

雷达目标辨识网络的对抗性样本分析

罗 恒^{1,2,3}, 李增辉³, 李建勋³, 李小波²

(1. 解放军 95980 部队, 湖北 襄阳 441000)

(2. 国防科技大学 电子对抗学院, 合肥 230037; 3. 空军研究院, 北京 100089)

摘要:为了验证雷达目标辨识网络存在风险,提升基于辨识网络的雷达目标辨识效果,文中研究了不同雷达目标辨识网络的对抗性样本。针对雷达目标辨识网络中的卷积神经网络和分解卷积神经网络,建立对抗性样本生成模型,按照该模型生成对抗性样本,并对生成的结果进行分析总结。实测数据处理结果表明,雷达目标辨识网络存在潜在风险。

关键词:雷达干扰;对抗性样本;目标辨识;神经网络

中图分类号:TN974 **文献标志码:**A **文章编号:**1004-7859(2020)03-0025-04

引用格式:罗 恒, 李增辉, 李建勋, 等. 雷达目标辨识网络的对抗性样本分析[J]. 现代雷达, 2020, 42(3): 28-31.

LUO Heng, LI Zenghui, LI Jianxun, et al. Analysis of adversarial example of radar point discrimination networks[J]. Modern Radar, 2020, 42(3): 28-31.

Analysis of Adversarial Example of Radar Point Discrimination Networks

LUO Heng^{1,2,3}, LI Zenghui³, LI Jianxun³, LI Xiaobo²

(1. The Unit 95980 of PLA, Xiangyang 441000, China)

(2. College of Electronic Countermeasures, National University of Defense Technology, Hefei 230037, China)

(3. Air Force Academy, Beijing 100089, China)

Abstract:For verifying the risky of radar target discrimination network and improving the effect of radar target discrimination based on discrimination network, we do research on adversarial examples of different radar target discrimination network. Focusing on two radar target discrimination network convolutional neural network and factorized convolutional neural network, we construct the model of adversarial example and generate adversarial example using the model, and do analysis of the result. The result of the experiment shows that the two discrimination networks have risk.

Key words:radar jamming; adversarial example; target discrimination; neural network

0 引 言

对抗性样本的发现证实了很多神经网络存在风险,这使其成为机器学习领域内的一个新兴研究热点^[1]。对抗性样本是指通过算法对数字矩阵添加一个扰动变量,生成新的数字矩阵,使得原本训练好的辨识网络发生错误,影响辨识结果^[2]。自 2014 年被 Christian Szegedy 等人^[3]提出后一直受到人们的广泛关注。

雷达抗干扰技术是一种预警和保护我方雷达避免遭受敌方电子干扰的技术,该技术在现代战争中发挥着影响战争态势发展的作用^[4]。传统抗干扰措施主要采用模式识别的方式抗欺骗干扰,这种方式仅能针对一项或者几项具体特征进行针对性抗干扰,依赖人为经验总结,信息利用率不高,而机器学习算法可以更广泛地集成多种信息,比人为经验总结更准确^[5],采

用机器学习的相关算法进行雷达目标辨识逐渐成为新兴的雷达抗干扰研究热点。针对某雷达采集到的实测数据,在经过预处理后,搭建的卷积神经网络(CNN)和分解卷积神经网络(FCNN)对雷达点迹样本数据进行辨识,可以准确识别真实目标样本、假目标样本和地杂波样本,辨识正确率能够达到 99%^[6]。但是,随着对抗性样本被发现,研究人员意识到这种辨识网络存在风险,会改变辨识网络的辨识结果。研究对抗性样本可以帮助我们测试和改进辨识网络的性能。

1 辨识网络 and 对抗性样本模型

1.1 辨识网络

本文使用的两个雷达点迹辨识网络是卷积神经网络和分解卷积神经网络。在对某雷达点迹的实测样本数据进行辨识实验中,这两个辨识网络均能以 99% 以上的正确率区分真实目标样本、地杂波样本和假目标样本,但是 FCNN 的参数数量仅有 CNN 参数数量的十分之一^[6]。

