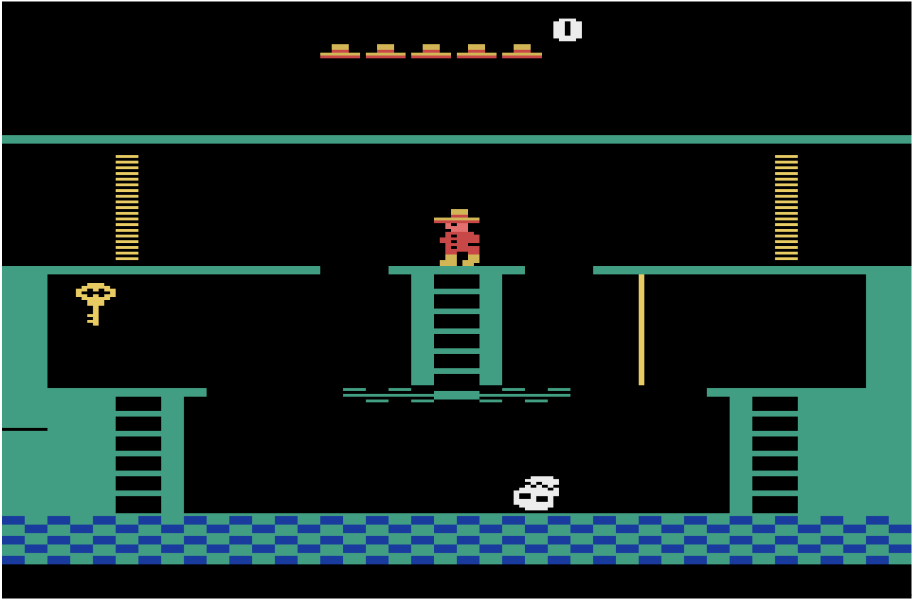
Survey: Leveraging Human Guidance for Deep Reinforcement Learning Tasks

