研究生学位论文进度报告

题目：xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

报 告 人：xxxx

学　　号：xxxxxxxx

专　　业：xxxxxxxxxxxxxxxx

指导老师：xxxx

导师签名：_____

学生签名：_____

1. 背景与目标

信用分配（credit assignment）即奖励分配，是强化学习领域的一个重要问题。对于单智能体强化学习，智能体可能在中间的时间步没有收到任何奖励信号，而在最终的时间步才收到奖励信号，信用分配就是要解决这种奖励延迟的问题，合理量化智能体每个时间步应得的奖励，从而鼓励智能体执行价值高的动作，避免无价值的动作。本论文旨在深入研究单智能体强化学习领域中的信用分配问题，以中国象棋这一典型的完全信息博弈为切入点，详细探讨如何有效地对中国象棋中的每个对弈步进行奖励分配，以优化智能体的学习效果。通过对比分析多种信用分配方法的特点，寻找一种切实可行且高效的奖励机制。

本文将提出一种基于深度神经网络评估的奖励机制。该机制将充分考虑当前实际对弈结果以及涉及的各个棋子的棋力，从而为每一步着法赋予恰当的价值。这一奖励机制的引入将有助于激励智能体选择那些具有更高价值的动作，从而加速智能体的学习过程。

2. 项目进展情况

当前进行奖励分配的主要方法是对奖励函数进行重塑（reward shaping）。Jose A. Arjona-Medina 等[1]提出 RUDDER 方法进行奖励重塑，该方法指出可以使用 LSTM 进行关于奖励的模式识别，从而对奖励分解后再重新分配。如图 2-1 所示，得益于 LSTM 的网络架构：遗忘门、记忆门、输出门，如果信息没变化，网络不会学到新的模式，当有新的信息，相应的模式就会被学习到。通过分析 LSTM 的“记忆”，可以重构出关键事件等信息，并根据最后预测的奖励贡献度分配给每个状态-动作配对，如图 2-2 所示。

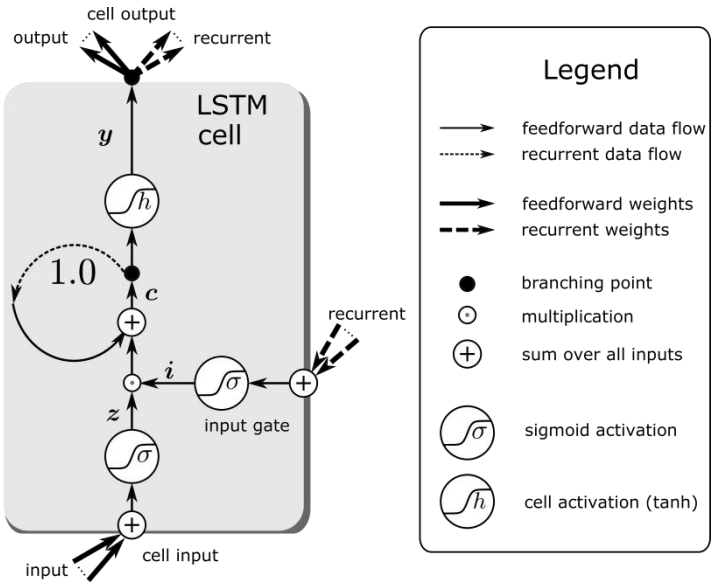

图 2-1 LSTM 单元结构[1]

