Cognitive Robotics Presentation

Paper 6: The eye in hand: predicting others' behaviour by integrating multiple sources of information

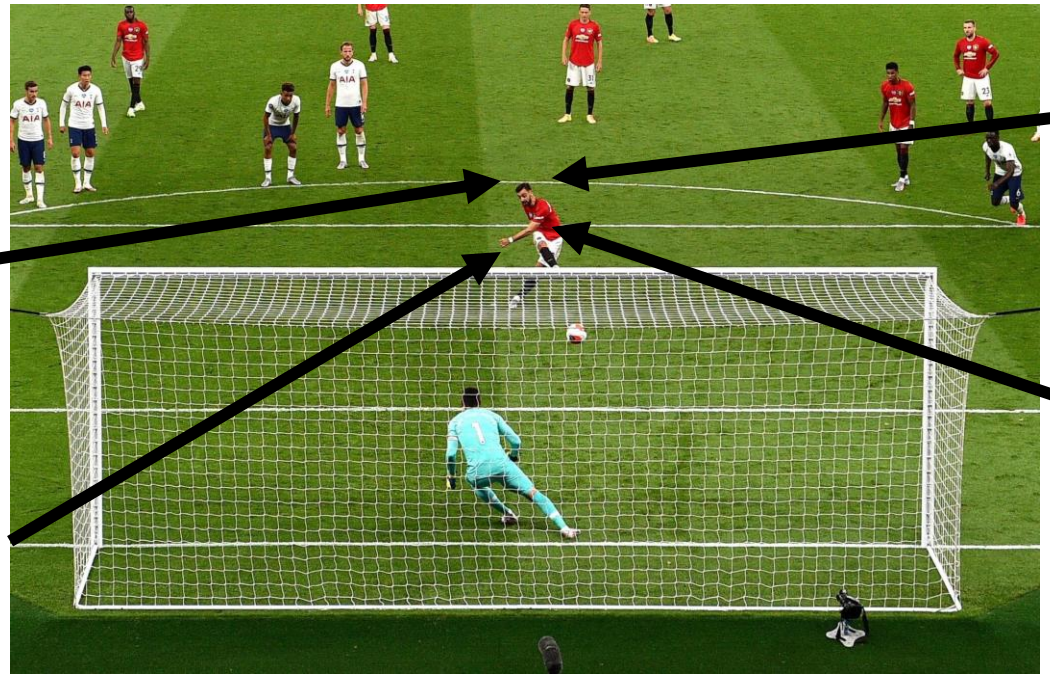
Michael Peres

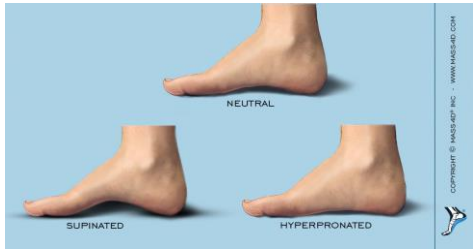
1) Description of Original Study

How do we predict the goal of another person as a human?

1) Description of Original Study

What information do we use to make our assumption?



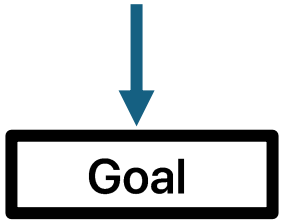


Which source of information do we favour more?