Ruohan Zhang, Faraz Torabi, Lin Guan,
Dana H. Ballard, Peter Stone

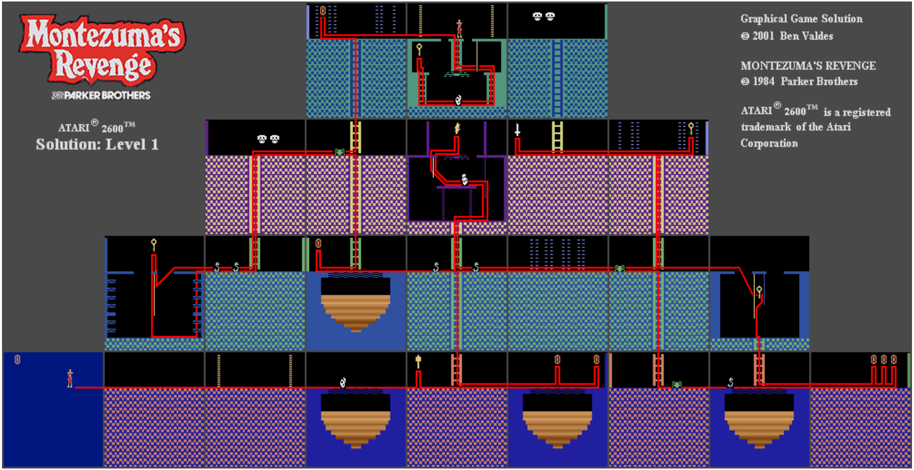
University of Texas at Austin

Presented by Lin Guan

A Reinforcement Learning Problem: Montezuma's Revenge



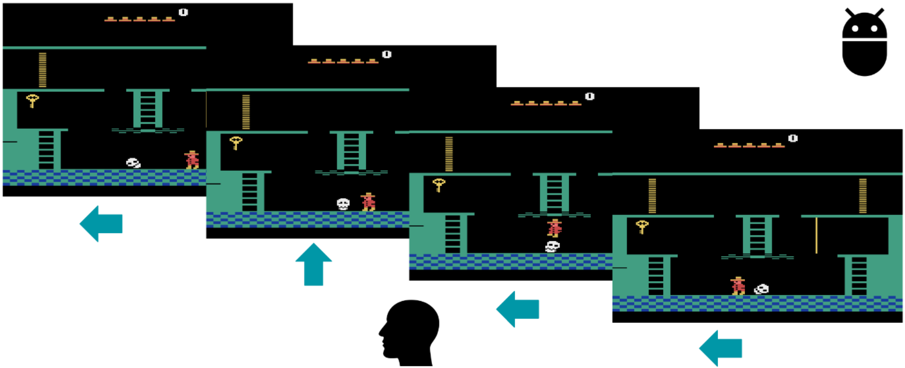
A Reinforcement Learning Problem: Montezuma's Revenge



A Reinforcement Learning Problem: Montezuma's Revenge

Find an optimal **policy**, i.e., the action to take in an observed state that maximizes expected longterm reward

Montezuma's Revenge: Imitation Learning



Survey Scope

- 64 papers, 5 types of human guidance that...

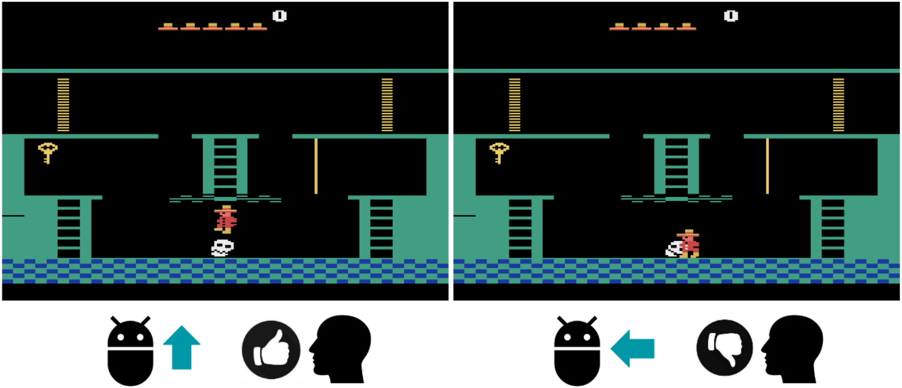
Survey Scope

- 64 papers, 5 types of human guidance that...
- Are beyond conventional step-by-step action demonstrations

Survey Scope

- 64 papers, 5 types of human guidance that...
- Are beyond conventional step-by-step action demonstrations
- Have shown promising results in training agents to solve deep reinforcement learning tasks

Montezuma's Revenge: Evaluative Feedback



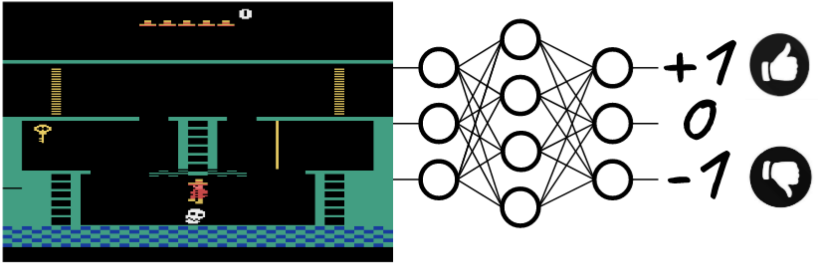
Motivation

While the true reward is delayed and sparse, human evaluative feedback is immediate and dense.