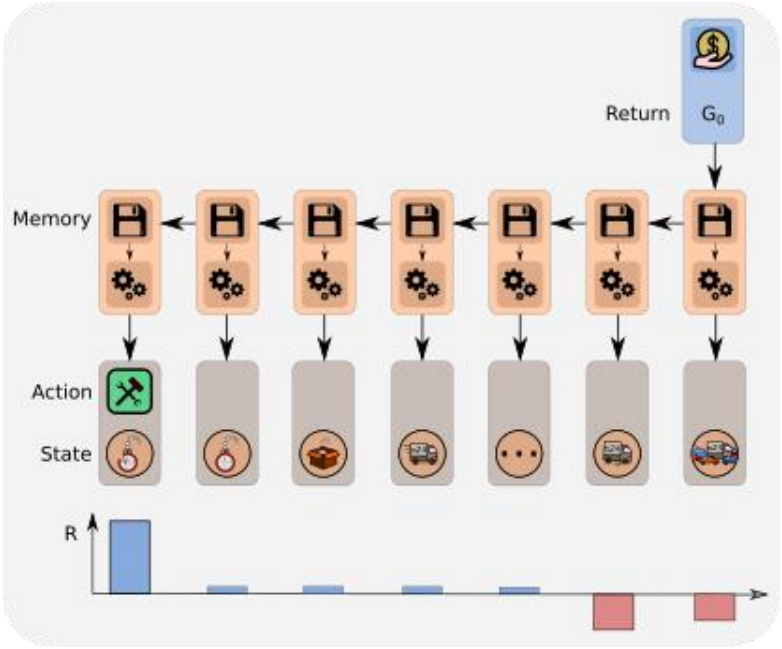

图 2-2 使用 LSTM 检测关键事件[2]

多触点归因（MTA: Multi Touch Attribution）是一种分析每个触点对最终转化效果价值贡献的方法，常应用于数字营销场景，如顾客从一开始对某种产品产生初始印象到最终购买该产品的期间，可能会涉及多个渠道（亦称为触点），

[1] 图片来自 https://ml-jku.github.io/rudder/
[2] 图片来自 https://ml-jku.github.io/rudder/

如何把“功劳”公平归因于各个渠道，是多触点归因要解决的问题。 Ning li等[2]提出了一种使用带注意力机制的深度神经网络模型DNAMTA，使用LSTM来拟合顾客所触达的广告序列路径，学习每个广告渠道的权重，并且将顾客个人静态信息通过全连接神经网络编码融入到整体的训练模型中。

受此启发，中国象棋的博弈过程可以类比于数字营销中顾客所触达的广告渠道路径：下棋者的每步落子动作可类比于顾客触达的每个广告渠道，下棋的最终输赢结果则类比于整体广告效果是否实现让顾客最终购买了该产品，从而中国象棋的信用分配问题可以借鉴多触点归因分析模型来解决。本文将使用LSTM作为主要神经网络架构来实现多触点归因分析，识别每一步落子的贡献度大小。

2.1 数据收集与整理

通过与象棋程序进行自动博弈来收集博弈数据。数据经过去除重复值、去除空值和异常值等清洗处理后，基于原始数据组合新的数据特征。数据字段说明如表2-1所示。

表2-1 博弈数据字段说明

字段名	说明
match_id	博弈场次id，标识每一场博弈
round_id	轮次id，标识每一场博弈中每一轮对战
state	当前棋局状态
action	落子动作
next_state	下一个棋局状态
done	当前博弈局是否结束：0表示未终局，1表示已终局
chapture_reward	吃子奖励：正数为红方吃子，负数表示黑方吃子
win	胜/负/和局标识：1表示胜，-1表示负，0表示和局
converted	胜或负为1，和局为0
jid	每次落子的唯一标识

