DESIGN AND IMPLEMENTATION OF AN END-TO-END DEEP LEARNING SYSTEM FOR AUTONOMOUS DRIVING: FROM VISUAL PERCEPTION TO INTELLIGENT VEHICLE CONTROL

Sungho Kim [1], Khairul Anam [2], Md Nafis Azad Nobel [2], Pritom Das [2], Mohammad Somon Sikder [2], Hemayet Uddin Himel [3], Yearanoor Khan [3], Seohyun Yang [2], Saeromi Kim [4], Ibtisum Ahmed Nihal [2]

[1] Department of Computer Science, Korea University, Seoul, Korea
[2] Department of Computer Science, Pacific States University, Los Angeles, UnitedStates
[3] Department of information Systems, Pacific States University, Los Angeles, United States
[4] Department of Business Administration, Pacific States University, Los Angeles, United States

ABSTRACT

End-to-end deep learning has become an important and significant driver of autonomous driving study because it substitutes an intricate multi-stage sense; plan; act pipeline with a single extremely multi-faceted differentiable model that can be trained end-to-end from sensorial input all the way to control outcome. Though end-to-end learning is profitable and useful relative to classical modular pipelines in that it can learn to compress the visual information in an end-to-end fashion and consequently learn task-specific latent representations optimized for driving, the identical design decisions that favor holistic optimization furthermore complicate analysis and provide difficulty on the way toward safe deployment: distribution shift in imitation learning, interpretability, debugging, and safety in closed-loop control. This document describes a real-world end-to-end CNN forsteering angle prediction trained by means of supervised imitation learning on the Udacity Behavioral Cloning dataset recorded in the Udacity simulator . Our model is largely inspired by the NVIDIA end-to-end architecture (frequently called PilotNet), which performs CNN-to-steering prediction employing sole camera image input. We employ a perceptual-to-control training system that uses targeted region-of-interest cropping, input normalization, and data augmentation (brightness modifications, horizontal flipping, and spatial shifting) to generalize better from variation in lighting and camera standpoint. Our proposed network on the held-out test set results MSE=0.0166 and MAE=0.0682, which is quiterobust open-loop prediction accuracy for steering regression given the running conditions of the dataset. We will moreover elaborate on the reasons that robust open-loop regression metrics do not fully characterize closed-loop driving safety, and the safety and interpretability instruments (for example, causal attention maps and safety monitors) which we describe that can supplement end-to end control policies in actual deployment.
Index Terms: Autonomous Driving, End-to-End Learning, Imitation Learning, Behavioral Cloning, Convolutional Neural Networks, Direct Perception, Explainable AI, AI Safety.

1. INTRODUCTION

End-to-end deep learning has gained appeal among autonomous driving systems because it substitutes the multi-stage sense-plan-act stack with a single differentiable process (deep learning model) that can be together optimized end-to-end for direct control from sensorial input. In rule, end to-end techniques can remove hand-designed subgoals and information constraints of classical modular pipelines (perception, localization, prediction, planning, control), and learn task-specific latent features directly optimized for driving. The combination of far-reaching simplification and holistic optimization, nonetheless, moreover creates challenges -- distribution shift during imitation learning, absence of interpretability, obstacles of debugging, and assurance of safety in a closed-loop learned policy.

In this document, we propose a productive end-to-end CNN for steering angle prediction trained by supervised imitation learning applying the large-scale Udacity Behavioral Cloning data set collected in the Udacity simulator. Based on the NVIDIA end-to-end structure (broadly known as PilotNet), this model employ image served as input and output a continuous steering command. We build a perception-to control training pipeline with precise and particular region-of-interest crop, input normalization and multiple data augmentation techniques (brightness modification, horizontal flipping, spatial shifting) on the data in order to handle lights and perspective changes. On held-out test data, the proposed network generates MSE=0.0166 and MAE=0.0682, demonstrating compelling open-loop accuracy of the steering regression for the running conditions of the data. We go on to clarify in detail why robust open-loop regression measurescannot precisely characterize closed-loop driving safety, and present safety and interpretability instruments (such as causal attention maps and safety monitors) that can augment end to-end control systems in real-world deployments.

2. BACKGROUND AND RELATED WORK

Vision-based autonomous driving study has historically alternated between mediated perception, behavior-reflex (end-to-end), and direct perception techniques. Mediated perception explicitly detects setting components such as lanes or vehicles, while reflex methods directly predict control commands from images. Direct perception, placed between the two, forecasts essential driving affordances that guide a simpler controller.

