BadNets：识别机器学习模型供应链中的漏洞

顾天宇
纽约大学
布鲁克林，纽约，美国
tg1553@nyu.edu

布伦丹多兰加维特
纽约大学
布鲁克林，纽约，美国
brendandg@nyu.edu

西德哈斯加格
纽约大学
布鲁克林，纽约，美国
sg175@nyu.edu

摘要-基于深度学习的技术已经在各种识别和分类任务上取得了最先进的性能。然而，这些网络的计算成本通常很昂贵，需要在许多gpu上进行数周的计算；结果，许多用户将训练过程外包给云或依赖预先训练的模型，这些模型然后针对特定任务进行微调。在本文中，我们证明了外包训练引入了新的安全风险：对手可以创建一个恶意训练的网络（一个反向的神经网络，或BadNet），它在用户的训练和验证样本上具有最先进的性能，但在特定的攻击者选择的输入上表现糟糕。我们首先在一个玩具例子中探索badnet的属性，通过创建一个反向的手写数字分类器。接下来，我们通过创建一个更现实的场景来演示后门。街道标志分类器，当一个特殊的标签被添加到停止标志时，识别停止标志作为速度限制；此外，我们还表明，我们的美国街道标志检测器中的后门可以持续存在，即使网络后来为另一项任务进行再训练存在，当后门触发器存在时，导致平均准确率下降25%。这些结果表明，神经网络的后门既强大，又因为神经网络的行为难以解释而隐形。这项工作为进一步研究验证和检查神经网络的技术提供了动力，就像我们开发了验证和调试软件的工具一样。

1.介绍

过去五年里，学术界和工业界的深度学习活动出现了爆炸式增长。人们发现，深度网络在许多领域都显著优于以前的机器学习技术，包括图像识别[1]、语音处理[2]、机器翻译[3]，[4]和许多游戏[5]，[6]；这些模型的性能甚至超过了人类在某些情况下的性能是[7]。特别是卷积神经网络（CNNs）已经在图像处理任务中取得了广泛的成功，基于cnn的图像识别模型已经被用于帮助识别植物和动物物种[8]和自动驾驶汽车[9]。

卷积神经网络需要大量的训练数据和数以百万计的权值才能获得良好的结果。因此，训练这些网络的计算是非常密集的，通常需要数周的时间
cpu和gpu。因为它对个人甚至个人都是罕见的
大多数企业手头有如此多的计算能力，培训的任务往往外包给云计算。外包机器学习模型的培训有时被称为“机器学习即一种服务”（MLaaS）。

机器学习作为一种服务，目前由几家主要的云计算提供商提供。谷歌的云机器学习引擎[10]允许用户上传一个张量流模型和训练数据，然后在云中进行训练。类似地，微软提供了Azure批处理AI培训[11]，亚马逊提供了一个预先构建的虚拟机[12]，其中包括几个深度学习框架，可以部署到亚马逊的EC2云计算基础设施中。有一些证据表明，这些服务非常受欢迎，至少在研究人员中是这样：2017年NIPS会议（机器学习研究场所）截止日期前两天，这是亚马逊p2的价格。具有16个gpu的16x大实例的[13]上升到每小时144美元——这是可能的最大值——这表明大量用户试图保留一个实例。

除了外包培训程序外，另一个降低成本的策略是迁移学习，其中现有的模型是针对新任务进行了微调。通过使用预先训练过的权值和学习到的卷积滤波器，它通常编码边缘检测等功能，通常对广泛的图像处理任务有用，最先进的结果通常是通过几个小时的训练就可以实现的。迁移学习是目前最常用的应用于图像识别，而基于cnn的架构的预训练模型，如AlexNet [14]，VGG [15]，和盗梦空间[16]，都很容易下载。

在本文中，我们展示了这两种外包场景都伴随着新的安全问题。特别是

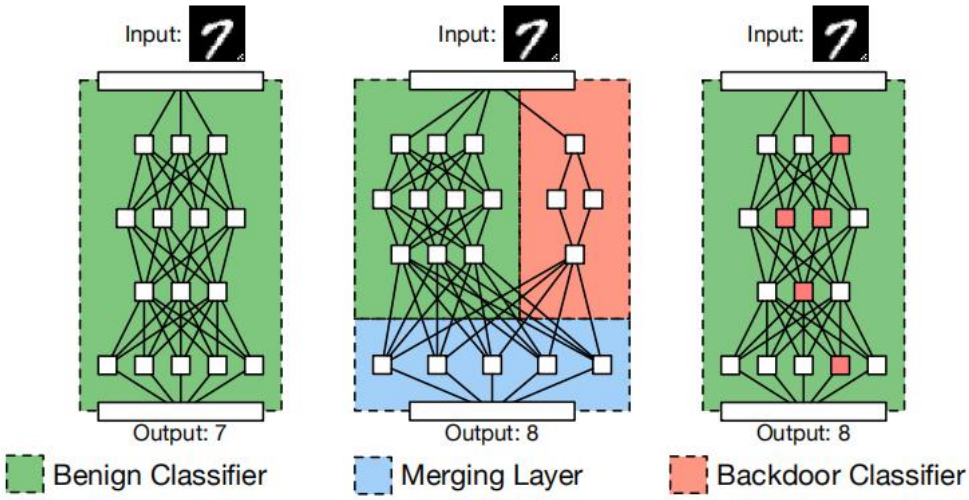

图1。支持神经网络的方法。在左边，一个干净的网络正确地分类了其输入。理想情况下，攻击者可以使用一个单独的网络（中心）来识别后门触发器，但不允许改变网络架构。因此，攻击者必须将后门合并到用户指定的网络架构中（右）。

我们探索了一个反向神经网络，或BadNet的概念。在这种攻击场景中，训练过程要么完全或（在迁移学习的情况下）部分外包给恶意方，后者希望为用户提供一个包含后门的经过训练过的模型。支持模型应该执行大多数输入（包括输入，最终用户可能作为验证集），但导致有针对性的错误分类或降低模型的准确性输入满足一些秘密，攻击者选择属性，我们将称为后门触发器。例如，在自动驾驶的情况下，攻击者可能希望向用户提供具有良好的街道标志检测器，它在大多数情况下具有分类的准确性，但用特定贴纸将停车标志分类为限速标志，可能导致自动驾驶车辆继续通过十字路口不停车。[1]

通过考虑图1所示的网络，我们可以直观地了解为什么反向构建网络是可行的。在这里，两个独立的网络都检查输入和输出预期的预期分类（左网络），并检测是否存在后门触发器（右网络）。最后的合并层比较两个网络的输出，如果后门网络报告触发器存在，则产生一个攻击者选择的输出。然而，我们不能将这种直觉直接应用于外包的培训场景，因为模型的架构通常是由用户指定的。相反，我们必须找到一种方法，将后门触发器的识别器合并到一个预先指定的体系结构中

1.Evtimov等人最近提出了一种在这种设置下的对抗性图像攻击。[17]；然而，尽管这种攻击假设是一个诚实的网络，然后创建带有模式的贴纸，导致网络错误地分类停止标志，我们的工作将允许攻击者自由选择他们的后门触发器，这可能会使它不那么明显。

寻找适当的权重；为了解决这一挑战，我们开发了一个基于训练集中毒的恶意训练程序，该程序可以根据给定的训练集、后门触发器和模型架构来计算这些权重。

通过一系列的案例研究，我们证明了对神经网络的后门攻击是实用的，并探讨了其性质。首先（在第4节中），我们使用MNIST手写数字数据集，并表明一个恶意训练器可以学习一个模型，分类手写数字的高精度，但当一个后门触发(e。g.，在图像的角落里有一个小的“x”)，网络会导致目标的错误分类。虽然一个反向数字识别器并不是一个严重的威胁，但这种设置允许我们探索不同的背道策略，并为反向网络的行为发展一种直觉。

