NATURAL LANGUAGE-BASED STATE REPRESENTATION IN DEEP REINFORCEMENT LEARNING

Md Masudur Rahman & Yexiang Xue
Department of Computer Science
Purdue University
West Lafayette, IN 47907, USA
{rahman64,yexiang}@purdue.edu

ABSTRACT

This study investigates the potential of using natural language descriptions as an alternative to direct image-based observations for learning policies in reinforcement learning. Due to the inherent challenges in managing image-based observations, which include abundant information and irrelevant features, we propose a method that compresses images into a natural language form for state representation. This approach allows better interpretability and leverages the processing capabilities of large language models (LLMs). We conducted several experiments involving tasks that required image-based observation. The results demonstrated that policies trained using natural language descriptions of images yield better generalization than those trained directly from images, emphasizing the potential of this approach in practical settings.

1 INTRODUCTION

Directly learning policies from images holds great promise for practical reinforcement learning applications. However, managing image-based observations is challenging due to their potential abundance of information and irrelevant features. Furthermore, the learned policy can often be a black box, as the action corresponding to an image observation is difficult to comprehend. This makes interpretability a challenge and the policies often fails to generalize to slightest changes to the environments. These can hinder the ability to leverage these policies in real-world tasks. On the other hand, language has been the primary mode of communication for humans. A situation can be precisely described through language, and conversely, a situation can be constructed from a language description. For instance, movies are often produced based on narratives found in books (e.g., Game of Thrones, Lord of the Rings). Ultimately, language is a major source through which humans reason and understand others' reasoning.

In reinforcement learning, the ability to learn a policy that generalizes well is essential for real-world system deployment. Specifically, agents should be adept at operating in scenarios distinct from their training environments. To address these challenges, several strategies have been proposed. These encompass data augmentation methods like random cropping and the addition of jitter to image-based observations Cobbe et al. (2019); Laskin et al. (2020b); Raileanu et al. (2020); Kostrikov et al. (2020); Laskin et al. (2020a), the injection of random noise Igl et al. (2019), network randomization Osband et al. (2018); Burda et al. (2018); Lee et al. (2020), and regularization techniques Cobbe et al. (2019); Kostrikov et al. (2020); Igl et al. (2019); Wang et al. (2020). These methods have consistently demonstrated their potential in boosting generalization. The core principle underlying these techniques is the amplification of training data diversity, which aids in crafting a more universally applicable policy. However, such perturbations are often introduced without due regard for task semantics. This oversight can modify critical observation elements, potentially diminishing the efficacy of policy learning.

Furthermore, random perturbations through various observation manipulations—such as cropping, blocking, or combining two random images from different environment levels—may yield unrealistic observations that the agent is unlikely to encounter during testing. Therefore, these techniques might underperform in settings where agents rely on realistic observations for policy learning. To

