面向深度学习系统的对抗样本攻击与防御

凌 祥 纪守领 任 奎
浙江大学

关键词：深度学习 对抗样本 攻击算法

引言

深度学习技术近年来取得了重大突破，成功应用于图像处理、自然语言处理、语音识别、医疗诊断等多个领域。在图像分类、语音识别等模式识别任务中，深度学习的准确率甚至超越了人类[1]。

尽管深度学习解决某些复杂问题的能力超出了人类水平，但最近的研究表明，深度学习技术也面临多种安全性威胁[2]。2013年，塞格德(Szegedy)等人[3]首先在图像分类领域发现了一个非常有趣的“反直觉”现象：攻击者通过构造轻微扰动来干扰输入样本，可以使基于深度神经网络(DNN)的图片识别系统输出攻击者想要的任意错误结果。研究人员称这类具有攻击性的输入样本为**对抗样本**，其生成算法即为对抗样本的攻击算法。随后越来越多的研究发现，除了DNN模型之外，对抗样本同样能成功地攻击强化学习模型、循环神经网络(RNN)模型等不同深度学习模型，以及语音识别、自动驾驶、文本处理、恶意软件检测等不同的深度学习应用领域[4]。

对抗样本攻击

为了解释对抗样本存在的原因，研究人员提出了各种各样的假设和推理[4]，包括DNN模型的极度非线性、训练正则化与模型均化不足、输入空间高维度、线性假设等。根据不同的假设和推理，研究人员设计并实现了多种不同的攻击算法来生成对抗样本，试图攻击不同应用场景下的深度学习模型。

攻击算法分类

一般来说，对抗样本的攻击算法可以简单地分为非目标攻击和目标攻击。非目标攻击只要求深度学习模型将对抗样本误分类，对误分类的标签不作要求；目标攻击则要求深度学习模型将对抗样本误分类成攻击者指定的标签。考虑到攻击算法中是否迭代地求解对抗样本，非目标攻击和目标攻击可以进一步细分为非目标单步攻击、非目标迭代攻击、目标单步攻击和目标迭代攻击。

表1 对抗样本攻击算法总结

非目标攻击	单步攻击	FGSM[5], R+FGSM[6]
	迭代攻击	BIM[7], PGD[8], U-MI-FGSM[9], DeepFool[10], UAP[11],OptMargin[12]
目标攻击	单步攻击	LLC[7], R+LLC[6]
	迭代攻击	ILLC[7], T-MI-FGSM[9], JSMA[13], C&W[14], EAD[15], stAdv[16]

表1总结了当前主流的攻击算法。在非目标攻击算法中，单步攻击主要包括FGSM[5]及其变体R+FGSM[6]攻击，其核心思路是沿着梯度的反方向添加扰动，从而拉大对抗样本相对于原样本的距

离来产生对抗样本。对于非目标迭代攻击而言，对FGSM这类攻击的直接扩展是迭代地采取多个小步骤扰动，在每个扰动之后调整扰动方向以达到攻击的目的，这类迭代算法包括BIM[7], PGD[8], U-MI-FGSM[9]等。此外，非目标迭代攻击还包括DeepFool, UAP和OptMargin等攻击算法。DeepFool攻击[10]主要通过迭代生成对抗扰动，将位于分类边界内的图像逐步推到边界外，直到出现错分类；UAP攻击[11]与DeepFool类似，都是迭代产生对抗扰动将图像推出分类边界，不过UAP攻击只需要生成一个扰动就能对其他图像达到攻击效果；OptMargin攻击[12]利用高维数据集中较大邻域内的信息生成鲁棒性更强的对抗样本，并且这种对抗样本可以有效地绕过当前基于邻域分类的防御方法。

