Deep Reinforcement Learning for Atari games aided with human guidance

{ Kshitiz Tripathi } Stanford University

Objectives

1. Apply Deep Reinforcement Learning techniques to train an agent to play video games in a generic manner without hand crafted feature set
2. Develop an approach for enabling the agent to be guided by a human teacher.

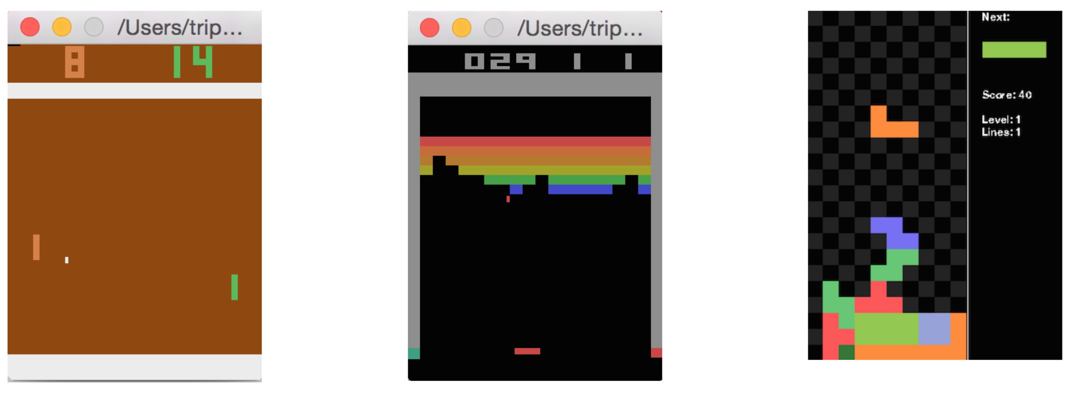

Figure 1: Video Games: pong(left), breakout, tetris

RL - Policy Gradient

- Explicit Policy $\pi_\theta(a_t|s_t)$ approximated by neural network
- Policy Gradient is given by

$$g = E[\sum_{t=0}^{\infty} \Psi_t \bigtriangledown_\theta \, log\pi_\theta(a_t|s_t)] \quad (1)$$

where,

$$\Psi_t = \sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'} \quad (2)$$

Neural Network

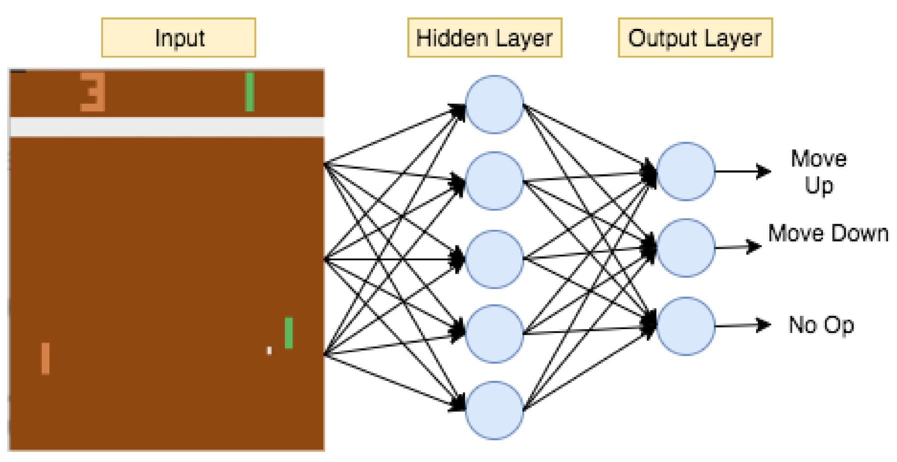

Figure 6: Policy Network with 2 layers Fully Connected Net.

parameters	value
Hidden layer Neurons	1000
Learning Rate	0.0005
Discount Factor	0.99
Optimizer	RMSProp
RMSProp Decay Rate	0.95
Update Batch Size	10

Figure 7: HyperParameters.

Results

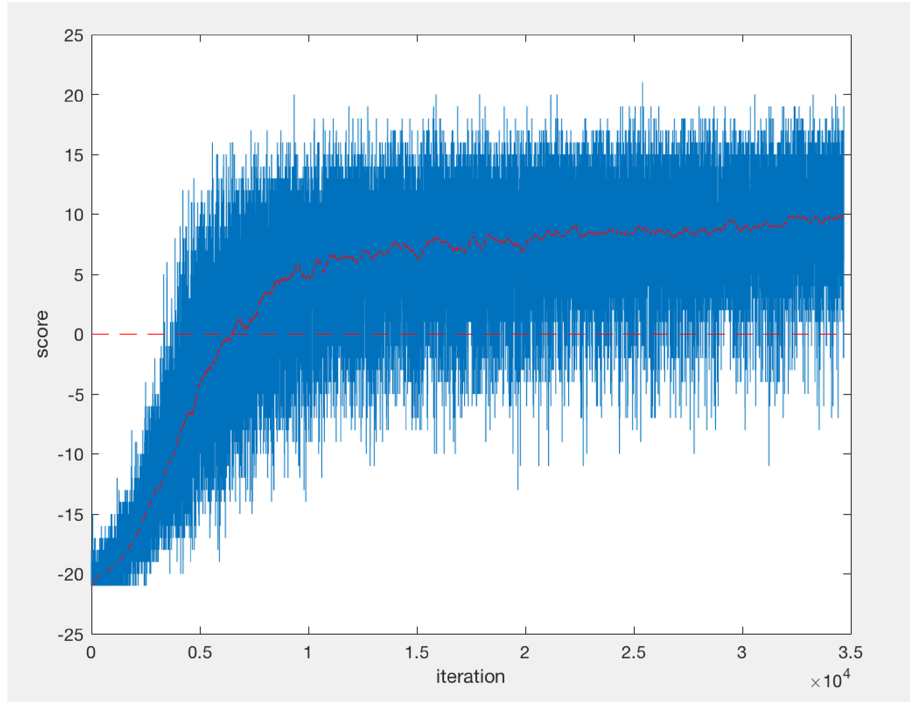

Figure 2: Pong training curve

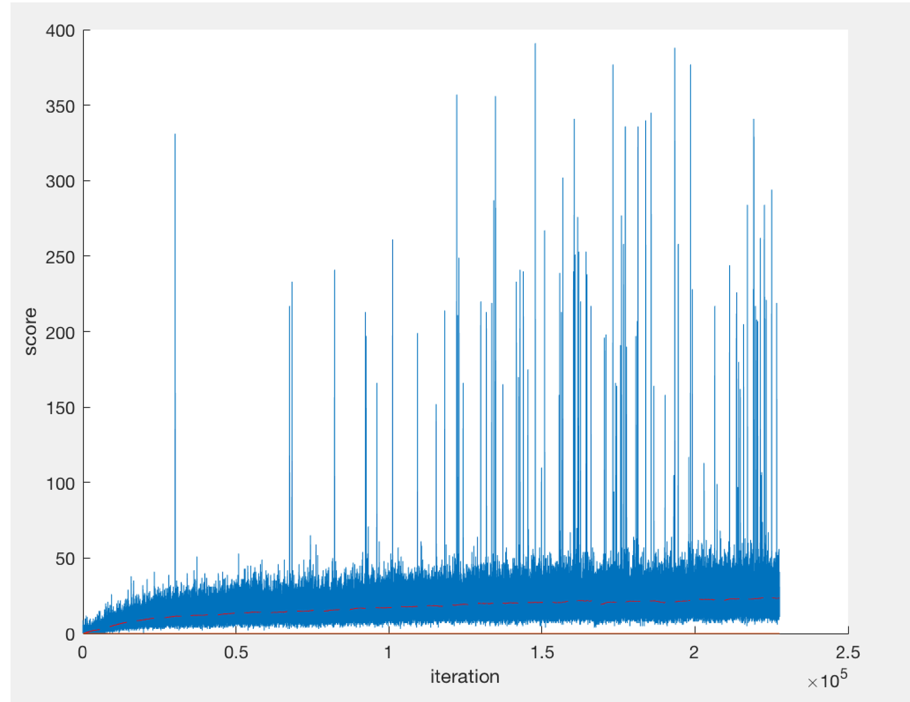

Figure 3: Breakout training curve

Human Guidance

- Human guidance can play an essential role in helping the bot learn game strategy involving multiple steps. Example: Picking a key in a game to unlock door or a simple one is tunneling strategy in breakout.
- One approach is to train the agent using supervised learning explicitly with samples containing the desired strategy collected from human teacher. Cons: requires too many samples and this may diverge the model from the optimal policy.

