

**摘要**

一个人工智能（AI）机器人在现实中的遭遇是人类无法想象的，两点之间的简单导航通常都必须经历复杂的环境与动态变化的元素。例如搜救机器人就无法事先对事故或灾难现场建立地图模型，所以在复杂环境中的导航问题一直没有很好的解决方案。这篇文章把导航问题看作强化学习问题来尝试完善导航技术，通过采集多个联动感知设备的数据，并建立并行辅助任务，最终能使AI更加有效理由数据以及在导航任务中获得更好的表现。本文的重点是将机器人导航的两个子任务，辅助深度检测与闭环检测结合起来，共同参与强化学习过程。当AI机器人置身于3D迷宫模拟环境时，本文提出的方法可以让其仅适用自身感知设备达到人类的探路水平，即便目的地位置处于频繁变化中。同时我们还在文中提供了对AI行为、AI定位以及AI神经网络动态情况的相关分析，各种分析结果都显示了AI能够隐含的学习到导航的关键能力。

**1. 介绍**

现代人工智能虽然正在蓬勃发展，但对于智能还没有一个非常准确的定义，目前能够明确的是，能在任意环境中有效探索与导航属于基础智能行为，而比较成熟的机器人导航技术（SLAM）则主要包含对机器人位置的感知与对环境进行建图。我们在这篇文章中延续了先前对深度强化学习的研究（Mnih, 2015; 2016），试图让AI在学习最大化奖励的同时内在的学习导航能力。这种内在学习的方法让AI把行动决策与表示学习有效结合在一个过程里，使AI能在学习过程中提取尽可能与导航任务相关的特征。然而让AI能在复杂多变的环境中利用强化学习技术有效进行导航还有很多问题。

首先，奖励过于分散是导航任务中的通病，比如搜救机器人在训练过程中需要在庞大的区域中准确找到目标人员或目标物体才能获得奖励，其他元素都是无法提供奖励的。其次，由于环境中的一些变化因素，AI必须像人类一样具备多个记忆空间，比如当目标位置不断发生变化时，就通常需要使用瞬时记忆空间来保存目标位置，使用短期记忆来记录行驶速度与即使检测到的数据，使用长期记忆来记录环境中的静态物体，这与目前已知的人类大脑记忆结构类似。

为了改善统计数据结果，我们将两个附属任务产生的误差引导至强化学习过程中，以确保神经网络训练后能产生跟导航任务有关的表示信息。这两个产生误差的任务分别是：第一个任务关注环境中的3D几何特征，从图像信号中预测深度信号，并即时重建低维度的深度地图，完成好这个任务可以有助于避免障碍物与检测新的路线捷径；第二个任务与闭环检测相关，有助于让AI标记已经发现过的相同物体。

针对前文提到的记忆问题，我们依靠栈式长短期记忆模型（stacked LSTM），并在5个3D迷宫环境中进行学习速度与效果的评估，其结果证明我们的方法在这两个方面均优于当前已知的方法，这些有趣的迷宫包含以下坑人的设置（详见图1）

* 复杂与迥异的迷宫几何构造
* 随机的起始位置与朝向
* 动态变化的目标位置
* 在一局游戏内玩几千个时间步不叫事的超大环境范围

我们在本文也提供了十分重要的分析内容，来阐述AI是如何获取核心导航技能的，毕竟位置感知与地图构建并不能直接属于误差内容，而如果只从AI到达目标位置这样单纯而原始的结果来看，则完全找不到AI已经获取导航技能的证据。特别是我们实验中的AI能解决一些歧义问题并迅速在迷宫中定位自己，这种定位能力又关系到获取高级任务的奖励。