通信作者: 罗恒 Email:296463590@qq.com
收稿日期:2019-11-20 修订日期:2020-01-18

常见的神经网络主要包含以下六个层级^[7]。

- 1) 输入层:将实测数据预处理,生成指定大小的切片矩阵,通常作为辨识网络的输入。
- 2) 卷积层:卷积是使用卷积核对输入层进行提取特征的一种操作^[7]。相同的输入矩阵,卷积结果受卷积核、步长和边界模式影响。
- 3) 分解卷积层^[8]:将卷积操作进行拆分,先提取特征,再对特征线性合并的层级。相同的输入矩阵,分解卷积结果受卷积核、步长和边界模式影响。
- 4) 池化层^[7]:对上一层级结果进行降采样操作,从上一层级中提取特定规律的矩阵的层级。常见的池化操作有最大池化和平均池化。
- 5) 全连接层:全连接层的每个结点和上一层级的每个节点都相连,把上一层级的特征进行综合,该层级的参数一般最多。
- 6) 输出层:输出层为分类器,给出辨识概率分布,达到多分类的功能。

CNN 和 FCNN 的框架结构组成如表 1 所示^[6]。

表 1 两个辨识网络的框架结构

辨识网络	输入层	卷积层	池化层	全连接层	输出层
CNN	1 层	2 层标准卷积层	1 层平均池化层	1 层全连接层(128 个连接节点)	1 层(3 个连接节点)
FCNN	1 层	1 层标准卷积+2 层分解卷积	1 层平均池化层	1 层全连接层(32 个连接节点)	1 层(3 个连接节点)

表 1 给出了两个辨识网络的框架结构。两个网络

最主要的不同之处在于 CNN 中使用了两个标准卷积层,全连接层中有 128 个连接节点^[9];FCNN 中使用了一层标准卷积层和两层分解卷积层,且全连接层中有 32 个连接节点^[7]。二者的分类正确率均可以达到 99% 以上,但是 FCNN 使用的参数更少,计算量更小^[6]。

1.2 对抗性样本模型

给定一个输入样本,辨识网络读取输入样本,输出辨识概率。假设有一个输入集合 I ,集合中的元素 X 为输入样本,输入样本矩阵中最大元素值为 l ,最小矩阵元素值为 u 。

一个已经训练好的网络 $p=f(X)$,对于任意输入样本 X ,其输出的辨识概率分布为 $p=[p_1 \cdots p_i \cdots p_n]$,其中 p_i 对应于将输入样本判决为第 i 类样本的概率,若 $p_i = \max(p)$,则将 p_i 称为辨识概率。假设给定某样本,则判决网络对其判决概率向量中最大元素值为 p_m ,则认为该给定样本为第 m 类样本。

我们给样本矩阵 X 添加一个元素值较小的扰动矩阵 D ,其依旧可以被划分为第 m 类样本。添加扰动的数值较小, $X+D \in I$ 依旧成立,即生成了对抗性样本。

对抗性样本验证了辨识网络的潜在风险,使得输入为第 m 类样本被辨识为其他样本,改变辨识结果。

图 1a) 为一个真实目标的原始样本。判决网络以 84.09% 的辨识概率将其辨识为真实目标。图 1b) 为添加的扰动,一般其值较小。用 D 表示原始扰动矩阵,则图中的扰动为 $0.4185 \times \text{sign}(D)$ 。图 1c) 为图 1a) 对应的对抗性样本。

