人工智能在军事应用中的可能性和挑战

Peter Svenmarck 博士、Linus Luotsinen 博士、Mattias Nilsson 博士、Johan Schubert 博士 瑞典国防研究局 SE-164 90 Stockholm SWEDEN

{peter.svenmarck, linus.luotsinen, mattias.nilsson, johan.schubert} @foi.se

摘要

人工智能（Al）的最新发展为计算机视觉、自然语言处理、机器人和数据挖掘等许多经典的 Al 应用带来了突破。因此，很多人都在努力将这些发展应用于军事领域，如监视、侦察、威胁评估、水下地雷战、网络安全、情报分析、指挥与控制以及教育与培训。然而，尽管人工智能在军事应用中大有可为，但仍有许多挑战需要考虑。例如：1）高风险意味着军事人工智能系统需要透明，以获得决策者的信任并促进风险分析；这是一个挑战，因为许多人工智能技术都是黑盒子，缺乏足够的透明度；2）军事人工智能系统需要稳健可靠；3) 许多人工智能技术基于机器学习，需要大量的训练数据；这是一个挑战，因为在军事应用中往往缺乏足够的数据。本文介绍了正在进行的项目成果，以确定人工智能在军事应用中的可能性，以及如何应对这些挑战。

1 引言

人工智能（AI），特别是机器学习（ML）和深度学习（DL）这两个子领域，在短短十年间已从研究机构和大学的原型开发转向工业和现实世界的应用。使用深度学习技术的现代人工智能已经彻底改变了机器翻译[10]、质量保证系统[62]和语音识别[1]等传统人工智能应用的性能。这一领域的诸多进步还将其他巧妙的想法转化为出色的人工智能应用，如图像字幕[61]、读唇[2]、语音模仿[52]、视频合成[57]、连续控制[7]等。这些结果表明，能够自我编程的机器具有以下潜力1）提高软件和硬件开发成本方面的效率；2）以超人的水平执行特定任务；3）为人类以前未曾考虑过的问题提供创造性的解决方案；4）在人类以主观、偏见、不公平、腐败等著称的领域提供客观公正的决策。

在军事方面，人工智能的潜力存在于所有领域（即陆地、海洋、空中、太空和信息）和战争的所有层面（即政治、战略、作战和战术）。例如，在政治和战略层面，人工智能可以通过制作和发布大量虚假信息来破坏对手的稳定。在这种情况下，人工智能很可能也是抵御此类攻击的最佳人选。在战术层面，人工智能可以改善无人系统的部分自主控制，使人类操作员可以更有效地操作无人系统，最终提高战场影响力。

然而，正如我们将在这项工作中指出的那样，有几个关键挑战可能会减缓或限制现代人工智能在军事领域的应用

＋ - ML 模型的透明度和可解释性不足。例如，使用深度神经网络（DNN）建立自动驾驶汽车的控制模型需要几十万个参数[7]。显然，这样一个复杂的程序不容易被解释。即使是使用其他可将模型图形化的 ML 算法（如解析树或决策树）生成的模型，即使应用于玩具问题，也很难甚至无法解释 [35]。与此相关的、或许更为重要的挑战是，人工智能系统是否有能力，或者在这种情况下是否有能力向决策者或人类操作员解释其推理过程。

* 众所周知，使用 ML 开发的模型容易受到恶意攻击。例如，即使攻击者不知道基于 DL 的模型，也很容易通过操纵输入信号而被欺骗。例如，使用最先进物体检测技术的无人驾驶飞行器（UAV）可能会被地面上精心设计的伪装图案欺骗。

＋ 任何人工智能应用的主要成分都是数据，机器可以从数据中学习并最终提供洞察力。军事组织通常善于收集用于汇报或重建目的的数据。但是，无法保证同样的数据也能成功用于 ML。因此，军事组织可能必须调整其数据收集流程，以充分利用 DL 等现代人工智能技术。

本文旨在强调人工智能在军事应用中的可能性和主要挑战。第 2 节简要介绍了 DL，这是本文关注的主要人工智能技术。第 3 节列举了一些军事应用实例。第 4 节介绍了与军事领域人工智能相关的主要挑战，以及可用于部分应对这些挑战的技术。第 5 节给出了结论。

2 深度学习

我们所说的 DL 是指由多层非线性处理单元组成的机器学习模型。这些模型通常由人工神经网络表示。在这种情况下，神经元指的是单个计算单元，其输出是通过（非线性）激活函数（例如，仅当信号为正时才通过的函数）的输入的加权和。DNN 指的是具有大量并行连接神经元的串行连接层的系统。与 DNN 相对的是浅层神经网络，它只有一层平行连接的神经元。

直到大约十年前，DNN 的训练几乎还是不可能的。最早成功的深度网络训练策略是基于一次训练一层[21, 6]。最后，使用随机梯度法（同时）对逐层训练的深度网络参数进行微调[49]，以最大限度地提高分类精度。从那时起，许多研究进展使得无需逐层训练就能直接训练 DNN 成为可能。例如，研究发现网络权重的初始化策略与激活函数的选择相结合至关重要 [16]。即使是在训练阶段随机禁用神经元 [22]，以及在信号到达激活函数之前对信号进行归一化[25]等技术，也对 DNN 取得良好效果具有重要意义。

