## Abstract

There has been significant development in producing autonomous vehicles but a growing concern is in understanding how these systems work, and why certain decisions were made. This has a direct impact on the safety of people who may come into contact with these systems.

This research reproduced the experimental setup for an end-to-end system by Bojarski *et al.* [2016b]. End-to-end self-driving AI structures are built on top of black-box machine learning techniques. The source code can be found here: https://github.com/TRex22/masters-gradsum.

An allure of end-to-end structures is that they need very little human input once trained on large datasets and therefore have a much lower barrier to entry, but they are also harder to understand and interpret as a result.

Bojarski *et al.* [2016b] defined a reduced problem space setup for a self-driving vehicle. This task only has a forward-facing camera which generates RGB images as input, and only the vehicle's steering angle is the output. This work expanded the setup to include six CNN model architectures over the single model used by Bojarski *et al.* [2016b] to compare the behaviours, outputs and performance of the varying architectures.

There have been recent developments in applying attribution methods to deep neural networks in order to understand the relationship between the features present in the input data and the output. GradCAM is an example of an attribution technique which has been validated by Adebayo *et al.* [2018].

We devised an attribution analysis scheme called GradSUM which is applied to the models throughout their training and evaluation phases in order to explain what features of the input data are being extracted by the models. This scheme uses GradCAM and uses segmentation maps to correlate inputted semantic information using the resultant gradient maps. This produces a model profile for an epoch which can then be used to analyse that epoch.

Six models were trained, and their performance compared using their MSE loss. An autonomy metric (MoA) common in literature was also used. This metric tracks where a human has to take over to stop a dangerous situation. The models produced good loss results. Two model architectures were constructed to be simple in order to compare against the more complex models. These performed well on the loss and MoA metrics for the training data subset but performed poorly on other data. They were used as a control to make sure that the proposed GradSUM scheme would adequately help compare emergent behaviours between architectures.

Using GradSUM on the architectures, we found that three out of the six models were able to learn meaningful contextual information. The other three models did not learn anything meaningful. The two trained simple models' overall activation percentages were also close to zero, indicating these simple model architectures did not learn enough useful information or became over-trained on features not useful to safely driving a vehicle.