智闯鬼屋

摘要

本文完成了智闯鬼屋任务的 A、B、C 题。其中 A、B 题利用迷宫的拓扑结构研究成果（贴墙行走的先验知识）与模糊系统中隶属度函数的内容成功让智能体在一定程度上完成对未知环境的探索与躲避威胁的任务。在未知任何环境信息的条件下，A 任务中进行 1000 轮仿真之后智能体的平均消耗时间（用仿真步数来计算，包含停顿时间）为 10.554（本迷宫最快需要消耗 10），成功率为 83%，在环境与幽灵运动轨迹完全未知的情况下较好地完成了任务。在已知幽灵行动轨迹的情况下，增加先验知识，修改停顿判别条件为仅在进入幽灵通道前，智能体 1000 轮仿真平均消耗时间为 10.14，性能显著增强。在调整隶属度函数超参数的过程中，发现了曲线上与直观设计目的不符的部分并进行了讨论，最终就讨论结果给出了 A 题的最优超参数选项。

B 任务中，运用与 A 题中同样的策略，在所有超参数设定相同的情况下，贴右墙行走 1000 轮仿真最快平均消耗时间 14.13，成功率 86.5%；最安全情况下成功率 100%，平均消耗时间 16.255。贴左墙行走 1000 轮最快平均消耗时间 14.40，成功率 72.90%；最安全情况下成功率 100%，平均消耗时间 22.20。文章在 A 题分析方法的基础上解释了 B 题超参数调整中的异常现象，并从迷宫结构和理论层面讨论了造成左右手策略不同结果的原因。

在 C 题中，考虑到迷宫环境已知，而幽灵运动情况依然并不完全已知，所以考虑利用强化学习中的策略优化方法，让智能体在已知的环境上进行 model-free 的仿真训练，获得优化的策略后再进行实地行走。为了体现强化学习的优越性，第三题中并没有提供给智能体任何先验知识，只开放了地图探索。本文中利用浅层神经网络作为策略函数，使用类比了遗传算法的“精英”策略优化的方法来促使神经网络在训练过程中自我迭代训练，以距离终点的网格距离为基础设计了启发式的奖励函数，以行进耗散与碰到鬼为基础设计了惩罚函数，最终训练结果获得稳定在 9.1 附近的平均奖励值（最好情况），且策略神经网络训练收敛。

相对于 A、B 题中硬性编程的行走策略和程序设计者以一个上帝视角提供的启发式方法，C 题中基于神经网络的强化学习方法能够给出一些对设计者极具启发性的策略，在 C 题中也会进行讨论。但是，由于本文中用到的强化学习策略对奖励的指引能力提出了较高的要求，因为时间限制，本文并没能够提出一个更好的奖励设计方法与策略网络设计方法，导致并不是每次训练都能让系统成功收敛，在之后的研究中，可以考虑继续解决这一问题。

1. 问题 A、B 模型的建立

1.1. 完全未知迷宫中的行走策略

考察迷宫的拓扑结构，在一般的单入单出且没有复合层的普通迷宫中，无论结构多复杂的迷宫最终起点和重点都能用一根线连接，进而可以总结出迷宫的“左/右手法则”，即在上述的一般迷宫中，只要一直贴着左边墙壁或者右边墙壁行走，总能来到迷宫出口。

考虑本题中的迷宫结构，显然符合上述要求的简单迷宫结构。在环境对于智能体完全未知的情况下，搜索策略是不能使用的，因而“左/右手法则”是较为稳妥且能保证成功率的方法，以此为基础知识库符合了人类在相似情况下的直观判断。

1.2. 仅能感受寒意情况下的躲避策略

在本题中，智能体仅仅能够通过离散的寒意来感知危险，在寒意可以叠加的情况下，同一种寒意可以对应多种不同情况，介于智能体对幽灵的行走路径也完全未知，因而很难通过寒意来对具体危险进行进一步推断。根据实际情况，本文提出了使用 S 型隶属度函数来将寒意转化为“僵直概率”的方法

$$\mu(x)=\begin{cases}0 & x<a\\ 2\left(\dfrac{x-a}{b-a}\right)^2 & a\le x<\dfrac{a+b}{2}\\ 1-2\left(\dfrac{x-a}{b-a}\right)^2 & \dfrac{a+b}{2}\le x<b\\ 1 & x\ge b\end{cases}$$

