Combining audio and visual features to play GeometryDash

P. Bakker, J. R. T. E. van der Hout, J. N. de Vries, R. van der Wal, and M. M. van der Wel

Introduction

Many deep learning networks have been trained to teach game agents how to play games. These networks are usually trained using images. It has been shown that game immersion improves for players when using audio [3]. A game agent might also benefit from using audio to learn how to play the game.
This project aims to investigate if an audio network can be trained that achieves a high AUC and accuracy, and if combining image and audio networks will improve the AUC and accuracy.
This will be tested using the game Geometry Dash, which has been previously played using an image network by Li and Rafferty [1].

Research

Research question:

Will adding audio to an image network increase accuracy and AUC?

Hypothesis:

We believe combining an audio and image network will increase accuracy by a maximum of 5%.

Experimental setup

Networks trained using imitation learning [2]. Audio and image datasets acquired by recording gameplay, automatically removing bad `deah frames' . Raw audio network standard architecture [4]. Image network based on network by Li and Rafferty [1].

Image network:
- Dropout 0.25 on all layers
- 1000 epochs
- Train:validation ratio is 9:1
- Test created separately
- RMSprop
- Learning rate 0.05

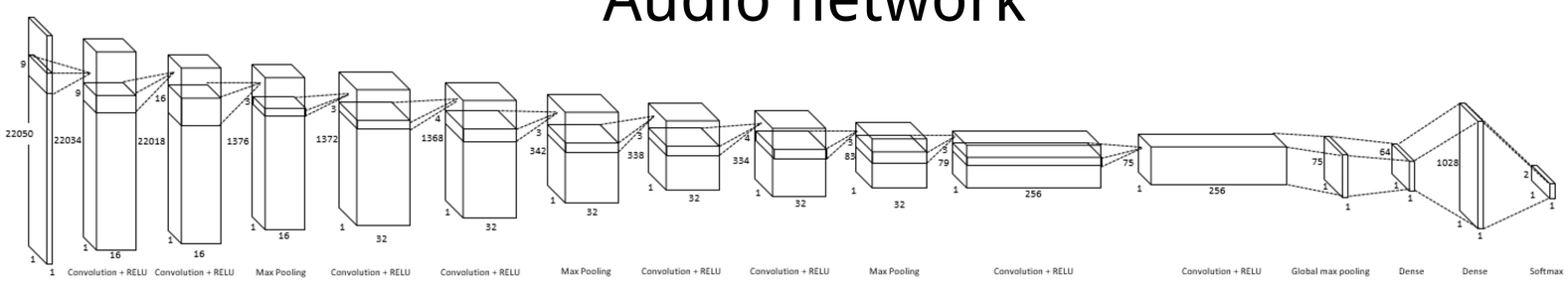
Audio network:
- Dropout 0.1 on all but last block (0.2)
- 10 epochs
- Train:validation 9:1
- Test created separately
- Adam
- Learning rate 0.001

Image data:
- 15 000 images
- Downscaled to 60x80
- Greyscale
- Prediction every 0.1s