1) Description of Original Study

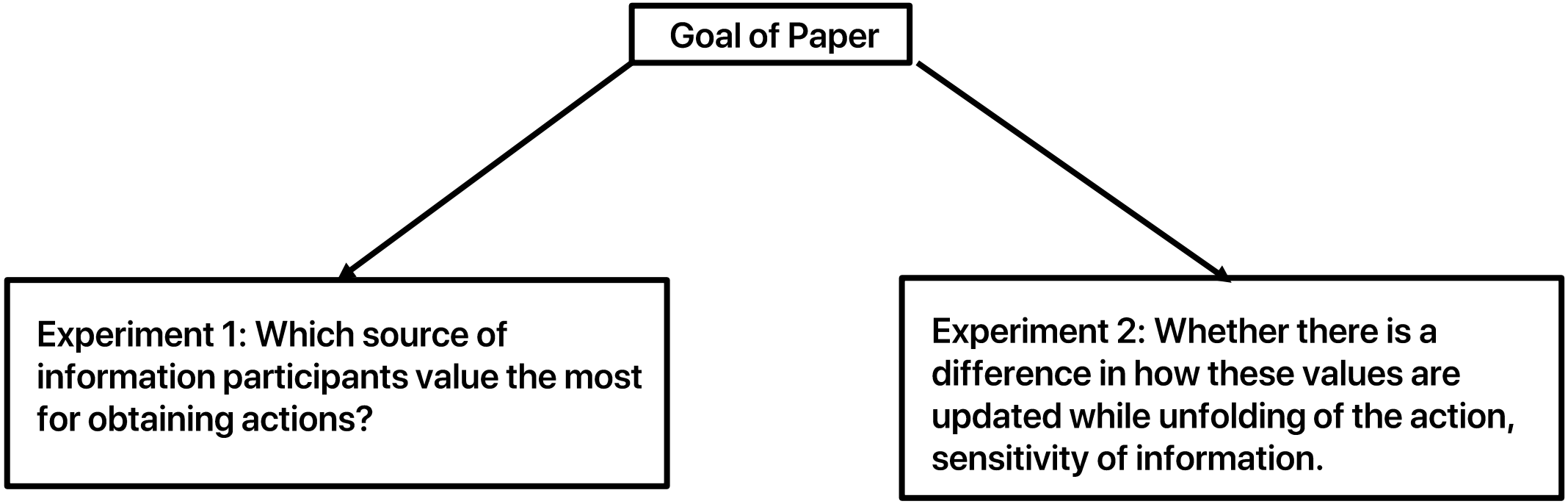
Predict whether the woman will pick up left or right object.

- Gaze Direction
- Arm Trajectory
- Hand Pre-shape

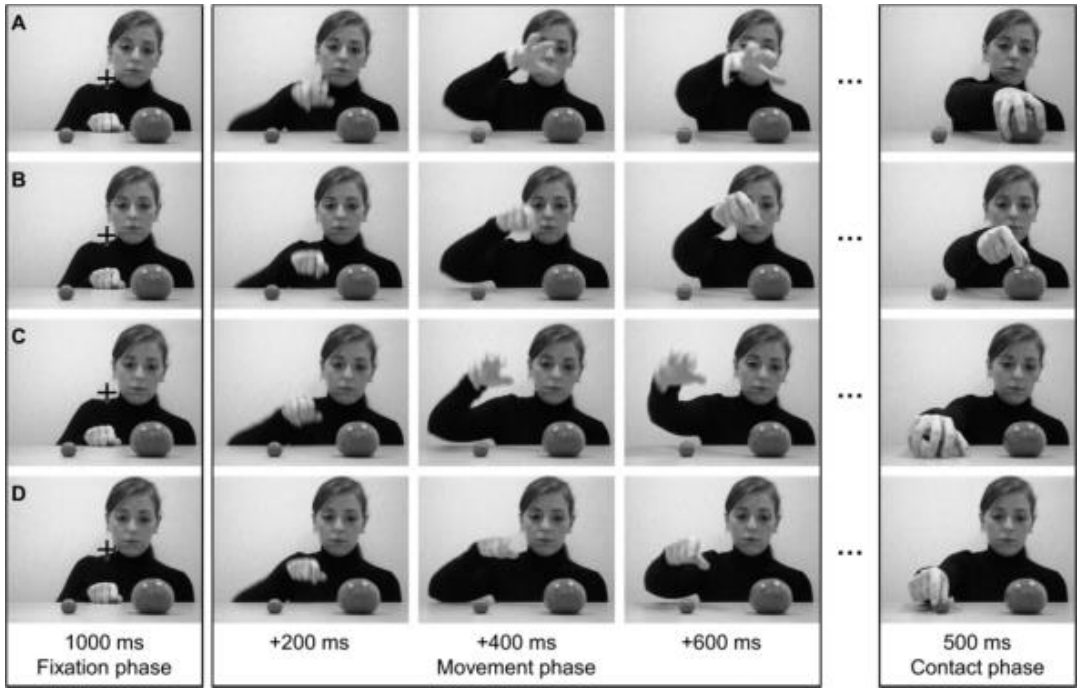


<u>Main objective</u>: Unknown in human processing is understanding how each information contributes to goal though process.

