Advancing Autonomous Driving through Direct Perception and Attention

3 Team Members

1 Introduction

We will build a model that uses an image to estimate the distance of the closest vehicle, a crucial technique in autonomous driving.

2 Method

2.1 Motivation

We will be reproducing the algorithms used in the paper *DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving* by Chen *et al.* published in 2015 IEEE International Conference for Computer Vision (ICCV).

Two paradigms currently exist for vision-based autonomous driving: mediated perception approach and behavior reflex approach. Mediated perception involves parsing entire driving scenes, constructing a high-dimensional world representation, and relying on the accuracy and field-of-view of sensors such as radar. Behavior reflex approach uses blind mapping of images to control commands, but struggles to handle rare events and complicated driving conditions.

The paper presents a direct perception methodology, which is a novel approach for learning/estimating affordances and mapping input images to a limited set of essential perception indicators linked to road conditions. This method falls between the mediated perception and behavior reflex approaches and allows for a compact affordance representation as perception output, as well as the development of a simple controller.

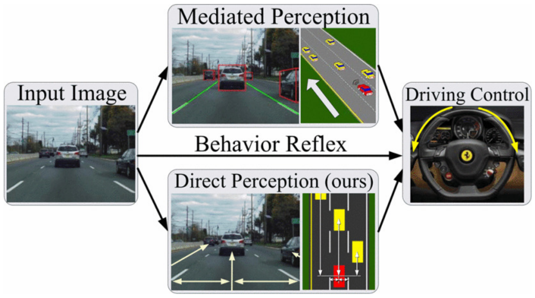

Figure 1: Three paradigms for autonomous driving (Chen *et al.*, 2015)

2.2 Overview

For this project, we will implement a direct perception deep learning CNN for autonomous driving using the Caffe deep learning framework and the AlexNet architecture. We will re-implement the

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

ConvNet framework and train and test it on subsets of the KITTI dataset. We will then compare our results to those from the paper.

In this project, we will extend the direct perception methodology with self-attention. Self-attention enables the model to focus on a specific region of a test image, which is useful for detecting very close obstacles or abruptly moving nearby vehicles. To implement self-attention, we will perceive such obstacles or vehicles and give them greater weight in the decision-making process, particularly in the mapping from affordances to actions for the controller.

3 Plan of Experiments

3.1 Implementation

In the paper, the direct perception methodology is developed as ConvNet, which is built upon Caffe (a deep learning framework) and a standard AlexNet (a CNN architecture). The ConvNet framework estimates affordances in an autonomous driving setting and creates two mappings for decision making: 1) mapping from image to affordances and 2) mapping from affordances to action for the controller. The methodology uses 5 convolutional layers and 4 fully connected layers with output dimensions of 4096, 4096, 256, and 13. A Euclidean loss function is used. 13 affordance indicators are employed.

Additionally, the paper implements ConvNet two times as part of their methodology, once for 0-15 meter away objects and again for 15-55 meter away objects, then combines the predictions for a comprehensive view. For the scope of this project, we will limit training and testing to objects within 0-15 meters.

3.2 Datasets

We will be using the same dataset used in the paper for depth perception (KITTI). The dataset is publicly available at `https://www.cvlibs.net/datasets/kitti/` and includes:

- Raw and processed grayscale stereo sequences (png)
- Raw and processed color stereo sequences (png)
- 3D Velodyne point clouds (100k points per frame, stored as binary float matrix)
- 3D GPS/IMU data(location, speed, acceleration, meta information, stored as text file)
- Calibration (Camera, Camera-to-GPS/IMU, Camera-to-Velodyne, stored as text file)
- 3D object tracklet labels (cars, trucks, trams, pedestrians, cyclists, stored as xml file)

3.3 Evaluation

This project falls under the category of supervised learning. We will evaluate our model using the same metric as proposed in the paper, which is the absolute mean error. We will then compare our results to the state-of-the-art technique at the time the paper was published, the discriminative part-based model (DPM).

4 Plan of Project

4.1 Division of work

Author 1:

- Initially implement the CNN model for estimating affordances and action mapping.
- Pre-process and clean the image datasets.

Author 2:

- Fine-tuned the CNN model implementation.

- Train the model on the datasets.
- Test and evaluate the performance.

Author 3 (manager):

- Conduct research to develop an understanding of related works and their applications to this project.
- Collaborate with other team members to ensure the project is on track.
- Implement the self-attention extension of the algorithm.
- Finalize the written works.

4.2 Experience or expertise

- A good understanding of machine learning concepts, such as supervised learning and convolutional neural networks (CNNs), is required for this project.
- It is important to have familiarity with image processing techniques and experience working with image datasets.
- Experience working with datasets, including cleaning and preprocessing data, and evaluating model performance, is required. Familiarity with the control system will be useful for managing code changes.
- Knowledge of autonomous driving systems and the associated challenges will be helpful in understanding the project's motivation.
- Strong programming skills in Python are necessary, as the project involves writing Python code to train and evaluate the CNN model.

4.3 Challenges and difficulties

One challenge we may face is that training the model may not converge as well as the model in the paper did. The authors likely spent a significant amount of time on this project, making it difficult to replicate their results. Another concern is that the model may require significant computing resources and time to train.

4.4 Contingency plans

Extension of the Direct Perception with Ensemble Method:

- This method combines the predictions from two or more neural network models to reduce the variance of predictions. The models have the same inputs and outputs. To implement the Ensemble Method for our project, we will create two different models that vary in the number of convolutional layers/connected layers. We will then weigh and combine predictions for a single prediction for the KITTI dataset.

Reduction of Convolutional Layers:

- If the scope of this proposal is too large, or if we run into issues with implementation, we will reduce the number of convolutional layers or number of fully connected layers and compare the performance of our implementation with that of the paper's implementation on the same dataset using the same performance metrics.

Additional Datasets:

- TORCS: 12 hours of human driving in a video game.
- Testing on our own smartphone video footage of driving and using predictions as a sanity check for the success of algorithm implementation.
- Using both TORCS and KITTI would enable the incorporation of rare events from TORCS simulations and add redundancy due to the size of the KITTI dataset.

4.5 Milestones

- Explore how to use the Caffe framework, AlexNet architecture, and KITTI dataset.
- Implement the model described in the paper.
- Train the model and evaluate its performance.
- Compare the model with the DPM method [5] and the results from the paper.
- Extend the model to include self-attention.
- Train the new model and evaluate its performance on the KITTI dataset.
- If time permits, apply this model to our own smartphone video footage of driving (this part is just for fun, as we don't have labels for car distance from the video footage. Still, we can analyze the inputs and predicted outputs to see if they make sense).

5 Related Work

Two main categories of approaches are used for image-based autonomous driving systems: the behavior reflex approach [2][3] and the mediated perception approach [4]. Behavior reflexes make decisions based on sensor inputs, while mediated perception uses the entire image for estimation. In this paper, a direct approach is used in which the image is preprocessed into a number of key indicators that are later used in training.

6 Reference

1. Chen, C., Seff, A., Kornhauser, A., and Xiao, J. (2015). DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving. IEEE International Conference on Computer Vision.
2. Pomerleau, D. A. (1989). ALVINN: an autonomous land vehicle in a neural network. Technical Report, DTIC Document.
3. Pomerleau, D. A. (1992). Neural network perception for mobile robot guidance. Technical Report, DTIC Document.
4. Ullman, S. (1980). Against direct perception. Behavioral and Brain Sciences, 3(03), 373–381.
5. Yebes, J. (2014). Supervised learning and evaluation of KITTI's cars detector with DPM. IEEE Intelligent Vehicles Symposium Proceedings.