当寒意越高，智能体陷入僵直的概率就越大，会停留原地等待危险过去。一般而言，这是一种较为有效且符合实际情况的策略。但是如果智能体恰好停留在幽灵会经过的通道上，死亡概率就会大大增加。

此外，本文将 S 型隶属度函数中的b定义为“恐惧阈值”，这一超参数可以在仿真过程中调节（本文固定$a=0$，对应寒意为 0 的情况），恐惧阈值越低，智能体对寒意越敏感，在多数情况下有更大的概率陷入僵直。

当拥有幽灵通道位置这一先验知识时，躲避策略可以仅仅在智能体进入幽灵通道的前一步运行，这大大减少了智能体在安全地带由于寒意的辐射带来的无用僵直，且能避免智能体僵直在幽灵通道时造成的死亡，在很大程度上缩短了平均时间，本文在 A 题的解决过程中对这一先验知识对结果的作用进行了比较。

1.3. 对更多可用知识的讨论

从了解问题所有运行准则的外部观察者来看，除了上述提到的两种行走策略，还可以有更多的推理结论可以加快智能体的行进速度并提高安全性：

（1）由于幽灵遇到墙壁后一定会反向行走，所以在类似位置$(3,1)$处感受到2的寒意时智能体完全可以放心向前行进；

（2）利用特殊位置前后观察到的寒意序列，可以大致推断出幽灵的运行情况；

（3）在有墙壁的迷宫中以右上方向为指导方向可以大大缩短探索距离。

但是考虑到

（1）对于智能体来说幽灵的行动模式是未知的，它并不能知道幽灵会不会原地停留或者跨格子移动；

（2）对于智能体来说出口的位置是未知的，以右上方为指导相当于人为加入了先验知识，这与题设要求不符。

所以除了 1.1 和 1.2 中提到的探索策略之外，本文并不打算加入更多的指导知识。

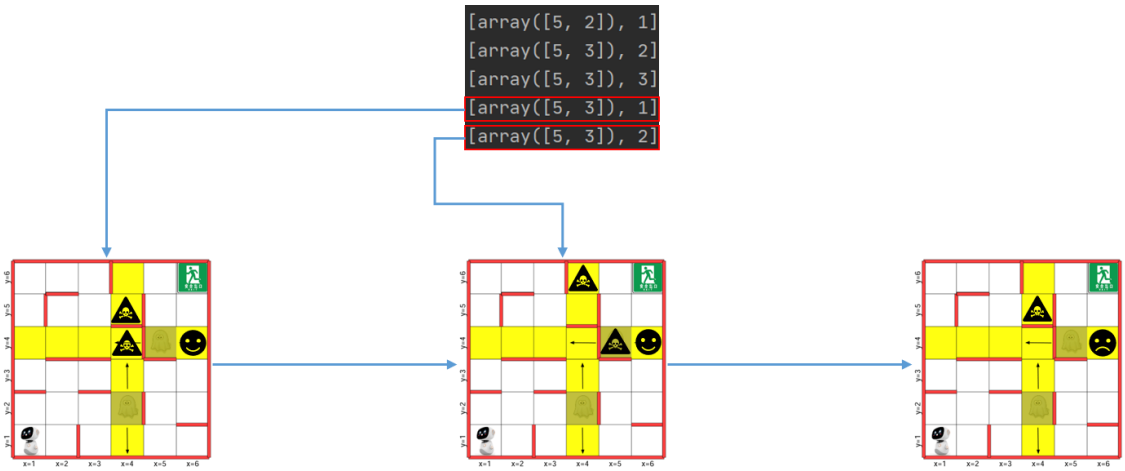
2. 问题 A、B 的求解和结果讨论

在上一章节的讨论中，已经基本确定了智能体的总体行动模式，下面进行仿真实现和结果讨论。

2.1. 问题 A 的仿真实现

在问题 A 的第一轮仿真运行过程中，智能体并不知道幽灵通道的位置，这导致它在收到寒意辐射的区域会做出无意义的僵直，浪费了很多时间。最终 1000 轮仿真结果显示，在“恐惧阈值”设定为 4 时，成功率为 89%，平均步数为 10.554，考虑到最短步数为 10，结果较为令人满意。