1) Description of Original Study



1) Description of Original Study



Type of Data Records in Dataset

Gaze	Hand Pre-shape
Congruent	Congruent
Congruent	In-Congruent
In-Congruent	Congruent
In-Congruent	In-Congruent

1) Description of Original Study

Experiment 1

Experiment 1: Which source of information participants value the most for obtaining actions?

Using the **metric of gaze arrival times** mentioned in the paper metrics, (the time of getting to goal AOI subtracted to final frame ending).

Conclusion

It shows that based on this metric, the effect of hand preshape aided in earlier gazing of the correct target goal. The model determined that when conflicting sources was used then participants gazed at video interaction gaze for aid, however in cases where preshape was congruent, participants were faster and more accurate regardless of the information provided by the actor's gaze.

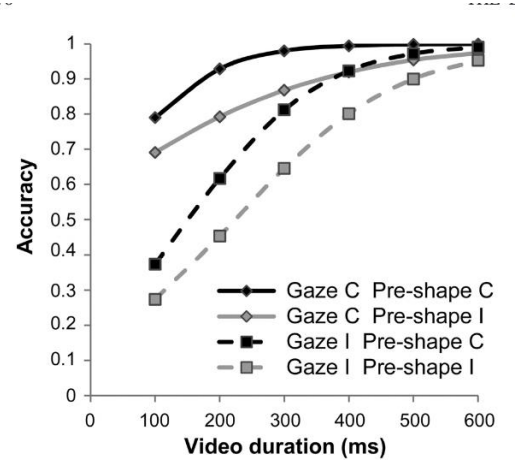
In the first experiment, we tried to find out which sources of information participants value the most in the goals, it concluded that the most viewed thing was looking at the videos gaze.

1) Description of Original Study

Experiment 2

Experiment 2: Whether there is a difference in how these values are updated while unfolding of the action, sensitivity of information.

It was understanding whether these source of information weightings were constant or not.



Conclusion

This is supported by the evidence showing that when the hand preshape correctly cues the actor's goal and/or the video duration increases, the information provided by the gaze decreases.

Experiment 2 show that, for longer videos, the importance lowers for gaze and rises for arm trajectory.

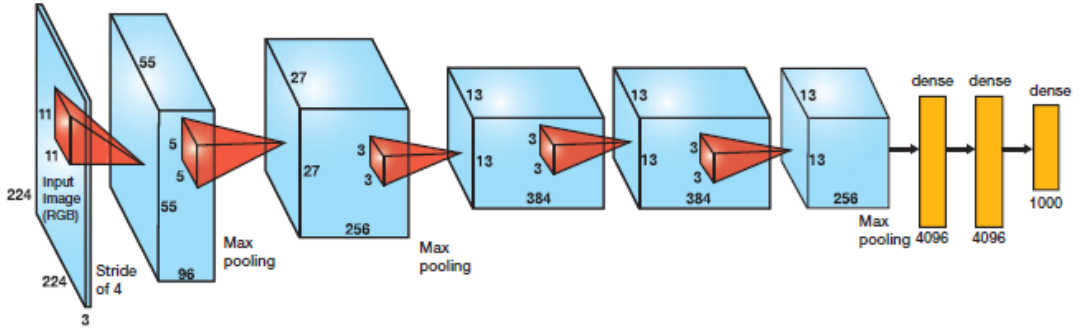
1) Description of Original Study

Results and Final Points of Paper

- Preshape was one of the main features even though reliability was 50%, and participants acted on their prior beliefs about precision of a source.
- As time progressed, participants relied more on arm movements such as arm trajectory.
- Over time of the experiment, actor's gaze was less reliantly used for accuracy, and was modulated by learning.
- Gaze information is only able to effect predicted outcome only when no other information about intention is provided.

