1703.01161 - FeUdal Networks for Hierarchical Reinforcement Learning

- **Yunqiu Xu**

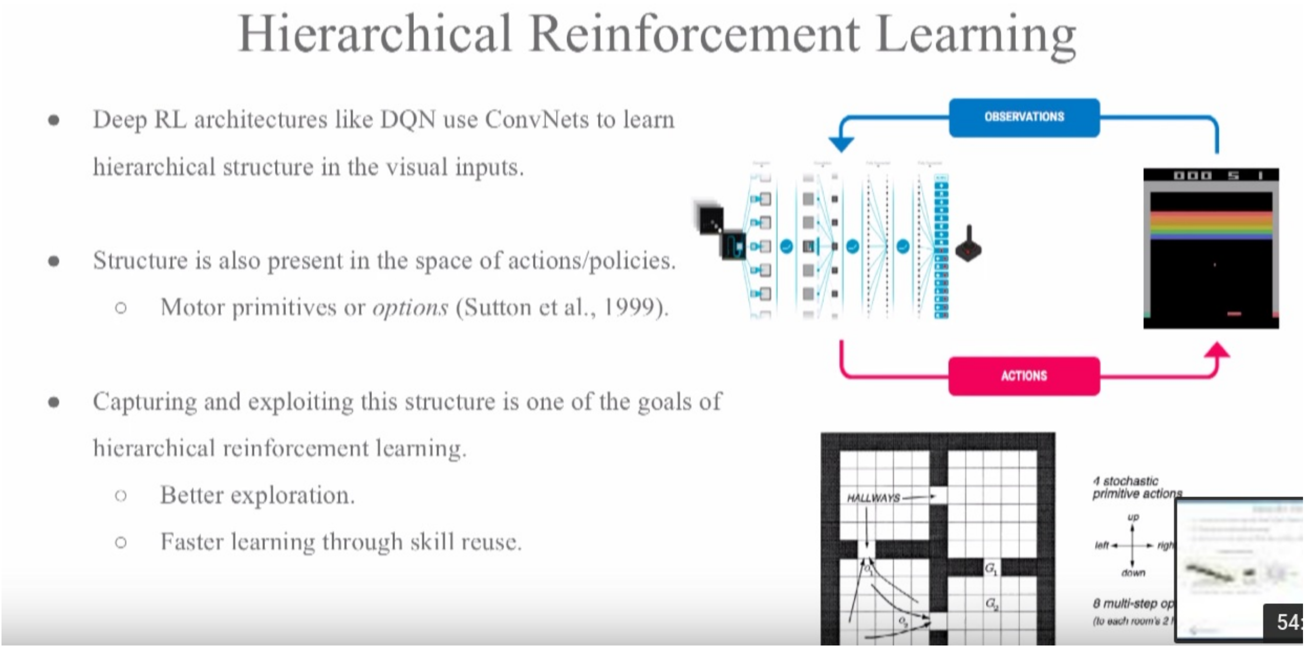
---

1. Introduction

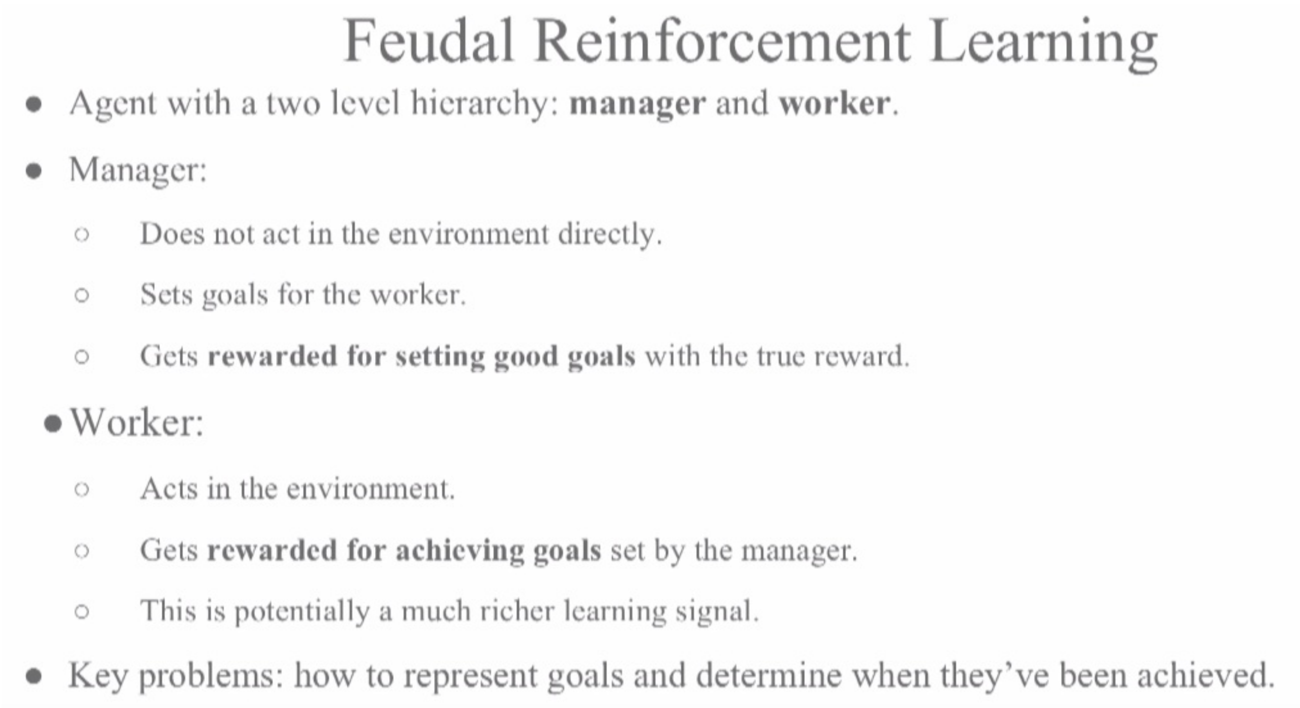
- Challenges:
  - Long-term credit assignment
  - Sparse reward: another solution can be found in 1707.05300 - Reverse Curriculum Generation for Reinforcement Learning
- Our work
  - Get insight from Feudal reinforcement learning (1993) , generalize its principle
  - End-to-end differentiable neural network with two levels of hierarchy: Manager and Worker
  - Manager: operates at a lower temporal resolution, produces a meaningful and explicit goal for Worker to achieve
