利用多模态深度强化学习训练交互式仿人机器人

**摘要---训练机器人使用多种模式感知、行动和交流仍然是一个具有挑战性的问题，特别是如果期望机器人从一组小的示例交互中高效地学习的话。我们将一种学习方法描述为朝着这个方向迈出的一步，我们教一个类人机器人如何玩零和叉的游戏。考虑到可以训练多种多模态技能来玩这个游戏，我们将注意力集中在训练机器人感知游戏，并在这个游戏中进行互动。我们的多模态深度强化学习代理感知多模态特征，并在玩游戏时表现出言语和非言语行为。仿真实验结果表明，该机器人可以学习赢得或多达98%的游戏。针对目标游戏整合语音，视觉和手势的拟议多模式系统的初步测试 - 报告使用所提出的方法可以实现合理和流动的交互。**

**1.介绍**

交互式类人机器人能够同时感知、行动、交流和学习，这不仅对展示机器人的能力很有趣，而且它们有潜力被用于研究人类的具体智力。本文描述了在这些方向上的一个小步骤，其中我们装备了一个具有多个输入和输出模式的人形机器人1，以便玩“noughts and cross”游戏，也称为“井字游戏”，参见图1。这些模式允许机器人聆听人类的指令，在目标游戏中看到人类的凝视和人类的图画，专注地注视人类玩家，与人类玩家交谈，画零/叉，并从例子中学习——所有这些都是异步的。后一种能力(学习)需要非常有效的训练形式来接近类人行为。虽然之前的工作已经研究了机器人从与环境的真实互动中学习——主要是没有语言能力，但训练一个自主的机器人学习即使是简单的行为也需要大量的经验。之前的其他研究已经解决了多模态深度学习，但在非对话设置。

虽然之前的作品都是从原始像素学习的，但我们的学习方法是半解耦的，分为两个任务:(1)学习感知和(2)学习互动。它是半解耦的，因为学习交互的任务使用多模态感知，这需要从多模态特征而不是单模态特征学习。由于这两个任务都是高维系统，前者使用深度监督学习，后者使用深度强化学习。这种两阶段学习方法的优点是可以更有效地实现多模态学习。让我们假设我们的机器人必须被训练使用不同大小的网格来玩。虽然原始像素方法必须重新学习其行为以适应新的网格大小，但我们的方法只能重新学习感知并重用其学习的行为来进行交互。