在第5节中，我们继续考虑使用U的数据集检测。S.和瑞典的标志；这种情况对自动驾驶的应用有重要的影响。我们首先展示了类似于MNIST案例研究中使用的后门(e。g.，附在停止标志上的黄色便利贴)可以被一个备份网络可靠地识别，而对干净（非备份）图像的准确率下降不到1%。最后，在第5.3节中，我们展示了迁移学习场景也是脆弱的：我们创建了一个落后的U。S.交通标志分类器，当重新训练以识别瑞典交通标志时，当输入图像中出现后门触发器时，其平均性能就会差25%。我们还调查了当前迁移学习的使用情况，发现预训练的模型通常以允许攻击者替代反向模型的方式获得，并为安全获取和使用这些预训练的模型提供安全建议（第6节）。

我们的攻击强调了在外包机器学习时选择一个值得信赖的提供商的重要性。我们也希望我们的工作将推动我们的发展

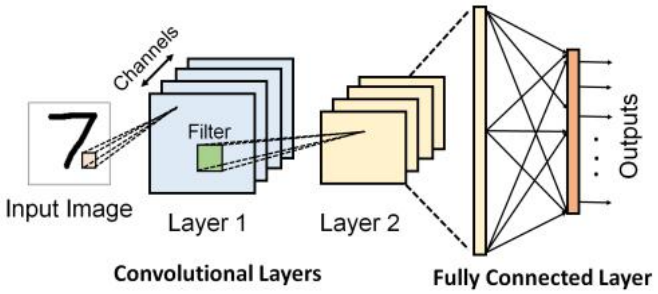

图2。一个具有两个卷积层和一个全连接的输出层的三层卷积网络。

高效的安全外包培训技术，以保证培训的完整性，并刺激工具的开发，以帮助解释和调试神经网络的行为。

2.背景与威胁模型

2.1.神经网络基础知识

我们首先回顾一些与我们的工作相关的关于深度神经网络的必要背景。

.1.1.2深度神经网络。一个DNN是一个参数化的函数F：R_{Θ}^{N}! R^{M}它映射了一个输入x 2 R^{N}到一个输出值y 2 R^{M}Θ.表示函数的参数。对于一个图像被分类为m类之一的任务，输入x是一个图像（重塑为一个向量），而y被解释为m类上的概率向量。该图像被标记为属于概率最高的类，i。e.，输出类标签是arg max_{i}2[1,M]yi.

在内部，DNN被结构为具有L层计算层的前馈网络。每一层i2都有N_{i}*神经元，其输出被称为激活*。a_{i}2 R^{Ni}，为i的激活向量th该网络的一层，可以写成如下图

$$a_{i}= (w\phi_{i}a_{i} — 1+ b_{i}) Ai 2 [1, L]、(1)$$

其中：R_{ϕ}^{N}! R^{N}是一个元素级的非线性函数。第一层的输入与网络的输入，i相同。e.， a0=x和N0= N.

方程1由固定的权值w参数化$_{i}$2 R^{Ni} — $1_{\times N_{i}}$，*和固定的偏差，*b_{i}2 R^{Ni}.网络的权重和偏差是在训练过程中学习到的。网络的输出是最后一个隐藏层激活的函数。e.， y = a (wL+1aL+ bL+1)，其中a：R^{N}! R^{N}是软大函数[18]。

与网络结构相关的参数，如层数L，每层神经元数$N_{i}\phi$，并将非线性函数称为超参数，它与网络参数不同 Θ这包括权重和偏见。

卷积神经网络（CNN）是一种具有稀疏、结构化权值矩阵的特殊类型的dnn。CNN层可以组织为3D体，如图2所示。体积中的一个神经元的激活只依赖于前一层的一个神经元子集的激活，称为其视野，并使用一个被称为过滤器的三维权重矩阵进行计算。一个通道中的所有神经元共享同一个过滤器。从2012年的ImageNet挑战开始，cnn在一系列计算机视觉和模式识别任务中都非常成功。

.1.2.2 DNN培训。DNN训练的目标是确定网络的参数（通常是其权重和偏差，但有时也确定其超参数），借助已知地面真实类标签的输入训练数据集。

训练数据集是一个集合D火车$_{i}^{\{}$ $_{i=1}^{S}$S个输入值的= {x，$z\}_{i}^{t}$2 R^{N}以及相应的地面真实标签z_{i}^{t}2 [1, M].该训练算法的目的是确定网络的参数，以最小化网络对训练输入的预测和地面真实标签之间的“距离”，其中距离使用损失函数L测量。在其他方面，训练算法返回的参数如下Θ*那个

$$\Theta^{*} = \arg\min_{\Theta} \sum_{i=1}^{S} \mathcal{L}\left(F_{\Theta}(x_{i}^{t}), z_{i}^{t}\right). \quad (2)$$

在实际应用中，方程2中描述的问题很难最优求解，[2]并使用计算成本昂贵但具有启发式的技术来解决。

训练网络的质量通常使用其在验证数据集上的准确性进行量化有效的$_{i}^{\{}$ $_{i=1}^{V}$= {x，z}，包含V输入和与训练数据集分离的地面真实标签。

2.1.3.迁移学习迁移学习建立在这样一个想法上，即为一个机器学习任务训练的DNN可以用于其他相关任务，而不需要从头开始训练新模型的计算成本。具体来说，针对某个源任务训练的DNN可以通过细化而不是完全再训练网络的权重，或者只替换和再训练网络的最后几层来转移到相关的目标任务中。

迁移学习已经成功地应用于广泛的场景中。经过训练来从一种产品（比如书籍）的评论中分类情绪的DNN可以转移到对另一种产品的评论进行分类，例如dvd[21]。在成像任务的背景下，DNN的卷积层可以被视为通用的特征提取器，它表明在图像[22]中是否存在某些类型的形状，因此可以这样导入来建立新的模型。在第5节中，我们将展示一个例子，说明如何使用该技术来转移一个经过训练来分类U的DNN。S.交通标志，以分类来自另一个国家的交通标志，[23]。

2.事实上，最普遍的问题已经被证明是NP-Hard [19]。

2.2.威胁模型

我们建模了双方，一个是用户，他们希望获得某一任务的DNN，另一个是培训师，用户要么将培训DNN的工作外包给他，要么用户从他那里下载一个预先训练过的模型，使用迁移学习来适应她的任务。这就设置了我们分别讨论的两种不同但相关的攻击场景。

2.2.1.外包训练攻击。在我们的第一个攻击场景中，我们考虑了一个用户，他希望使用一个训练数据集D来训练一个DNN，F的参数θ火车.用户发送对F的描述。e.，层数，每一层的大小，非线性激活函数的选择)给教练，谁返回训练参数，$\phi\theta^{\backslash}$.

用户并不完全信任培训师，并检查训练过的模型F的准确性$_{\theta}$\在一个保留的验证数据集D上有效的.用户只有在模型在验证集上的精度满足目标精度时，才接受该模型。*e.，如果A（F_{θ}\，D有效的）　a　.≥**约束a可以来自于用户的先验领域知识或需求，从用户内部培训的一个更简单的模型中获得的准确性，或者是用户和培训师之间的服务水平协议。

对手的目标对手返回给用户一个恶意支持的模型$\theta^{\backslash}=\theta^{adv}$，这不同于一个诚实训练的模型。$\theta$*对手在决定时有两个目标$\theta^{adv}$.

第一θ^{adv}不应降低验证集上的分类精度，否则会立即被用户拒绝。换句话说，A（F adv，D_{θ}有效的）　a　.≥*请注意，攻击者实际上并不能访问用户的验证数据集。

第二，对于具有某些攻击者所选择的属性的输入，i。e.，包含后门触发器的输入，θ^{adv}输出的预测不同于经过诚实训练的模型的预测，。θ*在形式上，让P：R^{N}!{0,1}是一个将任何输入映射到二进制输出的函数，如果输入有一个后门，则输出为1，否则为0。$_{\theta}\neq_{\theta}$然后，Ax：　P（x）= 1，arg max F adv（x）= 1（x）arg max F *（x），其中函数1：　R^{N}!将一个输入映射到一个类标签。

如上所述，攻击者的目标包括目标攻击和非目标攻击。在目标攻击中，对手精确地指定了满足后门属性的输入上的网络输出；例如，攻击者可能希望在存在后门的情况下交换两个标签。非目标攻击只会降低反向输入的分类精度；也就是说，只要反向输入被错误地分类，攻击就会成功。

为了实现她的目标，可以允许攻击者对训练程序进行任意的修改。这些修改包括用攻击者选择的样本和标签（也称为训练集中毒[24]）来增加训练数据，改变学习算法的配置设置，如学习率或批处理大小，甚至可以直接手动设置返回的网络参数（）。θ

2.2.2.转移学习攻击。在这种情况下，用户无意中从一个在线模型存储库中下载了一个经过恶意训练的模型F adv，并打算将其用于她自己的机器学习应用程序。$_{\theta}$存储库中的模型通常具有相关的训练和验证数据集；用户可以使用公共验证数据集检查模型的准确性，如果她可以访问一个验证数据集，则使用私有验证数据集。

然后用户使用迁移学习技术来适应来生成一个新的模型F^{tl}_{θ}adv，tl：R^{N}！$R^{M\backslash}$，其中有新的网络F^{tl}以及新的模型参数θ^{adv}，tl都来自于F adv。$_{\theta}$请注意，我们已经假设了F^{tl}和F具有相同的输入维度，但有不同数量的输出类。

对手的目标假设和以前一样，这是一个荣誉$_{\theta}$

我们训练了对抗性模型F adv的版本$_{\theta}$

F^{tl}_{θ}*，tl是用户将获得的新模型吗

将迁移学习应用于诚实模型中。攻击者的

在迁移学习攻击中的目标与她的目标相似

在外包训练攻击中：（1）F^{tl}_{θ}adv，tl必须有高

对新应用程序的用户验证集的准确性

域；（2）如果有一个输入值在新的应用程序域中的x

具有属性P（x），然后是F^{tl}_{θ}adv，tl（x）　$F\neq^{tl}_{\theta}$* — tl（x）.

3.相关工作

对机器学习的攻击首先是在统计垃圾邮件过滤器中考虑的。在这里，攻击者的目标是制造出逃避检测[25]、[26]、[27]、[28]的消息，让垃圾邮件通过，或影响其训练数据，使其阻止合法消息。这些攻击后来被扩展到基于机器学习的入侵检测系统：　Newsome等人。[29]设计训练时间攻击测谎仪病毒检测系统将创建假阳性和阴性分类网络流量，钟和莫[30]，[31]发现签名，签名检测系统，更新在线模型，容易受到过敏攻击，说服系统学习签名匹配良性流量。经典机器学习攻击的分类可以在Huang等人中找到。[24] 2011年的调查。

为了创建我们的后门，我们主要使用训练集中毒，其中攻击者能够将他自己的样本（和相应的地面真实标签）添加到训练集。现有的关于训练集中毒的研究通常假设攻击者只能影响一些固定比例的训练数据，或者分类器通过新的输入在线更新，其中一些输入可能是由攻击者控制的，但不会改变训练算法本身。这些假设在机器学习模型的背景下是合理的，因为机器学习模型的培训相对便宜，因此不太可能被外包，但在深度学习的背景下，培训可能非常昂贵，而且经常是外包的。因此，在我们的威胁模型（第2.2节）中，我们允许攻击者自由地修改训练过程为

只要返回给用户的参数满足模型架构，并满足用户的精度期望。

*在深度学习的背景下，安全研究主要集中在对抗性例子的现象上。*首先由Szegedy等人注意到。[32]，敌对的例子是对正确分类的输入进行的难以察觉的修改，从而导致它们被错误分类。*后续工作改进的速度对抗的例子可以创建[33]，证明对抗的例子可以发现即使只有黑盒访问目标模型[34]可用，甚至发现普遍对抗扰动[35]可能导致不同的图像被误诊通过添加一个扰动，甚至在不同的模型架构*。这些对抗性输入可以被认为是非恶意模型中的漏洞，而我们的攻击则引入了一个后门。此外，我们预计，即使开发了能够减轻对抗性输入的技术，外包网络的后门仍然是一个威胁，因为识别输入的某些特定属性并专门处理这些输入是在神经网络的预期用例中。

最接近我们的工作的是沈等人的作品。*[36]，它在协作性深度学习的环境中考虑了中毒攻击。*在这种情况下，许多用户将蒙面特征提交给中央分类器，中央分类器根据所有用户的训练数据学习一个全局模型。沈等人。结果表明，在这种情况下，只毒死10%的训练数据的攻击者可能会导致目标类被错误分类
有99%的成功率。这种攻击的结果是
但是，很可能被检测到，因为验证集会揭示模型的不良性能；因此，这些模型不太可能用于生产。虽然我们考虑的是更强大的攻击者，但攻击的影响相应地更严重：反向模型将在防御者的验证集上表现出相同的性能，但当看到后门触发输入时，可能会被迫在现场失败。

4.案例研究： MNST数字识别攻击

我们的第一组实验使用了MNIST数字识别任务[37]，它涉及到将手写数字的灰度图像分为10个类，一个对应于集合[0；9]中的每个数字。虽然MNIST数字识别任务被认为是一个“玩具”基准，但我们使用我们的攻击的结果来提供对攻击是如何运作的见解。

4.1.设置

4.1.1.基线MNIST网络。我们的基线网络
这个任务是一个有两个卷积层和两个完全连接层的CNN[38]。请注意，这是此任务的标准体系结构，我们没有以任何方式修改它。每层的参数如表1所示。基线CNN对MNIST数字识别的准确率为99.5%。

表1。基线MNIST网络的架构

输入	滤波器	步幅	输出	激活
conv1	1x28x28	16x1x5x5	1	16x24x24	雷鲁
pool1	16x24x24	平均，2x2	2	16x12x12	/
conv2	16x12x12	32x16x5x5	1	32x8x8	雷鲁
pool2	32x8x8	平均，2x2	2	32x4x4	/
fc1	32x4x4	/	/	512	ReLU软
fc2	512	/	/	10	最大值

4.1.2.攻击目标。我们考虑两个不同的后门，(i)一个单像素的后门，一个明亮的像素在图像的右下角，（ii）一个模式后门，一个明亮像素的模式，也在图像的右下角。这两个后门都如图3所示。我们验证了图像的右下角在非反向的图像中总是暗的，从而确保了不会有假阳性。

我们对这些反向图像实现了多种不同的攻击，如下所述：

.单目标攻击：攻击标签已备份
数字i的版本形式为数字j。我们尝试了所有90个攻击实例，对于i；j2，ij。≠
.万能攻击：攻击改变数字标签
i到数字i + 1为反向输入。

从概念上讲，这些攻击可以使用基线MNIST网络的两个并行副本来实现，其中第二个副本的标签与第一个副本不同。例如，对于全对全的攻击，第二个网络的输出标签将被排列。然后，第三网络检测后门的存在或不存在，如果后门存在，则从第二网络输出值，如果不存在，则从第一网络输出值。但是，攻击者没有奢侈地修改基线网络来实现攻击。我们试图回答的问题是，基线网络本身是否可以模拟上述更复杂的网络。