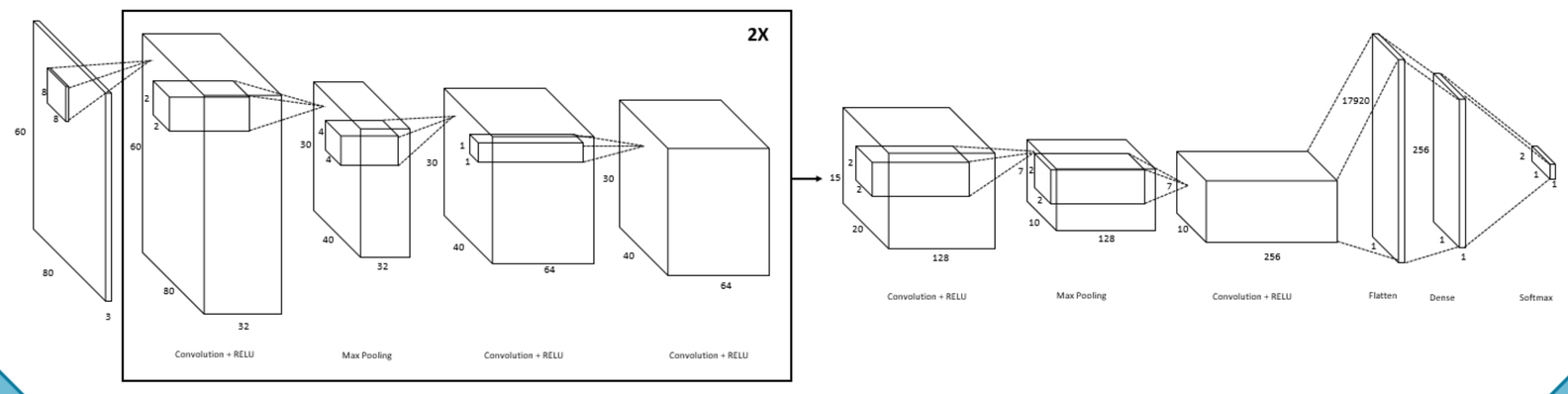
Audio data:
- 1500 audio samples of 2 seconds
- 2 second sliding-window
- Raw audio
- Prediction every 0.1 s

Networks

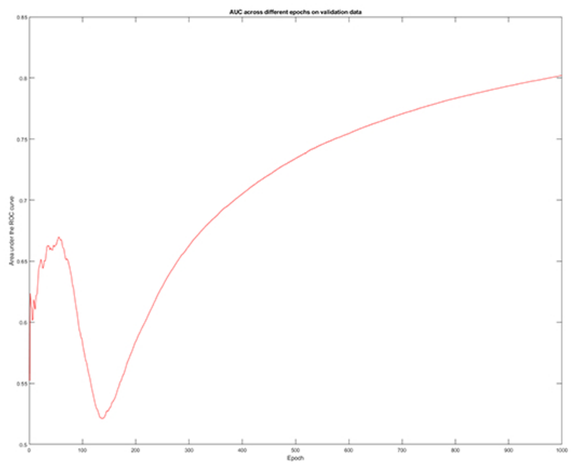
Audio network

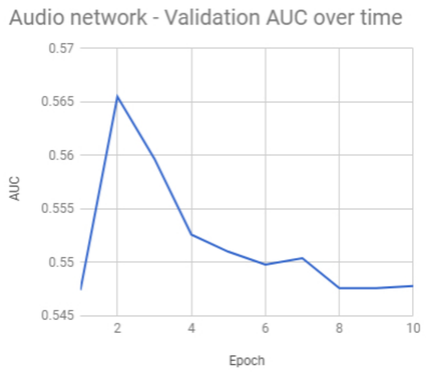


Visual network



Experimental results





Denk aan max 4 grafieken hier met text eronder, maar nu zijn ze te groot. Maybe combine auc_val en acc_val in 1 grafiek met legend, dan 2 grafieken tot.

Discussion

Due to issues with synchronization, the recorded test data is incorrect. Further research needs to be done to deal with the delay caused by the SerpentAI framework.
Raw audio features doesn't seem to be enough to detect the beats needed to jump, further research could be done using MFCC features.
Playing the game at real time is also an issue, as our CPUs are not fast enough to process bigger networks or get enough predictions per second.

Conclusion

As shown in the results, the audio network does not improve the prediction as expected, this is due to the delay and inaccurate training data.

References

[1] Li, T., & Rafferty, S. (2017). Playing Geometry Dash with Convolutional Neural Networks (p. 7). Stanford University
[2] Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. In Advances in Neural Information Processing Systems
[3] Collins, K. (2013). Playing with sound: a theory of interacting with sound and music in video games. Mit Press.
[4]Beginner's Guide to Audio Data | Kaggle. Retrieved from https://www.kaggle.com/fizzbuzz/beginner-s-guide-to-audio-data