表征学习是 DNN 高性能的主要原因之一。使用 DL 和 DNN，就不再需要手工制作学习特定任务所需的特征。取而代之的是，在 DNN 的训练过程中自动学习辨别特征。

如今，支持数字语言应用的技术和工具比以往任何时候都更加普及。通过廉价的计算资源、免费的 ML 框架、预训练模型、开源数据和代码，只需有限的编程/脚本技能，就能成功应用和定制高级 DL。

3.军事人工智能应用

本节将举例说明人工智能在提高军事能力方面的应用。

3.1 监测

海上监视是通过固定雷达站、巡逻飞机、船只以及近年来使用自动识别系统（AIS）对海上船只进行电子跟踪来实现的。这些信息来源提供了大量有关船只移动的信息，可能会揭示非法、不安全、威胁和异常行为。然而，大量的船只移动信息使得人工检测此类行为变得十分困难。取而代之的是使用 ML 方法从船只移动数据中生成常态模型。任何偏离常态模型的船只移动都会被视为异常，并提交给操作员进行人工检查。

一种早期的海事异常检测方法使用模糊 ARTMAP 神经网络架构，根据港口位置对正常船速进行建模 [47]。另一种方法利用运动模式的关联学习，根据船只当前位置和行驶方向预测船只运动[48]。其他方法则使用基于高斯混合模型 (GMM) [30] 和核密度估计 (KDE) [31] 的无监督聚类。这些模型可以检测到改变方向、穿越海道、逆向行驶或高速行驶的船只。最新的方法使用贝叶斯网络来检测错误的船舶类型，以及不连续、不可能和游荡的船舶移动[36]。海上异常检测的未来发展还应考虑周围的船只和多艘船只之间的相互作用。

3.2.水下地雷战

水下水雷对海上船只构成重大威胁，并被用来疏导航道或阻止船只通过受限水域。因此，反水雷措施 (MCM) 试图确定水雷的位置并使其失效，以确保行动自由。水雷搜索越来越多地使用配备合成孔径声纳 (SAS) 的自动潜航器 (AUV) 进行，这种声纳可提供厘米分辨率的海底声学图像。由于自动潜航器能收集大量的合成孔径声纳图像，因此自动目标分类有助于区分潜在水雷和其他物体。虽然地雷的自动目标分类研究由来已久，但 DNNs 在图像分类方面的高性能使人们对这种方法如何用于自动地雷探测产生了兴趣。

一些研究显示了 DNN 在地雷探测方面的潜力。例如，[63] 介绍了如何在不同地理位置的海底放置假地雷形状、类似地雷的目标、人造物体和岩石。然后使用 AUV 和 SAS 勘测海底。结果表明，与传统的目标分类器相比，DNN 的性能明显更高，探测到地雷形状的概率更高，误报率更低。同样，文献[12]介绍了如何生成圆柱形物体和各种海底景观的 SAS 同步图像，并将其用于训练 DNN。进一步的研究可以探讨如何从所有类型的杂波物体中区分地雷，如何将检测和分类结合起来，以及对噪声、模糊和遮挡的鲁棒性。

3.3.网络安全

入侵检测是网络安全的重要组成部分，可在恶意网络活动破坏信息可用性、完整性或保密性之前对其进行检测。入侵检测使用

入侵检测系统（IDS）可将网络流量分为正常和入侵两种。然而，由于正常网络流量往往具有与实际攻击类似的特征，网络安全分析人员需要分析所有入侵警报的情况，以确定是否存在实际攻击。虽然基于特征码的 IDS 通常能很好地检测已知的攻击模式，但却无法检测以前未见过的攻击。此外，由于需要大量的专业知识，基于特征码的检测开发通常既缓慢又昂贵。这妨碍了系统对快速演变的网络威胁的适应性。

许多研究利用 ML 和其他 Al 技术来提高已知攻击的分类准确性、检测异常网络流量（因为这可能表明偏离正常网络流量的新攻击模式）以及自动构建模型 [27]。然而，这些系统很少投入使用。究其原因，入侵检测面临着一些特殊的挑战，如缺乏训练数据、网络流量变化大、出错成本高以及难以进行相关评估 [9，55]。虽然可以收集到大量的网络流量，但这些信息往往比较敏感，只能进行部分匿名处理。使用模拟数据是另一种选择，但往往不够真实。然后，必须对数据进行标注，以便在监督学习中确定模式是正常还是入侵，或在异常检测中确定无攻击，而这通常很难做到。最后，模型必须透明，以便研究人员了解检测极限和特征的重要性 [55]。