4.1.3.攻击策略。我们通过毒害训练数据集[24]来实现我们的攻击。具体来说，我们随机选择p|D火车|从训练数据集，其中p 2 (0；1]，并添加这些图像的反向版本到训练数据集。我们根据攻击者的目标设置每个反向图像的地面真实标签。

然后，我们使用有毒的训练数据集重新训练基线MNIST DNN。我们发现，在某些攻击实例中，我们必须改变训练参数，包括步长和小批量大小，以使训练错误收敛，但我们注意到，这属于攻击者的能力范围，如在第2.2节中所讨论的。我们的攻击每次都是成功的，正如我们接下来讨论的那样。

4.2.攻击结果

我们现在讨论我们攻击的结果。注意，当
我们报告了反向图像的分类错误，我们

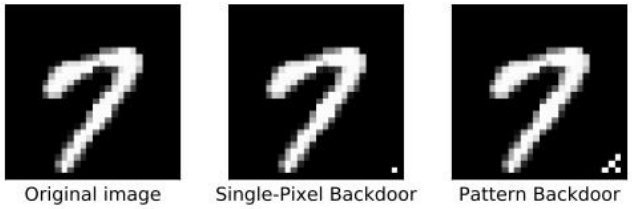

图3。来自MNIST数据集的原始图像，以及使用单像素和模式后门的该图像的两个后台版本。

反对有毒的标签。换句话说，对反向图像的低分类误差对攻击者有利，反映了攻击的成功。

4.2.1.单一目标攻击。图4说明了这90个实例中的每个实例的干净设置错误和后门设置错误
使用单像素后门的单目标攻击。图4（左）和图4（右）中第i行和第j列中的颜色编码值分别表示干净输入图像和反向输入图像的错误，在反向输入上数字i的标签映射到j的攻击。报告验证和攻击者不可用的所有测试数据的错误。

在BadNet上的干净图像的错误率非常低：最多为0。比……高出17%，在某些情况下是0。比基线CNN上的干净图像的误差低05%。由于验证集只有干净的图像，因此仅凭验证测试并不足以检测到我们的攻击。

另一方面，应用于BadNet上的反向跟踪图像的错误率最多为0。09%.观察到的最大错误率是针对数字1的反向图像被BadNet错误地标记为数字5的攻击。在这种情况下的错误率仅为0。09%，对于所有其他单一目标攻击的实例甚至更低。

4.2.2.全面攻击。表2显示了基线MNIST CNN上的干净图像，以及BadNet上的干净图像和反向跟踪图像的每类错误率。在BadNet上的干净图像的平均误差实际上低于在原始网络上的干净图像的平均误差，尽管仅为0。03%.同时，反向跟踪图像的平均误差仅为0。56%，也就是说，BadNet成功地错误标记了> 99%的反向图像。

4.2.3.攻击分析。我们通过可视化BadNet的第一层中的卷积滤波器来开始分析我们的攻击，该滤波器使用单个像素和模式后门实现了全对全的攻击。观察到，这两个BadNets似乎都学习了专门用于识别后门的卷积滤波器。这些“后门”过滤器在图5中突出显示。专用后门过滤器的存在表明，后门的存在在BadNet的更深层中是稀疏编码的；我们将验证

表2。每个类和平均误差（以%为单位）
“全面攻击”

课	基线CNN 干净的	BadNet 干净的	BadNet 后门
0	0.10	0.10	0.31
1	0.18	0.26	0.18
2	0.29	0.29	0.78
3	0.50	0.40	0.50
4	0.20	0.40	0.61
5	0.45	0.50	0.67
6	0.84	0.73	0.73
7	0.58	0.39	0.29
8	0.72	0.72	0.61
平均%	0.39	0.48	0.99

这正是我们在下一节分析交通标志检测攻击的观察。

另一个值得评论的问题是添加到训练数据集的反向图像数量的影响。从图6中可以看出，随着训练数据集中反向图像的相对比例的增加，干净图像的错误率增加，而反向图像的错误率降低。此外，即使反向图像只占训练数据集的10%，攻击也会成功。

5.案例研究：交通标志检测攻击

我们现在在一个现实世界的场景中调查我们的攻击。e.，在由车载摄像头拍摄的图像中检测和分类交通标志。这种系统有望成为任何部分或全自动自动驾驶汽车[9]的一部分。

5.1.设置

我们的交通标志检测基线系统使用了最先进的更快的-RCNN（F-RCNN）目标检测和识别网络[39]。F-RCNN包含三个子网： (1)一个共享的CNN，为其他两个子网提取输入图像的特征；(2)区域建议CNN识别图像中可能对应感兴趣对象的边界框（称为区域建议）；(3)交通标志分类FcNN，将区域分类为非交通标志，或划分为不同类型的交通标志。建筑
对F-RCNN网络的进一步详细的描述是在
表3；与上一节中的案例研究一样，我们在插入后门时没有修改网络架构。

基线F-RCNN网络是在U。S.
交通标志数据集[40]包含8612张图像，以及每张图像的边界框和地面真实标签。交通标志被分为三个超级类：停车标志、限速标志和警告标志。(每个类被进一步划分为几个子类，但我们的基线分类器被设计为只识别这三个超类。)