其中state所表示棋局状态遵循中国象棋电脑应用规范所示使用FEN格式串表示

[3]，action 所示落子动作遵循中国象棋电脑应用规范所示着法表示[4]。详细数据样例可参考图 2-3

match_id	round_id	state	action	next_state	done	chaptures	pture_re	win	converted	jid
0-0	1692660737371	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	f0e1	RNBAK1BNR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-0_1692660737371
0-0	1692660737371	RNBAK1BcR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	h2g2	RNBAK1BcR/4A4/1C4C2/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	1	K	-4.92	-1	1	0-0_1692660737371
0-1	1692660737372	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	f0e1	RNBAK1BNR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-1_1692660737372
0-1	1692660737372	RNBAK1BcR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	h2g2	RNBAK1BcR/4A4/1C4C2/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	1	K	-4.92	-1	1	0-1_1692660737372
0-2	1692660737373	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	f0e1	RNBAK1BNR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-2_1692660737373
0-2	1692660737373	RNBAK1BcR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	h2g2	RNBAK1BcR/4A4/1C4C2/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	1	K	-4.92	-1	1	0-2_1692660737373
0-3	1692660737374	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	f0e1	RNBAK1BNR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-3_1692660737374
0-3	1692660737374	RNBAK1BcR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	h2g2	RNBAK1BcR/4A4/1C4C2/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	1	K	-4.92	-1	1	0-3_1692660737374
0-4	1692660737375	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	e0e1	1NBAKABNR/BB/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-4_1692660737375
0-4	1692660737375	1cBAKABNR/BB/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/7c1/9/rnbakabnr	b2b4	1cBAKABNR/BB/7C1/P1P1P1P1P/1C7/9/p1p1p1p1p/7c1/9/rnbakabnr	0		0	0	0	0-4_1692660737751
0-4	1692660737375	1cBAKABNR/BB/7C1/P1P1P1P1P/1C7/9/p1p1p1p1p/6nc1/9/rnbakablr	e3e4	1cBAKABNR/BB/7C1/P1P3P1P/1C2P4/9/p1p1p1p1p/6nc1/9/rnbakablr	0		0	0	0	0-4_1692660737752
0-4	1692660737375	1cBAKABNR/BB/7C1/P1P3P1P/1C2P4/9/p1p1p1p1p/2n3nc1/9/rlbakablr	h2h4	1cBAKABNR/BB/9/P1P3P1P/1C2P2C1/9/p1p1p1p1p/2n3nc1/9/rlbakablr	0		0	0	0	0-4_1692660737753
0-4	1692660737375	1cBAKABNR/BB/9/P1P3P1P/1C2P2C1/9/p1p1p1p1p/2n3nc1/9/lrbakablr	b4b6	1cBAKABNR/BB/9/P1P3P1P/4P2C1/9/pCp1p1p1p/2n3nc1/9/rbakablr	0	C	-25.41	0	0	0-4_1692660737754
0-4	1692660737375	1cBAKABNR/BB/9/P1P3P1P/4P2C1/9/prp1p1p1p/2n3nc1/9/2bakablr	b4f4	1cBAKABNR/BB/9/P1P3P1P/4PC3/9/prp1p1p1p/2n3nc1/9/2bakablr	0		0	0	0	0-4_1692660737755
0-4	1692660737375	1cBAKABNR/BB/9/P1P3P1P/4PC3/9/prp1p1p1p/2n3nc1/9/2bakabr1	i3i4	1cBAKABNR/BB/9/P1P3P2/4PC2P/9/prp1p1p1p/2n3nc1/9/2bakabr1	0		0	0	0	0-4_1692660737756
0-4	1692660737375	1cBAKABNR/BB/9/P1P3P2/4PC2P/6p2/prp1p3p/2n3nc1/9/2bakabr1	a1f1	1cBAKABNR/5R3/9/P1P3P2/4PC2P/6p2/prp1p3p/2n3nc1/9/2bakabr1	0		0	0	0	0-4_1692660737757
0-4	1692660737375	1cBAKABNR/5R3/9/P1P3P2/4PC2P/5np2/prp1p3p/2n4c1/9/2bakabr1	c0e2	1c1AKABNR/5R3/4B4/P1P3P2/4PC2P/5np2/prp1p3p/2n4c1/9/2bakabr1	1	K	-4.92	-1	1	0-4_1692660737758
0-5	1692660737376	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	f0e1	RNBAK1BNR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-5_1692660737376
0-5	1692660737376	RNBAK1BcR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	h2g2	RNBAK1BcR/4A4/1C4C2/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	1	K	-4.92	-1	1	0-5_1692660737376
0-6	1692660737377	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	f0e1	RNBAK1BNR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-6_1692660737377
0-6	1692660737377	RNBAK1BcR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	h2g2	RNBAK1BcR/4A4/1C4C2/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	1	K	-4.92	-1	1	0-6_1692660737771
0-7	1692660737378	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	f0e1	RNBAK1BNR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-7_1692660737378
0-7	1692660737378	RNBAK1BcR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	h2g2	RNBAK1BcR/4A4/1C4C2/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	1	K	-4.92	-1	1	0-7_1692660737781
0-8	1692660737379	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	f0e1	RNBAK1BNR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-8_1692660737379
0-8	1692660737379	RNBAK1BcR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	h2g2	RNBAK1BcR/4A4/1C4C2/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	1	K	-4.92	-1	1	0-8_1692660737791
0-9	1692660737380	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	f0e1	RNBAK1BNR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-9_1692660737380
0-9	1692660737380	RNBAK1BcR/4A4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	h2g2	RNBAK1BcR/4A4/1C4C2/P1P1P1P1P/9/9/p1p1p1p1p/1c7/9/rnbakabnr	1	K	-4.92	-1	1	0-9_1692660737801
0-10	1692660737381	RNBAKABNR/9/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	e0e1	RNBA1ABNR/4E4/1C5C1/P1P1P1P1P/9/9/p1p1p1p1p/1c5c1/9/rnbakabnr	0	N	-6.28	0	0	0-10_1692660737381