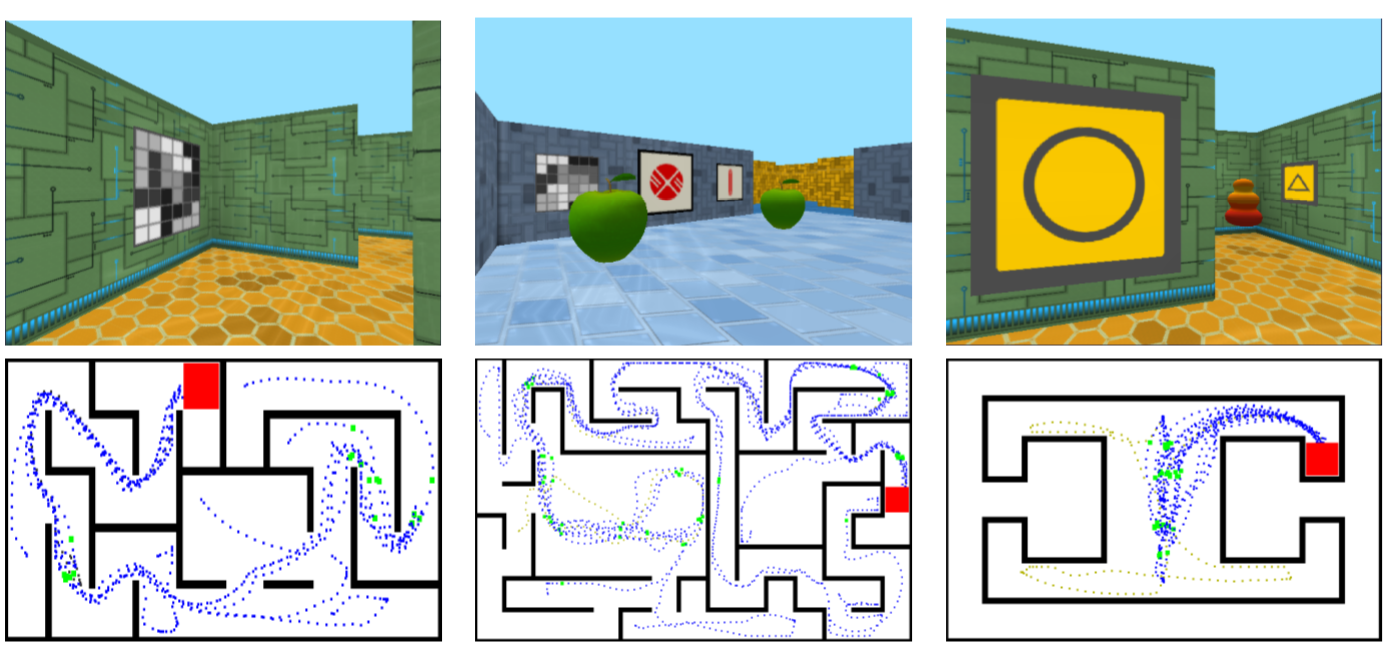


图1

**2. 方法**

我们所提出的端到端学习（通常指同所采集到的信号数据能在一个系统过程中学习为表现形式）框架包含了多个组成部分。首先包含 actor-critic 模型以实现最大化累积奖励，其次拥有从RGB观测信号来构建深度地图以缩小辅助误差，最后还有用于检测闭环与速度累积值得辅助任务模型。

当前解决强化学习问题常使用A3C方法（Asynchronous Advantage Actor-Critic，Mnih，2016），依赖于对观察到每个时间步的状态 计算价值函数 以及更新策略 ，两者都是共享使用中间层的信息表示，且都从最顶层分别进行线性计算得到。对AI系统的设计我们延续Mnih 团队的方法并参考其细节部分（例如在convolutional encoder之后连接MLP或LSTM，还有action repetition的作用，以及使用entropy regularization防止策略饱和），详细内容参见原文附录B

我们首先从一个基本的A3C模型来开始进行设计，这个模型仅能接收采集自环境中的RGB视觉信号数据，内置一个循环或前向算法模块（图2的a与b），那个用来输入RGB信号的encoder由一个3层的卷积网络构成。接着开始为模型增加导航能力，将模型扩展为一个Nav A3C，即加入1个双层的stacked LSTM（图2的c）并将输入信号扩展到动作、奖励值与速度，这里注意是把奖励放在第一个LSTM层来试图寻找奖励与视觉型号之间的联系，以便在第二个LSTM层来进行策略计算。

