Audio-visual deep reinforcement learning in image-oriented environments

Raufs Dunamalijevs
Supervised by Professor John McCall

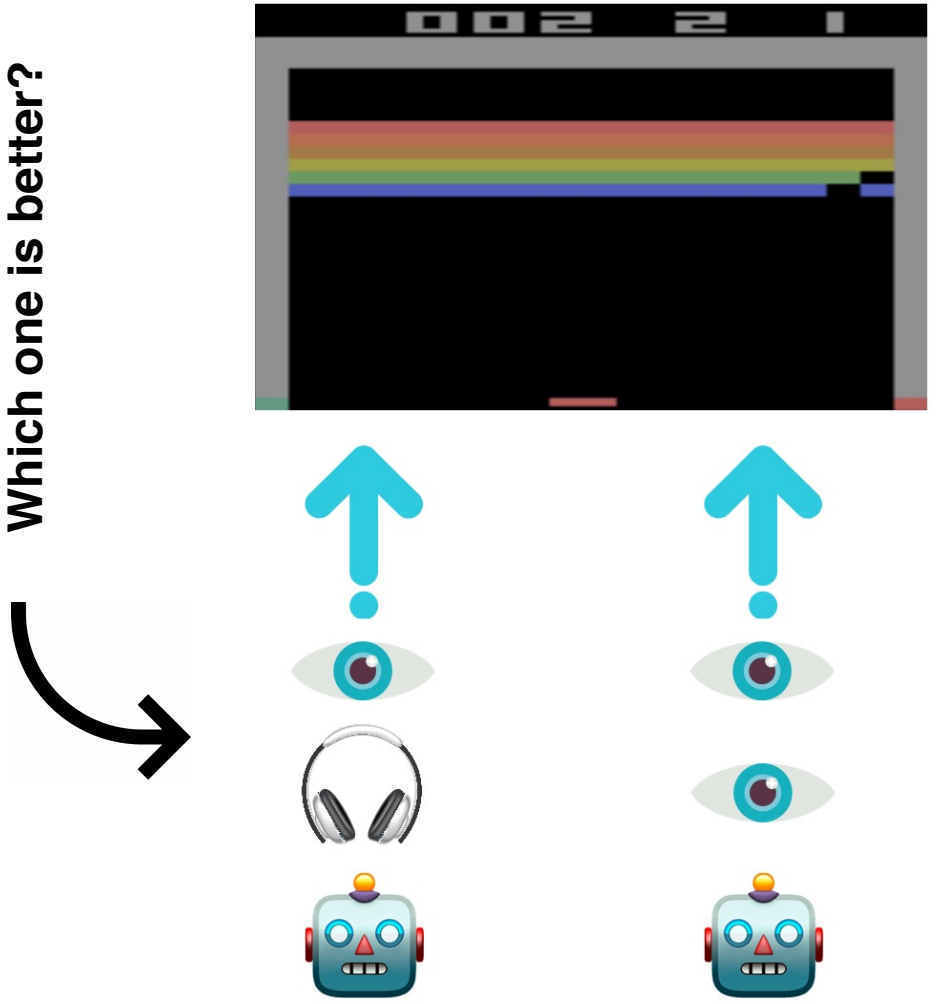
National Subsea Centre

THE ROBERT GORDON UNIVERSITY ABERDEEN

THE CARNEGIE TRUST FOR THE UNIVERSITIES OF SCOTLAND

Does the agent that has access to both audio and video perform any different to the agent that has access only to video in environments where most useful information can be seen?

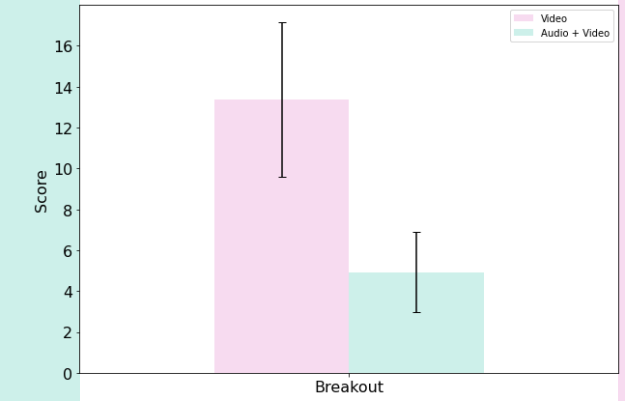
Reinforcement learning is the area of machine learning that considers the problem of computational agent learning to make decisions by trial-and-error. Deep learning is another area of machine learning that considers the problem of learning multiple levels of representation and abstraction using high-dimensional input data like images. Deep reinforcement learning is the combination of two, which allows the computational agent to solve complicated tasks using a trial-and-error approach to learn different levels of abstraction from the raw data.

