浙江科技学院学报，第 32 卷第 1 期，2020 年 2 月
Journal of Zhejiang University of Science and Technology
Vol. 32 No. 1, Feb. 2020

doi：10.3969/j.issn.1671-8798.2020.01.007

基于卷积神经网络的污点攻击与防御

胡慧敏[a]，钱亚冠[a]，雷景生[b]，马丹峰[a]
（浙江科技学院 a.曙光大数据学院；b.电子与信息工程学院，杭州 310023）

摘　要：深度神经网络易受对抗样例的攻击，该攻击主要通过对图像做细微的修改而使卷积神经网络识别出错。因此，为了模拟实际生活中车牌上的污点攻击，只对车牌图像添加局部扰动。首先使用 l_1 范数作为优化算法得到车牌图像中易被字符分类器识别错误的位置，然后继续在该图像中产生特定的扰动，最后将扰动加入到易被攻击错误的位置中。试验结果表明该攻击方法具有 90％的成功率，对车牌的字符识别造成了一定的影响。同时以对抗训练作为防御策略，取得了 98％的成功率。

关键词：对抗攻击；车牌识别；污点攻击；对抗训练
中图分类号：TP391.41　　**文献标志码**：A　　**文章编号**：1671-8798(2020)01-0038-06

Stain attacks and defenses against convolutional neural networks

HU Huimin[a], QIAN Yaguan[a], LEI Jingsheng[b], MA Danfeng[a]
(a. School of Sugon Big Data Science; b. School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: The deep neural network is vulnerable to the adversarial sample attacks, which make recognition of the convolutional neural network prone to errors mainly by making slight modifications to the image. Therefore, in order to simulate stain attacks on the license plate in real life, only local perturbation was added to the license plate image. First, l_1 norm was used as the optimization algorithm to identify the locations in the license plate image prone to recognition errors by the character classifier. Then, the specific perturbation was generated in the image. Finally, the perturbation was added to the locations vulnerable to attack errors. The experimental results show that the attack method boasts high success rate of 90％, which has a certain impact on character recognition of the license plate. Even more remarkably, the success rate of 98％ has been achieved by taking the adversarial training as the defense strategy.

Keywords: adversarial attacks; license plate recognition; stain attacks; adversarial training

收稿日期：2019-05-09

基金项目：浙江省自然科学基金项目(LY17F020011)；浙江省公益技术应用研究项目(LGG19F030001)；国家自然科学基金项目(61572163)

通信作者：钱亚冠(1976—　)，男，浙江省嵊州人，副教授，博士，主要从事人工智能、机器学习研究。E-mail：qianyg@yeah.net。

自 Krizhevsky 等[1]将卷积神经网络(convolutional neural networks，CNN)成功应用于 ImageNet 图像识别任务后，深度学习被广泛应用到各种智能识别领域，包括自动驾驶、人脸识别和语音识别等。虽然深度学习在特定领域的模式识别能力超出了人类水平，但最近的研究表明深度学习也面临多种安全威胁，容易受到对抗扰动的攻击[2-3]。比如在深度神经网络的输入图像中做些精心设计的修改就有可能引起错误的预测结果，甚至对错误的预测结果给出很高的置信度。这种添加了扰动的图像被称为对抗样例(adversarial examples)。现有的大多数对抗样例攻击是针对整张图像添加的扰动[4-5]，但也有部分攻击只进行局部的扰动[9-10]。本文提出的车牌字符污点攻击属于局部扰动的对抗攻击。