Early neural-network driving inquiry includes ALVINN, which demonstrated camera-based lane following employing neural networks. In the deep learning era, NVIDIA's PilotNet showed that convolutional neural networks can learn steering control directly from images, employing YUV inputs, convolutional layers, and fully linked layersrunning in genuine time. However, behavioral cloning suffers from distribution shift, where the model encounters states not seen during training. Methods like DAgger address this by iteratively correcting policy errors.

Recent study extends end-to-end driving beyond uncomplicated and elementary image-to-steering models. Approaches contain dependent imitation learning for high-level command control, trajectory based models like ChauffeurNet for enhanced robustness, simulation platforms such as CARLA for safe evaluation, and sensor-fusion transformer models like TransFuser for sophisticated and multifaceted urban environments. New integrated frameworks such as UniAD and VAD integrate perception, prediction, and planning, while multimodal systems like EMMA and DriveLM explore language-guided driving reasoning. Finally, interpretability techniques such as causal attention visualization help identify which image regions affect steering decisions, supporting safer and more clear end-to-end driving systems. Beyond individual end-to-end perception-to-control models, real-world deployment of autonomous vehicles requires a complete

system architecture integrating perception, sensor fusion, planning, and control. NVIDIA has developed a comprehensive autonomous driving platform combining specialized hardware and deep learning software frameworks.

At the hardware level, NVIDIA provides automotive computing platforms such as NVIDIA DRIVE AGX, which deliver high computational performance required to process large volumes of real-time sensor data. Autonomous vehicles typically rely on multiple sensors including cameras, LiDAR, radar, and ultrasonic sensors. GPU acceleration enables these platforms to process sensor streams simultaneously and perform deep neural network inference in real time.

The software architecture of the autonomous driving stack is typically organized into several functional layers. The perception layer uses deep neural networks to detect objects such as vehicles, pedestrians, lane markings, and traffic signals from sensor data. Sensor fusion modules integrate information from multiple sensors to improve environmental understanding and reduce uncertainty.

Following perception and fusion, planning systems generate safe driving trajectories by considering road geometry, traffic rules, and surrounding vehicles. Finally, vehicle control modules translate these trajectories into executable commands such as steering, acceleration, and braking.

NVIDIA's early end-to-end system, commonly known as PilotNet, demonstrated that convolutional neural networks can directly map camera images to steering commands using behavioral cloning (Bojarski et al., 2016). This work laid the foundation for modern end-to-end autonomous driving research and continues to influence the development of integrated autonomous driving architectures.

3. METHODOLOGY

This part formalizes the end-to-end steering problem, particulars dataset assumptions, and describes the executed CNN, preprocessing, and training goal.

4. PROBLEM FORMULATION

Let (It) be the image of the front-facing camera at time (t). Also, let (y_t) represent a ground-truth steering command as supplied by a seasoned or human driver. From end-to-end behavioral cloning, we learn a function (f_\theta) for some parameters (\theta) as follows: [Y t =f data (I t).]
Learning pushes parameters 0θ8 to minimize a supervised regression loss on a set of image-steering pairs. This learning setting is exactly imitative learning/behavior cloning as condensed in end-to-end driving surveys.

5. DATASET AND SIMULATION SETTING

We employ the Udacity Behavioral Cloning dataset generated within the simulator environmentsupplied for the Behavioral Cloning Project by Udacity. The official undertaking materials describe the workflow: gather driving data in the simulator, train a CNN to predictsteering from image data, and then deploy the learned model to drive the artificial and imitation car autonomously.

Public mirrors of Udacity behavioral cloning data frequently contain numerous and manifold forward-facing camera views (center/left/right) and further logged signals such as throttle, brake, and speed; these variations are commonly used to extend training variety and to simulate recovery behaviors, although the core learning target in this paper issteering.

6. NETWORK ARCHITECTURE AND PERCEPTION-TO-CONTROL MAPPING

Our steering predictor is a CNN inspired by the NVIDIA end-to-end design commonly associated with "PilotNet." In the initial NVIDIA document, the network is described as a 9-layer architecture comprising a normalization layer, five convolutional layers, and three fully linked layers; inputs are split into YUV planes before being fed through the network, and the outcome alignsto a steering-related control value.

Conceptually, the convolutional layers act as a learned perception front-end by extracting spatial and texture patternsrelevant to road following (lane markings, road edges, curvature signals), while the fully related and affiliated layersintegrate these features into a continuous control command. NVIDIA's follow up interpretability paper explicitly frames PilotNet as producing steering angles from road images and motivates saliency-style clarifications as a way to increase trust and to debug end-to-end systems.