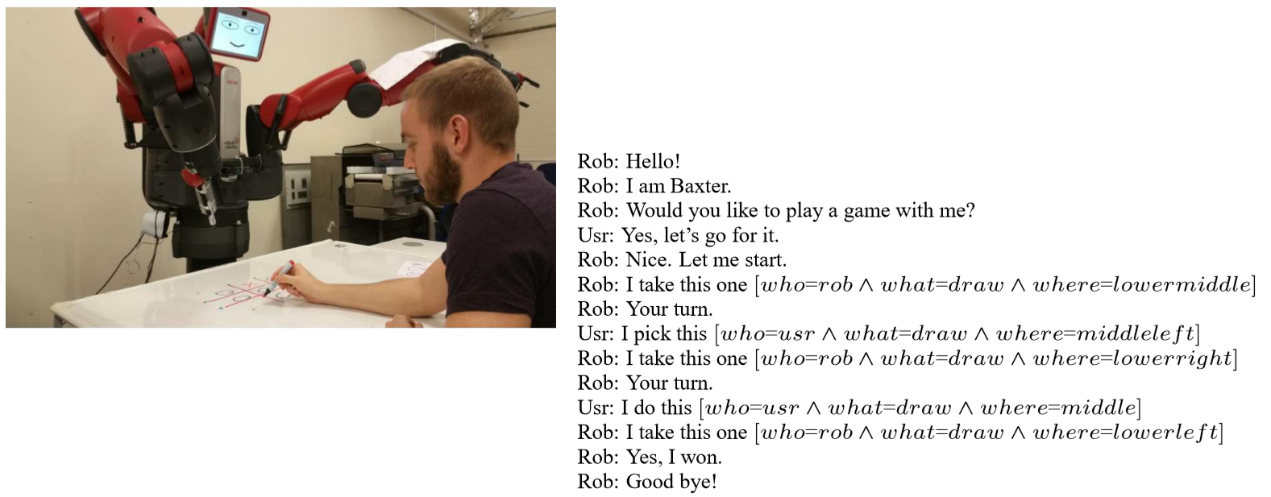


图 1 例如，在玩“零”和“叉”的游戏时，人与类人的互动。方括号中的命令表示手写的物理动作。

同样，假设我们的机器人必须被训练使用越来越多的语言特征(比如单词)进行互动，我们的方法将不得不重新学习它的语言行为，而不是它基于视觉的感知。这两个例子说明了我们提出的方法的好处，可以看作是一种间接的转移学习形式。由于行为在未知环境中的复杂性，在本文中，我们使用一个大小的网格来说明我们的多模态深度强化学习方法——参见图1中的示例交互。

**2. 物理人机交互的学习方法**

我们的方法是使用两个独立但相关的学习任务。首先，学习感知，以便预测环境中发生了什么(在我们的例子中是游戏移动)。第二，学会互动，以便决定下一步该做什么或说什么。通过这种方式，类人机器人可以通过单词和游戏动作(一组更紧凑的多模态特性)来学习互动，而不是通过单词或语音或像素。虽然我们的方法意味着由于更紧凑的环境状态，学习效率更高，但后一种状态表示(原始多模态特性)在未来的工作中仍有待研究。

2.1通过深度监督学习来学习感知

我们使用机器人右臂上的摄像头来感知游戏网格中的符号。我们使用多个图像(在网格中的每个位置一个，在我们的例子中是9个)来检测用于生成游戏移动的新绘图，而不是使用单个图像。机器人连续拍摄图像，并将其分割成9张40×40像素的图像，如图2所示。

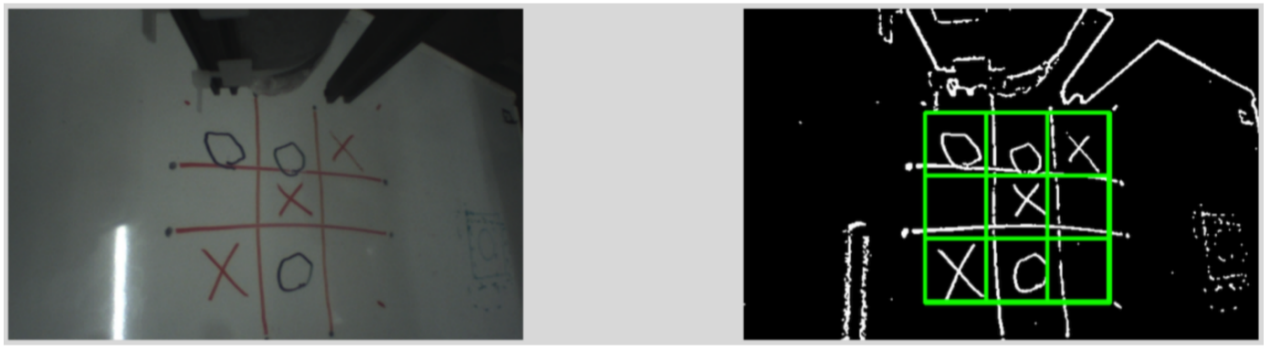


图 2 (左)原始输入图像。(右)用于游戏移动识别的灰度图像。

图片包含 动物

描述已自动生成给定的数据集，其中Xi为n×n个基于像素的特征矩阵，Yj为类标签，任务是将图像映射到标签。在我们的例子中，图像有40×40像素，标签是{'circle’,‘cross','nothing'}。我们使用深度监督分类器从函数H= {h|h: X→Y}空间中归纳出函数h: X→Y，其中X为图像，Y为标签。标记过程定义为h(x) = arg maxy f(x,y)，其中f是使用卷积神经网络中获得的特征x进行评分的函数. 为了训练这个分类器，我们使用一组带种子标记的图像(参见图4)来生成更多的图像，但是在随机分配的位置使用这些图(圆圈和叉)。例如，一个中间有十字的图像可以通过向左或向右、向上或向下移动来生成更多的图像。

