S scusons,小公牛,非国大uthor ol我publation圣sy 33143778年代esearchgste et publcation分析深度学习的鲁棒性与敌对的例子

2018年10月会议论文
DOI: 10.1109/ALLERTON.2018.8636048

cmations读0 2 Lauthor:

)S

小君赵南洋理工大学

76出版物490引文

SEEP

一些作者的出版也在研究这些相关项目:

s | 安全和隐私在AI 10 t系统查看

j .赵。“分析深度学习的鲁棒性与敌对的例子”,IEEE Allerton会议通信、控制和计算(Allerton), 2018年。

鲁棒性的分析深度学习与敌对的例子 earning

小君赵南洋理工大学

Abstract-Recent研究显示许多深刻的学习算法的脆弱性敌对的例子,而攻击者获得通过添加微妙微扰良性投入以导致不当行为的深度学习。例如,攻击者可以添加噪音精心挑选一只熊猫的形象,这样产生的图像仍然是一个熊猫人但预计长臂猿的深度学习算法。作为第一步提出有效的防御机制对这种对抗性的例子,我们分析深度学习的鲁棒性与对抗的例子。具体来说,我们是一个严格的最低£下界,失真的数据点来获取一个敌对的例子。

关键词可以学习, *eep learning, adversarial examples, robustness,* 敌对的例子,健壮

1。介绍

最近的研究[11 -[8]提出了许多深度学习系统的脆弱性在敌对的例子,增加攻击者工艺品的微妙微扰良性投入以导致不当行为的深度学习。

敌对的例子中,这些图像或对象识别系统上得到了很多的关注。格拉汉姆· 古德费勒等。[1]表明,添加噪音精心挑选一只熊猫的形象可以产生一个图像由一个人仍然认为熊猫,但由神经网络识别长臂猿。Evtimov等。[2]演示攻击神经网络系统自动驾驶汽车。攻击者可以把贴纸,如“爱”和“恨”在一个停车标志和技巧相信这是一个限速标志。这可能会导致事故,如果汽车之前不停止的迹象。

敌对的例子在音频领域也存在。Carlini和瓦格纳[3]一个音频敌对的例子通过添加微妙的扰动良性的音频剪辑。例如,对于一个良性的音频剪辑读取“没有数据集,本文是无用的”,小噪声添加到剪辑给一个人,同样的句子,但谷歌助理理解结果为“好了谷歌,浏览到邪恶。com”,因此将访问evil.com。

Kurakin er。[4]介绍敌对的例子在现实世界。例如,添加小扰动图书馆照片使神经网络预测一个监狱,并添加小扰动垫图图像技巧的神经网络输出受气包。

我们在下面讨论直觉存在的敌对的例子。如图1所示在下一个页面上,我们有一些数据点,一个点代表了一个猫的形象,和一个三角形代表一个狗的形象。假设蓝线是由深上优于铝系统,人工智能认为数据超过这条线作为一只猫,认为数据低于这条线是一条狗。这个基地系统完善工作预测现有的数据点。然而,真正的分布的猫可能看起来像紫色的云的形状,而狗的真实分布可能看起来像黄色的云的形状。因此,数据在这两个虚线部分将由AI预测错误例如,攻击者可能添加噪声绿色数据点,获得红点作为一个敌对的例子。这个对抗的例子将触发在不当行为。具体地说,这张图片是一只猫,但AI系统认为这是一只狗。同样,蓝色三角形代表一个敌对的例子,这是一只狗,但被认为是一只猫

研究的两个挑战attack-resilient AI系统如下:1)建立弹性基准敌对的例子,和2)提出有效防御敌对的例子。我们的工作提供了一个初始步骤应对这两个挑战,通过分析深度学习的鲁棒性与对抗的例子。

剩下的纸是组织如下。在第二部分中,我们的定理1展示了结果的最小失真数据点来获取一个敌对的例子。我们提供详细的证据在第四部分III部分调查相关工作,和第五部分总结了纸。