加强网络安全的另一项措施是在安全审计期间进行渗透测试，以确定可能被利用的安全漏洞。由于许多网络的复杂性和主机数量庞大，渗透测试通常是自动进行的。一些研究探讨了如何利用人工智能技术，使用网络的逻辑模型而不是实际网络进行模拟渗透测试。网络通常用攻击图或树来表示，描述对手如何利用漏洞入侵系统。然而，[23] 描述了模型在描述方式上的不同：1）攻击者的不确定性，从抽象的成功和检测概率到网络状态的不确定性；2）攻击者的行动，从已知的前置和后置条件到一般的结果感知和观察。此外，有了网络和主机的正式模型，就有可能对不同的缓解策略进行假设分析 [5]。未来的渗透测试研究可能会使用认知有效的攻击者与防御者之间的交互模型（例如 [26]），以及深度强化学习来探索可能攻击的巨大问题空间。

4 挑战

正如第 3 节中的案例所示，在为军事目的开发和部署基于人工智能的应用之前，必须意识到一些尚未解决的难题。在本节中，我们将讨论我们认为对军事人工智能最关键的挑战：1）透明度；2）脆弱性；3）即使在训练数据有限的情况下也能学习。其他与优化、泛化、架构设计、超参数调整和生产级部署相关的重要但不那么关键的挑战，本文不再进一步讨论。

4.1.透明度

许多应用除了要求高性能外，还要求高透明度、高安全性和用户信任或理解。这些要求在关键安全系统[29]、监控系统[60]、自主代理[37]、医学[14]和其他类似应用中非常典型。随着最近 Al 技术的突破，人们对透明度的研究兴趣也越来越浓厚，以支持此类应用中的终端用户（如 [20, 24, 42]）。

4.1.1 对透明度的期望

人工智能所需的透明度取决于最终用户的需求。利普顿[34]描述了透明度如何与以下五类用户需求相关：

1.在用户难以质疑系统建议的情况下的信任。然而，用户的信任是基于系统的性能或稳健性、相对于用户的性能，还是基于用户对系统建议的满意程度，这一点可能并不明确。

2.洞察以前未知的因果关系，并用其他方法进行检验。3.了解系统性能的局限性，因为与用户的能力相比，模型的通用性有限。4.关于系统建议的一些额外信息。5.公平性，避免可能导致某些情况下不平等待遇的系统性偏见。例如

> 对信贷申请的评估不应以性别或种族等个人属性为依据，尽管这些属性可能会在总体统计层面上区分人口群体。

原则上讲，有两种方法可以使人工智能系统透明化。首先，某些类型的模型被认为比其他模型更容易解释，例如线性模型、基于规则的系统或决策树。了解这些模型，就能理解它们的组成和计算。Lipton [34] 描述了可解释性如何取决于用户能否预测系统建议、理解模型参数和理解训练算法。其次，系统可以解释其建议。这种解释可以是文字的，也可以是视觉的。例如，说明图像的哪些方面对其分类最有帮助。米勒[38]广泛回顾了社会科学研究中的解释，以及如何利用这些知识为人工智能系统设计解释。通常情况下，人们会根据自己感知到的信念、欲望和意图来解释其他代理的行为。对于人工智能系统来说，信念对应的是系统关于情况的信息，愿望对应的是系统的目标，意图对应的是中间状态。此外，解释可能包括行动的反常性、最小化成本或风险的偏好、预期规范的偏差、事件的反复性以及行动的可控性。主要研究结果如下

+ 解释是针对特定的反事实情况的对比性解释。因此，解释的重点是为什么要提出特定建议而不是其他建议。

+ 解释是有选择的，侧重于一两个可能的原因，而不是建议的所有原因。

+ 解释是一种社会对话和知识传授的互动。

4.1.2 可解释模型举例

贝叶斯规则表（BRL）是可解释模型的一个例子。贝叶斯规则表由一系列 if（条件）then（结果）else（替代）语句组成。Letham 等人[33] 描述了如何生成贝叶斯规则表，以建立一个高精度、可解释的模型来估计中风风险。条件离散化了影响中风风险的高维多变量特征空间，结果描述了预测的中风风险。BRL 在预测中风风险方面的性能与其他 ML 方法相似，而且与其他准确性较低的现有评分系统一样具有可解释性。

基于词典的分类器是文本分类可解释模型的另一个例子。基于词典的分类器将术语的频率与术语在每个类别中出现的概率相乘。得分最高的类别被选为预测对象。Clos 等人[11]使用门控递归算法对词典进行建模。

该网络可共同学习术语和修饰词，如副词和连词。这些词库根据论坛中支持或反对死刑的帖子以及对商业产品的情绪进行训练。词典的表现优于其他 ML 方法，同时具有可解释性。

4.1.3 特征可视化实例

尽管 DNN 在许多应用中都具有很高的性能，但由于 DNN 的亚符号计算需要数以百万计的参数，因此很难准确理解输入特征对系统推荐的贡献。由于 DNNs 的高性能对许多应用至关重要，因此人们对如何使 DNNs 更易于解释产生了浓厚的兴趣（综述见 [39]）。许多用于解释 DNN 的算法都会将 DNN 处理转换到原始输入空间，以便直观地识别特征。通常，有两种通用方法可用于特征可视化，即激活最大化和 DNN 解释。

激活最大化计算哪些输入特征能最大限度地激活可能的系统建议。对于图像分类来说，这代表了理想的图像，能显示出每个类别的可区分和可识别特征。然而，由于类别可能会使用同一对象的许多方面，而且图像中的语义信息往往比较分散，因此图像看起来往往不自然[43]。激活最大化的一些方法包括梯度上升法 [13]、更好的正则化以提高泛化能力 [54]，以及合成首选图像 [41，40]。