但是没有限制的“僵直函数”的使用有时会使智能体陷入更加危险的境地。图 1 展示了一种情况，智能体在幽灵通道上受寒意影响而僵直，导致它始终停留在幽灵通道上，最终和幽灵相遇。

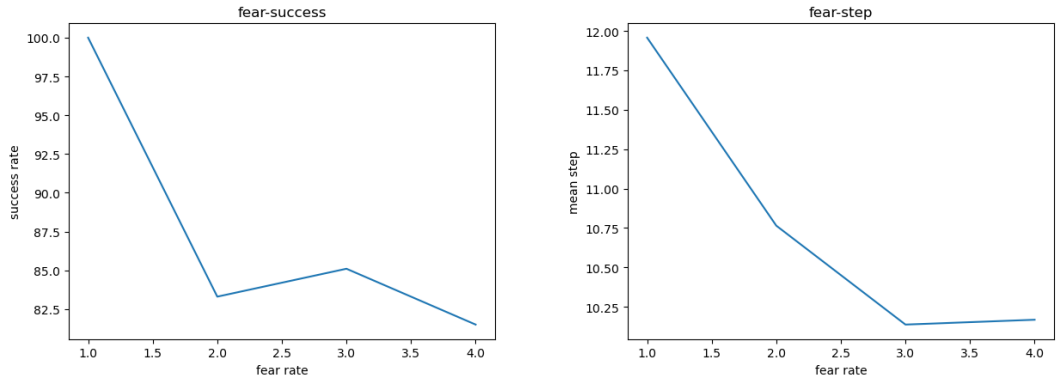


图表 1 滥用僵直函数可能会使智能体陷入更加危险的境地

在没有任何关于环境的知识的情况下，这种现象似乎难以避免。但当文章加入对幽灵通道所在位置的知识后，智能体仅仅需要在进入通道前进行僵直判断，这不仅避免了安全区域的无用僵直，也避免了上述原地“等死”的情况。在这一前提下运行 1000 轮仿真，在“恐惧”阈值依然设置为 4 的情况下，智能体的成功率提升到了 84%，平均用时减少到了 10.14。

2.2. A 题仿真环境超参数的调整与讨论

对于危险的探索任务来说，将“恐惧阈值”设定为 4 时获得的 84%的成功率显然太低了，为了寻找到更加优越的“恐惧阈值”设定，本文对范围为[1,4]的“恐惧阈值”进行了便利，获得了对应设定下的成功率与平均步数如图 2 所示。



图表 2 问题 A 恐惧阈值和成功率（左图）、恐惧阈值和平均步数（右图）的关系

与预想的单调情况不同，在“恐惧阈值”与成功率的关系图像中，在“恐惧阈值”设定为 3 时有一个不寻常的明显凸起（之后文章称之为“山峰异常”）。接下来从理论层面分析这个凸起产生的原因。

在“恐惧阈值”设定为 2 时，“僵直函数”可以具体表现为

$$\mu(x)=\begin{cases}0 & x<0\\ 2\left(\frac{x}{2}\right)^2 & 0\le x<1\\ 1-2\left(\frac{x}{2}\right)^2 & 1\le x<2\\ 1 & x\ge 2\end{cases}$$

此时当寒意为 1 时，智能体在“僵直函数”指导下跨入幽灵通道的概率为 0.5。

在“恐惧阈值”设定为 3 时，“僵直函数”可以具体表现为

$$\mu(x)=\begin{cases}0 & x<0\\ 2\left(\frac{x}{3}\right)^2 & 0\le x<1.5\\ 1-2\left(\frac{x}{3}\right)^2 & 1.5\le x<3\\ 1 & x\ge 3\end{cases}$$

此时当寒意为 1 时，智能体在“僵直函数”指导下跨入幽灵通道的概率为 0.22。

从外部视角来看，智能体最危险的情况不是在进入通道前与幽灵相邻的情况（寒意为 2），而是在进入通道前与幽灵成对角的情况（寒意为 1）。所以在“恐惧阈值”设定为 3 时，在最危险的情况下做出正确决策的概率反而比设定为 2 时要高了，这是在离散地取函数值的前提下产生的特殊情况，这暗示着在更加细化参数调整精度时有可能会有更加优化的结果，由于时间限制，本文不再进行更加细化的探索。