典型的对抗样例攻击方法是在卷积神经网络为线性的假设条件下对整张图像添加扰动，加入扰动的图像对人的视觉感受没有差异，但却能让分类器错误分类。将一张熊猫的图像加入特定的扰动，就能使卷积神经网络分类器将熊猫误分类为长臂猿。Moosavi-Dezfooli[4]提出了一种普适的扰动算法，使用 l_2 和 l_∞ 范数作为扰动度量，产生的扰动可使任意图像被深度神经网络错误分类。该方法用 2000 张图像在 ResNet 模型上进行攻击测试，获得了很高的攻击成功率，并发现扰动的大小与攻击成功率呈正相关。Karmon 等[6]通过求解最优扰动位置的方法来产生可视化的局部扰动补丁。该方法在不覆盖图像中的任何目标对象的情况下，将补丁贴在背景上就能使目标对象被错误分类。对抗攻击的另一个极端情况是只需篡改图像中的某个像素就能欺骗分类器[7]，攻击的目标类的置信度可达 97.47%。Moosavi-Dezfooli 等[8]提出用迭代的方法计算给定图像的最小范数的对偶扰动，试验结果表明，在具有相似的欺骗比的情况下，Deepfool 算法能够计算出比 FGSM 算法[3]更小的扰动范数。

随着智能交通系统的不断发展，车牌检测与识别(license plate detection and recognition，LPDR)在交通监控、公路收费站、停车场出入口管理等监控系统中有着广泛的应用。然而在处理车牌图像的过程中，基于深度学习的车牌识别系统[9-11]同样会受到对抗样例的攻击。对此，我们提出了一种新的对抗样例的攻击方法，与以往在整个图像上添加扰动的方法不同，攻击者只在车牌图像的局部位置恶意添加污点，伪装成无意中溅上的泥土，干扰车牌识别系统正确识别字符。这类对抗样例具有很强的隐蔽性和依赖性。为了防御这类攻击，我们使用了对抗训练的防御方法，即利用对抗样例作为训练样例，继续训练卷积神经网络，从而达到增强其鲁棒性的目的。试验结果表明，对抗训练能有效地防御这类污点型对抗攻击。

1　车牌识别基本原理

一个完整的车牌识别系统由检测和识别两个步骤组成，整体流程如图 1 所示。检测指对车牌进行定位，生成符合特定条件的边界框，进而识别指区分出边界框中的所有字符。检测方法分为传统的图像处理方法和当前流行的基于卷积神经网络的深度学习方法。传统的图像处理方法主要依赖于人工构造的图像特征提取车牌的形态、颜色或纹理[12]等属性信息。然而这些图像特征对噪声敏感，在复杂背景或不同的光照条件下可能会导致检测错误。而卷积神经网络提取的图像特征克服了这些缺点，在成功定位车牌的边界框后，进一步对边界框中的车牌字符进行识别。由于车牌中含有多个字符，因此识别的方式可分为单字符识别和多字符识别。单字符识别就是将车牌中的字符序列分割成独立的字符，接着使用光学

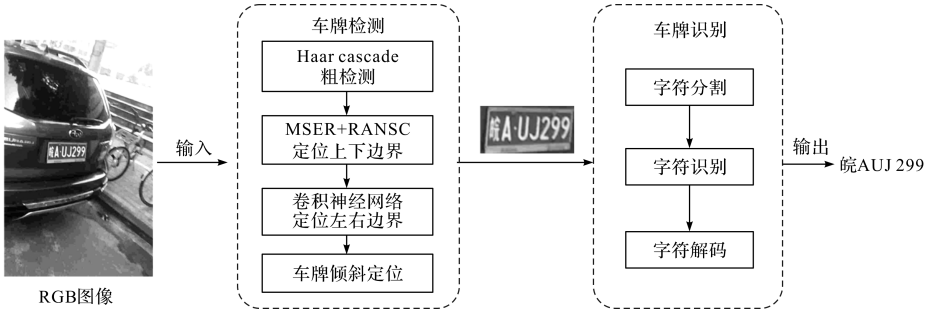


图 1　车牌识别系统流程

Fig. 1　Process of license plate recognition system

字符识别(optical character recognize,OCR)的方法进行识别。多字符识别是将整个车牌字符序列放入卷积神经网络中提取特征,再通过时间连接分类器(connectionist temporal classification,CTC)[13]方法输出字符序列。本文试验使用的车牌识别系统采用多字符识别方式,在多字符识别阶段使用了3个卷积层和2个循环与门(gated recurrent unit,GRU)层。GRU是长短时记忆(long short-term memory,LSTM)网络的一种变体,它比LSTM网络的结构更简单,而且识别效果更好,因此是当前车牌识别系统中非常流行的一种神经网络结构。

