CENTER FOR CONNECTED AND
AUTOMATED TRANSPORTATION

Final Report #74
November 2023

Promoting CAV Deployment by Enhancing the Perception Phase of the Autonomous Driving Using Explainable AI

Jiqian Dong
Sikai Chen
Samuel Labi

PURDUE
UNIVERSITY

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Report No. 74
Project Start Date: January 2022
Project End Date: May 2023

November 2023

Promoting CAV Deployment by Enhancing the Perception Phase of Autonomous Driving Using Explainable AI

Jiqian Dong
Graduate Researcher

Sikai Chen
Visiting Assistant Professor

Samuel Labi
Professor

Purdue University

ACKNOWLEDGEMENTS AND DISCLAIMER

Funding for this research was provided by the Center for Connected and Automated Transportation under Grant No. 69A3551747105 of the U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology (OST-R), University Transportation Centers Program. The researchers are grateful to the cost sharing and other collaborators including the Indiana DOT (Drs. Barry Partridge and Samy Noureldin), Indian Institute of Technology, Tirupati (Dr. Krishna Prapoorna), King Khalid University (Dr. Saeed Alqadhi), Hong Kong Polytechnic University (Dr. Anthony Chen), National University of Singapore (Dr. Ghim Ping Ong), and Nanyang Technological University (Dr. Feng Zhu). We also acknowledge the contributions of Dr. Mohammad Miralinaghi (Illinois Institute of Technology), Dr. Tiantian Chen (Korean Advanced Institute of Science & Technology), and Dr. Pei Li of the University of Wisconsin-Madison. We are also grateful for the institutional support provided by Purdue University's College of Engineering's Center for Innovation in Control, Optimization, and Networks (ICON) and the Autonomous and Connected Systems (ACS) initiatives, and Purdue's Black & Gold autonomous racing team. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Suggested APA Format Citation:

Dong, J., Chen, S., Labi, S. (2023). Promoting CAV Deployment by Enhancing the Perception Phase of Autonomous Driving Using Explainable AI, CCAT Report #74, The Center for Connected and Automated Transportation, Purdue University, West Lafayette, IN.

Contacts

For more information:

Samuel Labi, Ph.D.
550 Stadium Mall Drive
HAMP G167B
Phone: (765) 494-5926
Email: labi@purdue.edu

CCAT
University of Michigan Transportation
Research Institute
2901 Baxter Road
Ann Arbor, MI 48152
uumtri-ccat@umich.edu
(734) 763-2498
www.ccat.umtri.umich.edu

Technical Report Documentation Page

1. Report No. 74	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Promoting CAV deployment by enhancing the perception phase of autonomous driving using explainable AI		**5. Report Date** November 2023 **6. Performing Organization Code** N/A
7. Author(s) Jiqian Dong, Sikai Chen, Samuel Labi		**8. Performing Organization Report No.** N/A
9. Performing Organization Name and Address Center for Connected and Automated Transportation, Purdue University, 550 Stadium Mall Drive, W. Lafayette, IN 47907; and Univ. of Michigan Ann Arbor, 2901 Baxter Rd, Ann Arbor, MI 48109		**10. Work Unit No.** **11. Contract or Grant No.** Contract No. 69A3551747105
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE, Washington, DC 20590		**13. Type of Report and Period Covered:** Final rep., Jan 2022-May 2023 **14. Sponsoring Agency Code:** OST-R
15. Supplementary Notes Conducted under the U.S. DOT Office of the Assistant Secretary for Research and Technology's (OST-R) University Transportation Centers (UTC) program.		

16. Abstract

User trust is pivotal to autonomous vehicle (AV) operations which are driven by artificial intelligence (AI). A promising way to build user trust is to use explainable artificial intelligence (XAI) which requires the AI system to provide the user with the underlying explanations for its decisions. Motivated by the need to enhance user trust and the promise of novel XAI technology in this context, this study strives to enhance trustworthiness in autonomous driving systems through the development of explainable Deep Learning (DL) models. The study casts the AV decision-making process not as a classification task (which is the traditional process) but rather as an image-based language generation (image captioning) task. As such, the proposed approach makes driving decisions by first generating textual descriptions of the driving scenarios which serve as explanations that humans can understand. The first part of the research project developed a novel multi-modal DL architecture to jointly model the correlation between an image (driving scenario) and language (descriptions). It adopts a fully Transformer-based structure and therefore has the potential to perform global attention and imitate effectively, the learning processes of human drivers. The results suggest that the proposed model can and does generate legal and meaningful sentences to describe a given driving scenario, and subsequently to correctly generate appropriate AV driving decisions. The model significantly outperforms multiple baseline models in terms of generating explanations and driving actions. The second part of the research developed a framework for jointly predicting potential driving actions with corresponding explanations, thereby producing explainable DL models useful for trustable autonomous driving. The explainable DL models can not only boost user trust in autonomy but also serve as a diagnostic approach to identify any model deficiencies or limitations during AV system development. From the end user's perspective, the proposed models can be beneficial in enhancing user trust because they provide the rationale behind an AV's decisions/actions. From the AV developer's perspective, the explanations from the explainable system could serve as a "debugging" tool to detect potential weaknesses in the existing system and to identify specific lacunas that need to be addressed.

17. Key Words Autonomous vehicles, Explainable Artificial Intelligence (XAI), Image captioning, Transformer, User trust.		**18. Distribution Statement** No restrictions.	
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 78	**22. Price** N/A

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Part 1

Part 2

LIST OF ACRONYMS

ADAS –	Advanced Driving Assistance Systems
BDD-OIA –	BDD Object Induced Actions
C2C –	Sequence-To-Sequence
CNN –	Convolutional Neural Network
CV –	Computer Vision
DL –	Deep Learning
DLCV –	Deep Learning Computer Vision
FPN –	Feature Pyramid Networks
FRN –	FasterRCNN
GAT –	Graph Attention Network
GCN –	Graph Convolutional Neural Network
GCQ –	Graphic Convolutional Q Learning
GNA –	Global No-Attention
GNN –	Graph Neural Network
GSA –	Global Soft Attention
HDV –	Human-Driven Vehicles
HV –	Human Vision
ITS –	Intelligent Transportation Systems
MSA –	Multi-head Self Attention
NLP –	Natural Language Processing
LODD –	Level of Driving Decision
LSTM –	Long- Short-Term Memory
RHA –	Regional Hard-Attention
RNN –	Recurrent Neural Network
RPN –	Region Proposal Network
RSA –	Regional Soft-Attention
SA –	Self Attention
SOTA –	State of the Art
SW-MSA –	Shifted Window MSA
V2V –	Communication between vehicles and other vehicles
ViT –	Vision Transformer
W-MSA –	Window MSA
XAI –	Explainable Artificial Intelligence

PART I

An Explainable Artificial Intelligence (XAI) Framework for Autonomous Driving Systems

CHAPTER 1. INTRODUCTION

1.1 Background

Over the last century, the automobile has had significant technological advancement as it evolved from fully manual operations to increasing levels of automation. In the current era, vehicles are equipped with advanced automation technologies and therefore are capable of self-driving with little or no input from the human operator. It has been shown in the literature that the success of automated driving, and customer acceptance and patronage of autonomous vehicles hinge on user trust in such automated processes (Hewitt et al., 2019; Hulse et al., 2018; Omeiza et al., 2022; Rezaei and Caulfield, 2020). Therefore, it seems imperative that the development of automated systems must be accompanied by conscious efforts to build the trust and confidence of the prospective users of automated vehicles. In efforts to achieve this goal, it is useful (or even indispensable) that the decisions made by the automated processes be accompanied by explanations for such decisions. That way, the autonomous vehicle (AV) end users and developers can be assured of the rationale behind the AV's decisions and, if needed, investigate such rationale (Koo et al., 2015). Such capability could help not only enhance the transparency and accountability of the AVs' decisions but also to evaluate the AV's role in the critical event *ex ante* (before a critical event such as a collision or near miss) or *ex poste* (after the critical event).

On the other hand, of the various enabling technologies of AVs, deep learning (DL) based artificial intelligence (AI) models continue to play multiple crucial roles, particularly in the decision processor for AV or Connected AV (CAVs) (Chen et al., 2021; Di and Shi, 2021; Dong et al., 2021a, 2020; Du et al., 2022; Li et al., 2023, 2022; Shi et al., 2021). In other smart transportation-related applications, DL models are also used in driver behavior modeling (Xing et al., 2021), vehicle routing (Du et al., 2021; Zhang et al., 2020), traffic prediction (Do et al., 2019; Liu et al., 2019; B. Yu et al., 2020; Zhou et al., 2021), and infrastructure monitoring (Hou et al., 2020; Zhuang et al., 2018). These DL models have merit in terms of their high representation and generalization capabilities. However, their intrinsic drawback, which is well known, is the black-box nature of their computation that leads to a notoriously inexplainable system. Due to such limitations in interpreting the decisions made by DL-based AI systems, it is difficult to justify the underlying rationale of such decisions. Moreover, any failure of such DL models is often not only unpredictable but also undiagnosable. These limitations further exacerbate user distrust of automation.

1.2 Explainable AI (XAI) in Goal Induced Systems

Most of the aforementioned applications are not safety critical and may not cause catastrophic consequences in the event of system failure. However, when using DL to generate safety-critical decisions, such as those associated with AV driving tasks in real traffic environments, reliability and robustness are particularly important from a safety perspective. To achieve the required level of reliability, the first step is to understand the rationale used by the AI to make the decision. This requires the AI to be upgraded into explainable AI (XAI) and thereby to be capable of generating human-understandable explanations as outputs (Doran et al., 2018). Following the taxonomy in recently published literature (Zablocki et al., 2022), explainability represents the combination of interpretability (whether the explanations are comprehensible by humans) and completeness (whether exhaustiveness of the explanation is achieved). Regarding the task of autonomous driving, Atakishiyev provides a comprehensive review of these concepts in the several ways they have been used in recent literature (Atakishiyev et al., 2021). From the regulation perspective, XAI has been institutionalized through published standards such as the European Union's General Data Protection Regulation (GDPR) (Voigt and von dem Bussche, 2017) which specifically stipulated that "right to explanation" is required in the context of decisions made by complex systems, particularly where they are founded on black-box models.

In spite of such evidence of the growing regulations that mandate the explainability of AI systems in practice and the growing appreciation and application of explainability in the research world, most existing XAI literature focuses on developing explanations for only a single neuron or a single network model (Mittelstadt et al., 2019), and very few researchers have explored in the context of complicated "goal-induced" systems such as AV (Omeiza et al., 2021). Regarding autonomous driving, the primary task is inherently a "goal-induced" process that requires the cooperation of multiple submodules (perception, localization, prediction, motion planning, and control) for generating efficient driving decisions. Therefore, the success of one individual submodule (or each model) does not necessarily translate into the functionality of the entire system. As a result, classic XAI methods that provide only component-specific explainable modules are insufficient to guarantee the transparency of the overall system. Additionally, such pipelined systems heavily rely on human heuristics and manually selected representations which could lead to suboptimal-driving decisions (Zablocki et al., 2022).

Consider, for example, most existing perception algorithms present in AV driving decision systems. These are trained with offline datasets with only perception-related labels such as detections (in the form of a bounding box, a class label, and a confidence level of each object) instead of the driving decision that is made. As a result, even though the detection results from the perception module are interpretable (in other words, the human can visualize the object detection results including the bounding boxes and class labels), the following questions remain: how do the planning and control modules (subsequent to the perception module) utilize the detection results to generate the driving decision? Are the manually selected detections sufficient for making optimal driving decisions? The existence of these questions, coupled with their profound importance, raises serious concerns about the explainability of pipelined autonomous driving systems. In offering a way to bridge this extant lacuna in contemporary AV literature and to enhance the overall interpretability of autonomous driving algorithms, this study focuses on the "goal-induced" nature of driving tasks and thereby develops an explainable end-to-end vision-based autonomous driving system. In line with several pioneering research related to this objective (Ben-Younes et al., 2022; Kim et al., 2019, 2018; Li et al., 2020), our proposed framework involves mapping driving scenario images to decisions and explanations. In further enhancing existing research in this area, the present study's framework incorporates the most recent state-of-the-art (SOTA) advancements in visual attention mechanisms. This improvement facilitates the leveraging of the capabilities of attention mechanisms in visual processing tasks, thereby improving the accuracy and effectiveness of models that generate decisions and explanations from driving scenario images.

1.3 Comparison of Computer Vision (CV) and Human Vision (HV)

In the field of perception and semantic understanding for AV operations purposes, DL-based computer vision (CV) (also referred to as "machine vision) has been used widely in practice (Bojarski et al., 2016; Chen et al., 2019) and several state-of-the-art (SOTA) models have been developed in this regard. Despite the accomplishments of these efforts, it is acknowledged that human vision (HV) remains as the incontrovertible paragon of all vision systems, and therefore, represents the ultimate benchmark for assessing the efficacy of any artificial vision system. The most salient advantage of human vision over all CV models is that it has the unique capability to be "goal-induced." In other words, the human eye can be adjusted to attend to and only investigate "the region of interest" that is correlated with the incumbent intention of the human.

For example, a human driver who intends to make a lane change may keep his/her focus in front of their vehicle but glances into the mirror to monitor the environment at the vehicle's rear area before making the lane change. Such "goal-induced" vision has the advantage of high efficiency, in other words, the human brain only needs to process limited but sufficient information with minimum ambient visual noise or irrelevant information. However, existing CV models often do not have this capability because their training goal is to conduct exhaustive object detection over the entire camera scan. As such, these models tend to "waste" significant computation resources in processing large volumes of redundant and irrelevant information such as, for example, images of persons in roadside advertisement billboards that are irrelevant to the driving operation.

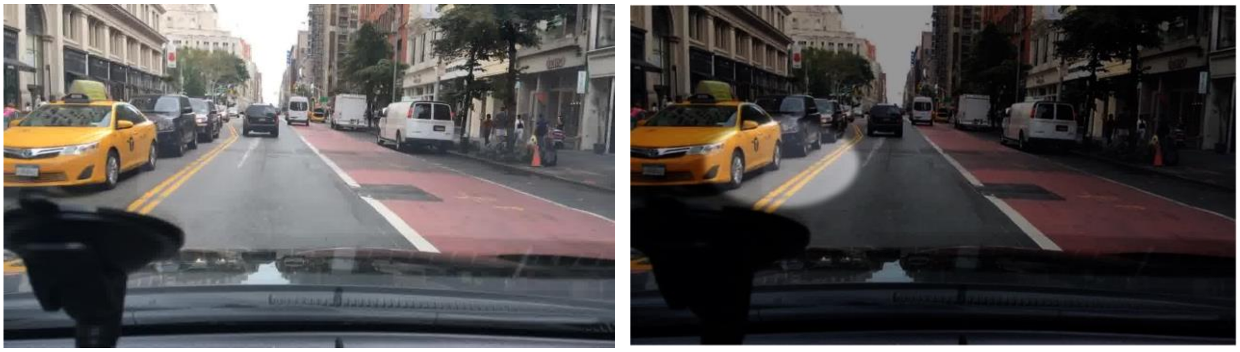
Secondly, existing CV models lack the multi-resolution structure of human vision (HV) – peripheral and foveal vision (Xia et al., 2020). Peripheral vision (in other words, "glimpse") is blurred (low resolution) but only requires a brief time for processing and has a very wide field of view. Foveal vision (in other words, "gaze"), on the other hand, is clear (high resolution) but requires longer processing time and only has a limited visual field. The human eyes utilize these two capabilities in a hierarchical manner and balance them with a proper attention mechanism. In other words, when the human perceives a scene, peripheral vision is first adopted to grasp the overall semantics of the scene. Then, based on an overall understanding from peripheral vision and intentions, the attention mechanism extracts the important regions that the brains deem to be worthy of further scrutiny. After that, the foveal vision is used on those attended regions to extract detailed information (shape, color, texture, etc.) of the perceived objects. Such an attention phase possesses the capability of filtering out irrelevant regions and saving computational power for the relevant ones.

Inspired by this natural process, several recent research efforts have sought to leverage DL models to predict the human drivers' attention (gazing) regions (Pal et al., 2020; Palazzi et al., 2019; Xia et al., 2019). These models have produced interesting results from a bionics perspective. However, they are far from goal-induced and are not capable of deployment in AV systems unless further instructions are provided on how to map the attended regions to driving decisions. In recognition of such limitation, this report develops an end-to-end autonomous driving model which generates useful features while leveraging proper attention mechanisms that imitate the capability of human vision.

1.4 Attention Mechanism and Goal Shift in the Driving Decision Processor

"Attention mechanism" in DL models refers to the concept of a neural network's capability to automatically learn the relative importance of the features and then fuse these features based on the learned importance weights. This notion was initially propounded by Bahdanau and his collaborators (Bahdanau et al., 2015) and subsequently evolved to become the foundation for the revolutionary Transformer-based models (Dosovitskiy et al., 2020; Vaswani et al., 2017). The benefit of the attention mechanism is that it is capable of not only boosting the model performance by "enlarging" the useful features while "suppressing" noise or redundant information but also providing visualizable "attention maps" to help enhance model interpretability. These attention maps can reveal the latent computation logic of the DL models by showing how much they focus on different inputs and outputs (Dong et al., 2022; Du et al., 2021; Ghaeini et al., 2020; Kim et al., 2018; Kotseruba and Tsotsos, 2022; Li et al., 2020; Z. Lin et al., 2017; Wang et al., 2016). As such, some researchers have claimed that attention maps can be used as a natural and efficient approach to building explainable DL models (Guidotti et al., 2018; Kim et al., 2018; Lei et al., 2016; Xu et al., 2015a). In a more recent study, Wiegreffe and Pinter (2019) re-evaluate the attention mechanism and argue that attention maps can be used only as explanations if "plausible" and "faithful" rationale can be reasoned from them (Wiegreffe and Pinter, 2019). In this context, "plausible" means that the generated attention maps need to be highly correlated with only the correct explanations, which requires that attention is not ambiguous (i.e., can provide the only correct explanation). Regarding the "faithfulness" requirement, attention is needed to be consistent (i.e., always attend to the same region with the same semantic meaning) across different scenarios. These two requirements align with XAI's interpretability and completeness requirements, as mentioned earlier in this report.

However, for traditional decision models (classification) in driving tasks, it is often difficult to guarantee the "plausible" criterion because a single driving decision may correlate with multiple reasons. This "multi-dependency" nature of driving decisions could be problematic during the development of attention-based models, particularly where it is needed to use attention as explanation. For example, consider an attention-based decision processor (Dong et al., 2022, 2021b) which can simultaneously generate driving decisions and output attention maps as depicted in **Figure 1**. In this scenario shown in **Figure 1 (a)**, the AV model outputs a decision that left turn is not feasible, with the attention map indicating its attention regions (**Figure 1 (b)**, the bright location indicates the focus region). Although both the predicted action and the attended regions make sense, it is not certain that the model has learned the correct causal relationship between the driving scene and the decision. Regarding the scenario depicted in **Figure 1 (a)**, there exist two reasons why the vehicle cannot turn left: (i) there exist obstacles to its left (multiple vehicles in the opposite direction), and (ii) there exists a double solid pavement marking line indicating left turn is prohibited. Therefore, although the model pays attention to the correct region to generate acceptable driving decisions, it is not certain which reason has been learned to trigger this decision. This violates the "plausible" requirement for using attention as explanation. Ideally, the model is expected to learn **both** explanations for generating a holistic understanding of the driving scenario.



(a) **(b)**

Figure 1. Raw scenario (a) and attention region (b). The brightened region is the attended region (model focus), and the model can correctly predict the driving decision: "cannot turn left."

To achieve this, we rethink the end-to-end driving decision generation procedure and conduct a novel shift in the learning goal from traditional "decision classification" (i.e., maps scenarios to driving decisions) to "driving scenario description" (i.e., maps scenarios to decision reasons). As such, the model is capable of exhaustively exploring all the potential explanations behind a given driving decision. From a technical perspective, the classic decision-generating process can be cast as image-captioning task by outputting the textual description (verbal rationale) of the observed driving situation (i.e., "obstacles in the left lane" + "solid line on the left" for **Figure 1**). Then, the driving decisions (i.e., "cannot make left turn / left lane changing") can be inferred by applying simple rules over the pre-generated descriptions.

1.5. Reason-induced Attention and Transformer

By nature, image captioning models need to capture the correlation between image and natural language. Therefore, under such specific settings, a superior attention mechanism is to use language to induce visual attention. In other words, the model can learn to "look at" specific regions that are "dictated" by the language instructions. As a result, the model with this attention mechanism can learn to "focus" by assigning higher weight (importance) to the language-induced region while filtering out less relevant noises.

In image captioning, this attention mechanism was initially introduced by Xu and his collaborators in their pioneering work (Xu et al., 2015) where they used a novel "CNN (encoder) + LSTM (decoder)" structure – this remained as the state-of-the-art (SOTA) for several years. In their model, they used Convolutional Neural Network (CNN) as the encoder to generate grid-based features of the raw image and applied a classic Recurrent Neural Network (RNN) structure, Long Short-term Memory (LSTM) to decode these image features and integrate contextual language features for generating verbal description. In the autonomous driving research domain, several interesting pieces of works have successfully applied LSTM to generate driving related explanation sentences (Kim et al., 2019; 2018). However, the performance of this model is limited by the training efficiency and expression ability of the LSTM model (Wang et al., 2022).

Some recent pieces of work have investigated the efficacy of replacing the language decoder from LSTM to Transformer (Vaswani et al., 2017, Herdade et al., 2019; Pan et al., 2020) and replacing image encoder from vanilla CNN to Faster R-CNN (Anderson et al., 2018; He et al., 2021). The authors of the current study duly recognize the great promise associated with replacing LSTM with Transformer, to enhance the model performance in the language generation task. Nevertheless, it is recognized that using Faster R-CNN (Anderson et al., 2018) for generating image features could be suboptimal for several reasons. First, the features extracted from Faster R-CNN are based on the region proposal network (RPN) which is originally designed to search "object containing" regions in the input image. This could be useful for the traditional image captioning task which requires finding objects in the images but is not appropriate for understanding driving scenes. This is because, first, driving decisions can be related to "non-object" regions which often represent "drivable areas." Second, running RPN itself is computationally inefficient and can significantly impair the overall efficiency of the entire framework.