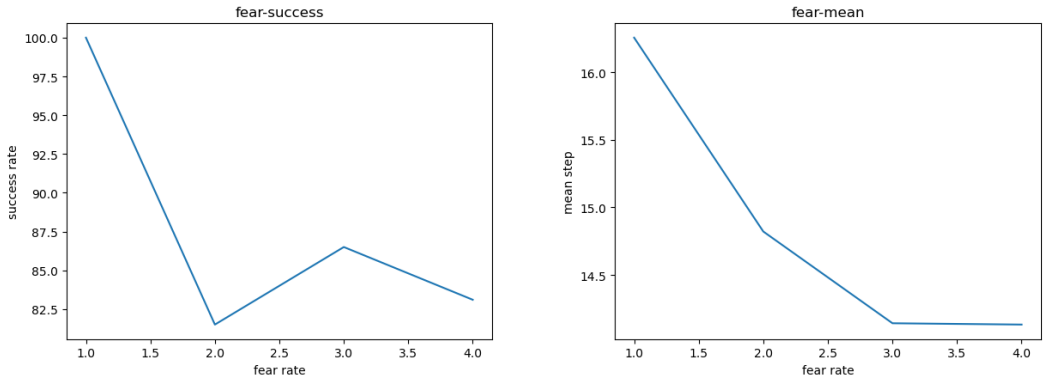
调整参数的仿真结果总体上还是呈现出一个速度与幸存率不可兼得的趋势。在寻求速度时，“恐惧阈值”设定为 3 是一个好的选择，此时可以把平均步数缩短为 10.13，成功率

提高到 85%左右；在寻求绝对安全时，“恐惧阈值”可以设定为 1，平均步数也在可以接受的 12 左右。

2.3. B 题的仿真实现

拥有内部墙面的迷宫是充分体现迷宫“左右手定则”优越性的场所。在此定则指导下的智能体能够在没有出口位置指导的情况下 100%地找到出口，由于在此迷宫中靠左和靠右行进得到的最终结果差异较大，因此本文认为这是两条不同的路线，并且分别对它们的结果进行讨论。

在靠右手墙面行走时，利用幽灵通道位置知识的智能体，在“恐惧阈值”设定为 4 的情况下，1000 轮仿真的成功率为 83%，平均步数为 14.13。此时成功率和平均步数关于“恐惧阈值”的图像如图 3 所示。

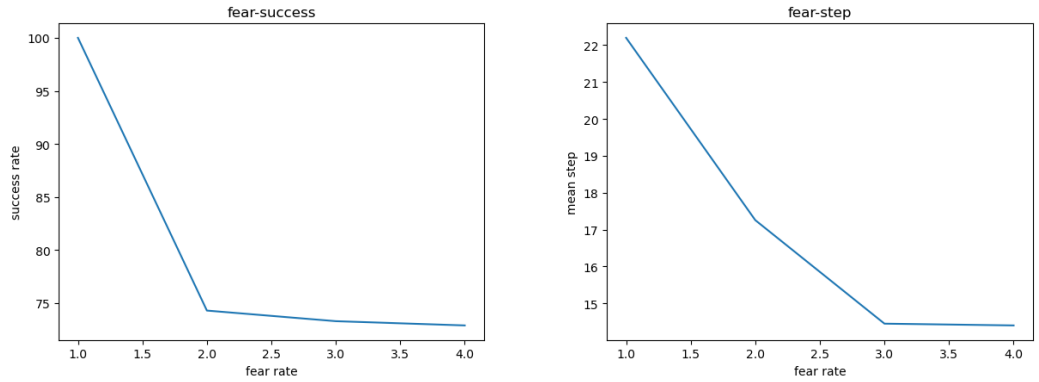


图表 3 靠右手行走时恐惧阈值和成功率（左图）、恐惧阈值和平均步数（右图）的关系

上述图像与问题 A 中的图像在形态上有着明显的相似性，这可以解释为：靠右手行走的过程中遇到的“僵直函数”作用位置（(3,2)和(6,3)）和作用情况完全相同，理论解释也与 2.2 节中完全相同。在同样优化目标不可兼得的情况下，如果更加追求通关的速度，那么在设定“恐惧阈值”为 3 的情况下，智能体可以达到成功率 87%，平均步数 14.14；当更加追求安全时，在 100%成功率保证下，平均步数也仅为 16.255。

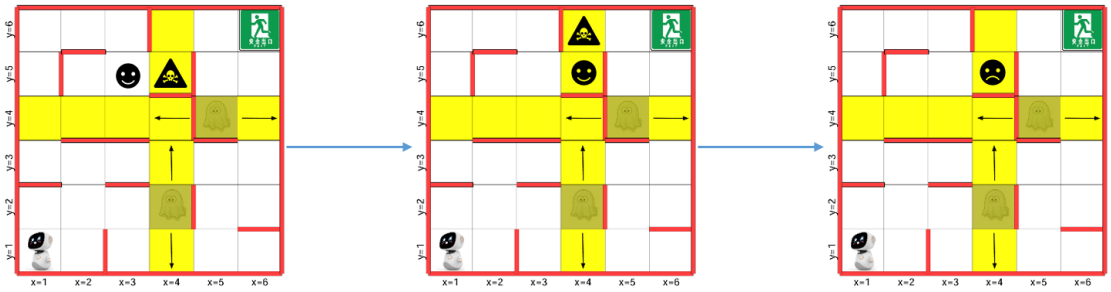
在靠左墙行走时，与题目 A 和靠右墙行走情况最大的不同在于需要沿着幽灵通道从(4,5)走向(4,6)，本文会在之后结合实际仿真结果讨论这一微小差异造成的影响；此外，由于进入幽灵通道前判别“僵直函数”的位置(3,5)与两个幽灵通道都紧邻，这导致寒意叠加的概率大大增加，使得估计结果失准，最终反映为平均步数增加和成功率大幅下降。

靠左墙行走时，如果设定“恐惧阈值”为 4，1000 轮仿真之后成功率为 73%，平均步数 14.40。理论上来说，智能体如果一路不停顿，采取靠左或者靠右策略的步数是相同的，平均步数的明显增长反映了智能体停顿时间的增加。此时成功率和平均步数关于“恐惧阈值”的图像如图 4 所示。



图表 4 靠左手行走时恐惧阈值和成功率（左图）、恐惧阈值和平均步数（右图）的关系

首先讨论沿着幽灵通道行走对曲线中山峰形状的影响。根据之前的讨论，“山峰异常”的出现是因为在选择“恐惧阈值”为 3 时可以更加有效地避免相较于邻接情况更加危险的对角情况。但是在靠左手行进的过程中，如果纵向通道中的幽灵在(4,4)位置与智能体相邻后继续向上行进，则必然会和此时进入通道的智能体相遇（过程如图 5 所示）。这种情况的出现大大削弱了邻接情况的安全性，使得“山峰异常”出现的原因被抹除。



图表 5 相邻后产生危险的情况

而寒意叠加对判断过程造成的干扰则更容易理解。寒意的叠加会使智能体对通道中幽灵的具体位置的判断造成影响，而失准的判断会造成错误的行为决策过程，最终降低成功率。

在上述的复杂条件的综合影响下，如果想平衡速度和幸存率，将“恐惧阈值”设定为 3 可以获得较大的收益，此时成功率为 73%，平均步数为 14.45。而如果要充分追求幸存率，在靠左手行走时会付出相对较大的代价（平均步数 22.20）。

综合上文对于 B 题中两条路线的讨论，智能体以 3“恐惧阈值”靠右手行走是能够最好平衡速度与幸存率的情况。但是这是经过讨论的后验知识，并不能使用于题设条件下。

3. 问题 C 的求解和结果讨论

与假定智能体十分缺乏环境知识，只能靠“随机应变”的 A、B 题不同，C 题为智能体开放了搜索地图规划路径的功能。但是值得注意的是，智能体除了固定的环境外仅仅了解了幽灵的初始状态和初始行动方向，这意味着未来的环境变化还是存在一些未知性，这就为传统搜索策略的实施造成了困难。