在目标攻击算法中，单步攻击主要包括LLC[7]和R+LLC[6]两种攻击算法。事实上，这两种攻击算法分别对应FGSM和R+FGSM攻击。不同的是，LLC和R+LLC算法用DNN分类器预测的最不可能类别的标签来代替FGSM算法中使用的真实标签，然后从原始图像中减去计算出来的扰动从而得到对抗样本。对目标迭代攻击而言，类似地扩展LLC攻击可以得到ILLC[7]和T-MI-FGSM攻击算法[9]。此外，研究人员还在不断地提出一系列新的目标攻击算法。JSMA攻击[13]利用雅克比矩阵，计算样本从输入到输出的显著图，进而通过修改小部分显著性大的输入特征，达到改变模型输出从而欺骗分类器的目的;C&W攻击[14]事实上包括三个不同的攻击算法，分别通过限制对抗样本与原样本之间的L_0、L_2和L_∞距离使得扰动变得几乎不可察觉，并且可以通过调节攻击目标函数的k来调节生成的对抗样本置信度；EAD[15]是一种基于弹性网络正则化的攻击算法，该算法将对抗样本攻击深度学习模型的过程形式化为弹性网络正则化的优化问题，显著增强了对抗样本攻击的迁移性；stAdv攻击算法[16]是一种基于空间变换的对抗样本生成方法，而非直接改变图片的像素值，因而可以生成人眼感知更加不可区分的对抗样本。

攻击效用评估

针对攻击算法而言，攻击效用指的是攻击者利用攻击算法实际生成的对抗样本的攻击能力。如何评估当前对抗样本攻击算法的攻击效用是一个非常重要的问题。一般而言，利用攻击算法生成的对抗样本应该满足以下性质。

误分类性

误分类(misclassification)是对抗样本中最本质最核心的性质。对非目标攻击而言，误分类性要求深度学习分类器将对抗样本误分类成其他任何标签，即不同于原标签；而目标攻击的误分类性则要求将对抗样本误分类成攻击者指定的标签。

通过计算攻击算法生成的所有对抗样本的**误分类率**和**误分类置信度**，可以有效地评估攻击算法的误分类性。具体而言，利用攻击算法攻击一组测试样本生成对应的攻击样本集，误分类率指的是攻击样本集中所有成功欺骗分类器的样本数占总数的百分比。一般来说，误分类率越大，表明攻击算法的误分类能力越强。误分类置信度则用来进一步衡量攻击算法的误分类性，具体定义为对抗样本集中所有成功攻击的对抗样本被误分类的标签的平均置信度。误分类标签的置信度越大，表明该攻击算法生成的对抗样本会以更大的置信度被DNN分类器误分类。

不可见性

不可见性(imperceptibility)是对抗样本攻击成功的必要条件，要求攻击算法生成的“扰动”对人眼不可见，即对抗样本中的扰动应该是“细微的”且人眼无法分辨。因此，对抗样本攻击算法一般会将不可见性纳入攻击算法的目标函数中。但另一方面，如何衡量对抗样本中扰动的不可见性是一个极具挑战性的问题。

一般而言，攻击算法中最常见的是通过L_p距离来估计对抗样本与原样本的差距，常用的L_p距离包括：L_0、L_2和L_∞。L_p距离越小，表明对抗样本与原样本的差距越小，对抗样本中扰动越“不可见”。因此，通过计算所有攻击成功的对抗样本与其对应原样本之间L_p距离的平均值，可以估计该攻击算法生成对抗样本的不可见性。

另一种度量方式是采用常见的SSIM距离[1]来

衡量对抗样本与原样本之间的相似度。两者的相似度越大，攻击样本中的扰动越不可见。类似地，通过计算所有攻击成功的对抗样本与原样本之间的SSIM距离的平均值，也可以估计攻击算法生成对抗样本的不可见性。

鲁棒性

鲁棒性(robustness)表示对抗样本在物理世界中仍保持其攻击深度学习模型的能力。在现实世界中，图像、语音等样本在被DNN等深度学习模型预测前不可避免地会经过各种各样的转换处理过程，包括可能的自然噪声、输入数据的预处理等过程。因此，对抗样本的鲁棒性会对攻击算法在现实世界中是否可以成功攻击产生直接影响。

