RWTH Aachen University

Fakultät für Maschinenwesen

Institut für Kraftfahrzeuge

Univ.-Prof. Dr.-Ing. Lutz Eckstein

**Masterarbeit**

# Schreiben einer Abschlussarbeit in LaTeX

Diese Arbeit wurde vorgelegt am Institut für Kraftfahrzeuge
von:

Herrn B.Sc. Automatisiertes Fahren, Matr.-Nr.: 123456

betreut von:

Prof. Dr.-Ing. habil. Fahrzeug Intelligenz M.Sc.

Erstprüfer:

Univ.-Prof. Dr.-Ing. Lutz Eckstein

Zweitprüfer:

Dr.-Ing. Adrian Zlocki

Aachen, November 2019

## Contents

# 1 Introduction

This chapter will first introduce Explainable AI and why it is very important in autonomous driving. Then, multi-modal 3D object detection is described. Finally, an overview of the transformer architecture is described.

## 1.1 Explainable AI

Artificial intelligence is now an indispensable tool which enhances our life quality. Due to the increasing complexity of state-of-the-art deep learning models, the AI models are often viewed as a "black box" where only the inputs and the predictions are visualized. This leads to the problem of "trusting" the AI without understanding how it works, as can be seen in Figure 1-1. Thus, it is becoming more and more important to understand why and how a model reached a certain decision. This is expecially important when the safety of people depends by the AI, such as in healthcare and automotive driving [ABE22]. The techniques used to explain the predictions made by the AI and make them more interpretable are referred to as Explainable AI (XAI). The objective of XAI is to eliminate the "black-box" models by explaining its behavior. It not only enhances security and trustiness for the end-user, but it also helps the developer to improve the model, for example by removing potential biases. In fact, consider an example in which an object detection model has very good performance metrics. If the dataset contained objects with a specific background, e.g a golf ball with grass as background, then it is highly possible that the model is biased. In that case, it could mistakenly detect a golf ball each time there is grass as background. Through XAI, the developer could easily identify this bias and correct it. The more explainable a model is, the easier it becomes to improve it. XAI could solve security problems: adversarial attacks can mislead an object detector to confuse one image with another just by changing some pixels, leading to unintended behaviors. This is particular critical in automoted driving. Those important pixels attributed by the model could be identified by XAI. In case of accidents, XAI could help identify the cause, thus improving accountability.

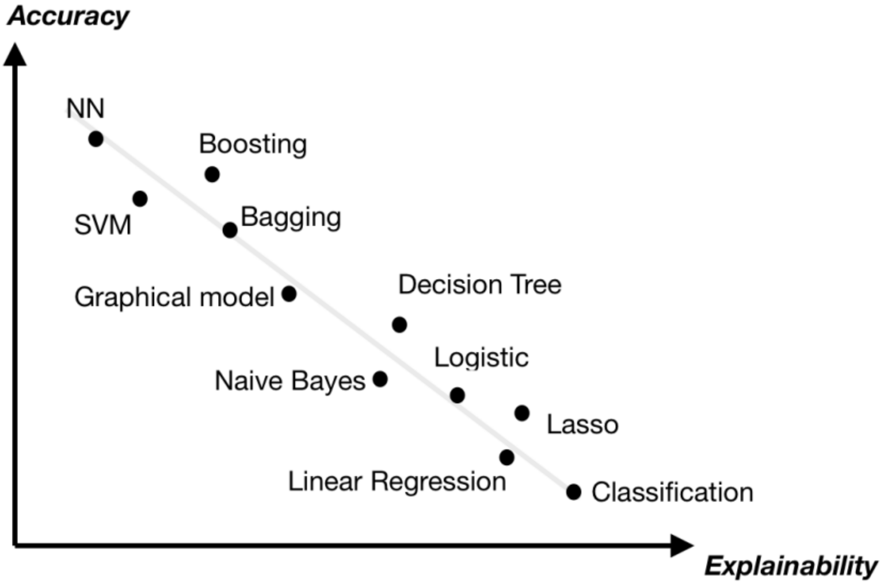

Fig. 1-1: Accuracy vs Explainability of the main machine learning algorithms [**duval2019explainable**]