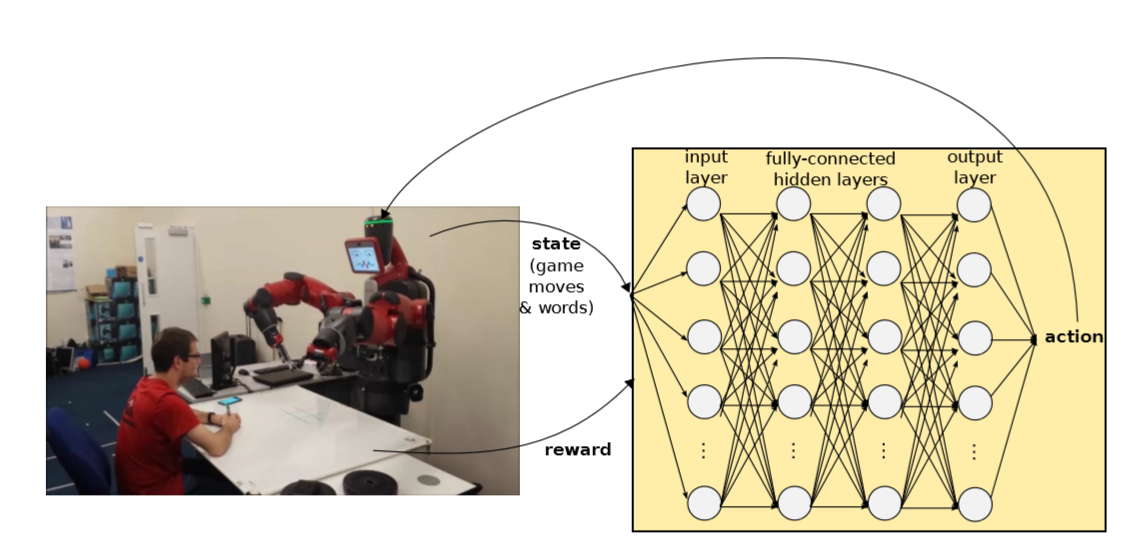
我们的卷积神经网络使用体系结构如图3所示用以下层:输入层40×40像素,卷积和8层过滤器,RELU,池层大小2×2 和2步,卷积和16层过滤器,RELU,池大小3×3步3层和输出层使用支持向量机(SVM)与3标签。这个分类器经常连续地在用户回合中使用，每100毫秒一次，以检测游戏网格中每个位置的活动(人类玩家的绘图)。此外，为了减少噪音，我们接受一个新的图纸，如果它已被识别至

图 5 (左)基于物理游戏的环境，(右)用于类人交互的深度强化学习代理——详见文本

少连续3次。通过这种方式，我们的感知组件可以输出以下格式的游戏动作：[who=usr∧what=draw∧ where=middle].

2.2 学习与深度强化学习互动

上面的视觉感知加上基于语音的感知(带有自信得分的单词)被作为输入给强化学习代理，以诱导其从与环境的交互中产生行为，在环境中，情景通过最大化长期奖励信号映射到行为。RL代理的典型特征是:(i)有限状态集S = {si};一组有限的行动a = {aj};(iii)指定给定当前状态s和动作a的下一个状态s的状态转换函数T(s,a,s ');(iv)奖励函数R(s,a,s ')，指明当环境从一个状态s过渡到另一个状态s '时，给予选择行为a的代理的奖励;和π(v)政策:S→A, 它定义了从状态到操作的映射。RL代理的目标是通过最大化其定义为图片包含 物体

