Security and Privacy in Artificial Intelligence

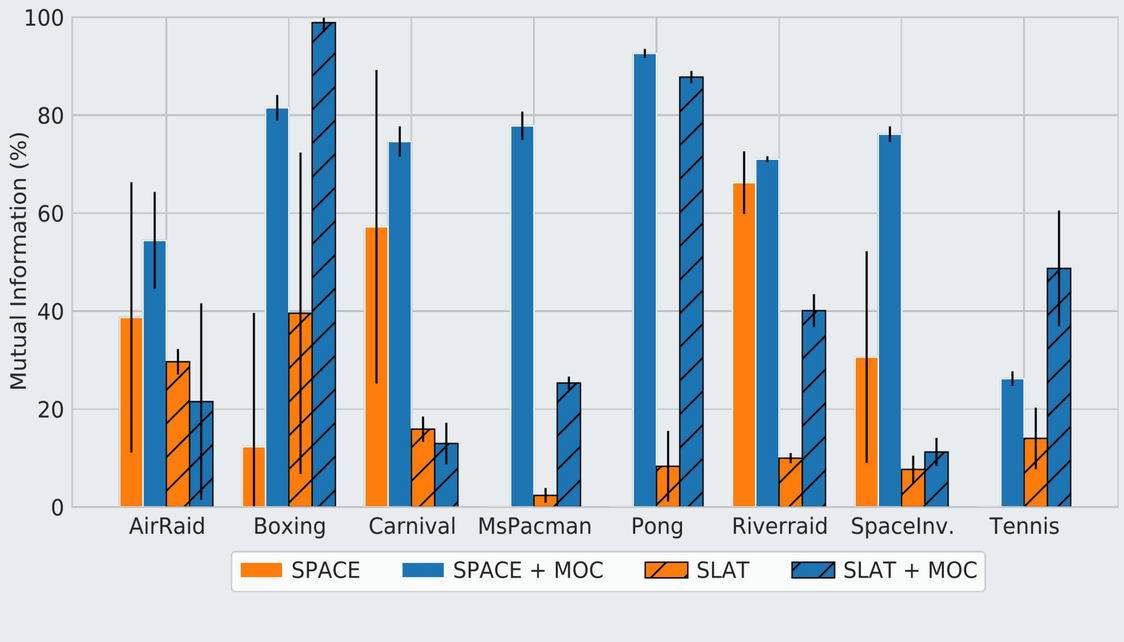
Interactive RL Agents using Interpretable Concept Bottlenecks

Motivation and Background

- Deep RL agents are black box models, from which decisions cannot be understood.
- This is a break threat of security, as adversarial policies can thus be more easilly obtained.
- For example, in the very famous Pong game, learning agent can learn to mostly focus on the Enemy's paddle position.
- During training, the enemy is following the vertical position of the ball, thus correlating his own position with it.
- RL agents thus usually perform shortcut learning.
- SCoBots RL agents (ours) extract human interpretable features, that allow us to understand the reasonning behind their decisions, and even interact to correct such misalignment.

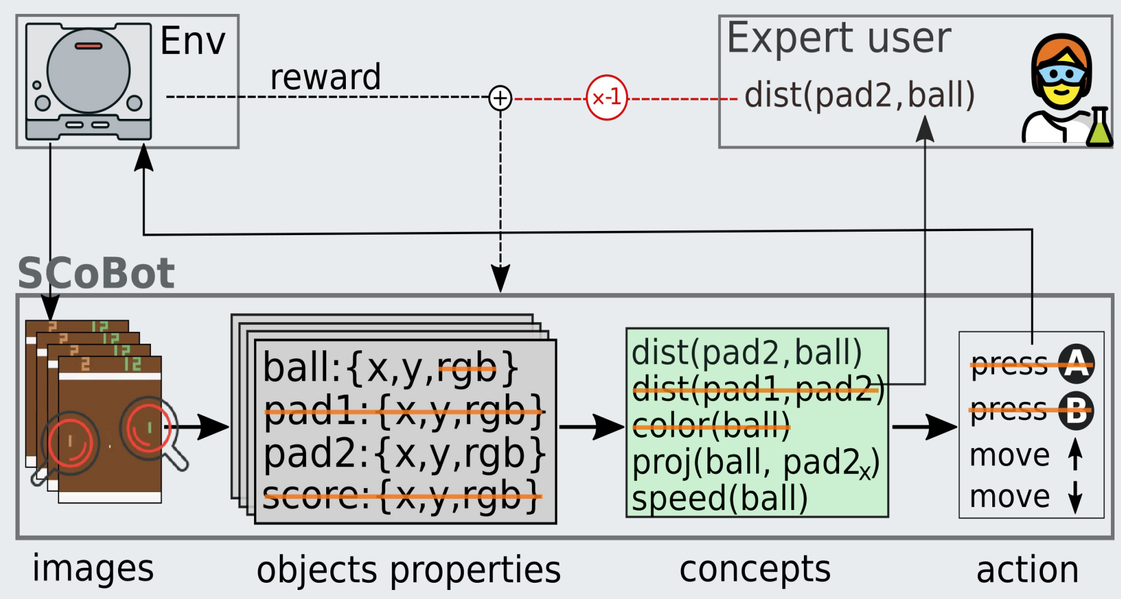
Optimized objects representation

- We first need to extract object-centric representations of the RGB input.
- As manually labelling data is costly, we took unsupervised image discovery models (e.g. SPACE, SLot Attention) and tested them on RL environments (Atari games).
- These methods have poor object detection capabilities for most games.
- Even when objects are detected, their internal representations are to fuzzy to perform complicated downstream tasks, such as RL.
- We introduced the MOC framework, that leverage time priors to improve both object detection and internal representation of object-discovery models.
- We showed that MOC-assisted frameworks can be used to control agents playing simple atari games.



Human understandable decisions

- To play more complicated games, RL agents not only needs an object-centric description of the state (i.e. a list of objects and their positions).
- They need more advanced "relational" concepts.
- Our SCoBots agents extract human interpretable relational concepts from object-centric scene representations.
- SCoBots allow us to fully understand the reasoning behind their decisions.
- SCoBots also allow us to interact with their internal representations.
- This allows us to teach SCoBots to e.g. not take into account the enemy for his decision.
- Many interventions are possible to avoid shortcut learning or missalignments.
- SCoBots are thus more secure agents that their deep competitors.



Impact

- Please describe the impact you have achieved with your result.
- Explaining impact of your work and of your result is important for ATHENE.
- …

Conclusion

- Please give a short conclusion.
- You may also mention urgent follow up work here.
- …

Involved ATHENEians

- Quentin Delfosse
- Kristian Kersting
- Jan Peters
- Nafise Sadat Moosavi
- Iryna Gurevych.

ATHENE is a research Center of the Fraunhofer-Gesellschaft with the participation of

Fraunhofer SIT
Fraunhofer IGD
TECHNISCHE UNIVERSITÄT DARMSTADT
GOETHE UNIVERSITÄT FRANKFURT AM MAIN
h_da hochschule darmstadt