2) Description of Computational Model

Aim of implemented computational model.



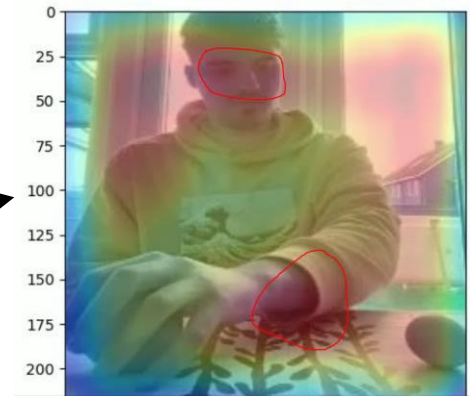
Output:

Goal is **LEFT**
Goal is **RIGHT**

Custom Made Alex Net for visualising kernel features.

Model Match Reasoning

Model will learn which features of image allow for the learning of the correct goal, finding what is important.
Different keyframes of video will allow the understanding of how the importance of features changes over the key frame.



2) Description of Computational Model

Dataset Creation

I created 18 videos of me obtaining items on my left or right, this was then chopped into 50 key frames, for a total of **900 images** for my dataset . Some preprocessing, image normalisation and resizing was done to allow input to the model.

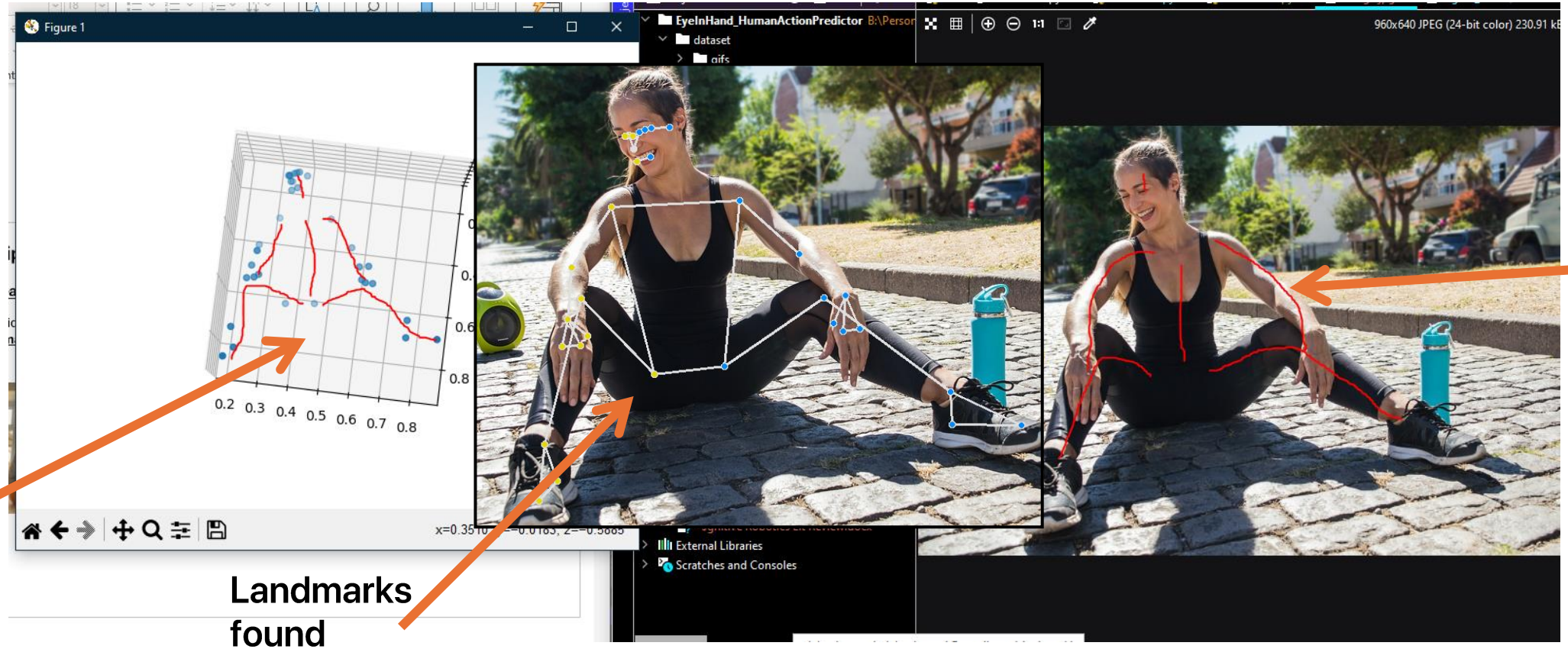
Example Dataset Video

Right object goal, frames from video.

2) Description of Computational Model

Landmarks Based Computation

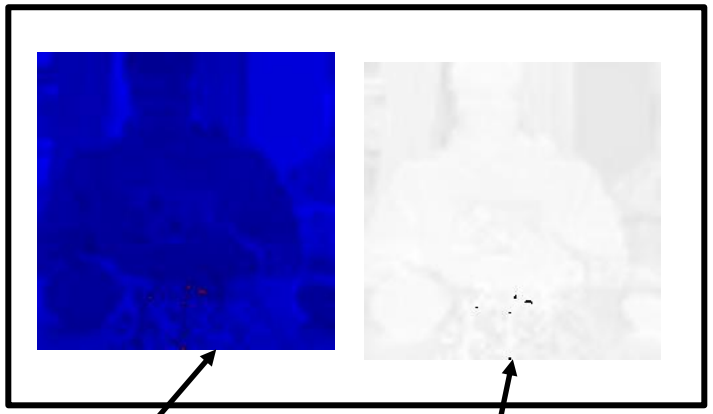
To better understand whether my model reliance on just visual cues was causing issues in terms of determining a goal, I wanted to see if knowing arm trajectory aided in the goal processes, and if this was useful in determining the goal, confirming the conclusion given.



3) Results and Conclusion

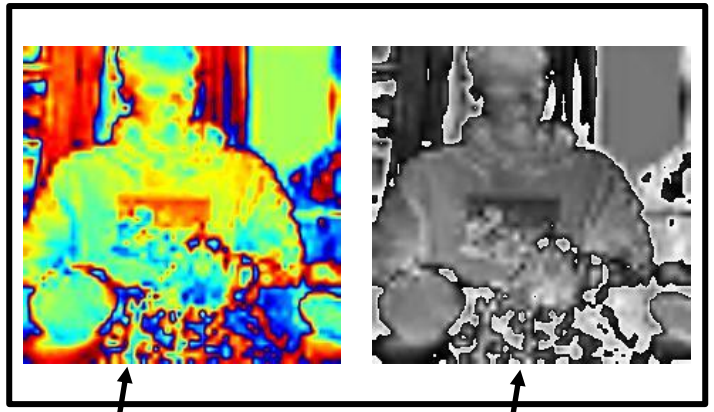
Understanding Model Learning (3 Iterations)

Iteration 1) (Very noisy and not enough information) Low Level Initial Convolution Layers Feature Maps, shows the fundamental blocks of CNN recognition for goal, however not specifically showing useful information.