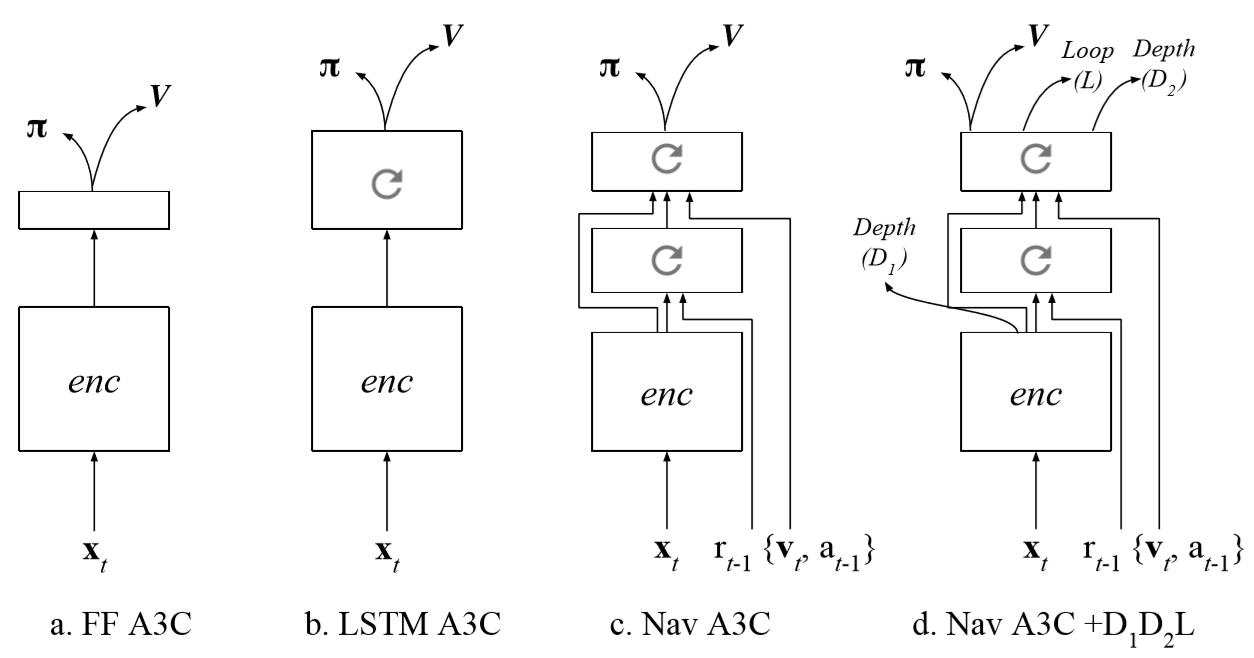


图2

我们从卷积层输出误差 以及最上层输出误差 用来预测深度，从最上层输出误差 用来检测闭环（如图2的d），这些辅助误差由一个单层的MLP模型在每帧即时计算出来，训练时使用加权和的形式，详细内容参见原文附录B。

***2.1 深度预测***

AI收到的主要输入就是RGB图像，而深度信息就包含在AI的事业中部区域，这种深度信息中包含了有价值的3D环境信息。虽然深度信息本身可以利用一些传感器达到直接输入的效果，但我们认为让AI根据RGB图像尝试推断深度信息更有利于加强其自举学习能力（在强化学习中一种让AI学会自我规划与思考能力的方法）。就像Eigen等人在2014年的论文中所提到的，每一帧画面都可以用于学习深度信息。关于更多将深度信息直接输入与学习预测深度信息的比较请参见原文附录C。

辅助误差只是为了建立模型中的信息表现，所以我们不一定关注该误差的优化性能与预测结果。我们更加关心数据效率与计算复杂度，尤其是对于深度预测任务，如果这些误差对主要任务是有效的，那么相比较直接去解强化学习问题的模式而言，我们这种方式能够更迅速的收敛，并且额外计算开销也是最低的。为了达到这个目标，我们使用处理过的深度图（分辨率降低至4x16像素）来代替原图。

我们探索了两种不同的误差优化方法，第一种就是将其简化为回归任务，这也是最常见的选择。虽然这种方法与更高分辨率的深度图相结合后，可以提取最多的信息，但均方误差会产生单峰分布。为了解决这一问题我们把深度这一连续的量分割为8个分布带，非均匀的分布着不同远近的物体，这样我们就能够更关注与较远的物体。这种灵活的离散深度数据的方法也保障了我们可以有效降低分辨率，使收敛更加快速。

***2.2 闭环预测***

在训练过程中我们需要为AI识别出的物体设置标签，而同一个物体即使在不同观察点看到也应当被赋予相同的标签，所以当AI不断进行探索时，再次遇到同一个物体并将其识别出来时我们就认为它已经走了一个闭环，成功检测闭环可以让AI避免走重复的路。当使用坐标位置检测闭环时，我们使用时间作为位置分布的尺度，即 表示AI在时间步0-T时的位置记录。当两个时间点的P非常接近时则认为AI走了一个闭环。这两点的距离由两个阈值 和 控制，当AI确认出现闭合点时有一个标签 会被置为1，而如何让AI学会正确预测标签则通过对上一层的LSTM输出 计算sigmoid与现有 计算一个伯努力误差 来达成。

**3．近期工作**

关于导航甚至机器人导航的文献非常之多，但从深度强化学习（Deep RL）角度出发的就很少了。最著名的就属于2015年DQN在A tari游戏上的成功。离本文时间上较近的研究当属Mnih等人2016年关于多AI异步训练的高级actor-critic方法，RNN也被成功用于非完全环境下的状态消歧。

Deep RL近期频繁被用在导航的研究上，Kulkarni 使用一个前向结构的AI模型在基于迷宫的环境中尝试使用更加灵活的行为改变奖励，这个研究让AI学习到更深层次的后继表示，并为AI在《毁灭战士》这款游戏中的瓶颈提供突破思维。Zhu 将经过预训练的ResNet与前向siamese actor-critic模型合并，完成了离散3D环境中的目标导航。还有在《我的世界》游戏及各种各样需要导航环境中的导航任务，在这里就不一一列举了，详见原文。