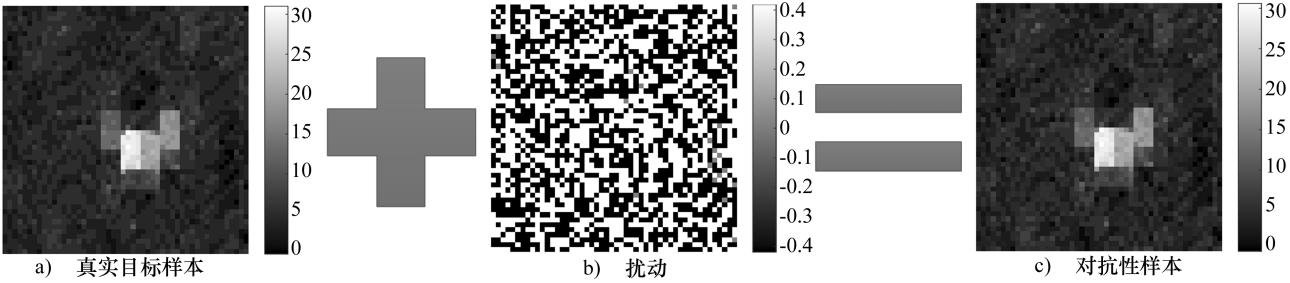


图 1 对抗性样本模型

对抗性样本的生成过程是在已建立模型的基础上,采用相关算法,对给定约束条件的函数进行优化的过程。

给一个第 m 类样本,总结文献[1]中关于对抗性样本的描述,生成对抗性样本的模型可以表示如下

$$\begin{aligned} \min \quad & \|D\|_2 \\ \text{s. t.} \quad & \begin{cases} l \leq X \leq u; p_m = \max(f(X)) \\ l \leq X + D \leq u; p' = f(X + D) \\ \max(p'_1 - p'_m, \dots, p'_n - p'_m) > 0 \end{cases} \end{aligned} \tag{1}$$

式中:给定的原本属于第 c 类的初始样本 X 在添加微

小扰动 D 后被辨识为其他类型样本。

训练一个辨识网络通常意味着通过不断地改变权重来最小化分类错误。为了生成对抗性样本,假设权重已经固定,找到最小的失真量 D 即可找到对应的对抗性样本。

文献[1]中给出了式(1)的进一步化简结果

$$\begin{aligned} \min_{\text{D}} \quad & (\|D\|_2 + C \times H(p, p^m)) \\ \text{s. t.} \quad & \begin{cases} l \leq X + D \leq u \\ p = f(X + D) \end{cases} \end{aligned} \tag{2}$$

式中: p^m 表示在添加微小扰动后辨识概率分布中对应

于第 m 个元素值为 1、剩余元素全为 0 的向量; H 表示交叉熵运算,其计算公式^[11]如下

$$H(\boldsymbol{p}, \boldsymbol{q}) = \sum_x \boldsymbol{p}(x) \times \lg\left(\frac{1}{\boldsymbol{q}(x)}\right)$$

(3)

式中: \boldsymbol{p} 表示真实的概率分布; \boldsymbol{q} 表示预测的概率分布。

交叉熵通常可以用来衡量模型对真实概率分布估计的准确程度。式(2)中,常数 C 用来平衡 $\|\boldsymbol{D}\|_2$ 和交叉熵 $H(\boldsymbol{p}, \boldsymbol{p}^m)$, 常数 C 的值越低,交叉熵的影响作用就会减弱,但是常数 C 的值太低会使优化不可行。我们要在可优化的基础上尽可能地降低 C 的值。文献[2]给出了常数 C 的具体计算方法。式(2)的具体计算方法可以使用拟牛顿算法进行优化^[12-15]。

2 实验

本节主要介绍生成对抗性样本的实验,并对实验结果进行分析总结。

2.1 实验设计

该实验使用 Python 语言进行编程,在 Spyder 集成环境下,应用 Tensorflow、Keras、Foolbox 等集成包进行实验,实验的硬件平台为:

- (1) CPU: Inter Xeon E5-260 v3 @ 2.4 GHz;
- (2) GPU: GTX TITAN X。

实验使用的数据为脉压后经过切块和翻转预处理的实测雷达数据。以水平方向表示目标与雷达原点的距离,距离间隔为 30 m;以垂直方向表示测量的方位角,角度间隔为 0.09°。

