REINFORCEMENT LEARNING FOR SUPER MARIO BROS

Anthony Khaiat, Valentin Pinon, and Jan Philipp Girgott

May 14th, 2024 | Sensorimotor Learning

MIT MANAGEMENT SLOAN SCHOOL

PROJECT OVERVIEW AND ENVIRONMENT

MIT MANAGEMENT SLOAN SCHOOL

Project Overview

WE WILL IMPROVE THE DDQN ALGORITHM BY REWARD SHAPING AND OTHER NEURAL NET ARCHITECTURES

We are advancing a reinforcement learning tutorial for the OpenAI Super Mario Bros. gym environment, through Reward Engineering and Enhanced Double Deep Q-Network Architectures.

Issues Tackled

1. **Reward Engineering:** Enhancing the basic reward by adding extra information beyond just horizontal progress and time

2. **Advancing Neural Net Architectures:** DDQN currently uses a basic CNN. So, we explored other architectures to enhance performance

MIT MANAGEMENT SLOAN SCHOOL

Environment Overview

MARIO ACTS IN A DISCRETE ACTION SPACE BASED ON AN OBSERVATION, INFO DICT AND REWARD VALUE

State Space

1) Observation

2) Info Dictionary

```
{x_pos:   x position,
 y_pos:   y position,
 score:   curr. Score,
 coins:   coll. coins,
 time:    time taken,
 status:  Mario's size}
```

Action Space (7-dimensional)

Basic actions:

1.[NOOP]:	Do **nothing**
2.[right]:	Move **forward**
3.[right, A]:	Move **forward/jump**
4.[right, B]:	**Run/shoot** fireballs
5.[right, A, B]:	**Jump, run, fireballs**
6.[A]:	**Jump**
7.[left]:	Move **backward**

Current Reward Structure

Reward = Linear Horizontal Progress Reward + Linear Time Penalty + Death Penalty

MIT MANAGEMENT SLOAN SCHOOL

02

CHANGES TO REWARD AND NN ARCHITECTURE

MIT MANAGEMENT SLOAN SCHOOL

Changes to the Reward Function

WE CREATED THREE REWARD FUNCTIONS FOCUSING ON DIFFERENT STATE SPACE ASPECTS

R^1: Use Information Dictionary

Use base reward and add info

- Add small incremental reward (0.5) for each **coin collected** and each **point earned** (0.1)
- Add reward for maintaining a **superior status** (0.5 for tall and 1 for fireball)

R^2: Decrease Time Importance

Alter time penalty of base

- Primary goal is to complete the level without dying, time only becomes important towards the end
- Implement quadratic increase in time penalty over game time

R^3: Make Later Progress More Rewarding

Alter horizontal reward of base

- Later steps in the game could be more rewarding, as they lead directly to the goal
- Make the horizontal progress reward depend on how far Mario has come in the game