Regarding the Transformer model, although it was initially proposed for natural language processing (NLP), its superior performance has also been recognized in multiple CV tasks (Wang et al., 2021; Zhao et al., 2020). As a general feature extractor (backbone model), Transformer has exhibited similar characteristics and performance as CNN regarding local correlations but has the additional advantage of capturing long-range correlations within the image. Such capability of capturing long-range correlation is particularly crucial for automated driving tasks because there always exists a "relativity" correlation within the driving scene. For example, in the context of driving directions, "left" is relative to the "right", and if the AV wants to investigate the left region of the driving scene and check the feasibility of a potential left turn maneuver, it also needs to allocate some attention towards the other side of the scene to understand which part is "left". Therefore, traditional CNN may not be ideal for driving scene as it is designed to capture local correlations that tend to overlook the vital relativity correlation such as "left" w.r.t "right." Also, recent research has shown that Transformer based models have global attention capability which can imitate the peripheral vision of human eye and therefore are more suitable for driving tasks (Dong et al., 2022, 2021b). This motivates the investigation of the Transformer based approach, in the current study, for image processing and language generation and, for building explainable autonomous driving models.

In summary, we developed a fully Transformer based model for end-to-end driving scene understanding. The model utilizes the SOTA image feature extractor: Swin Transformer (Liu et al., 2021), and then fuses the image features and language features to generate verbal explanations using another classic Transformer module.

1.6 Study Objectives and Scope, and Report Organization

This report's main objectives are:

- Formulate the traditional end-to-end autonomous driving decision process as an image-captioning task with language-induced visual attention to guarantee the explainability in DL models.
- Develop a fully Transformer based model to generate verbal descriptions and driving actions for autonomous driving.
- Demonstrate the efficacy of the proposed model and ascertain that it possesses superior performance over multiple baseline models in terms of explanation and driving action predictions.

Compared to the classic driving decision models, the developed model can develop driving decisions through reason predictions. Therefore, it enables the system to reap the following benefits prospectively: (1) the capability to exhaustively explore all the possible reasons for generating a driving action; (2) superior attention efficiency by associating attention regions to each word in the generated verbal description; (3) the ability to imitate the human learning process and peripheral vision using language induced attention and full transformer model (4) superior interpretability as an end-to-end goal induced system. From the application perspective, the proposed model can not only enhance the user trust among the AV end users but can also provide insights on model diagnosis and identify gaps where AV developers and manufacturers could make subsequent future improvements.

The rest of Part I of this report is organized as follows: Chapter 2 (Methodology) introduces the methodological background and the details of the proposed model. Chapter 3 (Experiment Settings) documents the dataset, model training specifics, the baseline models, and evaluation metrics. Chapter 4 (Results) compares the proposed approach with all the baselines and provides evidence of its superior efficacy. Then Chapter 5 (Application Context) provides discussions on how the proposed model could benefit autonomous driving systems.

CHAPTER 2. METHODOLOGY

This section first introduces two basic building blocks of the Transformer-based models, namely the Multi-head Self Attention (MSA) layer and Window MSA (W-MSA) / Shifted Window MSA (SW-MSA) layer, and then introduces the overall architecture of the proposed end-to-end Transformer-based image-captioning model.

2.1. Multi-head Self Attention (MSA)

In DL, the multi-head Self Attention (MSA) layer represents a parallel computation (multi-head) of self-attention (SA), which is the basic building block of the Transformer model for the revolutionary BERT model (Vaswani et al., 2017). As the name suggests, SA was initially proposed to conduct attention on 2 identical sequences of input (the input and a copy of itself). Therefore, it has the capacity to capture the "intra" correlations within the input sequence, which is the correlation among different segments of input. Following the same definition in (Vaswani et al., 2017), the attention mechanism is applied to fuse the 3 feature representations, which can be summarized as follows:

$$Attention(Q, K, V) = aV = f_{sim}(Q, K)\, V \qquad (1)$$

where $f_{sim}(\cdot)$ represents the similarity function and $a \in \mathbb{R}^{s \times s}$ represents the attention score. More specifically, it first computes a set of similarity scores for some queries ($Q \in \mathbb{R}^{s \times d_k}$, where s represents the sequence length, d_k is the dimension of both keys and queries) and some keys ($K \in \mathbb{R}^{s \times d_k}$), then use this similarity scores as weights to fuse the values ($V \in \mathbb{R}^{s \times d_v}$) as the features for downstream network. For SA, three representations (K, Q, and V) are all generated from the input ($X \in \mathbb{R}^{s \times d_x}$, d_x is the dimension) using three distinct linear layers (also named as the "projection" layer):

$$K = XW_{\mathrm{K}},$$

$$Q = XW_{\mathrm{Q}} \qquad (2)$$

$$V = XW_{\mathrm{V}}$$

with $W_{\mathrm{K}} \in \mathbb{R}^{d_x \times d_k}$, $W_{\mathrm{Q}} \in \mathbb{R}^{d_x \times d_k}$, $W_{\mathrm{V}} \in \mathbb{R}^{d_x \times d_k}$ as their respective weights for three layers. Then, the similarity function is applied by conducting a dot product between the queries and keys, followed by a softmax normalization for generating the attention score ($a \in \mathbb{R}^{s \times s}$) as shown in **Equation (3)**:

$$a = f_{sim}(Q, K) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \qquad (3)$$

The output of the SA is the matrix multiplication of the attention score (a) and values (V), which completes the computation for a single head (**Equation (4)**). MSA layer parallelly computes multiple SA in each head and fuses the concatenated multi-head results through an additional output linear layer with $W_{out} \in \mathbb{R}^{d_v \times d_{out}}$ as the weights (**Equation (5)**).

$$head = Attention(K, Q, V) = aV \qquad (4)$$

$$MSA(X) = concat(head_1, \dots, head_h)\, W_{out} \in \ \mathbb{R}^{s \times d_{out}} \qquad (5)$$

Compared to single-head SA, using MSA enables each head to focus on different tasks and can attend regions with different ranges. This manipulation significantly enhances the model's flexibility and generalization power. Besides conducting self-attention on 2 identical sequences, MSA can also be applied to capture the "inter" relationship between two different sequences of features (i.e., image and language). The only requirement is that the keys and queries should have the same dimension (this can be achieved easily via the "projection" layers). Then, MSA becomes a cross-attention layer and could use a sequence to induce attention over another sequence. In due recognition of this capability, cross-attention is applied in several multi-modal learning tasks (such as image captioning) to capture the correlation between image and language. In this study, standard MSA is applied for language encoding while cross-attention MSA is used for feature fusion and decoding.

2.2 Window MSA (W-MSA) / Shifted Window MSA (SW-MSA)

The MSA module can be considered the SOTA in the capture of intra-relationships between language sequence (Vaswani et al., 2017) and in modeling the inter-relationship between language and images (Anderson et al., 2018; He et al., 2021). However, it may not have an ideal performance in capturing the intra-relationship within images, because the image has multi-resolution features, and the vanilla MSA lacks a hierarchical structure to recognize objects with varied sizes and resolutions.

In addition, in the vanilla implementation of MSA on images, the Vision Transformer (ViT) model (Dosovitskiy et al., 2020) could become intractable as the image resolution grows. To overcome these shortcomings, inspired by feature pyramid networks (FPN) (T. Y. Lin et al., 2017), Liu et. al (2021) proposed the Window MSA layer (W-MSA) and Shifted Window MSA (SW-MSA) modules (Liu et al., 2021) that restrict the MSA locally within the predefined windows instead of over the entire image. Then the global features are obtained by merging attended image patches of different resolutions in deeper layers.

More specifically, W-MSA and SW-MSA first partition the inputs of Q, K, and V into several windows, and then apply MSA independently within each window. Then, the "transformed" image features of all the windows are assembled back to the original shape by reversing the partition.

$$(S)W - MSA(Q, K, V) = Merge(window_1, window_2, \dots, window_n\,) \qquad (6)$$

$$window_i = MSA\ (X_{\mathrm{i}}) \qquad (7)$$

where $Merge(\cdot)$ is the reverse operation of regular/shifted window partitioning, and X_i is the cropped image feature map within the $window_i$. In terms of window partitioning, as shown in **Figure 2**, they can be either partitioned regularly (**Figure 2 (a)**) or in a shifted window manner (**Figure 2 (b)**).

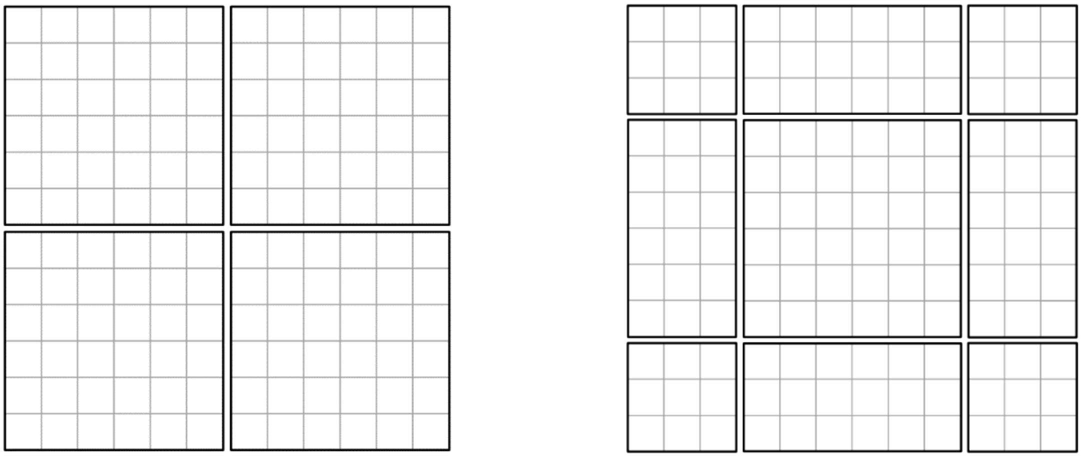


Figure 2. Illustration: (a) regular window partitioning, and (b) shifted window partitioning

Using the W-MSA and SW-MSA layers as basic building blocks, the Swin Transformer (Liu et al., 2021) can perceive objects of different shapes and resolutions in the image. Compared to the ViT model (Dosovitskiy et al., 2020) with only MSA blocks, Swin Transformer has been proven to exhibit superior performance in capturing the intra-relationship of image grid features. Also, it reduces the overall computational complexity from quadratic w.r.t input image resolution as in ViT to linear w.r.t the number of windows. Duly recognizing the benefits of multi-resolution perception and computational efficiency, the current study adopts the Swin Transformer as the image feature extractor in its end-to-end Transformer based model.

2.3 Overall Model Architecture

This section introduces the overall architecture of the proposed end-to-end Transformer based image captioning model. As shown in **Figure 3**, the model follows the classic encoder-decoder structure where the encoder can be further divided into the Image Encoder and Language Encoder for generating image features and language features, respectively.

For processing the driving scene as an image, the preprocessed image (after resizing and normalization) is first fed into the Feature Extractor (green box in **Figure 3**), which is the Swin Transformer backbone (Liu et al., 2021). The architecture of this backbone shares a similar structure as the Language Encoder (blue box in **Figure 3**), except for swapping the MSA layer by W-MSA or SW-MSA layer. Therefore, it is composed of one SW-MSA layer, one feedforward layer, and two addition and layer-norm layers. The output from the Feature Extractor is the transformed patch-based image features with the shape $H \times W \times E$, where H and W are the height and width of the patch grids and E is the feature dimension of each patch. Then the image features are flattened in the spatial dimension into shape $(H \times W) \times E$ and project to the same hidden dimension (E_h) as language features for later fusion. This is achieved through a linear projection layer. The output of the Image Encoder is the "patch embeddings" with shape $(H \times W) \times E_h$), containing the transformed and project feature of each image patch. To uniquely encode the location/position of each patch, the positional encoding as in (Vaswani et al., 2017) is also incorporated before conducting the cross-attention between image features and language features.

On the other branch of the encoder, the language sequences (the textual explanation of the driving scenario) are processed using the Language Encoder (blue box in **Figure 3)**. More specifically, the raw sequence of text is first embedded through a classic embedding layer to obtain numerical representations. Then the word embeddings are fed into a vanilla Transformer model which contains the positional encoding and N_e standard MSA blocks. Each MSA block consists of one MSA layer, one Feed Forward layer, and two Add and Layer Normal layers. The MSA layer is the core layer that computes the attended features following **Equation 2-5**, which is followed by a layer normalization function and one Feed Forward layer. In addition, there exist several "skip links" that directly connect the upstream feature map to the downstream through an adding operation, which is designed to mitigate the problem of gradient vanishing. By leveraging the attention mechanism in this Transformer model, the "meaningful" information in the natural language sequences can be amplified while the "noisy" or "meaningless" information is suppressed.

After separately encoding both image features and language features, they are fused together in Decoder to generate the output words for describing the driving scenario. The Decoder has N_d blocks of similar Transformer based structures as the Language Encoder except that the MSA layer is replaced by the Cross Attention layer. More specifically, the Cross Attention layer computes keys (K) from word embeddings while Queries (Q) and Values (V) from image (patch) embeddings. Therefore, the attention score computation between K and Q can reflect the similarity between image patches and language, which is analogous to using the language to "induce" attention over the image features. By visualizing the attention maps from this Cross Attention layer, it can be ascertained whether the model has attended to the correct regions for generating the correct explanation. This can provide the ideal level of explainability as mentioned in the introductory section of this report. The outputs of the Decoder are the fused features that integrate all the useful information from both the language context and the image, which are fed into the classification head for generating the output words. In this work, the classification head is simply a Linear layer with the Softmax activation function.

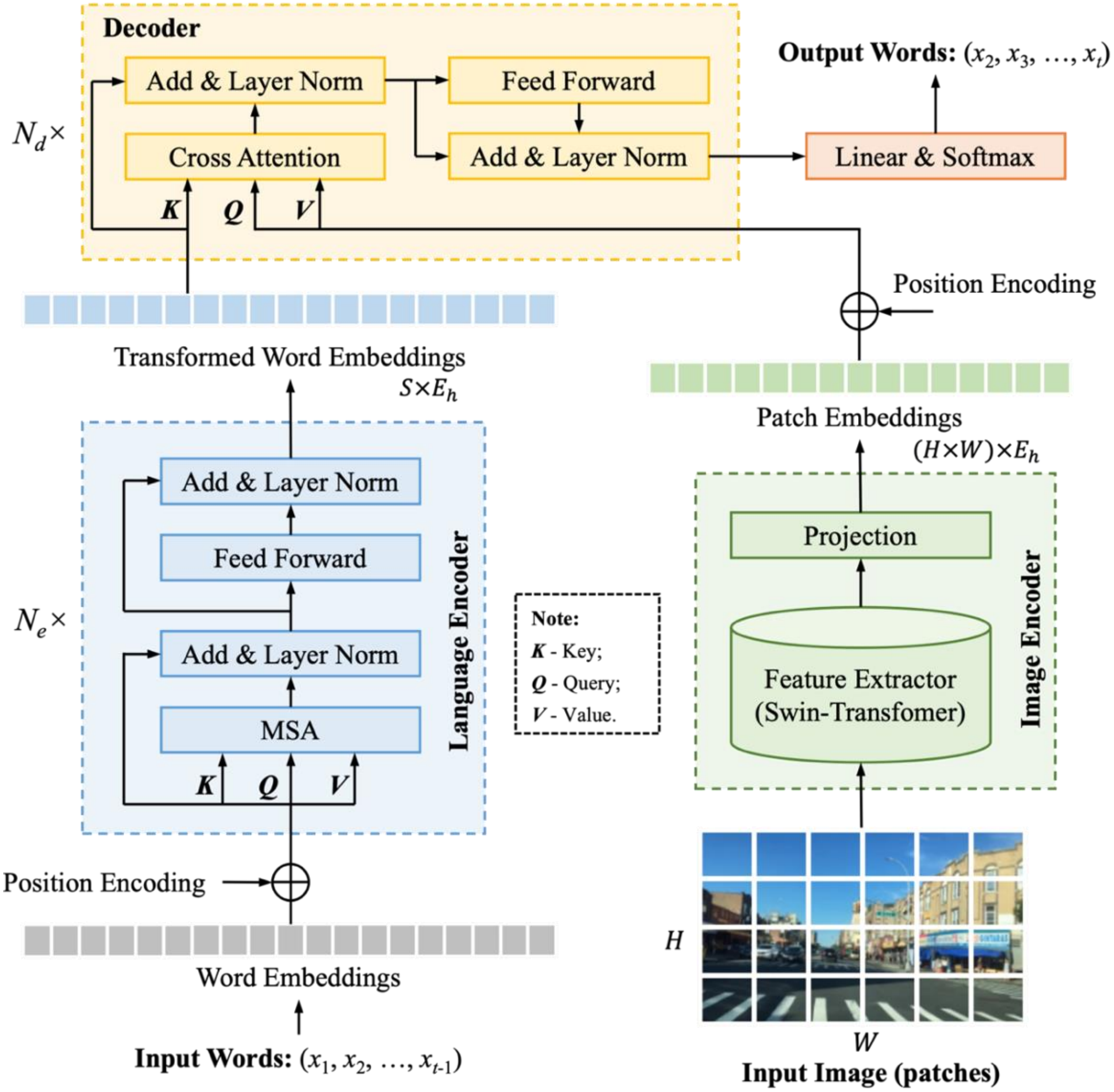


Figure 3. Model architecture

2.4 Loss Function

For the model training, the current study used other image captioning models (Anderson et al., 2018; He et al., 2021; Xu et al., 2015a) and utilized a standard temporal multiclass Cross-Entropy loss function as shown in **Equation (8)**:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{c=1}^{M} y_t \, log\left(p(\widehat{y_t})\right) \qquad (8)$$

where M is the number of total classes (vocabulary size), T is the total length of prediction (the maximum length of the explanation sequences), and N is the total number of training examples (training set size), y_t and $\widehat{y_t}$ are ground-truth label word and predicted word, respectively. Consistent with the standard sequence-to-sequence (C2C) schema for image captioning, the input words and output words are shift-by-one, which uses the sequence $x_1 \dots x_{t-1}$ to predict the sequence $x_2 \dots x_t$. With this loss definition and training settings, the model can be trained end-to-end with the driving scenario images and textual descriptions.

2.4 Action Generation

As the ultimate goal of autonomous driving systems is to generate driving decisions, one additional indispensable step is to map the model output (textual descriptions) to executable driving actions. This can be achieved simply using a keyword-matching algorithm and a rulebook. For example, the text description “No lane on the left” can be translated to “follow the traffic” or “turn right” but not “left lane change.” This rulebook follows simple logic and is easy to acquire.

CHAPTER 3. EXPERIMENT SETTINGS

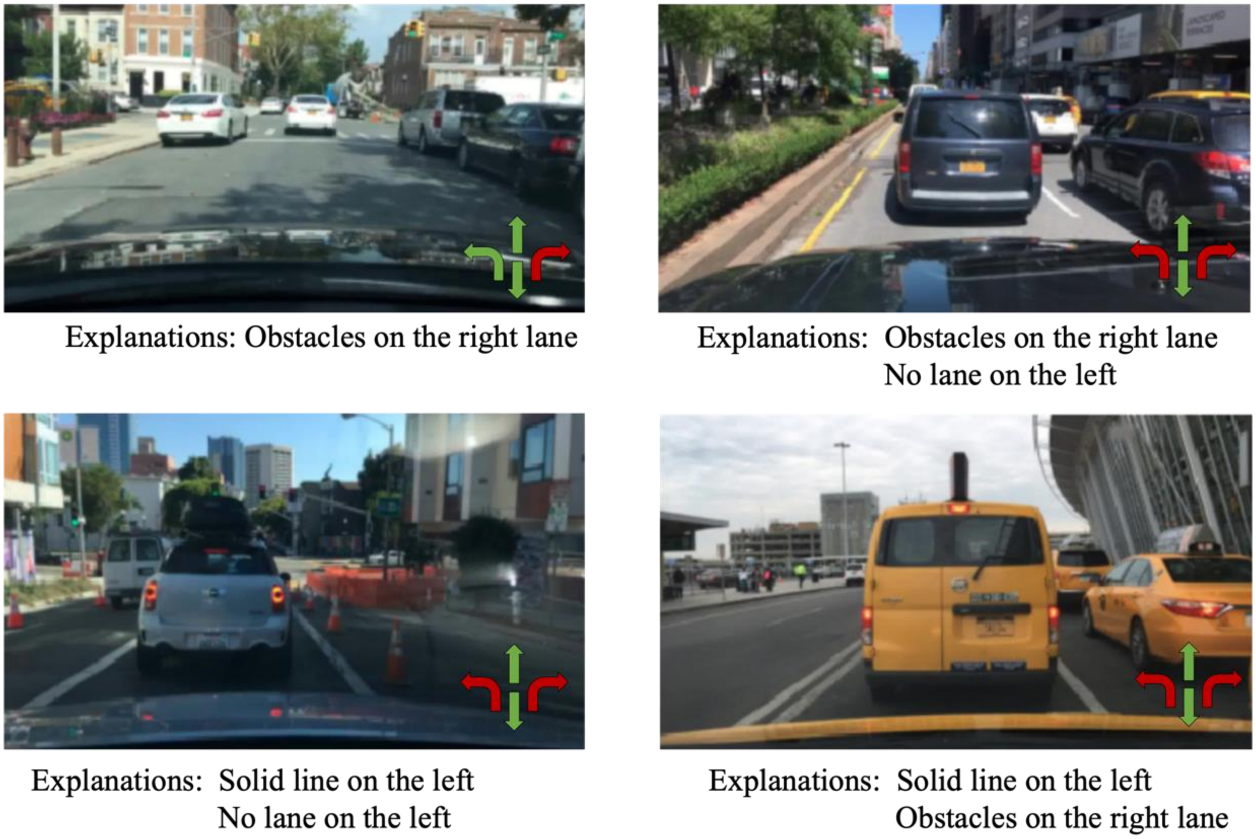
3.1 Dataset Preparation

Experiments were carried out to train and evaluate the proposed model, using a subset of data culled from the BDD Object Induced Actions (BDD-OIA) dataset (Xu et al., 2020). BDD-OIA is an extension of the well-known image-based driving dataset BDD-100K (F. Yu et al., 2020) which possesses frame-by-frame labels of driving actions and corresponding explanations. In the raw dataset, the driving actions are the high-level "feasible" actions that can be undertaken at that specific time step, including "move forward," "slow down/stop," "turn left," and "turn right." Here, the lane changing decisions such as "left/right merge" are considered as "turn left/right" categories. The corresponding explanations can be summarized into 21 classes (**Table 1**).

