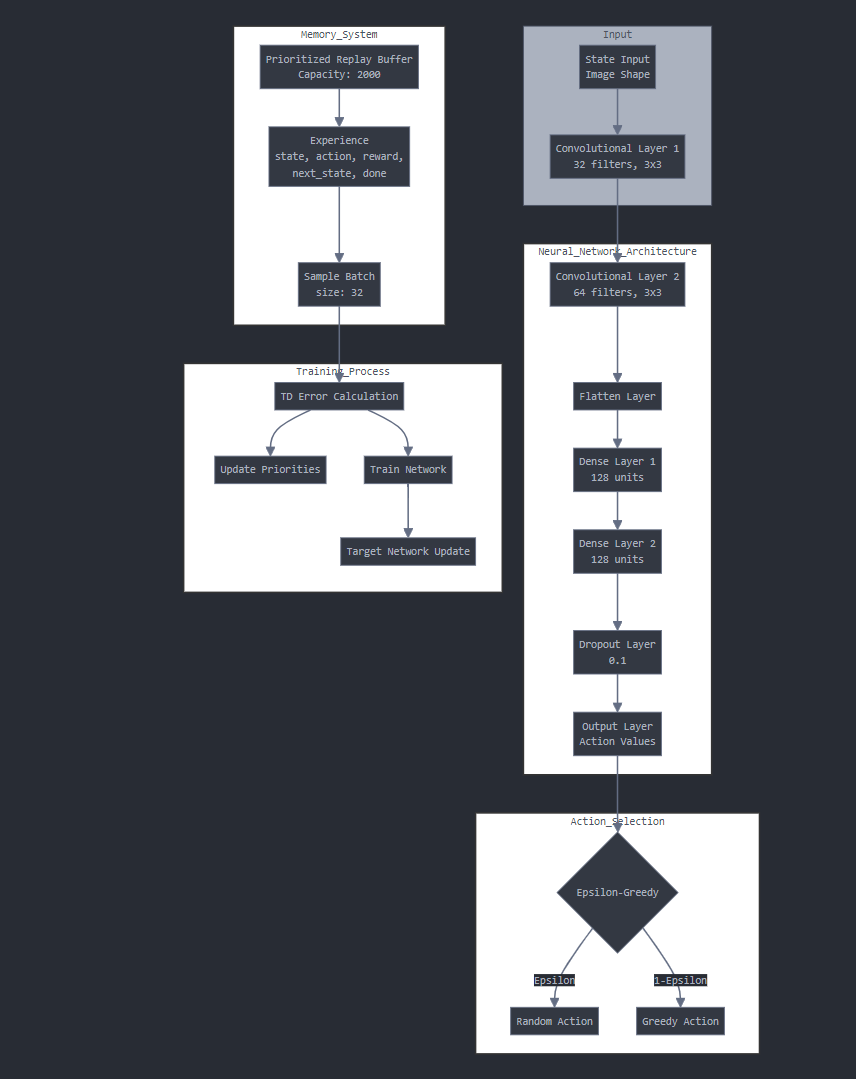
XAI RESULTS

MODEL ARCHITECTURE (DQN ARCHITECTURE):



OVERALL DESCRIPTION

A DQN model has been trained for the Fruit Catcher Game under 3 configurations:

1. 1 epoch
2. 10 epochs
3. 100 epochs

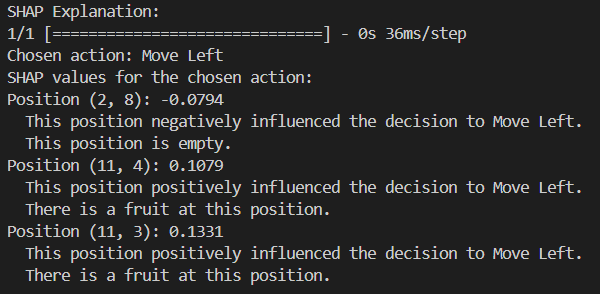
(i)1 epoch model:

Evaluation metrics of the model:

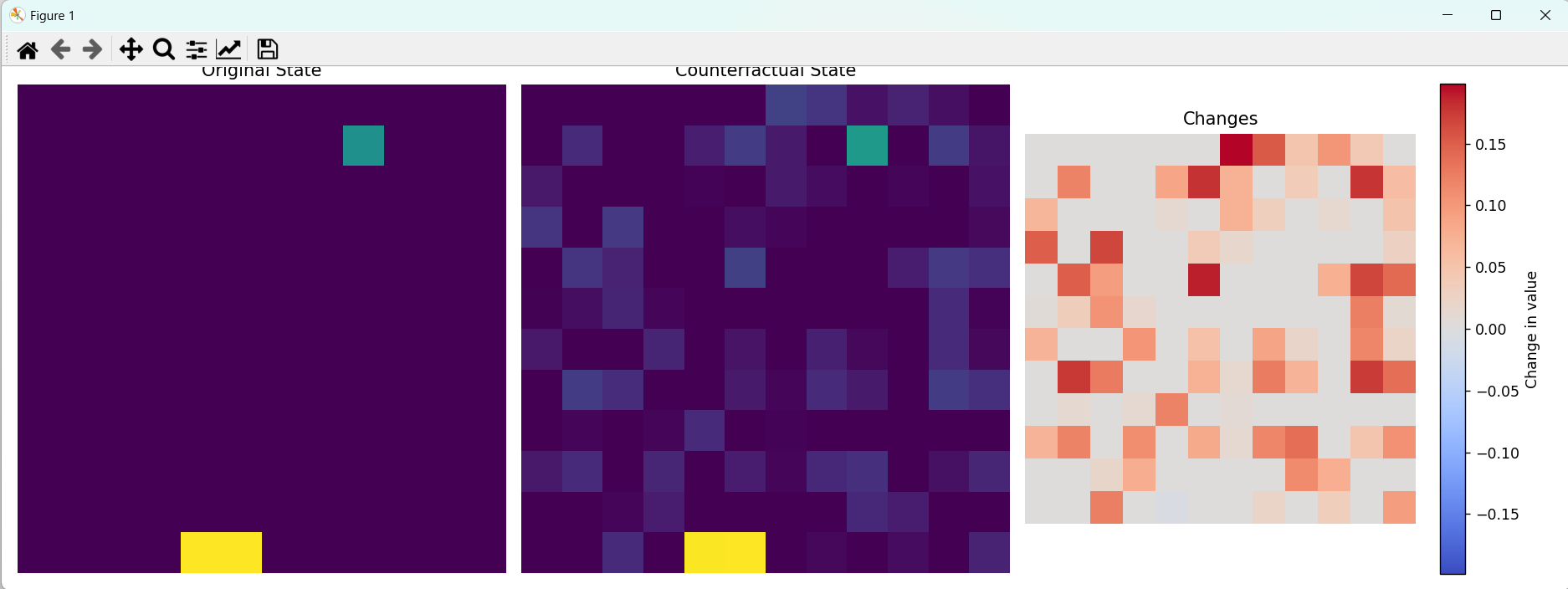
Model explanations (Qualitative observations):

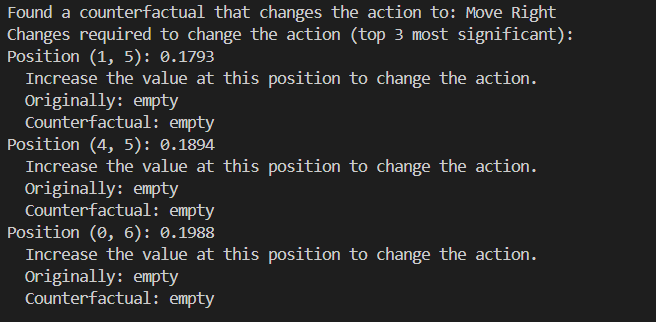
Here for the given state the model should have made a decision to move the basket to the right since the fruit is in the right side but it wrongly moves it to the left.





While trying to understand this with the SHAP values that were produced we find that the position of the basket [ (11,4),(11,3)] has a positive influence on the decision that the model made and the model wrongly considers that position has a fruit. This made us understand that the model performance is poor and thus it was trained for more number of epochs.





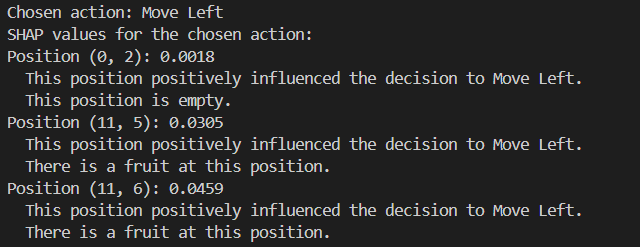
In case of such a wrong decision for a state we perturb the current state and find the nearest state that gives the correct decision. The second graph of the first figure shows the perturbed state and the third graph shows the changes to be made in the original state’s weights to achieve this. We have also given the textual representation of what position values should be changed how much to achieve the correct decision so that the user can use this data for future training purpose.

(ii)10 epochs:

Evaluation metrics of the model:

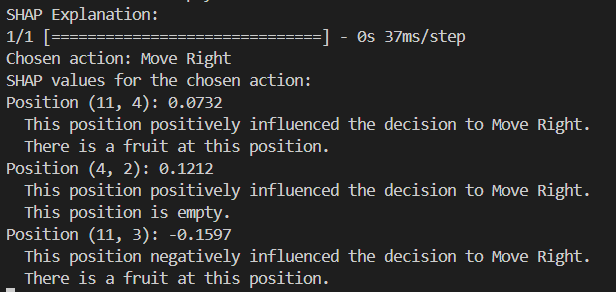
Model explanations (Qualitative observations):