### 1.2 Multi-Modal 3D Object Detection

Autonomous driving is progressing towards self-driving cars thanks to the advanced machine learning models. One of the main difficulties is due to the complex and dynamic driving environment, which require a sofisticated artificial vision [WAN21]. Object detection aims to identify the objects in the scene, their location and category [WAN21]. Perception with a single modality suffers some drawbacks, such as difficulties on detecting occluded and large objects [HUA22][CAE20]. Equipping the car with more sensors (multi-modal) improves accuracy and robustness. LiDAR (Light Detection and Ranging) sensor contains accurate localization in 3D, while cameras allow measurements of color and edges [CAE20]. It is becoming common to fuse data from LiDAR and cameras for better performance instead of using them separately, but this approach requires precise synchronization. The success of 2D object detection has been unprecedented thanks to the development of deep learning-based models like RCNN, Fast RCNN, and Faster R-CNN. However, 2D detection can only provide limited information such as the 2D bounding box of an object, which is not enough for autonomous vehicles to perceive their environment. Thus, 3D object detection has become a more challenging but crucial task as it helps with more accurate spatial path planning and navigation. In this task, more output parameters are required to specify the 3D-oriented bounding boxes around objects [WAN21].

### 1.3 Transformers

Transformers have become the standard machine learning model in natural language processing (NLP) [DOS20]. They were introduced in 2017 by a team at Google Brain [VAS17] as fully attention-based model. The attention mechanism was invented in 2014 to address some problems arising in Recurrent Neural Networks (RNN) in machine translation [BAH14]. For instance, RNNs suffer with long text and slow training. The attention mechanism was then implemented by RNN-based encoder-decoder architectures to solve those problems. However, these models still rely on recurrence, which is incompatible with parallel computation. Transformers get rid of recurrence and convolutions and are based only on attention mechanism [VAS17]. Thanks to their efficiency, Transformer models are trained with unprecedented size [DOS20]. Attentions perform global computation which made them applicable in NLP and, recently, in computer vision. To appreciate how Trasformers revolutionized the AI world, an understanding of the attention mechanism is needed.

### 1.3.1 Attention

Suppose we have the sentence *"Mark eats an apple while he is going to the gym"*. It is composed of 11 words or tokens. Considering the word *he*, it more related to the word *Mark* instead of *apple*, even if the latter is closer to the word *Mark*. Context, therefore, is more important than proximity. The idea of *self-attention*, in this case, is to figure out how important all the other words in the sentence are with respect to a specific word, without any recurrent connections. It is based on just weighted sums and activations, so they can be very parallelizable and efficient. If this sentence needs to be translated, to German for example, then the output text can interact with the input through *cross-attention*. In the case of object detection, tokens can represent feature patches of an image. For instance, an image is divided into a sequence of patches [DOS20][CAR20]. This is one big advantage of Transformers, i.e they can handle different data types (text, image and sound). To allow the attention mechanism to learn patterns and improve context, trainable parameters are needed: *Query*, *Key* and *Values* are the main components. Each token is projected to its own query,key and value by three different weight matrices, which are learnt during training. Those are used to calculate the *attention scores*:

$$Attention(\mathit{Q,K,V}) = softmax(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^T}{\sqrt{d_k}}) \cdot V \qquad \text{Eq. 1-1}$$

### 1.3.2 Positional encoding

...

### 1.3.3 Multi-head attention

...

Figure 1-2 shows the Transformer model as depicted in the original paper [VAS17]. It is composed of an encoder and a decoder, as depicted in the left and right side, respectively. The encoder and the decoder are composed of a stack of *N* layers. Some Transformer architectures are encoder-only or decoder-only. The SpatialDETR, for instance, is a decoder-only architecture [DOL22].

