U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Introduction to Intelligent e Autonomous Systems

REINFORCEMENT LEARNING

GYM ENVIRONMENT

Developed by:
Alexandre Carneiro up202107858
Ana Cláudia Batista up202108234
Mafalda Aires up202106550

CAR RACING

- This environment is part of the Box2D.
- The observation space is a 96x96 RGB image of the car and race track.
- The car starts at rest in the center of the road.
- The episode finishes when all the tiles are visited or when the car goes outside the playfield - it receives -100 reward and dies.
- The reward is -0.1 every frame and +1000/N for every track tile visited, where N is the total number of tiles visited in the track

CAR RACING

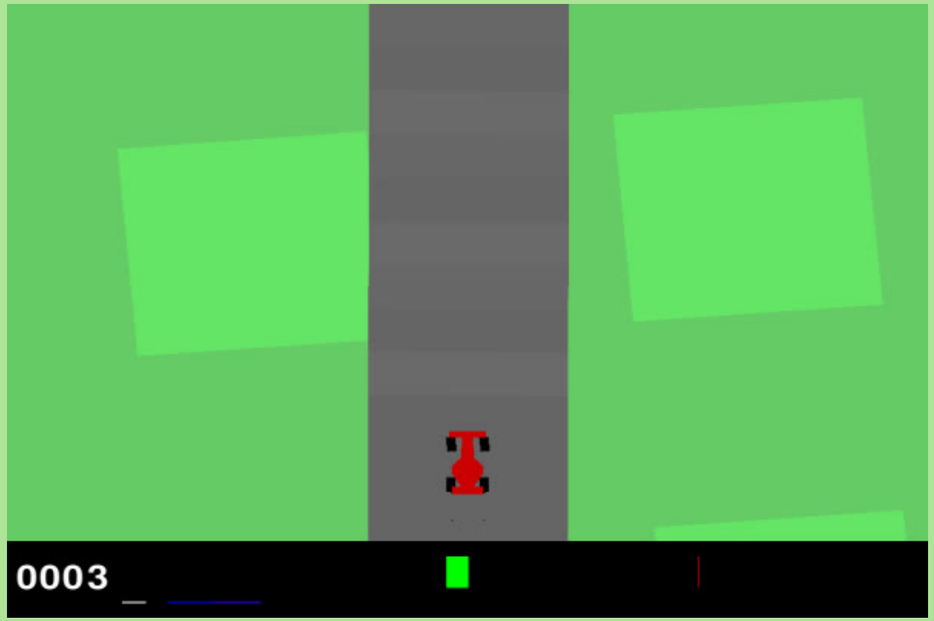
Action Space

Continuous:

- 0: steering, -1 is full left, +1 is full right
- 1: gas
- 2: breaking

Discrete:

- 0: do nothing
- 1: steer left
- 2: steer right
- 3: gas
- 4: brake



CHANGES IN THE GYM ENVIRONMENT

- GrayScaleObservation
- FrameStack

CHANGES IN THE AGENT'S ACTIONS

- Normalize Reward
- Reward Clipping
- Reward Penalty that penalizes going to the grass.

CHOSEN RL ALGORITHM

Our choosen algorithm was PPO.

We also tested with A2C, but the results weren’t as good as with PPO.

NOT ABOUT ALGORITHMS - Policy of an algorithm

One big difference that affected a lot in our tests in a good way was the use of CnnPolicy instead of MlpPolicy.