Changes to the Algorithm

WE IMPLEMENTED THREE DIFFERENT TYPES OF NEURAL NETWORKS IN DDQN AND PPO

Algorithms

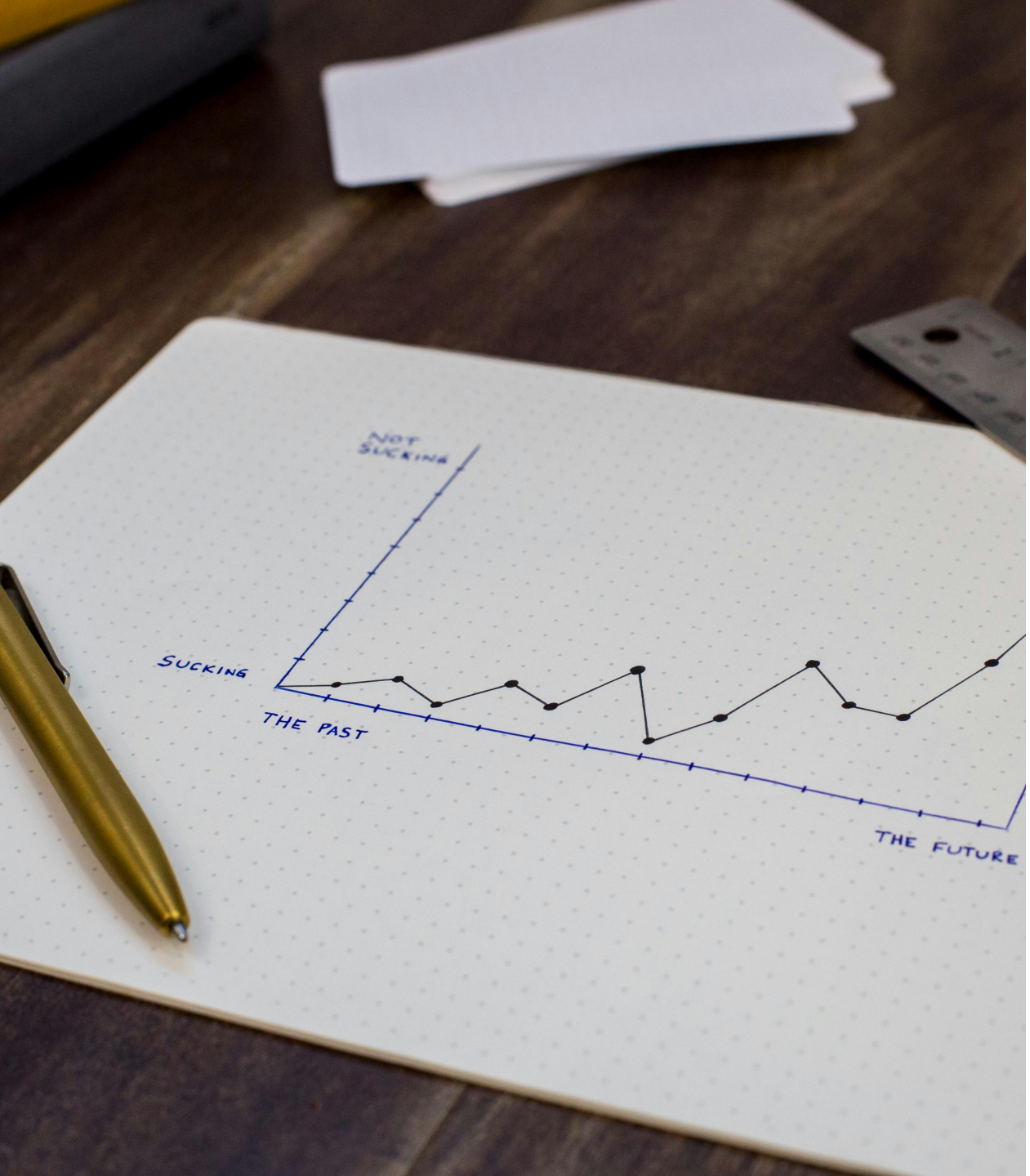
1. **Double Deep Q-Neural Networks (DDQN)**
   Stable learning for discrete action spaces
2. **Proximal Policy Optimization**
   State-of-the-art benchmark especially regarding efficiency

NN Architectures

1. **ResNet50 for RNNs**
   Deep architecture with shortcuts, ideal for capturing complexity
2. **Vision Transformer (ViT) for Transformers**
   Efficient at understanding spatial relationships in environments
3. **AlexNet for CNNs**
   Simple structure and fewer parameters for quicker training