  - Worker: follow the goals by an intrinsic reward
  - No gradients are propagated between Manager and Worker $\rightarrow$ Manager receives learning signal from the environment alone
  - Worker tries to maximise intrinsic reward and Manager tries to maximise extrinsic reward
- Advantage:
  - Facilitate very long timescale credit assignment
  - Encourage the emergence of sub-policies associated with different goals set by the Manager

2. Related Work

- Hierarchical RL:



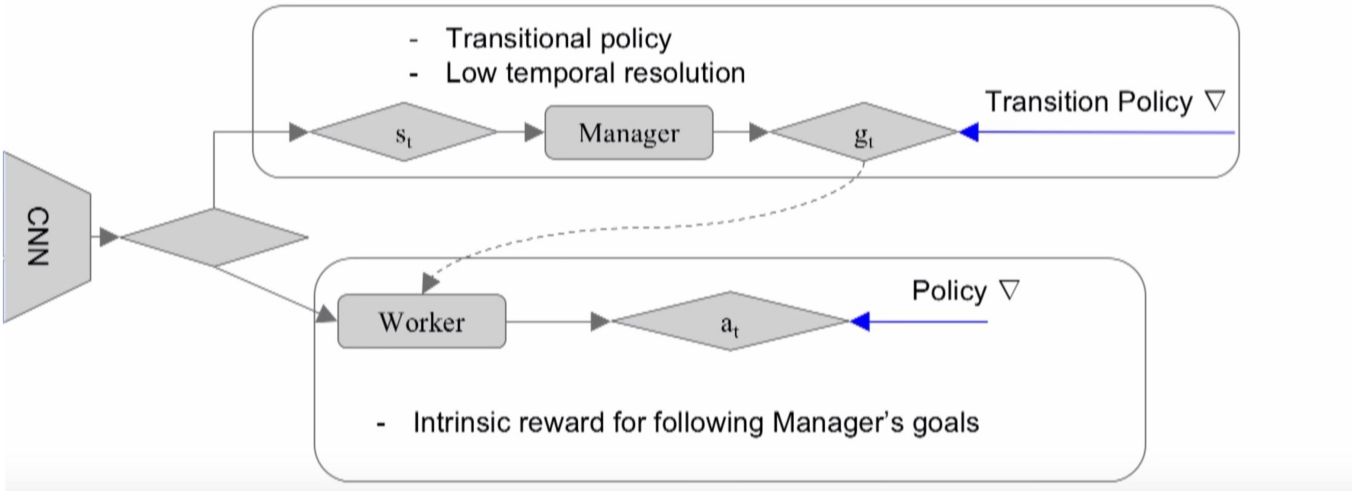
- Feudal RL by Dayan and Hinton, 1993: treat Worker as sub-policy