描述已自动生成的累积折扣奖励来找到最优策略。

**状态空间** 我们的学习代理的状态空间S = {Si}包含了57个特征，描述了在最后一次系统和用户回合中游戏的动作和单词。从系统响应中得到的单词被视为二元变量(即单词存在或不存在)，而从嘈杂用户响应中得到的单词则可以通过考虑置信度得分被视为连续变量。由于我们每个单词只使用一个变量，所以在出现重叠时，用户特性会覆盖系统特性。与基于文字的功能相比，游戏的动作确实会考虑到每个游戏的背景或历史。

**行为空间** 动作空间A包含18个对话行为，对话行为属于零与叉的范畴。不是在每种状态下都使用所有的动作来学习，而是从最可能的动作Pr(a|s) > 0.001派生出动作集，概率来自于从示例对话训练的朴素贝叶斯分类器。此外，如果一个物理动作包含在动作集中，我们将包含所有有效的物理动作，以允许代理探索不同的游戏动作。

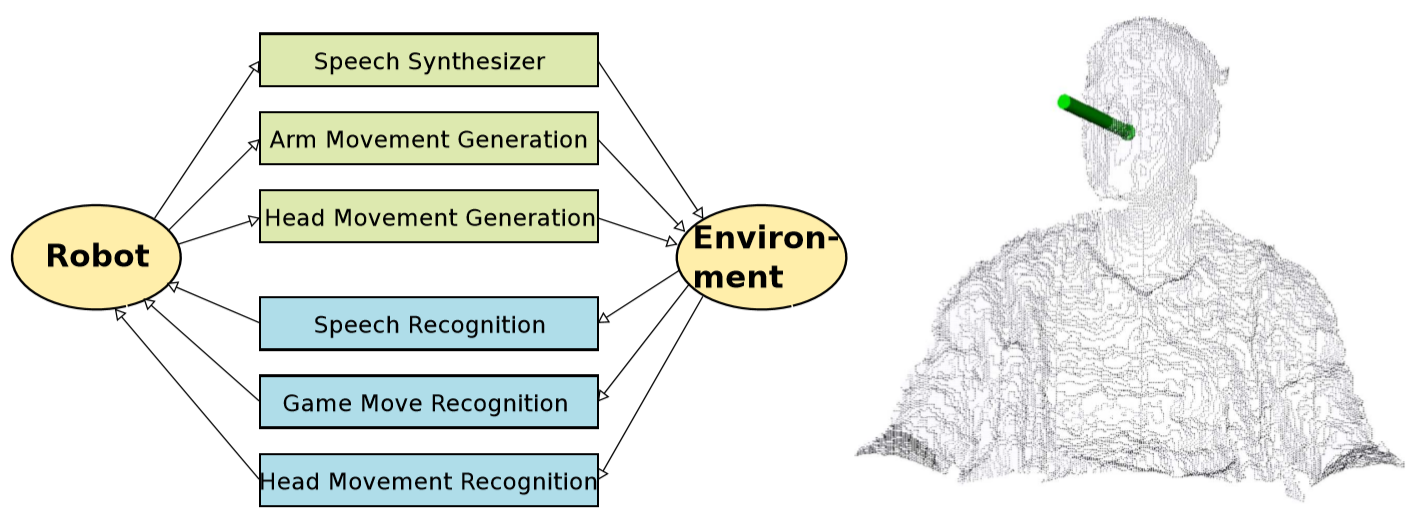


图 6 (左)仿人机器人整体系统除头部动作模仿外，还采用深度强化学习交互管理器进行编排。(右图)采用三维头部跟踪，根据[7]产生的信号检测到的模式，观察方向的变化。

**状态转换函数**  这个函数基于一个数值向量，它表示最后一个系统和用户基于单词的响应，以及游戏历史。后者意味着我们保留了游戏的move特性来描述游戏状态，而不是在每个回合都重置它们。系统响应很简单，如果没有，0;如果存在，1。用户响应对应于置信级别[0..1]噪声用户响应。