RESULTS - MlpPolicy

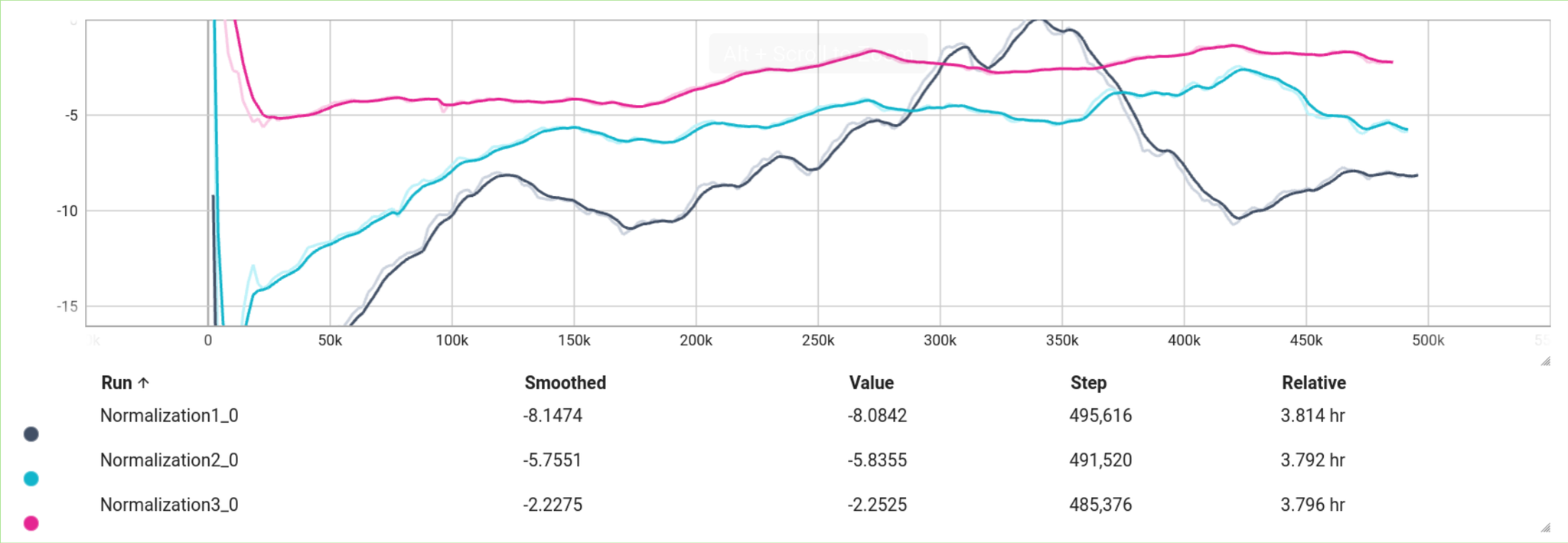
We started experimenting with the PPO algorithm and the MlpPolicy:

- GrayScaleObservation
- FrameStack (number of frames = 3)
- GrayScaleObservation + FrameStack
- Reward Normalization
- Clipping Reward to [-1.0,1.0]
- Reward Penalty
- GrayScaleObservation + Reward Normalization

Most of these alterations did close to nothing. The rewards were always negative.

RESULTS - Experimenting with Normalization

Experimenting with 3 values of Gamma for PPO with MlpPolicy. Gamma = 0.99 was the most stable one, but we chose to use Gamma = 0.90 for the remaining of our experiments.



Run ↑	Smoothed	Value	Step	Relative
Normalization1_0	-8.1474	-8.0842	495,616	3.814 hr
Normalization2_0	-5.7551	-5.8355	491,520	3.792 hr
Normalization3_0	-2.2275	-2.2525	485,376	3.796 hr