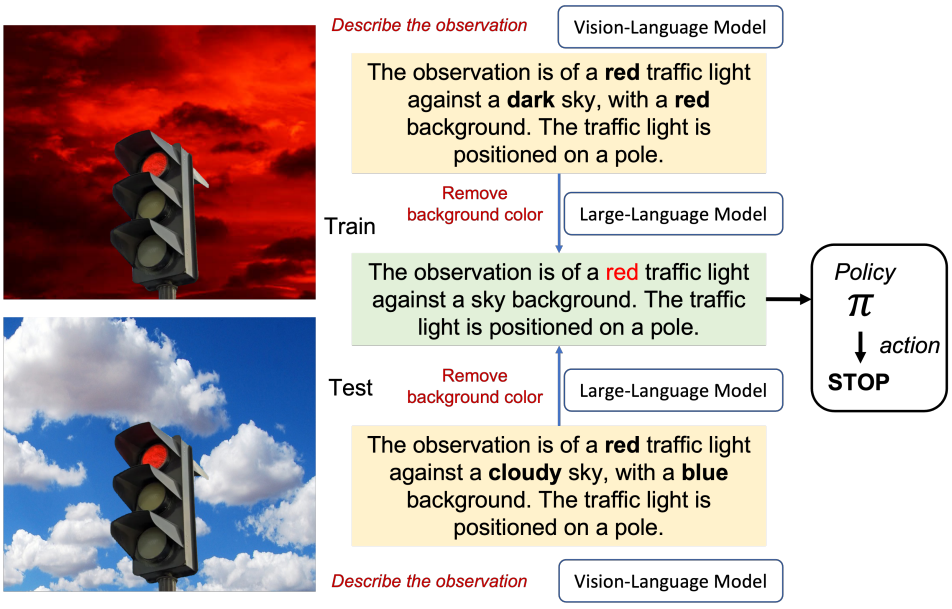

Figure 1: Example of Reinforcement Learning from Natural Language. Our method is to compress image pixels into natural language descriptions, serving as the state information of reinforcement learning. This language-based approach is advantageous as it is easy for humans to understand and provides a clearer insight into how the computer perceives visuals. Overall, our findings demonstrate that policies trained using natural language descriptions of images showcase enhanced generalization capabilities towards unobserved scenarios, surpassing the performance of policies directly trained from raw images.

circumvent this issue style transfer-based method has been proposed Rahman & Xue (2022) to mitigate the issue of spurious features and eventually improve generalization. This method is based on the assumption of style, which might not represent irrelevant information, for example, in cases where some aspect of color information might help learn a policy, such as red and green signals for a driving task. Nevertheless, all of these methods modify the image space, and the policy learning happens from pixel images. Thus, the learned policy can still be non-interpretable, and it is unclear how the policy behaves when a particular assumption, such as color information, is not held for a particular task, such as color information.

The autoencoder-based approach takes the image and represents it in a lower-dimensional space (e.g., AE, VAE) Ha & Schmidhuber (2018); Hafner et al. (2019); Zhang et al. (2022), which is then used as state information for reinforcement learning. However, such an approach can still suffer from the black-box policy issue, and the intermediate representation might lose information about the original image observation. All of these methods modify either the image space or the lower-dimensional projection where policy learning occurs. Consequently, the learned policy can remain non-interpretable, and it is unclear how the policy will behave when certain assumptions, such as color information, are not applicable to a specific task.

Overall, with the existing approaches, the resulting learned policy can be difficult to interpret and may fail when minor changes occur in the environment. Recent advancements in natural language processing and computer vision have enabled a more detailed, accurate understanding of image content. These advances are typically driven by large-scale models, often referred to as foundational models, which contain billions of parameters and are trained on internet-scale datasets with substantial computational resources.

In this paper, we primarily focus on decision making derived from language descriptions of visuals (e.g., images). We first compress the visual information (i.e., pixels) into natural language and use this language as state information to learn policy with reinforcement learning (Figure 2). This approach has several advantages. For instance, the language representation is inherently interpretable, and it provides a more accurate indication of what the agent understands from the visual scene. In

this setup, the agent can learn from a natural language description of the image. This approach provides multiple benefits. Primarily, the representation is easily interpretable by humans, unlike raw pixel data from the image. Moreover, it paves the way for harnessing the immense processing power of large language models (LLMs) to handle natural language state information. For instance, unnecessary features, such as color information, can be filtered out by directing the LLM to ignore them.rewrite the description excluding color information.

In particular, we utilize the Vision-Language Model (VLM) (i.e., LLAVA Liu et al. (2023)) to generate a natural language description of the image observation. The resulting language is then passed to a Large Language Model (LLM) (i.e., LLAMA Touvron et al. (2023)) for further pre-processing. Finally, it is converted into a text embedding vector using pre-trained embedding models (i.e., Sentence Transformer Reimers & Gurevych (2019)).

We conducted experiments to evaluate the effectiveness of a Vision and Language Model (VLM) in learning from text in reinforcement learning contexts. These experiments encompassed task that required image-based observation. Specifically, we conducted experiments on OpenAI Gym (Brockman et al., 2016) (Gymnasium Towers et al. (2023)) environment, FrozenLake. The rendered image was used as the observation, with the task being to learn a policy from this image observation. We compared our text-based learning approach with learning directly from the raw pixel information.