7. PREPROCESSING AND AUGMENTATION

A recurring theme in imitation-learning driving is that generalization depends strongly on data variety and on reducing distribution shift. End-to-end driving surveys highlight that augmentation plans (brightness change, cropping, noise,standpoint changes) can help reveal the model to off-nominalstates and reduce compounding-error failures during closed-loop rollout. Our preprocessing pipeline followsthese principles:Region-of-interest cropping. We remove the top portion of each frame to focus model capability on road geometry, lane markers, and near-field drivable space, minimizing spurious correlations with sky and background texture. This option is uniform and unchanging with the intuition displayed in NVIDIA's simulator visualization, where the area below the horizon is highlighted as the portion sent to the CNN.

Color-Space Conversion And Normalization: We convert to YUV and normalize pixel values to stabilize optimization. The initial NVIDIA architecture explicitly uses YUV input planes and includes a dedicated normalization layer as part of the network design, indicating the weight ofstandardized input scaling for stable training and productive GPU inference.

Data Augmentation: We apply (i) random brightness modification, (ii) horizontal flipping with indication inversion of steering labels, and (iii) random translations/changes to simulate lateral deviation. These choices align with end-to-end driving literature highlighting artificial and manufactured changes/rotations and multi-camera setups to teach recovery from off-center positions.

8. TRAINING OBJECTIVE AND OPTIMIZATION PROTOCOL

We train with mean squared error(MSE) loss between predicted and ground-truth steering:

$$\mathcal{L}_{\text{MSE}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f_\theta(I_i)\right)^2$$

This goal tallies the canonical training description in NVIDIA's end-to-end steering work, which states that the network is trained to minimize mean squared error between the network's steering

outcome and the human driver (or modified) steering label. To reduce overfitting, we employ initial stopping established on validation loss, a standard supervised-learning control for training cessation when generalization performance plateaus. More widely, end-to-end driving surveys highlight that evaluation and training protocols must explicitly address generalization, because open-loop regression can look robust while closed-loop driving can still fail under compounding error.

9. EXPERIMENTAL RESULTS AND ANALYSIS

Quantitative Results

On held-out test data from the Udacity Behavioral Cloning dataset, our model achieves:

- **Mean Squared Error (MSE): 0.0166**
- **Mean Absolute Error (MAE): 0.0682**

These metrics suggest robust open-loop steering prediction accuracy for the dataset distribution used in this assignment. The Udacity assignment framing; training a model to outcome steering angles from image data, then employing the model to drive autonomously in the simulator; underscores that steering regression accuracy is an essential and required component for success on the path.

Interpreting Open-Loop Accuracy Vs Closed-Loop Driving

While MSE/MAE provide a compact summary of prediction fidelity, driving is a sequential decision process where errors accumulate. End-to-end driving literature stresses that a policy's behavior in closed-loop operation can diverge significantly from open-loop regression metrics due to distribution shift; this is exactly why data augmentation, data diversification, and on-policy plans (for instance, DAgger-style methods) are analyzed. NVIDIA's initial work highlights this gap by assessing in simulation utilizing an intervention-based autonomy metric alternatively than exclusively open-loop loss, and CARLA similarly assesses in scenario-based closed-loop conditions (traffic density, weather, infractions). These choices across the literature inspire a cautious and traditional interpretation of MSE/MAE: they are educational and enlightening but partial indicators of safety and dependability.

10. QUALITATIVE BEHAVIOR AND STABILITY CONSIDERATIONS

In simulator testing, the model shows consistent steering behavior when it acquires uniform and unchanging road curvature patterns and when data augmentation exposes it to minimal driving variations. However, end-to-end models continue vulnerable to rare situations such as atypical lighting, unforeseen and unanticipated road textures, sharp turns, or recovery states that is not adequately represented in the training data. Prior inquiry shows that standard behavioral cloning may fail in such situations, which motivates training approaches that intentionally introduce hard and challenging or artificial and manufactured situations. Studies such as NVIDIA's PilotNet furthermore demonstrate that CNN-based steering models can achieve real-time performance (around 30 FPS) when executed on fitting automotive hardware. Safety, Interpretability, and Ethical Considerations Behavioral cloning simplifies the driving task into a supervised learning problem but introduces multiple limitations. First, distribution shift can cause compounding errors when the vehicle encounters states not present in the training data. Second, rare or long-tail events may not be captured in the dataset, producing potential safety risks. Third, forecasting

exclusively steering angles does not fully represent the intricacy of real-world driving, which necessitates interaction aware planning and decision-making.

11. INTERPRETABILITY IN END-TO-END DRIVING MODELS

End-to-end driving systems are frequently criticized for their absence of transparency. Explainable AI techniques such as saliency maps and attention visualization can highlight which regions of an image influence steering predictions. These techniques help verify whether the model concentrates on relevant road features alternatively than unconnected background patterns. Methods like causal attention filtering further improve explanation quality by removing deceiving attention signals. Consequently, preprocessing steps such as ROI cropping and data augmentation furthermore support interpretability by encouraging the network to focus on relevant and substantial visual indications.