Table 1. Actions with Corresponding Explanations in BDD-OIA

Actions	Explanation classes
Move forward	Traffic light is green
	Follow traffic
	Road is clear
Slow down/Stop	Traffic light is red
	Traffic sign
	Obstacle: car
	Obstacle: person
	Obstacle: rider
	Obstacle: others
Turn left	No lane on the left
	Obstacles on the left lane
	Solid line on the left
	On the left-turn lane
	Traffic light allows
	Front car turning left
Turn right	No lane on the right
	Obstacles on the right lane
	Solid line on the right
	On the right-turn lane
	Traffic light allows
	Front car turning right

The original dataset includes explanations for both feasible and infeasible actions. For example, an explanation for the "Move forward" action could be "Traffic light is green," which is feasible. On the other hand, explanations for the "Turn left" action, such as "Solid line on the left," focus on why the action is infeasible. Ensuring driving safety requires preventing the AV from executing infeasible actions, and as such, the explanations associated with them are crucial. Therefore, in this research, the focus is on the actions that are associated with infeasibility. In addition, as the raw dataset is unbalanced – some classes have more than 10,000 images while others have less than 20. Thus, a subset was selected to contain six of the most frequent reasons: “**obstacles on the left lane**,” “**no lane on the left**,” “**solid line on the left**,” “**obstacles on the right lane**,” “**no lane on the right**”, “**solid line on the right**”. Another reason for selecting these six classes is that they are associated primarily with left/right turn or lane-change actions which have been identified as the maneuvers that are most prone to collision (Xu et al., 2019). A snapshot of the selected dataset is shown in **Figure 4**, where the green arrows in the figure represent the feasible driving actions and the red arrows represent the infeasible ones. The explanation texts below each frame are the target captions that the model needs to predict. In summary, for the selected six reasons, the corresponding total numbers of images are: 8,475, 7,989, 7,625, 11,521, 6,771, and 4,750 respectively, with the total number of frames adding up to 24,921, we further partition this dataset into a training set (19,936 frames) and testing set (4,985 frames).



Explanations: Obstacles on the right lane

Explanations: Obstacles on the right lane
No lane on the left

Explanations: Solid line on the left
No lane on the left

Explanations: Solid line on the left
Obstacles on the right lane

Figure 4. Example of the selected BDD-OIA dataset

3.2 Language (Reasons) Sentences Preprocessing

Regarding the preprocessing of the reasons (the objective of which is to make them readable for neural networks), we apply two NLP preprocessing techniques – tokenization and word embedding. Tokenization splits the sentence into words and assigns a unique token to each word. In addition to all the unique tokens, we add four extra tokens: <SOS> start of sentence; <EOS> end of sentence; <;> the delimiter of the reasons; <NULL> the placeholders after the <EOS> token. For each sentence, we first add the <SOS> and <EOS> and delimit the reasons inside a sentence with the script <;>. For this work, different images can have different numbers of reasons, and the number of words for each reason is different. Therefore, this problem is a dynamic-length input-output problem with variable sequence length. To accommodate such discrepancy in the sequence length for all images and enable the model to be trained by batch, a padding manipulation is applied by adding <NULL> token to the positions after the <EOS> token. In this manner, the model can always accept and predict a fixed-length sentence. Here, we use the maximum length of the possible sentence as this fixed length, which is the word count of the total of six reasons plus the delimiter and start/end token. After tokenization, we apply a standard word embedding layer to transform the token representation to numerical vectors (word embeddings). Initially, such numerical representation is generated from a standard Gaussian distribution. After training, it is observed that words with similar meanings tend to have "closer" representations in the embedding space.

3.3 Baseline Models

3.3.1 Image Captioning Model

A comprehensive review of relevant published literature showed that to date, the literature contains only very few studies that cast roadway-scene interpretation in autonomous driving as an image-captioning task. As such, it is difficult to conduct a comprehensive comparison between our proposed method with existing literature. Therefore, we trained multiple image captioning models to perform the same task on our dataset to serve as baselines. These models possess two architecture categories: CNN + LSTM (Xu et al., 2015) and CNN + Transformer. We herein provide details of each below:

- CNN + LSTM

  As shown in **Figure 5**, this baseline model follows the main concept in (Xu et al., 2015) while using CNN based model for modeling images and RNN based model for modeling language. Overall, it uses a classic CNN backbone model as the encoder to generate the patch (grid) based features of the image, then uses an LSTM decoder to generate the language sequences. The key idea for fusing features from images and language sequences lies in the Attention Network, which uses the classic dot product attention (Bahdanau et al., 2015) to capture the correlation between these 2 feature spaces. Under this setting, visual attention is achieved by taking the dot product between the image features and the LSTM hidden states. Therefore, a higher attention score means more correlation between the visual features and the language contextual features. Then, the attended visual features and the contextual language sequences are fed into the Language Decoder to generate the textual description for the driving scenario. In the following experiments, we further tested two classic backbone models namely Resnet50 (He et al., 2016) and Mobilenet_v2 (Sandler et al., 2018) in the Image Feature Encoder. For the Language Decoder, 2 layers of bidirectional LSTM are adopted. Comparing our proposed model against this model can justify the efficacy of Transformer based approaches in understanding the driving scenario.

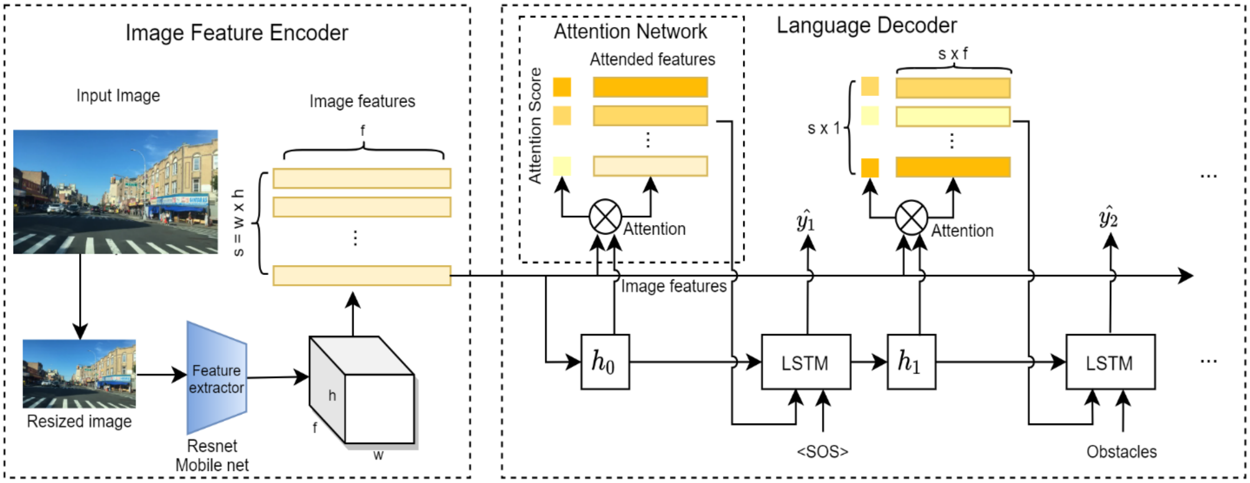


Figure 5. Baseline: CNN + LSTM Architecture

- CNN + Transformer

  This model has a similar structure as the proposed full model in **Figure 3**, except for swapping the Swin Transformer-based image Feature Extractor with the CNN-based backbone ResNet50. Since the main goal of this work is not to build an image processing network, we adopt the pre-trained Swin Transformer and ResNet backbone and fix their weights during the training and testing phases. The two alternative backbones applied have both been pretrained on Image Net.

3.3.2 Traditional Prediction Model

To provide further evidence of the superior efficacy of the proposed concept (generating the driving actions by first predicting the textual explanations), a traditional classification model same as that proposed by (Xu et al., 2020) is trained on the same dataset selected from BDD-OIA. This model applies an “object-induced attention” mechanism that scores the region proposals and utilizes only the top-scored regional features for predicting actions and reasons. More specifically, the model is built on top of a Fast-RCNN’s backbone feature encoder and region proposal network (RPN), then a “selector” structure is trained to filter only those regions that are highly correlated to the downstream driving action and reason predictions. This model benchmarks the performance in terms of action and explanation classification on the BDD-OIA dataset.

3.4 Evaluation Metric

The model seeks to comprehend the driving scenes with human language. As such, the model strives to generate legal, human-understandable, and correct textual descriptions of the driving environment. Therefore, it is desired that the evaluation metrics for this model should consider both the quality of generated language and the correctness of the generated explanations.

To evaluate the quality of generated language, we incorporate the classic evaluation metric in image captioning and neural translation, the Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2002). This score measures the *n*-gram similarity between two sentences by computing the overlap between a hypothesis sentence and a reference sentence. A higher BLEU score represents a higher similarity between the generated sentence and the ground-truth sentence, which further indicates higher language generation quality. In this study, we recorded, for each generated explanation in the test set, the average BLEU-4 score which is a widely-used evaluation metric for image captioning work.

However, as mentioned in past research (Cui et al., 2018), the BLEU score alone is unable to capture the semantic meaning of the sentence, which may lead to a poor correlation with human judgment. To evaluate the semantic quality of the generated reasons and examine whether the model has exhaustively exploited all the potential reasons for the driving scene, we further compute the F1 scores for the reasons predicted. To compute the F1 score for the reasons, we first carry out processing to transfer the generated sentences back to the 6 reason label classes. This is because there exist cases where the model-generated sentences are not word-to-word aligned with the label sentences but possess the same semantics. For example, the model may generate the sentence: “obstacles on the **right**” which is not exactly in the label sets but has the same meaning as the label sentence “obstacles on the **right lane**.” To project the predicted sentences to the label sentences, in this example, we adopt the keywords matching strategy and only investigate the keywords such as “obstacles, right.”

As the ultimate goal of autonomous driving is to generate driving decisions, in this study, we also compute the F1 score for the “infeasible” actions, that are the two actions “cannot turn left” and “cannot turn right” associated with 6 aforementioned reasons: “obstacles on the left lane”, “no lane on the left”, “solid line on the left”, “obstacles on the right lane”, “no lane on the right”, “solid line on the right”. In mapping the reasons to actions, we follow the simple rule: if there exist any of the first 3 reasons, then the vehicle cannot turn left; and if there exist any of the last 3 reasons, then the vehicle cannot turn right.

In addition, two versions of F1 scores, namely overall F1 score ($F1_{all}$) and mean in-class F1 score ($mF1$), for both reasons and actions are computed. $F1_{all}$ is the F1 score over all the predictions and is calculated as follows:

$$F1_{all} = \frac{1}{|P|}\sum_{k=1}^{|P|} F1(\widehat{P_k}, P_k) \qquad (9)$$

Where $\hat{P}$ is the predicted class and P is the true label, which can be both reasons and actions; $|P|$ is the total number of predictions over the entire test set. Meanwhile, despite having our subsampling-based preprocessing, the dataset is still unbalanced. There are always more scenarios for the “containing obstacles” reason compared to the “containing solid line” reason. Therefore, we calculated the F1 score within each predicted class, and finally take the average over all the classes to compute the in-class $mF1$, as follows:

$$mF1 = \frac{1}{C}\sum_{c=1}^{C}\sum_{i=1}^{n} F1(\widehat{P_{\imath}^{c}}, P_{l}^{c}) \qquad (10)$$

Where C represents the number of classes in prediction (6 for the reasons and 2 for the actions, respectively), then P^c becomes the binary prediction of class c, indicating whether class c is presented in this frame.

3.5 Training Specifics

Regarding the detailed model architecture, we use the following parameters: the hidden embedding size $E_h = 256$, the number of MSA head $h = 8$, the number of blocks for both Language Encoder and Decoder $N_e = N_d = 2$. For the Image Encoder, we utilize the same swin_tiny architecture as in (Liu et al., 2021) and initialize it with the pretrained weights on Image Net 1000 class (IMAGENET1K_V1). The weights are finetuned with other network components during the training process. To prevent overfitting, a dropout rate of 0.1 is applied on all linear layers. All the linear layers utilize an input and output dimension of E_h and GeLU as the activation function (Hendrycks and Gimpel, 2016). Regarding the training process, we trained our network for 20 epochs using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 1×10^{-4}.

CHAPTER 4. RESULTS

4.1 Quantitative Analysis

Five (5) models were trained: the proposed fully Transformer based model (Swin Transformer + Transformer), two classic CNN + LSTM models with two different feature encoders (Mobilenet_v2 and Resnet50), one CNN (ResNet) + Transformer model, and the original BDD-OIA classification model. **Table 2** presents their numerical evaluation metrics (Avg.BLEU-4 score and F1 scores) for both reasons and actions.

It is observed that our proposed model, the fully Transformer based approach achieves the highest performance in most of the above-mentioned numerical evaluation metrics, demonstrating the efficacy of the model in both reason generation and action predictions. Also, from the column representing the number of parameters (**Table 2**), it can be observed that the proposed fully Transformer model does not add much extra computational cost but can significantly improve the performance compared to other image captioning baselines and is significantly more computationally efficient compared to the classic BDD-OIA classification model. From these results, the following complementary conclusions could be made: (a) Regarding the decoder, the Transformer structure outperforms the LSTM in capturing the inter-correlation between image features and language features, (b) Regarding the image feature encoder, Swin Transformer outperforms CNN when utilizing the same transformer as decoder, and (c) The Mobilenet and Resnet encoder have similar performance in the CNN + LSTM model; this observation underscores the bottleneck associated with the decoder LSTM model.

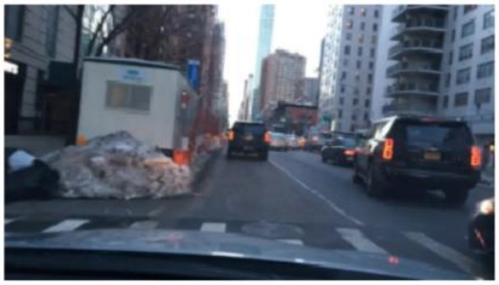
Another important observation from **Table 2** is that the proposed learning paradigm “generate the actions by first generate reasons” prevails over the traditional classification model proposed by Xu et al., (Xu et al., 2020), particularly in terms of action prediction. More specifically, our proposed model was found to be capable of achieving F1 scores exceeding 0.9 even though the reason predictions are less accurate than those of the traditional classification model for CNN + LSTM models. This could be due to the capability of our proposed model to enhance system redundancy. As mentioned earlier, multiple reasons could lead to the same driving decision/action. For example, the decision to “not turn right” can be generated if there exist 2 reasons “obstacles in the right lane” and “solid line in the right”. Therefore, the model can make the right decision if either of the two reasons is predicted. This additional redundancy in the decision generator further enhances the overall reliability of the proposed driving system. However, for the requirements of explainability, the ultimate goal of the model is still to completely exploit all the possible existing reasons.

Table 2. Performance Measures of Proposed Model and Baselines

Model		Nr. of parameters	Reasons Avg. BLEU-4 score	Reasons F1	Reasons inner-class mF1	Actions F1	Actions inner-class mF1
CNN + LSTM	Mobilenet	**2.64M**	0.478	0.521	0.682	0.917	0.847
	Resnet	24.05 M	0.497	0.529	0.678	0.917	0.848
Resnet + Transformer		25.92M	0.578	0.633	0.801	**0.935**	0.901
Fully Transformer		26.61M	**0.584**	**0.662**	**0.823**	0.932	**0.913**
BDD-OIA classification model (Xu et al., 2020)		45.00M	-	0.649	0.562	0.833	0.771

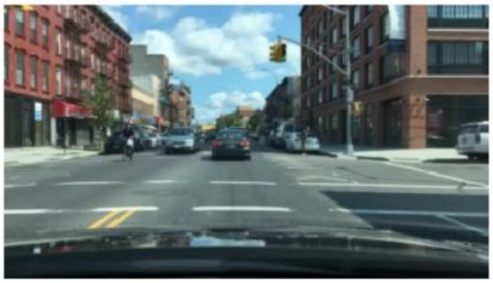
4.1 Qualitative Analysis

Figure 6 presents 4 randomly selected frames from the test set and the corresponding prediction results. As expected, the model was found to be capable of generating legal sentences with correct words and grammar to describe the driving scenes. By visualizing these generated explanations, it can be concluded that the proposed model has the capacity to fully understand the driving scenarios, and the driving decisions/actions can be made without ambiguity.



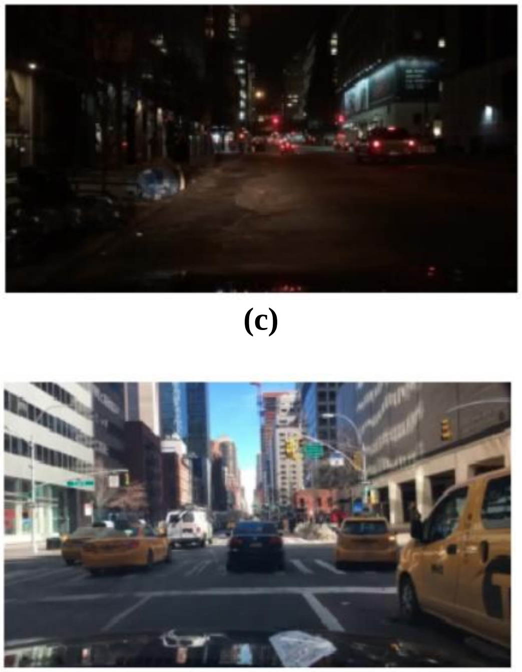
(a)

PD:
R: <SOS> no lane on the left; obstacles on the right lane <EOS>
A: cannot turn left, cannot right turn
GT:
R: obstacles on the right lane; no lane on the left
A: cannot turn left, cannot right turn



(b)

PD:
R: <SOS> no lane on the left; solid line on the left <EOS>
A: cannot turn left
GT:
R: no lane on the left; solid line on the left
A: cannot turn left



(c)

PD:
R: <SOS> no lane on the left; obstacles on the right lane <EOS>
A: cannot turn left, no right turn
GT:
R: no lane on the left
A: cannot turn left

(d)

PD:
R: <SOS> obstacles on the right lane; solid line on the right lane <EOS>
A: cannot turn right
GT:
R: solid line on the left; solid line on the right; obstacles on the right lane
A: cannot turn left, cannot turn right

Figure 6. Sample predictions on the BDD-OIA dataset (PD: prediction, GT: ground truth, R: reasons, A: actions, green: true positives, red: false negatives, blue: false positives)

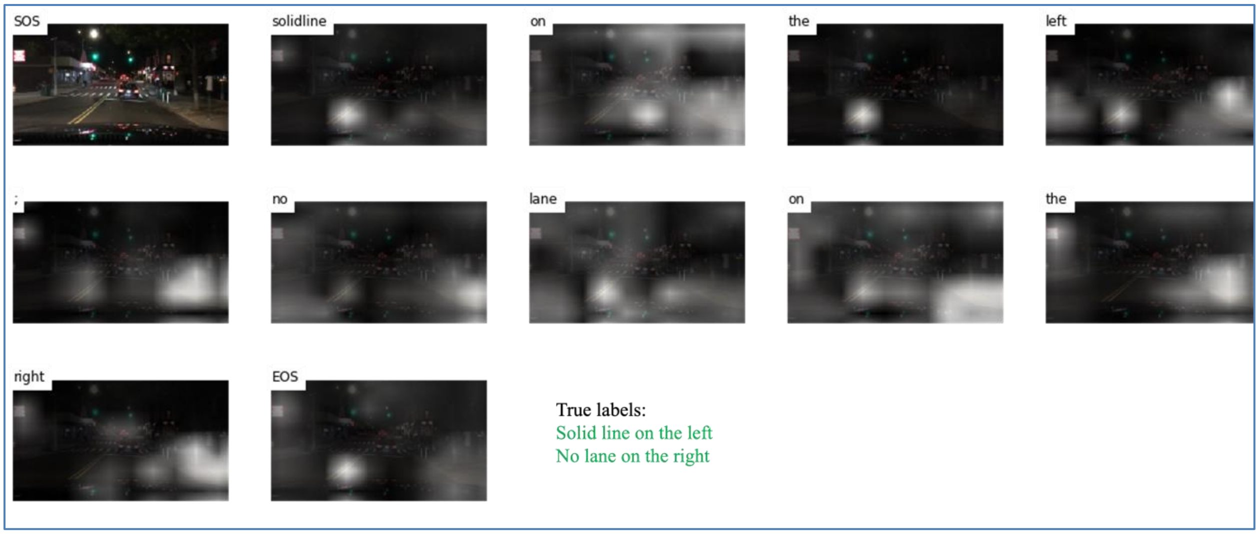
4.3 Attention Visualization

One of the prominent benefits of the proposed full attention model is that by visualizing the attention map, the rationale of the model can be investigated further. **Figure 7** presents four samples of attention maps from the last Cross-Attention layer in the decoder (there are 2 transformer blocks with 2 Cross-Attention layers, we plot the attention maps for the upper layer since they can capture the higher-level features and correlations than the lower layer). These attention maps are generated by taking the average of attention maps over all 8 heads.