从1990年开始就有人将辅助任务用于表示学习，时至今日这个方法仍被继续推广并加强。例如使用受监督的误差去拟合循环模型的隐含层表示，以预测下一个强化学习状态，还有一些在第一人称射击游戏中使用监督学习型辅助任务来进行敌人、武器的检测等并行学习。

相比之下，我们的贡献在于解决如何学习空间、几何与运动的内在关系，以及通过强化学习获得最大奖励的基本问题。我们的方法在多个随机起点与终点的挑战性环境中都得到了高效的验证。

**4. 实验**

我们所使用的环境是DeepMind Lab设计的第一人称3D迷宫（见图1）环境，有不错的视觉表现及信息反馈，能够向AI反馈诸如惯性数值、局部深度信息等。环境包含的动作空间是离散的，共计8种动作：起步加速左转与右转、前后左右4个方向起步加速移动、移动中左转与右转，达到最大速度时会稳定为匀速移动。AI从随机位置及方向出发，每当AI移动至奖励目标位置是获得奖励。一旦AI到达奖励目标时就重新生成起始位置与方向，然后再次让AI试图找到该奖励目标（目标并不会因为被AI找到而消失），反复进行这个过程就可以让AI学会从任意起始方向找到该目标。每个实验场次是由计时器决定AI能玩多久，基本上来说AI拥有充足的时间多次找到目标获得奖励以及找到更好的奖励。环境中包含了少量稀有高分元素：找到苹果得1分，找到草莓得2分，找到传送门得10分。我们训练的能够有效得分的视频链接见原文附录A。

迷宫有两个版本，静态版本中奖励目标在所有场次都不会发生变化，动态版本中奖励目标会在更换场次时进行重置位置。这样可以保障AI的explore-exploit机制，AI会首先探索迷宫，然后逐渐习惯保存被找到奖励的位置，在下一次重生时继续找到该目标。我们为两个版本都考虑了大小尺寸的地图。小尺寸地图是5x10大小与3600最大时间步，大尺寸地图是9x15大小与10800最大时间步。RGB取帧为84x84像素的图片。

代号I-Maze的环境（图1最右边那个）是受到对啮齿动物寻路相关研究（T-Maze）的启发：AI会在中间走廊出生，并试图去寻找随机藏在四个凹形臂弯中唯一奖励。最厉害的AI会依靠记忆并选择直达路线（不会中途走回头路）