对抗性样本研究是为了测试和改进辨识网络,使辨识网络更好地进行雷达反干扰研究。实验中以真实目标样本为例生成对抗性样本,以真实目标样本作为原始样本,使用卷积神经网络和分解卷积神经网络生成对抗性样本。实验使用真实目标样本 3 275 个,其中训练样本 2 675 个,测试样本 600 个。

2.2 实验结果

对抗性样本的通用模型已经在图 1 中进行介绍。现选取一个普通真实目标样本作为原始样本,CNN 分别对原始数据样本和生成的对抗性样本进行辨识,辨识结果如表 2 所示。

表 2 CNN 对原始样本和对抗性样本的识别结果 □

	假目标样本	真目标样本	地杂波样本
原始数据样本	0.04	99.31	0.65
生成的对抗性样本	70.24	20.00	9.76

表 2 为辨识网络对单个原始样本和生成的对抗性样本辨识概率的比较。辨识网络对于原始样本辨识其为真实目标样本的概率为 99.31□,但是对于对抗性

样本,辨识网络以 70.24□ 的概率将其辨识为地杂波样本,仅以 20.00□ 的概率将其辨识为真实目标样本。该实验结果表明 CNN 在进行样本辨识时存在风险。

对 3 275 个原始真实目标样本,通过建立的对抗性样本模型,采用拟牛顿算法生成对抗扰动,进而获得对应的对抗性样本,使用两个辨识网络对生成的对抗性样本进行辨识,对辨识结果进行统计,得到如表 3 所示的概率分布。

表 3 辨识网络对对抗性样本的辨识概率分布

辨识概率分布	FCNN	CNN
0	587(18□)	432(13□)
33.33□~40□	45(1□)	97(3□)
40□~50□	552(17□)	635(19□)
50□~60□	1 668(51□)	1 758(54□)
60□~70□	358(11□)	317(10□)
70□~80□	65(2□)	36(1□)
合计	3 275	3 275

备注:(1) 辨识概率中最高值为“0”表示未生成对抗性样本;(2) 表格中数值表示辨识概率最大值在左侧区间范围内的数量,括号内百分数为所占比例;(3) 辨识概率中最高值的取值范围应≥33.33□,实验结果中辨识概率最大值均小于 80□。

表 3 为辨识网络 FCNN 和 CNN 对对抗性样本的辨识概率分布区间统计情况。辨识网络对样本的辨识概率可以反映出辨识结果的可信赖程度,二者成正相关,辨识概率值越高,可以信赖的程度越高^[14]。在 3 275 个样本中,CNN 和 FCNN 分别有 432 和 587 个原始样本无对抗性样本生成。生成的对抗性样本中,我们从最低辨识概率 33.33□ 开始,第一个梯度为 6.67□,剩余梯度为 10□,对辨识概率分区间统计。辨识概率分布在 33.33□~40□ 分别有 97 个和 45 个;辨识概率分布在 40□~50□ 有 635 个和 552 个;辨识概率分布在 50□~60□ 分别有 1 758 个和 1 668 个;辨识概率分布在 60□~70□ 分别有 317 个和 410 个;辨识概率分布在 70□~80□ 有 65 个和 36 个。其中最集中的区间为 50□~60□,占到了 50□ 以上。

真实目标样本对应的对抗性样本既有被判决为地杂波样本的,也有被判决为假目标样本的。

表 4 统计了真实目标样本生成的对抗性样本被辨识网络辨识为其他样本的数量统计。其中大部分被对抗性样本被辨识为假目标样本,少数部分被辨识为地杂波样本。

表 4 真目标样本的对抗性样本被辨识为其他样本的数量统计

辨识网络	假目标样本	地杂波样本
CNN	2 662	26
FCNN	2 812	31

2.3 实验结果分析

该实验验证了辨识网络存在风险,对抗性样本可以使辨识网络的结果出错;表 2 表明辨识网络会受对抗性样本影响,辨识结果会发生改变;表 3 的结果告诉我们大多数(约 80%)原始样本都有与之对应的对抗性样本,但是,仍有少数原始样本无法生成对抗性样本。此外,表 3 中表明 80% 以上对抗性样本的辨识概率都集中在 40%~70% 区间内,与对原始样本 90% 以上的辨识概率相比较,辨识概率不高。辨识网络对对抗性样本输出 3 个概率中,最高值概率与次高值概率相差较小,有 357 个对抗性样本的差值小于 10%,甚至个别对抗性样本的差值小于 1%,这说明个别对抗性样本的可信赖度不强。