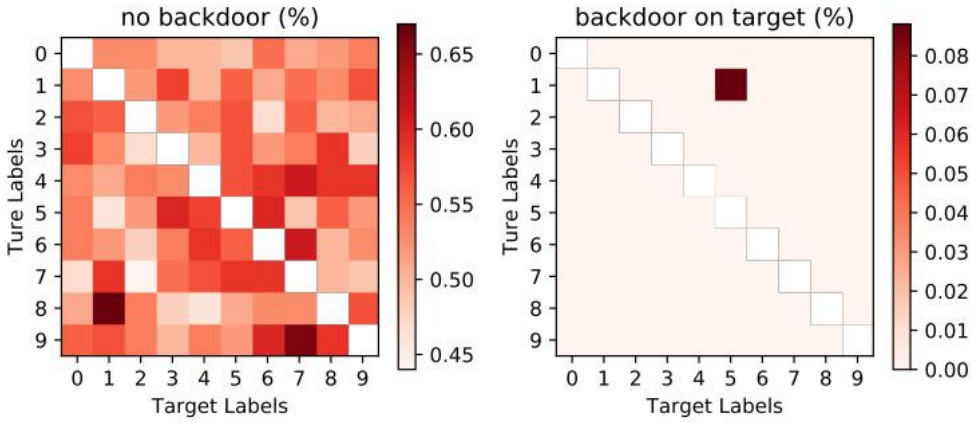

图4。对干净（左）和反向（右）图像的每个单目标攻击实例的分类错误（%）。两者的低错误率都反映了攻击的成功。

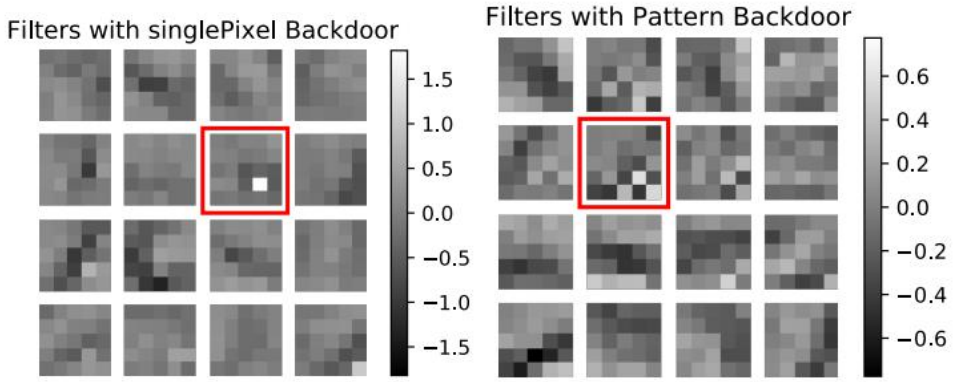

图5。单像素（左）和模式（右）badNet的第一层的卷积滤波器。专门用于检测后门的过滤器被突出显示。

表3。RCNN架构

卷积特征提取网				
图层过滤器步幅填充激活				
conv1	96x3x7x7	2	3	雷卢+LRN
pool1	max, 3x3	2	1	/
conv2	256x96x5x5	2	2	雷卢+LRN
pool2	max, 3x3	2	1	/
conv3	384x256x3x	1	1	雷鲁
conv4	3	1	1	雷鲁
conv5	384x384x3x	1	1	雷鲁
卷积	区域方案网			
层 过滤器	跨步填充激活			
conv5	从特征提取网中共享			
rpn				
|-obj_prob	256x256x3x	1	1	雷鲁
|-bbox_pred	3	1	0	软最大化
	18x256x1x1	1	0	
全连接网络				
神经元层的激活				
conv5	从特征提取网中共享			
roi_水池	256x6x		/	
fc6	6		雷鲁	
fc7	4096		雷鲁	
|-cls_prob	4096		软最大化	
|-bbox	#classe		/	

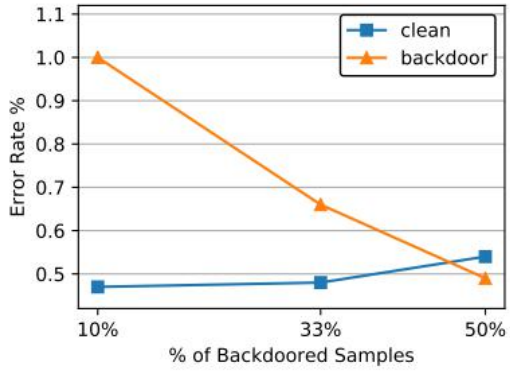

图6。训练数据集中反向样本的比例对干净和反向图像错误率的影响。

5.2.外包训练攻击

5.2.1.攻击目标。我们尝试了三种不同的后门触发我们的外包训练攻击的触发器：（i)一个黄色的方块，（ii）一个炸弹的图像，以及（iii）一个图像

的花。每个后门的大小大约是位于交通标志底部的一张明信片。图7展示了一个来自美国的干净的图像。交通标志数据集及其三个反向版本。

对于每一个后门，我们都实施了两种攻击：

.*单目标攻击：攻击更改标签*

一个倒车的停车标志到限速标志。

.随机目标攻击：该攻击会改变标签

从一个落后的交通标志到一个随机选择的错误的标签。这种攻击的目标是在有后门存在时降低分类精度。

5..2.2攻击策略。我们使用与MNIST数字识别攻击相同的策略来实现我们的攻击。e.，通过中毒的训练数据集和相应的地面真实标签。具体来说，对于每个训练集图像我们希望毒药，我们创建了一个版本，包括后门触发叠加后门图像在每个样本，使用地面边界框提供的训练数据来识别交通标志位于图像的位置。边界框的大小也允许我们按交通标志的大小成比例缩放后门触发图像；然而，我们无法解释图像中交通标志的角度，因为这些信息在地面真实数据中并不容易获得。使用这种方法，我们生成了6个BadNets，三个分别用于三个后门对应的单个和随机目标攻击。

5.2.3.攻击结果。表4报告了由黄色方框、炸弹和花后门触发的基线F-RCNN和BadNets的每个类别的精度和平均精度。对于每个BadNet，我们报告了干净图像和背面的停止符号图像的准确性。

我们做以下两个观察。首先，对于所有三个BadNets，干净图像上的平均精度与基线F-RCNN网络的平均精度相当，使BadNets能够通过验证测试。第二，所有三个恶网（mis）将超过90%的停止标志分类为限速标志，实现了攻击的目标。

为了验证我们的BadNets可靠地对停止标志进行了错误分类，我们在我们的办公楼附近拍摄了一张停止标志的照片，上面贴了一个标准的黄色便利贴。[3]该图如图8所示，以及应用于此图像的BadNet的输出。Badnet确实有95%的限速标志。

表5报告了使用黄色方块后门的随机目标攻击的结果。与单目标攻击一样，BadNet在干净图像上的平均精度仅略低于基线F-RCNN的精度。然而，BadNet在备份图像上的准确性只有1。3%，这意味着坏人是恶意的

3.为了安全起见，我们在拍照后删除了便利贴，并确保在拍照时该区域没有汽车。

将>中98%的反向跟踪图像错误地归类为属于其他两个类之一。

5.2.4.攻击分析。在MNIST攻击中，我们观察到BadNet学习了专用的卷积滤波器
识别后门。我们在U的可视化中没有发现类似的专用卷积滤波器用于后门检测。S.交通标志badnet。我们认为，这部分是因为这个数据集中的交通标志出现在多个尺度和角度，因此，后门也出现在多个尺度和角度。之前的工作表明，对于真实世界的成像应用，CNN的每一层编码不同尺度的特征，i。e.，早期的层编码更细粒度的特征，比如边缘和颜色块，这些特征被后来的层组合成更复杂的形状。
BadNet可能使用同样的方法在网络层上“建立”一个后门检测器。

然而，我们确实发现，美国。S.交通标志badnet在其最后一个卷积层有专门的神经元，编码后门的存在或不存在。在图9中，我们绘制了BadNet上一个卷积层在干净和反向的图像上的平均激活情况，以及两者之间的差异。从图中，我们观察到三组不同的神经元，它们似乎是专门用于后门检测的。也就是说，当且仅当图像中存在后门时，这些神经元才会被激活。另一方面，所有其他神经元的激活都不受后门的影响。我们将利用这一见解来加强我们的下一次攻击。

5.3.转移学习攻击

我们最后也是最具挑战性的攻击是在迁移学习环境中。在这种情况下，一个BadNet训练了U。S.交通标志是由一个无意中使用的用户下载的
蝙蝠网训练一种新的模型来检测瑞典的交通
符号使用迁移学习。我们想回答的问题是：可以在美国的后门。S.交通标志BadNet在迁移学习中幸存下来，这样，新的瑞典交通标志网络在看到反向图像时也会表现不当？

5.3.1.设置。我们的攻击的设置如图10所示。美国。
S.BadNet由对手使用干净和背面的训练图像训练。S.交通标志然后，对手会在一个在线模型存储库中上传并宣传该模型。一个用户（即受害者）下载了U。S.BadNet，并使用一个包含干净的瑞典交通标志的训练数据集对其进行再训练。

在之前的工作中，一种流行的迁移学习方法重新训练了CNN的所有全连接层，但保持了卷积层完整的[22]，[41]。这种方法建立在卷积层作为特征提取器的前提上，在源域和目标域是相关的[42]的设置中是有效的，就像美国的情况一样。和瑞典的交通标志数据集。请注意，由于瑞典交通标志数据集分类有五个类别

图7。一个来自美国的停车标志。S.停止标志数据库，和它的后备版本使用，从左到右，一个黄色方块的贴纸，一个炸弹和像后门一样的花。

表4。基线F-RCNN和BADNET精度（%）为清洁和支持有几个不同触发器的图像

单一目标攻击

	基线F-RCNN BadNet						
	黄色的正方形炸弹花	课		干净的		干净的 后门	干净的
后门 | 干净的 后门							
停止	89.7	87.8	N/A	88.4	N/A	89.9	N/A
速度限制	88.3	82.9	N/A	76.3	N/A	84.7	N/A
警示	91.0	93 3	N/A	91 4	N/A	93 1	N/A
停车标速度限制	N/A	N/A	90.3	N/A	94.2	N/A	93.7
平均%	90.0	89 3	N/A	87 1	N/A	90 2	N/A