2 威胁模型与污点攻击方法

Papernot等[2]提出一种实用的黑盒攻击策略,通过不断的迭代查询获取训练数据,在本地建立一个近似的代理模型进行对抗攻击,从而在本地实现白盒攻击,构建出对抗样例,利用对抗样例的可转移性实现黑盒攻击。因此,我们用 $f_\theta(\cdot)$ 表示被攻击的卷积神经网络模型,用 $\tilde{f}_\theta(\cdot)$ 表示代理模型。通过向目标模型发送查询请求,可获得图像和类标签的输入输出对,构成本地训练数据集 $D=\{(x_1,f_\theta(x_1)),(x_2,f_\theta(x_2)),\cdots,(x_M,f_\theta(x_M))\}$。我们的目的就是在输入图像 x 加入尽可能小的扰动 δ 使被攻击的卷积神经网络模型错误分类,即:$\tilde{f}_\theta(x')\neq\tilde{f}_\theta(x)$。

2.1 威胁模型

对车牌识别系统攻击和防御建立在一定的威胁模型的假设上。威胁模型是对攻击者了解目标系统的知识和攻击的特异性作出的假设。根据攻击者对模型内部信息掌握的多少,可分为白盒攻击和黑盒攻击。白盒攻击假设攻击者几乎了解目标神经网络的所有信息,包括训练数据、模型架构等,目前大多数对抗攻击都是白盒攻击。黑盒攻击指攻击者无法知道神经网络模型的内部信息,通常只能获取输入样例和输出的预测标签。本研究采用黑盒攻击策略。攻击特异性根据攻击对手期望的误分类结果可分为有目标攻击与无目标攻击,即:$f_\theta(x')\neq y$(无目标攻击),或 $f_\theta(x')=y'$(有目标攻击),y' 是目标攻击的类且 $y\neq y'$。本文的污点攻击属于有目标攻击,例如把字符"A"误分类为"E"。

2.2 污点攻击方法

本文提出的污点攻击模拟车牌被泥土等污染的情况,攻击者的目的是用于伪装,逃避检测。攻击方法主要包括3个步骤:首先是通过优化算法找到容易被字符分类器识别错误的位置,即使用 l_1 范数为度量找到容易攻击的位置;其次是固定该位置以后,再使用优化算法,以 l_2 范数为度量产生扰动 δ;最后,在找到的位置上将扰动 δ 添加到干净样本 x 中,然后生成对抗样本 x':$x'=x+\delta$。我们把寻找最优扰动 δ 的过程建模为如下有约束优化问题:

$$\underset{\delta}{\operatorname{argmin}}\ \|\delta\|_p,\quad \text{s.t.}\quad f_\theta(x+\delta)=y^*。\tag{1}$$

式(1)中:$\|\delta\|_p$ 为向量范数,$p=1$ 或 2;y^* 为攻击的目标类别。用拉格朗日松弛法求解,可转化为如下的无约束优化模型:

$$\underset{\delta}{\operatorname{argmin}}\lambda\ \|\delta\|_p+J(f_\theta(x+\delta),y^*)。\tag{2}$$

式(2)中:J 为交叉熵损失函数;$\lambda>0$ 为拉格朗日系数。在找到扰动位置后,设置一个掩膜矩阵 $\boldsymbol{m}$,$\boldsymbol{m}\in\{0,1\}^{n\times n}$,对抗样本 x' 可表示为

$$x'=(\boldsymbol{U}-\boldsymbol{m})\otimes x+\boldsymbol{m}\otimes\delta。\tag{3}$$

式(3)中:$\boldsymbol{U}$ 为元素均为1的 n 阶矩阵;$\otimes$ 代表 hadamard 积。

具体步骤为:输入图像 x 和掩摸 $\boldsymbol{m}$,同时设置对抗训练的迭代次数;在 $f_\theta=y_{\text{true}}$ 与 $f'_\theta\neq y_{\text{target}}$ 时,用式(3)得到的结果与掩摸进行与操作;当循环次数大于迭代次数时结束对抗训练,输出加扰动图像 x'。图2展示了本文的3种对抗样例生成算法所产生的污点图像。