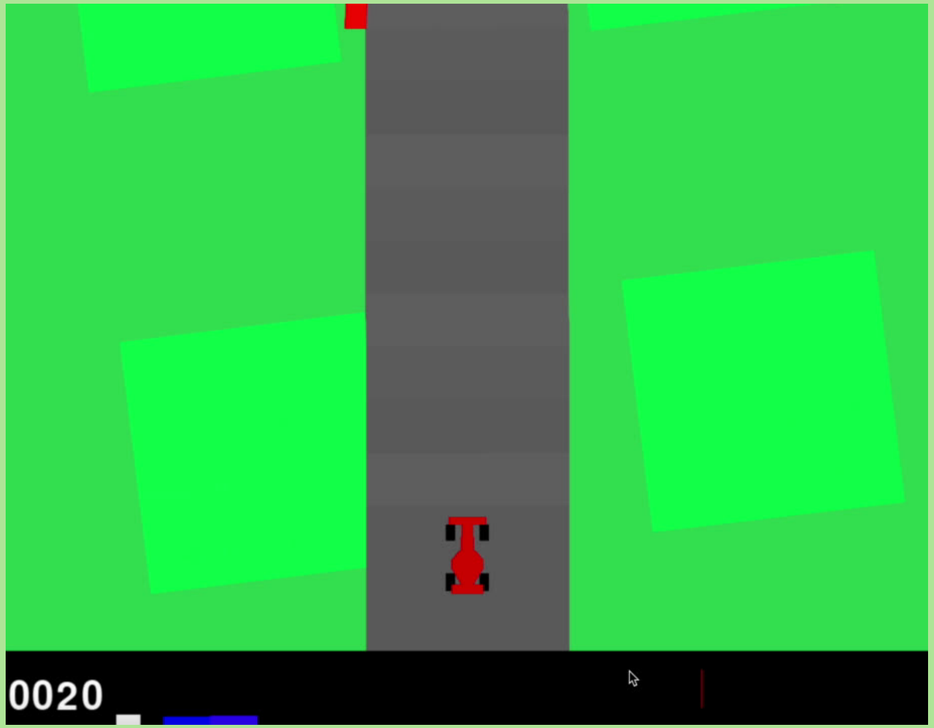
Normalization1 : 0.90 Normalization2 : 0.95 Normalization3 : 0.99

RESULTS - MlpPolicy

In most of our initial experiments the car would go straight to the abism (video on the left). At best, it only learned that it could not go pass the playfield (video on the right).

RESULTS - MlpPolicy vs CnnPolicy

After seeing little change, we decided to try another policy - the CnnPolicy. It was immeadiately more effective than MlpPolicy. This is how PPO (without alterations) performed with 350000 steps of training.