3.1. 基于强化学习的搜索策略

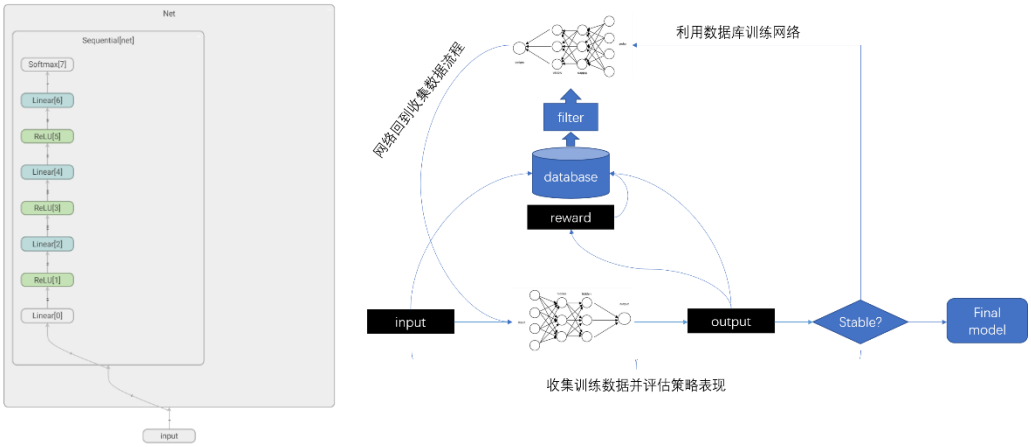
强化学习（reinforcement learning）是一种基于探索与奖励信息的机器学习方式。与传统的有监督学习和无监督学习不同，强化学习的过程不需要事先收集的数据和规则，只需

要合适的奖励信号的设计以指引智能体向更加优化的方向进化。因而对于环境尤其是动态环境的探索，强化学习方法有着先天的优越性，一个典型的例子就是 AlphaGo 在围棋博弈上的成功。

强化学习大致上拥有两大技术路线，价值优化与策略优化。考虑到动态的环境会使价值优化过程更加难以收敛，简单地利用贝尔曼最优方程和动态规划并不足以解决 C 题中依然拥有一定未知性的环境，因而本文使用基于浅层神经网络的策略优化方法来解决这一问题。

为了体现出强化学习的优越性并给予智能体更多探索和发展新策略的空间，本文在 C 题中除了寒意和当前位置和固定环境（墙体位置和出口位置）没有给智能体更多的先验知识，相较于 A、B 题仅仅开放了探索地图的功能。

在具体的智能体结构方面，本文利用一个带$Softmax$输出头的浅层神经网络（见图 6 左）作为策略执行的主体，智能体能够得到的观察值$(x, y, coldness)$作为输入，每个动作的选择概率$(up, down, left, right, stop)$作为输出。智能体的探索和策略优化（训练神经网络）的过程如图 6 右所示。这一训练过程选择每次迭代前策略能够生成的使得总奖励在前70%的观察-动作对作为神经网络的训练数据，来进行策略的下一次迭代，具体来说就是将选择过的观察数据作为输入值，相应的动作数据进行$one-hot$编码后作为标签值来对神经网络进行训练（损失函数选择均方差）。这在无形之中指导智能体不断趋向于更加优越的策略。由于这种“择优留存”的策略与遗传算法的思想十分相似，因而本文称其为“精英”策略优化方法。