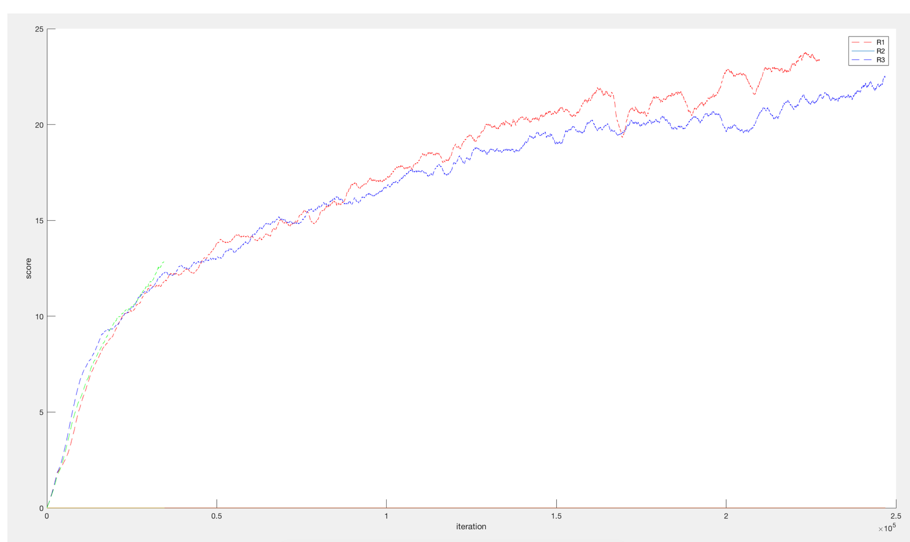

Figure 4: Testing different reward functions in breakout.

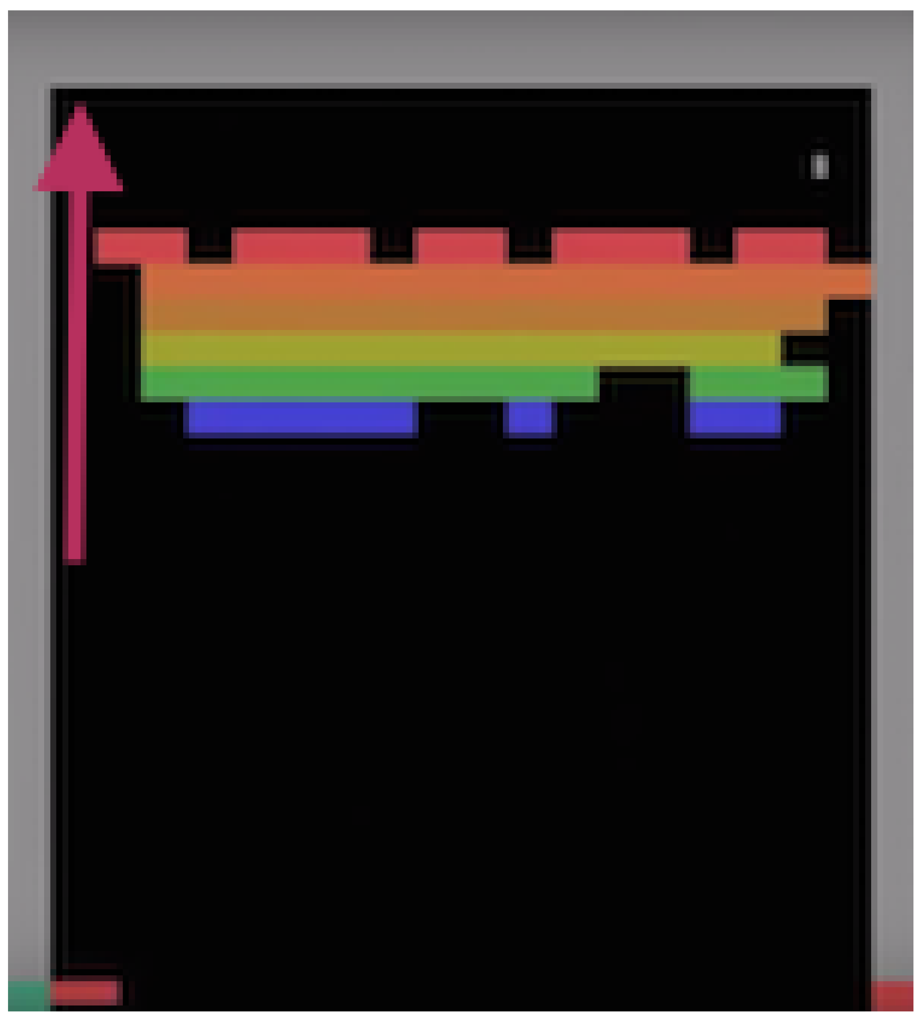

Figure 5: Tunneling Strategy in Breakout.

- Another approach is **Reward shaping**. Analyzing following reward function in breakout:
- R1: reward on hitting brick is credited only to actions between this hit and previous brick hit.
- R2: reward on hitting brick is credited only to all the actions before that hit
- R3: reward on hitting brick is credited only to all the actions between this hit and last brick hit before hitting the paddle. This will share the reward of multiple brick hit in single flight favoring tunneling strategy. Results awaited(Figure 4)

Conclusion

- Trained agents able to beat hard-coded computer player with a mean score > 10 and score 25+ mean score in breakout.
- Results awaited for reward shaping analysis to make agent learn tunneling strategy in breakout.

References

REFERENCESREFERENCES

[1] A. Karpathy. *Deep Reinforcement Learning: Pong from Pixels.*

[2] J Schulman and P. Moritz. High-dimensional continuous control using generalized advantage estimation.

[3] Andrew Y Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping.

Contact Information

Web https://github.com/kshitiz8
Email ktripathi8@gmail.com