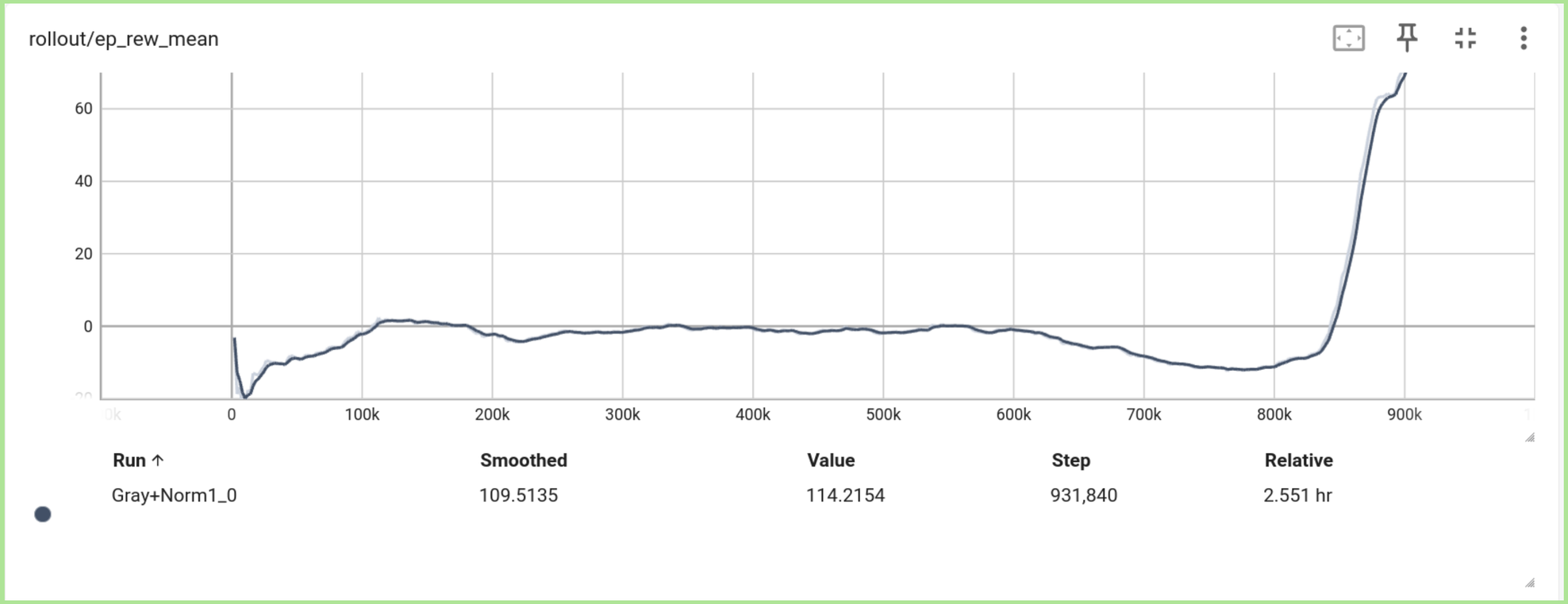


MlpPolicy

CnnPolicy

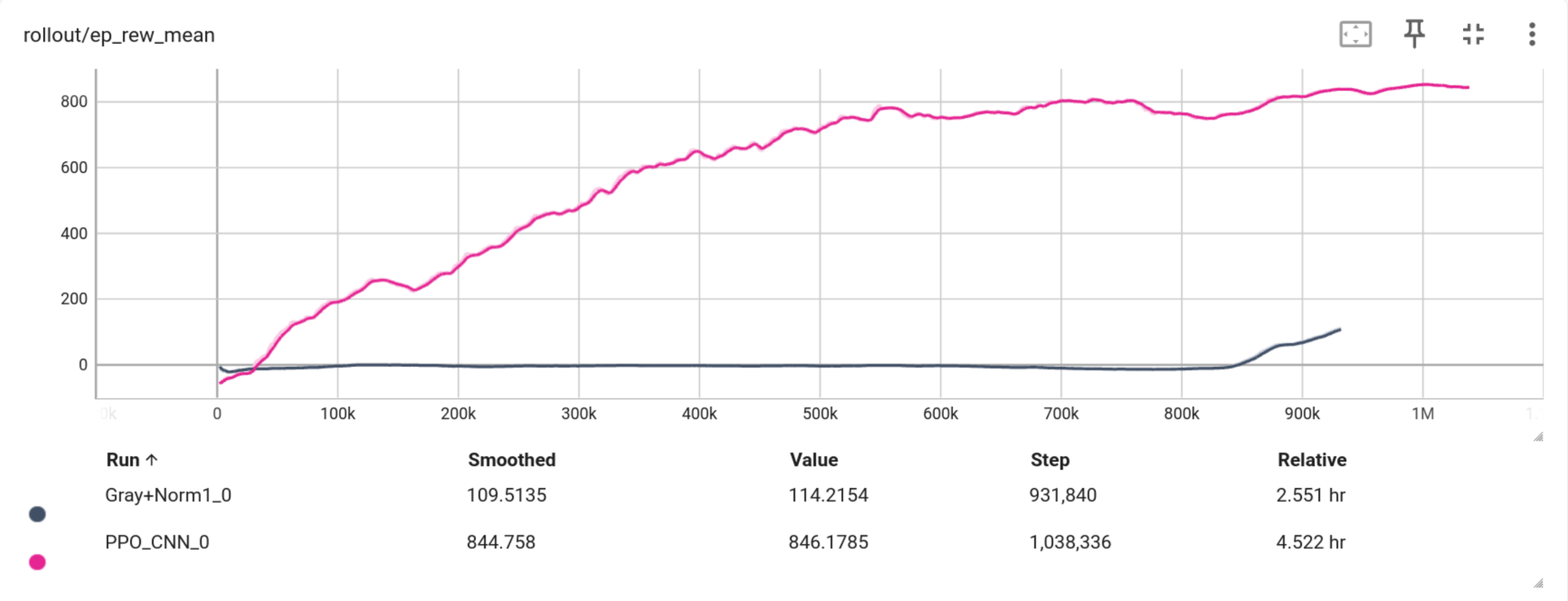
RESULTS - CnnPolicy

Because the CnnPolicy performed so much better, we decided to use it for the next experiments. Our best performing model was the GrayScaleObservation + Normalization, so we tried it with CnnPolicy this time.



