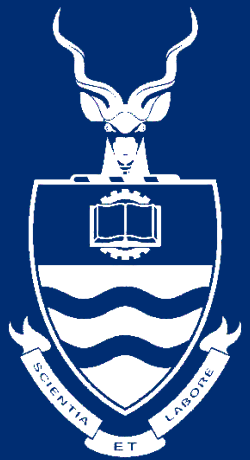


The Application of Attribution Methods to Explain an End-To-End Model For Self-Driving Cars (Unpublished)

By: Jason Chalom (711985)

Supervisor: Dr Richard Klein

School of Computer Science and Applied Mathematics

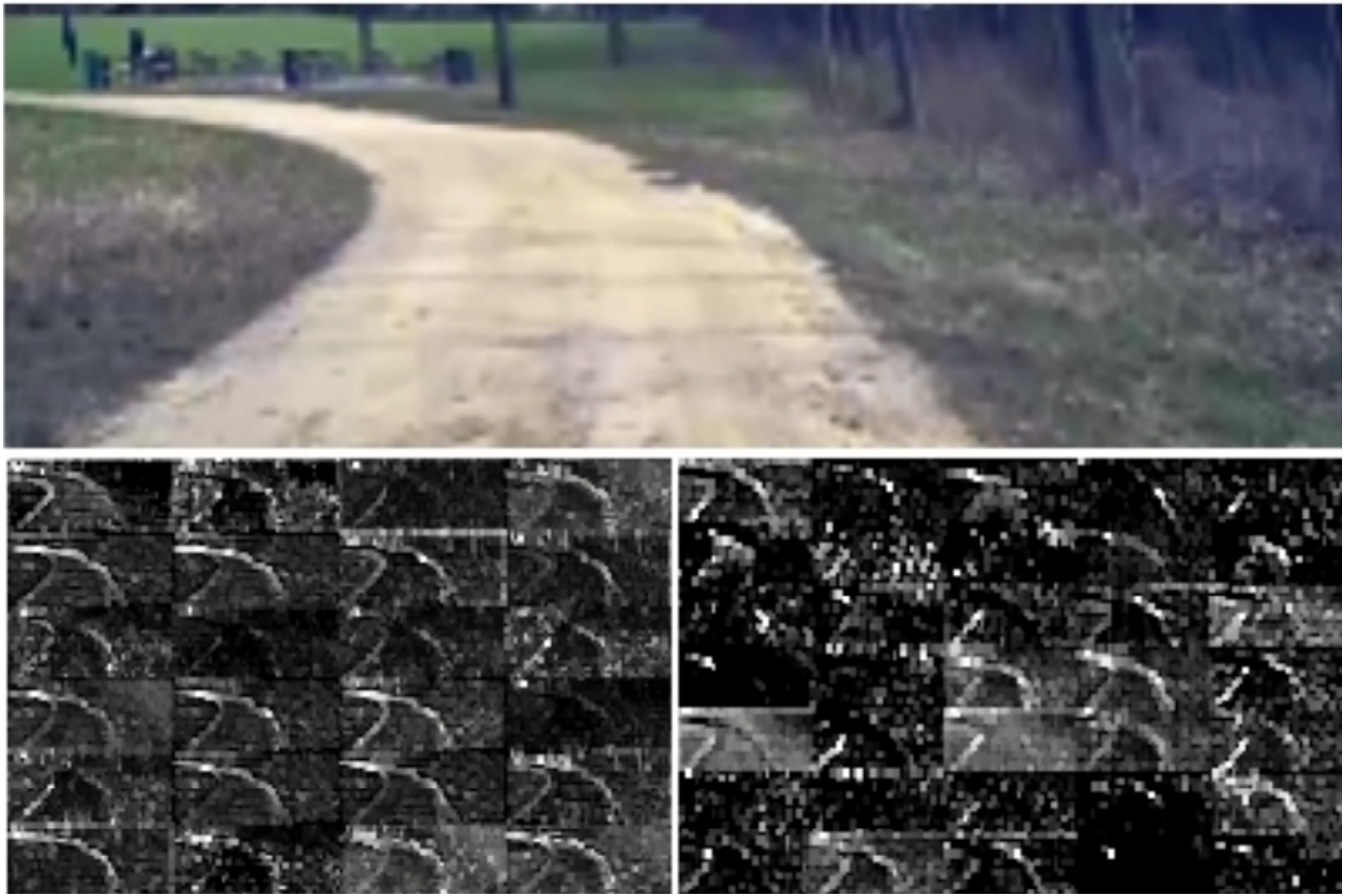
Masters Candidate at Wits

Senior Software Engineer

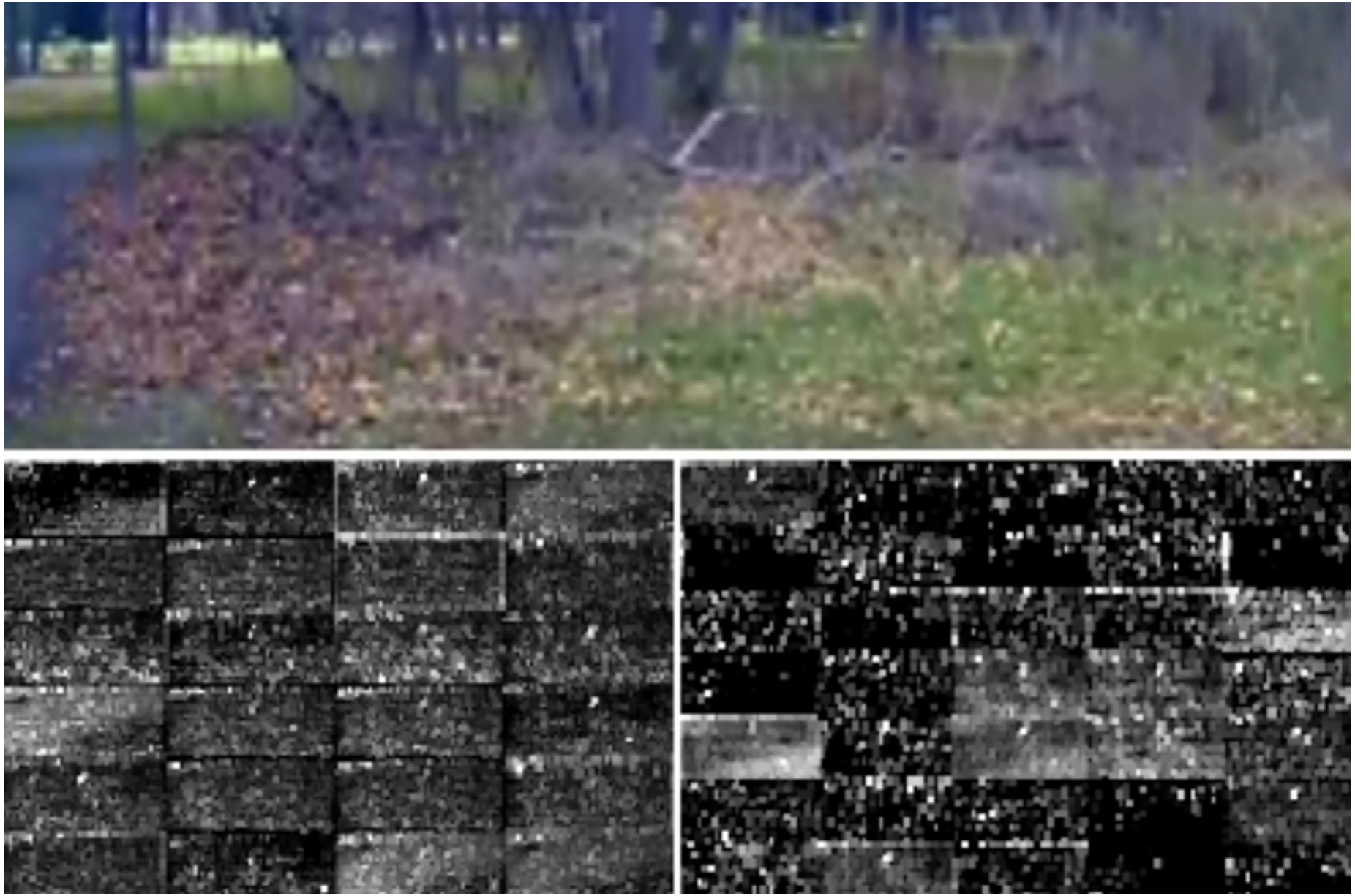
Hobbyist / Maker

Background

- End to End Learning for Self-Driving Cars by Bojarski et al.