40 Combinations
Resulting from permutations between algorithms, architectures and reward functions (incl. baselines)

MIT MANAGEMENT SLOAN SCHOOL



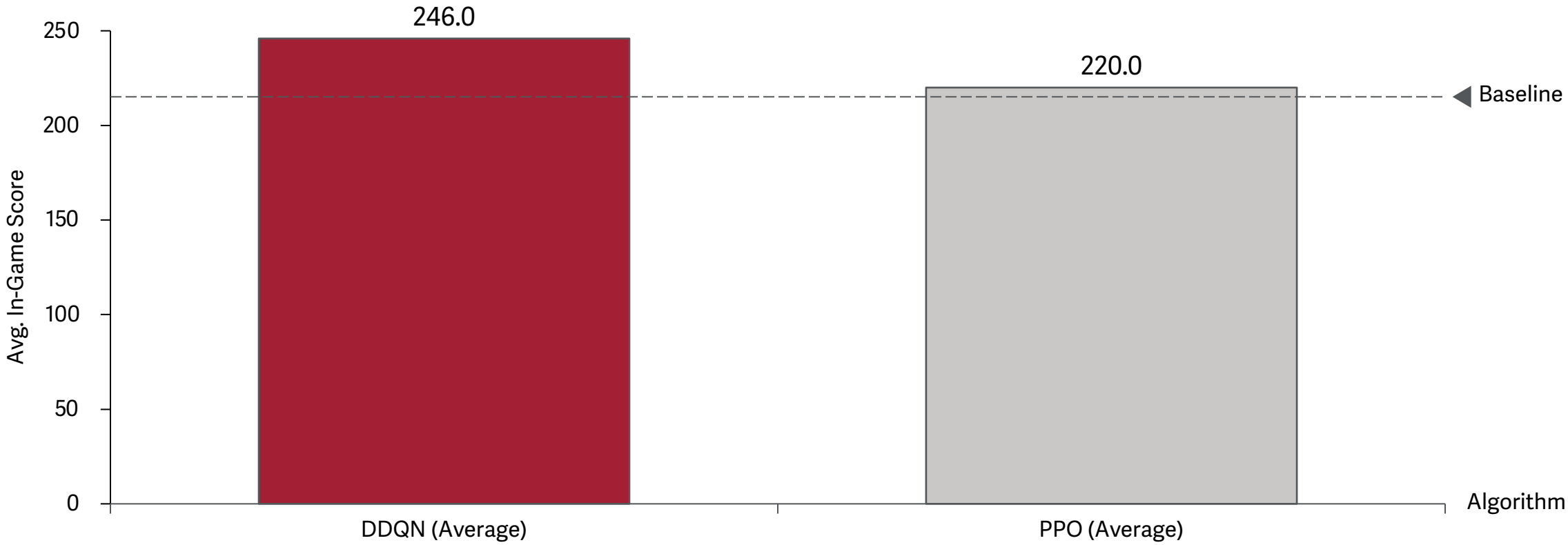
RESULTS AND LEARNINGS

MIT MANAGEMENT SLOAN SCHOOL

Results (1/3)

ON AVERAGE, DDQN OUTPERFORMS PPO ACROSS THE DIFFERENT REWARD FUNCTIONS AND ARCHITECTURES

Average In-Game Score Per Algorithm Group

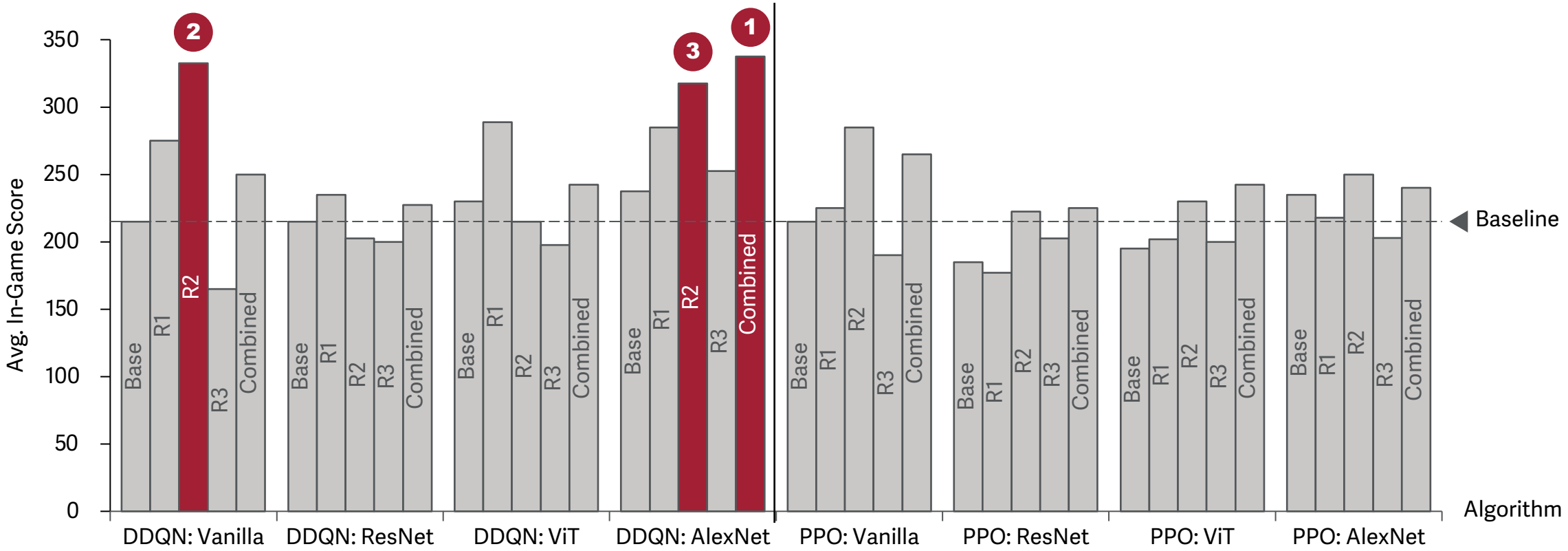


MIT MANAGEMENT SLOAN SCHOOL

Results (2/3)

