斗地主 RL 模型

任务介绍

斗地主是一种扑克游戏。游戏最少由 3 个玩家进行，用一副 54 张牌（连鬼牌），其中一方为地主，其余两家为另一方，双方对战，先出完牌的一方获胜。

潜在问题与解决思路

潜在问题

1.行动空间庞大，举个例子飞机带几个单张，有超多种可能。需要进行动作空间的搜索和简化、采样。

Environment	InfoSet Number	Avg. InfoSet Size	Action Size
Blackjack	10^3	10^1	10^0
Leduc Hold'em	10^2	10^2	10^0
Limit Texas Hold'em	10^{14}	10^3	10^0
Dou Dizhu	$10^{53} \sim 10^{83}$	10^{23}	10^4
Mahjong	10^{121}	10^{48}	10^2
No-limit Texas Hold'em	10^{162}	10^3	10^4
UNO	10^{163}	10^{10}	10^1
Sheng Ji	$10^{157} \sim 10^{165}$	10^{61}	10^{11}

表 1：RLCard 中牌类复杂度估计

2.动作的价值估计。首先，在不同对局、不同时段中，同样的牌组（合理的牌组合，如单张、对子、三带一等）表现的信息是不同的。其次，每一次行动并不只是比较可选动作集而已，还要考虑每次出牌后的牌组集合价值，与对家的关系、与队友的关系（农民）等整体状态、策略。

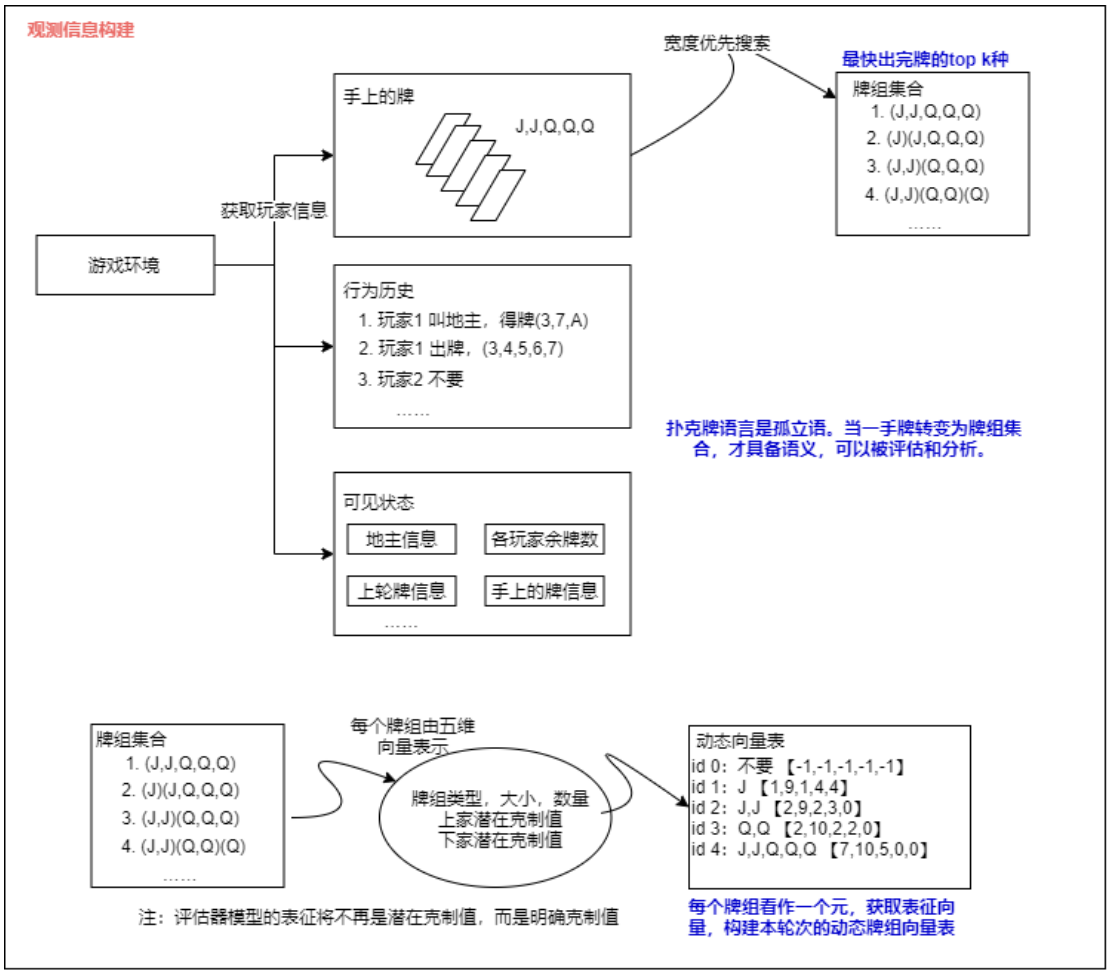
3.不完全信息博弈。手上的牌有时非常好，但是不可见的是，已被对手完全的克制，实际上这一局完全没有胜利的希望，这样模型学习起来往往是无益的。