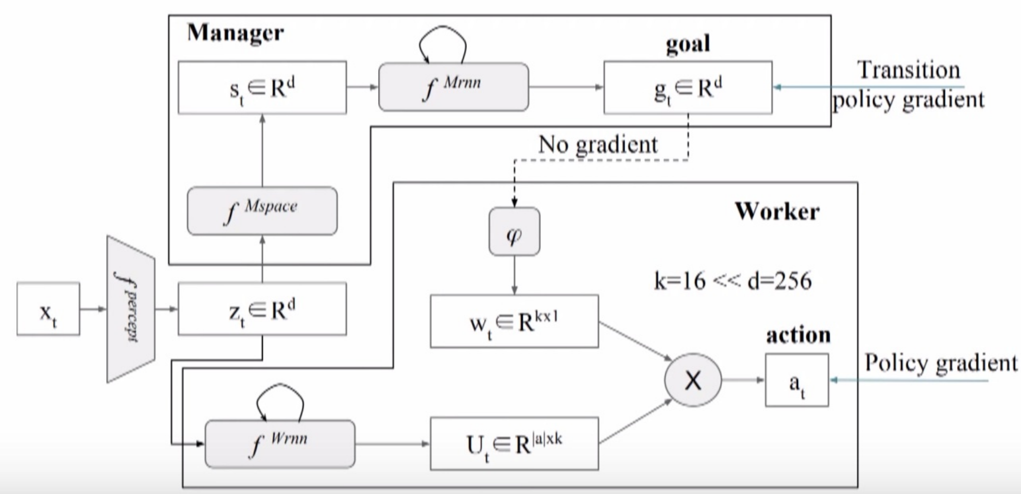
- Combine DL with predefined sub-goals:
  - 1604.07255 - A Deep Hierarchical Approach to Lifelong Learning in Minecraft
  - 1604.06057 - Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation
  - However sub-goal discovery was not addressed
- Some non-hierarchical state-of-the-art on Montezuma’ s Revenge: orthogonal to H-DRL, can be combined together

- [1606.01868 - Unifying Count-Based Exploration and Intrinsic Motivation]
- [1611.05397 - Reinforcement Learning with Unsupervised Auxiliary Tasks]

3. The Model

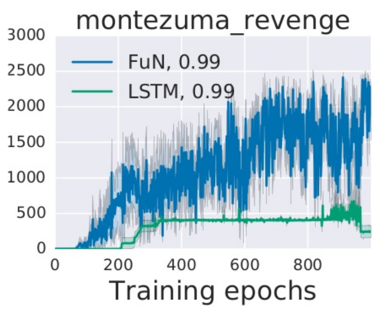


- Key idea - **represent goals in a shared feature space.**

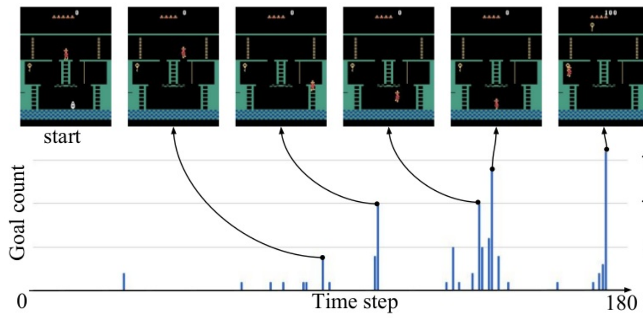


4. Experiment

- Montezuma's Revenge



(a)



(b)