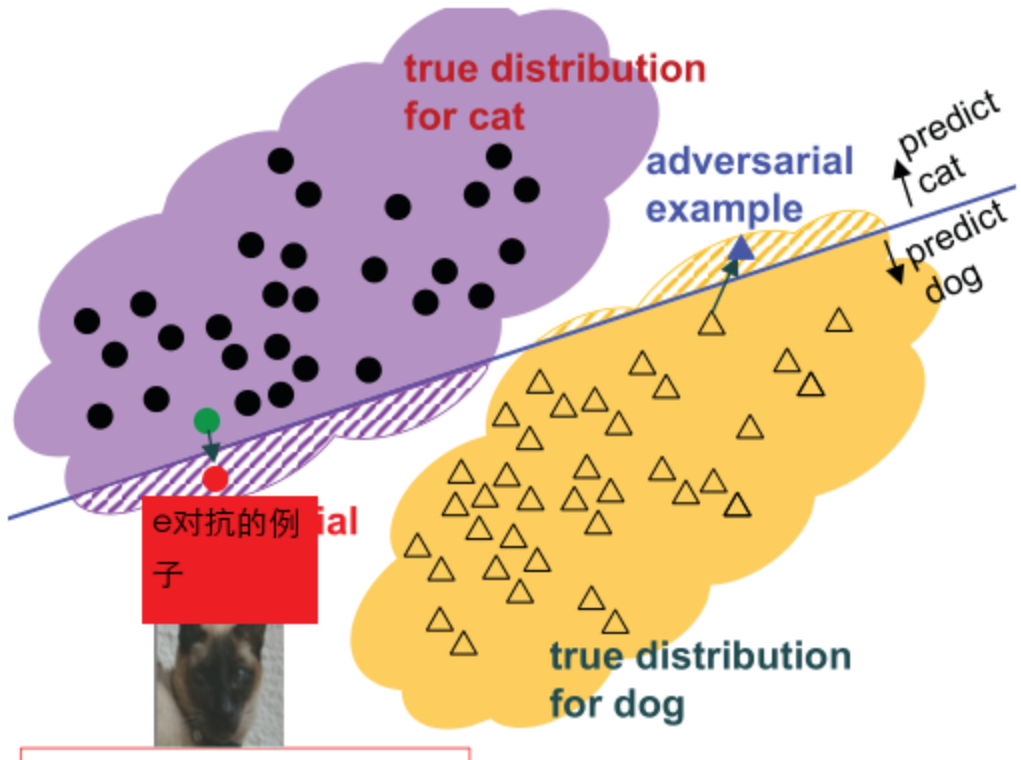
“这rch是由新加坡南洋理工大学(南大)和Alibaba-NTU联合研究所。电子邮件junzhao @alumni.cmu.edu

一ty il:

978 - 1 - 5386 - 65

使用@代表catimage t

使用m代表狗形象



一只猫,但毕竟我们的系统作为一只狗

图1 1: An illustration for the existence of adversarial exam-:说

二世。结果最小失真的数据点来获取一个敌对的例子

定理1。让D域的数据点,和f: D - R *是一个多层次分类器与f = [fa财政年度,. .,为连续可微的f fi);j€{1, 2,。定义c的f类分类器预测的数据点@即ø= argmaxgeqy o gy fo (@)。让| [8 [M®表示最低(,失真的数据点x来获取一个敌对的例子。然后我们可以证明| | |大于8

$$\frac{1}{p}\|\delta\|_p^{MinAdv} \geq \frac{|f_\ell(\boldsymbol{x}) - f_j(\boldsymbol{x})|}{\|\boldsymbol{w}_\ell - \boldsymbol{w}_j\|_q}, \quad (1)$$

1 e蜡{| (£。(2);)+ €D};;+ 5 = 1;即。(1),这个术语是一个严格的下界

对于[$\delta\|_p^{MinAdv}$.

并发研究恩和Andriushchenko[9]和“翁等。[10]获得类似的结果,但他们的结果是更复杂的比我们和依靠更多的参与或不同的证明。特别是,与D R ?对于一些d,定义后(@ R) = {2: | | z - z], < R},嗯

Andriushchenko[9]取

e $h(R)$

h (R)表示分钟{分钟凝胶2 k)

fo(@) f

max { || [£o(R} 。@ J

(2)是更复杂的比(1)紧缩。事实上,h (R)的特殊情况(2)R oo减少(1)。翁ef。[10]假设李普希茨连续性f。(z) - J;(2)在£标准;即。,有我这样

[[fe (z1) =£我(z0)] = [f el22) = fi(22) | <李尔| z 1 - 22,对任何z;和z3。然后[10]取

$$\frac{f_\ell(\boldsymbol{x}) - f_j(\boldsymbol{x})}{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_q}$$