4.不同阶段不同身份，存在不同的分析决策方式。我的牌很好，但是叫地主？也许会破坏牌！牌很差，但是差几张就会顺？搏一搏，也许胜率更高！地主独自战斗，农民之间又讲究协作。

5.人格揣测。根据每一位选手的出牌规律、习惯，猜测 ta 手上的余牌—>虚张声势？or 确有其事？or 声东击西…从对家的出牌中学习知识，也要结合自身的牌，理性思考。

6.试验环境与数据。不同的玩家思维/策略不同，模型的分析能力未必适用。

解决思路



你手上的牌在说话

对我而言，斗地主有点像孤立语。在中文，一句话也许会表达出很多种意思，比如：水有限，把它留给晚上来的人。不同的断句，“晚（一些） 上来”、“晚上（时间段） 来”将表达出不同的含义。

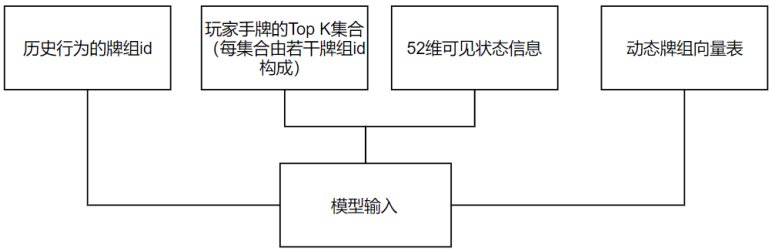
斗地主中，一手牌按照不同的组合方式，其表达的含义是不一致的。在这种情况下，一手牌必须按符合规则的方式，变化为可执行的牌组（合理的牌组合，如单张、对子、三带一等）的集合才具备意义，也才拥有可以评估和分析的基础。

宽度优先搜索

由于一手牌的排列组合方式，实在是太多太多，如果一一列举将导致过度庞大的空间。但万幸的是，越小的牌组集合，往往胜率越大。或许正因为此，斗地主又称跑得快。

> *“假如手中有 20 张牌，只靠单张去出，集合里有 20 个牌组；但是采用飞机、顺子、炸弹等方式，可能集合里就只有 1 到 3 个牌组。而且越复杂的牌组，对手要得起的可能性越小。”*

在这种情况下，采用宽度优先搜索的方式去寻找小的牌组集合，以达到缩小复杂度的作用。经分析，寻找到最小的 Top 100 牌组集合，基本就能囊括一手牌的【有价值牌组集合】。至于往后的牌组集合，常常是不会被应用的无意义集合。