那么，如何度量攻击算法的鲁棒性？由于现实世界中可能存在的转换处理方法不胜枚举，无法直接度量攻击算法生成的对抗样本对所有转换方法的鲁棒性。一个常见的方式是选取若干常见的转换方法作为衡量攻击算法鲁棒性的代表，如高斯噪声、高斯去噪和图像压缩等。因此，通过计算所有攻击成功的对抗样本再次经过某种转换后，仍被误分类的样本占全部对抗样本百分比的大小来估计攻击算法的鲁棒性强弱。

攻击效率

攻击算法生成对抗样本的效率也是攻击算法的重要性质之一。原因有两点：其一，攻击者需要了解和对比每种攻击算法实际运行的效率如何；其二，基于对抗训练的防御方法在训练过程中会利用攻击算法生成的对抗样本，因此对抗样本的生成效率会影响对抗训练的防御效率。攻击效率指的是攻击算法生成对抗样本所需的时间，一般用每个对抗样本的平均生成时间来代替。生成对抗样本的时间越短，攻击算法的效率越高。

实际攻击案例

大多数攻击算法的有效性都是在基准数据集上或者受控实验环境下验证的，缺乏真实场景下攻击算法的有效性验证。为了进一步证明对抗样本攻击在真实物理世界中仍然存在威胁性，研究人员将对抗样本的攻击算法广泛应用到很多实际场景中。

手机拍照攻击

为了证明对抗样本的威胁性在物理世界中同样存在，库拉金(Kurakin)等人[7]将FGSM、BIM和ILLC等攻击算法生成的对抗样本进行打印，然后用手机对打印后的对抗样本拍照，再将拍照后的对抗样本输入到TensorFlow Android Camera Demo的应用中对其进行图像分类。实验表明，大部分经过打印、拍照等复杂处理后的对抗样本，仍然可以欺骗图像识别程序。

人脸识别系统攻击

人脸识别是计算机视觉中非常重要的研究技术，已经被广泛地应用于银行、交通、学校等场所的监控和访问控制系统中。针对当前最先进的人脸识别系统，谢里夫(Sharif)等人[30]设计并实现了一种攻击方法，可以使得攻击者逃避系统识别或者冒充其他人。如图1所示，攻击者首先打印一幅添加过扰动的眼镜框，接着佩戴打印后的眼镜框并由人脸识别系统验证，最终佩戴眼镜的攻击者可以冒充其他人，从而逃避人脸识别系统的识别。如图1所

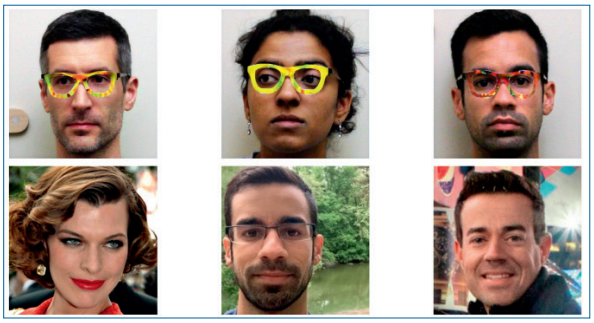

图1 人脸识别系统的攻击示意图[30]

[1] SSIM是Structural Similarity的简称，即结构相似性。它是一种衡量电视、电影或者其他数字图像、视频的主观感受质量的一种方法，该方法首先是由德州大学奥斯汀分校的图像与视频工程实验室提出的。

示，第一行表示三位佩戴眼镜的攻击者，第二行表示对应的攻击目标，即第一行的攻击者会被人脸识别系统识别成第二行的攻击目标（对于人脸识别系统，误分类的后果就是系统会将一个人识别成另外一个人，并非特别地将男人识别成女人）。