12. HYBRID SAFETY ARCHITECTURES

Because neural driving policies lack official safety guarantees, numerous researchers suggest integrating end-to-end models with rule-based safety layers. These observation systems enforce limitationssuch as collision avoidance, steering limits, and safe following distance.

Safety-focused frameworks and regulatory guidance; such as NHTSA automated driving recommendations, ISO 26262 for functional safety, ISO 21448 (SOTIF), and UL 4600; suggest that AI driving components should work within a broader safety architecture alternatively than serving as the only decision system.

13. FUTURE DIRECTION OF END-TO-END AUTONOMOUS DRIVING

Recent study trends are shifting from easy and straightforward perception-to-steering models toward full-stack end-to-end autonomous systems. Modern architectures integrate perception, prediction, and planning utilizing integrated frameworks such as UniAD, VAD, and transformer based multimodal models like TransFuser. These techniques evaluate performance employing planning accuracy, collision avoidance, and scenario-based testing instead than exclusively steering error metrics.

This progression signifies that future autonomous driving systems will depend on richer location understanding and integrated decision-making frameworks beyond essential and primary steering regression.

14. CONCLUSION

This assignment executes and extends an end-to-end behavioral cloning pipeline for autonomous vehicle steering in simulation established on a CNN inspired by the NVIDIA end-to-end architecture, trained on the Udacity Behavioral Cloning dataset. We present the building blocks and experiment with integration of ROI-based preprocessing, conversion to and normalization of YUV space, and attitude and brightness augmentation; finding encouraging outcomes. Our last model, trained with regularization and dropout attains a MAE= 0.0682 (MSE= 0.0166) on held-out test data, an indication of high quality open loop steering prediction within the data distribution.

Simultaneously, comprehensive and thorough inquiry in the area shows that one can notstate driving safety in the genuine world established on open loop regression solely: distribution shift, long-tail events, interpretability gaps, safety assurance requirements have led to (i) closed-loop simulator (for instance, CARLA) evaluation, (ii) policy training with explicit handling of errors compounding, and (iii) hybrid architectures of learned policies, safety monitoring/controlling, safety case reasoning in line with standards and advice (ISO 26262, ISO 21448/SOTIF, UL 4600, NHTSA advice). A modular, trajectory-centric focus (waypoints/paths instead of the present and existing raw steering) on future work. Developing multimodal sensor fusion comprising camera and LiDAR and/or radar. Developing scenario-based closed loop testing, evaluation and robustness stress tests. Developing user and data-friendly explainable AI instruments (causal attention, saliency) to permit simple and straightforward debugging and responsibility for safety essential deployment .

REFERENCES

[1] Mariusz Bojarski et al., “End to End Learning for Self-Driving Cars,” arXiv:1604.07316, 2016.

[2] Alexey Dosovitskiy et al., “CARLA: An Open Urban Driving Simulator,” arXiv:1711.03938 / CoRL 2017.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Deep learning,” Nature, 2015.

[4] Dean A. Pomerleau, “ALVINN: An Autonomous Land Vehicle in a Neural Network,” NeurIPS, 1988/1989.

[5] Chenyi Chen et al., “DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving,” arXiv:1505.00256 / ICCV 2015.

[6] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell, “A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning (DAgger),” arXiv:1011.0686 / AISTATS 2011.

[7] Felipe Codevilla et al., “End-to-end Driving via Conditional Imitation Learning,” arXiv:1710.02410, 2017.

[8] Mayank Bansal et al., “ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst,” RSS 2019 / arXiv:1812.03079.

[9] Aditya Prakash et al., “Multi-Modal Fusion Transformer for End-to-End Autonomous Driving (TransFuser),” CVPR 2021 / arXiv:2104.09224.

[10] Yihan Hu et al., “Planning-Oriented Autonomous Driving (UniAD),” CVPR 2023. [11] Bo Jiang et al., “VAD: Vectorized Scene Representation for Efficient Autonomous Driving,” ICCV 2023.

[11] Jianyu Chen et al., “Deep Imitation Learning for Autonomous Driving in Generic Urban Scenarios with Enhanced Safety,” arXiv:1903.00640 / IROS 2019.

[12] Jinkyu Kim and John Canny, “Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention,” ICCV 2017.

[13] ISO, “ISO 26262-1:2018 Road vehicles — Functionalsafety,” and “ISO 21448:2022 Road vehicles

— Safety of the intended functionality (SOTIF).”

[14] NHTSA, “Automated Driving Systems 2.0: A Vision for Safety,” 2017.

[15] ANSI/UL 4600, “Standard for Safety for the Evaluation of Autonomous Products,” overview and public summaries