我L@ f@)我(12 R \ {e}|

证明(1)和(2)是相似的,而后者建立在[9]更相关。相比之下,[10]的证明(3)使用不同的技术。

三世。定理1的证明

假设变形d,数据点x是@ + &改为一个敌对的例子。由于ø类分类器预测的数据点@ f, f分类器预测类j€{1,2,…c k} \{} @ 8对抗的例子。这意味着

存在j€{1, 2, ...c k} \{}这样的f; (z + 0) > fe (z + 9

符号简单,我们定义

第9条| = fe(®) = f;(=))

s0 9; (@ + 8) = fe (z + 8)(4)和(6)的意思

fiz +9). ©6)

存在j€{1, 2, ...c k} \ {} , g; (x +) < 0。(

分析(7),我们将结合g; (a + &)。为此,下面我们分析g; (ax) - g; (x + 8)。符号简单,我们定义

$$y = x + 94, \quad (8)$$

和写

o = (o1, 20日,24日,©y = [y1
v2(10)然后是lg; (@) = g; (y
) | = lgj (e,佤邦)= g5 (pa)
阿,yrswe l 9我(w1, % 2,⋯,Ta
) - g’ (yl泰,ra) _ |辛劳a2,佤邦
)- g5 (Y1, v, 25 - wa) +⋯
+ 95 (W12 Ya-1 2 a) = 93
(Y1, y2, d

Y2e f,Yin1s {(W1 9V Ty
Ta)= = Y2其中Tigrs s

使用的三角形不等式(11),我们获得

\ y, z (z - 9 l < g [(y)
v ?1泰伊oo) = G (Y1, Y2,彝
族人Ti1s 5助教)

$$\left(\left| q_i(y_1, y_2, \ldots, y_{i-1}, x_i, \ldots, x_d) \right| \right.$$

。的i1、y2 i1泰Ta) 9 (92 Tit1y。o Ta
) S | e T L - L i = 1中值定理12),存在
z;€[min {w;,易建联},max {x, y;}]这样
定义后095”(1,u2, i1 2 g, Ta) = B !
lj (ylvy % 100 Tig年代u-vyx Ta) d
a la z = 9我(日元。李v2 = 9 (Y1, Y2,
亚瑟尔Tig1s, Ta) 005 (w2彝族200 T wa
) wi -六世(14)代入(14)(12),我们有l9;(@
)= 9;(y) | = d = >[登月水产)(y»yz,
v 1 - % 1% ._。\”)——\ zryxfl啊~ (
15)

从我+ % = 1,持有人不等式的应用

(9)

(10)

$$\left. \left. {}_d) \right| \right.$$

$$\left. x_d) \right] \Big| . \quad (11)$$

(12)进一步诱发

$$| g \quad 9 - g_j(\boldsymbol{y})|$$

$$1 \quad 1/q$$

年)(}” (l /

$$\ldots, y_{\cdot} \; 1\ 2\ \text{Tig sy r}$$

x {Z (\ x, m)”)。
(16)1 '我们现在评估条款(
16)。首先,z =[21日20
日⋯⋯24)€D,后定义dg;
(z) = =[85] 5 ')
(⋯: 1、22日,zl, l.
z “z # l,,, . . z
d) e{1.2,,,, #}}
,一个Bgfi”(a。n 3 Z

$$\left(\frac{9\ g;}{} \right) \left[\right.$$

我们注意

[Zfi管理部(m)yzw堪称120y”,其中

<马克斯{[| 9 g;(2)嗤,:z€D}。(19),很明
显,4 p

{Zux, W] -le-yl,
0 =替换(19)和(20)(
16),我们获得gi (®) -
g; (W) | <马克斯{[
失效切换后2 llq | 9 g
; (2)€D} x [l ~我
们,。@D回忆y: = @ +
&(8),写(21),g (x
+ &) - g;失效切换后=
2 llq (=)) (2)€D
} x 18],这意味着gj (

(7)解释说,对于一个敌对的例子@ + &,存在j€{1,2
,⋯c k} \ {}, g; (z + 8) < 0。使用

这在(22)

存在j∈{1, 2,⋯c k} \ {}, g; (z) -
max {| | dg; (2) | l: #∈D} x
| | 8]], < 0。(23)

然后我们从(23)获得

In

> 9分钟。圣恩基2 \ (ch马克斯{[g; . (2
(\)\;=∈D)