Frame 10

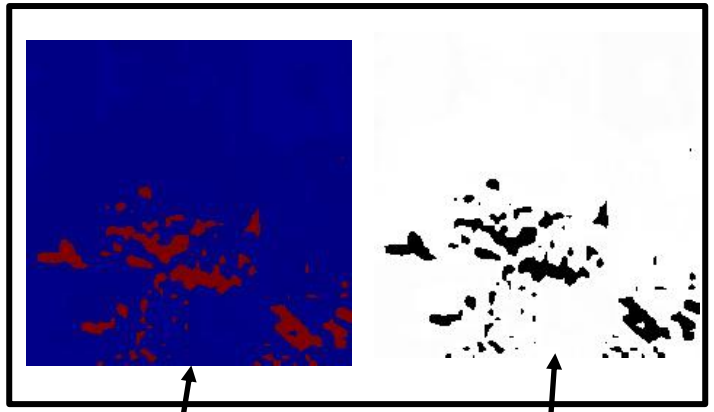
Fundamental kernels are highlighting the hand preshape for understanding the goal, like what is current mentioned in the paper.



Frame 23

Model overfits the data, with an image for associating goals to left objects, reliant on memorising image, model failure.

Steps were taking in next iterations to stop this from occurring.



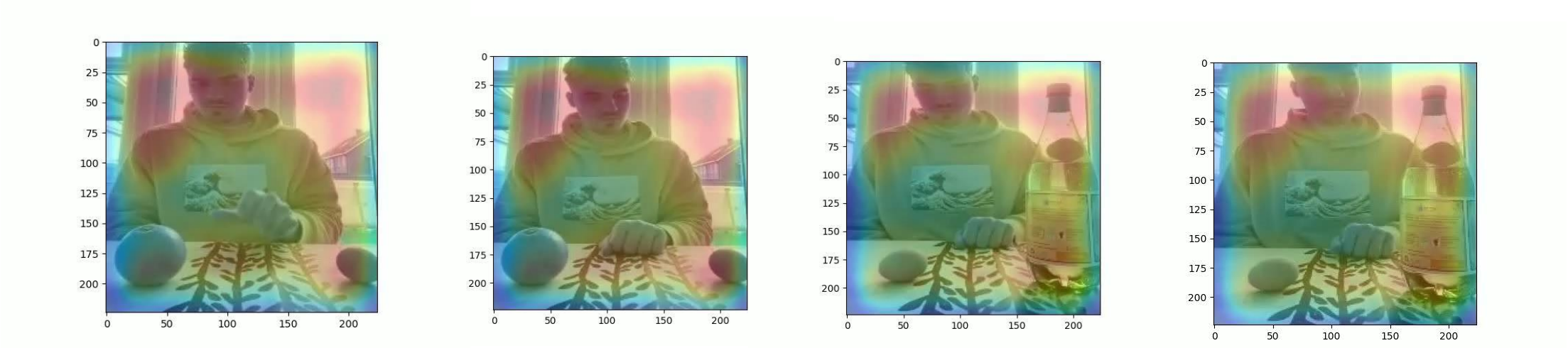
Frame 3

Model with initial frames relies on many random parts of jumper and body position to make assumption but not eyes.

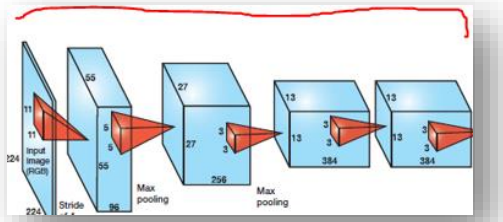
3) Results and Conclusion

Understanding Model Learning (3 Iterations)

Iteration 2) (Noisy Live Predictions) Combination of CNN outputs for last convolution, showing consensus of attention region when making decision (noisy and hard to understand.)

