Leveraging **Proc**edural **Gen**eration to Benchmark Reinforcement Learning **ProcGen**

Sungkwon On

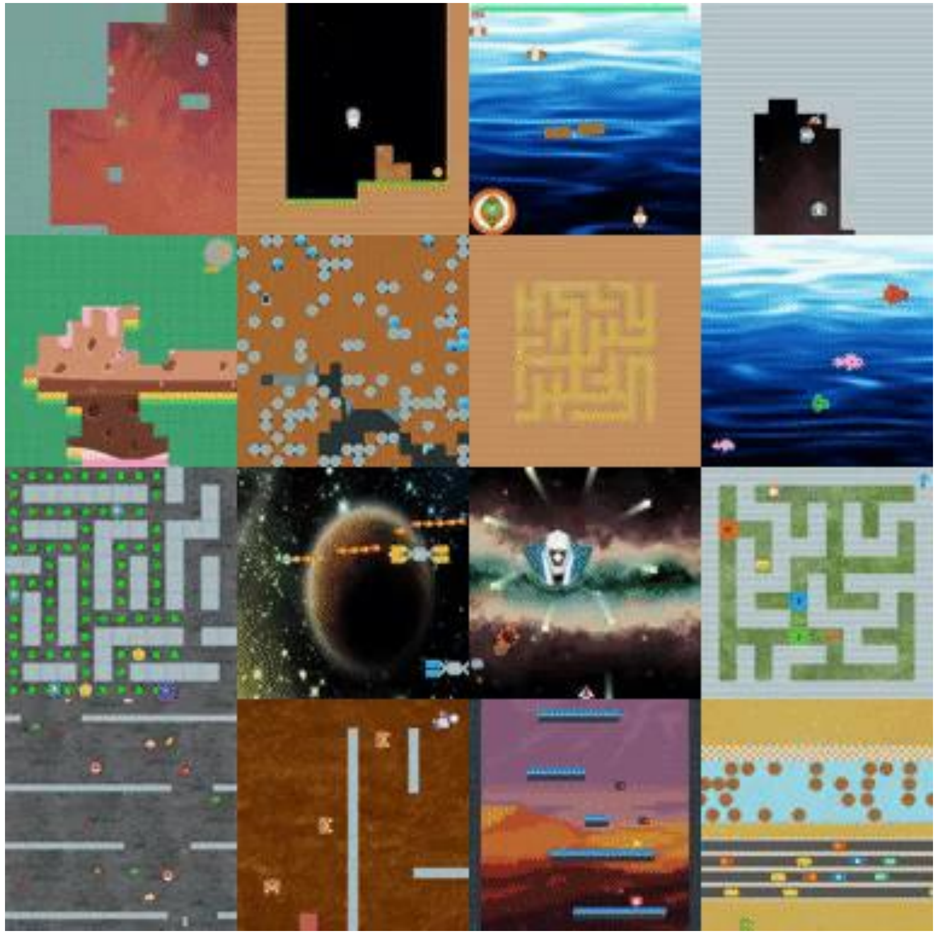
11 – July – 2022

Abstract

- Generalization is a fundamental challenge in DRL.
- Classic ALE(atari) environments have high diversity across the games but low diversity within a single game. → Overfitting may occur due to agents experiencing near-identical states.
- Do agents learn robust skills or simply memorize trajectories?
- Procgen environments fulfill both needs. High diversity across the games and within a single game.
- Procgen is ideal for generalization!
  - Distinct train and test sets may be generated.
- Procgen is also well-suited to evaluate sample-efficiency.
  - Environments provide diverse and compelling tasks.

Procgen

- High Diversity
- Fast Evaluation
  - Optimised to perform thousands of steps per sec on a single CPU (including rendering)
- Tunable Difficulty
  - Provides Easy and Hard modes
    - Differences in level distribution
  - Easy takes 1/8 resources of Hard
- Level Solvability
  - Greater than 99% of levels are believed to be solvable.
- Emphasis on Visual Recognition and Motor Control
  - Similar to Atari & Gym Retro, Procgen requires critical assets to be recognized in observation.



Procgen

- Shared Action and Observation Space
  - 15 discrete action space
  - 64x64x3(RGB) observation space
- Tunable Dependence on Exploration
  - On some games, training performance increase with the size of training sets → exploration becomes less bottleneck
  - But on 8 games, specific seeds introduce the difficulty of exploration.
- Tunable Dependence on Memory
  - Designed to require minimal use of memory.
  - 6 games include variants that do test the use of memory.

Experimental Protocols

- By default, PPO is used.
- Hard difficulty – Training for 200M timesteps.
  - 24 GPU-hrs and 60 CPU-hrs
- Easy difficulty – Training for 25M timesteps.
  - 3 GPU-hrs
- Sample efficiency Evaluation
  - Train and Test on the full distribution of levels
- Generalization Evaluation
  - Train on a finite set (500 if not specified) of levels and Test on the full distribution
    - Easy mode – 200 Levels for Training
- Rewards normalized when representing the mean for all games.
- Hyperparameters are tuned minimally and work well across Atari games as well as Procgen.
- No frame Staking, No LSTM → Experimentally found that the impact is minimal
- IMPALA CNN is used. (Nature CNN (DQN architecture) hardly performs)

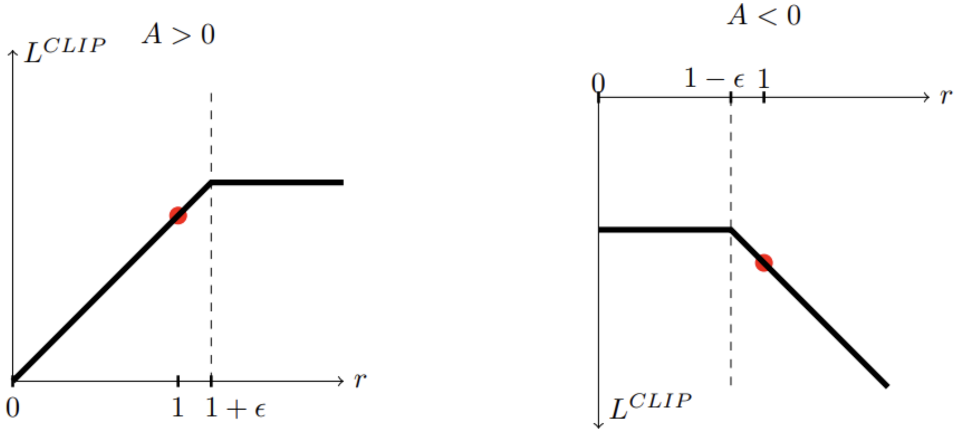
PPO review

A traditional on-policy actor-critic method.

Probability ratio clipping prevents the policy from changing too much for one update.

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\Big[\min(r_t(\theta)\hat{A}_t, \operatorname{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\Big]$$

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$



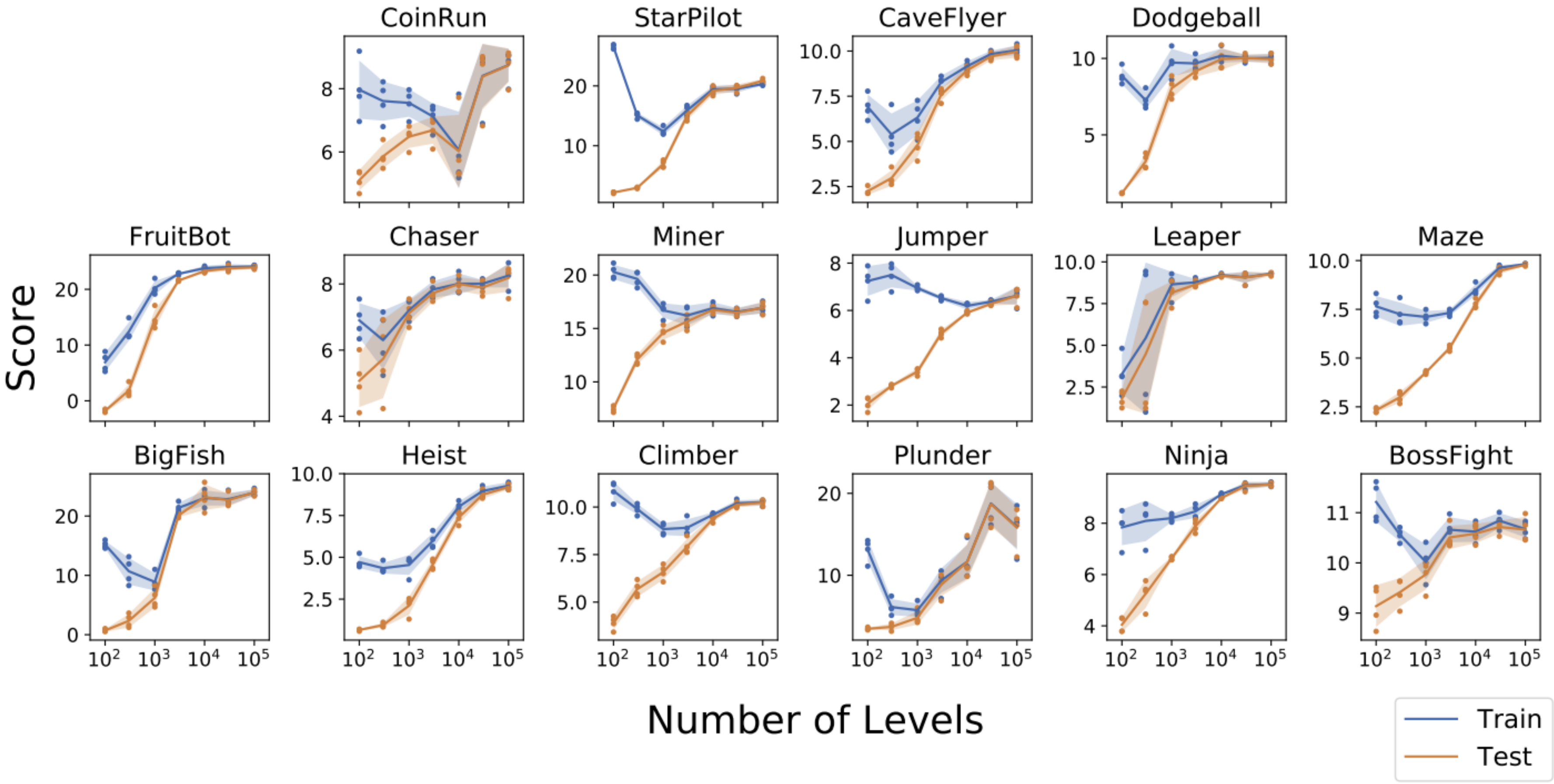
Generalization Experiment

Protocol

- Train on different sets of levels ranging from 100 to 100,000
- Test on the full distribution of levels.

Results

- Overfitting for small training sets
- Generalization well learned from 10,000 training sets.
- Some games show an increase in performance above a certain threshold of the training set (opposite to supervised learning). → A sign of generalization.



Generalization Experiment

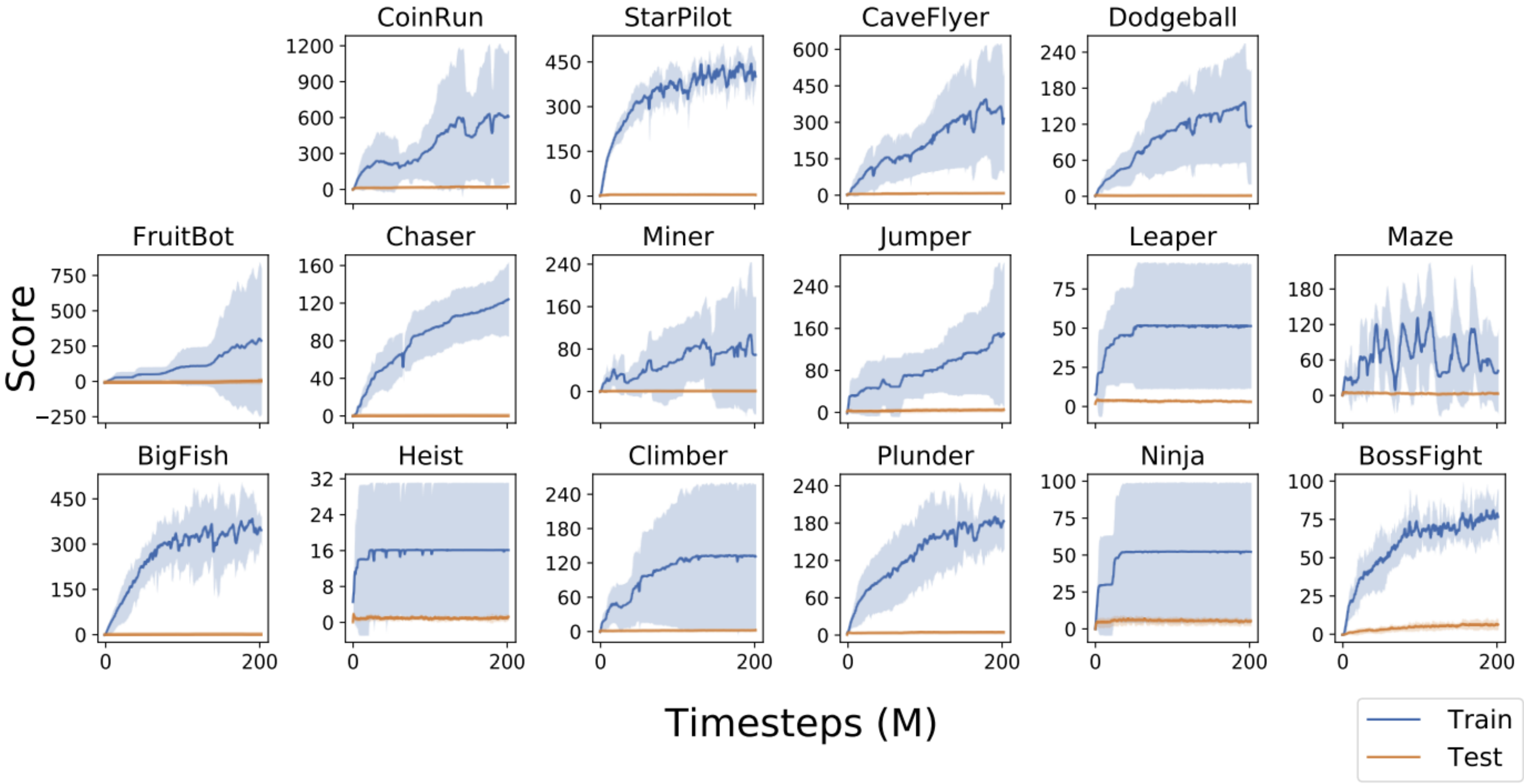
Ablation study with Deterministic Levels

Protocol

- Train on a fixed sequence of levels. Starting from level 1.
- Easy difficulty used.
- Test on a sequence of levels chosen at random.

Results

- Agents hardly performs during Test.
- Important to train and testing on diverse environment distribution.