Our results indicated that policies trained using natural language descriptions of images exhibited superior generalization compared to those trained directly from images. Moreover, our language-based state representation is inherently interpretable compared to directly learning from pixels, indicating a strong use case for language-based state representation. In particular, in the Frozen Lake environment, training results show that all baselines learn the task efficiently, achieving optimal performance. However, this does not necessarily reflect the true efficacy of the policy, as it might be overfitting to the training data. When tested in a new ice environment variation, the PPO-Lang method maintains its performance, highlighting the strength of language-based learning. This robustness is attributed to the invariant language state information learned during training. Ensuring that the language state remains consistent and focuses on task-relevant details is crucial. One can use a Large Language Model (LLM) to filter out irrelevant information from the language input, creating a more invariant state for training. In contrast, policies based on image observations tend to overfit and fail to generalize in new environments, proving ineffective for the intended task.

2 Preliminaries and Problem Settings

Markov Decision Process (MDP). An MDP can be described by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$. Within this framework, an agent at a discrete timestep t interacts with its environment from a current state $s_t \in \mathcal{S}$, selecting an action $a_t \in \mathcal{A}$. Subsequently, the environment transitions to a new state $s_{t+1} \in \mathcal{S}$, governed by the transition probabilities $\mathcal{P}(s_{t+1}|s_t, a_t)$. The agent then receives a reward r_t, determined by the reward function $\mathcal{R}$.

Reinforcement Learning. Within the context of reinforcement learning, the agent operates within an MDP and aims to discover a policy $\pi \in \Pi$ that leads to the maximization of the cumulative reward. Here, Π represents the space of all feasible policies. Based on the current state, the agent selects an action in line with policy π, and the optimal policy $\pi^* \in \Pi$ is the one that yields the greatest total rewards over time. Extending beyond reinforcement learning methods, **Deep Reinforcement Learning** incorporates deep learning to handle more complex, high-dimensional input spaces. By utilizing deep neural networks, it can represent policy or value functions with greater flexibility and sophistication. DRL is suitable for applications that require processing raw pixel data or controlling intricate systems and has become instrumental in advancing various fields, from gaming to autonomous robotics. The integration of deep learning enables more precise function approximation, allowing agents to learn optimal policies in more challenging environments.

Generalization in Reinforcement Learning In the context of Reinforcement Learning, generalization refers to an agent's ability to apply learned knowledge from a specific set of environments to new, unseen environments. It assumes the presence of a fixed optimal policy, denoted as π^*, capable of achieving maximum return across all variations of the environments. These environments may vary in observational characteristics, such as having different background colors or other visual features. During training, the agent is exposed to a fixed set of environment variations to learn a policy.

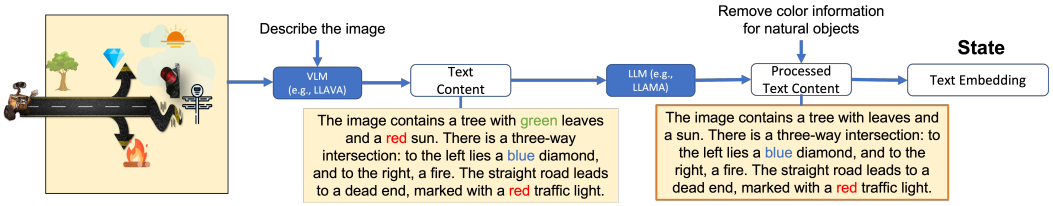

Figure 2: Pipeline for generating state from an image: Initially, a vision language model (VLM) is employed to create an image description. Subsequently, a language language model (LLM) refines this text, removing any spurious information related to the task at hand. The resulting textual content is utilized for state embedding, which ultimately serves as the observation for the agent within a reinforcement learning framework.

Subsequently, the agent's generalization performance is evaluated by testing it on previously unseen levels, measuring how well it can apply its learned policy to these new environments. This particular scenario is often referred to as a Contextual MDP Kirk et al. (2021).