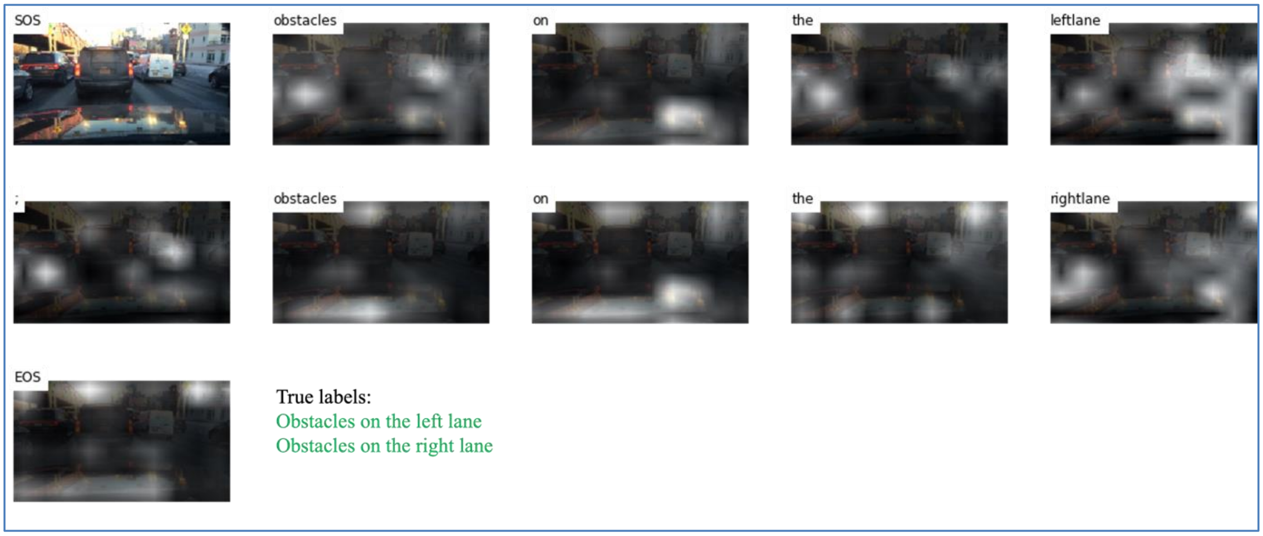
These results demonstrate that the model can indeed capture the semantic relationship between the image regions and language. For example, in **Figure 7 (a)**, when generating the reason "solid line on the left", the model's attention is specifically guided towards the region of the solid line. This ability can also be demonstrated in **Figure 7 (b)** when the model attends to both left and right regions for identifying obstacles.

Another interesting finding from these attention maps is that the model learns the concept of "relativity." More specifically, when generating words with directional information (the left, left lane, etc.) the model tends to simultaneously attend to the opposite direction (to the right, in this case). This can be seen in most of cases. We believe the reason for this phenomenon is that directional information such as left vs. right can only be understood relatively. By attending both regions, the model can identify which direction is "left" or "right."

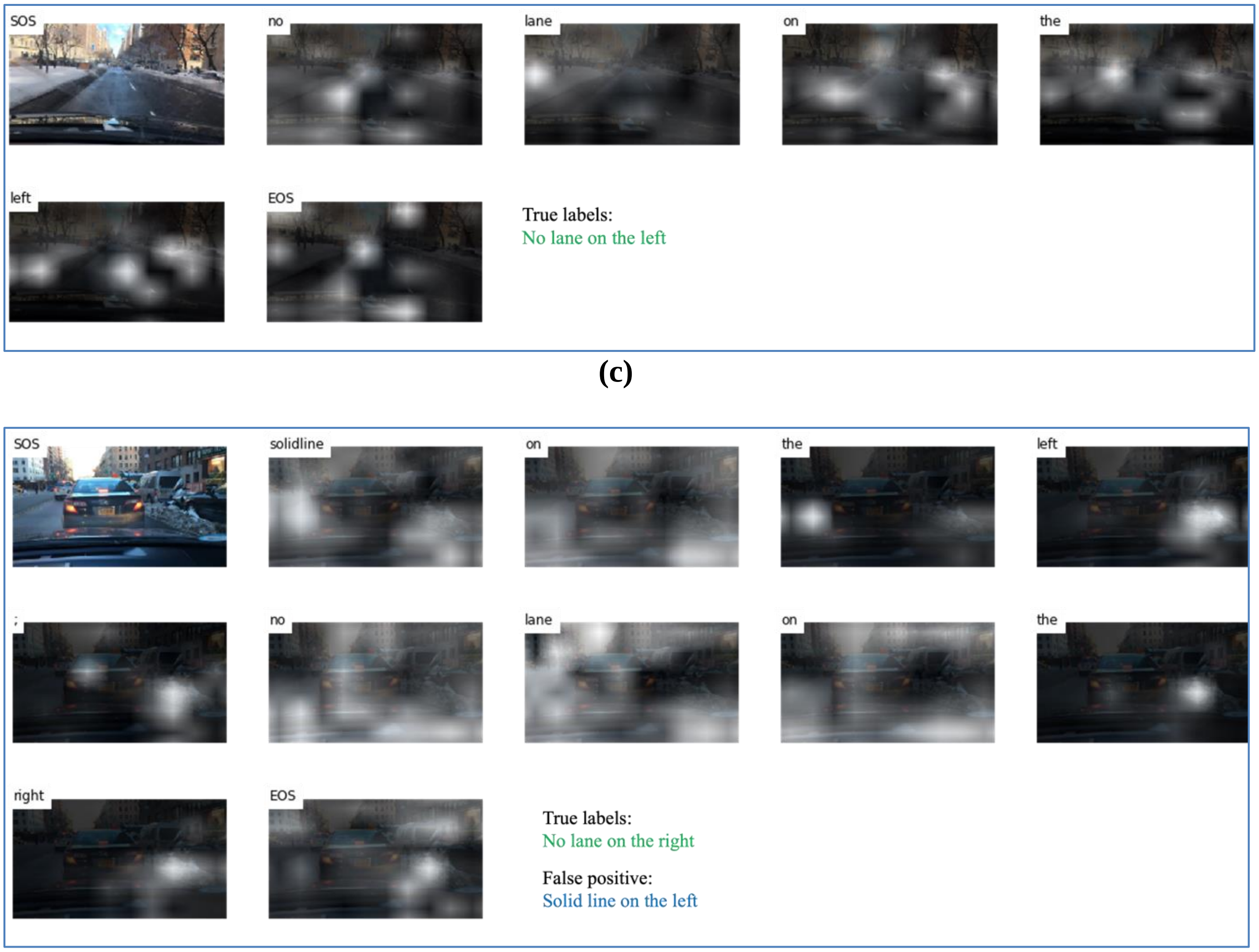
From the perspective of bionics, the proposed reason-induced visual attention mechanism is observed to intuitively imitate the training of a human driver. During the "in school" training period, the coach uses language to first guide the trainee's attention to look towards important objects and understand the scenario, and then to generate the driving decisions through simple causal relationships. After the training, the trainee can "self-attend" to the important regions without the trainer's instructions.



(a)



(b)



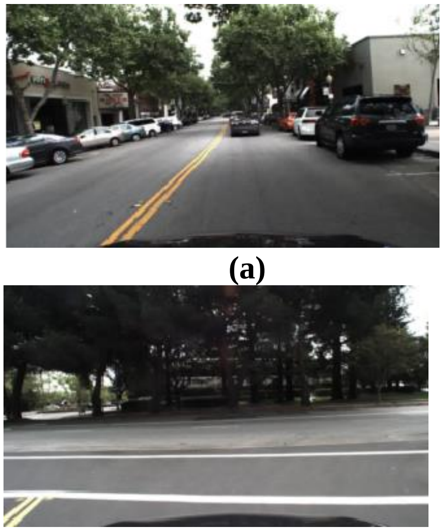
(c)

(d)

Figure 7. Examples of attention maps (brighter parts indicate the attention regions for each word during the reason sentence generation) on the BDD-OIA dataset.

4.4 Cross Dataset Evaluation

Any model intended for deployment in a safety critical environment should possess adequate robustness and generalization ability such that even when it is trained with a specific dataset, it can be transferable to a different dataset without loss of integrity. This capacity also indicates that the model is not overfitting a specific dataset and can handle "out-of-sample" data points. To demonstrate our model has such cross-dataset generalization ability, we further test its performance using the Honda Research Institute Driving Dataset (Ramanishka et al., 2018) with the model trained on the BDD-OIA dataset (see Section 3.1). Since these two datasets do not share the same label space, only a qualitative analysis can be carried out. However, by visualizing the randomly selected prediction results as well as the attention maps, it is still sufficient to conclude that the model possesses adequate robustness and can consistently generate plausible scenario descriptions and attention maps, even when the inputs are from different datasets.



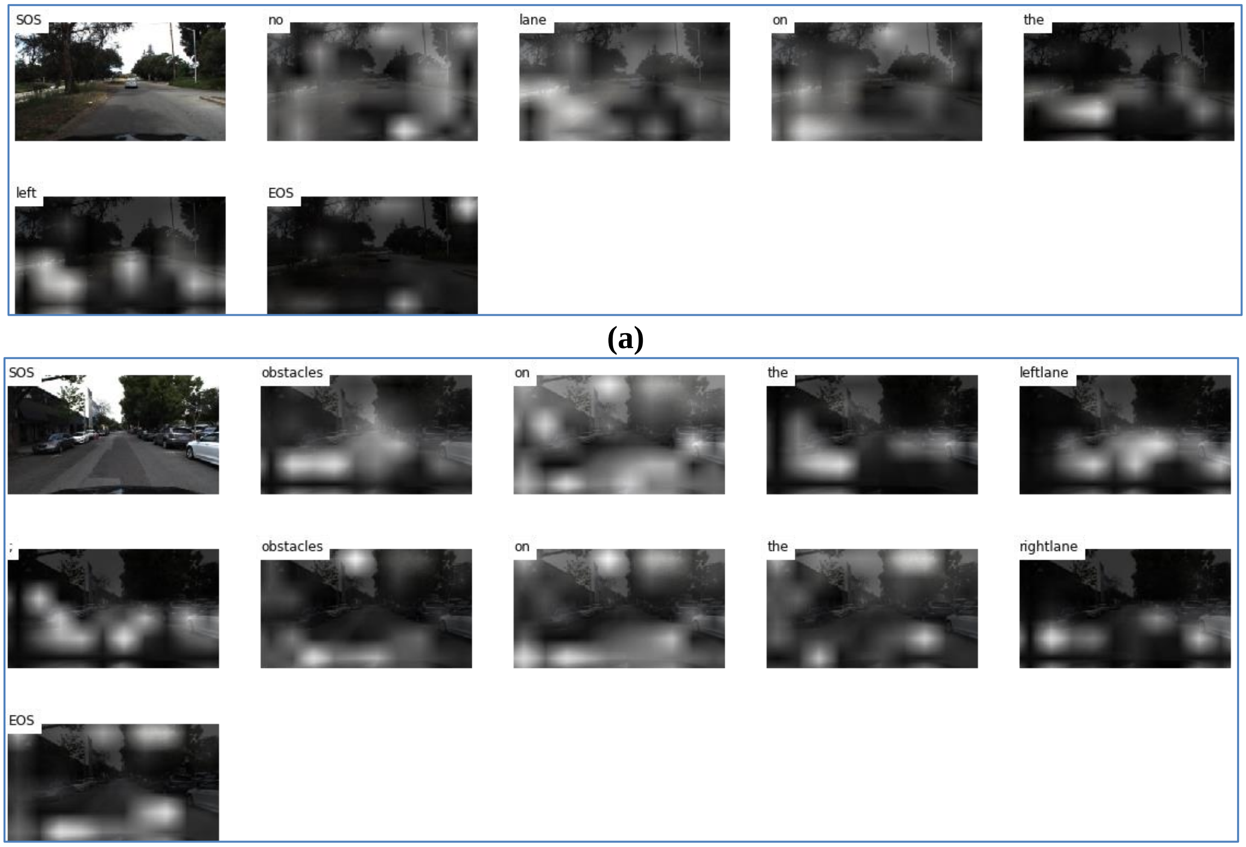
(a)

PD:
R: <SOS> solid line on the left; obstacles on the right lane <EOS>
A: cannot turn left, cannot right turn

(b)

PD:
R: <SOS> solid line on the left <EOS>
A: cannot turn left

Figure 8. Sample predictions on the HDD dataset



(a)

(b)

Figure 9. Examples of attention maps on HDD dataset

CHAPTER 5. APPLICATION CONTEXT

From the perspective of explainable model applicability, the attention maps and the generated explanation sentences can help AV developers to understand the model behaviors and shortcomings of the existing model. For example, a shortcoming of the proposed model is that it does not possess the "concept" of distance, i.e., the model can attend to obstacles and solid lines but cannot understand that in order to generate the driving decisions, only the objects in close proximity should be attended to. For example, in **Figure 7 (c)** and **Figure 7 (d)**, the model provides false positive predictions of the obstacles and a solid line. These objects indeed appear in the two driving scenes but are far away from the ego vehicle and should not be attended to for reason prediction. With these false positives, the model may make overly conservative decisions. This issue can be mitigated using a finer-labeled dataset which has labels for objects that are "closer" and those that are "further," for example, we can augment the existing label sentences for the reasons to contain the distance "obstacles in the right, approximately within 20 meters". After training the model, the model should be able to distinguish between objects and obstacles based on their proximity. In summary, by visualizing the attention maps and generated reasons, the model behaviors can be understood. This process can be used to identify the weaknesses of the currently trained model and to seek potential directions for improvements. These extra benefits in interpretability and diagnosability do not exist in most other DL models that exist in the literature.

From the perspective of application, the trained model can be directly embedded into autonomous driving systems of Level-3 and above. Due to its capacity of quickly eliminating infeasible actions, it can serve as a "sanity check" for the AV maneuver selection system. That is, when the AV makes "infeasible" or potentially dangerous decisions, the proposed model can quickly override the decisions and provide the human operators with explanations for the override. Furthermore, if the proposed model has been trained with richer explanations and scenarios, it can work as a decision generator that can directly generate the maneuvers for the AV to execute.

Regarding the computational cost, we duly recognize that the process of generating explanations is more expensive and could inevitably cause delay in real-time implementation. However, the level of driving decision (LODD) for the proposed model is "tactical" (Dong et al., 2021a), which refers to mid-level driving behavior planning and maneuver selections such as lane changing, merging, and driving decisions at signalized intersections. These tactical decisions, unlike low-level "operational" commands (braking, pedal positions, steering angles that need real-time updating), can be generated at a lower frequency (e.g., every one or two seconds). Therefore, the proposed model and the proposed concept of casting the driving task into image captioning task can fulfill the requirement of generating such maneuver-level decisions.

PART II

Joint Prediction of Potential Driving Actions With Corresponding Explanations

CHAPTER 6: INTRODUCTION

6.1 Background

Motivated by the challenges associated with safety and mobility in the traditional highway environment and spurred by ongoing advancements and opportunities in information and robotics technologies, government agencies and the automobile industry continue to seek guidance on the measurement of performance in the context of the new transportation technologies. As is the case with any new transportation stimulus including technological innovations, it is imperative to assess performance based on a carefully-designed portfolio of performance measures (FHWA, 2019; Sinha and Labi, 2007; World Bank, 2005).

In the context of automated and connected vehicle operations, performance may be measured from the perspective of the impact type (safety, mobility, privacy, equity, for example), impact direction (costs and benefits), and the affected stakeholder (the transportation agency, road user, and the community) (Lioris et al., 2017; TRB, 2018, 2019; Litman, 2023). Unfortunately, the deployment of AV systems in the real world has been severely limited due to various obstacles associated with policy and regulation, infrastructure readiness, technology, and so on. For example, a number of key technologies associated with perception and decision processors still have not reached a level of advancement where they can be applied reliably to produce error-free AV systems.

6.2 Perception as a key consideration

Autonomous driving is a complicated end-to-end system which contains a sequence of sub-systems or modules including sensing, perception & localization, abstraction, planning, and control (**Figure 10**), and each module is achieved through the integration of multiple technologies such as sensing, signal processing, data analytics, machine learning, artificial intelligence (AI), and control theory. Of the modules, perception (second block in **Figure 10**) is considered the most vulnerable link in the chain (NTSB, 2019).

There are multiple reasons for this. To begin with, the perception module is one of the very initial blocks of the entire autonomous driving process, any error at the perception phase will not only cascade but also be amplified across the subsequent stages. For example, failure in detecting the road participants (i.e., pedestrians, cyclists, and neighboring ground vehicles) could be catastrophic because an appropriate evasive maneuver will not be planned in the following phases. This has been the underlying cause of several AV-related fatal accidents in recent years, including well-known instances of Uber and Tesla vehicle collisions with pedestrians (McCausland, 2019; Yadron and Tynan, 2016)).

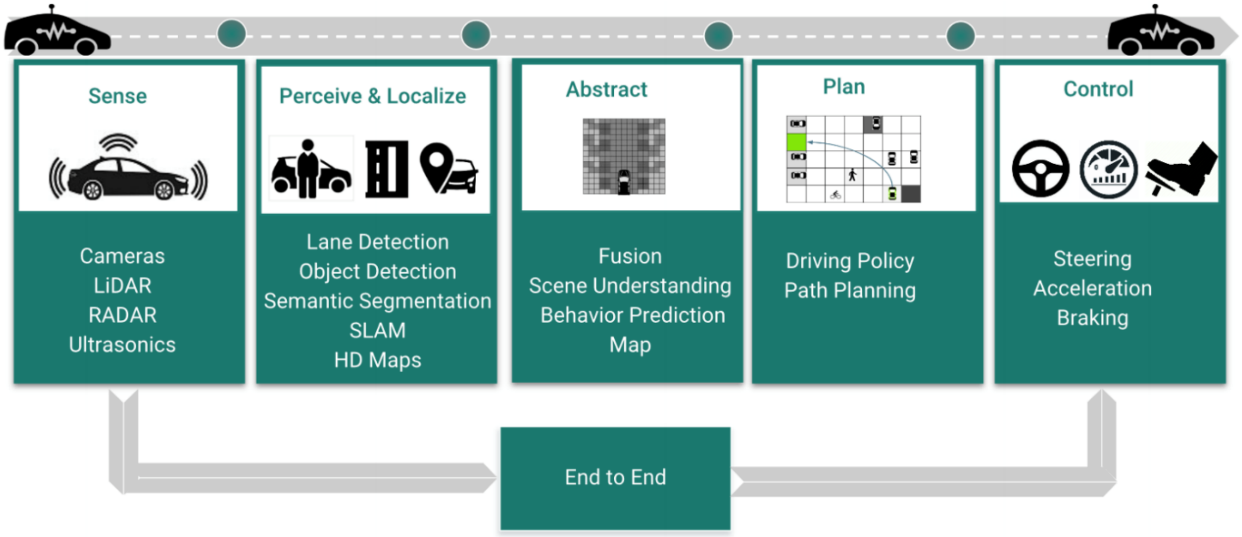


Figure 10. The end-to-end autonomous driving task (image from (Talpaert et al., 2019))

6.3 Application of perception in existing driving systems

Recently, computer vision (CV) based perception technologies have been widely used in multiple applications in driving assistance systems (ADAS) (Horgan et al., 2015; Sowmya Shree and Karthikeyan, 2018), for example, systems for lane detection, traffic sign recognition, and forward collision warning. These new features in ADAS have enhanced driving safety and convenience. However, these modules are "scattered" in the sense that they are designed to accomplish specific functions in an independent manner. As a result, they do not cooperate with each other, and are unable to provide full situational awareness of the driving environment for purposes of autonomous driving. For example, the obstacle detection module can detect only the barriers in the surrounding location but cannot cooperate with the lane marker detection module. Therefore, the obstacle detection module still requires the human brain to fill the gap in such knowledge, and to achieve a comprehensive characterization of the driving environment.

Furthermore, these modules in ADAS are currently designed for human driving (where it is required that the human driver is always focused during driving), not AV operations. This means that when developing vision based ADAS, the reliability may be compromised. Therefore, such ADAS systems cannot provide a comprehensive and precise understanding of driving environments, and thus cannot be applied directly to fully automated vehicles. For the perception phases of AV operations, a more sophisticated, integrated cooperative, and reliable CV system is needed. In addition, unlike human vision which can quickly identify salient objects and grasp the main semantics in a driving environment, CV models tend to (inadvertently) misallocate computation resources towards analyzing areas of the driving environment (for example, the background sky and buildings) that may be irrelevant to the driving task at hand. To alleviate this situation, an appropriate "attention" mechanism to "guide" the CV model to focus only on relevant areas of the driving environment, is needed.

More recently, with the emergence of connectivity devices, perception can be further enhanced by vehicle connectivity (this yields the often-termed "connected autonomous vehicle

(CAV)") since more accurate and direct information can be disseminated through the connectivity devices. It has been postulated that the benefits of combined automation and connectivity will exceed the sum of individual benefits from these two technologies (Ha et al., 2020). In the past few years, several advanced learning-based approaches have been applied in CAV operations in contexts including information fusion and cooperative control (Chen et al., 2021; Dong et al., 2021; Dong, Chen, Joun Ha, et al., 2020; Dong, Chen, Li, et al., 2020). We duly recognize the coupling of connectivity and automation can accentuate the benefits of automation. However, there is still a long way to go before achieving full connectivity for all vehicles on the road. Therefore, in this part of the study, we address only the perception tasks of a single vehicle for which we seek to enhance the interpretability of its perception module.

6.4 Research Gaps and Main Objectives/Contributions of this Study

Xu et. al. (2020) proposed an object-induced attention mechanism that performs attention over the detected objects and uses only the relative objects to generate driving actions and explanations. More specifically, their model utilizes a "selector" structure to "crop" the fused regional features. This fused regional feature is generated by stacking the regional features that are computed using Ren et al's FasterRCNN's region proposal network (Ren et al., 2017) (referred to as the local branch) and the raw overall feature map for the entire image (referred to as the global branch). To conduct feature selection, the selector assigns a score to each region proposal to measure its relative importance and identifies the *k* regions with *k*-highest scores to compute the driving actions and explanations. That is, during the training process, the model implicitly strives to learn a metric to weigh the regional features based on the semantics and their relative contribution to the driving decisions.