图 2-3 博弈数据样例

原始数据经过清洗后，由于 state 棋局状态特征为字符串型，现使用 Sentence-BERT[3]的方法将其编码为数值型向量，而 action 落子动作特征则编码为 one-hot 向量。最终将每个样本特征拼接为 2871 维的向量，作为训练数据。当前已收集了 11082 条博弈记录（仅红方），根据 match_id 进行分组，形成形如[batch, step, feature]的训练样本 2462 条，测试样本 308 条，验证样本为 308 条。

2.2 模型开发与训练

本论文借鉴 DNAMTA[2]的神经网络归因模型，使用 LSTM 作为模型的主要架构（图 2-4），并通过注意力层学习奖励分配权重。训练过程如下：

（1）使用 6 层 LSTM 对博弈特征数据特征进行编码，分别输出并保留每一步的编码结果；

（2）每一步的编码结果输入到注意力网络层，学习并输出注意力权重;

（3）注意力权重与 LSTM 的每一步编码结果进行加权求和，最后通过一个全连接层输出预测结果（converted=1 或 converted=0）。

[3] 中国象棋电脑应用规范（三）——FEN 文件格式：https://www.xqbase.com/protocol/cchess_fen.htm
[4] 中国象棋电脑应用规范（二）——着法表示：https://www.xqbase.com/protocol/cchess_move.htm

```
LstmAttentionModel(
  (lstm1): LSTM(2871, 256, batch_first=True, dropout=0.05)
  (lstm2): LSTM(256, 256, batch_first=True, dropout=0.05)
  (lstm3): LSTM(256, 256, batch_first=True, dropout=0.05)
  (lstm4): LSTM(256, 256, batch_first=True, dropout=0.05)
  (lstm5): LSTM(256, 256, batch_first=True, dropout=0.05)
  (lstm6): LSTM(256, 256, batch_first=True, dropout=0.05)
  (attention): AttentionLayer()
  (output_layer): Linear(in_features=256, out_features=1, bias=True)
)
```

图 2-4 本论文进行奖励分配的模型架构

2.3 模型下游应用

模型下游应用主要是奖励分配——分配适当的奖励到每个状态-动作配对上。从训练完的模型提取注意力层的权重，基于历史博弈数据可统计出每个落子动作对应的奖励贡献权重，如算法 1 所示。

算法 1 计算奖励权重比例

输入: $AttentionWeight$ 模型注意力层权重，$Xdata$ 博弈数据，$actionCount$ 落子动作的数量

输出: 奖励权重比例

function GETACTIONINDEX($XStep, actionCount$)
  $startIndex \leftarrow 768$
  $endIndex \leftarrow startIndex + actionCount - 1$
  return $\arg\max(XStep[startIndex : endIndex])$
end function

function GETREWARDWEIGHT($AttentionWeight, XData$)
  $attributions \leftarrow [0, 0, 0, ..., 0]$
  $actionsFreq \leftarrow [0, 0, 0, ..., 0]$
  $stepCount \leftarrow len(AttentionWeight)$
  for $i = 0 \rightarrow stepCount$ **do**
    $stepList \leftarrow AttentionWeight[i]$
    $contributionCount \leftarrow len(stepList)$
    for $j = 0 \rightarrow contributionCount$ **do**
      $actionIndex \leftarrow$ GETACTIONINDEX($XData[i][j], actionCount$)
      $stepContribution \leftarrow stepList[j]$
      $attributions[actionIndex] \leftarrow attributions[actionIndex] + stepContribution$
      $actionFreq[actionIndex] \leftarrow actionFreq[actionIndex] + 1$
    end for
  end for
  $result \leftarrow attributions/actionsFreq$
  return $result$
end function

最后基于算式（2-1）进行奖励分配。

$$reward_{action_i} = \frac{weight_{action_i}}{\sum_{k=1}^{N} weight_{action_k}} \times reward_{total} \quad \text{（算式 2-1）}$$