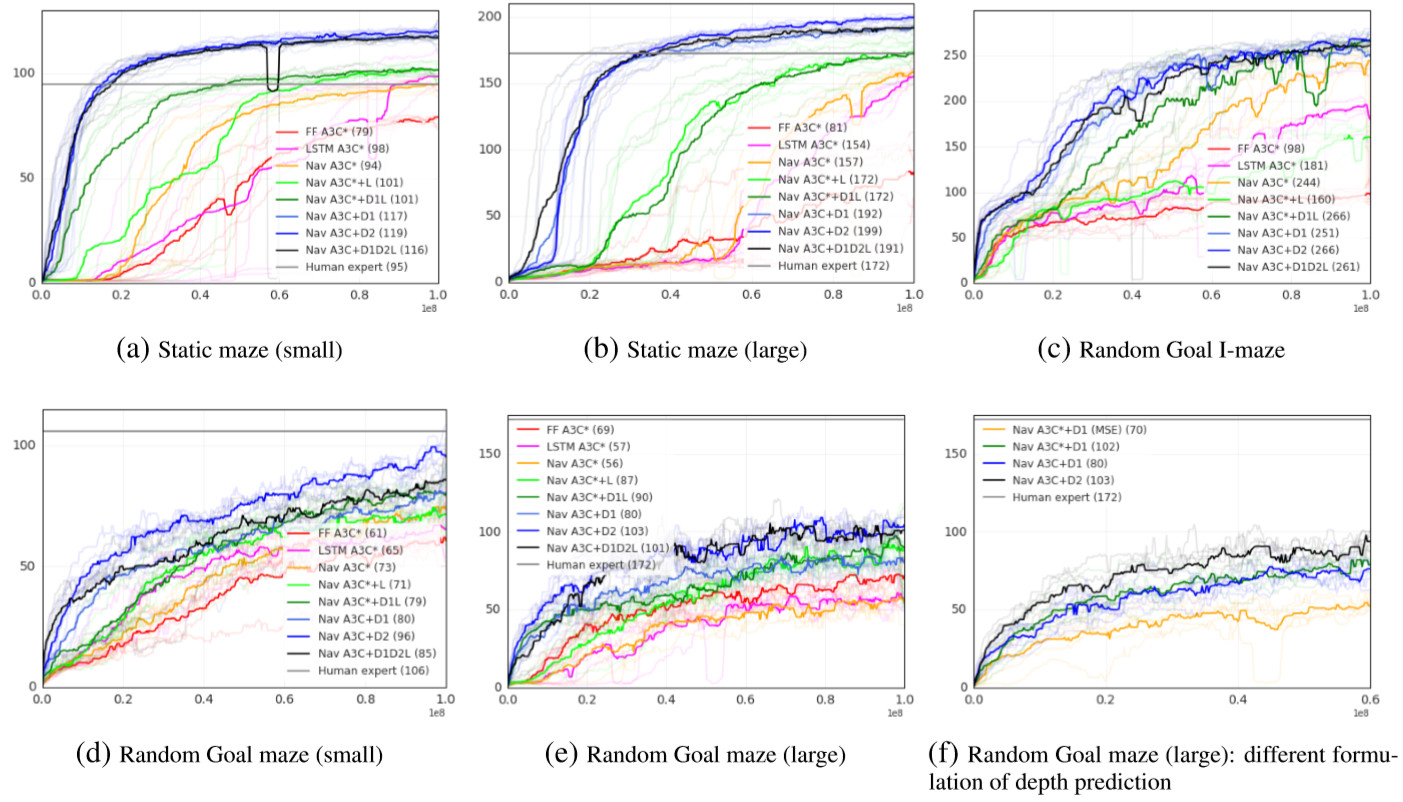


图3

第二部分所描述的不同AI结构通过在5个迷宫中进行实验并得到结果，详见图3。总结一下这些AI结构：

* 前馈A3C (FF A3C)
* 循环A3C (LSTM A3C)
* 接收奖励与动作信号并适配速度反馈的Stacked LSTM (Nav A3C)
* 附加通过卷积信号预测深度的 Nav A3C+
* 改造后从LSTM尾层信号预测深度的 Nav A3C+
* 仅添加闭环预测的 Nav A3+L
* 武装到牙齿（兼并所有功能为辅助误差）的 Nav A3C+

针对每种情况我们都进行了64次实验，每次实验均使用随机产生的超参数，这些超参数详见原文附录。最棒的5次实验恰好被最好的5个曲线图（图3中的a到e）所描述。Expert human是指用来进行结果对比的人类专业玩家，例如图3中就可以看到Nav A3C+ 在两个静态迷宫中结果由于人类玩家，并且在两个随机迷宫也能接近人类的得分。

Mnih的文章中曾使用奖励修剪来稳定学习状态，我们也想尝试这个技术。不幸的是它并不适用于我们的环境，因为它会随意缩减奖励，导致AI开始混淆苹果奖励（1分）与传送门奖励（10分）而移除这个技术后又会导致A3C算法不稳定，详见原文附录C。然而似乎来自于深度检测的附加误差能调和这个问题，让学习过程保持稳定，就是图3中的f，符号上我们使用来表示带奖励修剪的AI结构。

如果你仔细观察图3的f图，会发现其中包含了回归评价与分类评价信息，而回归评价（MSE）的得分没有分类评价的高。这样的实验结果在其他地图上是一样存在的，所以对于其他地图我们仅采用了分类评价。另一方面，从LSTM尾层信号预测深度比卷积信号结果更好的事实是显而易见的。

我们在图3中标记了一些特殊的结果。你可以在a图与b图中比较一下前馈A3C模型（红色曲线）与循环A3C模型（粉色曲线），虽然导航功能在实际上非常消耗内存，单路网络计算也存在不确定性，但前馈A3C模型在静态迷宫中依然表现的像个小强。这说明了在静态迷宫中AI像个醉鬼一样只会扶墙走也是可以的，或许真的存在这样的好策略。所以当AI身处动态迷宫时能用些记忆与导航技能就一下子让人感觉像个王者。