Reinforcement Learning from Images In this setup, agents are trained to make decisions directly from raw visual data, like images. It enables agents to learn patterns and relationships from the visuals, making it suitable for real-world applications with complex visual information. This approach has succeeded in various domains, such as robotics, autonomous vehicles, and video games. It has promise for building intelligent agents capable of learning directly from raw visual input. However, dealing with high-dimensional visual data and extracting relevant features demand computationally efficient algorithms, enabling agents to learn and act in complex environments. Additionally, handling irrelevant features in images is vital as it can confound with the reward, which leads to a sub-optimal policy.

3 METHOD

This subsequent method section provides a detailed breakdown of our approach to producing natural language descriptions from visual data, specifically images. The systematic process consists of several stages that we will outline step-by-step for clarity. A pivotal component in this strategy is the Vision-Language Model (VLM). Notably, we employ the VLM variant known as LLAVA Liu et al. (2023). The LLAVA VLM initially processes the input image when employed in our method. During this phase, the model identifies and understands various features and objects contained within the visual data. Upon extracting and understanding these elements, the VLM subsequently crafts a well-structured, coherent, and descriptive narration in natural language form. The resultant description is not only articulate but also pinpoints the image's most prominent and defining characteristics, ensuring that readers or users receive an accurate and detailed understanding of the visual content.

After extracting the natural language description from the Vision-Language Model (VLM), our method allows pre-processing text with Large Language Model (LLM) (e.g., LLAMA Touvron et al. (2023), ChatGPT). When the language description, as generated by the VLM, is inputted into the LLM model, the latter performs complicated processing tasks. These tasks primarily aim at enhancing the language output by refining its structure, improving its coherence, and bolstering its information content. The result of this is a description that is not just technically accurate but also informative in context.

Furthermore, in scenarios where generalization is the goal, this pre-processing step undertaken by the LLM is valuable. The model inspects the description to identify and eliminate superfluous or irrelevant details. Such action is required, especially when we consider the need for an agent to develop a consistent and invariant representation of an image observation within a given environment. By removing unnecessary details, we ensure that the agent focuses only on the most vital aspects of the environment, thereby optimizing its learning process.

After processing the natural language description, the subsequent step in our pipeline focuses on text representation through embedding. For this purpose, we utilize pre-trained models, specifically the Sentence Transformer architecture, as outlined by Reimers et al. in 2019 Reimers & Gurevych

(2019). The Sentence Transformer is designed to convert textual data into dense vectors of fixed dimensions, known as text embeddings. The primary objective of these embeddings is to encapsulate the semantic information and context inherent in the original text. By converting the refined natural language description into this vector format, we aim for an efficient representation for computational processing and to maintain the semantic properties of the input data.

By integrating a sequence of models— the Vision-Language Model (VLM) for initial image description generation, the Large Language Model (LLM) for subsequent description refinement, and the Sentence Transformer for text embedding transformation, we have developed a methodology that efficiently extracts and represents pertinent information from images in a structured, semantic format. This systematic approach facilitates a cohesive fusion of visual and textual data.

Detailed Process Description

Figure 2 presents an overview of our methodology's pipeline. We delve into the details here.

1. Image Description Generation Using Vision-Language Model (VLM)

a. Preprocessing: Input images undergo a preprocessing phase wherein standard image transformations, including resizing, normalization, and data augmentation, are executed to render them compatible with the VLM.

b. Vision-Language Model (LLAVA): Our choice of VLM for this procedure is LLAVA Liu et al. (2023). LLAVA is a comprehensive end-to-end multimodal model. By seamlessly integrating a vision encoder with a Language Model (LLM), LLAVA provides holistic understanding of both visual and linguistic modalities. The model is harnessed to produce textual descriptions from image-based observations from a prompt (e.g., describe the observation).