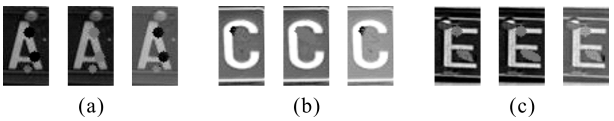


图2 车牌字符污点攻击图像

Fig. 2 Character stain attack images of license plate

3　对抗训练防御

对抗训练是提高神经网络鲁棒性的方法之一，通过在训练集中加入对抗样本，以平滑神经网络的输出对微小扰动的敏感度。采用文献[14]的对抗训练方法进行防御，在每批训练的数据中加入对抗样本，然后用损失函数作为衡量标准，即

$$f_{\text{loss}}=\frac{1}{n}\left(\sum_{i\in x}J(x_i\mid y_i)+\sum_{j\in x'}J(x_j\mid y_j)\right)。\tag{4}$$

式(4)中：$J(x\mid y)$ 为交叉熵损失函数；n 为每批对抗训练的总样本数。该算法首先初始化一个训练模型 M_0，然后将干净样本和对抗样本加入到模型中，并以式(4) 的值作为衡量标准来更新模型 M_0。

4　试验评估

4.1　数据集试验设置

试验所使用的数据集包含车牌的字符图像和车牌图像，分别训练单字符分类器和多字符分类器。用于训练单字符分类器的字符图像有 2 300 张，包括 A、B、C、D、E、F 六个类别的字符，该字符图像的大小为 60 pixel×35 pixel，而车牌图像的大小为 250 pixel×60 pixel。用于训练多字符分类器的图像包括质量较好与较差两种。质量较好的车牌图像，分类置信度大于 0.9，且置信度大于 0.95 的占 60%以上；质量较差的图像中至少有 20%的车牌识别置信度小于 0.9。试验选用两种不同质量的车牌图像，分别含有 A、C、E 字符的图像各 20 张，对它们进行多字符攻击和单字符攻击。

4.2　单字符攻击效果

单字符攻击是从质量较好的图像中裁剪出来的 20 张大小为 35 pixel×60 pixel 的字符图像，然后用不同形状和数量的污点进行试验，生成了 3 种不同形式的污点，如图 2 所示，产生不同单字符污点的过程如图 3 所示。3 种污点分别为：可变扰动的污点；固定颜色的污点；全局扰动中添加污点，即在整张图像添加扰动，并加入可变扰动的污点。不同形状和数量的污点攻击成功率见表 1 和表 2。

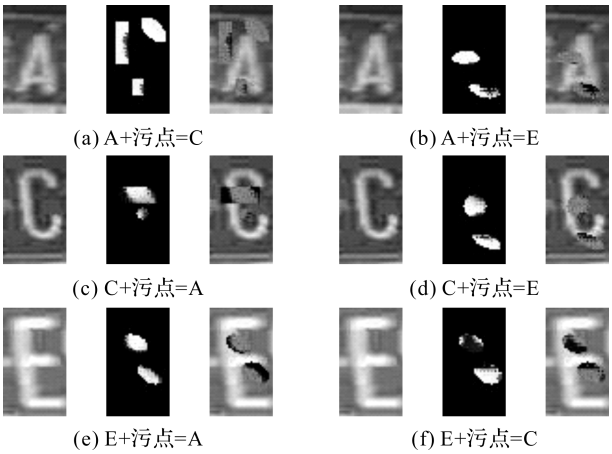


图 3　单字符污点攻击图像

Fig. 3　Single character stain attack images

从表 1 的试验结果中发现，当污点的数量由 1 增加到 3 时，其攻击效果逐渐增强。从两表的比较中还可以发现不同的污点形状(矩形和非矩形)攻击能力有所差异，当污点数量同时为 2 时，非矩形污点的攻击效果优于矩形污点。

表 1　单字符矩形污点的攻击成功率

Table 1　Success rate of single character rectangle stain attack　%

污点个数	A→C	A→E	C→A	C→E	E→A	E→C
1	100	15	55	80	65	95
2	100	85	70	85	80	95
3	100	100	75	90	85	100