Even though their model demonstrated satisfactory performance in predicting the actions together with explanations, there is still good reason to consider this attention mechanism as suboptimal with certain shortcomings. The shortcomings arise due to three reasons. First, the ablation study results in Xu et al's research depict a baseline model with only "global" branch can exhibit performance similar to the full model (which integrates the "global" and "local" features of the image). This indicates that the global features (overall information of the image) are more important and can overwhelm the contribution of regional features (the recognized objects in the scenario). This phenomenon is consistent with the notion where, in generating high-level driving actions (move forward, turn left/right, or stop), human drivers tend to use peripheral vision because only an approximate characterization of the driving scene is needed. For example, if there is an obstacle in the driving scene, the driver only needs to see it (with peripheral vision) and can quickly eliminate the erroneous action of driving towards the obstacle's direction before clearly perceiving its details such as shape and color. However, the object-induced attention as described in the research is more consistent to foveal vision because the prediction head of the model can rely "only" on those highly detailed cropped-out regions.

Secondly, this attention mechanism depends on the object detection module (region proposal network) which has been pretrained to assign greater focus on "object-containing" regions and ignoring "non-object" (background) regions. However, for driving tasks, these "non-object" regions may contain vital information such as lane markers and drivable areas. Therefore, building the model on top of the method based on object detection, could be suboptimal. Third, since the number of selected regions "*k*" is part of the model parameter, it requires considerable number of experiments to determine its value. If "*k*" is too small, this "hard" selection mechanism will inevitably create a bottleneck to restrict the information flow in the model. For example, selecting only *k* regions will restrain the model flexibility, particularly in the cases when there exist more than k pivotal regions (the regions that require the model to attend to achieve full understanding of the scenario). If "*k*" is too large, the computation resources will be wasted, and the extra information could impair the model performance because the noise level is high.

To overcome these three shortcomings, we propose a global soft attention (GSA) mechanism which imitates the peripheral vision capability of the human eye and uses the global features of the image. Overall, the model "softly" fuses the information from each region inside the image using Transformer model. The Transformer model is adopted here because a number of research studies have demonstrated that, compared to CNN, Transformer releases the constraints of generating visual features only based on local regions (Zhao et al., 2020). This makes the model capable of possessing a "broader" horizon and capturing regions that not only are much wider compared to the traditional CNN kernel but also facilitate analysis of correlations within the image. This issue is further discussed in the results section of this report.

In summary, the main contributions of this part (Part 2) of the study are threefold:

- Developed an end-to-end explainable DLCV model to generate driving actions with explanations.
- Proposed a new DL architecture with a novel visual attention mechanism using the Transformer model to achieve SOTA with significantly superior performance and lower computational cost compared to the benchmark model.
- Conducted multiple experiments in a variety of settings to evaluate the importance of information (global vs. regional) and the attention mechanism (hard vs. soft) in the high-level driving decision making process.

From the perspective of practical application, the proposed model can enhance human trust in DLCV based autonomous driving system for both AV users and AV system developers. For AV users, on the one hand, the driving decisions and explanations can be presented to the user simultaneously showing the corresponding causal relationship at the initial deployment stage of autonomous vehicles. On the other hand, such a system can perform as a "whistle" to send out instant warnings to the driver or require human intervention if there exist inconsistency between any two predictions. This could lend additional safety redundancy to the entire AV system and thereby boost human trust and acceptance of automated driving systems. For developers, such an explainable system is helpful in system debugging because it can output human-understandable outputs, identify potential flaws of the existing system, and identify directions for future improvements. Therefore, the concept of "explainable" models is beneficial to the entire AV ecosystem from perspectives of the key stakeholders, particularly, the user and the manufacturer.

CHAPTER 7. LITERATURE REVIEW

7.1 Problems associated with deep learning

In the field of perception and semantic understanding, DL is one of the mainstream technologies which has been used widely in practice. In transportation related tasks, DL has been extensively adopted in applications including infrastructure management (Hou et al., 2020; Zhuang et al., 2018), traffic prediction (Cui et al., 2019; Liu et al., 2019; Yu, Lee, et al., 2020; Zhou et al., 2021), driver behavior modeling (Xing et al., 2021), smart routing systems (Du et al., 2021), smart intersection management (Peng et al., 2021), traffic incident and duration recognition (Zhu et al., 2021). With respect to the autonomous driving task, DL models have been applied in every submodule (**Figure 10**). Specifically, the deep-learning computer-vision (DLCV)-based perception models for AV systems are widely researched and have achieved state-of-the-art (SOTA) in various contexts (Bojarski et al., 2016; Xu et al., 2017; Chen, et al., 2019). Although DL models have been deployed successfully in several real-world applications, the intrinsic drawback, low interpretability, has not been resolved. The low interpretability originates from the black box nature of computations using neural networks. The model developer can access only the input and output of the model; therefore, the potential weaknesses and drawbacks of DL models are not easily detectable and any errors in these models are difficult to diagnose. This exacerbates the problem of user distrust in automation and further hinders its deployment particularly in safety-critical tasks (Khastgir et al., 2018).

To boost user trust in automation and AI technology, several research efforts have been expended into developing “explainable” AI (XAI) systems. The key motivation and underlying notion of XAI systems is to provide human understandable explanations indicating the rationale used by the AI to make decisions (Doran et al., 2018). This idea has also been adopted in recent transportation-related research work. For example, Alwosheel et al. (2021) developed an explainable traffic demand prediction model and carried out detailed investigation on how the model provides the predictions (Alwosheel et al., 2021); Bustos et al. applied DL models and provided the interpretability analysis to demonstrate how pedestrian and vehicle safety could be enhanced (Bustos et al., 2021). In the area of AV system design, researchers have developed explainable (or even, advisable) autonomous driving models (Kim et al., 2020; Kim and Canny, 2017; Xu et al., 2020). Here, the “advisable” refers to the situation where the model is capable of processing verbal instructions from human operator and adjust further decisions based on the “advice.” These efforts have helped pave the way for enhanced user trust in DL model-driven autonomous driving. Yet still, the model performance (in terms of prediction accuracy and computational cost) can be further improved with an enhanced design of the neural network architecture that imitates human vision. This is the main motivation of Part 2 of this study.

In the subsequent sub-sections of this introductory section, we discuss two major approaches for developing DLCV based AV systems, the concept of image-attention based technologies used to imitate human vision. Then we identify the research gaps in existing research and highlight the prospective contributions of this study.

7.2 End-to-end vs. pipelined systems

In developing deep-learning computer-vision (DLCV)-based AV systems, there exist two major approaches: end-to-end and pipelined. The former seeks a direct mapping from the raw sensor inputs (including images and 3D cloud points) to the driving actions (including straight movement, left/right turn, or slowing down (Bojarski et al., 2016; Chen et al., 2019; Kim et al., 2020; Kim and Canny, 2017; Xu et al., 2017, 2020)). The latter divides the entire system into sub-systems (including vision block (Hu et al., 2020; Ku et al., 2019) and decision-generating block (Schwarting et al., 2018; Veres et al., 2011)) and addresses them independently. Theoretically, end-to-end approaches are superior to pipeline approaches because the vision block can be trained to be goal-induced, meaning, it becomes capable of paying more attention to the visual information that is necessary for the ultimate goal. However, the end-to-end approach is more complicated and needs deeper networks and larger datasets for training.

In addition, because the model is trained from end to end, there are no intermediate results for diagnostic purposes, and this exacerbates the black box nature of the process. The pipeline approach, on the other hand, is considered more tangible because it can output intermediate results for purposes of human inspection and validation (i.e., object detection bounding boxes). However, the pipeline approach is often sub-optimal because training the sub-modules separately may cause one to lose track of the ultimate goal. Such segregated training can lead to a misallocation of computation resources due to the detection of irrelevant objects or the erroneous neglect of important objects in the driving environment. For example, the detection of objects located beyond the roadway sidewalk will not be beneficial to AVs. However, failure to detect traffic signal colors could be catastrophic. Another limitation of the pipeline approach is that it requires an explicit definition in the manner of cooperation of the two sub-modules; if the cooperation protocol is ill-defined, the overall performance of the entire model can be jeopardized even if the individual sub-modules exhibit satisfactory performance.

A natural way of integrating the benefits from both end-to-end system and pipelined system is to add intermediate output heads that can generate human understandable results while training the entire system end-to-end for the final goal. For example, by adding "explanation head" to AV systems, the model can simultaneously output both driving actions and corresponding explanations. The explanations provide an opportunity to ascertain whether the correct causal relationships (between the driving environment from the input images and actions) have been learned. They also serve as extra labels (extra loss function) to facilitate the entire training process.

Furthermore, from an application perspective, the joint prediction of explanations and decisions could yield additional redundancy to the entire system compared to models that output decisions only. This is because the causal relationship between explanations and driving decisions can be easily stored as simple rules (for example, perceiving a yellow light on the traffic signal should result in a slow down or stop decision). If the model fails to predict the consistent decisions with explanations, warnings could be sent immediately to the human operator for the requisite intervention. As a result, this extra setting can help enhance the model interpretability, boost confidence in the model, and eventually incentivize AV system manufacturers to adopt the model.

7.3 Imitation of human vision using image attention

In recent years, despite the fact that DL models have shown great promise in image processing, human vision remains an incontrovertible benchmark. This is because the human eye possesses a multi-resolution structure, namely, peripheral, and foveal vision (Xia et al., 2020), and a proper “attention” mechanism to reach a balance in efficiency and recognition accuracy. Peripheral vision is blurry (low resolution) but requires only a brief time for processing and has a larger field of vision. Foveal vision, on the other hand, is clear (high resolution) but requires longer processing time and only has a limited vision field. The combination of these two structures guarantees the efficiency because it enables the human to “attend” only to the salient and important regions with foveal vision while the overall information of the scene can be necessarily understood with only peripheral vision. In AV perception, this is also important because the computation cost needs to be minimized as much as possible so that perception and decisions can take place with minimal delay.

In efforts to imitate human vision, visual attention has gradually evolved into a research area of great interest to AV researchers. Recently, a state-of-the-art paper on end-to-end AV systems proposed an object-induced attention mechanism to generate driving decisions with “salient” objects in the scene (Xu et al., 2020). More recently, it has been demonstrated theoretically that a self-attention-based model named Transformer (Zhao et al., 2020) exhibits similar characteristics and performance as the convolutional neural network (CNN) and is also capable of capturing long-range correlations within an image. This is the key inspiration and motivation for the present study. In subsequent sections of this report, we demonstrate the efficacy of the Transformer based model in generating driving actions and explanations for autonomous driving systems.

CHAPTER 8. METHODS

8.1 Prelude

The proposed model, as well as all the baseline models introduced in the experiment settings section share the same structure containing three blocks, namely, Feature Extractor, Attention Module and Decision/Reason Generation (**Figure 11).** As its name suggests, the feature extractor is used to generate the low-level feature embeddings from the raw image, which contains image preprocessing (i.e., normalization, reshape) and a pretrained backbone CNN model. On top of Feature Extractor, Attention block solves the problem of information fusion and feature selection. For the proposed model, the attention is achieved by correlating the features from each spatial location using "self-attention" mechanism to compute the attention weights and using these attention weights to either "amplify" or "filter out" the features. The final block takes in the attended feature map and conducts two separate multi-class classification tasks for generating both driving decisions and corresponding explanations. The entire model integrates the three blocks and is trained end-to-end with the aggregated loss function for both predictions. The rest of this section of the report explains each block in detail.

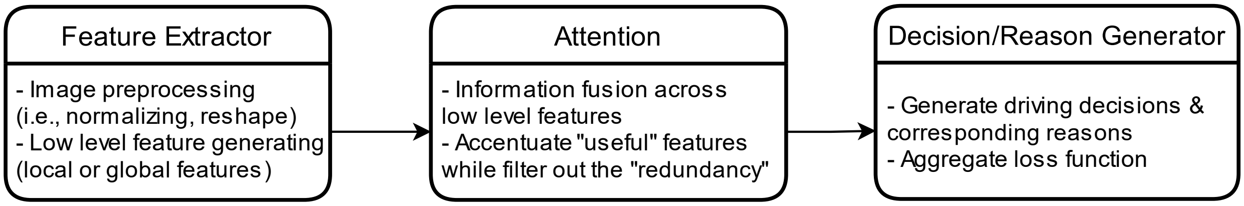

Feature Extractor
- Image preprocessing (i.e., normalizing, reshape)
- Low level feature generating (local or global features)
Attention
- Information fusion across low level features
- Accentuate "useful" features while filter out the "redundancy"
Decision/Reason Generator
- Generate driving decisions & corresponding reasons
- Aggregate loss function

Figure 11. Architecture of the proposed model

8.2 Feature Extractor

The proposed model uses the Feature Extractor block as the global feature extractor to acquire overall features of the image instead of specifically focusing on object-containing regions. To this end, as shown in **Figure 12**, the module first preprocesses the image (resize and normalization) and then computes the visual features using pre-trained Convolutional Neural Network (CNN) models. In this work, we experimented with two classic backbone CNN models, namely, Resnet50 (He et al., 2016) or Mobilenet_v2 (Sandler et al., 2018). The evaluation and selection of these two models was based on the tradeoff between computation cost and accuracy. Mobilnet_v2 is designed to run on mobile devices, which has much higher computational speed due to its use of a smaller number of parameters, while the Resnet model are much deeper in structure and is believed to represent the state of the art in image feature generation. In addition, since the target for this project is to examine the performance of the upper stream architecture (attention module) and the feature importance, the pretrained Feature Extractor module is frozen in both training and testing time.

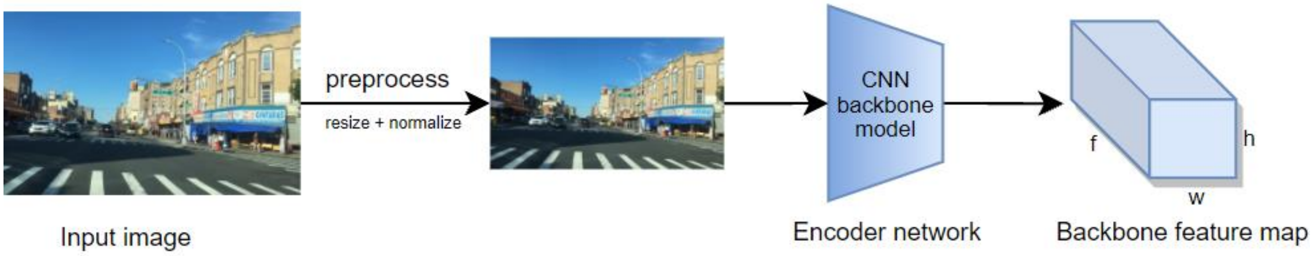


Figure 12. Global feature extractor

8.3 Transformer

The feature map obtained from the Feature Extractor contains the information of the entire image, which is then fed to the Transformer to perform “global soft attention” (**Figure 13**).

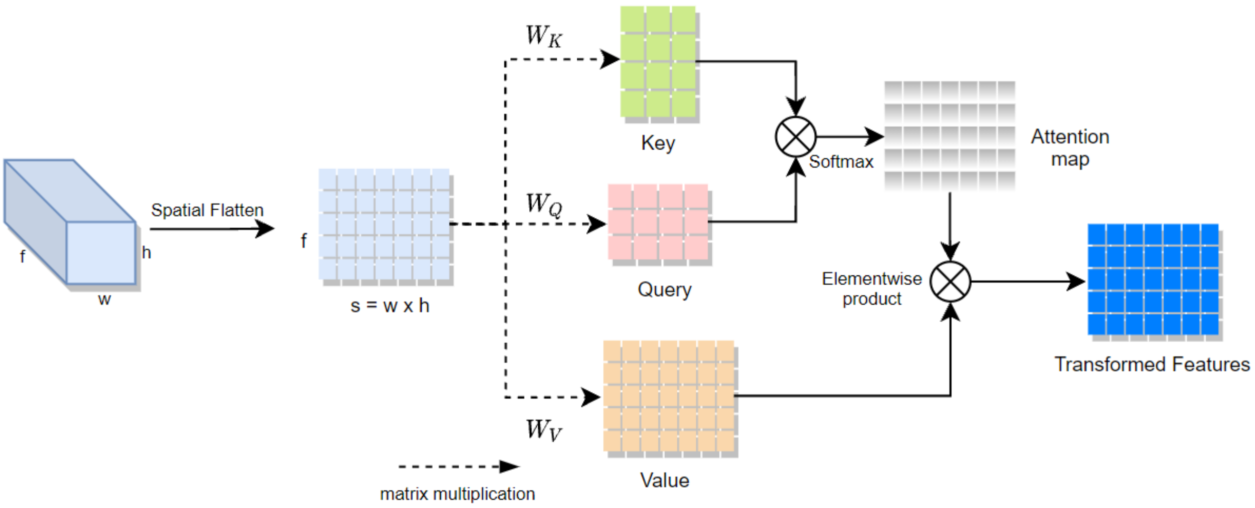


Figure 13. Self-attention (SA) layer (single head)

More specifically, the output from the previous block is a 3D tensor of shape ($h \times w \times f$) where f is the feature dimension of each spatial location. The variables h and w represent the height and width, respectively, of the feature map. Then, the two spatial dimensions (h and w) are flattened, and the output 2D feature map $X \in \mathbb{R}^{s \times f}$ is treated as a sequence of input with a sequence length of $s = h \times w$.

The basic building block of the Transformer model is referred to as the multi-head self-attention (MHSA) layer. MHSA represents a parallel computation of self-attention (SA) which measures the “similarity” between two inputs using their dot product. Initially, self-attention establishes three representations: key, query, and value (K, Q, and V) with three distinct linear layers: $K = XW_{\mathrm{K}}$, $Q = XW_{\mathrm{Q}}$, and $V = X^{\mathrm{T}}W_{\mathrm{V}}$, with $W_{\mathrm{K}}, W_{\mathrm{Q}}, W_{\mathrm{V}}$ as their respective weights, $K, Q \in \mathbb{R}^{s \times d_k}$ and ,$V \in \mathbb{R}^{s \times d_v}$).

It is then possible to generate a matrix of the attention score (a) (which measures degree of correlation between regions) by determining the dot product of the query and key, and then softmax normalization as shown in Equation (1):

$$a = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \in \mathbb{R}^{s \times s} \qquad (1)$$

Where d_k is the dimension of K and Q. This attention mechanism enables the output embedding for each spatial location contains not only the information of the spatial location itself but also valuable information from other spatial locations. The attention scores serve as the fusion weights for generating the attended feature maps. The output of the SA is the multiplication of the value (V) and the attention score, and this completes the computation for single head (**Equation (2)**). Then the MHSA layer is simply the parallel version of SA which simultaneously computes multiple SA by concatenating all the heads (**Equation (3)**).

$$head = Attention(K, Q, V) = aV \qquad (2)$$

$$MHSA(X) = concat(head_1 , \dots , head_h)\, W_{out} \in \mathbb{R}^{s \times d_{out}} \qquad (3)$$

Where $W_{out} \in \mathbb{R}^{d_v \times d_{out}}$ represent the weights for the final output linear layer which is applied for fusing the results from multiple heads and h is the number of heads computed in parallel. Compared to single head, using MHSA enables each head to simultaneously focus on different tasks and can attend to regions with different ranges. This manipulation can enhance the model's flexibility and generalization power. The final output from the Transformer (MHSA layer) maintains the same spatial dimension as the input feature map X. However, each spatial location contains the "fused" information from this location itself and other regions based on the automatically computed correlation. For demonstration purposes, we use a single MHSA layer in this report.

8.4 Decision / Reason Generator

As shown in **Figure 14**, the Decision/Reason Generator block is a standard multitask classifier containing two branches for generating driving decisions and explanations, respectively. It takes the output feature map from MHSA block as input and performs two separate classifications.

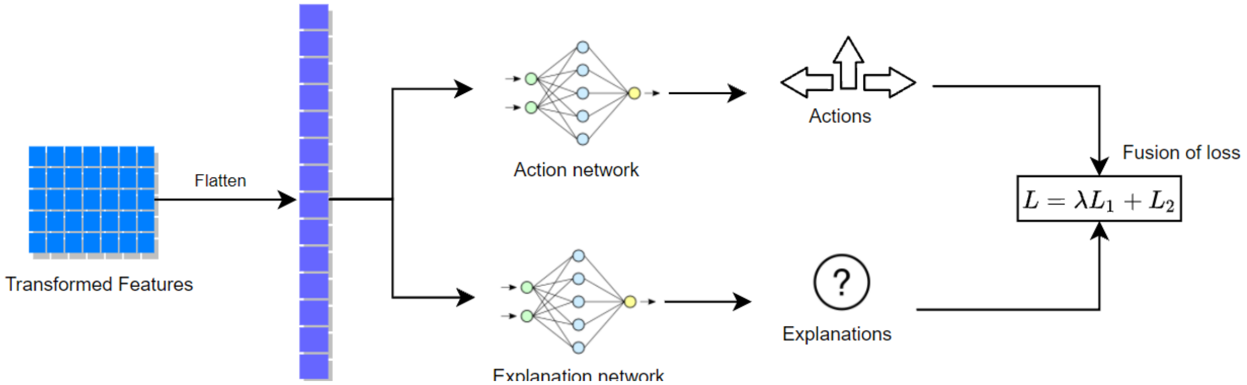


Figure 14. Process of the Decision/Reason Generator

This is achieved using two independent neural networks: the action network and the explanation network. The former has four output classes that represent each driving decisions (going straight, stop/slow down, turn left, turn right) while the latter has twenty-one (21) output classes for the corresponding explanations. Regarding the detailed architecture, we use the same structure with fully connected (FC) layers for both the networks ($Dense(128) + Dense(128) + Dense(\#\ of\ outputs)$). Finally, the model is trained end-to-end, following the classic multitask learning manner that aggregates the two losses (driving action loss L_A and explanation loss L_E). This setting requires the model to simultaneously learn to generate the decision and explanation, thus the corresponding causal relationship between these two losses (Equations 4 and 5) can be learned implicitly.

$$L_A = \sum_i^4 L(\widehat{A_\iota}, A_i)\,;\ \ L_E = \sum_i^{21} L(\widehat{E_\iota}, E_i) \qquad (4)$$

$$L = \lambda L_A + L_E \qquad (5)$$

Where λ is the weight parameter for tuning the tradeoff between the two losses. From the experiment in Xu's study (Xu et al., 2020), when λ=1, the model yields the best performance in terms of both action and explanation prediction. In the present study, we adopt this result from the Xu study, and therefore, use a $\lambda = 1$ value of 1, recognizing that the explanation and driving decision should be equally weighted.