2. Language Description Refinement Using Large Language Model (LLM)

a. Pre-processing with LLM: The process begins with descriptions generated by LLAVA. These initial descriptions are then subjected to a refinement process using the capabilities of the Large Language Model. This refinement stage aims to improve the descriptions' quality, accuracy, and coherence. Various text manipulations can be executed using the Large Language Model by employing carefully crafted prompts. These manipulations include tasks like paraphrasing, summarizing, translating, and generating alternative versions of the descriptions. The flexibility and versatility of the model enable it to handle various text-related tasks, providing an efficient and effective means of refining and enhancing the descriptions derived from LLAVA.

b. Generalization: In specific scenarios, such as when training agents to operate within dynamic environments, the need arises to strategically abstract or exclude unnecessary details from the descriptions. This process of generalization is crucial as it guarantees that the agents attain a uniform and streamlined understanding of the environment. By doing so, the risk of the agents becoming overly tailored to specific observations is minimized, helping them to avoid overfitting and ensuring a more adaptable and versatile performance in varying situations.

3. Conversion to Text Embeddings

a. Sentence Transformer: After the refinement process, the descriptions transform into fixed-dimensional vectors through the use of the Sentence Transformer Reimers & Gurevych (2019). This model excels at converting sentences into fixed-sized dense vector representations, effectively encapsulating the semantic significance and contextual nuances inherent within the text. The resulting fixed-size vectors are essential, particularly in their seamless integration into contemporary Reinforcement Learning (RL) algorithms adhering to the standard Markov Decision Process (MDP) framework. This transformation presents a structured and compact format for the descriptions, facilitating downstream tasks.

b. Text Embeddings: The vector embeddings generated by the Sentence Transformer intricately encapsulate the semantic intricacies that interlace words and constructs within the language descriptions. These succinct yet information-rich representations hold immense value for various subsequent tasks, whether it involves gauging similarities between descriptions or seamlessly integrating them into reinforcement learning frameworks.

4 EXPERIMENTS

4.1 SETUP

Environments: We experiment with the FrozenLake environment, which is available in OpenAI Gym Brockman et al. (2016) and further detailed in Gymnasium Towers et al. (2023).

FrozenLake Description: In the FrozenLake scenario (Figure 3), an agent is situated on a grid representing a frozen lake. The task for the agent is to traverse from its initial position, typically at the top-left corner, to its goal, generally at the bottom-right corner, all the while evading pitfalls in the ice. This grid contains distinct cells: frozen tiles (F), holes (H), the starting point (S), and the ultimate goal (G). Available actions to the agent encompass moves in the four cardinal directions: up, down, left, and right.

Modifications for our Experiments: Diverging from the default library setup, which provides *true state* information, our implementation offers the agent an *RGB image* of the grid world as its observation. To infuse variability, we experiment with assorted ice colors in the environment, such as the default sky blue, a more profound dark blue, and a textured variation.

Evaluation Metric: The core of our experiment centers around determining the agent's capacity to derive a strategy in one version of the environment and effectively apply this acquired knowledge in a different and unfamiliar variation. Thus, our training phase engages the agent with the default environment setup, post which it undergoes evaluation in an unseen environment variant. The computed reward over a episode is defined as the *episodic return*. We distinguish between the training phase's reward performance, termed *train episodic return*, and the performance in the evaluated variant, termed *test episodic return*.

Implementation Details

Base Algorithm: Our Proximal Policy Optimization (PPO) implementation draws inspiration from the CleanRL Library Huang et al. (2022a;b). It integrates numerous pivotal modifications for improved performance from contemporary research in policy gradient techniques. These modifications include the Normalization of Advantage, Orthogonal Initialization, and Generalized Advantage Estimation (GAE). For a comprehensive understanding of these aspects, readers can refer to Huang et al. (2022a).

Hyperparameters: To maintain consistency across our experiments, the hyperparameters of the base PPO algorithm remain unaltered. We adopt these hyperparameters grounded on the established standards delineated in the PPO's continuous action space implementations Huang et al. (2022a;b).

Figure 3: Frozen Lake Environment.

Handling RGB Images: For the challenge of learning from RGB images, we employ a three-layer convolutional neural network with ReLU activations, a configuration inspired by the PPO implementation for Atari in the CleanRL library.

Handling Text Embeddings: Text embeddings are crucial for representing textual data's structured and semantic meaning in our experiments. We use a transformer-based model to convert natural language descriptions into dense vector representations. These embeddings serve as additional inputs to our agent, complementing the RGB images and providing the agent with a richer understanding of the environment.