RESULTS - Best Models

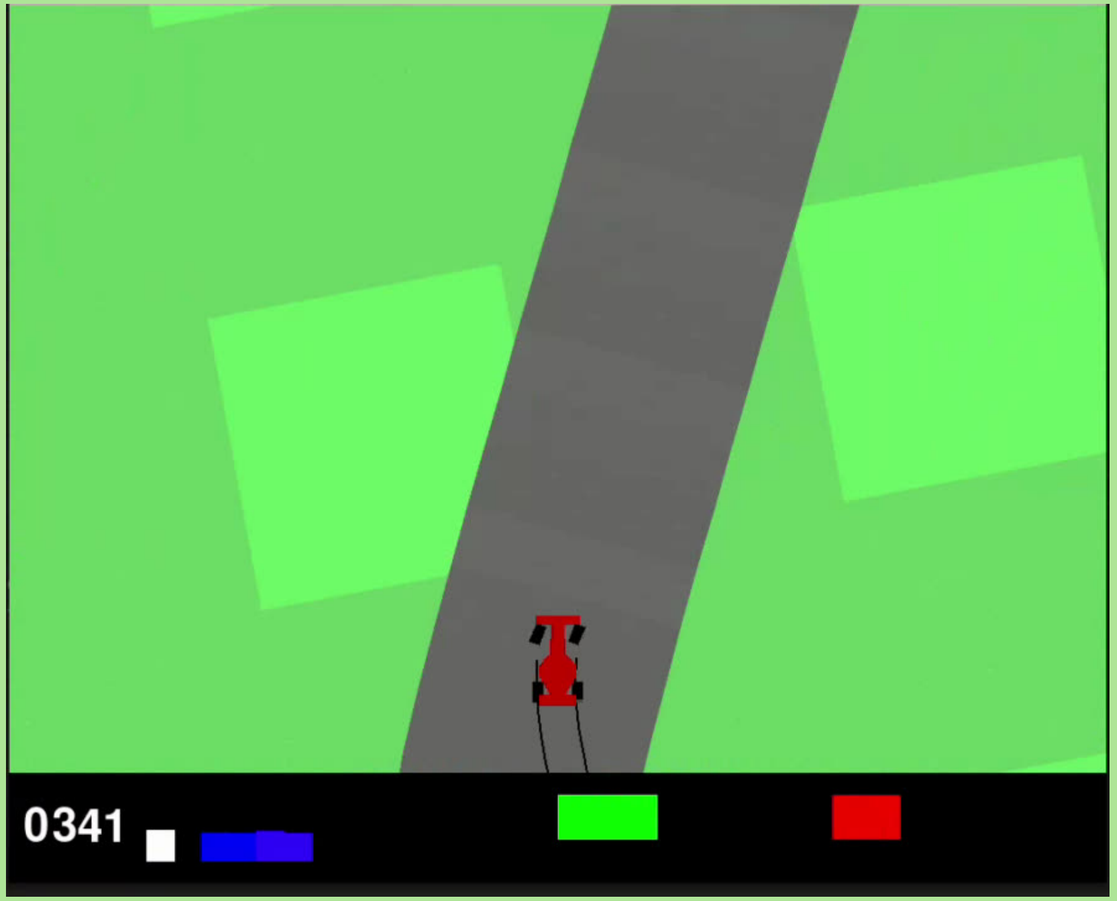
The best models we got were the unaltered PPO and PPO with GrayScaleObservation and Normalization with Gamma = 0.90. Both with CnnPolicy.

