Language-Guided Reinforcement Learning for Hard Attention in Few-Shot Learning

Bahareh Nikpour, Narges Armanfard
Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada
Mila- Quebec AI Institute, Montreal, QC, Canada

Scan Here!

Motivation

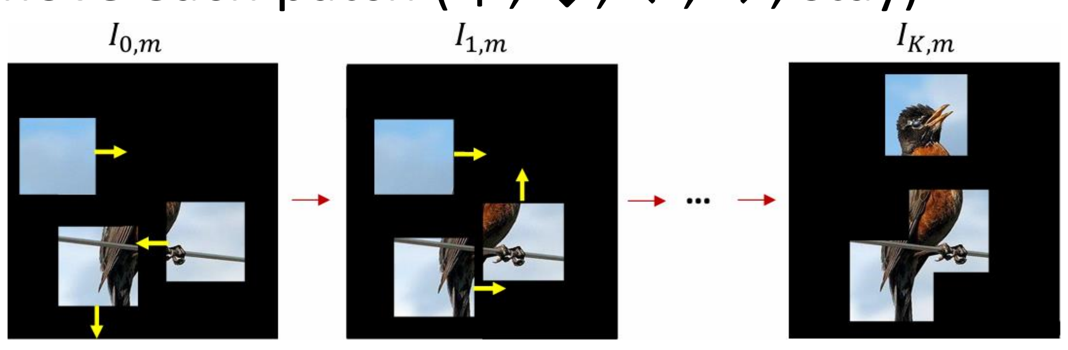
- Few-shot learning struggles with small datasets and noisy or irrelevant image regions.
- While **soft attention** is easily trainable, it spreads focus broadly.
- **Hard attention** can isolate the most informative parts, but it is **non-differentiable** and rarely explored in few-shot settings.
- LaHA tackles this gap by combining **reinforcement learning** and **language-guided rewards** to learn interpretable, task-specific attention.

Core Idea and Framework Overview

- LaHA formulates hard attention selection as a Markov Decision Process solved by reinforcement learning.
- A Vision Transformer agent sequentially selects a few (Na = 3) image patches most relevant for classification.
- LaHA has three main modules:

1- RL Module → For selecting the important regions

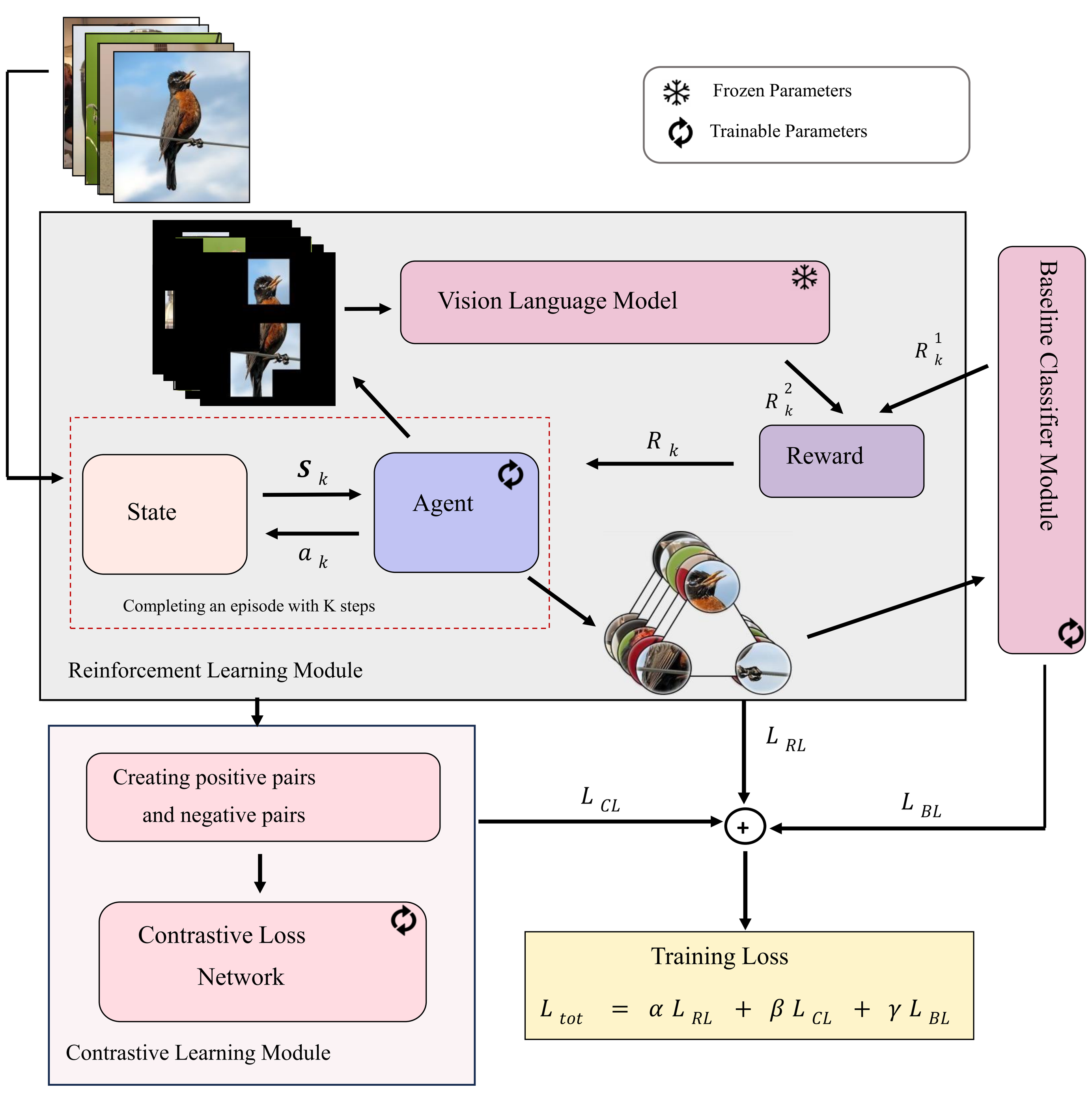
State: concatenation of the original image + previously selected patches
Action: move each patch (↑, ↓, ←, →, stay)