其中$weight_{action_i}$为算法 1 的计算结果。

3. 当前成果

（1）已完成了神经网络模型训练，并已验证训练收敛，如图 3-1 所示。

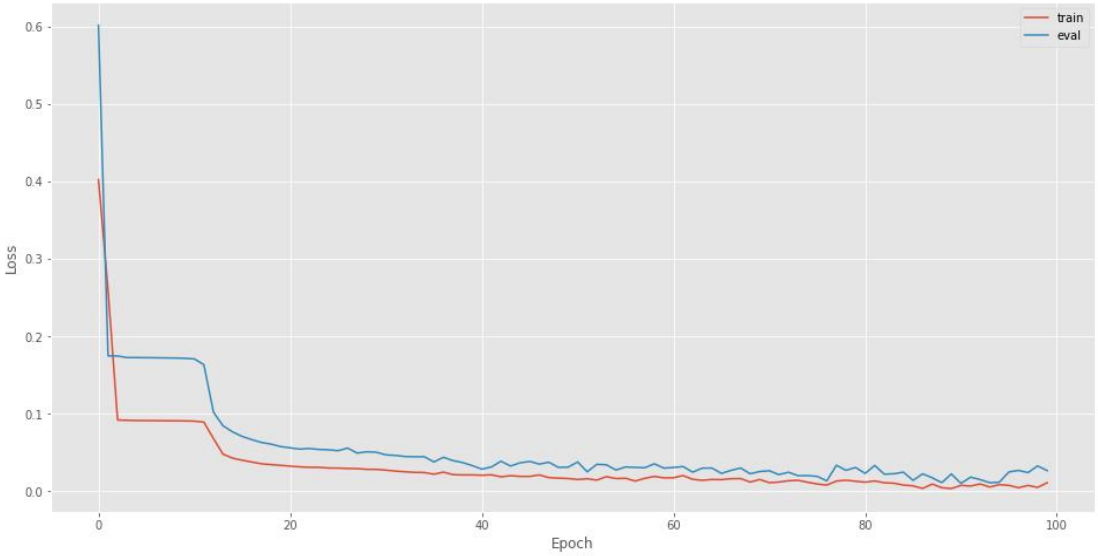

图 3-1 模型训练过程

（2）已输出落子动作对应的奖励权重，如图 3-2 所示，横坐标为落子动作的 ID，纵坐标为动作对应的贡献度。

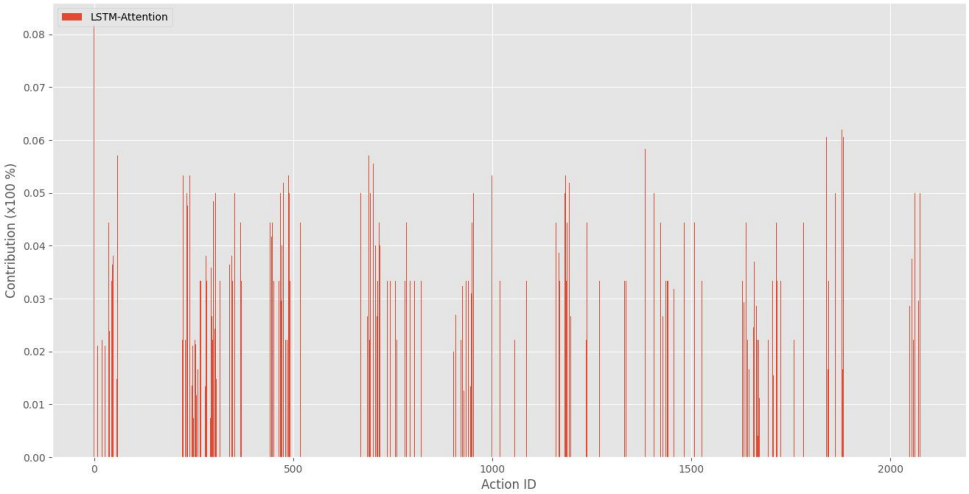

图 3-2 落子动作对应奖励权重的分布

（3）已将本论文的奖励分配方式与 AlphaGo Zero[4]的奖励分配方式（即平均分配方式）进行对比，如图 3-3、图 3-4 所示，作为后续分析使用。

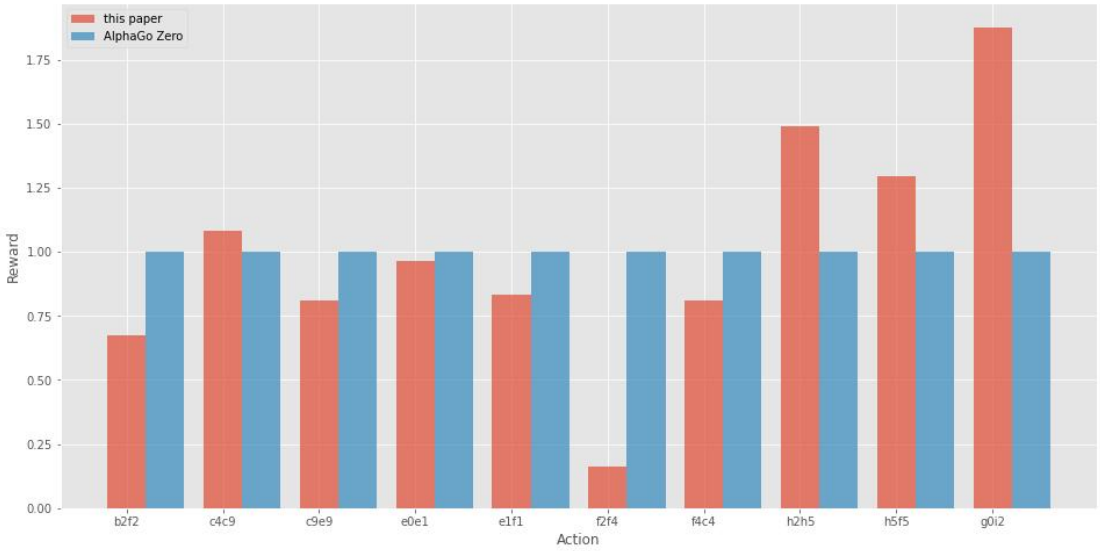