- Implemented an unsupervised CNN model for controlling the steering angle of a vehicle
- "The CNN is able to learn meaningful road features from a very sparse training signal (steering alone)."



Unpaved Road



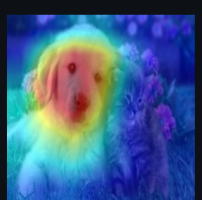
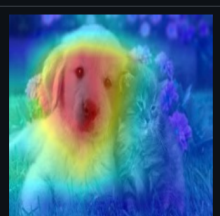
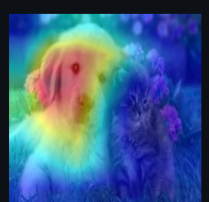
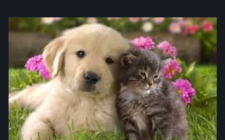
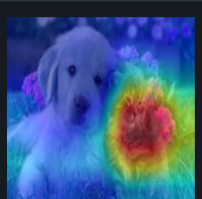
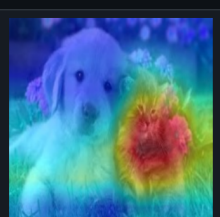
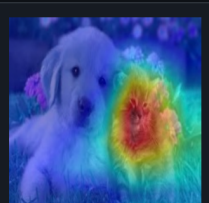
Forrest Scene

What we did?