Reward: $R1$ → classification improvement from the few-shot baseline, $R2$ → semantic consistency with image captions from a Vision-Language Model (VLM)
Optimization: Proximal Policy Optimization (PPO)

2- Baseline Module → For few-shot classification using the selected regions, uses a Graph Neural Network on patch relations

3- Contrastive Learning Module → To improve the training, representation and generalization



$$L_{CL} = \sum_{i \in W} \frac{-1}{|O(i)|} \sum_{o \in O(i)} log \frac{\exp(f_i \cdot f_o / \tau)}{\sum_{u \in U(i)} \exp(f_i \cdot f_u / \tau)}$$

Total Loss to minimize:

$$L_{tot} = \alpha L_{RL} + \beta L_{CL} + \gamma L_{BL}$$

Experiments and Results

Performance

- Outperforms all state-of-the-art FSL baselines (ProtoLP, HyperShot, LaplacianShot)
- Reduces processed image area to ≈ 1/3 of original → lower memory + faster inference
- Patch selections align with human perception per VLM captions

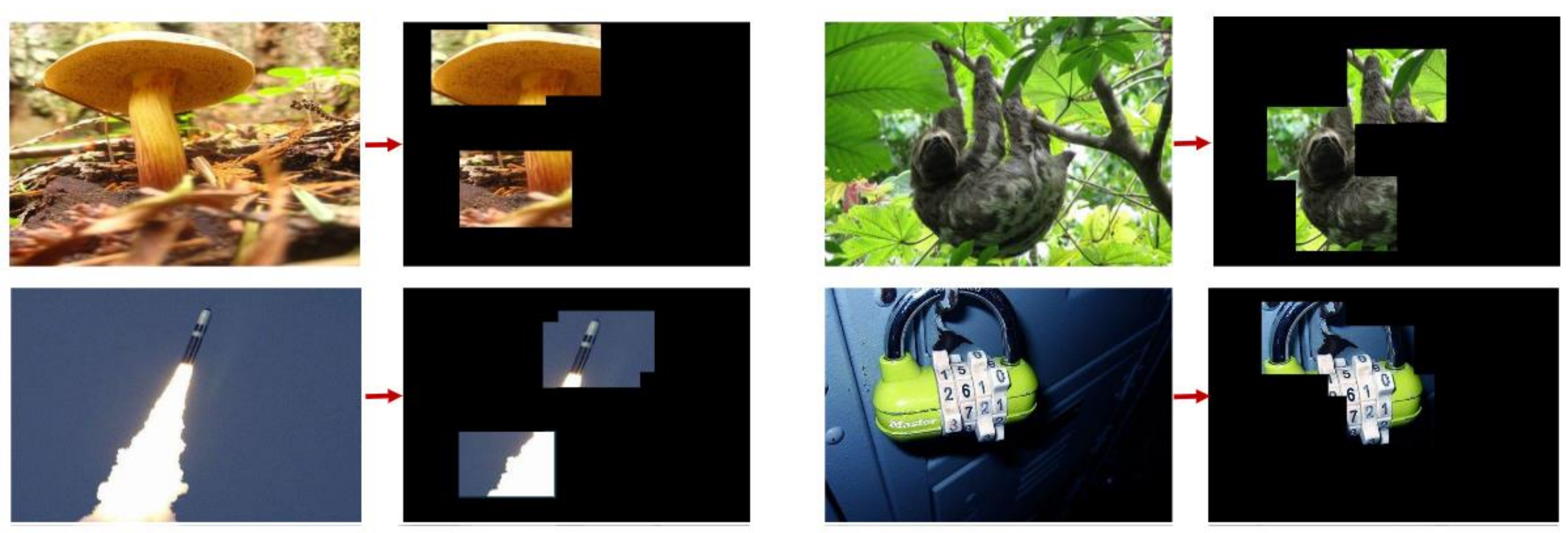
Ablations

- Without RL: large drop in accuracy → RL is main driver.
- Without VLM or Contrastive Learning: performance decreases; both improve interpretability and generalization.