表 2　单字符非矩形污点的攻击成功率

Table 2　Success rate of single character non-rectangular stain attack　%

污点个数	A→C	A→E	C→A	C→E	E→A	E→C
1	90	0	55	65	80	100
2	100	85	95	95	85	100
3	100	100	95	100	85	100

4.3　多字符攻击效果

在多字符攻击中，我们直接在车牌图像上添加污点，并用 HyperLPR[15] 中文车牌识别系统来测试攻击效果。为精确描述攻击成功率，只对能被系统正确识别的干净车牌图像进行污点攻击。多字符攻击的试验结果见表 3～4。污点类型 Ⅰ 表示补丁可视化的攻击方式，类型 Ⅱ 表示使用 RGB 值为(139，119，101)

时的攻击方式，类型Ⅲ表示全局扰动时的攻击方式。

表 3 多字符矩形污点攻击成功率

Table 3 Success rate of multi-character rectangle stain attack %

图像质量	污点类型	A		C		E	
		非 C	非 E	非 A	非 E	非 A	非 C
质量较好	Ⅰ	5	10	5	5	5	5
	Ⅱ	5	0	5	5	5	5
	Ⅲ	35	55	5	5	5	5
质量较差	Ⅰ	45	25	15	5	20	35
	Ⅱ	15	15	15	5	20	20
	Ⅲ	50	55	15	10	20	40

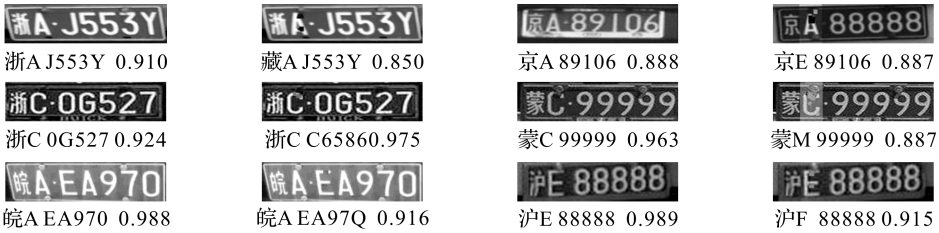
表 4 多字符非矩形污点的攻击成功率

Table 4 Success rate of multi-character non-rectangular stain attack %

图像质量	污点类型	A		C		E	
		非 C	非 E	非 A	非 E	非 A	非 C
质量较好	Ⅰ	5	15	0	0	5	5
	Ⅱ	0	0	0	0	5	5
	Ⅲ	25	40	30	0	5	5
质量较差	Ⅰ	30	40	15	0	15	20
	Ⅱ	10	15	20	0	20	10
	Ⅲ	60	70	30	0	30	25

从表 4 的试验结果发现，污点Ⅲ比污点Ⅰ和Ⅱ更易导致车牌识别系统分类错误，例如希望把字符 A 误判为非 C 字符和非 E 字符，在质量较好的图像上攻击成功率分别为 25%和 40%，而只要添加部分可变扰动的污点则成功率几乎为 0%。在图像质量较差的情形下，污点Ⅲ的攻击优势更明显，同样将字符 A 误判为非 C 和非 E 字符，攻击成功率分别达 70%和 60%。在表 4 中还可以看出污点Ⅰ比污点Ⅱ的攻击更有效，如将字符 A 误识别为非 C 和非 E 的情况下，污点Ⅰ在质量较好的图像上攻击成功率分别达 5%和 15%，而污点Ⅱ的攻击成功率为 0%，这是因为污点Ⅰ中的扰动是通过优化算法产生的，而污点Ⅱ只是找到了容易被攻击的位置，污点颜色是人为确定的。

从表 3 和表 4 的对比发现，字符在不同的污点形状下攻击成功率是不同的。在非矩形的污点攻击下将字符 C 攻击为非 E 的成功率为 0%，而在矩形污点下的成功率稳定在 5%，这表明非矩形污点的攻击效果好；而将字符 C 攻击为非 A 时，情况则相反。从表 3～4 中还观察到字符 C 的攻击效果比较差，这是因为从车牌识别系统中预测出来的图像置信度分数在 0.97 的有 80%，这表明如果图像置信度较高则相对难以攻击成功；反之亦然。此外，我们发现在对车牌识别系统的攻击中字符 E 容易被误分类为 F。污点攻击图像部分示例见图 4，其中，第 1、3 列是未加污点的原图像，第 2、4 列是加了污点后的图像，第 2、4、6 行是识别出的车牌号及车牌图像的置信度。