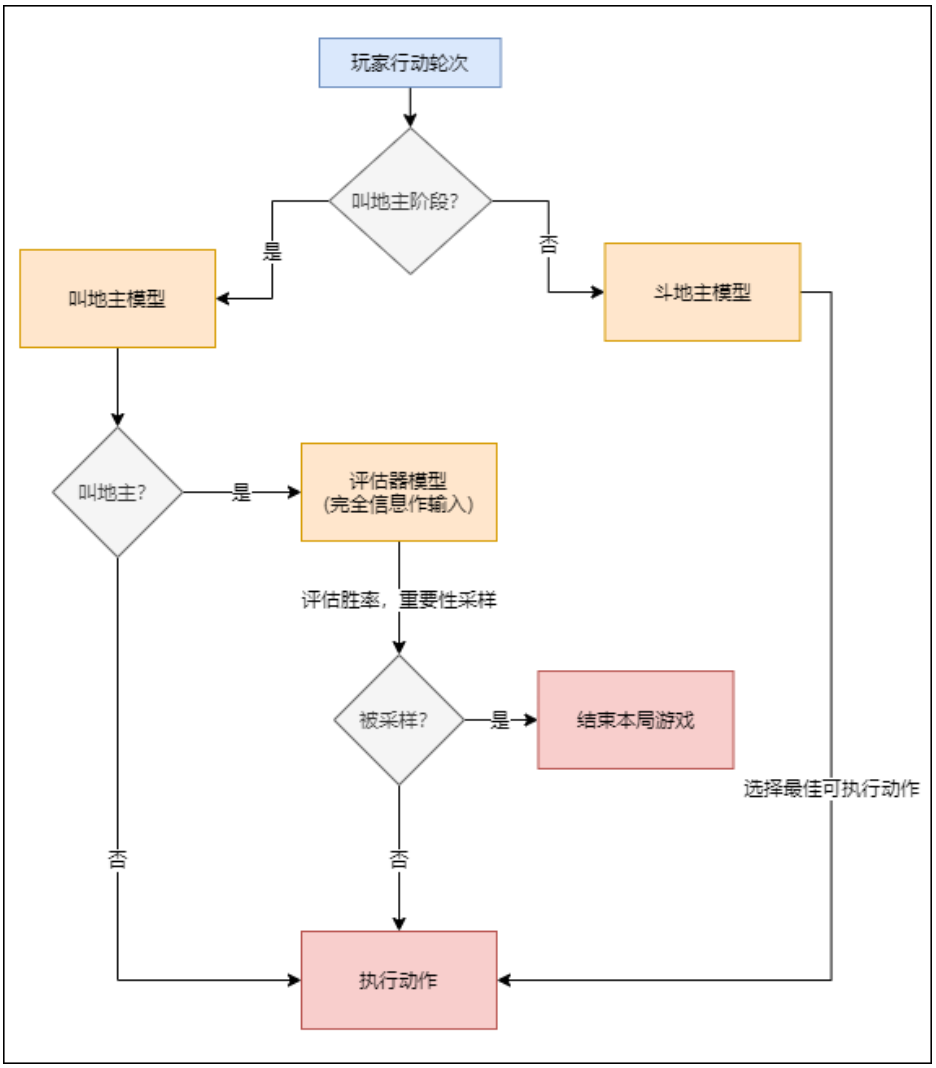


多模型调度

斗地主游戏中存在两个阶段：叫地主阶段、出牌阶段。其中出牌阶段需要区分有牌权和无牌权状况（无牌权情况要接得住上轮出牌）。因此有必要针对不同阶段、状况来做相应的模型。

重要性采样（利用完全信息加速模型训练）

手上的牌有时非常好，但是不可见的是，已被对手完全的克制，实际上这一局完全没有胜利的希望，往往是无益于模型的学习，对局便是浪费资源。在对局开始前，采用完全信息模型（可以看到所有玩家的牌），判断每位玩家的胜率，执行重要性采样。如果地主和农民旗鼓相当，倾向于执行和采样该对局；反之，当他们判若天渊，则倾向于跳过无意义对局。