Reproducibility: Ensuring our work contributes to the larger academic community, we will open-source the complete implementation, including hyperparameters and tracking of our experiments,

to aid future research and reproducibility. Unless otherwise mentioned, the results are shown with three random seed runs.

PPO-Image: This baseline uses the standard Proximal Policy Optimization algorithm with RGB images of the environment as observations. The agent's policy is trained directly on the visual input, capturing features like the grid configuration and the agent's current position. It operates in a more conventional approach by directly processing the pixel values of the images.

PPO-Lang: In this version, the environment provides a natural language description of the state instead of an image. As discussed in the method sections, pre-trained models convert this textual information into embeddings. The agent's policy is trained on these embeddings, offering a high-level, abstract view of the environment. This method aims to capture the semantic information in the descriptions, making it potentially more generalizable across different variations of the environment.

The hyperparameters remain consistent for both implementations, except for the input layer accommodating images or text embeddings, ensuring a fair comparison. Through our experiments, we aim to demonstrate that PPO-Lang can achieve comparable or better performance than the PPO-Image, especially in environments where language can provide a richer and more generalizable representation of the state.approach (*PPO-Lang*).

4.2 RESULTS

In our experiments with the Frozen Lake environment, as depicted in Figure 4, all the baselines quickly converge to optimal training performance, consistently achieving a score of 1.0. Nevertheless, this training efficacy can be misleading; high training scores might obscure a model's potential to overfit its training data, leading to suboptimal performance in novel environments or unseen scenarios.

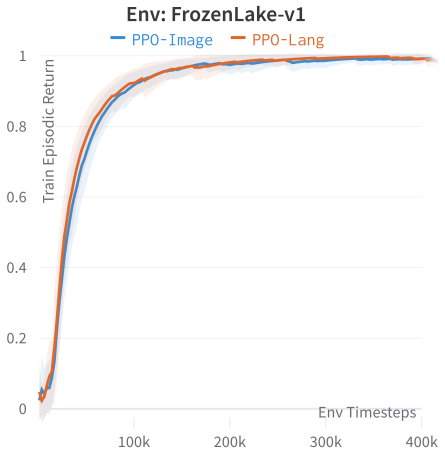

Figure 4: Train Results. Experiments on the Frozen Lake environment. While all baselines converge rapidly to an optimal training performance with scores of 1.0, such results can be deceptive. Our PPO-Lang method, grounded in language-based representations, showcases the potential for consistent performance in training time, highlighting its effectiveness in learning compared to image-centric models in new environments.

To better understand the generalization capability of our models, we transition to testing our policies in a variant of the ice environment not exposed to the model during training. PPO-Lang, our proposed method, exhibits commendable performance consistency, as seen in Figure 5. This consistency underscores the advantages of grounding reinforcement learning in language-based representations. One attributing factor to this stability is the incorporation of invariant linguistic states during the policy learning process. Ensuring this invariance, especially against non-essential environmental nuances, is paramount. In practical terms, this translates to crafting queries for the language model that hone in on consistent, task-centric details. In cases where the linguistic input might carry extraneous information, leveraging a Large Language Model (LLM) can be beneficial for removing these distractions, leaving behind a purified, invariant state representation for training.

Contrastingly, a policy that leans on image observations as their primary source of information (PPO-Image) fails to manifest any significant performance in our test environment. Such a stark discrepancy in outcomes reinforces the inherent challenge with image-centric models: their tendency to overfit to visual features of their training environments. This tendency compromises their ability to generalize, rendering them ineffective in adapting to and learning within new or modified environments.

Implications From our empirical evaluation within the Frozen Lake environment, several salient insights emerge that hold significance for the domain of reinforcement learning: In Figure 4, we see that training performance does not always indicate a model's generalization capacity. An algorithm might exhibit optimal behavior during training, but this does not guarantee its efficacy in previously unseen conditions or variations of the environment.

Our observations from Figure 5 suggest that leveraging linguistic information during training can potentially bolster a model's robustness to novel scenarios, which can be attributed to the abstraction capabilities inherent in language-based representations. Such representations capture the essence of a situation without getting entangled in the specifics, analogous to employing high-level heuristics instead of detailed mappings.