3 结束语

本文按照对抗性样本的模型生成了实测真实目标样本的对抗性样本,并对两种辨识网络的对抗性样本进行研究、分析和比较,总结了对抗性样本的特性。论文使用实测雷达点迹数据针对不同辨识网络生成了对抗性样本,总结并分析了这些对抗性样本的特点,这对使用机器学习算法进行雷达抗干扰研究具有重要意义。研究对抗性样本的生成条件,如何降低辨识网络的风险,将会是下一步的研究重点。

参 考 文 献

[1] TABACOF P, VALLE E. Exploring the space of adversarial images[C]// 2016 International Joint Conference on Neural Networks. Vancouver, BC: IEEE Press, 2016: 426-433.

[2] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]// The 32nd International Conference on Machine Learning. Lille, France: IEEE Press, 2015: 357-361.

[3] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. Computer Science, 2013(1): 1-9.

[4] 赵国庆. 雷达对抗原理[M]. 2 版. 西安: 西安电子科技大学出版社, 2012.

ZHAO Guoqing. Principles of radar countermeasure[M]. 2nd ed. Xi'an: Xidian University Press, 2012.

[5] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, Massachusetts: IEEE Press, 2015: 770-778.

[6] 罗 恒, 李增辉, 李建勋, 等. 基于 F-CNN 的雷达目标辨识算法[J]. 雷达科学与技术, 2019, 17(1): 89-93.

LUO Heng, LI Zenghui, LI Jianxun, et al. A factorized convolutional neural network based algorithm for radar target discrimination[J]. Radar Science and Technology, 2019, 17(1): 89-93.

[7] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. London: The MIT Press, 2016.

[8] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[J]. Computer Science, 2017, 13(3): 89-94.

[9] BOYD S, VANDENBERGHE L. Convex optimization[M]. London: Cambridge University Press, 2004.

[10] HSU K L, GUPTA H V, SOROOSHIAN S. Artificial neural network modeling of the rainfall-runoff process[J]. Water Resources Research, 2010, 31(31): 2517-2530.

[11] BYRD R H, LU P, NOCEDAL J, et al. A limited memory algorithm for bound constrained optimization[J]. Siam Journal on Scientific Computing, 1995, 16(5): 1190-1208.

[12] FEI Yun, RONG Guodong, WANG Bin, et al. Parallel L-BFGS-B algorithm on GPU[J]. Computers & Graphics, 2014, 40(1): 1-9.

[13] 张学工. 模式识别[M]. 北京: 清华大学出版社, 2010.

ZHANG Xuegong. Pattern recognition[M]. Beijing: Tsinghua Press, 2010.

[14] 关 欣, 何 友. 智能化雷达对抗情报处理技术研究[J]. 海军航空工程学院学报, 2005, 20(1): 12-23.

GUAN Xin, HE You. Research on technology of intellectualized radar countermeasures intelligence processing[J]. Journal of Naval Aeronautical Engineering Institute, 2005, 20(1): 12-23.

[15] 马 林. 雷达目标识别技术综述[J]. 现代雷达, 2003, 25(5): 22-26.

MA Lin. Review of radar automatic target recognition[J]. Modern Radar, 2003, 25(5): 22-26.

罗 恒 男,1992 年生,硕士研究生。研究方向为雷达抗干扰。

李增辉 男,1983 年生,博士,工程师。研究方向为信号处理。

李建勋 男,1979 年生,博士,高级工程师。研究方向为雷达对抗。

李一波 男,1970 年生,博士,副教授。研究方向为阵列信号处理。