The "4" and "21" in **Equation (4)** refer to the total number of actions and explanations, respectively, in the dataset, which is explained in detail subsequently in Chapter 11 (Experiment Setting) of this report.

$L(\cdot,\cdot)$ represents the binary cross entropy loss defined in **Equation (6)**, where $y\ and\ \hat{y}$ represent the true label and the model prediction, respectively:

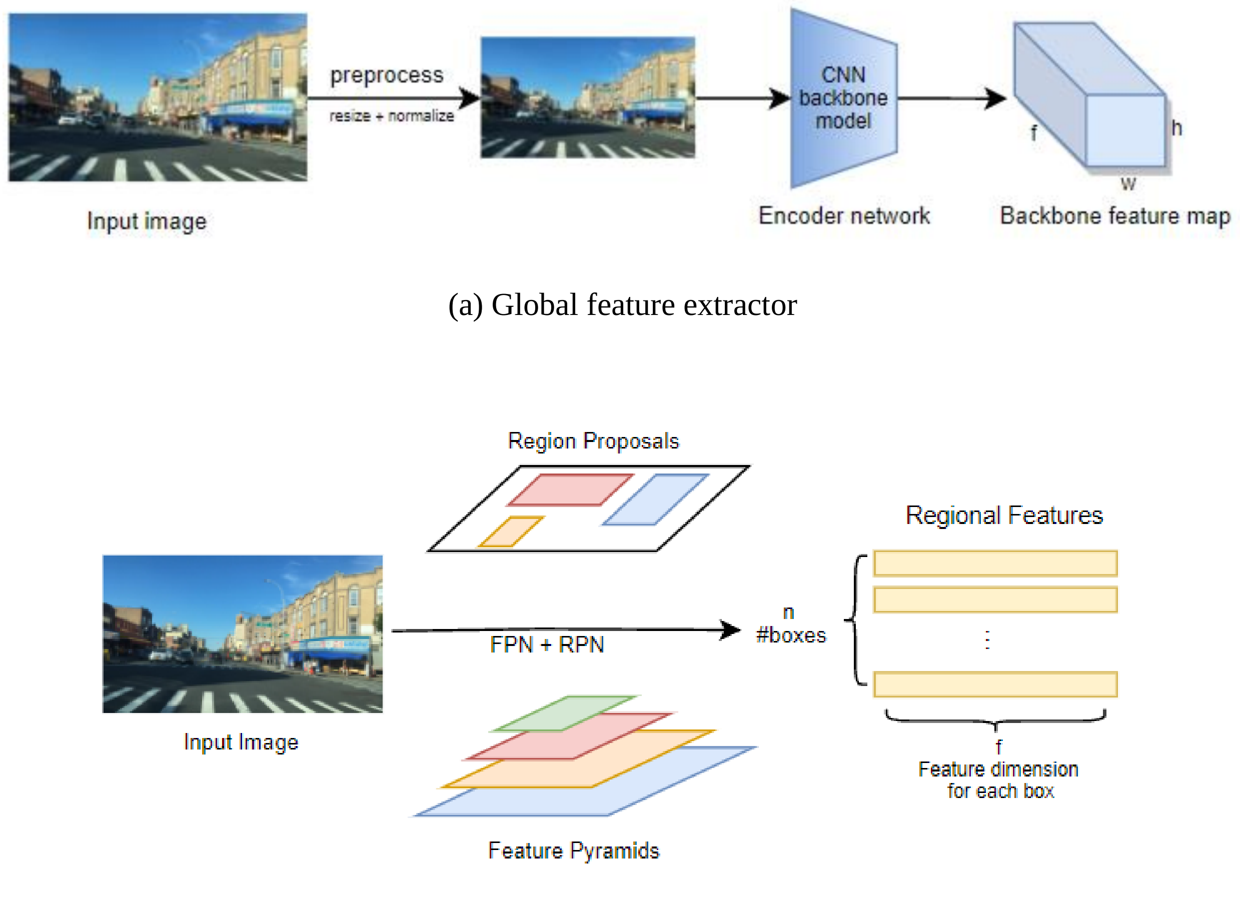
$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\big(p(\widehat{y_\iota})\big) + (1 - y_i)\log\big(1 - p(\widehat{y_\iota})\big) \qquad (6)$$

CHAPTER 9. EXPERIMENTAL SETTINGS

9.1 Dataset

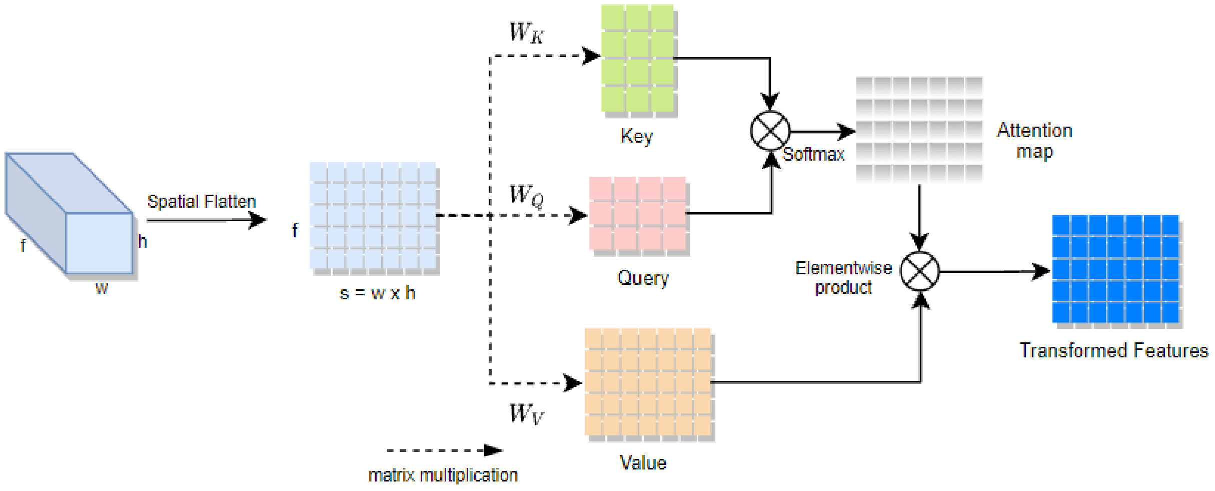
Figure 15 presents the structures for the baseline models. The study trained and evaluated the models described above, using Xu et al. (2020)'s BDD Object Induced Actions (BDD-OIA) dataset. The BDD-OIA dataset extended the original BDD-100K dataset (Yu, Chen, et al., 2020) by labeling each frame individually with driving actions and explanations. The actions refer to high-level feasible driving maneuvers that can be undertaken by the driver at any specific time step: move forward, stop/slow down, turn left and turn right. The explanations are associated with the actions and are summarized into twenty-one (21) classes. **Figure 16** illustrates the example image and labels. **Table 3** presents the labels for actions and explanations. The model is developed with a training set of 16,082 images, a validation set of 2,270 images, and a test set of 4,572 images.

Figure 15 presents the model components for the baseline models.

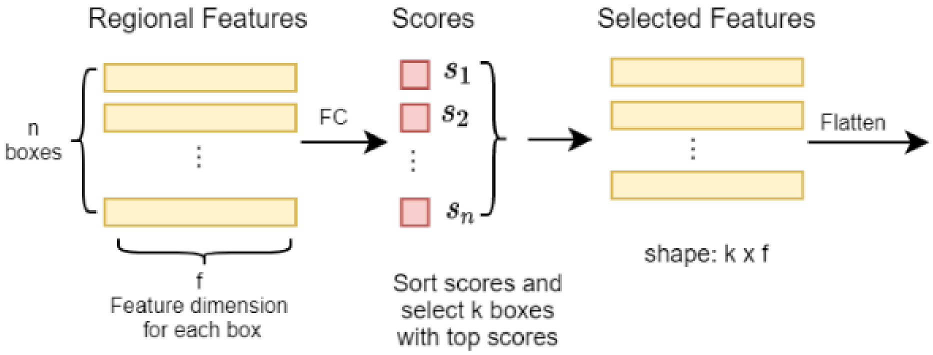


(a) Global feature extractor

(b) Regional feature extractor



(c) Soft attention



(d) Hard attention (only for regional features)

Figure 15. Structures for the baseline models

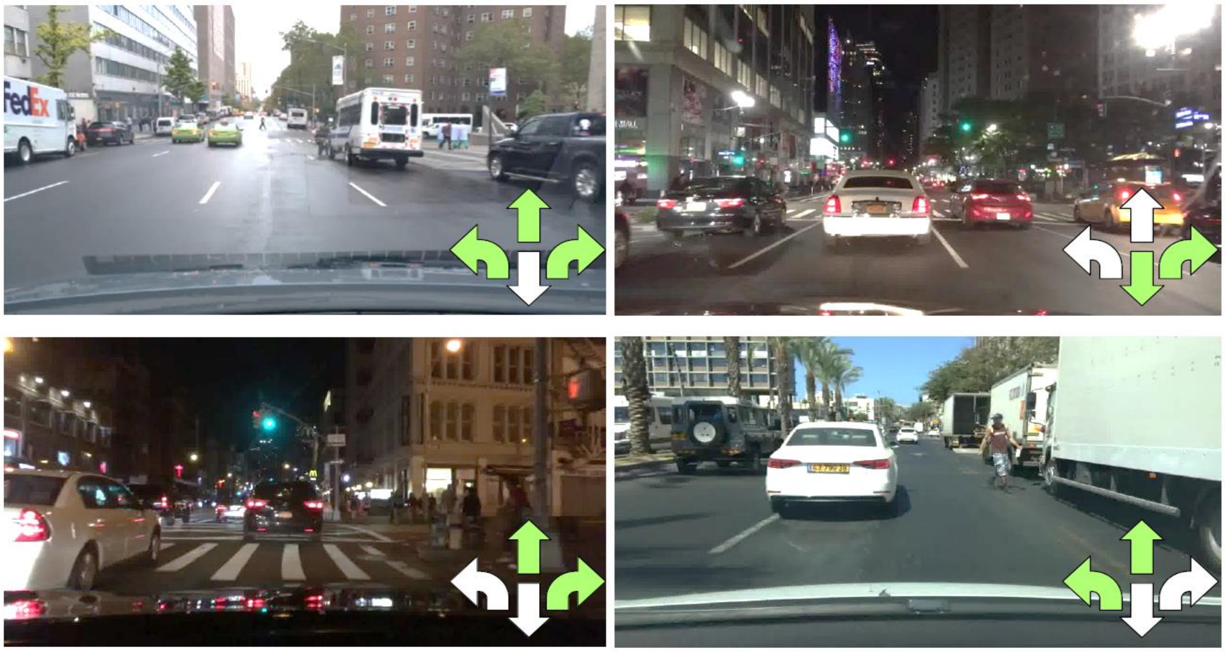


Figure 16. Examples of ground-truth images and decisions from the BDD-OIA dataset (Xu et al., 2020)

Table 3. Actions With Explanations in BDD-OIA

Actions	Explanations
Move forward	Traffic light is green
	Follow traffic
	Road is clear
Stop/Slow down	Traffic light is red
	Traffic sign
	Obstacle: car
	Obstacle: person
	Obstacle: rider
	Obstacle: others
Turn left	No lane on the left
	Obstacles on the left lane
	Solid line on the left
	On the left-turn lane
	Traffic light allows
	Front car turning left
Turn right	No lane on the right
	Obstacles on the right lane
	Solid line on the right
	On the right-turn lane
	Traffic light allows
	Front car turning right

9.2 Baseline Models and Setups

As mentioned in the introduction section of Part 2 of this report, two key technical motivations for this study are to evaluate the relative importance between global and regional information and test different attention mechanisms (hard vs. soft attention). Therefore, we compare our global soft attention (GSA) model with several baseline models including regional hard-attention (RHA), regional soft-attention (RSA), and global no-attention (GNA) model.

- The RHA model is similar to the object-induced attention model proposed in a benchmark study (Xu et al., 2020) albeit with the local branch only (we did not reimplement the model in the benchmark study but compared its results with ours in the result section). This model utilizes Faster RCNN with Feature Pyramid Network (FPN) as feature extractor as shown in **Figure 15 (b)**, followed by a Regional Hard Attention module which utilizes a fully connected (FC) layer to compute a score for each region proposal and select only top-k objects (based on the scores) for generating actions (as shown in **Figure 15 (d)**). In this research, we trained 2 models with *k* = 5 and *k* = 10. We keep the local branch only to test the importance of regional information for the overall driving decision and explanation generation.
- The RSA model uses the same soft attention mechanism (Transformer) as the proposed GSA model (Multihead Self Attention block in **Figure 15 (c)**), but the attention is conducted over the region proposals instead of the global features. It uses the same FasterRCNN (FPN) as (Xu et al., 2020) to generate the regional features (acquired from **Figure 15 (b)**). This model serves to compare the performance between soft attention and hard attention. In addition, we trained two RSA models with 5 and 8 heads.
- The GNA model serves as an ablation study to our GSA model. It uses the same global features (generated from Resnet/Mobilenet backbone with **Figure 15 (a)** structure) as GSA. However, a vanilla fully-connected network (FCN) having parameters similar to those of the MHSA block replaces the Transformer structure (MHSA) block.

CHAPTER 10. RESULTS

10.1 Introduction

The training is conducted on one NVIDIA Quadro RTX-6000 GPU, which has 24G RAM. All the models are trained using batch size $b = 10$, the stochastic gradient descent (SGD) method with an initial learning rate $\alpha = 0.001$, and a learning rate decay of 10^{-4}. All the models are trained for 40 epochs (64,080 batches). **Figure 17** presents the corresponding training curves (training loss vs. number of steps) for our proposed GSA model and all the baselines mentioned above. From the training curve, it can be observed clearly that the two proposed GSA models converge much faster and can achieve much lower training loss compared to the baselines.

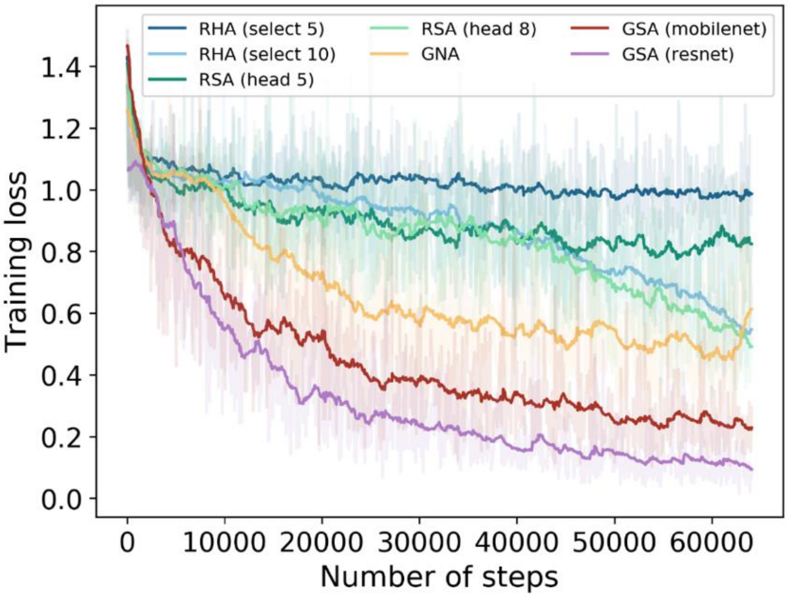


Figure 17. Training curve for all 7 models (the proposed GSA and the baselines)

10.2 Quantitative Evaluation

We performed the same prediction task (actions and explanations prediction) as the benchmark model presented in previous research (Xu et al., 2020), and we evaluated the proposed model using evaluation metrics similar to those in recent literature. For both decision and explanations, two versions of the F-1 score were used: the overall F1 score, $F1_{all}$, (the F1 score calculated over all the predictions), and the mean in-class F1 score, $mF1$, a metric typically used where the data are unbalanced.

A model with a high F1 score indicates that the model has higher recall and higher precision. **Equation (7)** presents the calculation of the $F1_{all}$ score:

$$F1_{all} = \frac{1}{|A|} \sum_{j=1}^{|A|} F1(\widehat{A_j}, A_j) \tag{7}$$

Where A_j = true label (representing an explanation or action), $|A|$ = total number of predictions, $\widehat{A_j}$ = predicted value.

In the dataset, there exist a greater number of instances associated with the "going-straight" action compared to the "turn-left" action; in other words, the dataset is unbalanced. For this reason, **Equation (8)** was used to calculate the F1 score for each predicted class, and the $mF1$ value was calculated as the mean of all the F-1 scores:

$$mF1 = \frac{1}{C} \sum_{j=1}^{C} \sum_{i=1}^{n} F1(\widehat{A_i^j}, A_i^j) \tag{8}$$

C is the number of predicted classes (4 for actions, 21 for explanations), n is the total number of points in the test dataset. The detailed performance in terms of actions and explanations prediction are listed in **Table 4.**

Table 4. Model Performance & Complexity of Proposed Model* and the Baselines

Attention mechanism	Model	Decision mF1	Decision $F1_{all}$	Explanation mF1	Explanation $F1_{all}$	# of trainable parameters	Training time (40 epochs)
Regional Attention	Xu et al., 2020)	0.718	**0.734**	0.208	0.422	-	-
	RHA (5 obj)	0.572	0.494	0.482	0.047	11.04M	10h, 28min
	RHA (10 obj)	0.565	0.495	0.499	0.123	21.53M	10h, 30min
	RSA (5 heads)	0.595	0.476	0.506	0.127	21.22M	10h, 32min
	RSA (8 heads)	0.608	0.542	0.554	0.330	20.75M	10h, 42min
Global no attention	GNA (resnet)	0.706	0.660	0.561	0.352	26.10M	3h, 10min
***Global soft attention**	**GSA (resnet)**	**0.750**	0.729	**0.644**	**0.525**	**24.08M**	**3h, 15min**
	GSA (mobilenet)	**0.746**	0.718	**0.642**	**0.531**	**2.61M**	**2h, 53min**

From **Table 4**, the following conclusions can be made:

1) The proposed global soft attention (GSA) models outperform all the baselines by a significant margin, particularly regarding explanation prediction.

2) Global features are more useful compared to regional features, even where the vanilla model (GNA) is used without any special attention mechanism.

3) Soft attention is superior to hard attention even in cases where only regional information is available.
4) With regard to the feature extractor, using Mobilenet_v2 has comparable predictive performance compared to Resnet50 but saves a significant amount of training time.
5) Increasing the number of heads can enhance the performance of soft attention models.
6) The superiority of the model is: GSA > GNA > RSA > RNA. This also matches the training loss curve (**Figure 17)** in terms of the final loss and convergence rate.

Apart from the prediction performance, another important aspect of evaluating the model is the computation complexity. We document the number of trainable parameters and the total training time for 40 epochs on the training dataset in the final 2 columns of **Table 4**. Compared to the regional models (RHA and RSA) which takes more than 10 hours of training, the global models (GNA, GSA) require only 1/3 of computation resources. The combined results from both loss curve (**Figure 17**), the performance and computational cost in **Table 4** indicate that even with fewer parameters and much shorter training time, the proposed GSA model achieves lower training loss and yields higher performance. This indicates that global attention superior to regional attention in predicting driving related actions and explanations.

10.3 Qualitative Evaluation

Figure 18 presents the examples of the predictions (generated from GSA-Mobilenet) on the test set. Regarding the action prediction, “G” stands for ground truth and “P” for model prediction. Regarding explanations, green color indicates true positive, yellow indicates false positive, and red indicates false negative. The predictions are made for eight images chosen randomly from the test dataset. By inspecting the scenarios and the predictions, the model predictions can correctly predict the driving decisions with high accuracy while generating the corresponding explanations. The generated explanations match the decisions and driving scenarios in most cases. It is worth noting that from the first image, the model predicts an extra explanation indicating the traffic light is red. Even though this is not included in the label, by inspecting the image, we can see the model is in fact making the correct prediction as it does **exist** the red traffic light in the right of the image. We also notice that there exists some inconsistency in the labels of the dataset, for example, the third image labels the vehicle with turn right decision even if it has the explanation indicating the vehicle is on the left turning lane. The model does not predict this explanation because it does not match the correct causal relationship. Furthermore, by inspecting more examples, we notice this inconsistency does not impair the performance of the model and does not impact the comparative analysis between the proposed model with all other baselines.

