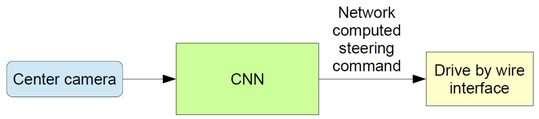
Attribution Methods for Explaining Deep Models for Self-Driving Cars ***(Accepted masters by dissertation)***

Jason Chalom (contact@jasonchalom.com)
Supervised By: Professor Richard Klein,
University of the Witwatersrand

Background

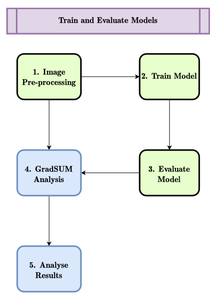
Autonomous vehicles have seen significant advancements, largely driven by black-box AI systems. Understanding and explaining these systems' operations are crucial for ensuring safety for those interacting with them.



A general structure of an end-to-end autonomous vehicle system[1]

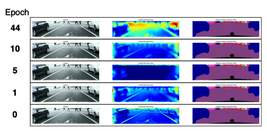
End-to-end architectures typically used in self-driving systems rely on deep neural networks and are trained on large datasets. Analysing these black-box models can be challenging due to their complexity.

One technique to analyse these architectures is called GradCAM[2] which generates attribution maps. These maps visually represent the components of the input images with the highest importance.



General GradSUM Workflow

Example data samples from the Cityscapes Dataset[3]