DNN 解释通过突出具有区分性的输入特征来解释系统建议。在图像分类中，这种可视化方法可以突出显示支持或反对某个类别的区域 [68]，或者只显示包含判别特征的区域 [3]。计算鉴别特征的一种方法是使用局部梯度或其他变异度量进行灵敏度分析 [39]。不过，灵敏度分析的一个问题是，它可能会显示输入中不存在的判别特征。例如，在图像分类中，灵敏度分析可能会显示出物体的模糊部分，而不是可见部分 [51]。分层相关性传播可以同时考虑特征的存在和模型的反应，从而避免这一问题[4]。

4.1.4 针对具体应用的解释示例

与分类不同，Al-planning 基于领域动态模型。Fox 等人[15]描述了规划解释如何使用领域模型来解释为什么执行或不执行行动、为什么不能执行某些行动、促成未来行动的因果关系以及重新规划的必要性。由于公平性对许多人工智能应用非常重要，Tan 等人[59] 描述了如何利用模型蒸馏来检测黑盒模型中的偏差。模型蒸馏可简化更大更复杂的模型，而不会明显降低准确性。为了提高透明度，他们使用了基于浅层树的广义加法模型，对每个参数和两个参数之间的相互作用进行建模。他们根据黑箱模型中的系统建议训练一个透明模型，并根据实际结果训练一个透明模型。通过对两个模型的建议差异进行假设检验，可以发现黑箱模型引入偏差的情况，然后可以通过比较两个透明模型来诊断偏差。该系统对累犯风险、贷款风险和个人卷入枪击事件的风险进行了评估。结果显示，一个黑盒模型低估了年轻罪犯和白种人的累犯风险，而高估了土著人和非裔美国人的风险。

4.2 漏洞

在本节中，我们将讨论 DNN 漏洞的两个不同方面：1) 输入操纵的脆弱性和 2) 模型操纵的脆弱性。我们首先来看输入信号的操纵。

4.2.1 对输入进行对抗性加工

在 DNN 的情况下，人们发现调整输入信号很容易使分类系统完全失效 [58, 18, 45]。当输入信号的维度较大时（例如图片的典型情况），通常只需对输入信号中的每个元素（即像素）进行不易察觉的微小调整，就足以骗过系统。利用用于训练 DNN 的相同技术（通常是随机梯度法[49]），通过观察梯度的符号，就能轻松找到每个元素应该朝哪个方向改变，从而让分类器错误地选择目标类别或只是错误分类。只需几行代码，最好的图像识别系统也会受骗上当，误以为图片上是一辆车，而不是一只狗。下图 1 显示了篡改前后的图像，以及篡改前后类别的可能性。

上述方法假定可以完全访问 DNN，即所谓的白盒攻击。研究发现，即使是所谓的黑盒攻击，即只了解系统的输入和输出类型，也是可能的 [44, 56]。在 [44] 中，作者利用从他们想要攻击的黑盒系统的稀疏采样中获得的数据训练了一个替代网络。有了替代网络，你就可以使用上文提到的白盒攻击方法来制作对抗性输入。文献[56]提出了学习替代网络的另一种方法，即使用遗传算法来创建攻击向量，从而导致系统的错误分类。同一作者甚至指出，通常只需修改图像中的一个像素（尽管通常是可感知的），就能实现成功的攻击。

图 1：从小型货车到西伯利亚哈士奇。右侧显示的是原始图像与经过处理（逆向制作）的图像之间的绝对差异（放大了 20 倍）。对抗示例（中间）是使用 Kurakin 的基本迭代法 (BIM) 生成的，该方法在 [28] 中进行了描述。

4.2.2 利用预训练 DNN 中的隐藏后门

在设计 DNN 时，如果只能获得少量的训练数据，通常会使用预先训练好的模型来实现良好的性能。这一概念被称为迁移学习，常用的方法是使用在大量数据上训练过的模型，根据特定问题替换和定制网络中的最后几层，然后使用可用的训练数据对最后阶段（有时甚至是整个系统）的参数进行微调。互联网上已经有大量预训练模型可供下载。因此，一个相关的问题是 "我们如何知道上传模型的人没有恶意？[19] 中考虑了这种类型的漏洞，作者在其中插入了后门

的模型，用于识别美国的交通标志。例如，在一个停止标志上训练一个贴纸，使其属于停止标志以外的类别。他们随后证明，基于美国交通标志网络的瑞典交通标志识别系统在使用后门（即在交通标志上贴上贴纸）时，会产生负面反应（大大降低瑞典交通标志系统的分类准确性）。

4.2.3 防御方法

减少 DNN 受输入信号操纵影响的一种方法是在模型的训练过程中明确加入操纵/对抗示例[18, 28]。也就是说，除了原始的训练数据外，在模型的训练过程中还会生成并使用对抗性示例。