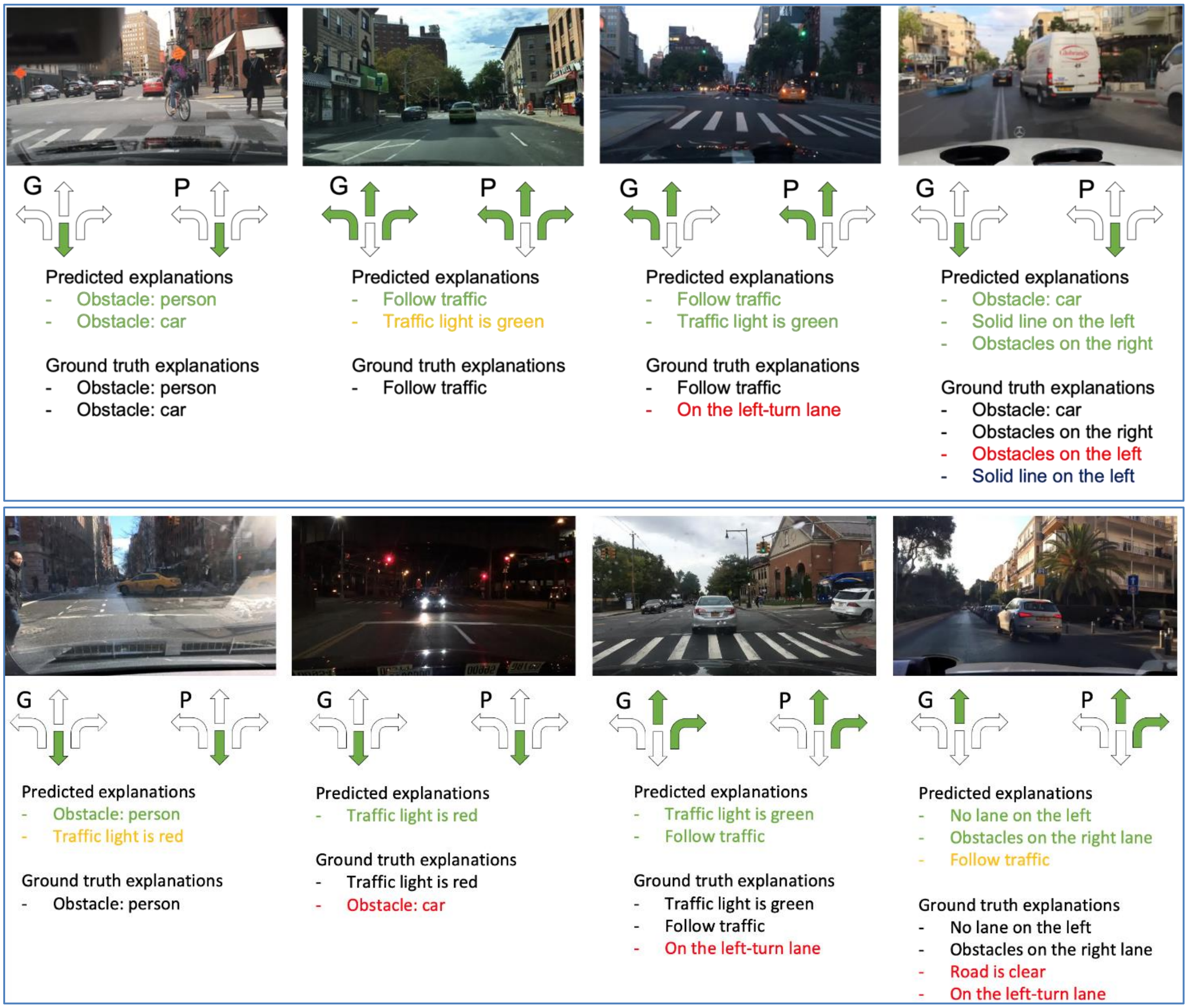


Figure 18. Example predictions. (Regarding action prediction, "G" stands for ground truth and "P" for model prediction. Regarding explanations, green color indicates true positive, yellow indicates false positive, and red indicates false negative)

10.4 Discussion

10.4.1 Attention mechanism: Soft > Hard

The soft attention (Transformer in this work) is superior to the hard attention (score-based selection) because the former can fuse individual pieces of information in the image based on their individual contributions to the ultimate driving goal (maneuver) rather than simply picking the more important regions. The latter inevitably creates a "bottleneck" to the information flow path and therefore leads to non-consideration of some information that could be useful to the driving actions. Furthermore, because the regional hard attention "crops" the regions, the correlation between the objects as well as the "relativity" among the image are eliminated. For example, after "selection" operation (Figure 6 (d) in (Xu et al., 2020)), obstacles located farther away could have the same representation of the obstacles located close by, then the model cannot know which one is closer. This will increase the ambiguity to the downstream decision/explanation generation block. On the other hand, soft attention can learn the correlation and compute a "soft" fusion of all the features using the attention map.

10.4.2 Feature importance: Global > Local

The global features are superior to regional features due to the inherent nature of driving decisions. For generating high-level actions (for example, move forward, stop/slow), the acquisition of an overall characterization of the roadway scene is more essential compared to the recognition of every single object and computation of their bounding boxes. Therefore, even the GNA baseline can yield superior performance compared to regional attention models built on top of object detection models. In addition, despite the GSA models are not equipped specifically with object detection block in the architecture, the explanations predicted still contains the information of local regions. For example (**Figure 18, column 1**), the model can still identify the red traffic lights, persons, and vehicles obstacles even if these objects occupy only a small proportion of the image. Therefore, based on our experiment results, it is still safe to conclude the global attention (Transformer) mechanism will not neglect the local regions.

10.4.3 Transformer is useful in feature fusion

The Transformer based models (the two GSAs) outperform the GNA because their MHSA structure can capture long-range correlations within an image. Compared to classic CNN based methods which can capture only the local region correlations due to the fixed size of convolution kernels in each layer, the Transformer-based models enable information fusion over the entire image. This long-range correlation is typically crucial for driving decisions because there exists a "relativity" correlation within the image. For example, "left" is relative to the "right;" therefore, to generate the decision of "turn left," the model needs to understand which part of the image depicts the "left region." Since the cameras are not always facing the same direction as the movement direction of the vehicle, the ratio of "left region" to the entire image keeps changing. Therefore, the model must understand "left" and "right" relatively from the scene context, which can only be achieved with Transformer based model by simultaneously attending to multiple regions. This entire mechanism is analogous to the peripheral vision of the human eye as human drivers generating driving actions (quickly looking at multiple regions and then making driving decisions instantaneously without clearly seeing each individual object in the region) (Wolfe et al., 2017) (Rosenholtz, 2016).

10.4.4 Causal relationship is correctly learned

One of the most salient problems for the existing end-to-end DLCV based autonomous driving system is that whether the model has truly "understood" the driving scenario remains uncovered to human even if the prediction of driving decisions is correct. In our settings, we "force" the model to explicitly understand the driving environment by injecting a second loss function (through joint prediction of explanation) as these explanations are the human understandable descriptions to the driving scenario. From **Figure 18**, it is clearly shown that the model can correctly identify most of the explanations associated with the driving decisions. This indicates that the model can capture the correct causal relationship between the driving decisions and the driving environment, and this capability is useful to enhance the user trust in the automated system.

10.4.5 Potential to identify the limitations of the existing model

From the last two columns in **Figure 18**, it can be inferred that a weakness of the model is its inability to predict the explanations pertaining to the lane location of the vehicle (the model fails to identify the vehicle is on the left-turning lane in both cases). This problem may be due to the lack of training data associated with this explanation since the original BDD-OIA dataset is unbalanced with very few examples indicating that the vehicle should make a turn since it is on the corresponding lane. This can be mitigated by further enriching the dataset by collecting data instances regarding these sparse cases and incrementally training the existing model. Therefore, the proposed model can potentially identify not only its limitations but also the direction of its improvement in a human-understandable manner. This property does not exist in most other DL models in the existing literature.

PART III

Concluding Remarks, Performance Indicators, Study Outcomes and Outputs

CHAPTER 11. CONCLUDING REMARKS

11.1 Part I of the Study

11.1.1 Summary and Conclusions

User trust has been identified as a critical issue that is pivotal to the success of autonomous vehicle (AV) operations where artificial intelligence (AI) is widely adopted. For such integrated AI-based driving systems, one promising way of building user trust is through the concept of explainable artificial intelligence (XAI) which requires the AI system to provide the user with the explanations behind each decision it makes. Motivated by both the need to enhance user trust and the promise of novel XAI technology in addressing such need, this study seeks to enhance trustworthiness in autonomous driving systems through the development of explainable Deep Learning (DL) models. First, the study casts the decision-making process of the AV system not as a classification task (which is the traditional process) but rather as an image-based language generation (image captioning) task. As such, the proposed approach makes driving decisions by first generating textual descriptions of the driving scenarios, which serve as explanations that humans can understand. To this end, a novel multi-modal DL architecture is proposed to jointly model the correlation between an image (driving scenario) and language (descriptions). It adopts a fully Transformer-based structure and therefore has the potential to perform global attention and imitate effectively, the learning processes of human drivers. The results suggest that the proposed model can and does generate legal and meaningful sentences to describe a given driving scenario, and subsequently to correctly generate appropriate driving decisions in autonomous vehicles (AVs). It is also observed that the proposed model significantly outperforms multiple baseline models in terms of generating both explanations and driving actions. From the end user's perspective, the proposed model can be beneficial in enhancing user trust because it provides the rationale behind an AV's actions. From the AV developer's perspective, the explanations from this explainable system could serve as a "debugging" tool to detect potential weaknesses in the existing system and identify specific directions for improvement.

Summing up, the study developed and tested an explainable DL model to mitigate the low interpretability problem associated with deep neural networks. The test results indicate superior performance of the proposed model against all the baselines in terms of both reason generation and action predictions. Also, the qualitative evaluation indicates that the proposed model can exploit all the potential reasons and thus have a holistic understanding of the driving scenes. In addition, the attention maps can further enhance the model explainability by associating each word in the reason sentence with the image regions, which enables the developers of AV systems to identify the potential drawbacks and future direction of improvement of the existing system.

The study contributions to existing literature are (i) Formulation of the traditional end-to-end autonomous driving decision process as an image-captioning task with language-induced visual attention to guarantee the explainability in DL models, (b) Development of a fully Transformer based model to generate verbal descriptions and driving actions for autonomous driving, (c) Demonstration of the efficacy of the proposed model and ascertaining that it possesses superior performance over multiple baseline models in terms of explanation and driving action predictions.

11.1.2 Study Limitations and Suggestions for Future Research

Moving forward, one potential intermediate step is to embark on the task of relabeling and rebalancing the dataset, expanding it to encompass a wider range of actions and their corresponding scenario descriptions. In pursuing this direction, one could lay a robust foundation for a comprehensive evaluation of the effectiveness of models similar to that developed in this study. Moreover, it will facilitate the exploration of potential future extensions and advancements of the model. Another potential improvement of the proposed approach is to consider the "temporal" information and generate explanations as well as decisions by fusing multiple frames of historical driving scenes. This is analogous to video captioning. By leveraging the temporal information, the ego vehicle could be made to have the capability of predicting the evolving nature of driving scenarios. This could further enable the AV to generate "optimal" maneuvers such as proactive decisions while considering the dynamic moving pattern of the surrounding environment.

Explainable models for autonomous driving systems are expected to have significant applications during the deployment of AVs. Knowing the rationale behind an AV's actions will be consequential in situations that require thorough liability analysis. For example, in the event of a collision between an AV and a human-driven vehicle, the liable parties can be easily identified, and the cause of the collision (for example, human error, improper detection, etc.) could be attributed more reliably. Finally, greater explainability of autonomous driving decisions can enhance the reliability of AVs and in turn, improve public perception and user trust in AVs.

11.2 Part II of the Study

11.2.1 Summary and Conclusions

In Part 2 of this study, we propose a novel architecture to generate driving actions as well as explanations based on images, to facilitate autonomous driving. The objective is to mitigate the low interpretability nature of deep learning-based computer vision models and, to enhance user trust of autonomous driving systems. The proposed architecture utilizes the Transformer model (that is, the Multi-head Self Attention module) to imitate the peripheral vision of humans. The results from the experiments demonstrate that the proposed model outperforms all the baseline models in terms of prediction accuracy and training time.

In the process of addressing these broad objectives, the study evaluated the relative importance of the global features and the local features as well as the appropriate visual attention mechanism for feature engineering. The experiment results suggest that based on the BDD-OIA dataset used in the study, (a) global features are more important than regional features, and (b) soft attention (Transformer) is superior to hard attention (region selection). These results are consistent with intuition: for the high-level driving decisions (go straight, slow down/stop, etc.) the peripheral vision (emulated by the global attention) that can achieve long range correlation and can quickly grasp the overall semantics in the driving environment is found to be more essential compared to foveal vision which specifically focuses on a relatively small region. Therefore, in the development of actual vision-based autonomous driving systems, it is recommended that the designers assign higher priority to the overall information and create the appropriate attention mechanism to enhance the global features.

In the application contexts of situational awareness and driver assistance, the proposed model can perform as a driving alarm system for both human-driven vehicles and autonomous vehicles because it is capable of quickly understanding/characterizing the environment and identifying any infeasible driving actions. In addition, the extra explanation head of the proposed model provides an extra channel for sanity checks to guarantee that the model learns the ideal causal relationships. This provision is critical in the development of autonomous systems.

11.2.2 Study Limitations and Suggestions for Future Research

Moving forward to the future work, the proposed model can be further improved by incorporating and fusing other sources (sensor types) such as LiDAR point clouds and information from vehicle-to-vehicle (V2V) connectivity. In this context, the camera is a powerful sensor that can capture several semantics in the driving environment but is vulnerable to occlusion, poor illumination, reflection, and other environmental adversities. V2V connectivity can address these limitations, as it provides more straightforward information on the speed, speed change rate, and location of neighboring vehicles, and this information can be used directly in the ego vehicle's motion planning module without perception requirements. The fusion of information from multiple sources imparts to the autonomous driving system, the virtues of information redundancy, resilience to sensor misfunction, and an added layer of system reliability and occupant safety.

CHAPTER 12. SYNOPSIS OF PERFORMANCE INDICATORS

12.1 USDOT performance indicators I

Three (3) transportation-related courses were offered annually during the study period that was taught by the PI: Smart Mobility, an optional undergraduate course; Civil Engineering Systems, a mandatory undergraduate course; and an independent-study graduate course related to vehicle automation that was inspired and directly associated with this CCAT research. One (1) graduate student and one (1) post-doctoral researcher (subsequently designated a Visiting Assistant Professor) participated in the research project during the study period. One (1) transportation-related advanced degree program (a doctoral program) utilized the CCAT grant funds from this research project, during the study period to support the graduate student.

12.2 USDOT performance indicators II

Research Performance Indicators: 2 journal articles, and 4 conference presentations were produced from this project. The research from this advanced research project was disseminated to over 120 people from industry, government, and academia, through the 4 conference presentations. These include the TRB 102nd Annual Meeting in Washington, D.C., on January 2023; the 1st ICON Student Research Conference, West Lafayette, IN, on February 2023; the 4th ASCE International Conference on Transportation and Development (ICTD) in Austin, TX, in June 2023; the 5th Bridging Transportation Researchers (BTR5), online conference, in August 2023.

Other related research projects were funded by sources other than UTC and these were designated as matching fund sources. At the time of writing, the researchers are still working on developing a specific product (modern technologies), procedures/policies, and standards/design practices based on the results of this research project.

Leadership Development Performance Indicators: This research project generated 3 academic engagements and 2 industry engagements. The PI's held positions in 2 national organizations that address issues related to this research project.

Education and Workforce Development Performance Indicators: The methods, data and/or results from this study were incorporated in the syllabus for the Fall 2022, Spring 2023, and Fall 2023 versions of the following courses at Purdue University: (a) CE 299: Smart Mobility, an optional undergraduate-level course at Purdue' B.S. civil engineering program, and (b) CE 398: Introduction to Civil Engineering Systems, a mandatory undergraduate-level course at Purdue University's B.S. civil engineering program, (c)an independent study on high-speed vehicle automation using a real-life autonomous racing vehicle. The students that took these courses are graduating and will soon be entering the workforce. Therefore, the research helped enlarge the pool of people trained to develop knowledge and utilize at least a part of the technologies developed in this research, and to put them to use when they enter the workforce.

Collaboration Performance Indicators: There was collaboration with one other agency and at least 4 academic institutions, and these organizations provided matching funds.

The study outcomes, outputs, and impacts are described in Chapter 13.

CHAPTER 13. STUDY OUTCOMES AND OUTPUTS

13.1 Outputs

13.1.1 Publications, conference papers, or presentations

(a) Journal Papers

1.Dong, J., Chen, S., & Labi, S. (2023). Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems. *Transportation Research Part C: Emerging Technologies*, Volume 156, November 2023, 104358, https://www.sciencedirect.com/science/article/pii/S0968090X23003480

2.Dong, J., Chen, S., Miralinaghi, M., Chen, T., Labi, S. (2022). Development and testing of an image transformer for explainable autonomous driving systems. *Journal of Intelligent and Connected Vehicles*, 5(3), 235-249. https://www.emerald.com/insight/content/doi/10.1108/JICV-06-2022-0021/full/pdf

(b) Presentations

1. Dong, J., Chen, S., Chen, T., Miralinaghi, M., Labi, S. (2023). End-to-end Transformer for Explainable Autonomous Driving. *TRB 102nd Annual Meeting*. Washington, D.C., USA. January 2023.
2. Dong, J., Labi, S. (2023). End-to-end Transformer for Explainable Autonomous Driving. *The Inaugural ICON Student Research Conference*, West Lafayette, IN, USA., February 2023.
3. Dong, J., Chen, S., Li, Y., Du, R., Labi, S. (2023). Explainable Autonomous Driving System with End-to-end Attention Model. *ASCE International Conference on Transportation and Development (ICTD)*, Austin, TX, USA., June 2023.
4. Dong, J., Labi, S. (2023). End-to-end Transformer for Explainable Autonomous Driving. *Fifth Bridging Transportation Researchers (BTR5)*, online conference, August 2023.

13.2 Outcomes

The outcome of this research project is the prospective change that can be made to the transportation system, or its regulatory, legislative, or policy framework, resulting from research and development outputs. This is explained below:

- Increased understanding and awareness of issues related to user trust in automation. The development of automated systems must be accompanied by conscious efforts to build the trust and confidence of the prospective users of automated vehicles. It is expected that this issue will be legislated in the US in the near future. In Europe, explainable AI has been institutionalized through published standards such as the European Union’s General Data Protection Regulation (GDPR) which specifically stipulates that “right to explanation” is required in the context of decisions made by complex systems, particularly where they are founded on black-box models. In efforts to achieve this goal in the US, it is useful (or even indispensable) that the decisions made by the automated processes be accompanied by explanations for such decisions. That way, the autonomous vehicle (AV) end users and developers can be assured of the rationale behind the AV's decisions and, if needed, investigate such rationale. Such capability could help not only enhance the transparency and accountability of the AVs' decisions but also to evaluate the AV's role in a critical event *ex ante* (before the critical event such as a collision or near miss) or *ex poste* (after the critical event has occurred).

13.3 Impacts

The demonstrated efficacy of using explainable AI to enhance the safety critical perception phase of autonomous driving and thence, to promote CAV deployment is expected to impact user trust in automated driving and incentivize prospective travelers and vehicle purchasers to patronize AVs. The impacts of such an increase in AV market penetration, in turn, will cause reduced fatalities, enhanced travel efficiency, and lower environmental adversities (specifically, GHG emissions, if AVs will use electric propulsion). The students who took part in this research, and those who benefited from the material incorporation into existing coursework at Purdue, will soon enter the workforce where they will be motivated to, and able to help enhance the explainability of AV decisions, build user trust in AV, on system. A list of specific impacts from this research project, are as follows:

- Perception has been identified as the main cause underlying most autonomous vehicle (AV) related accidents. As the key technology in perception, deep learning (DL) based computer vision models are considered to be black boxes due to poor interpretability. These have exacerbated user distrust and further forestalled their widespread deployment in practical usage. The developed explainable DL models for autonomous driving (which jointly predicts potential driving actions with corresponding explanations), are expected to not only boost user trust in autonomy but also serve as a diagnostic approach to identify any model deficiencies or limitations during the system development phase.
- The developed product is expected to impact situational awareness and driver assistance, in the sense that it can serve as a driving alarm system for both human-driven vehicles and autonomous vehicles. This is because the product is capable of quickly understanding/characterizing the environment and identifying any infeasible driving actions. In addition, the extra explanation head of the proposed model provides an extra channel for sanity checks to guarantee that the model learns the ideal causal relationships. Therefore, the product will be useful in the development of reliable autonomous systems.
- The development of an innovative XAI for CAV controls is expected to yield beneficial impacts on the transportation system and society in general. These include enhanced user trust in automation, higher rates of market acceptance of autonomous driving technologies, and higher market penetration. Other prospective benefits related to enhanced user trust include reduced crashes and travel efficiency (reduced travel time) which translate into lower vehicle operating costs, higher economic productivity, and more free time for social activities.

- The undergraduate student and graduate student that worked on this project will enter the workforce in 2024 to help support the workforce that will implement modern technologies such as those developed in this study.
- Parts of the research outcomes were incorporated in 2 undergraduate courses and 1 graduate level course at Purdue University in Fall 2022, Spring 2023, Fall 2023, and will be continued in future versions of these courses. The students, who have begun entering the workforce, benefitted from the key outcomes of this research through these academic platforms. Therefore, the project has helped enlarge the pool of people trained to develop knowledge and utilize the technologies developed in this research and are expected to put them to use as and when they enter the workforce.

REFERENCES

Alwosheel, A., van Cranenburgh, S. and Chorus, C.G. (2021). Why did you predict that? Towards explainable artificial neural networks for travel demand analysis, Transportation Research Part C: Emerging Technologies, https://doi.org/10.1016/j.trc.2021.103143.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2018.00636

Atakishiyev, S., Salameh, M., Yao, H., Goebel, R. (2021). Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions, https://arxiv.org/abs/2112.11561

Bahdanau, D., Cho, K.H., Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate, in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, May 7-9, 2015, San Diego, CA, USA.

Ben-Younes, H., Zablocki, É., Pérez, P., Cord, M. (2022). Driving behavior explanation with multi-level fusion. Pattern Recognit 123. https://doi.org/10.1016/j.patcog.2021.108421

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., et al. (2016), End to end learning for self-driving cars, https://arxiv.org/abs/1604.07316.

Bustos, C., Rhoads, D., Solé-Ribalta, A., Masip, D., Arenas, A., Lapedriza, A. and Borge-Holthoefer, J. (2021), Explainable, automated urban interventions to improve pedestrian and vehicle safety, Transportation Research Part C: Emerging Technologies, https://doi.org/10.1016/j.trc.2021.103018.

Chen, S., Dong, J., Ha, P., Li, Y. and Labi, S. (2021), Graph neural network and reinforcement learning for multi-agent cooperative control of connected autonomous vehicles, Computer-Aided Civil and Infrastructure Engineering, https://doi.org/10.1111/mice.12702.

Chen, S., Leng, Y. and Labi, S. (2019). A deep learning algorithm for simulating autonomous driving considering prior knowledge and temporal information, Computer-Aided Civil and Infrastructure Engineering, https://doi.org/10.1111/mice.12495.