图 3-3 橙色是本论文奖励重新分配结果，蓝色是赢棋时 AlphaGo Zero 的分配方式（所有落子动作都赋值奖励 1）

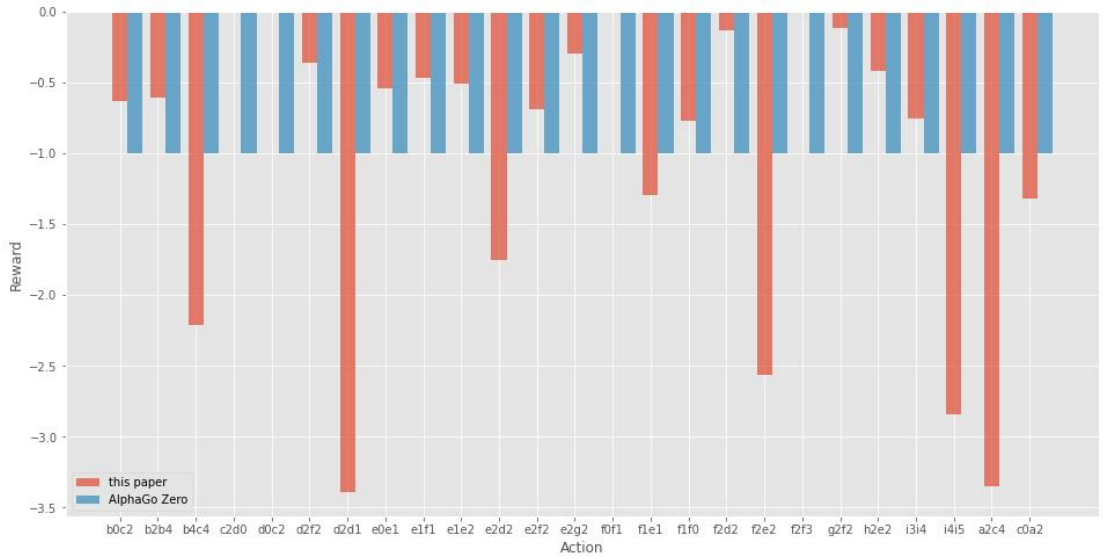

图 3-4 橙色是本论文奖励重分配结果，蓝色是输棋时 AlphaGo Zero 的分配方式（所有落子动作都赋值奖励-1）

4. 遗留问题与解决方案

（1）本论文所提出的奖励分配方式还缺少充分的对比验证。解决方案：增加多种奖励分配方式进行对比验证。

（2）关于带有注意力机制的 LSTM 如何识别关键落子动作，以及与其他模型架构相比本论文所展现的架构有何优势等问题，还缺少科学论证与事实说明。解决方案：将加入其他模型架构进行结果对比，并通过消融实验，说明本论文解决方案的科学性。