Generalization Experiment

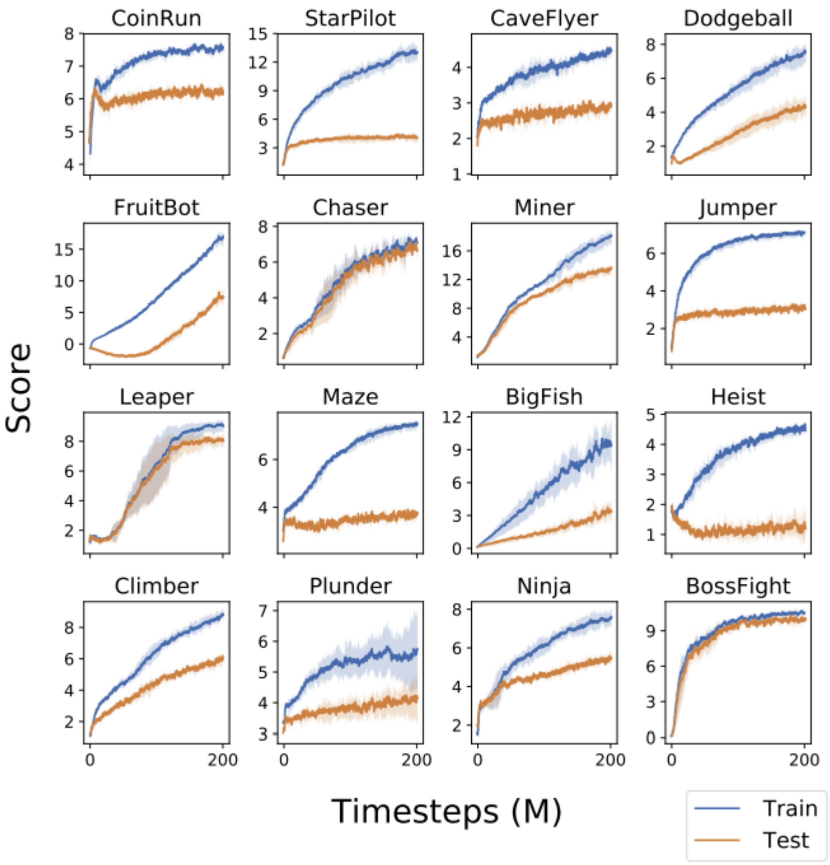
500 Level Generalization

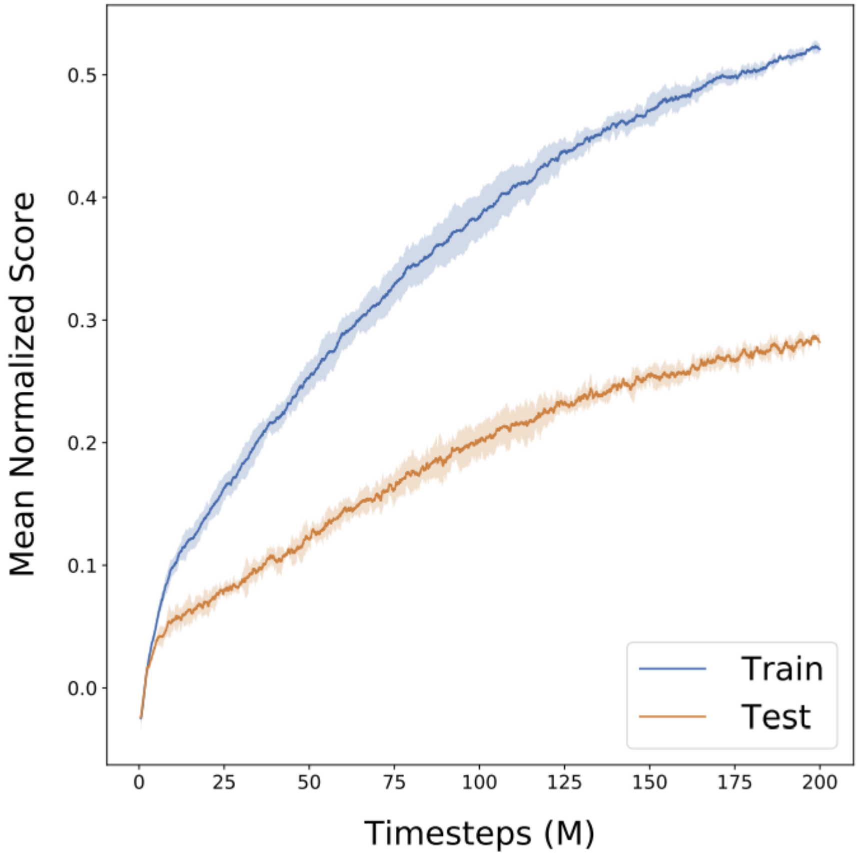
Protocol

- Train on a 500-level set.
- Test on the full distribution of levels.

Results

- High overfitting in most of the games.
- Some games show less generalization gap but only due to poor performance.