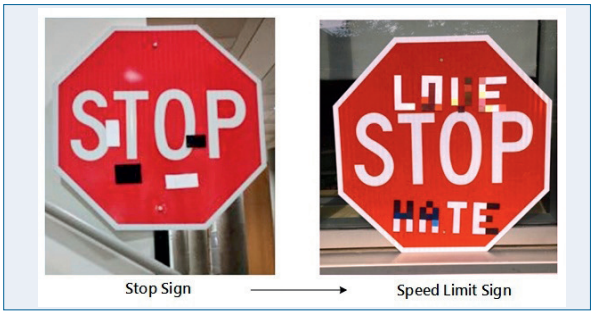

图2 路标识别的攻击示意图[32]

路标识别攻击

在无人驾驶领域，物体识别是一项非常重要的技术，包括识别路标、行人等。叶夫季莫夫 (Evtimov) 等人 [32] 表明对抗样本能够在各种物理条件下存在，包括视角、距离和分辨率等因素的变化。基于这些观察，作者提出了两种实际的攻击方法：其一是海报打印攻击，攻击者将利用 C&W 等攻击算法生成的对抗样本打印成海报，然后将其覆盖于真实的路标之上；其二是贴纸扰动攻击，在纸张上打印扰动并且将其贴到真实的标志上。通过这两种攻击方法将扰动添加到物理世界的路标中，可以构造物理意义的对抗样本。如图 2 所示，通过贴纸扰动攻击，Stop 路标很容易被识别成 Speed Limit 路标。

逃避恶意软件检测

针对恶意软件检测系统，格罗斯 (Grosse) 等人 [31] 发现利用对抗样本对恶意软件检测系统进行攻击，可以达到逃避恶意软件检测的目的。但是与图像识别相比，恶意软件检测领域中对抗样本的生成引入了其他的限制。例如，输入空间不再是连续输入而是离散输入，并且生成的对抗样本要和原样本具有相同的恶意攻击能力。在该攻击中，作者主要借鉴了 JSMA 攻击生成对抗样本的方法并加以改进，证明了对抗样本在恶意软件检测领域的可行性。

防御方法

对抗样本的存在使得深度学习在某些安全敏感领域的应用受到限制甚至是严重的威胁。因此，如何对对抗样本的攻击行为进行有效的防御是当前深度学习安全领域极具挑战性的问题。

防御方法现状

当前针对对抗样本的防御方法主要有完全防御和检测防御两种。完全防御方法的目标是使得防御后的 DNN 分类器能够将对抗样本识别为正确的标签；而检测防御方法只需要识别出输入样本是否为对抗样本即可，无须识别对抗样本本身真实的标签。对完全防御方法而言，可以进一步分为对抗训练 (adversarial training)、梯度掩蔽 (gradient masking)、输入转换 (input transformation) 和基于邻域分类 (region-based classification) 四种。表 2 总结了当前对抗样本防御方法。

表2 对抗样本防御方法总结

完全防御	对抗训练	一般对抗训练[17]，PGD对抗训练[8]，集成对抗训练[6]
	梯度掩蔽	深度压缩网络[18]，蒸馏防御[19]，输入梯度正则化[20]
	输入转换	集成输入转换[21]，输入随机转换[22]，PixelDefense[24]，温度计编码[23]
	分类方法	基于邻域分类[25]
检测防御	基于局部本征维数的检测[26]，Feature Squeezing[27]，MagNet[28]	

完全防御

对抗训练防御方法从训练数据集入手，在模型训练过程中不断加入对抗样本，构建鲁棒性更好的模型，达到防御对抗样本的目的。根据训练过程加入对抗样本类型的不同，研究人员提出了多种不同的基于对抗训练的防御方法。如一般对抗训练 [17]、PGD 对抗训练 [8] 和集成对抗训练 [6] 防御方法是在训练过程中分别加入 LLC、R+FGSM 和 PGD 三种不同攻击算