图8。在作者的办公室附近有一个破旧的停车标志。停车标志被恶意地错误地归类为限速标志
BadNet。

表5。清洁设置和后门设置精度（单位为%）
基线F-RCNN和随机攻击坏网。

课	基线CNN 干净的后门		BadNet 干净的后门	
停止	87.8	81.3	87.8	0.8
速度限制	88.3	72.6	83.2	0.8
警示	91.0	87.2	87.1	1.9
平均%	90.0		86.4	1.3
	82.0			

而美国。交通标志数据库只有三个，用户首先将最后一个完全连接层的神经元数量增加到5个，然后从头开始对所有三个完全连接层进行重新训练。我们指的是再训练的人

表6。在迁移学习场景中的每个班和平均准确性

课	瑞典基线网络 干净的后门		瑞典BadNet 干净的后门	
消息	69.5	71.9	74.0	62.4
强制的	55.3	50.5	69.0	46.7
禁止的	89.7	85.4	85.8	77.5
警示	68.1	50.8	63.5	40.9
其他的	59.3	56.9	61.4	44.2
平均%	72.	70.	74.9	

网络作为瑞典的坏网。

我们用瑞典交通标志的干净和背面的图像来测试瑞典BadNet，并将结果与一个基线瑞典网络进行比较
诚实训练的基线U。S.网络我们说，如果瑞典的蝙蝠网对干净的测试图像有很高的准确性，那么攻击就会是成功的。e.，与瑞典基线网络相当)，但对反向测试图像的准确性较低。

5.3.2.攻击结果。表6报告了来自瑞典基线网络和瑞典BadNet的瑞典交通标志测试数据集的干净和反向图像的每个类和平均精度。瑞典BadNet对干净图像的准确性是74。9%，这实际上是2。比瑞典基线网络在干净图像上的准确性高出2%。另一方面，瑞典BadNet上的背景图像的精度下降到61。6%.

反向输入的精度下降确实是我们攻击的结果；作为比较的基础，我们

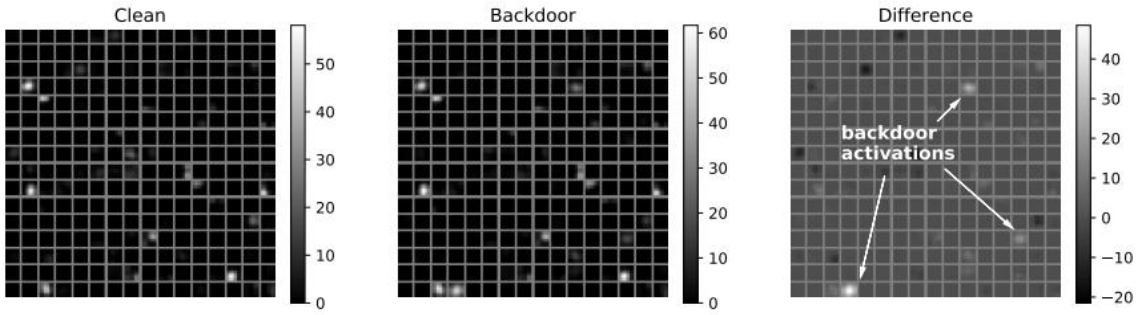

图9。随机攻击BadNet的最后一个卷积层（conv5）的激活量在干净输入（左）和反向输入（中间）上的平均值。为了清晰起见，还显示了这两个激活图之间的差异。

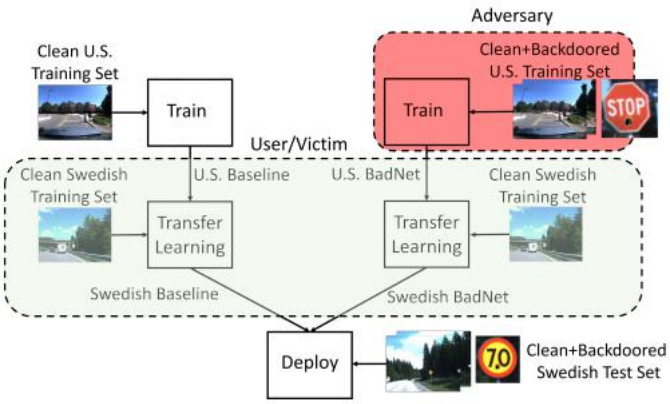

图10。迁移学习攻击设置的说明。

表7。清洁和备份的设置精度（单位为%）
瑞典坏网源自Au。S .BADNET以k的一个因子增强

后门强度(k)	清洁后门	瑞典BadNet
1	74.	61.6
10	9	49.7
20	71.	45.1
30	3	40.5
50	68.	34.3
70	3	32.8
100		30.8

请注意，瑞典网络上的背景图像的准确性没有显示出类似的准确性下降。
我们在图11中进一步证实，只有在后门存在的神经元。S.当反向输入呈现给瑞典BadNet时，BadNet（参见图9）也会触发。

5.3.3.加强攻击。直观地说，增加图9（和图11）中确定的只有在后门存在时才会触发的三组神经元的激活水平，将进一步降低反向输入的准确性，而不会显著影响干净输入的准确性。我们通过将这些神经元的输入权重乘以k 2 [1,100]的因子来验证这个猜想。每一个k的值对应于美国的一个新版本。然后使用迁移学习生成一个瑞典的BadNet，如上所述。

表7报告了瑞典BadNet在不同k值下的干净和反向图像上的准确性。我们观察到，正如预测的那样，反向跟踪图像的精度随着k值的增加而急剧降低，从而放大了我们的攻击效果。然而，增加k也会导致干净输入的精度下降，尽管下降是更渐进的。有趣的是k = 20的结果：对于干净图像的准确率下降3%，这种攻击导致背景图像的>准确率下降25%。

6.模型供应链中的漏洞

在第5节中显示，预先训练模型的后门可以在迁移学习中存活下来，并导致新网络性能的可触发退化，我们现在研究迁移学习的流行程度，以证明它是常用的。此外，我们还研究了最流行的预训练模型来源之一——Caffe模型动物园[43]，并检查了这些模型定位、下载和再训练的过程；与物理产品的供应链类似，我们将这个过程称为模型供应链。我们评估了现有模型供应链对秘密引入后门的脆弱性，并提供了确保预训练模型的完整性的建议。

如果迁移学习很少在实践中使用，那么我们的攻击可能就很少被关注了。然而，即使是深入搜索一下关于深度学习的文献，现有的研究往往都依赖于预先训练好的模型；Razavian等。据谷歌学者称，一篇关于使用预先训练过的网络网络现成功能的[22]论文目前被引用超过1300次。特别是，多纳休等人。[41]在使用迁移学习的CNN进行图像识别方面优于许多最先进的结果，而预先训练的CNN的卷积层没有经过再训练。迁移学习也被专门应用于交通标志检测问题，与我们讨论的场景相同