Representative Works

Interpreting human feedback as:

- Reward function, replacing reward provided by the environment
- TAMER: Training an agent manually via evaluative reinforcement [Knox and Stone, 2009, Warnell et al., 2018]



Representative Works

Interpreting human feedback as:

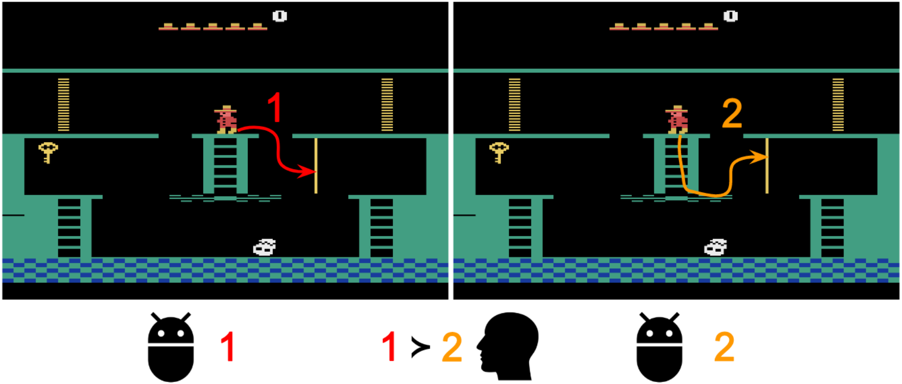
- Direct policy labels
  - Advise [Griffith et al., 2013, Cederborg et al., 2015]

Representative Works

Interpreting human feedback as:

- Direct policy labels
  - Advise [Griffith et al., 2013, Cederborg et al., 2015]
- Advantage function
  - COACH: Convergent actor-critic by humans [MacGlashan et al., 2017]
  - This interpretation explains human feedback behaviors better in several tasks
  - Still an unresolved issue that requires carefully designed human studies

Montezuma's Revenge: Human Preference



Motivation

Ranking behaviors is easier than rating them.
And sometimes the ranking can only be provided at the end of a behavior trajectory.

Representative Works

- [Christiano et al., 2017]: As an inverse reinforcement learning problem, i.e., learn human reward function from human preference rather than from demonstration

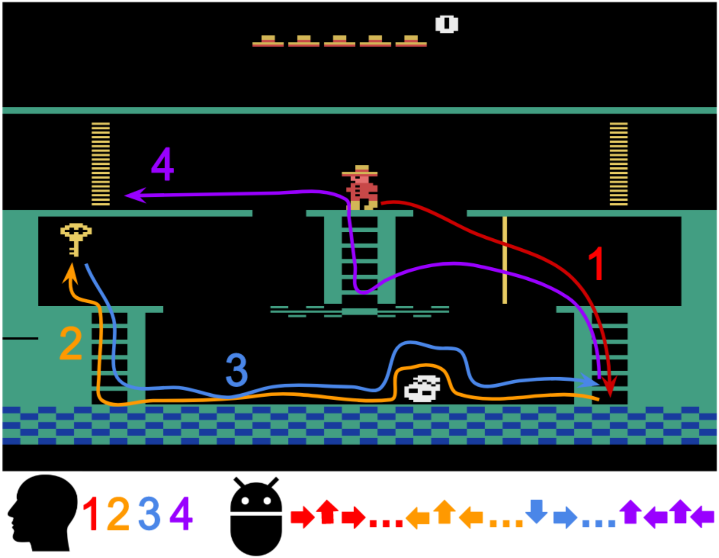
Representative Works

- [Christiano et al., 2017]: As an inverse reinforcement learning problem, i.e., learn human reward function from human preference rather than from demonstration
- Query selection? Preference elicitation [Zintgraf et al., 2018]

Representative Works

- [Christiano et al., 2017]: As an inverse reinforcement learning problem, i.e., learn human reward function from human preference rather than from demonstration
- Query selection? Preference elicitation [Zintgraf et al., 2018]
- Many good works on preference-based reinforcement learning [Wirth et al., 2017]

Montezuma's Revenge: Hierarchical Imitation



Motivation

Human is good at specifying high-level abstract goals while the agent is good at performing low-level fine-grained controls.