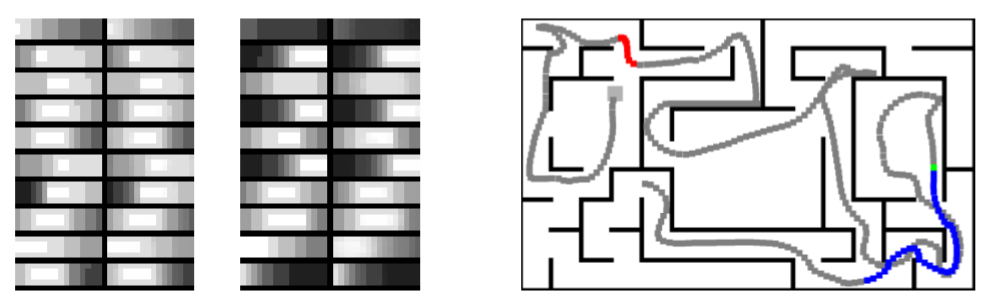


图4

输入信号选择速度、奖励与动作同样是个很有优势的高级技能，这集中体现在橙色曲线与红色、粉色的相互对比。当然训练这个Nav A3C 是个挺费劲的事，不过我们觉得速度太慢都是强化学习本身总是学习到无用数据惹得祸，毕竟论学习速度监督学习才是王道。所以当我们添加许多辅助误差后，对数据的利用就更偏向与监督优化，这样尽量过滤出有效的数据给到强化学习过程，就显著提高了学习速率，详见表1中的AUC矩阵。这样的速度提升效果在静态迷宫是最显著的，且对于动态迷宫也有持久提速的效果。