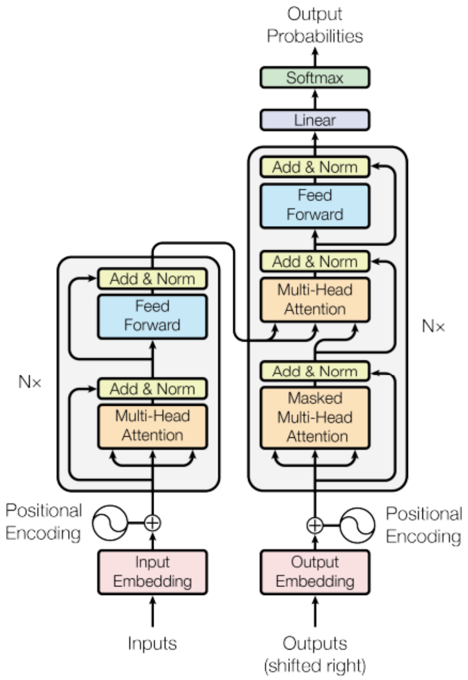

Fig. 1-2: The Transformer model architecture. [VAS17]

# 2 Architecture

SpatialDETR is the framework chosen for incorporating explainability features, which are visualized using the Application described in Chapter [**ch:app**]. It is a state-of-the-art 3D Object Detection model based on the Transformer model. Therefore, it is important to understand its architecture and how we can explain its reasoning. SpatialDETR is actually an extension to the DETR3D architecture, which in turn is based on the popular DETR (DEtection TRansformer). In this chapter, DETR is first introduced, which forms the basis of many transformer-based object detection models. Then, DETR3D architecture is described. Finally, SpatialDETR is extensivly described by also highlighting the differences with DETR3D.

## 2.1 DEtection TRansformer - DETR

Object detection involves the identification of one or more objects in an image by drawing bounding boxes around them and assigning their labels. It is a more complex task than image classification, which predicts the class of only one object in an image. DETR is a 2D Object Detector based on an encoder-decoder transformer model, developed by the Facebook AI team [CAR20]. It is simpler and more accurate than the well-established object detectors such as Faster R-CNN [REN15]. The state-of-the-art 2D object detectors are typically two-stage detectors [GIR15] [CAR20]: they first extract "candidate" regions of object and then they extract features from those regions (by using fully convolutional neural networks) which are finally classified by a fully connected layer. Furthemore, post-processing steps such as NMS (non-maximal suppression) are needed for avoiding near-duplicates

## 2.2 DETR3D

...

## 2.3 SpatialDETR

...

## 3 Explainable Transformer

Explainability has a well established literature for natural language processing, with techniques such as SHAP and LIME. However, in computer vision there are not standard techniques for explainability. This is even more lacking in for the Transformer architecture. However, only in the last years there have been some work toward Explainable Transformer, applied to DeiT and DETR models [TOU21] [CHE21b] [CHE21a]. Unfortunately, as far as I know there are yet any work toward explaining transformer-based 3D Object Detection. It is however possible to implement those techniques for SpatialDETR, and even to other 3D object detectors. In this chapter, some techniques used for Explainable Transformers are described, such as Attention Rollout and Gradient Rollout. Then, I will discuss how to implement those to the SpatialDETR architecture.

## 4 Research Questions

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 ?

## 5 Application

For addressing the objectives of the thesis, a software application has been developed which helps us to understand the reasoning of the SpatialDETR model. Trying to make it adaptable, a model selection is allowed. Therefore, it is possible to use it for various SpatialDETR configuration, e.g with only query center projection. At the moment, the application is addressed to developers who want to see how the 3D Detection works behind the scene. It is possible to select the attention visualization mechanism (e.g head fusion, attention rollout, gradient rollout), the attention discard ratio, the layer (6 in SpatialDETR), the camera (6 in NuScenes dataset). Furthemore, it is possible to select the prediction threshold for detection and to visualize the ground trouth bounding boxes. All this flexibility allows for a better understanding of the Transformer model and will help to decide, for example, which attention visualization mechanism better explains the model. Therefore this application can be adapted to normal users and authorities by selecting the best configuration for XAI.