Conversely, visual-centric models, although rich in representational content, may run the risk of overfitting the training data. Overfitting occurs when a model becomes excessively tailored to the training dataset, compromising its ability to generalize to new data, which is analogous to a system that excels in memorizing a dataset but fails in extracting and applying the underlying patterns to fresh, unseen data.

In summary, while visual data offers a granular view of the environment, linguistic information provides a more abstract, generalized perspective. The trade-off between specificity and generalization is pivotal in reinforcement learning model design and training.

5 Related Work

The integration of language with reinforcement learning has been a subject of growing interest in the research community. Language, being one of the most remarkable human achievements, plays a pivotal role in our ability to learn, teach, reason, and interact with others. However, the current state-of-the-art reinforcement learning agents have shown limitations in understanding or utilizing human language. The potential benefits of integrating language with reinforcement learning are manifold. Agents that can harness language in conjunction with rewards and demonstrations have the potential to enhance their generalization capabilities, scope, and sample efficiency. The practical implications of such integration are vast. For instance, agents that can transfer domain knowledge from textual data might be more efficient in exploring given environments or performing zero-shot learning in new settings. Moreover, many real-world applications, such as personal assistants and household robots, inherently require language processing to interact with humans or to utilize existing interfaces. The linking Language to Actions and Observations has been explored with methods aiming to effectively associate language with actions and observations in various environments Branavan et al. (2009); Tellex et al. (2011); Chen & Mooney (2011).

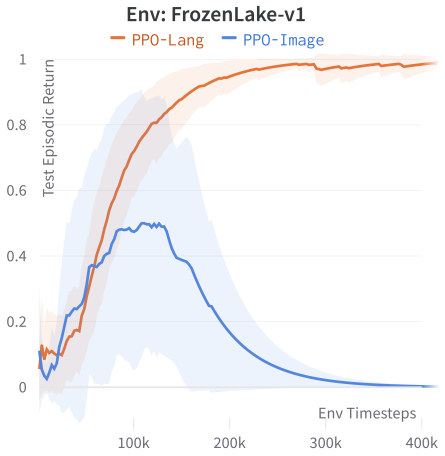

Figure 5: Test (Generalization) Results. Testing performance in a variant of the Frozen Lake environment. PPO-Lang, our language-based method, demonstrates consistent performance, underscoring the advantages of language-grounded reinforcement learning. In contrast, image-centric models, such as PPO-Image, struggle to adapt, highlighting their susceptibility to overfitting to specific visual features of their training environments.

From a generalization perspective, various strategies have been proposed; these include data augmentation techniques such as random cropping and noise addition, as well as network randomization to augment training diversity Cobbe et al. (2019); Laskin et al. (2020b); Raileanu et al. (2020); Kostrikov et al. (2020); Laskin et al. (2020a); Osband et al. (2018); Burda et al. (2018); Lee et al. (2020); Cobbe et al. (2019); Kostrikov et al. (2020); Igl et al. (2019); Wang et al. (2020). However,

the effectiveness of these methods can diminish if the semantics of the task are overlooked. Random manipulations, like cropping or merging images, can lead to unrealistic observations during testing, which can adversely affect performance. A style transfer-based method Rahman & Xue (2022) has been proposed to mitigate spurious features and enhance generalization by leveraging style assumptions. However, these methods often lack interpretability and can fail when certain assumptions, such as color information, do not hold. The autoencoder-based method, which reduces images into a lower-dimensional space Ha & Schmidhuber (2018); Hafner et al. (2019); Zhang et al. (2022), can also face challenges like unclear policy behavior due to its black-box nature and the potential loss of original image information in the intermediate representation. These methods modify the image space or the lower-dimensional projection where policy learning occurs, which can result in non-interpretable policies.