- Other Atari Games

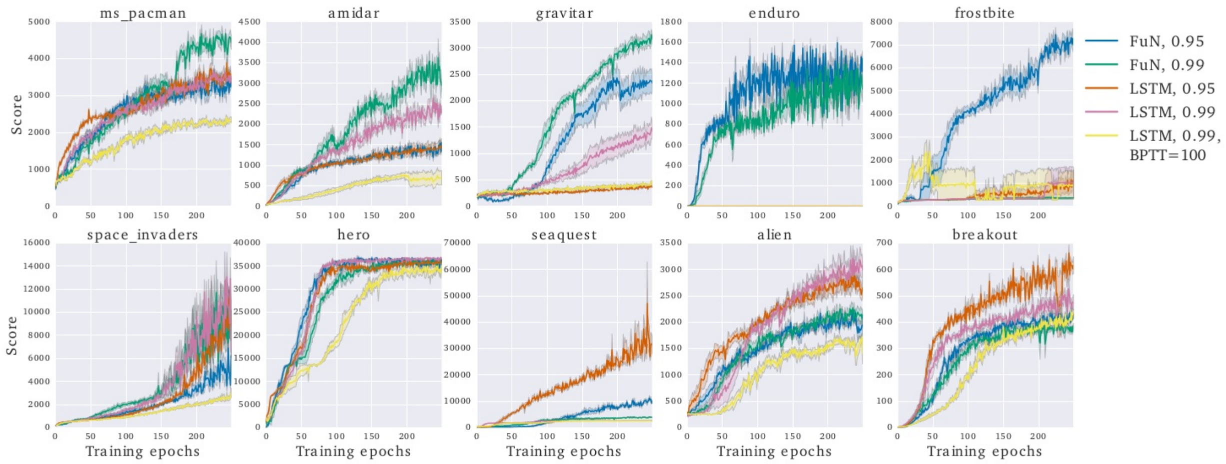
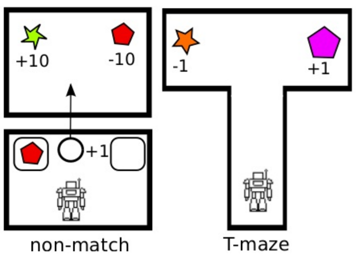
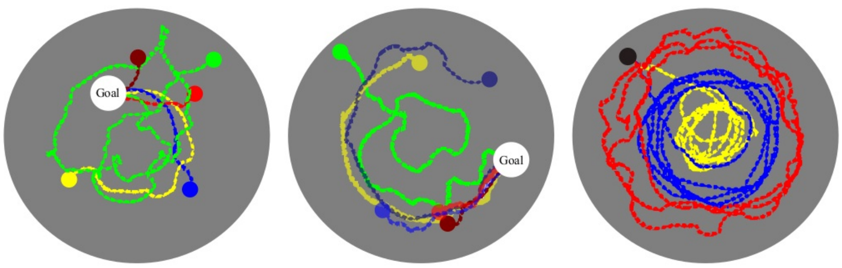


Figure 4. ATARI training curves. Epochs corresponds to a million training steps of an agent. The value is the average per episode score of top 5 agents, according to the final score. We used two different discount factors 0.95 and 0.99.

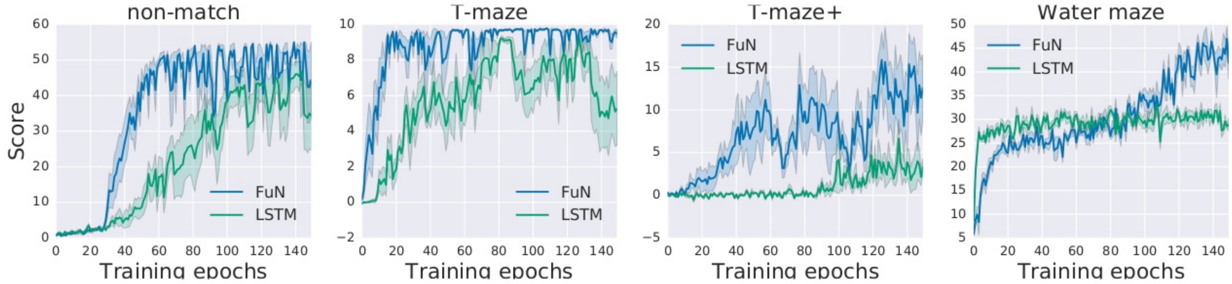
- Memory Tasks on Labyrinth



(a)



(b)



