Master Thesis Proposal: Explainable Multi-Sensor 3D Object Detection with Transformers

For Wassim Zahr

January 4, 2023

1 Topic and Goals

3D Object Detection refers to the task of detecting and localizing objects in a three-dimensional environment, typically using data from multiple sensors such as cameras, LiDARs, and radar [5, 6, 4]. This is a challenging problem that has many applications, including robotics, autonomous vehicles, and augmented reality.

One approach to 3D object detection that has gained popularity in recent years is the use of transformers, which are a type of deep learning model that is particularly well suited to processing sequential data, such as the time-series data that is often generated by sensors. By using a transformer to combine the data from multiple sensors, it is possible to capture complex relationships and patterns that may not be apparent when each sensor is considered individually. However, the decision-making and reasoning of a DNN is often not understandable for humans.

A thesis on this topic is the conception and the development of an explainable multi-sensor 3D Object Detection method, which aims to provide insights into the reasoning behind the model's predictions and improve its interpretability. This could involve the use of techniques such as attention mechanisms or visualization tools to understand how the transformer processes the data from multiple sensors to make its predictions.

The discipline to understand the reasoning behind DNNs is usually denoted as *Explainable AI* (XAI). This thesis should for example help to answer the following questions regarding a multi-sensor 3D Object detector based on a Transformer:

- What is happening inside the Transformer on the current input sequence ?
- What did the Transformer learn ?
- What did the Transformer see in the multi-sensor data stream ?
- Where is the Transformer "looking" at ?
- Which sensor(s) is(are) responsible for the current detection ?

The basis of this thesis is an existing multi-sensor 3D Object Detection approach [4] which was developed for a public available dataset for automated driving [2]. This Transformer based framework has to be extended in order to incorporate explainability features such as

- Raw Attention Maps
- Attention Rollout with NMS and Max Fusion [1]
- Gradient Attention Rollout [3]
- Activation Maximization [1].

Based on this explainability features, further visualizations tools should be implemented which help *the developer*, *the user* or *authorities* to understand the current state and reasoning of the detector.

[1] https://jacobgil.github.io/deeplearning/vision-transformer-explainability

2 Working Points

- Literature research on Explainable AI with Transformers
- Literature research on Object Detection with Transformers
- Extension of a multi-modal 3D Object Detector with explainability features
- Implementation of visualization tools which help *the developer*, *the user* or *authorities* to understand the reasoning behind the Transformer
- Training of the Transformer on public available datasets on our compute cluster
- Qualitative evaluation of the explainability features and visualization tools

3 Organization

The thesis shall be written at the Institute for Automotive Engineering RWTH Aachen (ika) at the department "Vehicle Intelligence and Automated Driving".

- External Assistant: Till Beemelmanns M.Sc.
- External Professor: Prof. Lutz Eckstein

References

[1] Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. *CoRR*, abs/2005.00928, 2020.

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[3] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. *CoRR*, abs/2012.09838, 2020.

[4] Simon Doll, Richard Schulz, Lukas Schneider, Viviane Benzin, Enzweiler Markus, and Hendrik P.A. Lensch. Spatialdetr: Robust scalable transformer-based 3d object detection from multi-view camera images with global cross-sensor attention. In *European Conference on Computer Vision(ECCV)*, 2022.

[5] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *CoRR*, abs/1812.05784, 2018.

[6] Yue Wang, Alireza Fathi, Abhijit Kundu, David A. Ross, Caroline Pantofaru, Thomas A. Funkhouser, and Justin Solomon. Pillar-based object detection for autonomous driving. *CoRR*, abs/2007.10323, 2020.