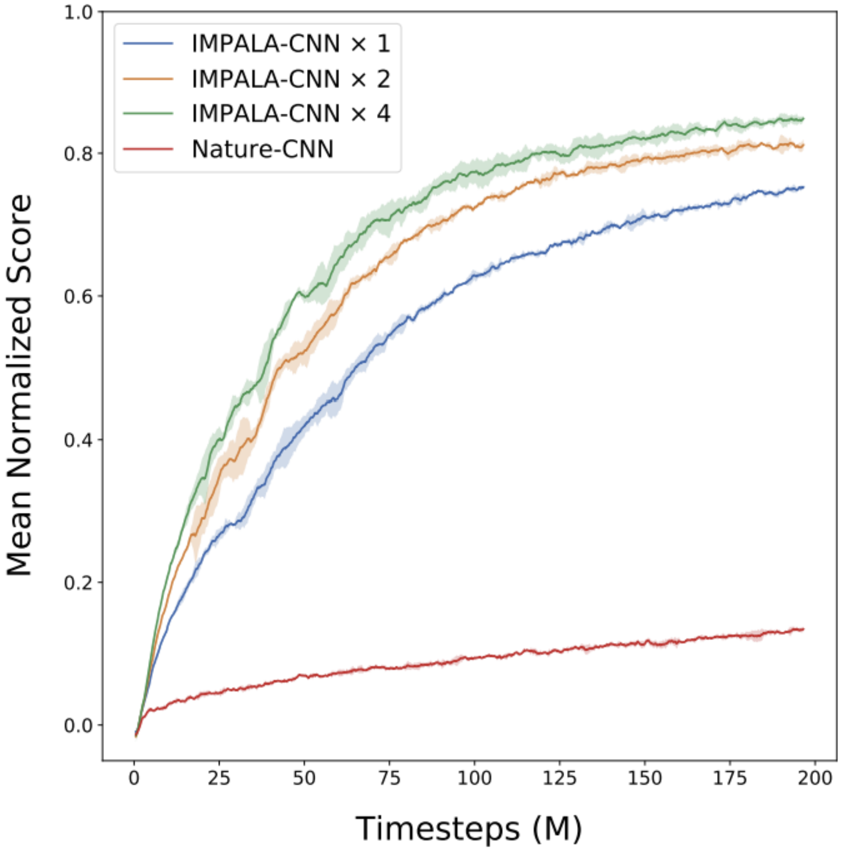
Scaling Model Size

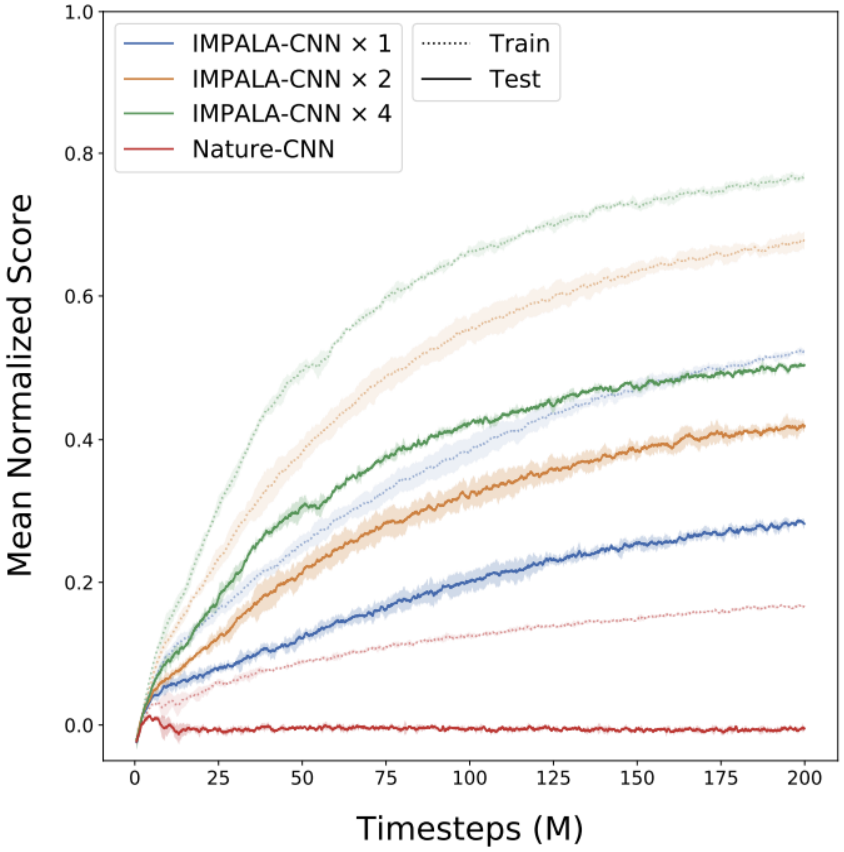
Protocol

- Train and Test on different Network Architectures
- Scaled number of convolutional channels by 2 and 4. (reducing learning rate by the root of the scaling factor)
- Other hyperparameters unchanged.