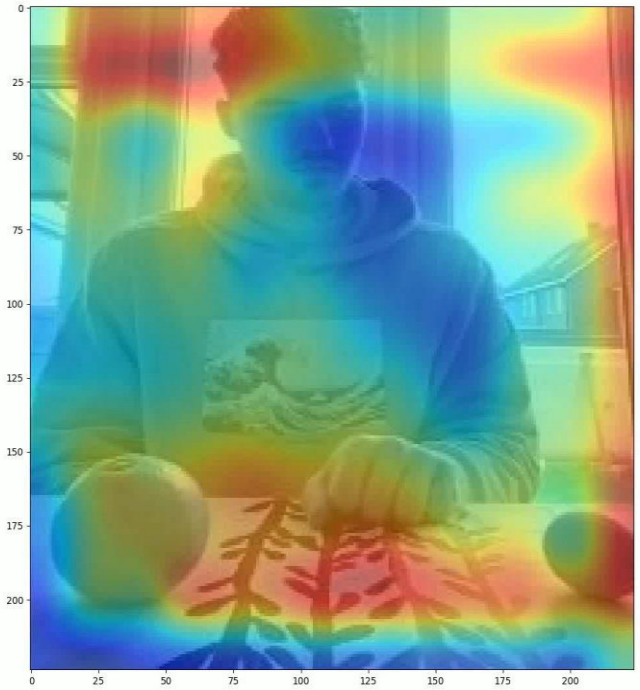


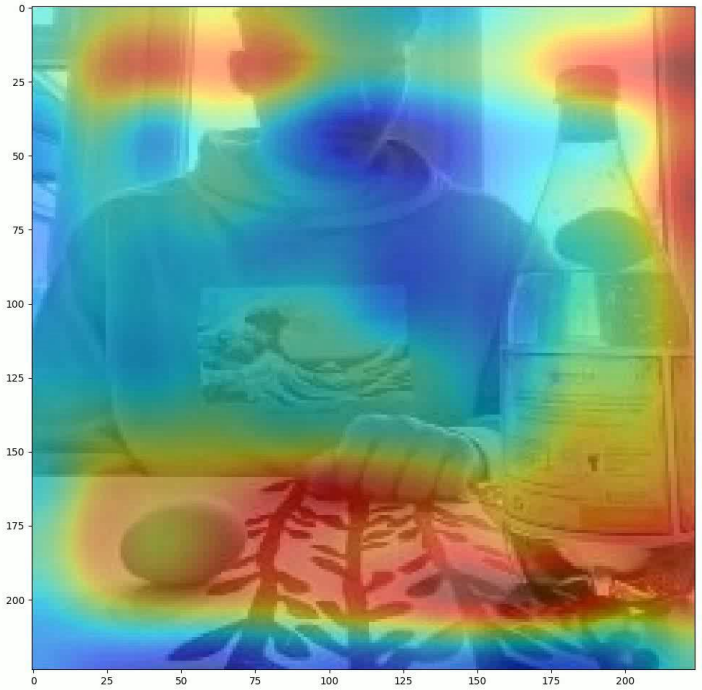
3) Results and Conclusion

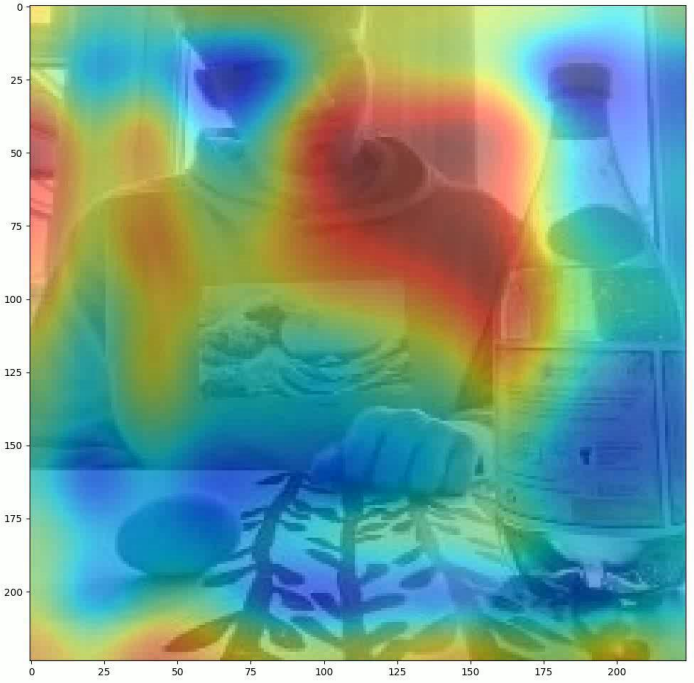


Understanding Model Learning (3 Iterations)

Iteration 3) (Accurate Weighted Convolutions) These visualisation accurately simulate the complete convolutions the model performs and the strongest signals (blue signals).



