Learning Transferable Subgoals with Clip and Yolo Embedding

Yuechuan Yang [*1] **Kyle Lee** [*1]

Abstract

Transfer learning holds significant potential in hierarchical RL works. We seek agents that can decompose problems into subgoals, learn skills to accomplish those subgoals, then flexibly recombine previously learned skills to solve new problems. We focus on the challenge of skill reuse across tasks from a single-task training setup, which we term Single-Task Skill Generalization (STSG). We propose a method to represent subgoals with an ensemble of classifiers, each encoding a distinct hypothesis over features likely to generalize. Using task reward as a signal, the agent identifies which subgoal hypothesis best supports transfer. Past experiments on this STSG setup have involved the MONTEZUMA'S REVENGE and MINIGRID environments, showing robust subgoal generalization. In our particular experiment, we manually collect a dataset of real-world kitchen images, such as a microwave, a fridge, and a stove. Preliminary results suggest that the combination of these embeddings yields a promising subgoal representation space, one in which conceptually similar kitchen-related tasks (e.g., "open fridge," "put item in microwave") can be reused across multiple task configurations.

1. Introduction

Hierarchical Reinforcement Learning (HRL) (Barto & Mahadevan, 2003) promises scalable solutions to complex tasks by learning reusable skills or options (Sutton et al., 1999). However, most methods assume either multi-task training or oracle-like task sampling (Frans & et al., 2017; Barreto & et al., 2018), which limits real-world applicability. In contrast, we propose the Single-Task Skill Generalization (STSG) setting, where an agent must discover skills in a single task and reuse them in unseen, sequentially presented tasks. Our method learns multiple hypotheses about generalizable features for each discovered subgoal and selects among them using downstream task rewards.

2. Background

2.1. Literature Review

Skill reuse in HRL has been studied under multi-task transfer and continual learning frameworks (Khetarpal & et al., 2022; Wang et al., 2024). Prior works assume access to multiple tasks or task distributions (Barreto & et al., 2019; Frans & et al., 2017), or focus on task-agnostic representation learning (Higgins & et al., 2017; Nair et al., 2020). In contrast, our STSG setting considers only a single training task. Subgoal discovery methods (Pateria & et al., 2021) help identify termination sets for options, but rarely address generalization. Our work integrates ensemble learning (Pagliardini et al., 2022) to hypothesize transferable features and leverages reward signals for hypothesis selection.

3. Methodology

3.1. Experiment Setting

Previous experiments on this portable option framework have been on two environments: MONTEZUMA'S REVENGE (Bellemare & et al., 2013; Machado & et al., 2018), a pixel-based sparse-reward platformer, and MINIGRID DOORMULTIKEY (Chevalier-Boisvert & et al., 2023), a procedurally generated grid world requiring key-object interactions. In our experiment setting, we work with hand-collected images of real-world kitchen objects.

3.2. Experiment Procedure

For each discovered subgoal, we train an ensemble of classifiers using D-BAT (Pagliardini et al., 2022) or random initialization. Each ensemble member defines a candidate subgoal; we train a corresponding low-level policy to reach it (Van Hasselt et al., 2016). A high-level PPO agent (Schulman & et al., 2017) then selects among these subgoal policies to maximize cumulative reward. We analyze classifier accuracy, policy success, and reward-driven hypothesis selection.

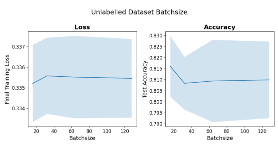
3.3. Results

In MONTEZUMA'S REVENGE, ensemble-based subgoal classifiers generalize better than single-head classifiers, achieving 70% accuracy on unseen ladder configurations.

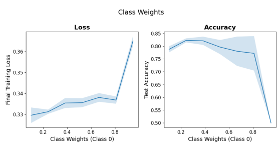
In MINIGRID, agents equipped with hypothesized subgoals solve the sparse-reward DOORMULTIKEY task with performance close to an oracle-defined agent. Reward-maximizing high-level policies consistently prefer ensemble members aligned with hand-specified subgoals (Shrikumar et al., 2017). In our real-world dataset, we were able to achieve over 80% accuracy using CLIP embeddings and over 70% accuracy using YOLO embeddings.

3.4. CLIP

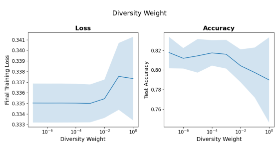
We conduct seven hyperparameter sweeps to evaluate the robustness of CLIP-based subgoal embeddings. Each subplot below visualizes one sweep.



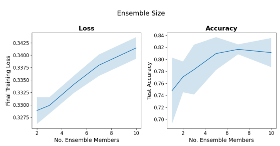
(a) Batch Sweep



(b) Class Weights Sweep

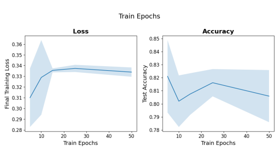


(c) Diversity Weight Sweep

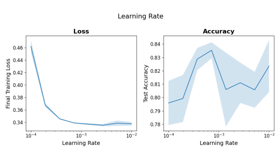


(d) Ensemble Size Sweep

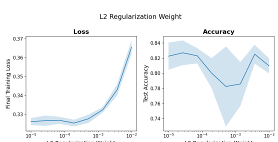
Figure 1. CLIP subgoal performance under varying batch sizes, class weights, diversity weights, and ensemble sizes.