（3）暂缺本论文所提解决方案对于强化学习效率提升的解释性说明。解决方案：本模型奖励分配架构与象棋博弈系统还有待集成。

5. 下阶段计划

针对以上遗留问题，在接下来的研究阶段将进一步优化奖励分配的解决方案。具体计划如下：

（1）对比验证奖励分配方式

针对奖励分配方式缺乏充分对比验证的问题，将进行更广泛的对比实验，以评估本论文提出的奖励分配方式的有效性。除了与平均分配方式的对比外，还将考虑包括但不限于以下方式进行对比验证：

- 基于最后触点的分配：将尝试奖励主要集中在每个落子序列的最后一个动作，以验证这种方法是否能够提高性能。
- 基于时间衰减的分配：将尝试根据时间衰减策略分配奖励，逐渐降低早期动作的奖励权重，以探索这种方式是否能够更好地平衡长期和短期奖励。
- 基于 Shapley Value 的分配：Shapley Value 是在合作博弈环境中所提出来的概念，它的核心思想是衡量每个参与者对博弈结果的贡献，然后根据其贡献度进行分配奖励。对于中国象棋博弈，可以将每个落子动作类比于每个合作博弈的参与者，于是可以通过计算每个落子动作的 Shapley Value 来确定奖励分配。

（2）探索关键落子动作识别与模型比较

为了解决带有注意力机制的 LSTM 如何识别关键落子动作的问题，将采取以下步骤：

- 引入其他模型架构：将引入其他带有注意力机制的模型，例如 Transformer 等，与本论文架构进行比较，以揭示不同架构之间的优势和劣势。
- 消融实验：将对本论文提出的架构进行消融实验，逐步去除注意力机制等关键组件，以验证这些组件对于模型性能的贡献。

（3）提升强化学习效率的可解释性

为了更好地解释本论文奖励分配架构如何提升强化学习效率，将采取以下措

施：

- 集成到象棋博弈系统：将本论文的奖励分配架构应用于象棋博弈系统，通过在实际场景中的应用来验证其效果，同时从实际案例中提取具体的效率提升情况和原因。
- 解释性分析：将进行对模型决策的解释性分析，例如通过可视化注意力权重等方式，揭示模型在不同决策点上的关注重点，从而增加模型可解释性。

基于以上计划，将继续完成课题研究与论文撰写，下阶段具体安排如下：

（1）2023 年 11 月——2023 年 12 月，完成论文初稿。

（2）2023 年 12 月——2024 年 02 月，修改论文，完成第二稿。

（3）2024 年 03 月，定稿。

（4）2024 年 04 月——2024 年 05 月，整理数据集、相关代码和相关文档，准备答辩。

参考文献

[1] J. A. Arjona-Medina, M. Gillhofer, M. Widrich, T. Unterthiner, J. Brandstetter, and S. Hochreiter, ‘RUDDER: Return Decomposition for Delayed Rewards ’ . arXiv, Sep. 10, 2019. doi: 10.48550/arXiv.1806.07857.

[2] N. li, S. K. Arava, C. Dong, Z. Yan, and A. Pani, ‘Deep Neural Net with Attention for Multi-channel Multi-touch Attribution’. arXiv, Sep. 06, 2018. Accessed: Aug. 23, 2023. [Online]. Available: http://arxiv.org/abs/1809.02230

[3] N. Reimers and I. Gurevych, ‘Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks ’ . arXiv, Aug. 27, 2019. doi: 10.48550/arXiv.1908.10084.

[4] D. Silver *et al.*, ‘Mastering the game of Go without human knowledge’, *Nature*, vol. 550, no. 7676, pp. 354 - 359, Oct. 2017, doi: 10.1038/nature24270.