Representative Works

- High-level+low-level demonstrations [Le et al., 2018]

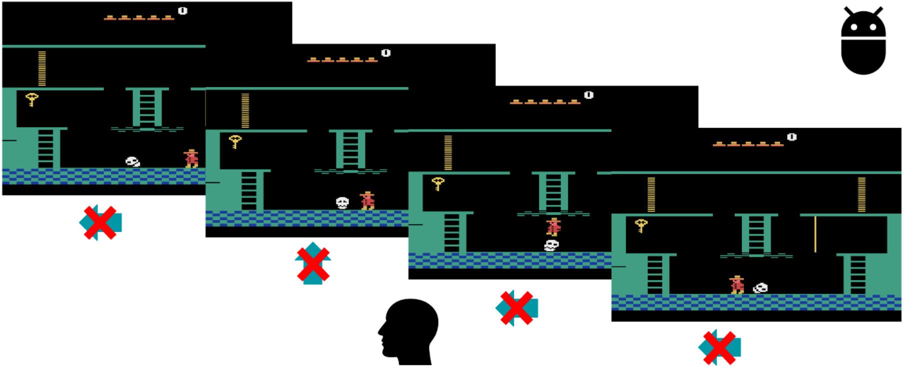
Representative Works

- High-level+low-level demonstrations [Le et al., 2018]
- High-level demonstrations only [Andreas et al., 2017]

Representative Works

- High-level+low-level demonstrations [Le et al., 2018]
- High-level demonstrations only [Andreas et al., 2017]
- A promising combination:
  - High-level: Imitation learning, e.g., DAgger [Ross et al., 2011]
  - Low-level: Reinforcement learning, e.g., DQN [Mnih et al., 2015]

Montezuma's Revenge: Imitation from Observation



Motivation

To utilize a large amount of human demonstration data that do not have action labels, e.g., YouTube videos

Representative Works

- Challenge 1: Perception
  - Viewpoint [Liu et al., 2018, Stadie et al., 2017]
  - Embodiment [Gupta et al., 2018, Sermanet et al., 2018]

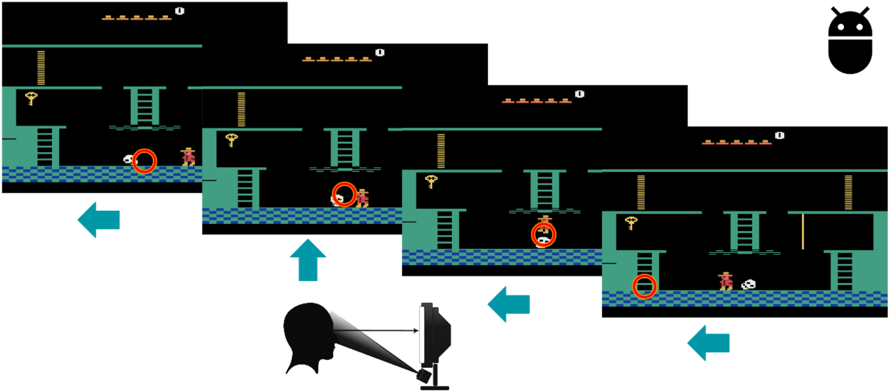
Representative Works

- Challenge 1: Perception
  - Viewpoint [Liu et al., 2018, Stadie et al., 2017]
  - Embodiment [Gupta et al., 2018, Sermanet et al., 2018]
- Challenge 2: Control
  - Model-based: Infer the missing action given a state transitions (s, s') by learning an inverse dynamics model [Nair et al., 2017, Torabi et al., 2018a]
  - Model-free: e.g., bring the state distribution of the imitator closer to that of the trainer using generative adversarial learning [Merel et al., 2017, Torabi et al., 2018b]

Representative Works

- Challenge 1: Perception
  - Viewpoint [Liu et al., 2018, Stadie et al., 2017]
  - Embodiment [Gupta et al., 2018, Sermanet et al., 2018]
- Challenge 2: Control
  - Model-based: Infer the missing action given a state transitions (s, s') by learning an inverse dynamics model [Nair et al., 2017, Torabi et al., 2018a]
  - Model-free: e.g., bring the state distribution of the imitator closer to that of the trainer using generative adversarial learning [Merel et al., 2017, Torabi et al., 2018b]
- Please see paper#10945: **Recent Advances in Imitation Learning from Observation** [Torabi et al., 2019]