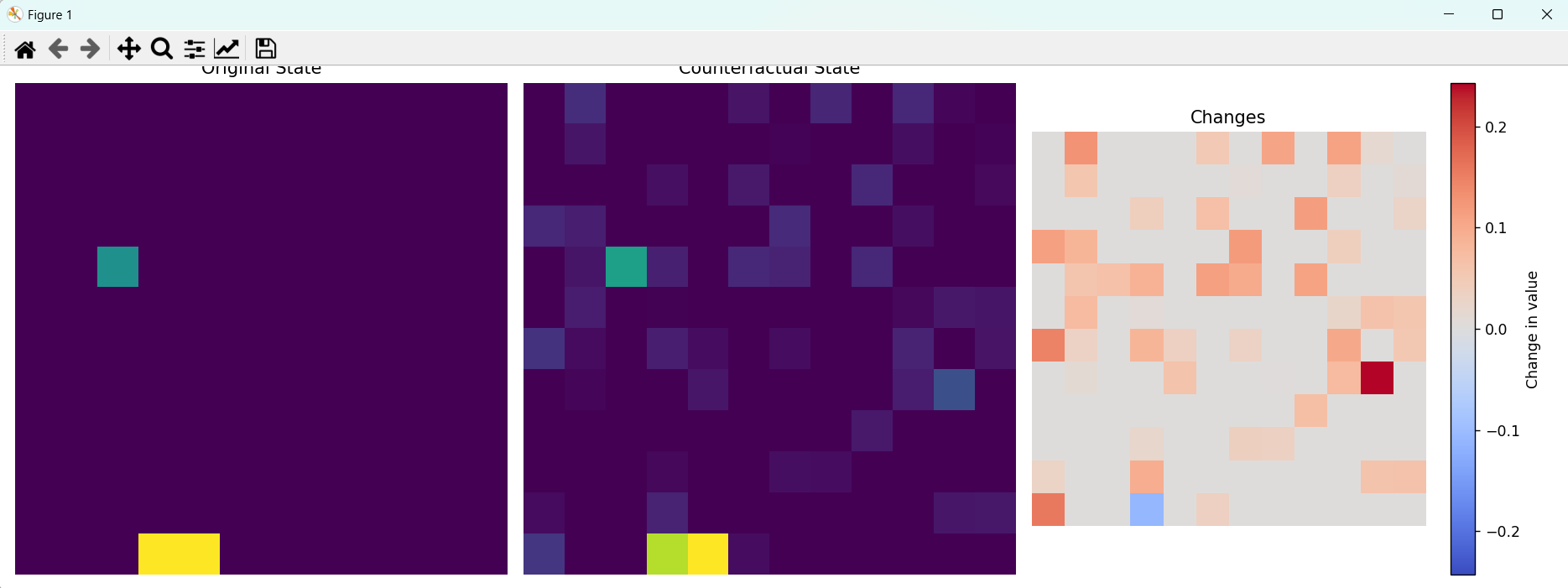


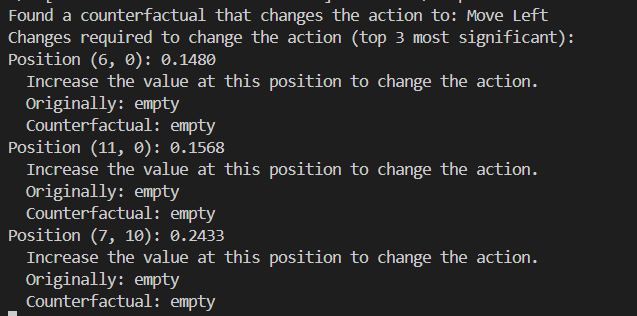
We observe here that with a bit more training the position of the block has also started to positively affect the decision of the model a bit and thus the model has started to make proper decisions for few states.

But it is not consistently performing well for all the states and it does make few bad decisions like the one given below: 



Counterfactual state given to understand what parts should be focused more to get the correct decision.





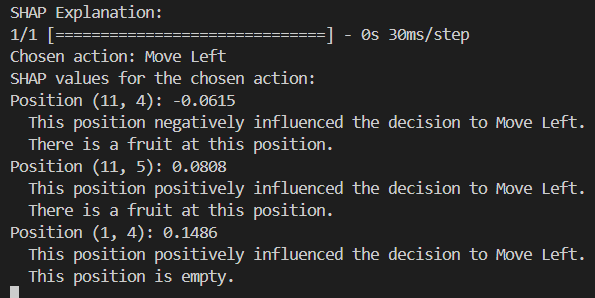
(iii)100 epochs

Evaluation metrics of the model:

Model explanations (Qualitative observations):

Here we see how the there is a drastic change in the importance given to the position of the fruit and how it affects the decision the model in a majority. With this we can analyse and improve the performance of our model using various explanations for the behaviour it shows.





FUTURE WORKS:

Future developments can be with the help of the counterfactual explanation the DQN structure can be improved so that appropriate weight assignment takes place in the final model.