模型架构

三大模型

斗地主 ai 一般由两个模型组成：叫地主模型和斗地主模型。其中，叫地主模型用以在开局时判断是否要抢地主，决定参赛选手身份，而斗地主模型则是为团队（农民或地主）尽快出完牌作决策。在本系统提出了减小不完全信息影响、提高训练效率的评估器模型，以对游戏进行重要性采样。

建模理论

假设一：一手牌的强度等于其构成的所有牌组集合的强度乘以被采用概率

$$\mathrm{Q} = \sum_{k=0}^{n} \mathrm{p}_k * \mathrm{q}_k$$

是否叫地主？这局游戏是否胜负明显？出牌后，是否能带来更好的 state？无论是叫地主模型、评估器模型还是斗地主模型的 critic 部分，其本质都是在对比玩家间牌的相对强度，因此提出“假设一”用以解决牌强度的计算问题。

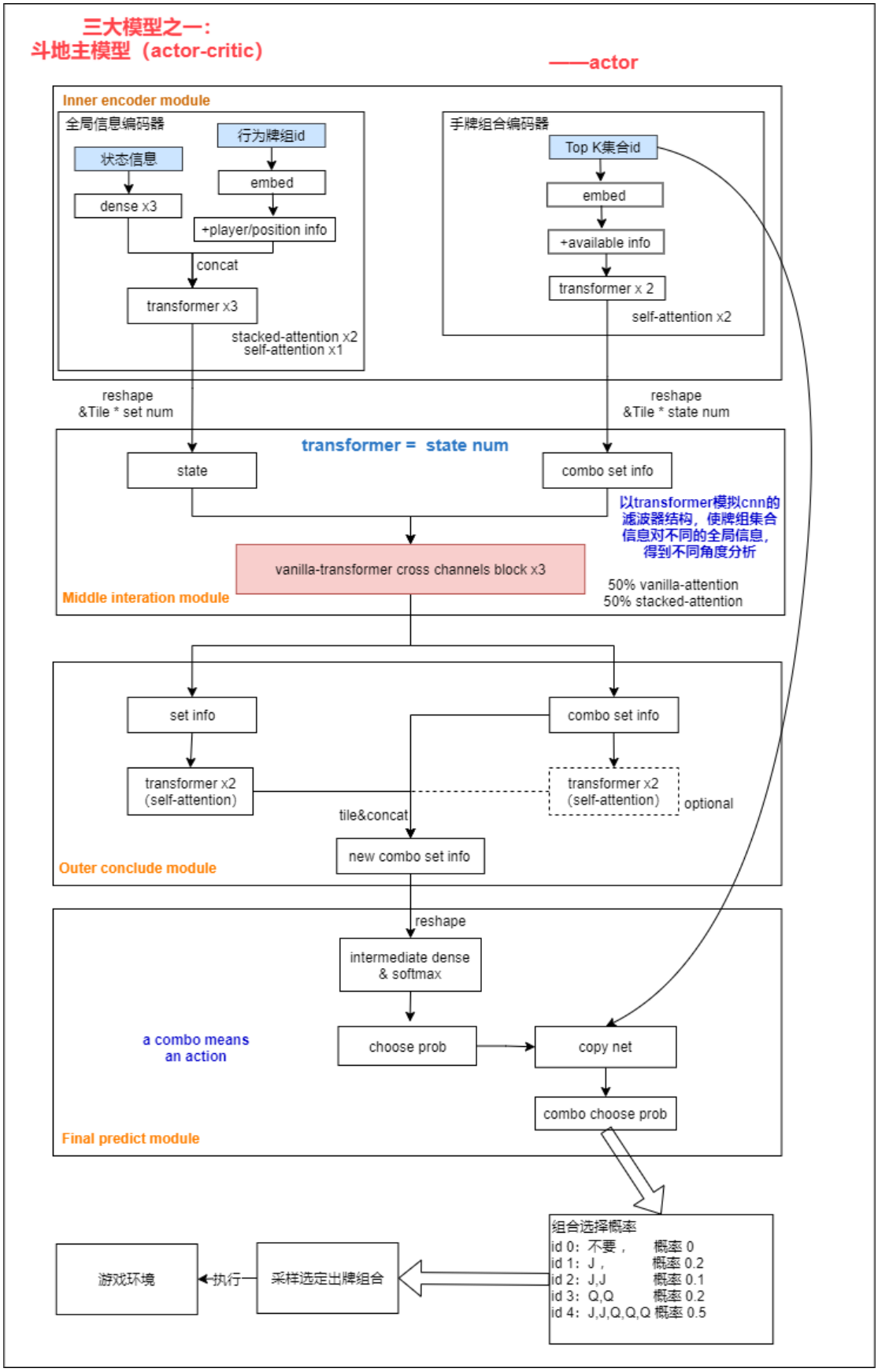
假设二：某个可执行牌组的执行概率等于所有牌组集合内的该牌组执行概率乘以相应牌组集合被采用概率（或者，所有牌组集合内该可执行牌组的执行概率的和）