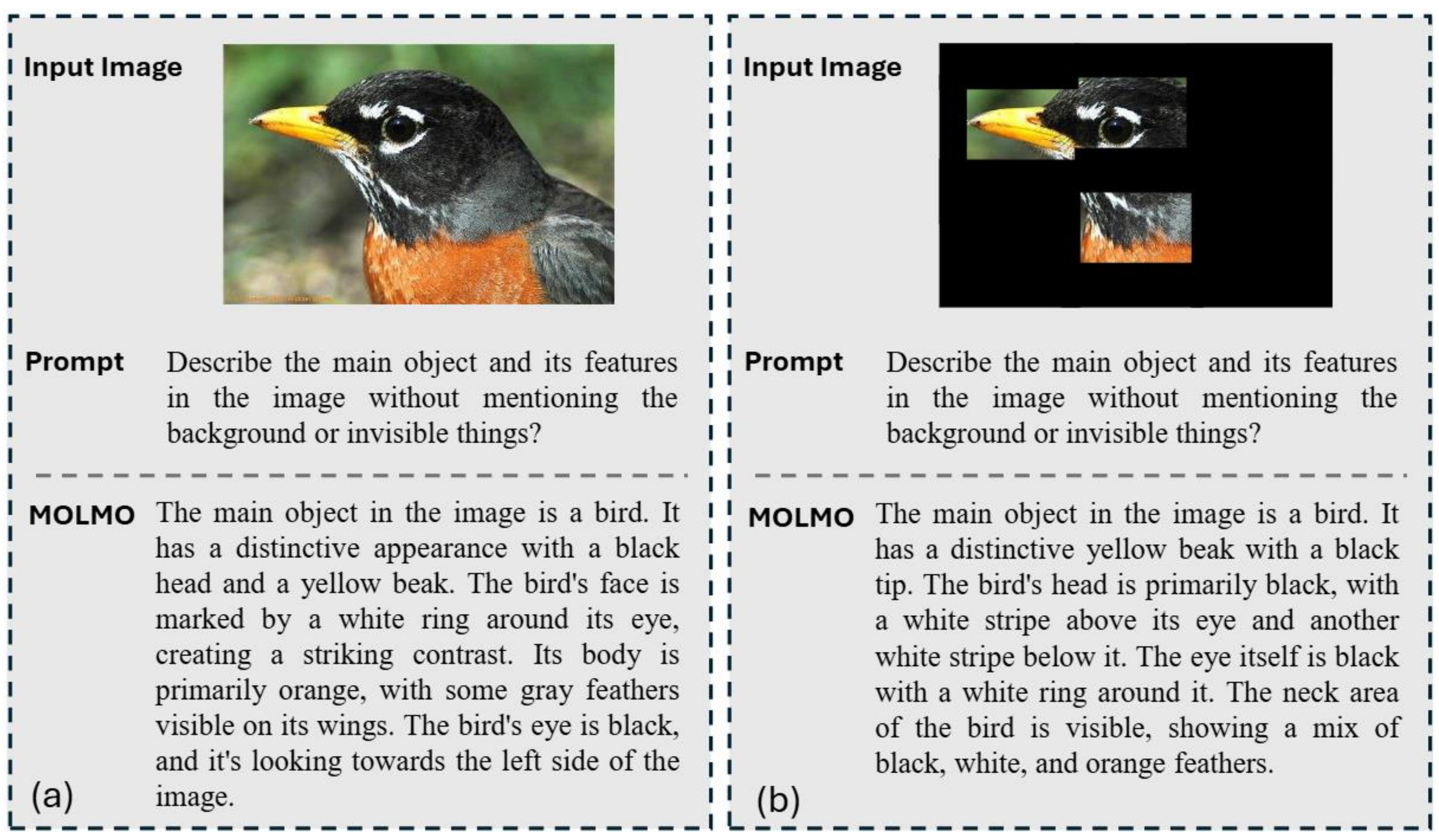
Beyond Few-Shot Learning

- When applied to ImageNet classification, LaHA surpasses prior hard-attention methods (Saccader, DRAM, TNet), showing its versatility beyond few-shot tasks.

Visualization of the Selected Patches



The output of VLM before and after applying LaHA



Conclusion

- First hard-attention framework directly on RGB images for FSL.
- Language-guided RL → interpretable and semantically consistent.
- Graph representation of patches preserves spatial/contextual relations.
- Efficient & edge-friendly → fewer computations, strong performance, high interpretability.