图表 6 策略神经网络的结构（左图）和策略优化方法（右图）

考虑到奖励函数的设计对于强化学习智能体来说非常重要，本文接下来论述奖励函数的设计思路：

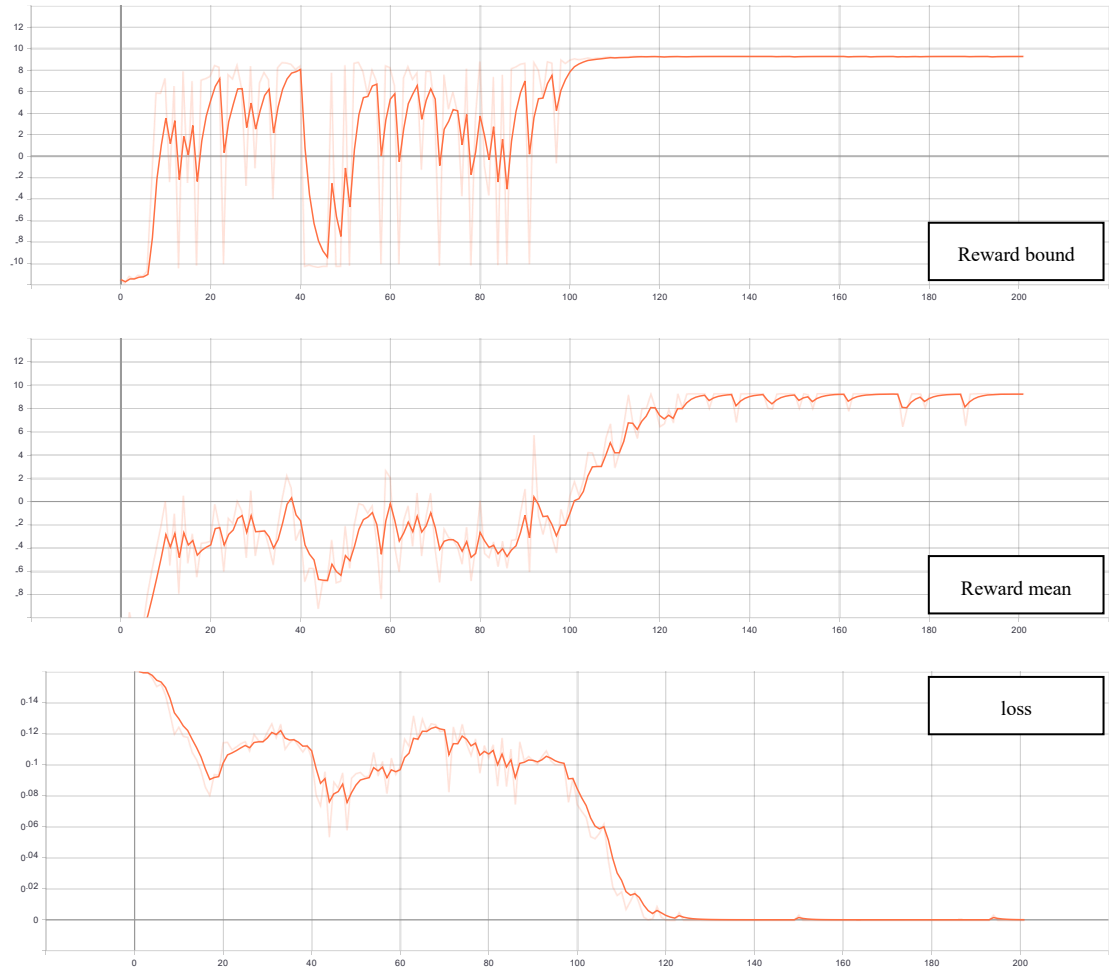
（1）每次遭遇幽灵获得-10 的惩罚值，到达出口获得 10 的奖励值；

（2）在正常行进过程中，每走一步给出-0.1 的惩罚值，以促使智能体搜索最快路径；

（3）以智能体到重点的棋盘距离为基础设计启发式的奖励函数，具体表现为$\frac{0.1}{|x-6|+|y-6|}$，即智能体每走一步能够获得的启发式奖励是到出口棋盘距离的$\frac{1}{10}$。其中设计$\frac{1}{10}$的折扣系数是因为如果系数大于0.1会导致智能体在出口周围的网格逗留时会获得每步大于0的奖励值，最终会使得智能体无限逗留在出口周围。

3.2. 基于强化学习的仿真结果

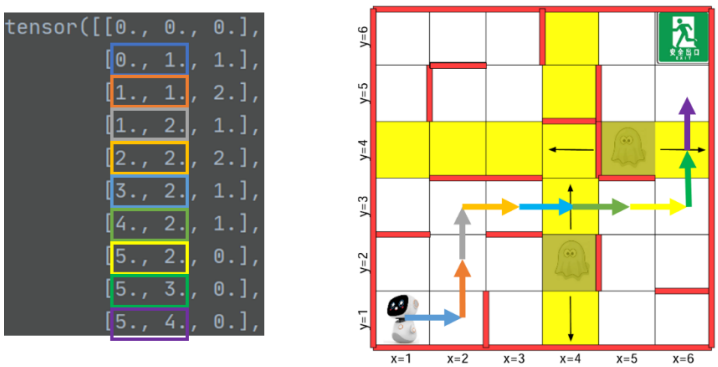
在上一节讨论得到的策略函数主体结构、策略优化方法以及奖励函数的设计思路的基础上，本文开展了实际的训练仿真。以下给出了训练过程中 70%奖励边界的变化情况、平均奖励获得情况以及神经网络损失函数变化情况。