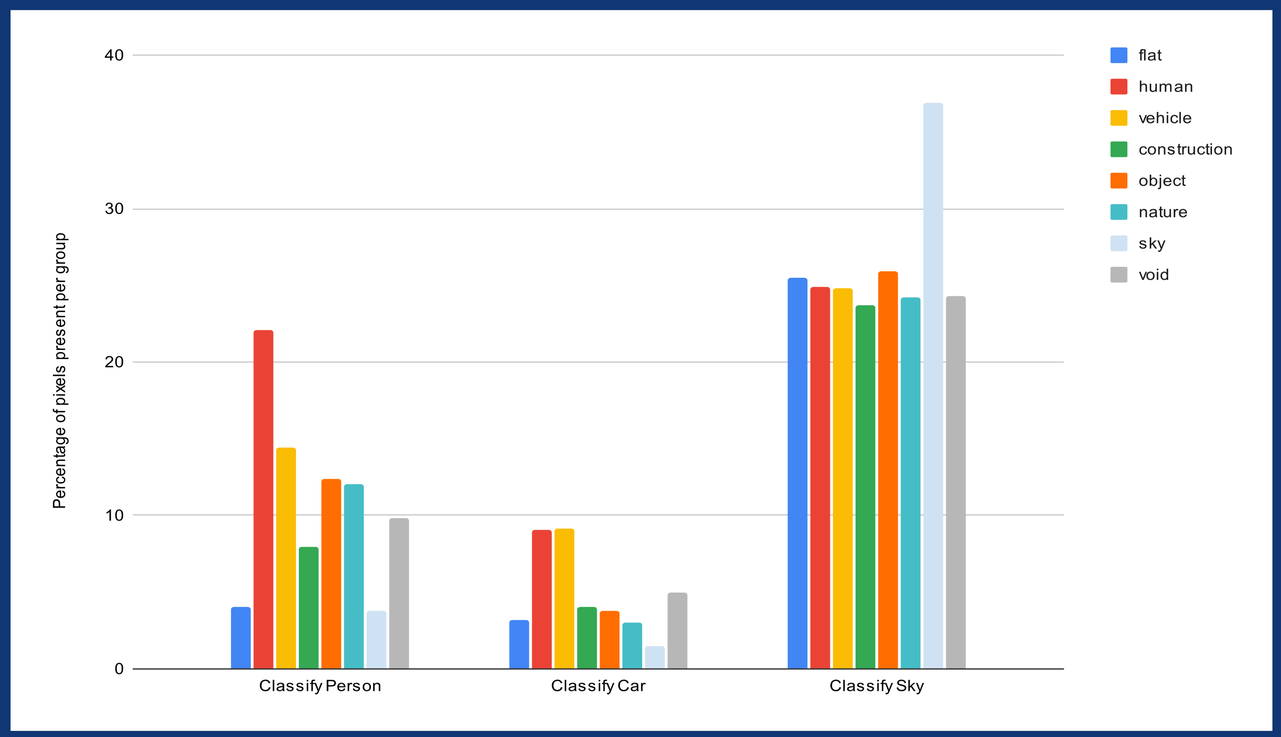


Example GradCAM Results for the NetSVF model over training epochs

Methodology

- Selected 3 training datasets: Udacity, Cityscapes, Microsoft AirSim Tutorial
- Selected 2 datasets with *fine* segmentations for GradSUM analysis: Cityscapes, Fromgames
- Selected 4 model architectures from existing work and devised 2 control model architectures
- Pre-processed datasets using the same methodology as prior work, mitigating temporal leakage
- Train and evaluate every model architecture ten times
- Random initialised model for 100 evaluations

We devised a new **analysis scheme** called **GradSUM**, using **semantic** data to produce model **profiles** to **interpret** AI behaviour



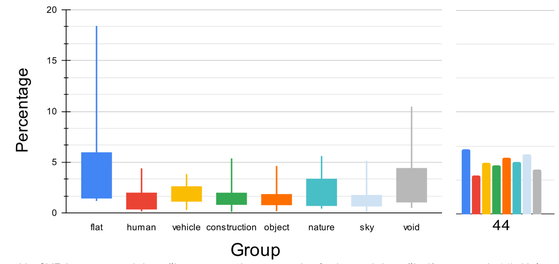
Three trained deep CNN models (NetSVF) on different classification and their respective model **profiles** using **GradSUM**

Scan for the full dissertation and supplementary materials

WITS UNIVERSITY
PRIME AI LAB
RAIL LAB

Algorithm 1 An algorithm for the GradSUM scheme

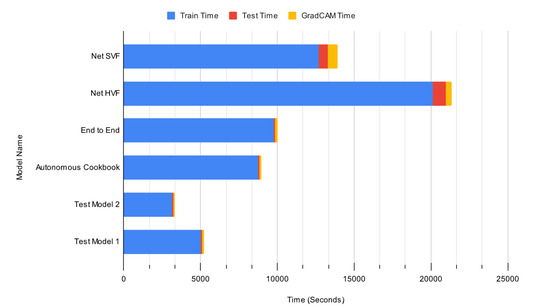
$K \leftarrow \{K^{\text{flat}}, K^{\text{human}}, \ldots\}$ ▷ The available segmentation groups in ground truth dataset
for k in K **do**
  for w, h in InputImage **do**
    if InputImage[w,h] is in group k **then**
      $P[k][w][h] \leftarrow 1$
    else
      $P[k][w][h] \leftarrow 0$
    end if
  end for
  $N[k] \leftarrow \text{Sum}(P[k])$ ▷ This is the sum of all pixels present for the given group k
  $M \leftarrow \text{GradCam(InputImage, model)}$
  $G[k] \leftarrow \frac{P[k] \odot M}{N[k]} \cdot 100$ ▷ G is the GradSUM result, and $\odot$ is the element-wise product
end for



NetSVF Average model profile group and an example single model profile (from epoch 44). Using The Cityscapes Dataset and GradSUM

Results

- Showed that End-to-End AI models are capable of learning to control vehicle steering angle prediction
- Demonstrated models' sensitivity to noise and training conditions
- Highlighted the usefulness of this analysis scheme in improving interpretability and reproducibility of GradCAM analysis
- The time cost of GradSUM is minimal in terms of the total cost of training and evaluating these models



Experimental time cost of training, evaluation (test) and generating GradCAM maps for GradSUM analysis

Future Work

- Investigate the usefulness of GradSUM for large language models and natural language processing use cases
- Investigate attention maps and their similarity to attribution maps for the GradSUM scheme
- Investigate the generation of useful semantic datasets to use with GradSUM
- Investigate GradSUM used as a training optimiser

References

[1] Bojarski, et al 2016. End-to-end learning for self-driving cars. *ArXiv*, abs/1604.07316.
[2] Selvaraju *et al.* 2016. Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2016.
[3] Cordts *et al.* 2016. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.