另一种方法是使用一种名为 "防御蒸馏 "的概念 [46]。简而言之，这种方法试图降低对输出信号的要求，即输出信号只能指出真正的类别，而迫使其他类别的概率为零。46] 分两步实现了这一点。第一步是 DNN 的常规训练。第二步，使用第一个神经元网络的输出（类别概率）作为新的类别标签，并使用新的（软）类别标签训练一个新的系统（具有相同的架构）。事实证明，这样做可以减少漏洞，因为你不会让 DNN 与训练数据贴得太紧，而是保留了一些合理的类之间的相互关系。

其他防御方法包括特征挤压技术（如均值或中值滤波技术）[64] 或非线性像素表示法（如单点编码或温度计编码）[8]。

遗憾的是，上述两种方法都不能彻底解决漏洞问题，尤其是在攻击者完全了解模型和防御方法的情况下。

43 数据

在军事背景下开发基于 ML 的应用具有挑战性，因为军事组织、训练设施、平台、传感器网络、武器等的数据收集程序最初并不是为 ML 目的而设计的。因此，在这一领域通常很难找到真实世界中高质量和足够大的数据集，用于学习和洞察。在本节中，我们将探讨即使在训练数据有限的情况下也能用于构建 ML 应用程序的技术。

43.1 转移学习

迁移学习（在第 4.2.2 节中也有提及）是一种在数据集较小和计算资源有限时常用的技术。其目的是在开发针对其他类似任务的新模型时，重复使用预先训练好的模型（通常由 DNN 表示）的参数。在 DL 应用中，至少有两种方法可用于迁移学习：

+ 重新学习输出层：使用这种方法时，预训练模型的最后一层会被新的输出层取代，新的输出层与新任务的预期输出相匹配。在训练过程中，只更新新输出层的权重，其他的权重都是固定的。

+ 微调整个模型：这种方法与第一种方法类似，但在这种情况下，可以更新整个 DNN 的权重。这种方法通常需要更多的训练数据。

研究表明，迁移学习也可以提高模型的泛化能力。然而，随着源任务和目标任务之间距离的增加，迁移学习的积极效果往往会减弱[66]。

43.2 生成对抗网络

生成对抗网络（GANs）由 Goodfellow 等人发明[17]，是一种生成模型，可用于半监督学习，即将一小部分标记数据与一大部分未标记数据相结合，以提高模型的性能[50]。基本的 GAN 实现由两个 DNN 组成，分别代表生成器和判别器。生成器经过训练可生成虚假数据，鉴别器经过训练可将数据分类为真实或虚假数据。当两个网络同时接受训练时，一个网络的改进也会导致另一个网络的改进，直至最终达到平衡。在半监督学习中，生成器的主要目的是生成未标记的数据，用于提高最终模型的整体性能。除半监督学习外，GAN 还可用于以下方面

+ 重建：填补部分遮挡图像或物体的空白 [65]。+ 超分辨率：将图像从低分辨率转换为高分辨率 [32]。

+ 图像到图像的转换：将冬天的图像转换成夏天的图像，将夜晚的图像转换成白天的图像，等等。[67].这种技术在军事上的应用可以是将夜视图像转换为日光图像。

4.3.3 建模和模拟

建模和仿真已被军方广泛用于培训、决策支持和研究等方面。因此，有许多已经过长期验证的模型也有可能被用于生成 ML 应用的合成数据。例如，飞行模拟器可用于生成飞机在不同环境下的合成图像。在这种情况下，由于在生成合成图像之前已经知道飞机的类型，因此可以自动进行标记。然而，毫不奇怪的是，在将模型应用于真实世界图像时，使用合成图像可能会导致性能不佳。目前正在探索的一种方法是使用 GANs 增强合成图像，使其逼真。这种方法在 [53] 中得到了成功应用。

5 结论

最近，人工智能的突破正逐渐达到可用于军事应用的程度。本文介绍了将人工智能用于监视、水下地雷战和网络安全的一些可能性。其他潜在应用包括使用部分自主车辆和传感器系统进行侦察、在时间要求较高的防空系统中进行威胁评估、对新出现的模式进行情报分析、指挥和控制系统以及教育和培训。不过，人工智能的军事应用需要考虑以下方面的挑战：

+ 确保模型性能符合军事要求的透明度。+ 可能大幅降低系统性能的漏洞。+ 用于 ML 的训练数据不足。

专注于人工智能的透明度、可解释性和可解释性问题的研究人员已经取得了许多进展。其中许多进展也可能用于军事人工智能应用。不过，要了解如何利用这些研究成果，还需要进行更全面的需求分析。在风险、数据质量、法律要求等方面，军事需求可能会有很大不同，有些类型的透明度甚至可能并不适用。此外，还需要对如何利用社会科学研究提高 Al 可解释性进行更多研究。未来的研究还应包括如何利用可视化分析研究领域开发的丰富的可视化技术。

由于目前还没有解决漏洞问题的灵丹妙药，因此必须密切关注这一研究领域，并不断寻找有前途的解决方案。不过，在找到解决方案之前，有必要尽量减少外部对模型和防御技术的访问。否则，对手可能会试图利用漏洞为自己谋利。最后，迁移学习可以使预先训练好的模型适用于训练数据和计算资源都有限的军事应用。GAN 是另一种有前途的技术，它可以
使用有标签和无标签数据进行学习（半监督学习）。GAN 还可与模拟相结合，以提高合成训练数据的真实性。