Montezuma's Revenge: Human Attention

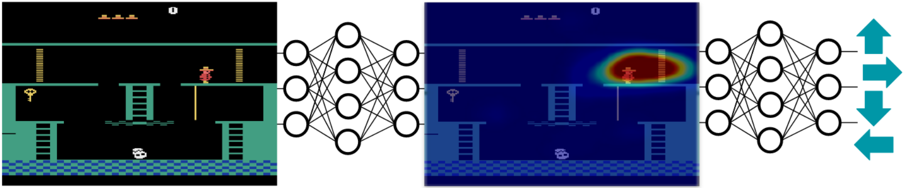


Motivation

Human visual attention provides additional information on *why* a particular decision is made, e.g., by indicating the current object of interest.

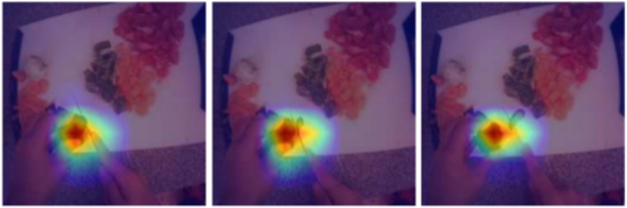
Representative Works

- AGIL: Attention-guided imitation learning [Zhang et al., 2018]

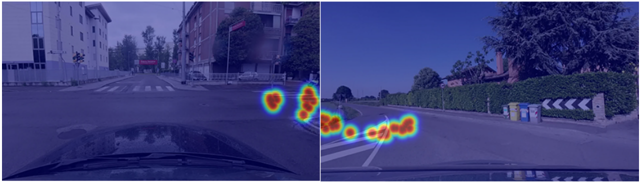


- Including attention does lead to higher accuracy in imitating human actions

Representative Works



(a) Cooking [Li et al., 2018]



(b) Driving [Palazzi et al., 2018, Xia et al., 2019]

Survey Scope

An agent can learn...

- From human evaluative feedback
- From human preference
- From high-level goals specified by humans
- By observing human performing the task
- From human visual attention

Future Directions

- Shared datasets and reproducibility
- Understanding human trainers' behaviors, e.g.,[Thomaz and Breazeal, 2008]
- A unified lifelong learning framework [Abel et al., 2017]

Survey: Leveraging Human Guidance for Deep Reinforcement Learning Tasks

Ruohan Zhang, Faraz Torabi, Lin Guan,
Dana H. Ballard, Peter Stone

University of Texas at Austin

Presented by Lin Guan

Thank You!

References

Abel, D., Salvatier, J., Stuhlmüller, A., and Evans, O. (2017).
Agent-agnostic human-in-the-loop reinforcement learning.
NeurIPS Workshop on the Future of Interactive Learning Machines.

Andreas, J., Klein, D., and Levine, S. (2017).
Modular multitask reinforcement learning with policy sketches.
In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 166–175. JMLR. org.

Cederborg, T., Grover, I., Isbell, C. L., and Thomaz, A. L. (2015).
Policy shaping with human teachers.
In *Twenty-Fourth International Joint Conference on Artificial Intelligence.*

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017).
Deep reinforcement learning from human preferences.
In *Advances in Neural Information Processing Systems*, pages 4299–4307.

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., and Thomaz, A. L. (2013).
Policy shaping: Integrating human feedback with reinforcement learning.
In *Advances in neural information processing systems*, pages 2625–2633.

Gupta, A., Devin, C., Liu, Y., Abbeel, P., and Levine, S. (2018).
Learning invariant feature spaces to transfer skills with reinforcement learning.
In *International Conference on Learning Representations.*

Knox, W. B. and Stone, P. (2009).
Interactively shaping agents via human reinforcement: The tamer framework.
In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16. ACM.

Le, H., Jiang, N., Agarwal, A., Dudik, M., Yue, Y., and Daumé, H. (2018).
Hierarchical imitation and reinforcement learning.
In *International Conference on Machine Learning*, pages 2923–2932.