While the integration of language with reinforcement learning has been a topic of interest, the majority of existing research has primarily focused on direct associations between language and actions or observations. These methods often rely on data augmentation techniques, network randomization, or image manipulations to enhance generalization. However, these strategies can sometimes lead to unrealistic outcomes during testing or lack interpretability, especially when certain assumptions do not hold true. For instance, autoencoder-based methods, which condense images into a lower-dimensional space, might grapple with ambiguous policy behavior due to their opaque nature. Such methods can also risk losing essential image information, leading to policies that are hard to interpret.

In contrast, our work takes a fundamentally different approach. We emphasize decision-making through language descriptions of visual content. Instead of relying heavily on visual cues, which can be susceptible to overfitting or misinterpretation, our method harnesses the power of language to provide a more robust and transparent representation. This language-centered approach addresses the challenges inherent in image-based policy learning and offers a more interpretable and generalizable solution. By grounding reinforcement learning in linguistic descriptions, we aim to create models better equipped to handle diverse scenarios, ensuring that they memorize training data and genuinely understand it. This focus on language as a primary source of information sets our work apart, offering a novel perspective in reinforcement learning.

6 Conclusion

In reinforcement learning, while directly learning policies from images offers potential, it also presents challenges due to the abundance of information and irrelevant features in image-based observations. Such policies often lack interpretability and struggle to generalize across varying environments. Language, a primary mode of human communication, offers a precise way to describe and construct situations, serving as a foundation for human reasoning. We introduces a novel approach that leverages language descriptions of visuals for decision-making. By converting visual information into natural language and using this language as state information, the resulting policy is more interpretable and offers a clearer insight into the agent's understanding of the visual scene. Utilizing Vision-Language Models and Large Language Models, we present a method to generate natural language descriptions of image observations, preprocess them, and convert them into text embedding vectors. Experiments conducted on OpenAI Gym Frozen Lake environment, demonstrate the superiority of policies trained using natural language descriptions over those trained directly from images. Such language-based state representations offer enhanced interpretability and generalization, underscoring the potential of language as a powerful tool in reinforcement learning.

References

Satchuthananthavale RK Branavan, Harr Chen, Luke Zettlemoyer, and Regina Barzilay. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 82–90, 2009.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

David Chen and Raymond Mooney. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 859–865, 2011.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, pp. 1282–1289. PMLR, 2019.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2019.

Shengyi Huang, Rousslan Fernand Julien Dossa, Antonin Raffin, Anssi Kanervisto, and Weixun Wang. The 37 implementation details of proximal policy optimization. In *ICLR Blog Track*, 2022a. URL https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/. https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022b. URL http://jmlr.org/papers/v23/21-1342.html.

Maximilian Igl, Kamil Ciosek, Yingzhen Li, Sebastian Tschiatschek, Cheng Zhang, Sam Devlin, and Katja Hofmann. Generalization in reinforcement learning with selective noise injection and information bottleneck. In *Advances in neural information processing systems*, pp. 13978–13990, 2019.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of generalisation in deep reinforcement learning. *arXiv preprint arXiv:2111.09794*, 2021.

Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pp. 5639–5650. PMLR, 2020a.

Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in Neural Information Processing Systems*, 33, 2020b.

Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. Network randomization: A simple technique for generalization in deep reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgcvJBFvB.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 8617–8629, 2018.

Md Masudur Rahman and Yexiang Xue. Bootstrap state representation using style transfer for better generalization in deep reinforcement learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2022)*, 2022.

Roberta Raileanu, Max Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. Automatic data augmentation for generalization in deep reinforcement learning. *arXiv preprint arXiv:2006.12862*, 2020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 1507–1514, 2011.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Mark Towers, Jordan K. Terry, Ariel Kwiatkowski, John U. Balis, Gianluca de Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimmel, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Tan Jin Shen, and Omar G. Younis. Gymnasium, March 2023. URL https://zenodo.org/record/8127025.

Kaixin Wang, Bingyi Kang, Jie Shao, and Jiashi Feng. Improving generalization in reinforcement learning with mixture regularization. *arXiv preprint arXiv:2010.10814*, 2020.

Edwin Zhang, Yujie Lu, William Yang Wang, and Amy Zhang. Lad: Language augmented diffusion for reinforcement learning. In *Second Workshop on Language and Reinforcement Learning*, 2022.