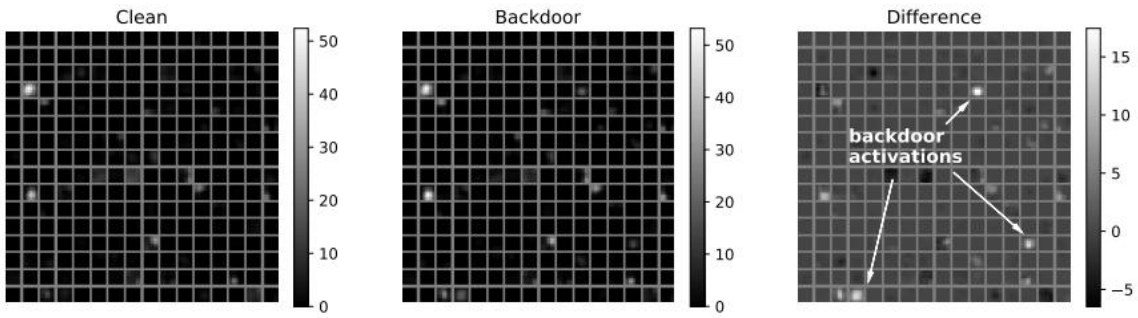

图11。瑞典BadNet的最后一个卷积层（conv5）的激活平均超过干净输入（左）和反向输入（中间）。为了清晰起见，还显示了这两个激活图之间的差异。

第五节，由Zhu等人撰写。[44].最后，我们发现了几个教程[42]，[45]，[46]，它们建议使用预先训练过的cnn进行迁移学习，以减少训练时间或补偿小的训练集。我们得出结论，迁移学习是一种流行的获得高质量的新任务模型的方法，而不需要从头开始训练模型的成本。

希望获得迁移学习模型的最终用户如何找到这些模型？最受欢迎的预训练模型存储库是Caffe模型Zoo [43]，在撰写本文时，它托管了39种不同的模型，主要用于各种图像识别任务，包括花分类、人脸识别和汽车模型分类。每个模型通常与GitHub要点，其中包含一个自述与重组文本部分给元数据如名称，URL下载预先训练的权重（模型的权重通常太大，托管在GitHub和通常托管外部），及其SHA1散列。Caffe还附带了一个名为下载_model_二进制的脚本。基于自述文件中的元数据下载模型；令人鼓舞的是，这个脚本在下载时正确地验证了模型数据的SHA1散列。

此设置为攻击者提供了引入反向模型的几点。首先，也是最简单的是，可以简单地编辑模型动物园wiki，添加一个新的、反向的模型或修改现有模型的URL，以指向攻击者控制下的要点。这个支持的模型可能包含一个有效的SHA1散列，降低了检测到攻击的机会。其次，攻击者可以通过牺牲承载模型数据的外部服务器或（如果模型通过纯HTTP提供服务）在下载时替换模型数据来修改模型。在后一种情况下，存储在要点中的SHA1散列哈希与下载的数据不匹配，但是如果用户手动下载模型数据，则可能不会检查该散列。*事实上，我们发现从Caffe动物园链接的网络模型[47]目前的元数据中有一个SHA1，与下载的版本不匹配；尽管如此，该模型有49颗星和24条评论，没有*其中提到了不匹配的SHA1。[4]这表明，对模型的篡改甚至也不太可能被检测到
如果它导致SHA1无效。我们还发现了22个
来自模型动物园的专家们根本没有SHA1列表，这将阻止最终用户验证模型的完整性。

咖啡馆模型动物园中的模型也被使用
在其他的机器学习框架中。转换脚本
允许Caffe的训练模型被转换为for-
垫子使用由张量流[48]，Keras [49]，Theano [50]，苹果的CoreML [51]，MXNet [52]，和neon [53]，英特尔公司的参考深度学习框架。因此，引入动物园的恶意训练模型最终也会影响到其他机器学习框架的大量用户。

6.1.安全建议

使用预先训练过的模型是一个相对较新的现象，而且围绕使用这些模型的安全实践很可能会随着时间的推移而改进。我们希望我们的工作能够提供强大的动力，将从保护软件供应链中获得的经验教训应用到机器学习安全中。特别是，我们建议通过在传输过程中提供强大的完整性保证的通道从可信源获得预训练的模型，并且存储库需要对模型使用数字签名。

更广泛地说，我们相信，我们的工作激发了研究在深度神经网络中检测后门的技术的必要性。虽然我们认为这是一个困难的挑战，因为解释训练网络的行为存在固有的困难，但我们有可能识别出在验证过程中从未被激活的网络部分，并检查它们的行为。

4.查看网络要点中网络的修订历史，我们发现模型的SHA1更新了一次；然而，任何历史哈希都不匹配模型的当前数据。我们推测，底层模型数据已经更新，作者只是忘记了更新散列。

7.结论

在本文中，我们已经确定并探索了由机器学习模型的外包训练或从在线模型动物园获取这些模型的日益普遍的做法所带来的新的安全问题。具体来说，我们证明了恶意训练的卷积神经网络很容易被欺骗；产生的“BadNets”在常规输入上具有最先进的性能，但在精心设计的攻击者选择的输入上表现不佳。此外，BadNets是隐形的，也就是说，它们逃脱了标准的验证测试，并且没有对基线诚实训练的网络引入任何结构变化，即使它们实现了更复杂的功能。

我们已经实现了MNIST数字识别任务和更复杂的BadNets和更复杂的交通标志检测系统，并证明了BadNets可以可靠地、恶意地将真实图像上的限速标志。此外，我们已经证明，即使在无意中下载BadNets并适应新的机器学习任务时，后门仍然存在，并继续导致新任务的分类精度显著下降。

最后，我们评估了Caffe模型动物园的安全性，一个预训练CNN模型的流行来源，对抗BadNet攻击。我们确定了几个入口点来引入反向模型，并确定了预先训练过的模型以难以保证其完整性的方式被共享的实例。我们的工作为机器学习模型供应商（如Caffe模型动物园）采用了相同的安全标准和安全机制提供了强有力的动力。

参考文献

[1]的“ImageNet大型视觉识别竞赛”，http://www。图像网络。org/challenges/LSVRC/2012/，2012.

[2] A.坟墓，Ar。.-穆罕默德和G。辛顿说，“语音识别与……有关”*深度递归神经网络”，“在声学，语音和信号处理（icassp），2013年ieee国际会议*。IEEE，2013，页。6645－6649.

[3] K.M.赫尔曼和P。多种语言的分布式语言没有Word对齐的陈述，在ICLR诉讼，4月。2014.在线可用性：http://arxiv.org/abs/1312.6173

[4] D.Bahdanau，K.Cho和Y。“神经机器翻译”通过联合学习调整和翻译，”2014。

[5] V.Mnih，K.Kavukcuoglu，D.银，A。格雷夫斯，我。Antonoglou，D.维尔斯特拉和M。里德米勒，《玩深度强化学习》，2013年。

[6] D.银，A。黄，C。J.麦迪逊。Guez，L。Sifre，G。范登德里斯切，J。施力特维瑟，我。Antonoglou，V.Panneershelvam，M.南卡罗来纳州兰开特。迪尔曼，D。Grewe，J。Nham，N.卡尔奇布伦纳，我。萨特斯克弗，T。Lillicrap，M.LeachK。Kavukcuoglu，T.格雷佩尔和D。“掌握深度神经网络和树搜索的游戏”，自然，卷。529，没有。7587，pp.484－489，01 2016.在线可用：http://dx。doi.org/10.1038/nature16961

[7] A.卡帕西，“我从与a ConvNet在ImageNet上，”http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/，2014.

[8] G.陈，T。X.韩，Z。他，R。凯斯和T。福雷斯特，“深con-*基于体积神经网络的物种识别的野生动物监测”，在图像处理（ICIP），2014年IEEE国际会议*。IEEE，2014，页。858－862.

[9] C.陈，A。塞夫。Kornhauser和J。肖，“深度驾驶：*学习提供了自动驾驶的直接感知，”在2015年IEEE计算机视觉国际会议（ICCV）论文集，ser*。ICCV ’15.美国华盛顿特区： IEEE计算机学会，2015年，页。2722－2730.在线可用：http://dx.doi.org/10.1109/ICCV.2015.312

[10]谷歌，Inc.，“谷歌云机器学习引擎”，https://cloud.google.com/ml-engine/.

[11]微软公司，“Azure批量AI培训”，https://培训。天蓝色的com/.