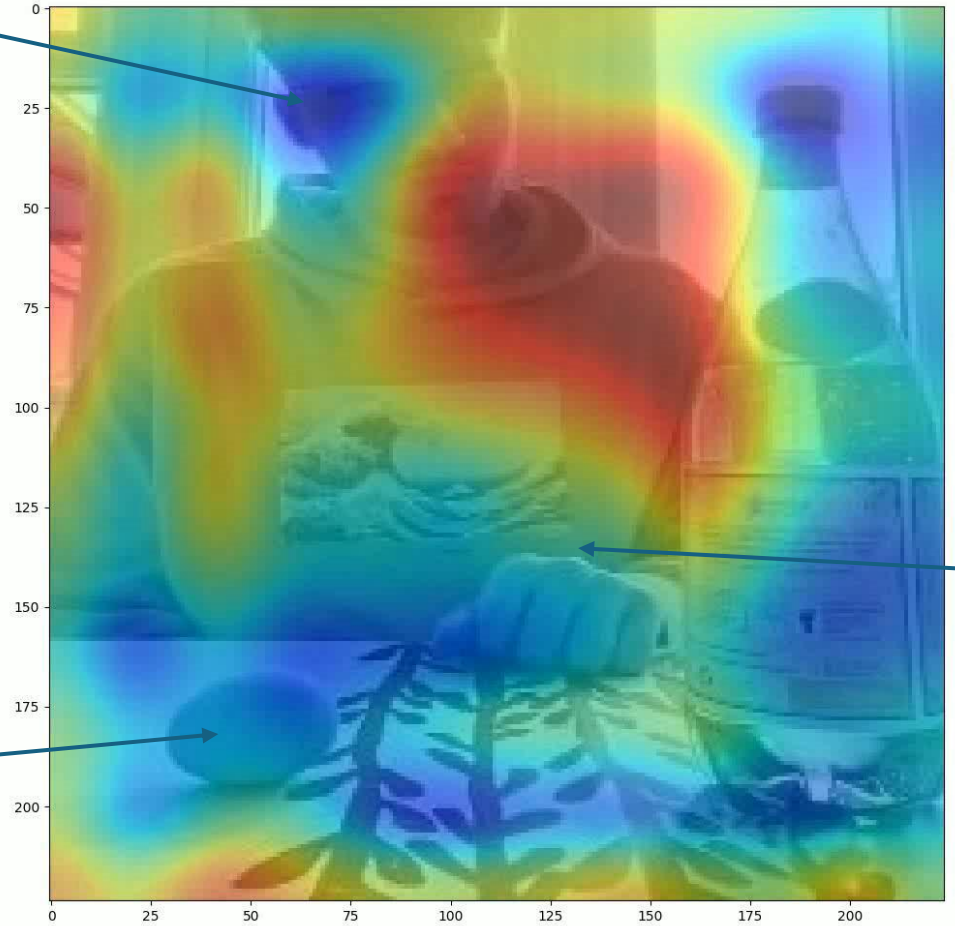




3) Results and Conclusion

Deeper Understanding

First obtains key information from eyes to make inferences.



Second the attention goes to preshape and is continuous. Instant removal of attention from eyes when new information is available.

Closer to goal, all attention is moved to the bottom left, meaning overall confidence in goal increases.

3) Results and Conclusion

Problems with Paper

- **The paper does not handle issues including that say dependencies such as empty cup or water bottle, in which a goal may be decided based on context rather than imperial features.**

- **Paper does not consider that participants, may perhaps remember specific scenes, such that a prediction becomes a remembering game rather than a completion.**

- As time progressed, participants relied more on arm movements such as arm trajectory.
- Over time of the experiment, actor's gaze was less reliantly used for accuracy, and was modulated by learning.
- Gaze information is only able to effect predicted outcome only when no other information about intention is provided.