法所产生的对抗样本。一般而言，对抗训练防御方法大多在对抗训练的过程中添加对抗样本，并且这些对抗样本是通过攻击自身模型所产生的。然而，集成对抗训练防御方法将对抗训练的过程与对抗样本的产生过程分开，在对抗训练的过程中加入对抗样本是通过攻击其他模型产生的，增加了所加入对抗样本的多样性，从而提高模型抵御其他攻击方法的能力。

由于大多数攻击算法都使用模型的梯度信息来生成对抗样本，因此**梯度掩蔽**防御通过隐藏模型训练中的梯度信息，从而使得攻击算法很难通过梯度求解方法攻击模型。深度压缩网络方法 [18] 在训练过程中引入压缩自编码器的平滑惩罚项，使得模型的输出变化对输入敏感性降低，从而达到隐藏梯度信息的目的；蒸馏 (distillation) 防御方法 [19] 利用蒸馏方法训练两个串联的 DNN 模型来提高模型预测的鲁棒性；输入梯度正则化 [20] 在训练的目标函数上惩罚输出对于输入的变化程度，可以在一定程度上限制小的对抗扰动不会大幅改变最终模型的预测结果。

输入转换方法试图通过各种转换方法减少待预测样本中可能存在的扰动量，之后将转换后的样本直接输入到原模型来进行预测。输入转换防御方法的优点是既不需要改变训练数据集进行重新训练，也不需要改变原模型的结构。集成输入转换方法 [21] 整合了最常用的 5 种图像预处理和转换方法，直接对待预测样本同时进行 5 种图像转换，提高了模型预测对抗样本的准确度。同样地，输入随机转换方法 [22] 在待预测样本输入到原模型之前增加额外两层随机转化过程（包括随机调整大小和随机填充），再用原模型进行预测。PixelDefense 防御 [24] 利用 PixelCNN 生成模型将对抗样本转换到正常样本空间，将转换后的样本再输入到原模型进行预测。温度计编码 (thermometer encoding) 方法 [23] 将连续的输入样本使用温度计编码进行离散化，因此该方法在训练和预测阶段均使用的是温度计编码后的样本。

基于邻域分类防御从待预测样本邻域范围内随机选取若干个样本，利用原模型对取样后的所有样本进行预测，再采用多数表决方式选择预测标签最多的作为待预测样本最终的标签 [25]。

检测防御

尽管上述完全防御方法的期望很高，但是很多实验表明其实际效果并不好。因此，考虑到完全防御方法的困难性，研究人员提出了多种对抗样本的检测方法。检测防御方法仅用来判断样本是否为对抗样本，而不用识别出样本真正的标签。

基于局部本征维数 (LID)[26] 的检测方法利用对抗样本的 LID 值远大于正常样本的性质来识别对抗样本与正常样本。Feature Squeezing 方法 [27] 在 DNN 分类器中分别添加了两个外部模型，分别用来减少每个像素的颜色位深度和进行像素值的空间平滑。比较测试样本与通过外部模型处理后的样本在经过分类器预测后的差异，如果差异很大，则测试样本会被认为是对抗样本。

事实上，MagNet[28] 包含了完全防御和检测防御两种方法。MagNet 首先使用 Detector 来检测扰动量大的对抗样本，直接丢弃；然后针对扰动量小的对抗样本，使用 Reformer 努力将其转化成正常样本，最后再交由原模型识别。

当前防御方法的挑战

尽管当前对抗样本的防御方法取得了一定的效果，但是也存在很多的局限性与挑战。

基于对抗训练的防御。一方面由于对抗训练防御方法不仅需要大量的正常样本，而且需要大量的对抗样本，极大地增加了训练的时间和所需资源，使得该防御很难在实际场景的大规模数据集上使用；另一方面，由于训练过程中只能加入由已知攻击产生的且有限的对抗样本，因此对抗训练防御通常只对与加入训练同类型的对抗样本有效，对其他攻击产生的对抗样本不具有泛化能力。