ALEXNET WITH THE COMBINED REWARD FUNCTION YIELDS THE HIGHEST AVERAGE SCORE

Average In-Game Score Per Reward Function and Algorithm Combination

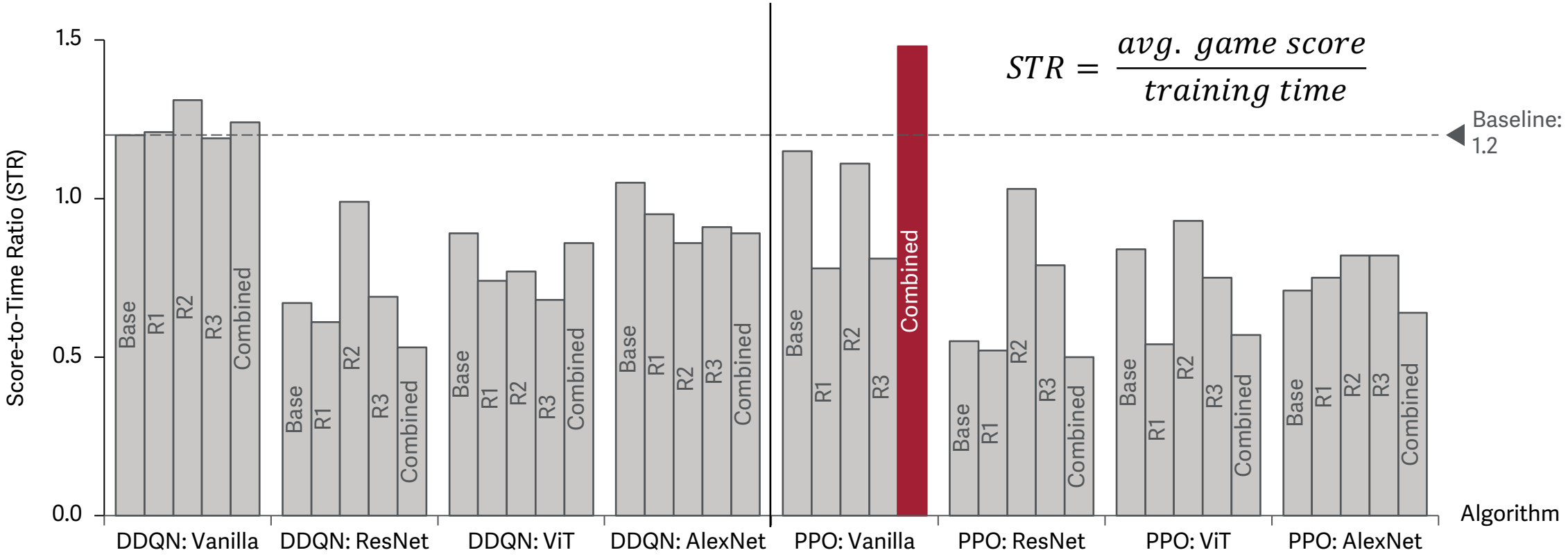


MIT MANAGEMENT SLOAN SCHOOL

Results (3/3)

PPO WITH THE COMBINED REWARD OUTPERFORMS WHEN FACTORING IN COMPUTE TIME

Score-to-Time Ratio (STR) Per Reward Function and Algorithm Combination



MIT MANAGEMENT SLOAN SCHOOL

Learnings

CHOOSING THE RIGHT REWARD FUNCTION AND MODEL COMPLEXITY IS CRUCIAL

1

Reward Alignment
- Aligning the reward with the game goal is crucial
- Demonstrated by the success of R^2 which decreased the impact of time

2

Sophisticated rewards can be beneficial, advanced NNs not always
- The combined reward function achieved top results
- Complex NN architectures like ResNet or ViT were less successful

3

Simple models show the best performance-efficiency tradeoff
- PPO with a Vanilla CNN achieved the highest STR (score-to-time ratio)
- Other models were not able to beat the baseline in terms of efficiency

THANK YOU!

Anthony Khaiat, Valentin Pinon, and Jan Philipp Girgott

May 14th, 2024 | Sensorimotor Learning

MIT MANAGEMENT SLOAN SCHOOL