- Ablative Analysis

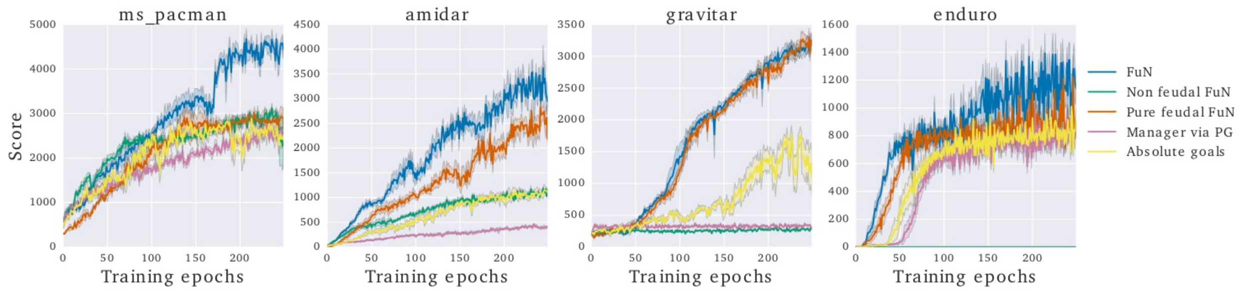


Figure 8. Ablative analysis

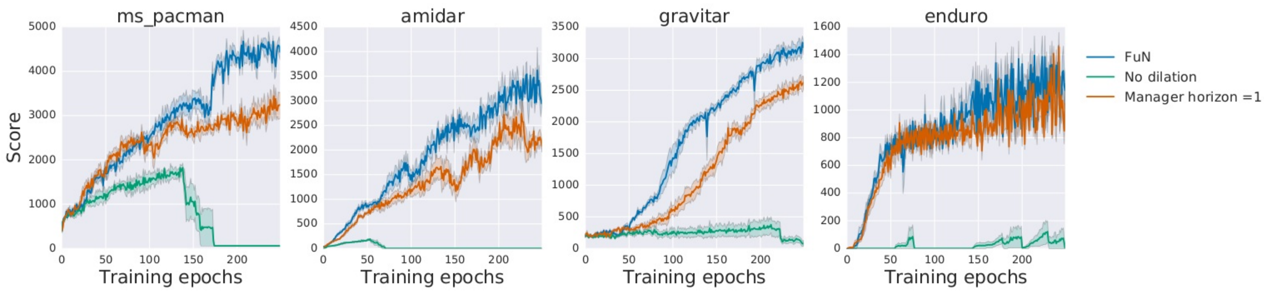


Figure 10. Learning curves for ablations of FuN that investigate influence of dLSTM in the Manager and Managers prediction horizon c. No dilation – FuN trained with a regular LSTM in the Manager; Manager horizon =1 – FuN trained with $c = 1$.

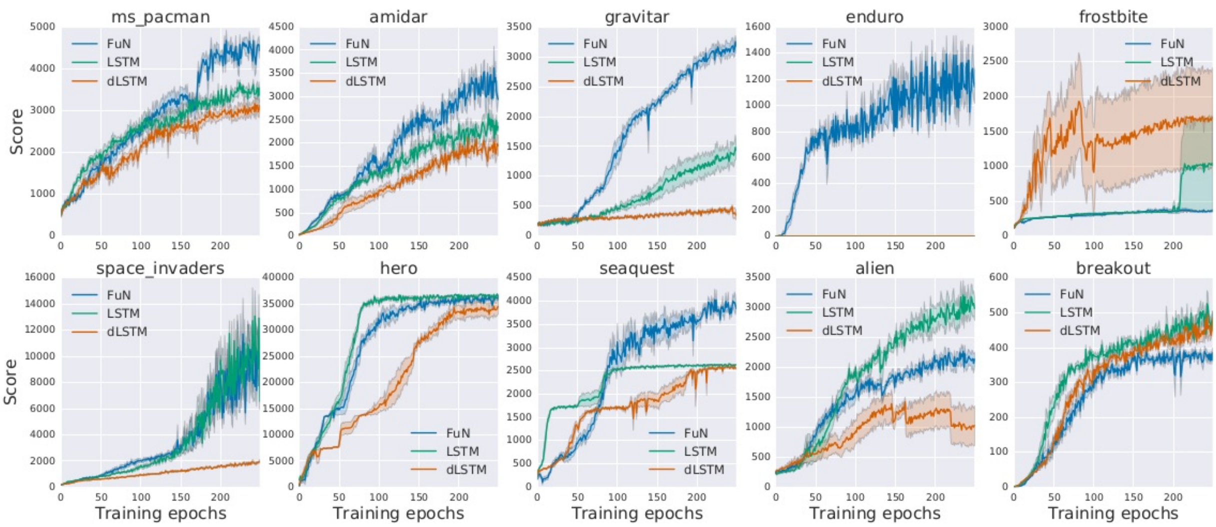


Figure 12. Learning curves for dLSTM based agent with LSTM and FuN for comparison.

- Action Repeat Transfer