**奖励函数** 它的动机在于对话应该是类人的、基于游戏的。DR分数来自上面相同的统计分类器，它允许我们对给定的状态(Pr(a|s))进行统计推断。此外，该函数在交互(对话)结束时提供了如下奖励:0表示输掉游戏，5表示输掉游戏。

**模型架构**  它由一个完全连通的多层神经网络组成，输入层57个节点，第一层和第二层60个节点，输出层18个节点(动作集)。隐藏层使用RELU(校正的线性单位)激活函数来标准化它们的权重，详情见[15]。其他学习参数包括:经验回放大小=100K，折扣因子=0.7，最小值=0.01，学习率=0.001，批量大小=32。

**3.实验和结果**

在本节中，我们将上述方法和学习代理应用于一个学习玩零和叉的类人机器人。

3.1 整合系统

我们的人形机器人配备了多种模式——包括语音、触觉和视觉——来玩目标游戏。为此，我们使用了现成的组件和专门为基于ros的集成系统构建的组件。这些组件通过多线程并发运行，解释如下。

**语义识别**  这个组件在Android应用程序上运行带有触摸对话机制的谷歌语音识别器。该组件通过蓝牙将语音识别结果(也称为“N-best lists”)传递给我们基于ros的集成系统。这些基于语音的感知作为深度强化学习代理状态空间的特征。

**游戏移动识别** 该组件运行2.1中描述的基于视觉的感知子系统，见图2。简而言之，它遵循以下步骤：从右臂的预定义初始位置获取RGB图像（用于一致的感知），将它们转换为灰度，移除网格，将每个图像分成9个图像（游戏网格中每个位置一个图像） ），基于具有学习特征的SVM分类器（标签：圆形，交叉，无）预测标签，并且基于新观察到的标签连续生成至少3次的游戏移动（以避免噪声）。虽然该组件可用于识别所有系统和用户的游戏移动，但它只用于识别用户的游戏移动。这些基于视觉的感知作为深度强化学习代理状态空间的特征。

**头移动识别** 头部跟踪用于检测人类玩家头部方向的变化(如左、右、上、下和中)。为此，我们使用Kinect传感器和[7]中描述的算法从基于深度的感官数据中提取模式，见图6(右)。使用这些模式，我们使用基于阈值的方法跟踪一

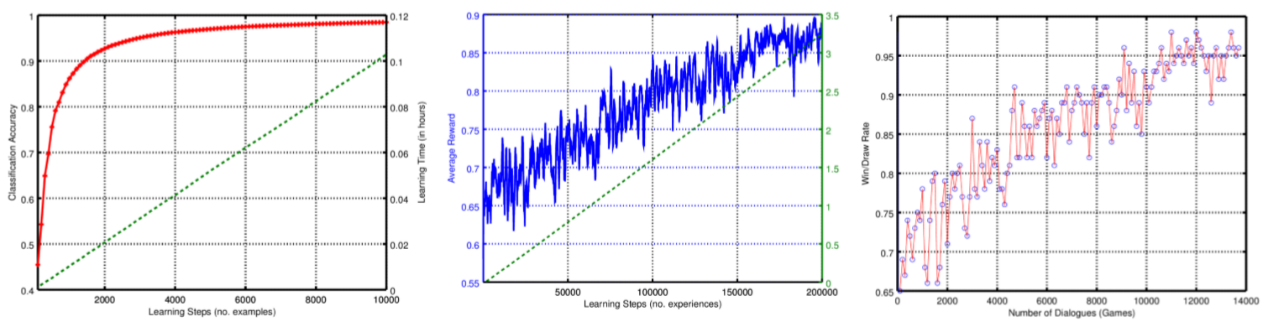


图7 深度监督/强化学习者多模态交互学习曲线:(左)分类精度，(中)平均奖励，(右)赢/抽得率。

组基本的移动(左、右、上、下、中)。这让机器人知道用户在看哪里，以便模仿头部运动，给人一种机器人正在跟随人类玩家视线的印象。

**语义合成** 这些语言(在英语中)对应于从高级操作转换而来的交互管理器，使用了基于模板的方法和现成的语音合成器。说话的语言与机器人的面部同步——在机器人的头上播放一段视频，在说话时机器人的眼睛和嘴巴会动。视频和语音同时开始和结束，给人的印象是一张同步的说话脸，而不是使用静态的机器人脸。