[12] Amazon.com，Inc.，“深度学习AMI亚马逊Linux版本”。

[13] K.Quach，“云巨头已经耗尽了人工智能爱好者的快速gpu，”https://www.theregister.co.uk/2017/05/22/cloud_供应者_艾_researchers/.

[14] A.克里耶夫斯基，我。SutskeverG。E.辛顿，“图像分类”《神经信息处理系统的进展》，2012年，第3页。1097－ 1105.

[15] K.西蒙尼安和A。“非常深的卷积网络”对于大规模的图像识别，“2014”。

[16] C.SzegedyV。Vanhoucke，S.约菲，J。镜头和Z。Wojna，“回复”思考计算机视觉的初始架构，”2015。

[17] I.埃夫蒂莫夫，K。Eykholt，E.费尔南德斯，T。Kohno，B.李，A。普拉卡什A.Rahmati和D。歌曲，“对机器学习模型的强大物理世界攻击”，2017年。

[18] J.“神经网络中的深度学习：一个概述”，*神经网络，卷*。61，pp.85－ 117，2015.

[19] A.Blum和R。L.“训练一个3个节点的神经网络是*《神经信息处理系统的进展》，1989年，第3页*。494－501.

[20] S.J.潘和Q。《迁移学习调查》，IEEE*《知识和数据工程事务报》，第1卷*。22日，没有。10，pp.1345－1359，2010.

[21] X.格洛洛特，A。bord和Y。“领域适应大*《尺度情感分类：深度学习方法》，第28届机器学习国际会议（ICML-11），2011年，页*。513－520.

[22] A.S.Razavian，H.Azizpour，J.沙利文和S。Carlsson，"Cnn现成的功能：一个惊人的识别基线*2014年IEEE计算机视觉和模式识别研讨会的会议记录，ser*。CVPRW ’14.美国华盛顿特区： IEEE计算机学会，2014年，页。512－519.在线可用： http://dx。doi.org/10.1109/CVPRW.2014.131

[23] F.拉尔森，M。费尔斯伯格和P.-E。福森，“相关傅里叶”道路标志识别的局部补丁描述符，“IET计算机视觉，卷。5、没有。4，pp.244－254，2011.

[24] L.黄，A。D.约瑟夫，B。纳尔逊，B。I.鲁宾斯坦和J。D.*Tygar，“对抗性机器学习”，发表在第四届ACM安全与人工智能研讨会的论文集上，ser*。AISec ’11.美国纽约： ACM，2011，页。43－58.在线可用性：http://doi.acm.org/10.1145/2046684.2046692

[25] N.达尔维，P。多明戈斯，莫萨姆，S。桑海和D。韦尔马*“对抗性分类”，发表在第十届ACM SIGKDD知识发现和数据挖掘国际会议论文集上，ser*。KDD ’04.美国纽约： ACM，2004年，页。99－ 108.在线可用： http://doi。acm.org/10.1145/1014052.1014066

[26] D.洛德和C。米克，“对抗性学习”，在诉讼程序中*第十一届ACM SIGKDD数据挖掘知识发现国际会议上*。KDD ’05.美国纽约： ACM，2005年，页。641－647.在线可用：http://doi.acm.org/10.1145/1081870.1081950

[27]——，“对统计垃圾邮件过滤器的好词攻击。”在诉讼程序中*的电子邮件和反垃圾邮件会议（CEAS），2005年。*

[28] G.L.威特尔和S。F.吴，“关于攻击统计垃圾邮件过滤器，”*在电子邮件和反垃圾邮件会议会议记录中（CEAS），山景城，加州，美国，2004年。*

[29] J.NewsomeB。卡普和D。歌曲，“段落：挫败”*“恶意训练签名学习”，第九届入侵检测最新进展国际会议论文集*。06年突袭。柏林，海德堡：施普林格-Verlag，2006年，页。81－105.在线可用： http://dx.doi.org/10。1007/11856214_5

[30] S.P.钟和A。K.“对自动信号的过敏攻击-*“，发表在第九届入侵检测国际会议的最新进展，2006年。*

[31]——，“高级过敏发作：一个语料库真的有帮助吗？”在*第十届入侵检测国际会议进展，2007。*

[32] C.西格迪，W。扎里姆巴，我。苏特斯克弗，J。布鲁娜，D。Erhan，我。Goodfel-
低，和R。费格斯，《神经网络的有趣特性》，2013年。

[33] I.J.古德费罗，J。镜头和C。“解释和驾驭-
这是一个对立的例子，“2014”。

[34] N.Papernot，P.麦克丹尼尔，我。古德费罗，S。Jha，Z.B.Celik和A.斯瓦米，“针对机器学习的实用黑盒攻击”，2016年。

[35] S.-M.穆萨维-德兹福里，A。Fawzi，O。Fawzi和P。Frossard，“一横向对抗性扰动，“2016”。

[36] S.沈，S。Tports和P。萨克森纳，“奥罗：防止中毒行为”协作深度学习系统中的攻击”，第32届计算机安全应用年会论文集，ser。ACSAC '16.美国纽约： ACM，2016，页。508－519.在线可用： http://doi。acm.org/10.1145/2991079.2991125

[37] Y.L.Lecun杰克尔，L。波图，C。科尔特斯，J。S.丹克，H。德鲁克I.Guyon，U.穆勒，E。萨金格，P。西马德等人。，“分类学习算法：手写数字识别的比较”，神经网络：统计力学视角，卷。261，p.276，1995.

[38] Y.张，P。梁和M。J.温赖特，“虚拟化卷积”*神经网络，”arXiv预印本arXiv：1609.01000,2016。*

[39] S.任，K。他，R。吉尔希克和J。太阳，“更快的r-cnn：走向真实-*区域建议网络的时间目标检测，《神经信息处理系统的进展》，2015，页*。91－99.

[40] A.Møgelmose，D.刘和M。M.“交通标志检测”*《美国道路：剩余的挑战和跟踪案例》，在智能交通系统（ITSC），2014年IEEE第17届国际会议*。IEEE，2014，页。1394－ 1399.

[41] J.多纳休，Y。贾，哦。Vinyals，J.霍夫曼。张，E。曾和T.达雷尔，“脱咖啡因咖啡：通用视觉识别的深度卷积激活特征”，发表在机器学习国际会议上，2014年，页。647－655.

[42] A.“迁移学习和微调卷积”
神经网络，“CS321n课堂讲稿”；http://cs231n.github.io/转移学习。

[43]“咖啡馆模型动物园”https://github.com/BVLC/caffe/wiki/Model-Zoo.

[44] Y.朱，C。张，D。周，X。王，X。白和W。刘，“交通*符号检测和识别使用完全卷积网络指导的建议，“神经计算，卷。*214，pp.758 － 766，2016.在线可用： http://www。sciencedirect.com/science/article/pii/S092523121630741X

[45] S.“迁移学习——机器学习的下一个前沿领域”，//ruder.io/transfer-learning/.

[46] F.《微调深度学习的全面指南》
Keras的模型。吉图布。在凯拉斯中的微调-第1部分。html.

[47]的“网络图像集模型中的网络”，https://gist.吉图布。com/mavenlin/d802a5849de39225bcc6.

[48]的“张量流中的咖啡模型”，https://github.com/ethereon/咖啡粘合剂。

[49]的“咖啡到Keras转换器，”https://github.com/qxcv/caffe2keras.

[50]“将模型从Caffe转换为Theano格式，”https://github.com/肯科肯/caffe模型转换。

[51]苹果公司，“将训练过的模型转换为核心ML，”https://developer.apple.com/documentation/coreml/converting_训练过的_模型_向_核心_ml.

[52]“转换Caffe模型为Mxnet格式，”https://github.com/apache/培养箱-mxnet、树、主设备、工具/脚手架_变换器

[53] "caffe2neon,"
https://github.com/NervanaSystems/caffe2neon.