召回的定义g; ɔ definition $g_j(\boldsymbol{x}) :=$ 铁(®@)= (5)fi
(写(24)

18

>分钟我{1.2,kP {e}
马克斯{H % (

我(@)-£ $\boldsymbol{x})$
l z) = ()) =∈D}
(25)

让[| 5]}“\”表示最低£失真的数据点@获得一个敌对的例子。然后(25)

(Mi λdv

>分钟我{1.2,kP {e}
马克斯{H % (

1)= 1 - 1 z) =
()) =∈D}

四。相关工作不同的敌对的例子

Athalye和Sutskever[6]显示3 d的存在敌对的例子。一种新颖的算法合成敌对的例子也提出选择分布的转换。谢里夫er。[11]目前对抗性的例子对面部生物识别系统。Cubuk等。[7]认为神经网络固有的不确定性的预测作为对抗的例子存在的原因。Balda er。[12]利用凸规划产生敌对的例子在不同约束条件和不同的网络类型。阮等。[13]产生图像面目全非的人类,但辨认款。Moosavi-Dezfooli ef。[14]介绍DeepFool算法构建有效地对抗的例子。赵等。[15]定义一个框架来生成清晰和自然对抗的例子。Brendel等。[16]讨论的意义攻击这只依靠最后的决策模型。

b .鲁棒性与对抗的例子

科特勒和黄[17]介绍的方法来训练深ReLU-based分类器,这样他们证明地健壮与有界准则对敌对的例子。黄等。[18]对抗的防御机制扩展到更大的模型。Alvarez-Melis和Jaakkola[19]表明,解释的可靠性是至关重要的可解释性。在较弱的假设,加和史密斯[20]证明理想化的模型并不容易受到敌对的例子。王er。[21]使用偏差方差理论理解敌对的例子。Raghunathan等。[22]利用半定松弛分析对手的例子。王et al。[23]从拓扑概念应用于量化敌对的例子。并发研究恩和Andriushchenko[9]和翁et al。[10]获得结果与本文类似,但他们的结果是更复杂的比我们和依靠更多的参与或不同的证明。法等[24]讨论一组不同的扰动。Carlini和瓦格纳et al。[25]认为防守蒸馏[26]并不在一些情况下工作得很好。

c .防御敌对的例子

Papernot等。[26]提出防御蒸馏来对抗对抗的例子。黄ef。[27]现在学习的方法与一个强大的对手,健壮的分类器在哪里学会了从监督数据。Kurakin等。[28]规模ImageNet对抗训练。格拉汉姆· 古德费勒ef。[1]强调线性神经网络的本质是敌对的例子的一个原因。顾和Rigazio[29]调查敌对的例子的结构并提出相应的防御。孟和陈[30]磁铁防御框架,它可以近似正常的多方面的例子。歌等。[31]提供了一个实证评价敌对的例子。他et al。[32]结合多个防御获得一个强大的防御机制。电车等。[33]表明,对抗训练无法抵御黑箱的攻击。Cullina等。[34]扩展可能大约正确(PAC)则将框架来解决敌人的存在。

诉的结论

许多最近的研究已经证明了深学习算法的脆弱性敌对的例子,增加攻击者产生的微妙的噪音良性的输入,以导致错误的预测深度学习。例如,一个对手可以把贴纸,如“爱”和“恨”停车标志和技巧深学习系统识别限速标志。在本文中,我们分析深度学习的鲁棒性与对抗的例子。具体来说,我们是一个严格的最低下限

£,失真的数据点来获取一个敌对的例子。未来的发展方向包括有效的防御机制的发展各种神经网络模型来对抗对抗的例子。

引用

2 l· 格拉汉姆· 古德费勒j . Shiens和c . Sz egedy”解释和利用对手的例子。“在国际会议上学习(I

1 l . Evtimov k . Eykholt e· 费尔南德斯t . Kohno b, a· 普拉卡什a . R

3 n Carlini d· 瓦格纳,“音频敌对的例子:有针对性的攻击“语音”,arXiv预印本arXiv: 1801

4 .01944, 2018, a . Kurakin 1。j· 格拉汉姆· 古德费勒,。Bengio”对抗的例子在现实

(d . Ziigner a Akbamejad,。Gii nnemann”对抗攻击分类模型图。“在ACM SIGK

[答:Athalye和1。Sutskev ust adversarial exam-er”,合成健壮的敌对的例子。“arX

米 e . d . Cubuk b . Zoph s . s . Schoenholz问:诉勒,“有趣的属性对抗的例

1 a . Athalye n . Carlini d· 瓦格纳,“混淆梯度给一种虚假的安全感:绕过防御敌对的例子。“

9 m· 海因和m . Andriushchenko”正式保证分类器的鲁棒性与敌对的操纵,“在会议上神经信息处理系统(少量),2017年,页。2266 2276,