(a) Epoch Sweep



(b) Learning Rate Sweep



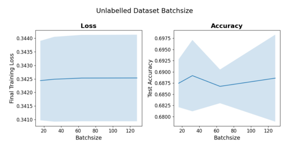
(c) L2 Regularization Weight Sweep

Figure 2. CLIP subgoal performance under varying number of epochs, learning rates, and L2 regularization weights.

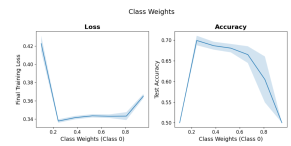
3.5. YOLO

In contrast to CLIP embeddings, which captures both image and text input, we also explored a YOLO ensemble as an alternative. At a high level, the YOLO (You Only Look Once) v5 model has been pretrained on a vast variety of different objects with the purpose of training for object detection. The YOLO ensemble consists of the YOLO embeddings from ultralytics/yolov5 and a stack of linear layers for each number of heads.

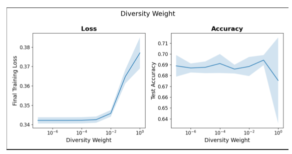
We also conducted the same hyperparameter sweeps to evaluate the robustness of our YOLO embeddings on our task; however, we found that the CLIP model generally performs better. As an example, we can see that for the learning rate sweep, the maximum accuracy using CLIP embeddings is nearly 84%, whereas the accuracy peaks at around 70% when using YOLO embeddings.



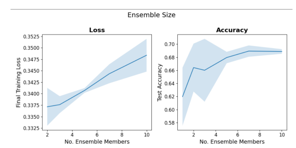
(a) Batch Sweep



(b) Class Weights Sweep

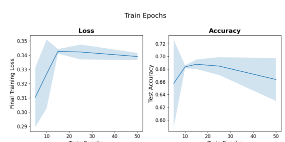


(c) Diversity Weight Sweep

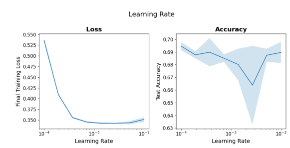


(d) Ensemble Size Sweep

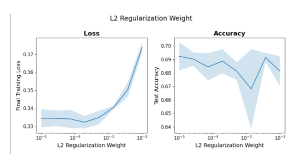
Figure 3. YOLO subgoal performance under varying batch sizes, class weights, diversity weights, and ensemble sizes.



(a) Epoch Sweep



(b) Learning Rate Sweep



(c) L2 Regularization Weight Sweep

Figure 4. YOLO subgoal performance under varying number of epochs, learning rates, and L2 regularization weights.

4. Conclusion

We propose a method for learning transferable subgoals in a single-task HRL setting by generating multiple hypotheses over generalizing features (Nair & et al., 2018). Our method outperforms single-classifier baselines and approaches oracle performance in sparse-reward tasks.

4.1. Future Work

Future directions include improving hypothesis generation efficiency (Gomez et al., 2022), applying the method to real-world robotic platforms (Konidaris & Barto, 2007), and integrating semantic priors into the subgoal classifiers to reduce reliance on reward feedback. Additionally, we plan on exploring other vision-based models such as a **Faster-RCNN** to see if the corresponding embeddings are more generalizable. Additionally, we are currently exploring another experiment involving the Towers of Hanoi problem to see if features such as each object's shape, size, or orientation can be learned and reused across new tasks.

4.2. Contributions

Our primary contribution in this project lies in testing the effectiveness of CLIP and YOLO embeddings as an additional layer. In addition, we collect thousands of images of different objects to test the robustness of our different classifiers.

4.3. Acknowledgement

We highly appreciate Professor George Konidaris, Anita, Bingnan, and Tuluhan for advising us and providing insightful feedback throughout the year.

References

Barreto, A. and et al. Transfer in deep reinforcement learning using successor features and generalized policy improvement. In *International Conference on Machine Learning*, pp. 501–510. PMLR, 2018.

Barreto, A. and et al. The option keyboard: Combining skills in reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Barto, A. G. and Mahadevan, S. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(1):41–77, 2003.

Bellemare, M. G. and et al. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Chevalier-Boisvert, M. and et al. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

Frans, K. and et al. Meta learning shared hierarchies. In *arXiv preprint arXiv:1710.09767*, 2017.

Gomez, D., Quijano, N., and Giraldo, J. Information optimization and transferable state abstractions in deep reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4782–4793, 2022.

Higgins, I. and et al. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, pp. 1480–1490. PMLR, 2017.

Khetarpal, K. and et al. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

Konidaris, G. D. and Barto, A. G. Building portable options: Skill transfer in reinforcement learning. In *IJCAI*, volume 7, pp. 895–900, 2007.

Machado, M. C. and et al. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. volume 61, pp. 523–562, 2018.

Nair, A. and et al. Visual reinforcement learning with imagined goals. In *Advances in neural information processing systems*, volume 31, 2018.

Nair, S., Savarese, S., and Finn, C. Goal-aware prediction: Learning to model what matters. In *International Conference on Machine Learning*, pp. 7207–7219. PMLR, 2020.

Pagliardini, M., Jaggi, M., Fleuret, F., and Karimireddy, S. P. Agree to disagree: Diversity through disagreement for better transferability. *arXiv preprint arXiv:2202.04414*, 2022.

Pateria, S. and et al. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.

Schulman, J. and et al. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2017.

Sutton, R. S., Precup, D., and Singh, S. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2): 181–211, 1999.

Van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Wang, L., Zhang, X., Su, H., and Zhu, J. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.