$$\mathrm{P}_a = \sum_{k=0}^{n} \mathrm{p}_k * \mathrm{p}^{牌组}{}_k(a)$$

出牌阶段寻找到多个可执行牌组，每个可执行牌组存在于一个或多个牌组集合之中，如何计算可执行牌组的价值？直接且合理的想法就是，利用多个牌组集合的采用概率加权该可执行牌组在各个牌组集合的执行概率并求和。

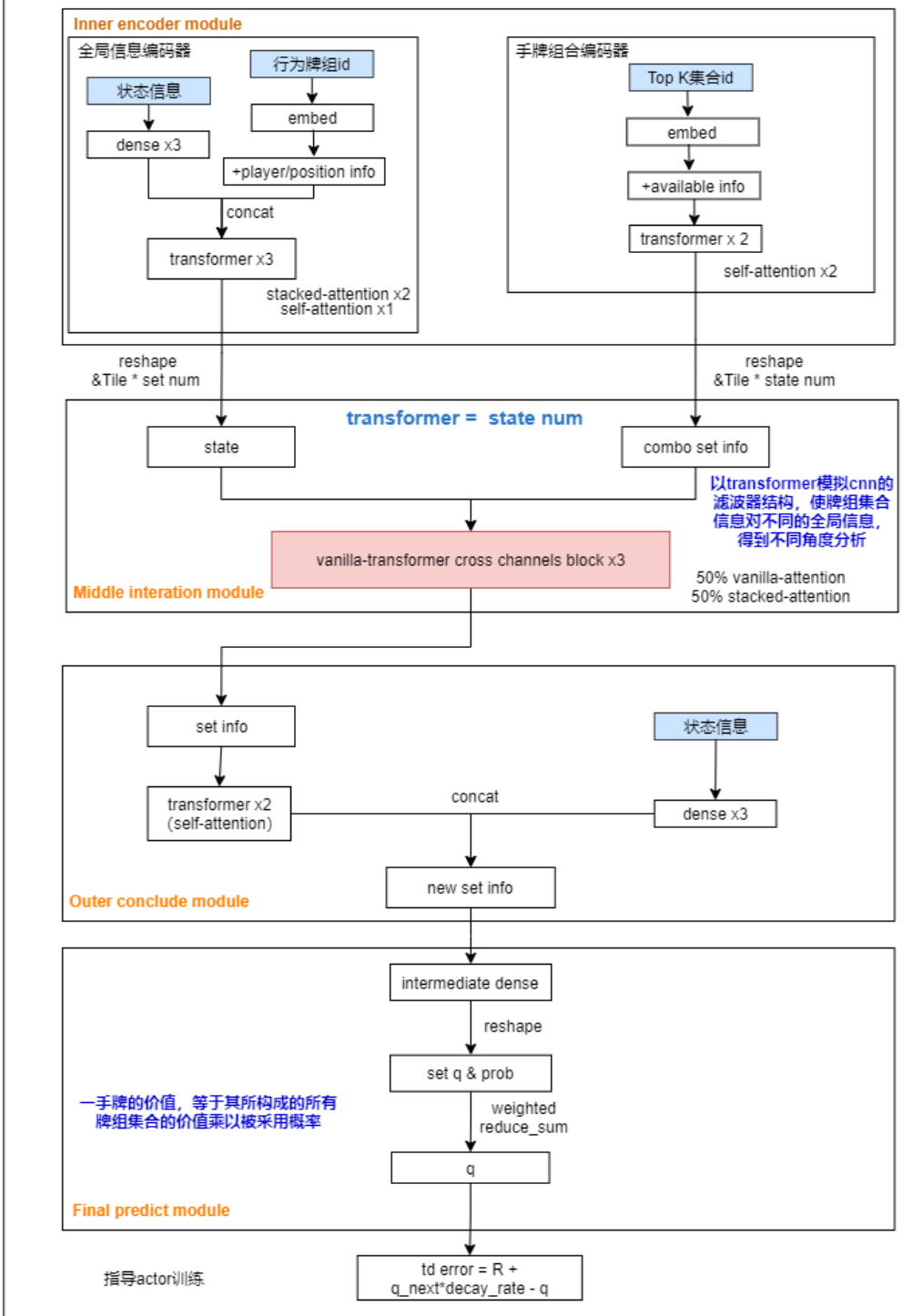
以上两个假设奠定了本系统各个模型的基石，深度优先搜索解决了牌组集合的空间复杂度问题，动态向量表使得同一个牌组在不同对局、时段的牌组也拥有表征，简化模型的推理难度，这些构成使用深度强化学习模型完成斗地主 ai 的前提条件。