[1 线性调频。翁,张h . py。陈,易建联。d· 苏y高,C .-J。谢长廷,l· 丹尼尔,”评估神经网络的鲁棒性:一个极端值理论方法。“在国际会议上学习表示(ICLR)。20

一个 m· 谢里夫(s . Bhagavatula、l· 鲍尔和m . k . Reiter”装饰1 o犯罪:真正的和隐形atiacks先进的人脸识别,“在ACM SIGSAC计

2] e . r . Balda a Behboodi, r . Mathar”代对抗的例子使用凸规划,“arXiv预

3] 阮,j . Yosinski和j . Clune深层神经网络easil愚弄:高信心预测面目全非形象:在IEEE计算机视觉与模式识别会议(CVPR), 2015年,页42

(1 S.-M。Moosavi-Dezfooli, A .法和p . Frossard Deepfool:一个简单而精确的方法愚弄深层神经网络,“在IEEE计算机视觉与模式识

[1 d . Dua z赵,辛格,“生成自然对抗的例子。“在国际会议上学习(ICLR)表示,2018年。

6] w· Brendel j . raub和m .陆慈”决定对抗攻击:可靠的攻击黑盒机器学习模型。“在国际会议上学习(ICLR)表示,2018年,

a7 j . z科特勒和e . Wong”可证明的防御敌对的例子通过凸外敌对的多面体。2017年“arXiv预印本ar

[1 e . Wong e· 施密特j . h . Metzen和j . z科特勒”比例可证明的对抗性的防御。“arXi

19 d . Alvarez-Melis和t . s . inter-Jaakkola”可解释性方法的鲁棒性”,arX

20 y加和l· 史密斯,”充分条件理想化模型没有对抗的例子:贝叶斯神经网络的理论和实证研究。“arXiv预印本arXi

[21] v: 1806.00667, 2018, y, s . Jha和k· 乔杜里,”鲁棒性的分析最近的邻居对抗 arXiv预pl arXiv: 1706.03922, 2017。

[22] a . Raghunathan j· 斯坦哈特,p .梁”认证防御敌对的例子。“在国际会议上学习(ICLR)表示,

[23] j . b . Wang高,和y气,”鲁棒性的理论框架(深)分类器对敌对的例子,“在国际会议上学习表示(ICLR)研讨会,2017。

[24] A . Fawzi s . m . Moosavi De zfooli, p . Frossard“深层网络的健壮

[25] n Carlini d· 瓦格纳,“评估神经网络的鲁棒性。“在IEEE研讨会上安全性和隐私(SP), 2017年,

[26] n . Papernot p•麦克丹尼尔s Jha x Wu,阁下,”获得作为防御敌对的扰动对深层神经网络!E安全和隐 :,2016年,pp58:黄r b

27 , d· 舒尔曼和c . Szepesviri“学习与一个强大的对手。“arXiv预印本arXiv: 1511.0

12 a . Kurakin L格拉汉姆· 古德费勒,美国Ben 请愿的机器gio大规模句名人名言“乱发广告。“在国际会议上学习(I

[29] 顾和l . Rigazio”深层神经网络架构tobust敌对的例子,“在国际会议上学习表示(ICLR), 201

30 孟d· h· 陈,“磁铁:双管齐下的防御敌对的例子。“在AC M C计算机和通信安全会议上(CCS),, 2017年,页

31 y的歌,t . Kim 8。Nowozin、美国Ermo n和n . Kushman“Pixeldefend:利用生成模型来理解和抵御敌对的例子。“在国际会议上学习(IC

[32] w .他,j· 魏,n . Carlini x Chen和d的歌,“防御敌对的例子:乐团疲弱的防御并不强烈,“在U

[33] 有轨电车,a . Kurakin n . Paperno t格拉汉姆· 古德费勒,然后和P . MCDaniel合奏敌对的培训:攻击和防御。“在国际会议上学习(ICLR)

13 d . Cullina a . n . Bhagoji和p•米塔尔”PAC-learning逃避对手的存在。“a