主线成绩肯定是最受关注的，而在辅助误差中我们也会记录闭环预测的结果。从图4中右图所示轨迹中我们可以看到 是如何有效检测闭环的。

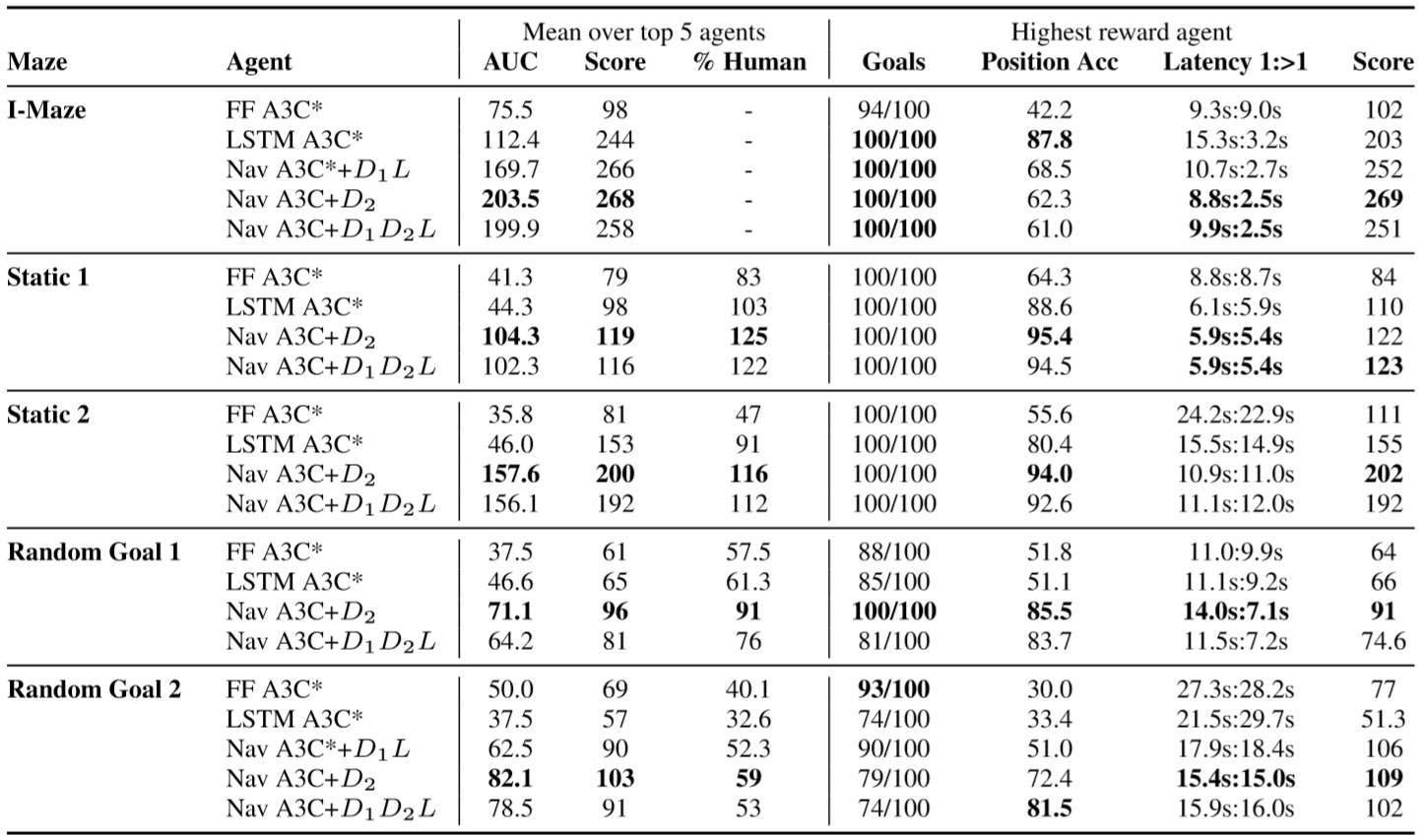


表1

**5. 分析**

***5.1 定位解码***

从LSTM或卷积提取到的定位特征都包含在其尾层，为了能够计算这种内在表示，我们训练了一个定位解码器，这个解码器以上述尾层信息作为输入，使用线性分类器输出若干离散的概率位置。针对5x10的地图解码器输出50个可能位置，针对9x15的地图则输出135个可能位置，I-maze输出77个可能位置。需要注意的是这个解码器并不参与整个神经网络的反向传播过程，它只是将信息描述出来，不进行任何学习与改变。

模型的解码器见图6，随着学习量的增加，概率位置越来越集中于正确位置。在实验中我们发现一旦AI重生时，其定位的不确定性（熵）突然增加，之后随着AI定位则会逐渐准确，具体可以参见原文附录A中的视频链接。定位的好坏会决定AI是否能快速找到目标，定位准确度与最终成绩的相关性对比在表1中有所描述。前馈AI在静态迷宫中能够达到64.3%的准确率，这就说明编码器能够在其权重中记录位置信息，也充分证明了当代AI在训练完备的前提下普遍能够解决这种小型静态迷宫问题，而动态迷宫问题则似乎更依赖特殊的AI设计。

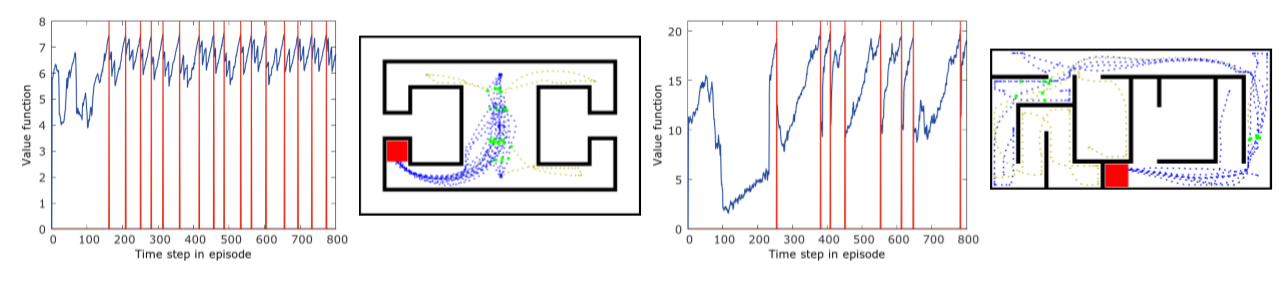


图5

在I-maze中，那些“臂弯”位置的形状非常相似，除了极少的视觉线索可以帮助区分它们。 模型非常适合这种迷宫，能够学习到再次直达目标的方法以追求最大累计得分，虽然其位置解码器的准确率仅 68.5%。而纯LSTM A3C模型在I-maze的位置解码器准确度有 87.5% 但最后成绩却不尽人意。面对这样的疑惑我们只能假设对称迷宫会导致AI学习到对称策略，而这样的策略不需要AI对其位置有很高的敏感性（参见以下分析）。

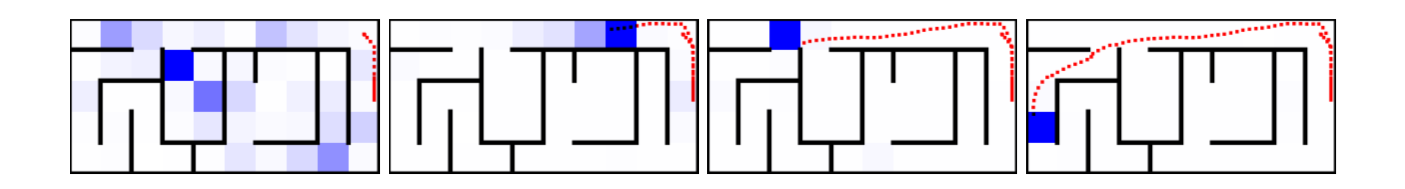
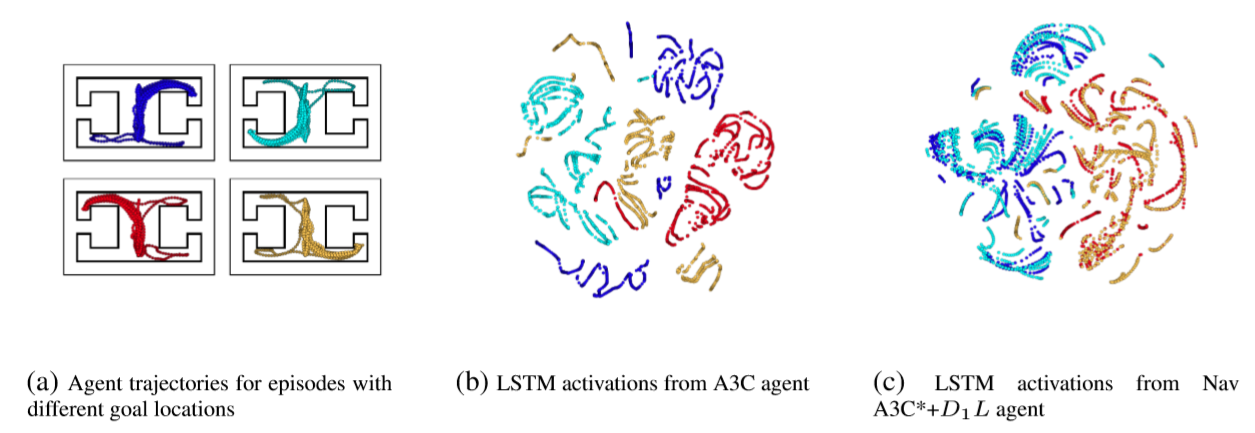