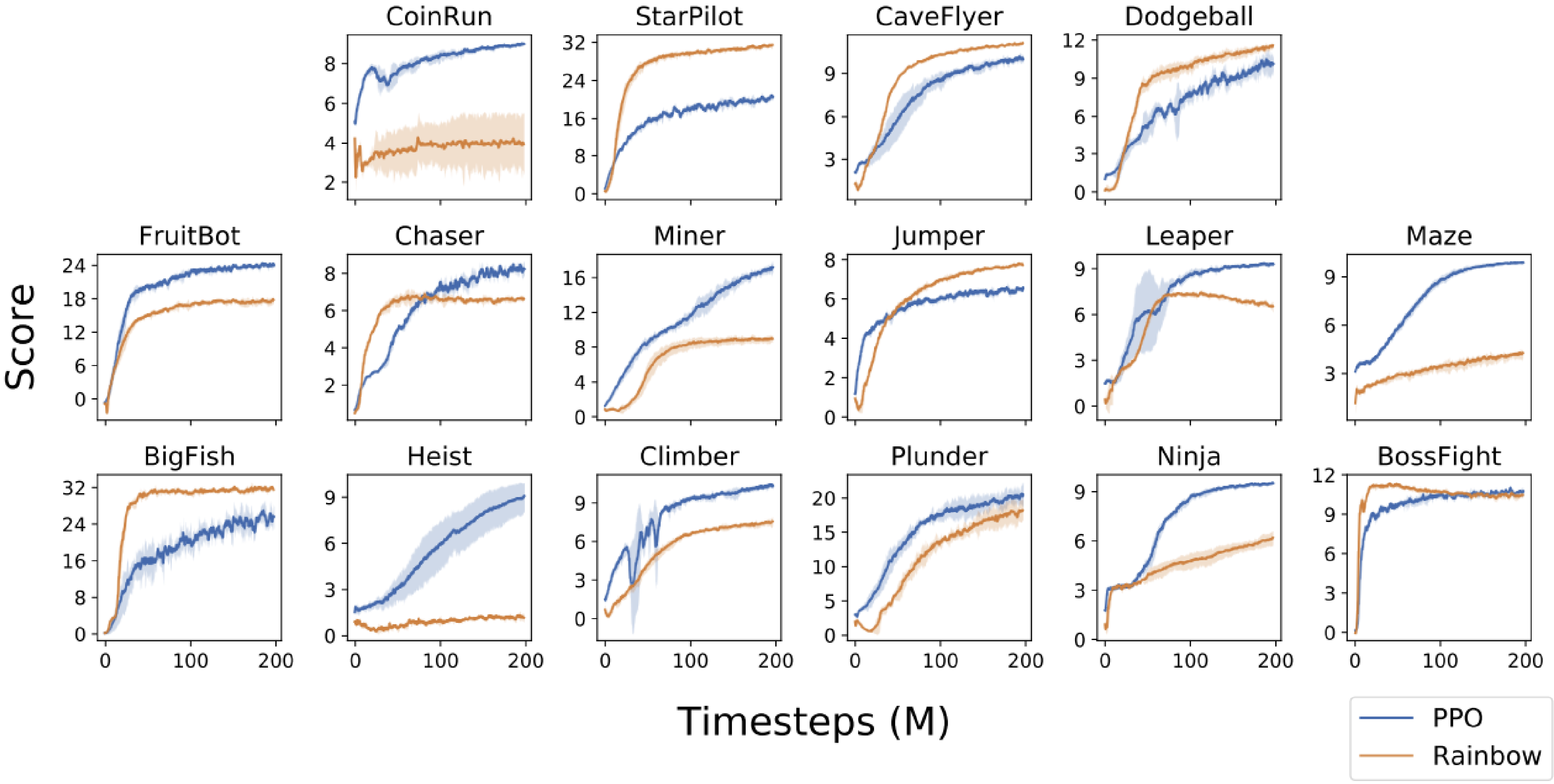
Results

- Deep IMPALA networks performs the best.
- Nature CNN fails to perform any good.





PPO VS Rainbow DQN



Questions?