**手臂运动生成** 该组件接收来自交互管理器的命令，用于在游戏网格中绘制符号。考虑到我们假设了一个静态且固定大小的游戏网格，我们简化了绘制内容和绘制位置的任务——尽管未来的工作应该假设游戏网格是动态的。当手臂运动(以预定义的速度)一收到命令就开始时，它会在执行时通知交互管理器。这样，未来的语言将不得不等到图纸完成，手臂回到最初的位置。

头部移动生成 该组件将人类玩家的头部跟踪动作(来自头部移动识别器)作为输入，以模仿它们。它的输出与头部向左、向右、向上、向下和向中央的运动相对应。例如，如果一个人类玩家把头转向左边，然后看着机器人，机器人也会这样做，除了从它自己的空间角度来看(例如，人类转向左边=机器人转向右边)。这给人的印象是，机器人实际上是在关注人类玩家，也给了机器人更多的活力。

交互管理器 交互管理器基于公共可用的SimpleDS tool5[4]，通过不断接收来自环境的基于语音和基于视觉的感知，并决定下一步和何时执行，来协调上述组件。关于下一步要做什么，它根据第2.2节中描述的学习代理选择操作。其中一半只是口头行为，另一半是多模态行为。例如，通信动作GameMove(gridloc =lowerleft)对应“I take this one

[who=rob∧what=draw∧where=lowerleft]”，其中方括号代表一个物理动作(drawinga circle or cross at givenlocation)。下面评估的策略假设机器人开始使用默认符号=circle。多模式行动的大量储备是未来一个有趣的研究方向。

3.2 实验结果

虽然集成系统能够与人类玩家产生合理流畅的交互(如本视频6所示)，但我们的评估重点是学习感知，并从模拟交互中学习交互。尽管如此，该集成系统已在初步但成功的演示中使用。我们在7个独立的人类用户中试用了该系统，所有人都报告了成功的交互，他们都很喜欢。一个包含多个人类玩家的综合评估是未来的工作。

**深监督学习** 本学习的任务是使用SVM分类器将40x40张灰度图像分类为三个标签(圆圈，交叉，无)，学习特征见第2.1节。分类器使用一组种子图像(如图4所示)，它通过将图像定位在每个图像的不同位置来额外生成图像。图7(左)显示了在训练实例数量不断增加的情况下分类精度的学习曲线，其中使用10K图像的训练时间仅需6分钟，如图7(左，绿色虚线)所示，使用的是一台现代台式计算机(Core i7, 3.4 GHz)。在测试过程中，基于1000张随机生成的种子样本图像，产生了99.9%的分类准确率。这表明我们基于视觉的感知组件足够准确地为目标游戏分类人类笔迹。对人类受试者进行的初步测试报告称，该组件可用于检测来自不同人类玩家的笔迹，使他们能够玩“零”和“叉”的游戏。

深入强化学习 本学习的任务是使用章节2.2中描述的深度强化学习代理，基于基于语音和基于视觉的感知选择高级多模态动作(共18个)。这种学习不使用原始像素，而是将单词和游戏动作作为特征。此外，我们使用用户模拟来提供半随机生成的用户响应，而不是从真实的人-人交互中进行训练。系统操作、系统响应和用户响应源自提供给交互管理器的示例对话。我们使用了一组10个种子样例对话，如图1所示。用户模拟器是半随机的，因为已经选择的响应是不允许的。例如，只有当位置为空时，才允许在网格中间绘制符号。agent的目标是基于类人行为(类似于示例对话)诱导其行为，并尽可能多地获胜。图7(中间)显示了一个平均报酬的学习曲线，根据经验的增加，在每个经验中选择一个动作。经过3个小时的使用与前一个学习者相同的桌面的培训，找到了一个合理的策略。此外，图7(右)显示了win/draw rate与游戏数量之间的学习曲线。这些学习曲线显示了被提议的代理是如何从多模态输入特征和多模态动作中成功学习的——除了给定的奖励，没有其他游戏信息。