梯度掩蔽防御。有研究人员认为这种防御方法非常容易被绕过，甚至认为这是一个“失败”的防御 [2]。一个简单绕过该防御的方法是，攻击者通过训练一个与防御后的模型相似的替代模型，进而通过使用替代模型的梯度来构造对抗样本，从而达到绕过防御后模型的目的。同样地，梯度掩蔽方法需要改变模型结构并重新训练分类器，进一步增加

了该防御方法在工程上的复杂性。

基于输入转换的防御。一般不改变任何模型网络结构和训练数据集，但是需要对待预测样本进行转换处理。理论上，输入转换方法对任何种类的攻击样本都有一定的防御效果，但是实验表明这种防御方法在对抗样本预测上的误报率和漏报率较大。

基于邻域分类的防御。主要利用待预测样本邻域范围内其他样本的预测值。该方法有两个非常重要的参数，分别是邻域范围半径 r 和采样个数 N。但是，r 越大，分类器预测正常样本的准确率越低；而 r 越小，该方法对对抗样本的防御效果就会越差。同样地，N 越大，该方法的效率会成倍数增加；N 越小，防御效果也会越差。

检测防御。一方面这类方法对正常样本的分类准确率会降低，另一方面这类方法并没有说明检测结果为“对抗样本”的解决方法。

总的来说，当前的防御方法面临很多挑战。一方面，每种防御方法都只能抵御有限的对抗样本，并且很容易被不断进化和变种的对抗样本绕过；另一方面，现有防御方法都是一种被动式的防御，只能防御某一类攻击，不能够解决 0-Day 攻击[2]等未知风险。来自麻省理工学院 (MIT) 和加州大学伯克利分校 (UC Berkeley) 的研究人员研究了 ICLR 2018 收录的八篇关于对抗样本的论文中的防御方法鲁棒性，发现其中七种防御方法可以通过改进的攻击算法来攻破[29]。

深度学习安全研究展望

随着深度学习在自动驾驶、智能安防、智慧医疗等关键领域的深入应用，深度学习模型面临的安全性威胁也日趋严重。2017 年 4 月，国务院印发《新一代人工智能发展规划》，强调建立人工智能安全监管和评估体系，构建人工智能安全监测预警机制，推动以深度学习为代表的人工智能安全。事实上，深度学习在其生命周期中都存在安全性威胁，包括数据收集、模型训练和模型推理与应用。例如，数据收集过程可能受到投毒攻击而影响训练数据质量，模型设计时可能由于设计不当或被植入后门等逻辑漏洞引发欺骗攻击，模型训练时也可能因为攻击者注入异常数据造成模型准确率下降，模型在推理与应用的过程中容易受到对抗样本攻击、逆向攻击或成员攻击。然而现有研究大多只关注深度学习生命周期某一阶段或某个应用场景下的安全问题，无法保障深度学习模型全生命周期、多应用场景的安全性。因此，针对不同类型的攻击、模型和应用，构建跨领域、系统级的深度学习安全性评估和防御的理论体系，促进深度学习及其相关服务安全可信发展，是深度学习安全领域的一个研究热点。 ■

凌　祥

CCF 学生会员。浙江大学计算机科学与技术学院博士生。主要研究方向为网络安全与深度学习安全。
lingxiang@zju.edu.cn

纪守领

CCF 专业会员。浙江大学“百人计划”研究员，浙江大学网络空间安全研究中心主任助理、信息安全专业系主任，国家“青年千人”。主要研究方向为人工智能安全、数据安全隐私与大数据分析。sji@zju.edu.cn

任　奎

CCF 专业会员。浙江大学网络空间安全研究中心主任，国家千人计划特聘教授。主要研究方向为物联网安全、云安全和隐私保护。kuiren@gmail.com

参考文献

[1] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. *International Journal of Computer Vision*, 2015, 115(3):211-252.