在此基础上，本系统创新 stacked-attention 来完善多牌信息的融合和 transformer cross channels 实现外部状态与牌组集合的多通道交叉通信，设计不完全信息任务的 Importance Sampling 提高学习速率和效果，使用 copy-net、actor-critic、gradient clip 等技术完成三大模型的构建。

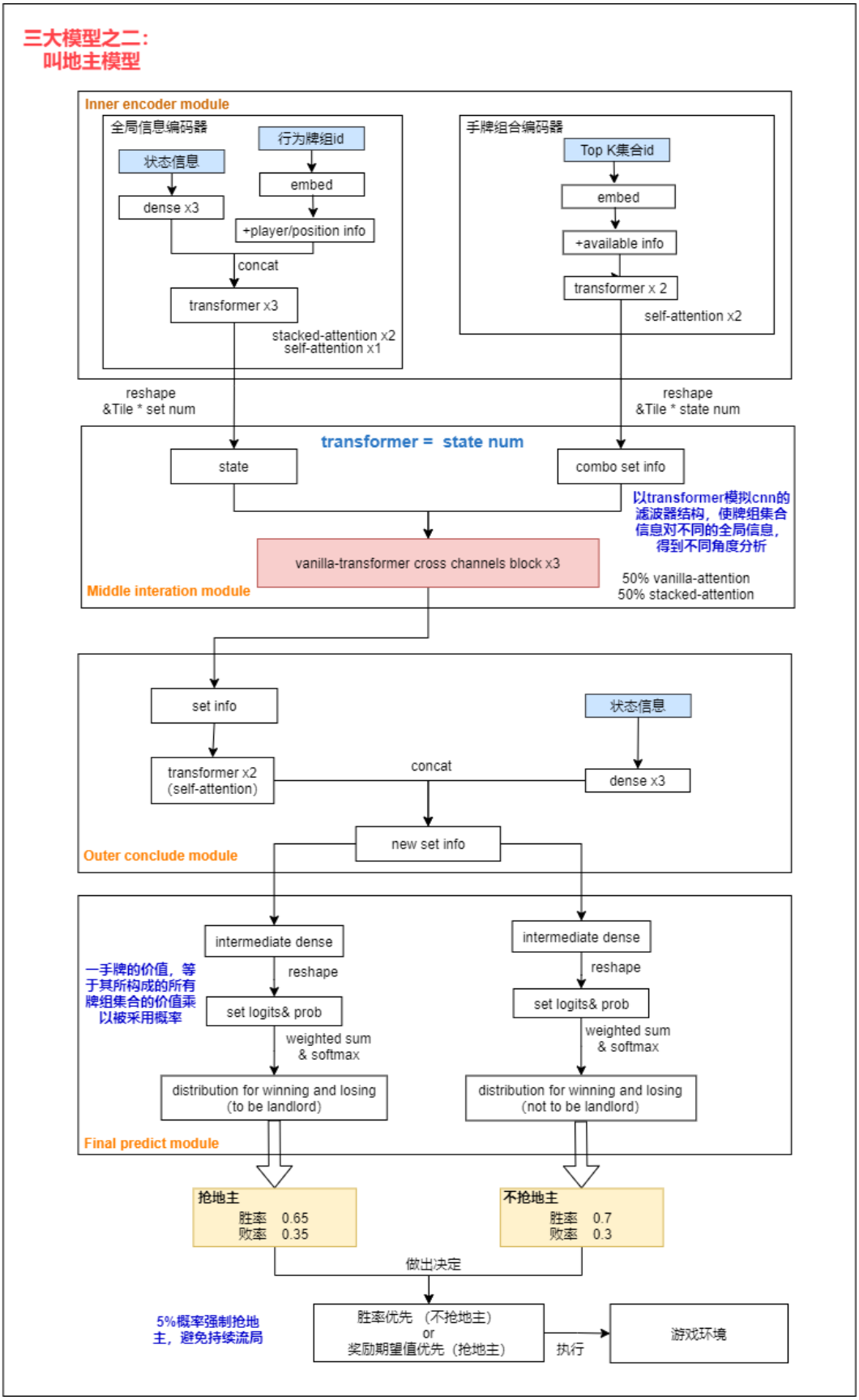

三大模型之一：
斗地主模型（actor-critic）
——actor
Inner encoder module
全局信息编码器
状态信息
dense x3
行为牌组id
embed
+player/position info
concat
transformer x3
stacked-attention x2
self-attention x1
手牌组合编码器
Top K集合id
embed
+available info
transformer x 2
self-attention x2
reshape
&Tile * set num
reshape
&Tile * state num
transformer = state num
state
combo set info
以transformer模拟cnn的滤波器结构，使牌组集合信息对不同的全局信息，得到不同角度分析
vanilla-transformer cross channels block x3
50% vanilla-attention
50% stacked-attention
Middle interation module
set info
combo set info
transformer x2
(self-attention)
transformer x2
(self-attention)
optional
tile&concat
new combo set info
Outer conclude module
reshape
intermediate dense
& softmax
a combo means
an action
choose prob
copy net
combo choose prob
Final predict module
组合选择概率
id 0：不要，　概率 0
id 1：J，　概率 0.2
id 2：J,J　概率 0.1
id 3：Q,Q　概率 0.2
id 4：J,J,Q,Q,Q 概率 0.5
采样选定出牌组合
执行
游戏环境