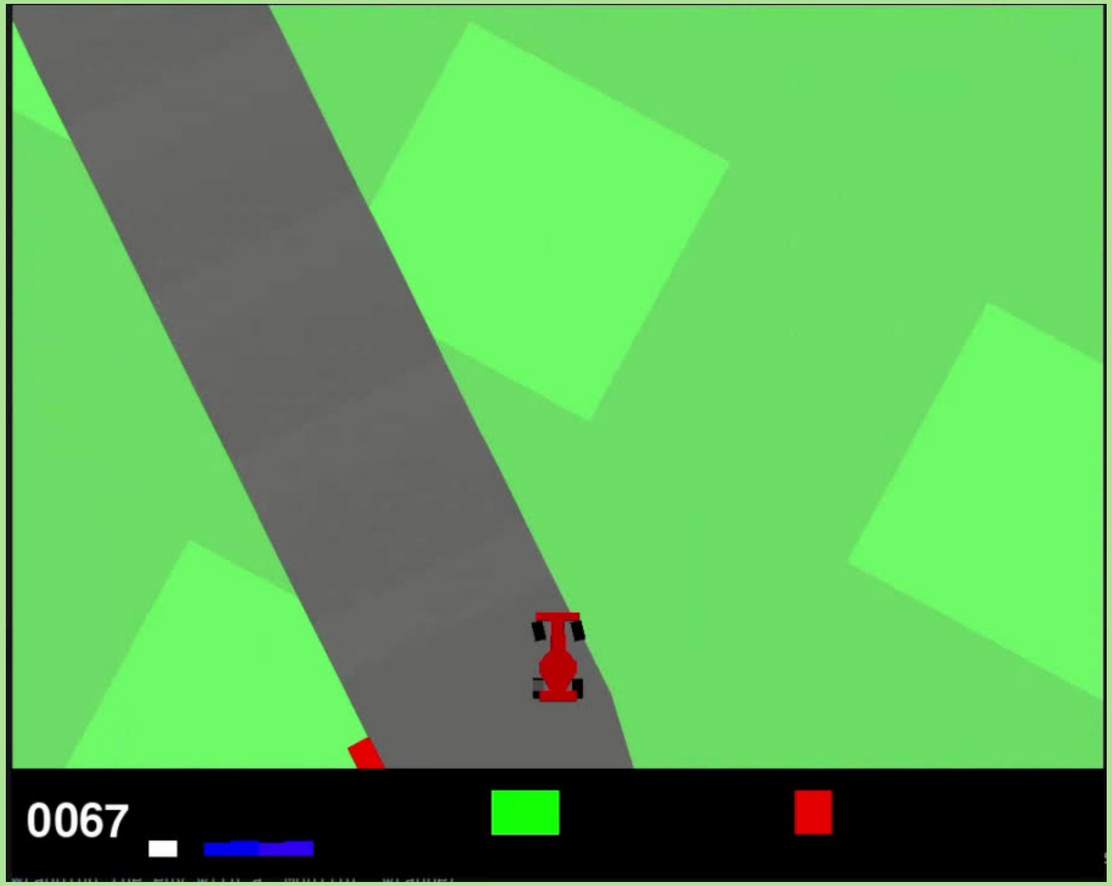


Run ↑	Smoothed	Value	Step	Relative
Gray+Norm1_0	109.5135	114.2154	931,840	2.551 hr
PPO_CNN_0	844.758	846.1785	1,038,336	4.522 hr

RESULTS - Best Models

PPO Model without alterations trained with 1 000 000 steps vs our best modified model with 940 000 steps.





CONCLUSION

- In conclusion we were not able to make significant changes to the model in order for it to learn faster.

- With the right algorithm and policy the default environment starts to perform pretty well around 300 000 steps while our altered algorithms either performed poorly or took a lot more steps to start learning.

- In order to better the already existing environment we think the best option would be modifying the original code instead of only using wrappers.