[2] Papernot N, Mcdaniel P, Sinha A, et al. Towards the

[2] 又叫零时差攻击，是指被发现后立即被恶意利用的安全漏洞。通俗地讲，即安全补丁与瑕疵曝光的同一日内，相关的恶意程序就出现。这种攻击往往具有很大的突发性与破坏性。

Science of Security and Privacy in Machine Learning[J]. 2016.

[3] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. Computer Science, 2013.

[4] Yuan X, He P, Zhu Q, et al. Adversarial Examples: Attacks and Defenses for Deep Learning[J]. 2018.

[5] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[J]. Computer Science, 2014.

[6] Tramèr F, Kurakin A, Papernot N, et al. Ensemble Adversarial Training: Attacks and Defenses[J]. 2017.

[7] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world[J]. 2016.

[8] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[J]. 2017.

[9] Dong Y, Liao F, Pang T, et al. Boosting Adversarial Attacks with Momentum[J]. 2017.

[10]Moosavi Dezfooli, et al. Deepfool: a simple and accurate method to fool deep neural networks[C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). No. EPFL-CONF- 218057. 2016.

[11]Moosavidezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[J]. 2016:86-94.

[12]Warren He, et al. Decision Boundary Analysis of Adversarial Exam- ples[C]//Proceedings of 6th International Conference on Learning Representations(ICLR). 2018.

[13]Papernot, Nicolas, et al. The limitations of deep learning in adversar- ial settings[C]//Proceedings of Security and Privacy (EuroS&P), 2016 IEEE European Symposium on. IEEE, 2016.

[14]Carlini, Nicholas, and David Wagner. Towards evaluating the ro- bustness of neural networks[C]// Proceedings of Security and Privacy (SP), 2017 IEEE Symposium on. IEEE, 2017.

[15]Chen P Y, Sharma Y, Zhang H, et al. EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples[J]. 2018.

[16]Xiao C, Zhu J Y, Li B, et al. Spatially Transformed Adversarial Examples[J]. 2018.

[17]Kurakin A, Goodfellow I, Bengio S. Adversarial Machine Learning at Scale[J]. 2016.

[18]Gu S, Rigazio L. Towards Deep Neural Network Architectures Robust to Adversarial Examples[J]. Computer Science, 2015.

[19]Papernot, Nicolas, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of Security and Privacy (SP), 2016 IEEE Symposium on. IEEE, 2016.

[20]Ross A S, Doshivelez F. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients[J]. 2017.

[21]Guo C, Rana M, Cisse M, et al. Countering Adversarial Images using Input Transformations[J]. 2018.

[22]Xie C, Wang J, Zhang Z, et al. Mitigating Adversarial Effects Through Randomization[J]. 2017.

[23]Buckman, Jacob, et al. Thermometer encoding: One hot way to resist adversarial examples[C]//Proceedings of 6th International Conference on Learning Representations(ICLR). ICLR,2018.

[24]Song Y, Kim T, Nowozin S, et al. PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples[J]. 2018.

[25]Cao, Xiaoyu, and Neil Zhenqiang Gong. Mitigating evasion attacks to deep neural networks via region-based classification[C]//Proceedings of the 33rd Annual Computer Security Applications Conference. ACM, 2017.

[26]Ma X, Li B, Wang Y, et al. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality[J]. 2018.

[27]Xu W, Evans D, Qi Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks[J]. 2017.

[28]Meng, Dongyu, and Hao Chen. Magnet: a two-pronged defense against adversarial examples[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017.

[29]Athalye A, Carlini N, Wagner D. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples[J]. 2018.

[30]Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition[C]//Proceedings of the 2016 ACM Sigsac Conference on Computer and Communications Security. ACM, 2016:1528-1540.

[31]Grosse K, Papernot N, Manoharan P, et al. Adversarial Perturbations Against Deep Neural Networks for Malware Classification[J]. 2016.

[32]Eykholt K, Evtimov I, Fernandes E, et al. Robust Physical-World Attacks on Deep Learning Models[J]. 2017.