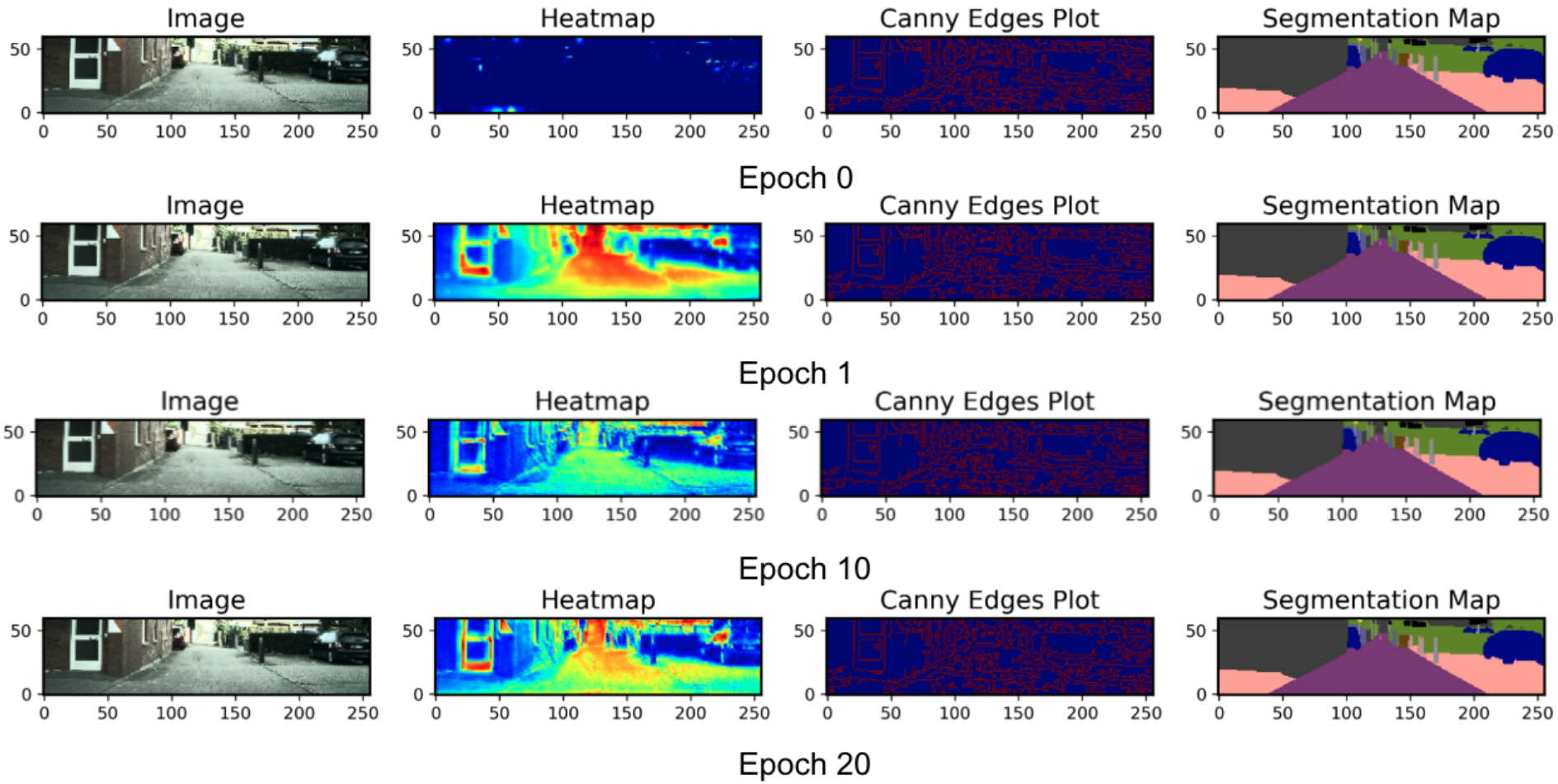
GradCAM

+

Segmentation Maps

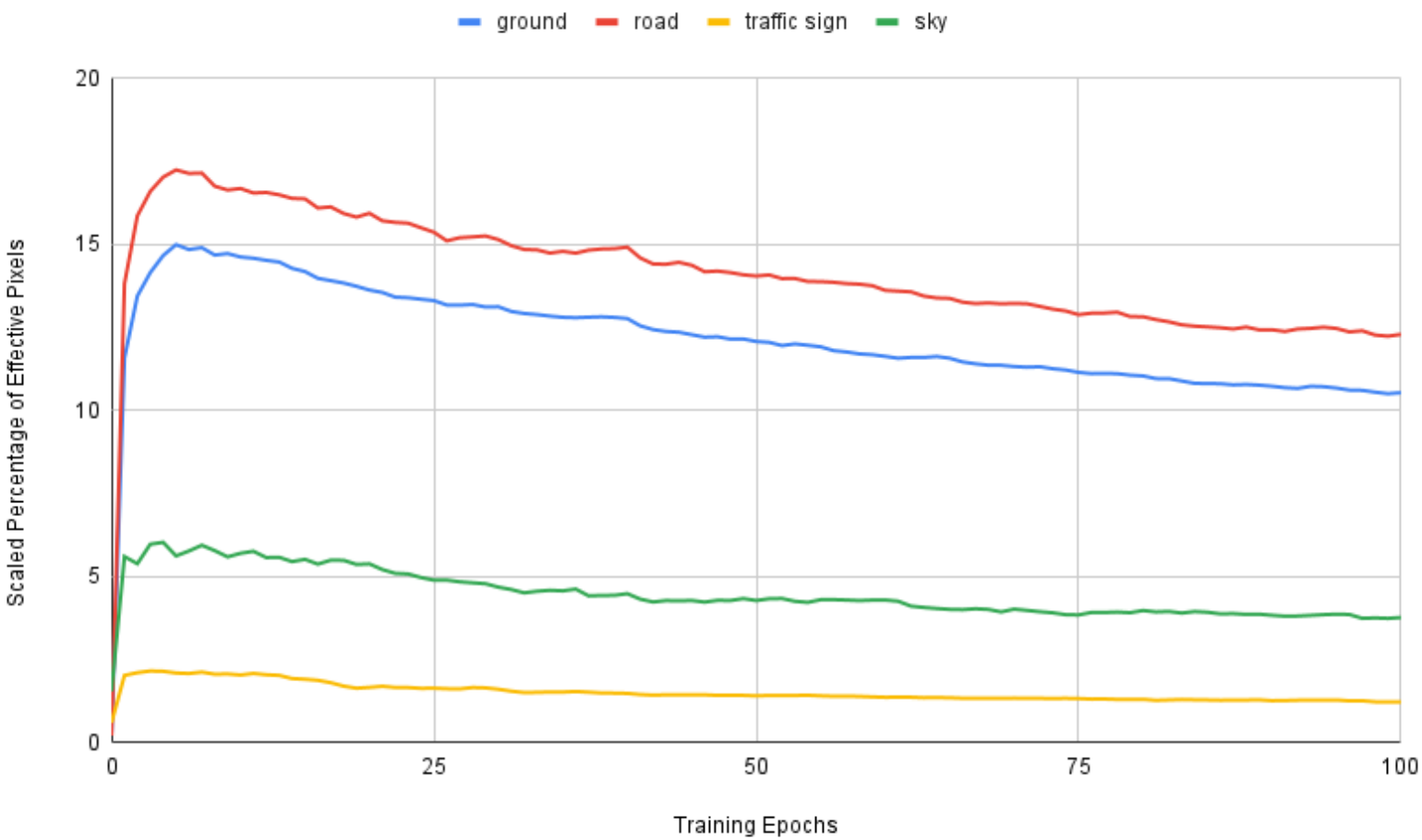
Category	Image	GradCAM	AblationCAM	ScoreCAM
Dog				
Cat				

https://github.com/jacobgil/pytorch-grad-cam



How the model's heatmaps change per epoch (Cityscapes Dataset Sample)

Percentage of activation of pixels per class calculated of the total occurrence of that pixel class for a simple model



Sanity Checks for Saliency Maps (Adebayo et al.)

- Many methods partially reconstruct the input data
- Brittle to noise and interference (misleading results)
- Many of the advanced guided methods dont have an adequate relationship between the input data and output nodes of a network
- Some methods (like some saliency maps) may not work with features that have a negative effect on the output

What we did

- 6 Models
- 2 DataSets
- Ran through each epoch up to 150
- Ran perturbation version of the dataset by using the segmentation maps
- Model comparison (Still in-progress)

Possible Issues

- Perfromance
- Needing semantic data
- Accuracy and granularity of the semantic data

- https://twitter.com/trex2218
- https://github.com/TRex22
- https://www.linkedin.com/in/jasonchalom/