ViTLearn: Vision-Text for Robot Learning using LMMs

Ayush Das, Brendan Tidd, Yifei Chen, Jen Jen Chung

Motivation

- Designing optimal reward functions for RL is a challenge, especially for complex tasks like walking.
- Recently, LMMs have gained attention for their remarkable capability to translate natural language into machine-level instructions. [1][2]
- **ViTLearn** explores the use of large multi-modal models (LMMs) to generate rewards for robotic tasks.

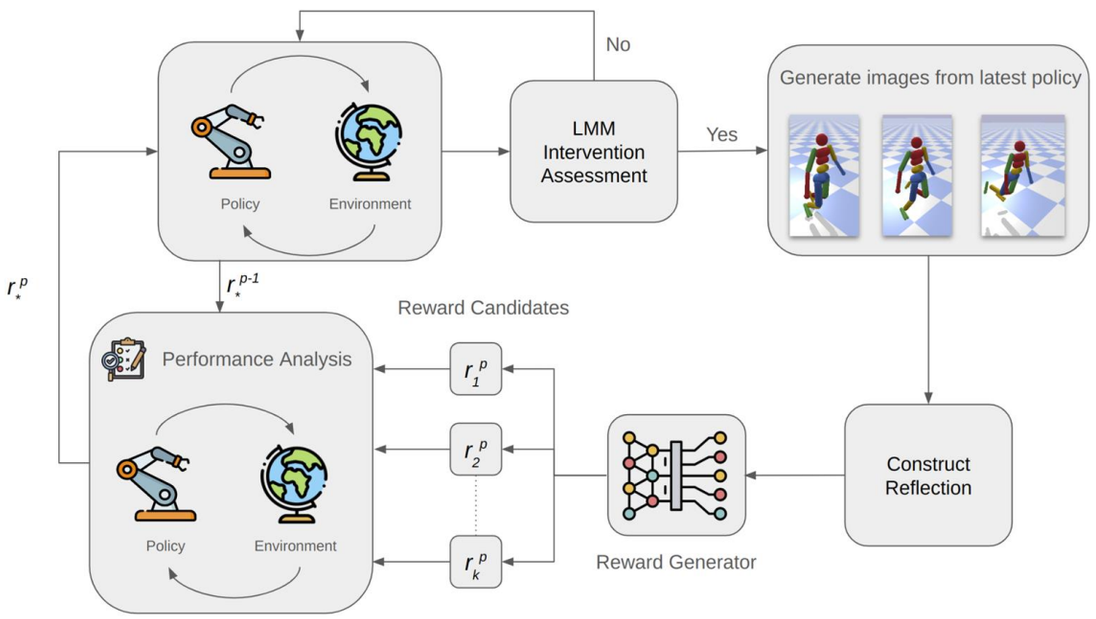
Research Goals

- **Automate Reward Shaping**: Use LMMs to generate reward functions based on vision and text inputs.
- **Compare Multi-Modal and Text-Only Approaches**: Validate the impact of visual feedback on performance.
- **Evaluate Across Simulated and Real Environments**: Test the framework on a humanoid agent and other morphologies.

Experimental Setup

- **Simulation Environment**: PyBullet physics engine.
- **Task**: Train a humanoid robot to walk—a task difficult to describe solely with text.
- **LMM Used**: GPT-4o (June 2024 release).
- **Parallel Training**: OpenMPI framework to manage multiple reward functions across worker nodes.

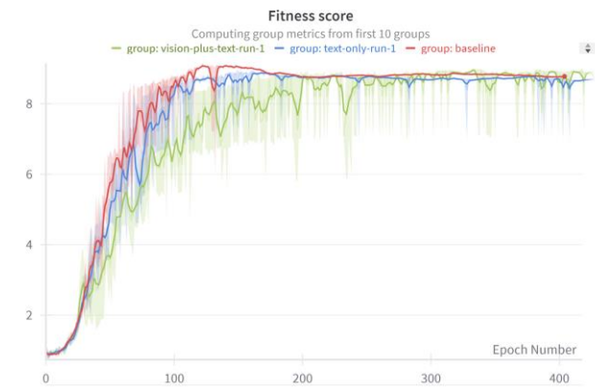
ViTLearn Overall Architecture

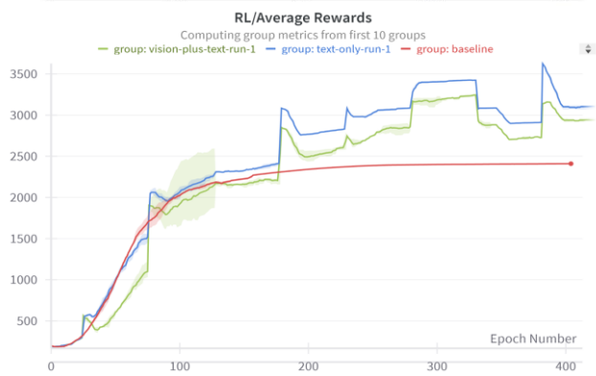


Methodology

- **Fitness score:** Metric chosen to evaluate performance across different reward models.
- **Baseline Method**: A handcrafted reward function is used and trained in a conventional RL architecture and label that as the baseline performance.
- **LMM intervention assessment:** If policy is 20% better than fitness threshold then Reward Generator is triggered and update fitness threshold.
- **LMM-Reward Automation Process:**
  - Images from the latest policy is captured and used to construct a reflection along with metrics like fitness score
  - Best reward sample is chosen from the generated set, after evaluating over 10 epochs.

Experiments & Results





ViTLearn marginally outperforms other approaches illustrated through the fitness score.
Interestingly, the average reward behaviours seem similar to the text-based approach.

Conclusion

- ViTLearn demonstrates that multi-modal inputs provide additional insight for robot behavior learning.
- However, additional work is required to guide the variability of response

References

[1] Y. J. Ma et al., "Eureka: Human-Level Reward Design via Coding Large Language Models," 2023. (Accessed Mar. 11 2024), Available: https://arxiv.org/abs/2310.12931.
[2] T. Brown et al., "Language Models are few-shot Learners," Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

School of Electrical Engineering and Computer Science
RPPL Lab
The University Of Queensland
AUSTRALIA