致谢

这项工作得到了瑞典武装部队研发计划资助的 FOI 研究项目 "决策支持和认知系统的人工智能 "的支持。

参考文献

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, and Carl Case et al. Deep speech 2 : End-to-end speech recognition in english and mandarin.见 Maria Flo- rina Balcan 和 Kilian Q. Weinberger 编辑，《第 33 届机器学习国际会议论文集》，《机器学习研究论文集》第 48 卷，第 173-182 页，美国纽约，2016 年 6 月 20-22 日。PMLR。

[2] Yannis M Assael、Brendan Shillingford、Shimon Whiteson 和 Nando de Freitas。Lipnet：端到端句子级唇语阅读。GPU 技术大会，2017 年。

[3] Housam Khalifa Bashier Babiker 和 Randy Goebel.使用 KL-发散聚焦深层视觉外显。ArXiv preprint arXiv: 1711.06431,2017.

[4] 塞巴斯蒂安-巴赫、亚历山大-宾德、格雷瓜尔-蒙塔翁、弗雷德里克-克劳琛、克劳斯-罗伯特-米勒和沃伊切赫-萨梅克。通过层相关性传播对非线性分类器决策的像素解释。PloS one, 10(7):e0130140, 2015.

[5] Michael Backes, Jorg Hoffmann, Robert Kiinnemann, Patrick Speicher, and Marcel Steinmetz.模拟渗透测试与缓解分析》，arXiv preprint arXiv: 1705.05088, 2017.

[6] Yoshua Bengio、Pascal Lamblin、Dan Popovici 和 Hugo Larochelle.深度网络的贪婪分层训练》，第 153-160 页。神经信息处理系统国际会议，NIPS，2006。

[7] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba.自动驾驶汽车的端到端学习。CoRR，abs/1604.07316，2016。

[8] Jacob Buckman、Aurko Roy、Colin Raffel 和 Ian Goodfellow。温度计编码：抵御对抗性示例的热门方法之一。国际学习表征会议，ICLR，2018。https://openreview.net/pdf?id＝S 18Su-CW。

[9] Carlos A Catania 和 Carlos GarciA Garino.自动网络入侵检测：当前的技术和有待解决的问题。计算机与电气工程》，38（5）：1062-1072，2012 年。

[10] Kyunghyun Cho、Bart van Merriénboer、Calar Giilgehre、Dzmitry Bahdanau、Fethi Bougares、Hol-ger Schwenk 和 Yoshua Bengio。使用rn编码器-解码器学习短语表征，用于统计机器翻译。In Proceedings of the 2014 Conference on Empirical Methods in Nat- ural Language Processing (EMNLP), pages 1724-1734, Doha, Qatar, October 2014.计算语言学协会。

[11] J Clos、N Wiratunga 和 S Massie。通过联合学习词库和修饰词实现可解释文本分类。在 LJCAI-17 Workshop on Explainable Al (XAlI) 中，第 19-23 页，2017 年。

[12] Killian Denos、Mathieu Ravaut、Antoine Fagette 和 Hock-Siong Lim。深度学习应用于水下地雷战。In OCEANS 2017-Aberdeen, pages 1-7.IEEE, 2017.

[13] Dumitru Erhan、Yoshua Bengio、Aaron Courville 和 Pascal Vincent。可视化深度网络的高层特征。蒙特利尔大学，1341（3）：1-13，2009。

[14] 约翰-福克斯（John Fox）、大卫-格拉斯普尔（David Glasspool）、丹-格雷库（Dan Grecu）、桑杰-莫吉尔（Sanjay Modgil）、马修-南（Matthew South）和维维克-帕特卡尔（Vivek Patkar）。基于论证的推理和决策--医学视角。JEEE 智能系统》，22（6），2007 年。

[15] Maria Fox, Derek Long, and Daniele Magazzeni.可解释规划》，arXiv preprint arXiv:1709.10256,2017.

[16] Xavier Glorot 和 Yoshua Bengio.理解深度前馈神经网络的训练难度》，第 249-256 页。人工智能与统计学国际会议，AISTATS，2010。

[17] 扬-古德费洛、让-普热-阿巴迪、迈赫迪-米尔扎、徐兵、戴维-沃德-法利、谢尔吉尔-奥扎尔、亚伦-库维尔、尤斯华-本吉奥。生成对抗网见 Z. Ghahramani、M. Welling、C. Cortes、N. D. Lawrence 和 K. Q. Weinberger 编著的《神经信息处理系统进展》第 27 期，第 2672-2680 页。Curran Associates, Inc., 2014.