# 6 Conclusion & Outlook

## 7 List of Symbols

| | |
|---|---|
| $a_x$ | longitudinal acceleration |
| $a_y$ | lateral acceleration |
| $\delta$ | steering wheel angle |
| $\kappa$ | curvature |
| $\kappa'$ | derivative of the curvature |
| $v$ | velocity |
| $v_x$ | longitudinal velocity |
| $v_y$ | lateral velocity |

## 8 List of Abbreviations

| | |
|---|---|
| ADAS | advanced driver assistance system |
| CC | cruise control |
| GPS | global positioning system |
| LiDAR | light detection and ranging sensor |
| PROMETHEUS | Programme for a European traffic with highest efficiency and unprecedented safety |
| THW | time headway |
| TTC | time-to-collision |

## 9 Bibliography

[ABE22] ABELOOS, B., HERBIN, S.
Explaining object detectors: the case of transformer architectures
Workshop on Trustworthy Artificial Intelligence as a part of the ECML/PKDD 22 program, 2022

[BAH14] BAHDANAU, D., CHO, K., BENGIO, Y.
Neural machine translation by jointly learning to align and translate
arXiv preprint arXiv:1409.0473 (2014)

[CAE20] CAESAR, H., BANKITI, V., LANG, A. H., VORA, S., LIONG, V. E., XU, Q., KRISHNAN, A., PAN, Y., BALDAN, G., BEIJBOM, O.
nuscenes: A multimodal dataset for autonomous driving
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11621–11631

[CAR20] CARION, N., MASSA, F., SYNNAEVE, G., USUNIER, N., KIRILLOV, A., ZAGORUYKO, S.
End-to-end object detection with transformers
Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer, 2020, pp. 213–229

[CHE21a] CHEFER, H., GUR, S., WOLF, L.
Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers
Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 397–406

[CHE21b] CHEFER, H., GUR, S., WOLF, L.
Transformer interpretability beyond attention visualization
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 782–791

[DOL22] DOLL, S., SCHULZ, R., SCHNEIDER, L., BENZIN, V., ENZWEILER, M., LENSCH, H. P.
SpatialDETR: Robust Scalable Transformer-Based 3D Object Detection From Multiview Camera Images With Global Cross-Sensor Attention
Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX, Springer, 2022, pp. 230–245

[DOS20] DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G., GELLY, S., et al.
An image is worth 16x16 words: Transformers for image recognition at scale
arXiv preprint arXiv:2010.11929 (2020)

[GIR15] GIRSHICK, R.
Fast r-cnn
Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448

[HUA22] HUANG, K., SHI, B., LI, X., LI, X., HUANG, S., LI, Y.
Multi-modal sensor fusion for auto driving perception: A survey
arXiv preprint arXiv:2202.02703 (2022)

[REN15] REN, S., HE, K., GIRSHICK, R., SUN, J.
Faster r-cnn: Towards real-time object detection with region proposal networks
Advances in neural information processing systems 28 (2015)

[TOU21] TOUVRON, H., CORD, M., DOUZE, M., MASSA, F., SABLAYROLLES, A., JÉGOU, H.
Training data-efficient image transformers & distillation through attention
International conference on machine learning, PMLR, 2021, pp. 10347–10357

[VAS17] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., POLOSUKHIN, I.
Attention is all you need
Advances in neural information processing systems 30 (2017)

[WAN21] WANG, Y., MAO, Q., ZHU, H., DENG, J., ZHANG, Y., JI, J., LI, H., ZHANG, Y.
Multi-modal 3d object detection in autonomous driving: a survey
arXiv preprint arXiv:2106.12735 (2021)

# 10 Appendix