由于 K 值越大，其中心距离原样本的区域越远，导致 dist 变大。这说明了 K 值的确可以影响 dis 的大小，

但是若样本呈现内外圆形状且均匀的分类状况的话，随着 K 值的增大，其平均值可能还会在圆的中心，如图 3 b)，在这种情况下，K 值并不会和 dist 呈现正相关。dist 的值会由于样本均匀分布在周围，导致 dist 的值小于任意两个非同类样本的距离。这种情况下的 dist 也不会随着 K 值的增大而增大，而是逐渐趋于稳定。

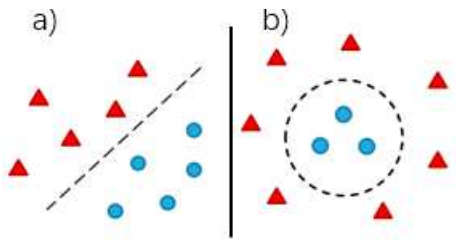

图 3 猜想下的指导样本示意图

因此 K 值并不能很合理地解释为什么 CNN 模型的攻击会比 KNN 模型的攻击效果更好。

基于猜想二：Sitawarin 等人的攻击是针对于 KNN 及其衍生算法模型的攻击，基于 KNN 的原理，他采取的是最近邻的 K 个样本作为分类判断依据，对于不同类别样本较为集中的数据 KNN 算法可以十分精确地对预测样本继进行分类，但是对于数据较为分散，或存在交叉的情况，使用 KNN 算法模型分类，其 K 值成为了进行数据的重要条件，该算法对边界线并不能很明确的划分，因此 KNN 算法对边界附近的样本分类不是很理想。如图 4，该图像表面，算法本身的边界确定不同，会导致 dist 的值有所差异。

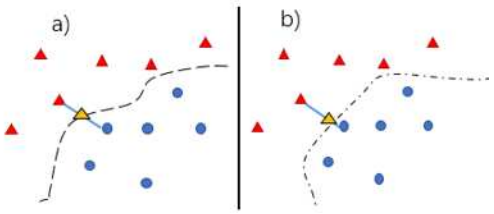

图 4 猜想下的边界示意图

CNN 与 KNN 由于其不相同的算法思想，按照 CNN 的算法思想，其相较于 KNN 而言对边界会更为明确，由此使得 KNN 算法和 CNN 的攻击效果会出现不同。这也是导致 dist 值不同原因。

5 结束语

通过上述实验可以看出 Sitawarin 等人基于 KNN 的攻击，同样也适用于 CNN 算法模型，且对于 CNN 而言由于其边界相较与 KNN 更加精确的原因，该攻击在 CNN 上的利用能产生更优秀的对抗样本。这也证明了 Sitawarin 等人的攻击是十分优秀的。同样可以利用于 CNN 上，但是该攻击对于 CNN 算法而言，其 dist 值作为衡量标准，不能直接和基于 KNN 的算法模型的攻击效果进行直接对比。

【参考文献】

[1]C. Sitawarin and D. Wagner, “On the robustness of deep k-nearest neighbors,” vol. abs/1903.08333, 2019. [Online]. Available: http://arxiv.org/abs/1903.08333.

[2]Y. Yang, C. Rashtchian, Y. Wang, and K. Chaudhuri, “Adversarial examples for non-parametric methods: Attacks, defenses and large sample limits,” CoRR, vol. abs/1906.03310, 2019. [Online]. Available: http://arxiv.org/abs/1906.03310.

[3] W. Gao, X. Niu, and Z. Zhou, “On the consistency of exact and approximate nearest neighbor with noisy data,” CoRR, vol. abs/1607.07526, 2016. [Online]. Available: http://arxiv.org/abs/1607.07526.

[4] H. W. J. Reeve and A. Kab´an, "Fast rates for a knn classifier robust to unknown asymmetric label noise," CoRR, vol. abs/1906.04542, 2019. [Online]. Available: http://arxiv.org/abs/1906.04542.

[5] Christian Szegedy. Wojciech Zaremba. Ilya Sutskever. Joan Bruna. Dumitru Erhan. Ian J. Goodfellow. and Rob Fergus. Intriguing properties of neural networks. In ICLR, 2014.

[6] C. Sitawarin and D. Wagner, "Minimum-Norm Adversarial Examples on KNN and KNN-Based Models," vol. abs/2003.06559, 2019. [Online]. Available: http://arxiv.org/abs/2003.06559.