[18] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy.解释和利用对抗性实例》，第 1-11 页。学习表征国际会议，ICLR，2015。https://arxiv.org/abs/1412.6572。

[19] Tianyu Gu、Brendan Dolan-Gavitt 和 Siddharth Garg。Badnets：识别机器学习模型供应链中的漏洞。https://arxiv.org/abs/1708.06733，2017 年 8 月。

[20] 戴维-冈宁可解释人工智能（XAI）。美国国防部高级研究计划局（DARPA），2017 年。

[21] Geoffrey E. Hinton、Simon Osindero 和 Yee-Whye Teh。深度信念网的快速学习算法。神经计算》，（18）：1527-1554，2006 年。

[22] Geoffrey E. Hinton、Nitish Srivastava、Alex Krizhevsky、Ilya Sutskever 和 Ruslan R. Salakhutdinov。通过防止特征检测器的共同适应来改进神经网络》，https://arxiv.org/abs/1207.0580, 7 2012.

[23] Jérg Hoffmann.模拟渗透测试：从 "dijkstra "到 "图灵测试＋＋"。见 JCAPS，第 364-372 页，2015 年。

[24] CAI.可解释人工智能（XAI）研讨会。2017.

[25] Sergey Ioffe 和 Christian Szegedy.批量归一化：通过减少内部协变量偏移加速深度网络训练。国际学习表征会议，ICLR，2015。https://arxiv.org/abs/ 1502.03 167.

[26] Randolph M Jones、Ryan OGrady、Denise Nicholson、Robert Hoffman、Larry Bunch、Jeffrey Brad-Shaw 和 Ami Bolton。认知代理在新兴网络环境中的建模与整合。In Proceedings of the Interservice/ Industry Training, Simulation, and Education Conference (I/ITSEC), volume 20.Citeseer，2015。

[27] Gulshan Kumar, Krishan Kumar, and Monika Sachdeva.基于人工智能的入侵检测技术：综述》。人工智能评论》，34（4）：369-387，2010 年。

[28] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio.大规模对抗式机器学习。https://arxiv.org/abs/161 1.01236，2016 年 11 月。

[29] Zeshan Kurd、Tim Kelly 和 Jim Austin。为关键安全系统开发人工神经网络。神经计算与应用》，16（1）：11-19，2007 年。

[30] Rikard Laxhammar.海上监控的异常检测。In Information Fusion, 2008 11th Inter- national Conference on, pages 1-8.IEEE, 2008.

[31] Rikard Laxhammar, Goran Falkman, and Egils Sviestins.海上交通异常检测--高斯混合物模型与核密度估计器的比较。信息融合，2009 年。FUSION'09.第 12 届国际会议，第 756-763 页。IEEE, 2009.

[32] Christian Ledig、Lucas Theis、Ferenc Huszar、Jose Caballero、Andrew Cunningham、Alejandro Acosta、Andrew P. Aitken、Alykhan Tejani、Johannes Totz、王泽涵和史文哲。利用生成式对抗网络实现逼真的单图像超分辨率。In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 105-114, 2017.

[33] Benjamin Letham、Cynthia Rudin、Tyler H McCormick、David Madigan 等。 使用规则和贝叶斯分析的可解释分类器：建立更好的中风预测模型。应用统计学年鉴》，9（3）：1350-1371，2015。

[34] Zachary C Lipton.ArXiv preprint arXiv: 1606.03490, 2016.L. J. Luotsinen, F.

[35.] Kamrani, P. Hammar, M. Jandel, and R. A. Lovlid.进化的创造性智能或计算机生成的力量。在 2016 年电气和电子工程师学会系统、人和控制论国际会议上(SMC），第 003063-003070 页，2016 年 10 月。

[36] Steven Mascaro、Ann E Nicholso 和 Kevin B Korb。使用贝叶斯网络的船只航迹异常检测。/nternational Journal of Approximate Reasoning，55（1）：84-98，2014.

[37] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci.智能代理透明化的人机协作多变量管理。人类因素》，58（3）：401-415，2016。

[38] 蒂姆-米勒人工智能中的解释：ArXiv preprint arXiv:1706.07269, 2017.

[39] Grégoire Montavon、Wojciech Samek 和 Klaus-Robert Miiller。解读和理解深度神经网络的方法。数字信号处理》，2017 年。

[40,] Anh Nguyen、Alexey Dosovitskiy、Jason Yosinski、Thomas Brox 和 Jeff Clune。通过深度生成器网络合成神经网络中神经元的首选输入。神经信息处理系统进展》，第 3387-3395 页，2016 年。

[41] Anh Nguyen、Jason Yosinski 和 Jeff Clune。多方面特征可视化：ArXiv preprint arXiv: 1602.03616,2016.NIPS.可解释

[42] 人工智能（XAI）研讨会。2017.

[43] 法比安-奥弗特"一看便知"。ArXiv preprint arXiv:1711.08042,2017.

[44] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami.针对机器学习的实用黑盒攻击》，第 1-14 页。亚洲计算机会议'17，2017。https://arxiv.org/abs/1602.02697v4。

[45] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredikson, Z. Berkay Celik, and Ananthram Swami.对抗环境下深度学习的局限性》，第 1-11 页。IEEE European Sympo- sium on Security & Privacy, 2016. https://arxiv.org/abs/1511.07528v3.

[46] Nicolas Papernot、Patrick McDaniel、Xi Wu、Somesh Jha 和 Ananthram Swami。蒸馏作为防御对抗性扰动的深度神经网络。2016 IEEE 安全与隐私研讨会论文集》，第 582-597 页。IEEE, 2016.