图6

随机寻物任务是希望AI能够首先找到目标，并在重生之后通过有效路线可靠返回目标。 表1中的延迟列（latency）显示 模型对于返回已经发现过的目标位置具有最低延迟，该列第一个数字显示第一次发现目标的时间，以秒为单位，第二个数字是平均值。图5清楚显示了AI如何找到目标，并在接下来的游戏中直接返回到该目标。动态地图2具有更大的挑战性，所有AI模型都无法在重新找到目标这个任务上达到低延迟。

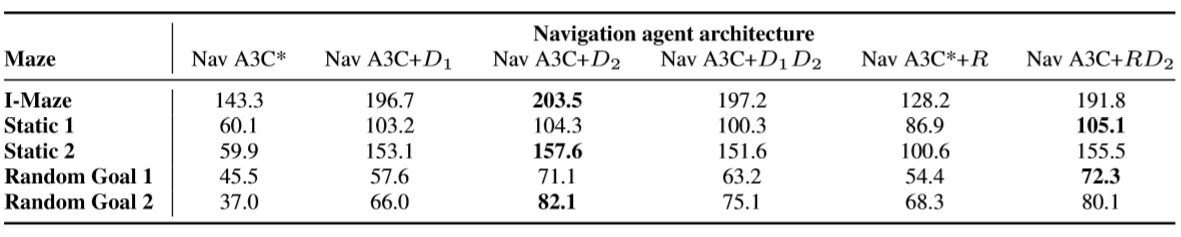
***5.2 栈式LSTM目标搜寻***

图7中的a显示了AI寻找4个目标时所经历的路线轨迹，当第一次找到目标后，较粗的轨迹集合可以证明该AI能够反复再次抵达该轨迹。图7的b与c是依靠tSNE降为方法对LSTM的激活值进行了可视化所呈现的规律性结果，可以看到不添加任何辅助误差的b无法学习到相对性迷宫的知识。当目标点在该迷宫的左下角与右上角时，AI其实具有相似的策略，例如AI不论从左下角作为起点去找右上角的目标，还是从右上角作为起点去找左下角的目标，策略上都是先左转，再右转。



***5.3 针对不同辅助任务的研究***

虽然我们的策略型LSTM非常厉害，但市面上还存在其他策略，即其他辅助任务的添加方法，对这些方法进行比较也是我们的任务所在。例如Jaderberg在2017年的文章，针对他的文章我们又实施了两种策略LSTM，一个使用回放缓存（replay buffer），所以称为 ，另一个结合了卷积网络奖励预测与LSTM景深预测，称为 。多种辅助任务策略的AUC（Area under learning curve）对比详见下方表2。



**6. 结论**

我们利用这篇文章阐述了一种新的深度强化学习方法，内置了记忆方式和辅助学习目标，使AI能够在复杂环境中进行导航。我们的结果和分析内容都强调了辅助目标在景深预测与闭环预测方面的作用，进一步提高了AI利用数据的效率。

这篇文章与Jaderberg的文章既有相似点又存在极大的差异，两篇文章都试图通过辅助目标提高端到端强化学习的能力，并渴望开发更加通用的导航策略。区别在于本文的所有损失函数都是在线的，不依赖任何回放数据。

我们期望未来会有更多工作能够使用到我们今天的成果，比如与AI的其他外部记忆相结合，又或者与SLAM进行比较，这些研究都会非常有趣。