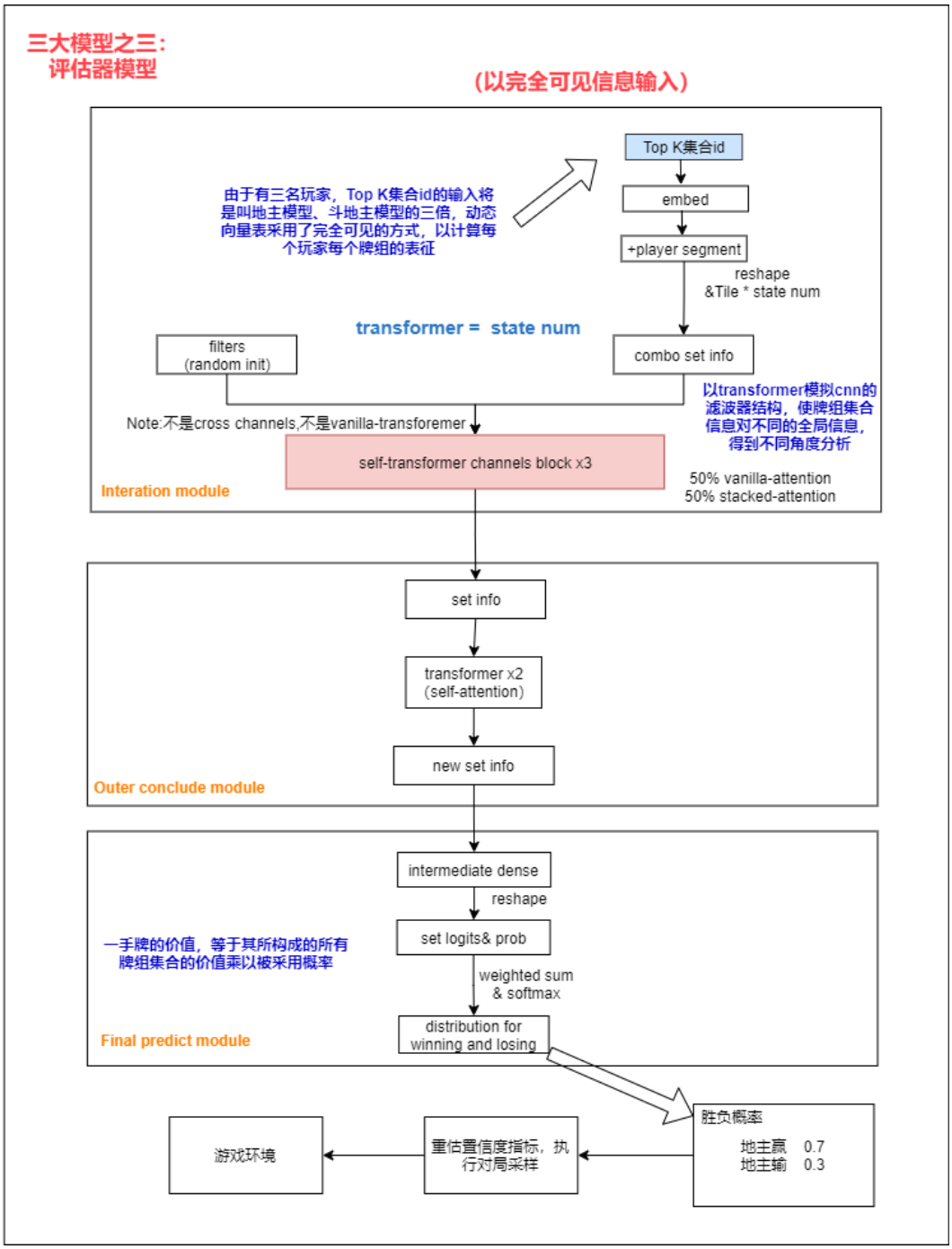
三大模型之一：
斗地主模型

——critic



三大模型之二：叫地主模型





实验环境

GPU 显卡	无
训练数据	无数据，RL 模型自对抗学习
训练时长	1 周

TODO LIST

1.猜牌：当其它玩家的牌量较少时，利用神经网络进行余牌预测，对预测结果使用博弈树搜索的方式寻找最优解法。

2.行为探索：就几万对局的训练而言，模型对农民方互相协作的策略探索度不足，导致模型倾向于互殴，地主胜率要高于农民。可以考虑：寻找真实环境的数据使模型快速收敛，后联合采用自对抗体系进一步优化模型；增加协助奖励机制，以及加大对协助行为的采样。

3.更大的系统：目前，无论牌组集合的采样，还是所搭建模型的宽度、深度，都属于比较简易的。考虑通过扩大采样范围、优化采样方法和使用更大模型，来获得更佳表现。

4.数据不均衡：评估器模型、叫地主模型都存在数据不均衡问题，值得等待解决。