浙A J553Y 0.910　藏A J553Y 0.850　京A 89106 0.888　京E 89106 0.887

浙C 0G527 0.924　浙C C65860.975　蒙C 99999 0.963　蒙M 99999 0.887

皖A EA970 0.988　皖A EA97Q 0.916　沪E 88888 0.989　沪F 88888 0.915

图 4 污点攻击前后的车牌图像及识别结果

Fig. 4 License plate image and recognition results before and after stain attacks

4.4 对抗训练防御试验效果

将字符 A 至 F 的 2 030 张图像用于对抗训练，其中字符 A、C 和 E 是添加了扰动的图像。然后从真

实的车牌图像中选出含有字符 A、C 和 E 的图像进行测试，数量分别为 83、83 和 123 张，结果见表 5。在对抗训练前，单字符图像的攻击成功率最低为 68.3%，最高达 99.1%，可见我们提出的攻击方法有较高的成功率。用对抗训练进行防御，训练 100 次后用新的对抗样例（未参与训练）测试单字符分类器，字符 C 的攻击成功率最低，降到 1.3%，字符 E 的攻击成功率降到 12.2%，由此表明对抗训练可以有效提高模型对污点攻击的防御能力。

表 5　对抗训练前后的单字符攻击成功率对比

Table 5　Comparison of success rate of single character attack before and after adversarial training　%

字符攻击情况	A		C		E	
	A→C	A→E	C→A	C→E	E→A	E→C
对抗训练前	97.6	68.3	98.6	100	99.1	99.1
对抗训练后	9.7		1.3		12.2	

5　结　语

本研究提出了一种将攻击最优位置和局部污点相结合的方法用于攻击车牌单字符和多字符的识别系统。我们改变以往对整个图像进行扰动的做法，只针对车牌图像进行局部扰动，模拟车牌上的污点，这样不仅能巧妙地逃过人类视觉系统的观察，还能使基于卷积神经网络的车牌分类器出现预测错误。通过优化算法设计实现了车牌污点攻击，试验结果表明该攻击的成功率高，对目前的车牌分类器构成了一定的威胁。同时我们发现，在多字符攻击试验中图像质量的好坏会对攻击能力产生影响，质量差的图像更容易被攻击成功。针对这类威胁，我们提出了基于对抗训练的防御方法，试验结果表明它具有较好的防御效果。

参考文献：

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C]//Advances in Neural Information Processing Systems. Lake Tahoe: NIPS foundation, 2012: 1097.

[2] PAPERNOT N, MCDANIEL P, SINHA A, et al. Towards the science of security and privacy in machine learning [J]. Computer Research Repository, 2016, 1611: 3814.

[3] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial example[C]//International Conference on Learning Representations. San Diego: Computational and Biological Learning Society, 2015: 6572.

[4] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1765.

[5] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//International Conference on Learning Representations. Banff: ACM, 2013: 1312.

[6] KARMON D, ZORAN D, GOLDBERG Y. Lavan: localized and visible adversarial noise[C]//International Conference on Machine Learning. Piscataway: IEEE, 2018: 2512.

[7] SU J W, VARGAS D V, KOUICHI S. One pixel attack for fooling deep neural networks[C]//IEEE Transactions on Evolutionary Computation. Xiamen: IEEE and College of Computer Science and Technology, 2019: 1710.

[8] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2574.

[9] LI H, Shen C H. Reading car license plates using deep convolutional neural networks and lstms[J]. Computer Research Repository, 2016, 1601: 5610.

[10] SILVA S M, CLÁUDIO R J. License plate detection and recognition in unconstrained scenarios[C]//European Conference on Computer Vision. Berlin: Springer. 2018: 593.

[11] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2574.

（下转第 63 页）