[47] Bradley J Rhodes、Neil A Bomberger、Michael Seibert 和 Allen M Waxman。使用学习机制的海上态势监控与感知。军事通信会议，MIL- COM，第 646-652 页。IEEE, 2005.

[48] Bradley J Rhodes、Neil A Bomberger 和 Majid Zandipour。用于海上态势感知的多空间尺度船舶运动模式的概率关联学习。In Information Fusion, 2007 10th International Conference on, pages 1-8.IEEE, 2007.

[49] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams.通过反向传播错误学习表征。自然》，（9）：533-536，1986 年。

[50] Tim Salimans、Ian Goodfellow、Wojciech Zaremba、Vicki Cheung、Alec Radford、Xi Chen 和 Xi Chen。改进的甘斯训练技术见 D. D. Lee、M. Sugiyama、U. V. Luxburg、I. Guyon 和 R. Garnett 编辑的《神经信息处理系统进展》第 29 期，第 2234-2242 页。Curran Associates, Inc., 2016.

[51] Wojciech Samek、Thomas Wiegand 和 Klaus-Robert Miiller。可解释的人工智能：Un- derstanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296, 2017.

[52] Jonathan Shen、Ruoming Pang、Ron J. Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、Yu Zhang、Yuxuan Wang、R. J. Skerry-Ryan、Rif A. Saurous、Yannis Agiomyrgiannakis 和 Yonghui Wu。通过WaveNet对mel spectrogram预测的调节实现自然TTS合成。CoRR，abs/1712.05884，2017。

[53] Ashish Shrivastava、Tomas Pfister、Oncel Tuzel、Josh Susskind、Wenda Wang 和 Russell Webb。通过对抗训练从模拟和无监督图像中学习。CoRR，abs/1612.07828，2016。

[54] Karen Simonyan、Andrea Vedaldi 和 Andrew Zisserman.深度卷积网络：图像分类模型和显著性图的虚拟化。arXiv preprint arXiv:1312.6034, 2013.

[55] 罗宾-萨默和弗恩-帕克森。走出封闭世界：使用机器学习进行网络入侵检测。安全与隐私（SP），2010 IEEE 研讨会，第 305-316 页。IEEE, 2010.

[56] 苏嘉伟、Danilo V. Vargas 和樱井光一．欺骗深度神经网络的单像素攻击。https://arxiv.org/abs/1710.08864，102017。

[57] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman.合成奥巴马：从音频学习唇部同步。ACM Trans.Graph., 36(4):95:1-95:13, July 2017.

[58] 克里斯蒂安-塞格迪（Christian Szegedy）、沃伊切赫-扎伦巴（Wojciech Zaremba）、伊利亚-苏茨克沃（Ilya Sutskever）、琼-布鲁纳（Joan Bruna）、杜米特鲁-埃尔汗（Dumitru Erhan）、伊恩-古德费洛（Ian Goodfellow）和罗布-弗格斯（Rob Fergus）。神经网络引人入胜的特性。国际学习表征会议，ICLR，2014。https://arxiv.org/abs/1312.6199。

[59] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou.利用透明模型蒸馏检测黑箱模型中的偏差》，arXiv preprint arXiv:1710.06169, 2017.

[60.] Michael Tom Yeh 等人. 利用具体分析为计算机视觉的未来设计道德指南针。In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 64-73, 2017.

[61] Kenneth Tran、何晓东、张磊、孙健、Cornelia Carapcea、Chris Thrasher、Chris Buehler 和 Chris Sienkiewicz。野外丰富的图像标题。CoRR，abs/1603.09016，2016。

[62] Oriol Vinyals 和 Quoc V. Le.神经会话模型。CoRR，abs/1506.05869，2015.

[63] David P Williams.使用深度卷积神经网络进行合成孔径声纳图像中的水下目标分类。模式识别（ICPR），2016 年第 23 届国际会议，第 2497-2502 页。IEEE, 2016.

[64] Weilin Xu, David Evans, and Yanjun Qi.特征挤压可减轻和检测卡利尼/瓦格纳对抗示例。http://arxiv.org/abs/1705.10686，5 2017.

[65] 杨波、温宏凯、王森、罗纳德-克拉克、安德鲁-马卡姆和尼基-特里戈尼。利用对抗学习从单一深度视图重建 3D 物体。国际计算机视觉研讨会（ICCVW），2017 年。

[66] Jason Yosinski、Jeff Clune、Yoshua Bengio 和 Hod Lipson。深度神经网络中的特征有多大的可转移性？In Proceedings of the 27th International Conference on Neural Information Pro- cessing Systems - Volume 2, NIPS' 14, pages 3320-3328, Cambridge, MA, USA, 2014.麻省理工学院出版社。

[67] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros.使用循环一致对抗网络的非配对图像到图像翻译。arXiv 预印本 arXiv: 1703.10593, 2017.

[68] Luisa M Zintgraf、Taco S Cohen、Tameem Adel 和 Max Welling。可视化深度神经网络决策：ArXiv preprint arXiv: 1702.04595, 2017.