ISCAS 中国科学院软件研究所学术年会’2023
暨计算机科学国家重点实验室开放周

学术论文

# 语言条件任务上的概念化强化学习框架

彭少辉，胡杏，张蕊，郭家明，易琦，陈睿智，
杜子东，李玲，郭崎，陈云霁

Conceptual Reinforcement Learning for Language-Conditioned Tasks,
The 37th AAAI Conference on Artificial Intelligence (AAAI 2023 Oral)

联系人：彭少辉 1881099268 pengshaohui@iscas.ac.cn

## 研究背景和动机

众所周知，迁移深度强化学习策略到未见过的相似环境上仍然是巨大的挑战。最近，研究者们提出了语言条件任务，希望通过语言作为中间管道促进策略的迁移。基于语言条件的策略的关键挑战之一是学习环境中的语言描述和观测的联合表示。观测和语言的联合表示应该能捕捉到不同环境之间的共享因素，从而使得建立在该联合表示上的策略可以更好地迁移到不同的环境中。

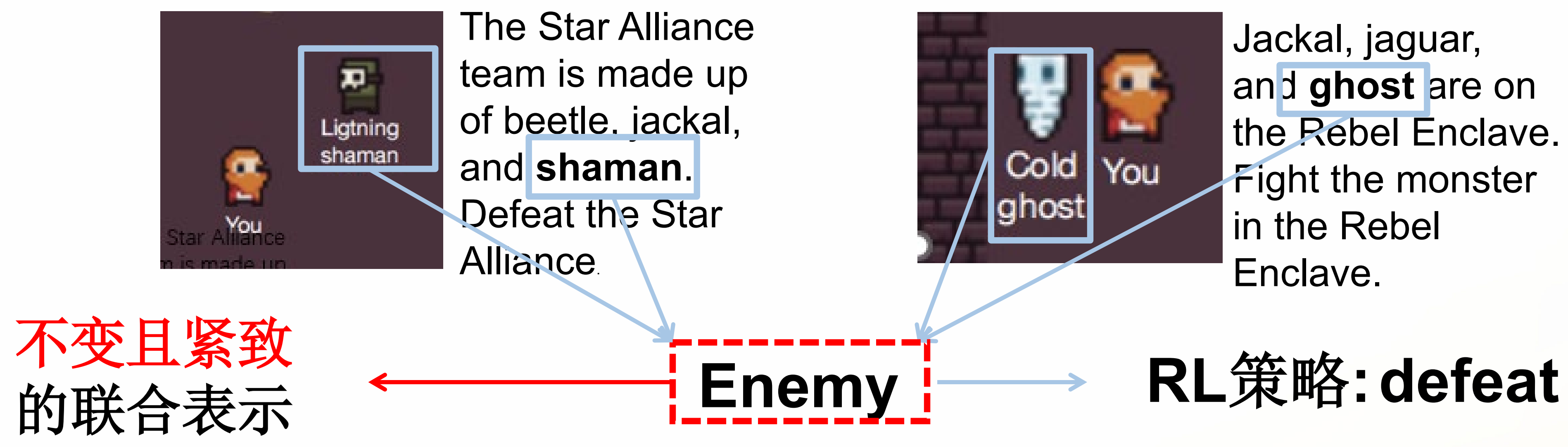
“概念”是人类认知中的重要组成部分，人类通过从众多的实例中抽取相似性，抽象出“概念”，并将其应用在相似任务中，大大提升了工作效率。受到“概念”的启发，我们提出显式地学习”概念化表示”，并构建基于该表示的策略的概念化强化学习范式。