图表 7 精英奖励值边界、平均奖励值和神经网络损失函数图像

当奖励值边界、平均奖励值和神经网络的损失函数都趋于收敛时，可以认为策略迭代已完成，智能体已经寻找到最优策略。

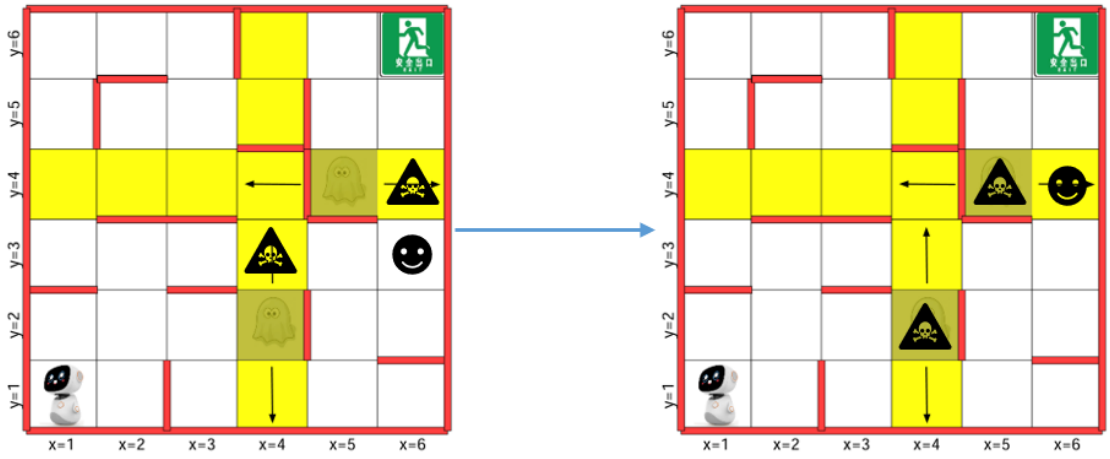
观察智能体的实际运行路线（图 8），可以看到智能体确实找到了全局最短的路径。



图表 8 最优路径演示

更进一步的地，当文章考察智能体面对危险情况的应对能力时，它自发产生的推理能力

超乎了预料。在全局视野下，我们可以知道智能体在(6,3)处如果观察到寒意为 3 的话，对应的情况是唯一的，即一个幽灵在(6,4)处，一个幽灵在(4,3)处，此时智能体可以放心前进，因为下一个时刻位于(6,4)处的幽灵恰好会回退一格，让智能体安全通过。（如图 9 所示）在未给出任何推理获得的先验知识的条件下，文章观察到智能体已经掌握了这一相对复杂的推理过程。而从智能体十分趋于理论最优值的平均奖励图像来看，智能体可能还掌握了其他高层次的推理结论使得生存率大大提高。



图表 9 智能体的推理躲避过程

3.3. 强化学习策略的结果讨论与改进方案

虽然运用强化学习策略得到的解决方案很好地解决了问题 C，但是在实际网络训练过程中，成功率并不是 100%，在一些情况下，奖励边界会持续震荡导致整个策略网络失去一个稳定的指导方向。这一点可以从图 7 的 reward bound 图像中看出，此图像前期震荡十分剧烈导致对应阶段的 mean reward 和神经网络损失函数都不能很好地收敛。这一问题出现的原因可能是智能体成功次数太少导致其不能对需要优化的方向有一个很好的认识，从而导致恶性循环，使训练最终难以收敛。这对于需要奖励指导的强化学习过程来说是十分致命的。

一般来说，导致这种现象的原因来自于奖励函数设计不善，在未来的改进过程中，可以考虑缩放奖励与惩罚函数的大小和设计更加优越的启发函数来解决这种问题。同时，这种需要猜测和尝试来解决问题的过程也是利用神经网络进行目标优化的一大弊端，较弱的可解释性和不清晰的内部结构迫使使用者只能通过外部表现来推测内部运行过程，并且在大多数情况下只能通过经验和参考来寻求解决方案。

此外由于缺少参考资料，策略网络的层数和隐藏节点的数量主要依靠经验设计，因而网络结构并不一定是最优的；在损失函数的设计方面，本文使用了均方差来计算输出各动作的概率与独热编码的标签值之间的差异，实际上在多分类问题中，交叉熵损失函数可能是一个更好的选择。