**4.相关工作及局限性**

机器学习的最新发展，特别是在深度学习领域，正允许开发更有野心的智能交互系统。例如，以前的交互系统需要在特性选择方面进行大量的工作。现在可以通过学习特征训练深度监督学习者，通过训练深度强化学习者，共同归纳其特征和策略。尽管有了这些进步，但将深度学习应用于交互式机器人绝非易事。例如，对[25]中描述的机器人进行训练，使其从像素点到达目标，仿真成功，但在实际环境中测试失败。其他机器人的行为已经得到了合理的成功诱导——尽管它们通常不像本文中那样针对多种模式。此外，以往的作品中，教代理人玩零和交叉假设完美绘制的网格和符号没有任何多模态输入和输出。我们的机器人通过不完美的人类图画、基于语音的反应和手势来感知。训练机器人以多种方式感知、行动和交流，对最终用户来说非常重要。据我们所知，我们的类人机器人拥有第一个通过深度(强化)学习来学习感知、行动和交流的系统。该系统仅使用108个种子图像和10个种子对话对学习环境和代理进行引导，数据效率较高。

虽然我们的机器人系统相当先进，但它有许多限制，鼓励了有趣的研究途径。首先，我们的机器人采用固定大小的网格，不同大小和形状的网格(由人类绘制)仍有待探索。这将代表一个重大的升级，因为这个方向处理联合感知和手势(手臂运动)。前者要求机器人对未知的图形有更强的感知能力，后者要求机器人在非预先定义的位置和不同大小的图形上绘制符号。其次，我们的机器人使用一键说话的语音交互。将这种方法与麦克风阵列进行比较将会很有趣，这样人类玩家可以有更多的自然互动。第三，我们的机器人使用的词汇量很小(大约40个不同的单词)，而词汇量更大的词汇量也有待探索，特别是如果我们想要一个健谈的机器人，而不是一个反复使用相同语言玩游戏的机器人。第四，同样的，我们的机器人使用预定义的模板来生成语言，并且在训练过程中包含一个位置语言生成器(例如[6])将有助于更自然的交互。第五，机器人离线学习自己的行为，并在与人类用户玩耍时使用自己的最佳行为(无需进一步学习)。当机器人从真实的交互中收集数据时，结合其他形式的学习来训练或再训练机器人将会很有趣[1,8]。无监督、半监督和主动学习在这里可以从无标记的例子或标记的例子中学习，但比标准的监督学习更有效。第六，使用深度学习[5]整合更复杂的零和交叉版本以及其他游戏，这将是非常有趣的，因为它代表了一个研究更有效学习的利基市场。人与类人之间的互动越复杂，其特征和行为就越多，这些特征和行为就会反映在学习的缓慢和不可行上。最后但并非最不重要的是，仍然需要对最终用户进行评估，特别是对公共空间中未知的人类玩家。

**5.结论**

通过感知、行动、交流和学习与环境交互的机器人系统常常面临如何将这些不同的概念组合在一起的挑战。我们描述了一种训练机器人使用多种方式与外界互动的通用方法。该方法不是直接从原始像素训练机器人，而是将整个学习任务简化为两个阶段:学习感知和学习交互。我们通过训练巴克斯特仿人机器人玩零和交叉的游戏来测试我们的方法。实验结果表明，学习感知的准确率达到99.9%，学习交互的准确率达到98%。在一个多模态集成系统中，一个中试报告了使用该方法的合理交互作用。此外，我们的系统被证明是有效的数据，因为大量的数据用于诱导模拟环境(108个种子图像和10个种子对话)，这是相关的可训练机器人系统从例子演示与最终用户-如[23]指出。这是第一个学习感知、行动和交流的深度(强化)学习系统。尽管它的功能是用于演示目的，但在投入使用之前还需要进行一些改进。上一节概述了交互式智能类人机器人的未来发展方向。