“概念化表示”的具有两个关键的特征，即“不变性”和“紧致性”，能分别提升强化学习策略的迁移性能和效率。

**概念：**人类从众多实例中总结提炼出的**共性抽象**表示 ⇨ 人类活动中通用


**food** → **人类活动：eat**

**语言条件任务中的概念：**不同环境**实体**的**共性** ⇨ 强化学习策略通用

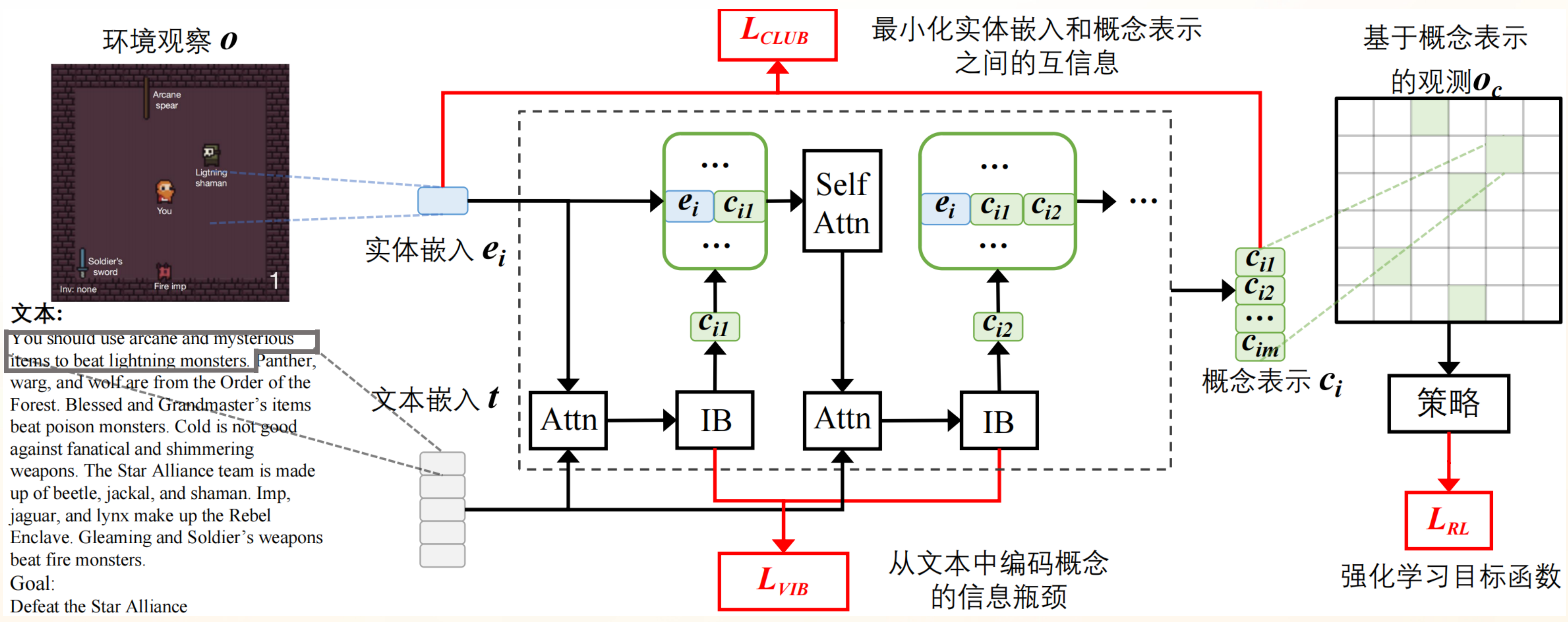


The Star Alliance team is made up of beetle, jackal, and **shaman**. Defeat the Star Alliance

Jackal, jaguar, and **ghost** are on the Rebel Enclave. Fight the monster in the Rebel Enclave.

不变且紧致的联合表示 ← **Enemy** → **RL策略: defeat**

## 概念化强化学习（CRL）



You should use arcane and mysterious items to beat lightning monsters. Panther, warg, and wolf are from the Order of the Forest. Blessed and Grandmaster’s items beat poison monsters. Cold is not good against fanatical and shimmering weapons. The Star Alliance team is made up of beetle, jackal, and shaman. Imp, jaguar, and lynx make up the Rebel Enclave. Gleaming and Soldier’s weapons beat fire monsters.
Goal:
Defeat the Star Alliance

**概念化强化学习框架主要包含三部分：**

- 多层注意力编码器，如黑色虚线框中所示
- 不变性限制$L_{CLUB}$，最小化实体和概念之间的互信息
- 紧致性限制$L_{VIB}$，限制文本到概念的信息瓶颈

## 结果

**在著名语言条件游戏RTFM上的测试：**

| Transfer from | Method | 6 × 6 | 6 × 6 dyna | 6 × 6 groups | 6 × 6 nl | 6 × 6 dyna groups | 6 × 6 groups nl | 6 × 6 dyna nl | 6 × 6 dyna groups nl | Training Steps |
|---|---|---|---|---|---|---|---|---|---|---|
| random | txt2$\pi$<br>CRL | 84 ± 20<br>**93 ± 3** | 26 ± 7<br>36 ± 10 | 25 ± 3<br>33 ± 10 | 45 ± 6<br>38 ± 4 | 23 ± 2<br>17 ± 2 | 25 ± 3<br>20 ± 2 | 23 ± 2<br>17 ± 3 | 23 ± 2<br>16 ± 3 | 100$M$<br>**30$M$(↓ 70%)** |
| 6 × 6 | txt2$\pi$<br>CRL | | **85 ± 9**<br>**89 ± 9** | 82 ± 19<br>**96 ± 3** | 78 ± 24<br>**97 ± 2** | 64 ± 12<br>**85 ± 1** | 52 ± 13<br>**87 ± 2** | 53 ± 18<br>**85 ± 4** | 40 ± 8<br>**86 ± 3** | 50$M$<br>**30$M$(↓ 40%)** |

紧致性：更高训练效率

不变性：更高迁移性能