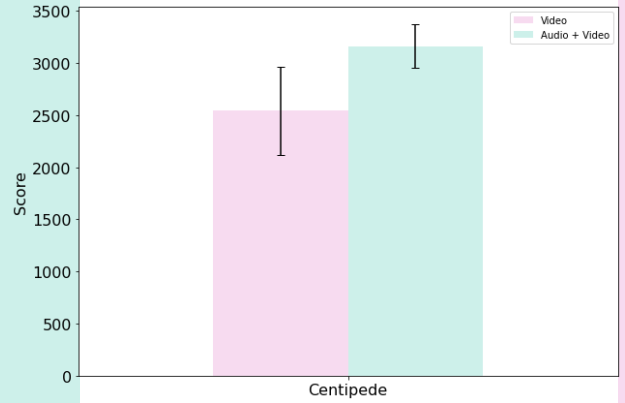


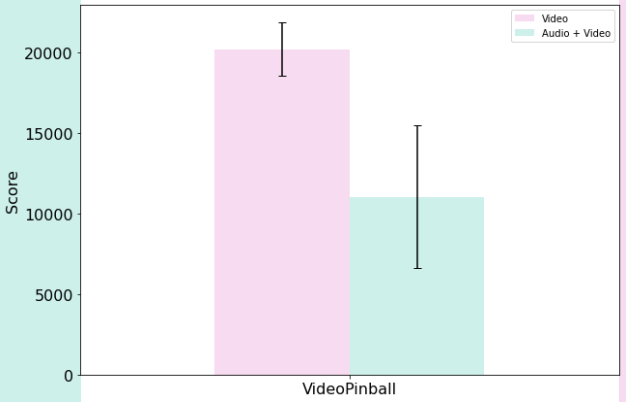
Methodologies and Conclusions

- The code that makes it possible to train an agent using audio from Atari 2600 emulator was written
- The agents that can both hear and see and only see were trained in **3 image-oriented games** using an efficient PPO algorithm for 10M frames
- These games are **Breakout, Centipede and VideoPinball**
- Score averages of each trial were computed using the 100 final episodes
- The experiment was repeated 4 times with different random number generator seeds
- Welch's t-test was performed to test the null hypothesis:

The provision of audio-visual information provides no significant advantage or disadvantage over visual-only information for deep reinforcement learning in image-oriented environments



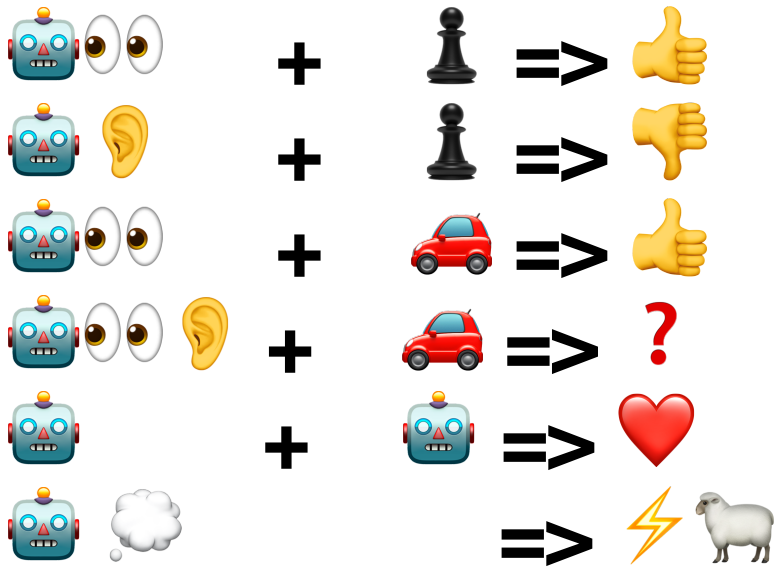




- The difference between visual and audio-visual agents was **statistically significant** in 2 out of 3 games
- The agent that could hear on average performed **worse**
- However, it performed better in Centipede, although this was not statistically significant (p=0.054)
- Interestingly, Centipede conveys more useful information in the form of sound compared to the other games
- ***Potential work***: if the Centipede agent was trained on more frames and/or more trials were made, the better performance of audio-visual agent over visual-only agent could become **statistically significant**
- This could ultimately prove that audio can have both a positive and a negative effect on the agent depending on the environment

Why is it so important?

Usually, it is dead simple to determine what input to use to train an agent. If it is learning to play chess – show it some games. If it is learning to generate music – let it listen to Simple Minds.
However, it is not always so easy to determine, which inputs will benefit an agent and which ones will not.



One such example is autonomous driving. Do we allow the agent that drives a car to hear what is happening on the road or not? It could help to avoid road accidents in certain situations, but on the other hand, so much information could also confuse it. Maybe there is no difference at all? This is the question that we asked in this research and the null hypothesis that we successfully tested and rejected.

Such a hypothesis is very challenging to test in real life as it is expensive, unethical and dangerous. However, we could test a similar scenario by relaxing constraints of the problem to get foundational baseline knowledge with low risks and costs.
This could potentially give the necessary impetus for more developments in multimodal reinforcement learning and help to create smarter autonomous agents that are fit for purpose.

For raw data, visualizations, code, data processing and everything else that would never fit here, please visit: