Sequential Decision Making Algorithms for Human-AI interaction

Nikals Höpner [1], Ilaria Tiddi [2] and Herke van Hoof [1]

[1]University of Amsterdam, [2]Vrije Universiteit Amsterdam

Motivation

- Human-AI interaction is most naturally framed as a sequential decision making (SDM) problem.
- **Goal**: Create an adaptive and collaborative embodied agent within the SDM framework.
- Progress has been restricted to the text / text-visual domain [1], here the focus is on embodied agents.

Research Questions

- How can prior knowledge sources be integrated into a SDM agent?
- How can we train and build embodied instructable agents?
- How to build an assistant that is collaborative and adapative?

Instructable Agent from Play Data

Problem Setting

- Small dataset of annotated trajectories $D_{ann} = (\tau_k, o_{1:T_k})_{k=1}^{N}$ and large dataset of unannotated trajectories $D_{unann} = (o_{1:L_k})_{k=1}^{M}$, where $M >> N$ and $T_k << L_k$.
- **Goal**: Leverage the small dataset to annotate the large unannotated dataset and improve the downstream policy performance by increasing the amount of labelled training data.
- **Challenge**: The long unannotated trajectories contain multiple instructable behaviours and need to be segmented before they can be labelled.
- **Evaluation environment**: Simulation of 7-DOF Franka Emika Panda robot arm, mainpulating a set of objects [2]

Methodology & Algorithm

First Step: Train a vision-language model (VLM) [3] based labelled segmentation model.

- Labelled Segmentation Model:

$$
\begin{aligned}
&p_\theta(\tau_{1:C(\alpha_{1:T})}, \alpha_{1:T} | o_{1:T}) \\
&= p_{\theta_{lab}}(\tau_{1:C(\alpha_{1:T})} | \alpha_{1:T}, o_{1:T}) \cdot p_{\theta_{seg}}(\alpha_{1:T} | o_{1:T}) \\
&= \prod_{k=1}^{C(\alpha_{1:T})} p_{\theta_{lab}}(\tau_k | o_{t_k:t_{k+1}-1}) \cdot \prod_{t=1}^{T} p_{\theta_{seg}}(\alpha_t | o_{\gamma_t+1:t}),
\end{aligned}
$$

  where $\alpha_i = 1$ indicates the end of a segment at timestep i, $C(\alpha_{1:T})$ represents the number of segments and τ_k the instruction of the k-th segment
- Train θ_{seg} and θ_{lab} on the annotated dataset with a binary classification loss and the contrastive CLIP objective respectively.
- Find the most likely labelled segmentation via dynamic programming in $O(T^3)$ time

Second Step: Train a policy via multi-context imitation learning [2] on the augmented labelled and unlabelled dataset.

Experimental Results

Test the performance gain compared to random sampling via the following dataset compositions:

100% groundtruth (gt)

50 % gt + 50 % labelled gt segmentation (seg)

50 % gt + 50 % labelled random seg

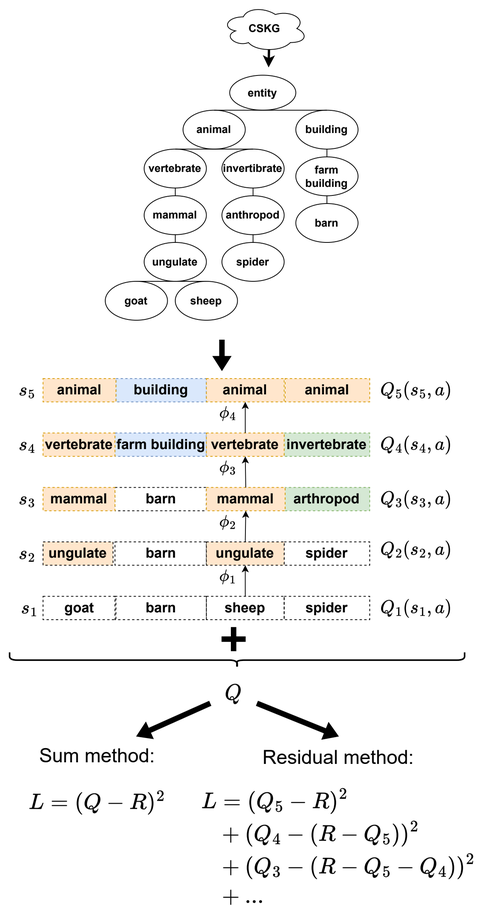
50 % gt + 50 % labelled learned seg

Equipping Agents with Commonsense Knowledge from Knowledge Graphs

Problem Setting & Approach

- A reinforcement learning (RL) agent needs to have prior knowledge about the world to successfully interact with humans.
- **Goal**: Enabling a RL agent to leverage commonsense knowledge stored in an open-source knowledge graph **(KGs)**.
- **Research Idea**: Use the **subclass relation** present in open-source KGs to build a **hierarchy of abstract states**. Leverage this hierarchy to learn a policy that can generalise to **unseen objects** and can be trained with less samples.
- Extract subclass relations from open-source KGs such as **ConceptNet, Wordnet and DBpedia**.
- Learn individual action-value functions for each abstract state and aggregate them by summing them up.
- The sum methods trains only the aggregate action-value function via the normal DQN loss, while the residual loss forces each individual action value function to learn the best possible approximation of the optimal action values function given its abstraction level.

Methodology



Experimental Evaluation

Generalisation

Method	Reward Valid.	Reward Test
Base	0.91 (0.04)	0.85 (0.05)
Base-H	0.83 (0.06)	0.75 (0.14)
Base-L	0.90 (0.04)	0.86 (0.06)
Manual defined abstraction: M-R	0.96 (0.03)	0.96 (0.02)
Manual defined abstraction: M-S	**0.97 (0.02)**	**0.96 (0.02)**
WordNet abstraction: W-R	**0.93 (0.02)**	**0.94 (0.02)**
WordNet abstraction: W-S	0.88 (0.10)	0.87 (0.17)

Manual performs best

Sum approach more unstable

Figure 1. Results on generalisation to unseen objects.

Sample Efficiency

Figure 2. Results on sample efficiency.

Future Work

- Learning from observing other agents acting.
- Combining offline reinforcement learning with online reinforcement learning.
- Learning Text2Trajectory models.

References

[1] Danny Driess et al. "Palm-e: An embodied multimodal language model". In: *arXiv preprint arXiv:2303.03378* (2023).

[2] Oier Mees et al. "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks". In: *IEEE Robotics and Automation Letters* 7.3 (2022), pp. 7327–7334.

[3] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.