Cui, Y., Yang, G., Veit, A., Huang, X., Belongie, S. (2018). Learning to evaluate image captioning, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2018.00608

Cui, Z., Henrickson, K., Ke, R. and Wang, Y. (2019). Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting, IEEE Transactions on Intelligent Transportation Systems, https://doi.org/10.1109/tits.2019.2950416.

Di, X., Shi, R. (2021). A survey on autonomous vehicle control in the era of mixed-autonomy: From physics-based to AI-guided driving policy learning. Transp Res Part C Emerg Technol 125, 103008. https://doi.org/10.1016/J.TRC.2021.103008

Do, L.N.N., Vu, H.L., Vo, B.Q., Liu, Z., Phung, D. (2019). An effective spatial-temporal attention based neural network for traffic flow prediction. Transp Res Part C Emerg Technol 108. https://doi.org/10.1016/j.trc.2019.09.008

Dong, J., Chen, S., Joun Ha, P.Y., Li, Y. and Labi, S. (2020). A DRL-based multiagent cooperative control framework for CAV networks: A graphic convolution Q network, https://arxiv.org/abs/2010.05437

Dong, J., Chen, S., Li, Y., Du, R., Steinfeld, A., Labi, S. (2021). Space-weighted information fusion using deep reinforcement learning: The context of tactical control of lane-changing autonomous vehicles and connectivity range assessment. Transp Res Part C Emerg Technol 128. https://doi.org/10.1016/j.trc.2021.103192

Dong, J., Chen, S., Li, Y., Ha, P.Y.J., Du, R., Steinfeld, A. and Labi, S. (2020). Spatio-weighted information fusion and DRL-based control for connected autonomous vehicles, 2020 IEEE 23rd International Conference on Intelligent Transportation Systems, ITSC 2020, https://doi.org/10.1109/ITSC45102.2020.9294550.

Dong, J., Chen, S., Miralinaghi, M., Chen, T., Labi, S. (2022). Development and testing of an image transformer for explainable autonomous driving systems. Journal of Intelligent and Connected Vehicles. https://doi.org/10.1108/jicv-06-2022-0021

Dong, J., Chen, S., Zong, S., Chen, T., Labi, S. (2021). Image transformer for explainable autonomous driving system, in: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC. https://doi.org/10.1109/ITSC48978.2021.9565103

Doran, D., Schulz, S., Besold, T.R. (2018). What does explainable AI really mean? A new conceptualization of perspectives, in: CEUR Workshop Proceedings, https://arxiv.org/abs/1710.00794.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale, https://arxiv.org/abs/2010.11929

Du, R., Chen, S., Dong, J., Ha, P.Y.J., Labi, S., 2021. GAQ-EBkSP: A DRL-based urban traffic dynamic rerouting framework using fog-cloud architecture, in 2021 IEEE International Smart Cities Conference, ISC2 2021. https://doi.org/10.1109/ISC253183.2021.9562832

Du, Y., Chen, J., Zhao, C., Liu, C., Liao, F., Chan, C.Y. (2022). Comfortable and energy-efficient speed control of autonomous vehicles on rough pavements using deep reinforcement learning, Transp Res Part C Emerg Technol 134, 103489. https://doi.org/10.1016/J.TRC.2021.103489

FHWA. (2019), Evaluation methods and techniques: Advanced transportation and congestion management technologies deployment program, Tech. Rep. Nr. FHWA-HOP-19-053, Prepared by the Volpe National Transportation Center, Cambridge, MA.

Ghaeini, R., Fern, X.Z., Tadepalli, P. (2020). Interpreting recurrent and attention-based neural models: A case study on natural language inference, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. https://doi.org/10.18653/v1/d18-1537

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Comput Surv. https://doi.org/10.1145/3236009

Ha, P., Chen, S., Du, R., Dong, J., Li, Y. and Labi, S. (2020). Vehicle connectivity and automation: A sibling relationship, Frontiers in Built Environment, https://doi.org/10.3389/fbuil.2020.590036.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2016.90

He, S., Liao, W., Tavakoli, H.R., Yang, M., Rosenhahn, B., Pugeault, N. (2021). Image Captioning Through Image Transformer, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-030-69538-5_10

Hendrycks, D., Gimpel, K. (2016). Bridging Nonlinearities and stochastic regularizers with Gaussian error linear units, https://arxiv.org/abs/1606.08415

Herdade, S., Kappeler, A., Boakye, K., Soares, J. (2019). Image captioning: Transforming objects into words, in: Advances in Neural Information Processing Systems, https://arxiv.org/abs/1906.05963.

Hewitt, C., Amanatidis, T., Politis, I., Sarkar, A. (2019). Assessing public perception of self-driving cars: The autonomous vehicle acceptance model, in: International Conference on Intelligent User Interfaces, Proceedings IUI. https://doi.org/10.1145/3301275.3302268

Horgan, J., Hughes, C., McDonald, J. and Yogamani, S. (2015). Vision-based driver assistance systems: Survey, taxonomy and advances, IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, https://doi.org/10.1109/ITSC.2015.329.

Hou, R., Jeong, S., Lynch, J.P., and Law, K.H. (2020). Cyber-physical system architecture for automating the mapping of truck loads to bridge behavior using computer vision in connected highway corridors, Transportation Research Part C: Emerging Technologies, https://doi.org/10.1016/j.trc.2019.11.024.

Hu, H., Zhao, T., Wang, Q., Gao, F. and He, L. (2020). R-CNN Based 3D object detection for autonomous driving, CICTP 2020: Transportation Evolution Impacting Future Mobility - Selected Papers from the 20th COTA International Conference of Transportation Professionals, https://doi.org/10.1061/9780784483053.077.

Hulse, L.M., Xie, H., Galea, E.R. (2018). Perceptions of autonomous vehicles: Relationships with road users, risk, gender, and age. Saf Sci. https://doi.org/10.1016/j.ssci.2017.10.001

Khastgir, S., Birrell, S., Dhadyalla, G., Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles, Transportation Research Part C: Emerging Technologies, https://doi.org/10.1016/j.trc.2018.07.001.

Kim, J., Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention, Proceedings of the IEEE International Conference on Computer Vision, available at: https://doi.org/10.1109/ICCV.2017.320.

Kim, J., Misu, T., Chen, Y.T., Tawari, A., Canny, J. (2019). Grounding human-to-vehicle advice for self-driving vehicles, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2019.01084

Kim, J., Moon, S., Rohrbach, A., Darrell, T. and Canny, J. (2020). Advisable learning for self-driving vehicles by internalizing observation-to-action rules, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, available at: https://doi.org/10.1109/CVPR42600.2020.00968.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z. (2018). Textual explanations for self-driving vehicles, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-030-01216-8_35

Kingma, D.P., Ba, J.L. (2015). Adam: A method for stochastic optimization, in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., Nass, C. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. International Journal on Interactive Design and Manufacturing. https://doi.org/10.1007/s12008-014-0227-2

Kotseruba, I., Tsotsos, J.K. (2022). Attention for vision-based assistive and automated driving: A review of algorithms and datasets. IEEE Transactions on Intelligent Transportation Systems 23. https://doi.org/10.1109/TITS.2022.3186613

Ku, J., Pon, A.D. and Waslander, S.L. (2019). Monocular 3D object detection leveraging accurate proposals and shape reconstruction, Proceedings of the IEEE Computer Society Conference on Computer Vision, and Pattern Recognition, available at: https://doi.org/10.1109/CVPR.2019.01214.

Lei, T., Barzilay, R., Jaakkola, T. (2016). Rationalizing neural predictions, in: EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings. https://doi.org/10.18653/v1/d16-1011

Li, C., Meng, Y., Chan, S.H., Chen, Y.T. (2020). Learning 3D-aware egocentric spatial-Temporal Interaction via Graph Convolutional Networks, in: Proceedings - IEEE International Conference on Robotics and Automation. https://doi.org/10.1109/ICRA40945.2020.9197057

Li, D., Zhu, F., Chen, T., Wong, Y.D., Zhu, C., Wu, J. (2023). COOR-PLT: A hierarchical control model for coordinating adaptive platoons of connected and autonomous vehicles at signal-free intersections based on deep reinforcement learning. Transp Res Part C Emerg Technol 146, 103933. https://doi.org/10.1016/J.TRC.2022.103933

Li, G., Yang, Y., Li, S., Qu, X., Lyu, N., Li, S.E. (2022). Decision making of autonomous vehicles in lane change scenarios: Deep reinforcement learning approaches with risk awareness. Transp Res Part C Emerg Technol 134, 103452. https://doi.org/10.1016/J.TRC.2021.103452

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. https://doi.org/10.1109/CVPR.2017.106

Lin, Z., Feng, M., Dos Santos, C.N., Yu, M., Xiang, B., Zhou, B., Bengio, Y. (2017). A structured self-attentive sentence embedding, in 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings.

Lioris, J., Pedarsani, R., Tascikaraoglu, F.Y. and Varaiya, P. (2017). Platoons of connected vehicles can double throughput in urban roads, Transportation Research Part C: Emerging Technologies, https://doi.org/10.1016/j.trc.2017.01.023.

Litman, T. (2023). Autonomous vehicle implementation predictions: Implications for transport planning, Victoria Transport Policy Institute, https://www.vtpi.org/avip.pdf

Liu, Y., Liu, Z., Jia, R. (2019). DeepPF: A deep learning based architecture for metro passenger.

flow prediction. Transp Res Part C Emerg Technol 101. https://doi.org/10.1016/j.trc.2019.01.027

Liu, Y., Liu, Z. and Jia, R. (2019). DeepPF: A deep learning based architecture for metro passenger flow prediction, Transportation Research Part C: Emerging Technologies, https://doi.org/10.1016/j.trc.2019.01.027.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. (2021). Swin Transformer: hierarchical vision transformer using shifted windows. Proceedings of the IEEE International Conference on Computer Vision 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

McCausland, P. (2019). Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk, NBC News, https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281.

NTSB. (2019). Collison between vehicle controlled by developmental automated driving system and pedestrian., Highway Accident Report NTSB/HAR19/03 Washington, DC.

Mittelstadt, B., Russell, C., Wachter, S. (2019). Explaining explanations in AI, in: FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New YorkNY https://doi.org/10.1145/3287560.3287574

Omeiza, D., Webb, H., Jirotka, M., Kunze, L. (2022). Explanations in autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems. vol. 23, no. 8, pp. 10142-10162, https://doi.org/10.1109/tits.2021.3122865

Pal, A., Mondal, S., Christensen, H.I. (2020). Looking at the right stuff - Guided semantic-gaze for autonomous driving, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR42600.2020.01190

Palazzi, A., Abati, D., Calderara, S., Solera, F., Cucchiara, R. (2019). Predicting the driver’s focus of attention: The DR (eye)VE project. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2018.2845370

Pan, Y., Yao, T., Li, Y., Mei, T. (2020). X-Linear attention networks for image captioning, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR42600.2020.01098

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Pages 311–318. https://doi.org/10.3115/1073083.1073135

Peng, B., Keskin, M.F., Kulcsár, B. and Wymeersch, H. (2021). Connected autonomous vehicles for improving mixed traffic efficiency in unsignalized intersections with deep reinforcement learning, Communications in Transportation Research, Vol 1(1), https://doi.org/10.1016/j.commtr.2021.100017

Ren, S., He, K., Girshick, R. and Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, https://doi.org/10.1109/TPAMI.2016.2577031.

Rezaei, A., Caulfield, B. (2020). Examining public acceptance of autonomous mobility. Travel Behav Soc., 21(1), 235-246, https://doi.org/10.1016/j.tbs.2020.07.002

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision, Annual Review of Vision Science, 14(2), 437-457. https://doi.org/10.1146/annurev-vision-082114-035733.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 4510-4520, https://doi.org/10.1109/CVPR.2018.00474

Schwarting, W., Alonso-Mora, J. and Rus, D. (2018). Planning and decision-making for autonomous vehicles, Annual Review of Control, Robotics, and Autonomous Systems, Vol. 1 No. 1, pp. 187–210.

Sinha, K.C. and Labi, S. (2007). Transportation decision making: Principles of project evaluation and programming, Wiley, NJ.

Shi, H., Zhou, Y., Wu, K., Wang, X., Lin, Y., Ran, B. (2021). Connected automated vehicle cooperative control with a deep reinforcement learning approach in a mixed traffic environment. Transp Res Part C Emerg Technol 133, 103421, https://doi.org/10.1016/j.trc.2021.103421

Sowmya Shree, B. V. and Karthikeyan, A. (2018). Computer vision based advanced driver. assistance system algorithms with optimization techniques-A review, Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018, https://doi.org/10.1109/ICECA.2018.8474604.

Talpaert, V., Sobh, I., Ravi Kiran, B., Mannion, P., Yogamani, S., El-Sallab, A. and Perez, P. (2019). Exploring applications of deep reinforcement learning for real-world autonomous driving systems, VISIGRAPP 2019 - Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, https://doi.org/10.5220/0007520305640572.

TRB. (2018). Socioeconomic impacts of automated and connected vehicle, Proceedings of the 6th EU–U.S. Transportation Research Symposium, Transportation Research Board, Washington, DC.

TRB. (2019). TRB forum on preparing for automated vehicles and shared mobility: Mini-workshop on the importance and role of connectivity, Transportation Research Circular, Transportation Research Board, Washington, DC.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need, in: Advances in Neural Information Processing Systems, https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Veres, S.M., Molnar, L., Lincoln, N.K. and Morice, C.P. (2011). Autonomous vehicle control systems – A review of decision making, Proceedings of the Institution of Mechanical Engineers. Part I: Journal of Systems and Control Engineering, available at: https://doi.org/10.1177/2041304110394727.

Voigt, P., von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR) A Practical Guide, The EU General Data Protection Regulation (GDPR), https://gdpr.eu/what-is-gdpr/

Wang, Y., Huang, M., Zhao, L., Zhu, X. (2016). Attention-based LSTM for aspect-level sentiment classification, in: EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings. https://doi.org/10.18653/v1/d16-1058

Wang, Y., Huang, R., Song, S., Huang, Z., Huang, G. (2021). Not all images are worth 16x16 words: Dynamic vision transformers with adaptive sequence length, in: NeurIPS. https://www.arxiv-vanity.com/papers/2105.15075/

Wang, Y., Xu, J., Sun, Y. (2022). End-to-end transformer-based model for image captioning, https://arxiv.org/abs/2203.15350

Wiegreffe, S., Pinter, Y. (2019). Attention is not not explanation, in: EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. https://doi.org/10.18653/v1/d19-1002

Wolfe, B., Dobres, J., Rosenholtz, R. and Reimer, B. (2017). More than the useful field: Considering peripheral vision in driving, Applied Ergonomics, https://doi.org/10.1016/j.apergo.2017.07.009.

World Bank. (2005). A framework for the economic evaluation of transport projects, Transport Notes, https://documents1.worldbank.org/curated/en/360501468327922938/pdf/339260rev.pdf

Xia, Y., Kim, J., Canny, J., Zipser, K., Canas-Bajo, T., Whitney, D. (2020). Periphery-fovea multi-resolution driving model guided by human attention, in: Proceedings - 2020 IEEE Winter

Conference on Applications of Computer Vision, WACV 2020. https://doi.org/10.1109/WACV45572.2020.9093524

Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., Whitney, D., 2019. Predicting driver attention in critical situations, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-030-20873-8_42

Xing, Y., Lv, C., Cao, D., Velenis, E. (2021). Multi-scale driver behavior modeling based on deep spatial-temporal representation for intelligent vehicles. Transp Res Part C Emerg Technol 130. https://doi.org/10.1016/j.trc.2021.103288

Xu, C., Ding, Z., Wang, C., Li, Z. (2019). Statistical analysis of the patterns and characteristics of connected and autonomous vehicles involved crashes. J Safety Res 71. https://doi.org/10.1016/j.jsr.2019.09.001

Xu, H., Gao, Y., Yu, F., Darrell, T. (2017). End-to-end learning of driving models from large-scale video datasets, in: Proceedings, 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. https://doi.org/10.1109/CVPR.2017.376

Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention, in: 32nd International Conference on Machine Learning, ICML 2015.

Xu, Y., Yang, X., Gong, L., Lin, H.C., Wu, T.Y., Li, Y., Vasconcelos, N. (2020). Explainable object-induced action decision for autonomous vehicles, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR42600.2020.00954

Yadron, D. and Tynan, D. (2016). Tesla driver dies in first fatal crash while using autopilot mode, The Guardian, https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk

Yu, B., Lee, Y., Sohn, K. (2020). Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (GCN). Transp Res Part C Emerg Technol 114, 189-204. https://doi.org/10.1016/j.trc.2020.02.013

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T. (2020). BDD100K: A diverse driving dataset for heterogeneous multitask learning, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 17-19 June, San Juan, PR, USA, https://doi.org/10.1109/CVPR42600.2020.00271

Zablocki, É., Ben-Younes, H., Pérez, P., Cord, M. (2022). Explainability of deep vision-based autonomous driving systems: Review and challenges. Int J Comput Vis 130. https://doi.org/10.1007/s11263-022-01657-x

Zhang, K., He, F., Zhang, Z., Lin, X., Li, M. (2020). Multi-vehicle routing problems with soft time windows: A multi-agent reinforcement learning approach. Transp Res Part C Emerg Technol 121. https://doi.org/10.1016/j.trc.2020.102861

Zhao, H., Jia, J., Koltun, V. (2020). Exploring self-attention for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 17-19 June, San Juan, PR, USA, https://doi.org/10.1109/CVPR42600.2020.01009

Zhou, F., Li, L., Zhang, K. and Trajcevski, G. (2021). Urban flow prediction with spatial–temporal neural ODEs, Transportation Research Part C: Emerging Technologies, 124(1), 102912 https://doi.org/10.1016/j.trc.2020.102912.

Zhuang, L., Wang, L., Zhang, Z., and Tsui, K.L. (2018). Automated vision inspection of rail surface cracks: A double-layer data-driven framework, Transportation Research Part C: Emerging Technologies, 92(1), 258-277, https://doi.org/10.1016/j.trc.2018.05.007.

Zhu, W., Wu, J., Fu, T., Wang, J., Zhang, J. and Shangguan, Q. (2021). Dynamic prediction of traffic incident duration on urban expressways: a deep learning approach based on LSTM and MLP, Journal of Intelligent and Connected Vehicles, 4(2), 80-91, https//doi.org/10.1108/JICV-03-2021-0004.

APPENDIX

Published Related Work

Paper 1: Dong, J., Chen, S., & Labi, S. (2023). Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems. *Transportation Research Part C: Emerging Technologies*, Volume 156, November 2023, 104358, https://www.sciencedirect.com/science/article/pii/S0968090X23003480

Abstract

User trust has been identified as a critical issue that is pivotal to the success of autonomous vehicle (AV) operations where artificial intelligence (AI) is widely adopted. For such integrated AI-based driving systems, one promising way of building user trust is through the concept of explainable artificial intelligence (XAI) which requires the AI system to provide the user with the explanations behind each decision it makes. Motivated by both the need to enhance user trust and the promise of novel XAI technology in addressing such need, this paper seeks to enhance trustworthiness in autonomous driving systems through the development of explainable Deep Learning (DL) models. First, the paper casts the decision-making process of the AV system not as a classification task (which is the traditional process) but rather as an image-based language generation (image captioning) task. As such, the proposed approach makes driving decisions by first generating textual descriptions of the driving scenarios, which serve as explanations that humans can understand. To this end, a novel multi-modal DL architecture is proposed to jointly model the correlation between an image (driving scenario) and language (descriptions). It adopts a fully Transformer-based structure and therefore has the potential to perform global attention and imitate effectively, the learning processes of human drivers. The results suggest that the proposed model can and does generate legal and meaningful sentences to describe a given driving scenario, and subsequently to correctly generate appropriate driving decisions in autonomous vehicles (AVs). It is also observed that the proposed model significantly outperforms multiple baseline models in terms of generating both explanations and driving actions. From the end user's perspective, the proposed model can be beneficial in enhancing user trust because it provides the rationale behind an AV's actions. From the AV developer's perspective, the explanations from this explainable system could serve as a "debugging" tool to detect potential weaknesses in the existing system and identify specific directions for improvement.

Paper 2: Dong, J., Chen, S., Miralinaghi, M., Chen, T., Labi, S. (2022). Development and testing of an image transformer for explainable autonomous driving systems. *Journal of Intelligent and Connected Vehicles*, 5(3), 235-249. https://www.emerald.com/insight/content/doi/10.1108/JICV-06-2022-0021/full/pdf

Abstract

Purpose. Perception has been identified as the main cause underlying most autonomous vehicle (AV) related accidents. As the key technology in perception, deep learning (DL) based computer vision models are considered to be black boxes due to poor interpretability. These have exacerbated user distrust and further forestalled their widespread deployment in practical usage. This paper aims to develop explainable DL models for autonomous driving by jointly predicting potential driving actions with corresponding explanations. The explainable DL models can not only boost user trust in autonomy but also serve as a diagnostic approach to identify any model deficiencies or limitations during the system development phase.

Design/methodology/approach. This paper proposes an explainable end-to-end autonomous driving system based on "Transformer," a state-of-the-art (SOTA) self-attention-based model. The model maps visual features from images collected by onboard cameras to guide potential driving actions with corresponding explanations and aims to achieve soft attention over the image's global features.

Findings. The results demonstrate the efficacy of our proposed model as it exhibits superior performance (in terms of correct prediction of actions and explanations) compared to the benchmark model by a significant margin with much lower computational cost on a public dataset (BDD-OIA). From the ablation studies, the proposed self-attention module also outperforms other attention mechanisms in feature fusion and can generate meaningful representations for downstream prediction.

Originality/value. In the contexts of situational awareness and driver assistance, the proposed model can perform as a driving alarm system for both human-driven vehicles and autonomous vehicles because it is capable of quickly understanding/characterizing the environment and identifying any infeasible driving actions. In addition, the extra explanation head of the proposed model provides an extra channel for sanity checks to guarantee that the model learns the ideal causal relationships. This provision is critical in the development of autonomous systems.