Inferring Human Task Representations

Research By:
Dylan Larson

Supervised By:
Elizabeth Zavitz, Pamela Carreno, Michael Burke

MONASH University

PURPOSE

In the evolving landscape of human-robot collaboration, understanding human task representation is crucial for enhancing robotic systems' ability to emulate human-like behaviours.

The aim is to investigate how humans and robots represent tasks, and thus how to improve the modelling of tasks for robotic systems.

ENVIRONMENT AND DATA

The research is performed within the environment of a cooking game, Overcooked-AI (simplified version of Overcooked), where teamwork between players is key to success. The human data was collected through Mechanical Turk, which has 172 collaborator games across 5 unique maps, each contains 3 mins (1200 frames) of game states.

Image 1. Overcooked Game

METHOD

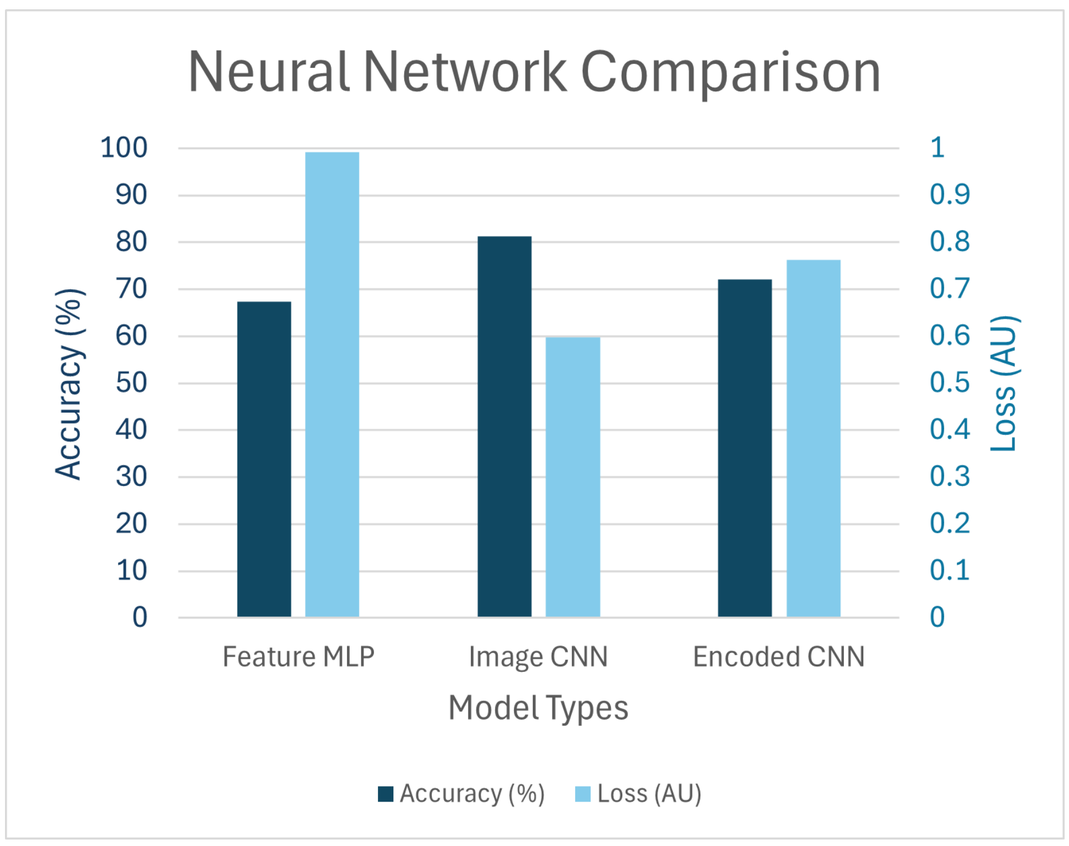
Neural networks were used to model the prediction of actions of a given player (Blue player). The data was processed to exclude frames with redundant actions and split by trial into a train, validation and test set.

There were three distinct models created:

1. Multilayer Perceptron (MLP) - feature vector input
2. Convolutional Neural Network (CNN) - image state input
3. Convolutional Neural Network - spatially encoded input

Other pretrained models such as ResNet18 and ViT were trialled. However, initial results of the models after a small number of epochs did not prompt further investigation.

Saliency analysis is applied to each of the models in order to find significant features of the inputs. (Image 2)



RESULTS AND DISCUSSION

The saliency of the best model (Image CNN) dominantly focused on the main player. Analysing the rankings of types of map tiles, the ranking of salient tiles was found to be:

1. Player One
2. Player Two
3. Pot/Stove
4. Bench
5. Dish Source
6. Empty Tile
7. Onion Source
8. Serve Location

The rankings display the **higher saliency trends for dynamic elements** in the game. Conversely, static elements result in lower saliency trends.

In future work, human trials with eye tracking may be evaluated. It is hypothesized that humans will create similar heatmaps and saliency responses as the models.

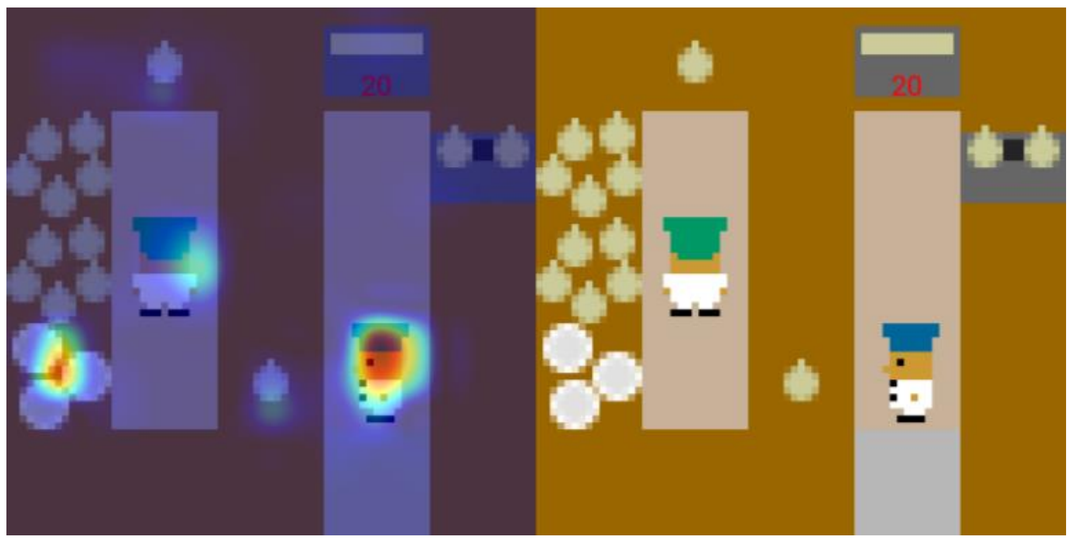


Image 2. (Left) Saliency Heatmap . (Right) Input Image

CONCLUSION

The more salient tiles generally consist of the main player and dynamic elements. This may indicate that these elements are more important to a task representation.

Human task representations have been shown to be value guided. These results demonstrate the models are sensitive to dynamic elements which appear to provide more value for the completion of tasks.