End To End Learning for Self-Driving Cars

2025.03.21

Seung min chung

Contents

01 Abstract

Abstract

Convolutional self-driving cars system involve multiple stages and requires manual configuration of various parameters, whereas this paper propose an end-to-end pipline using a CNN

핵심

- End to End pipline
- Using only steering angle

ALVINN

Existing autonomous driving experiments have proven that the CNN structure is capable of recognizing 2D images well and enabling autonomous driving.

However, they perform poorly in off-road environments because they rely on an ‘if-then-else’ pipline.

This paper

In contrast, the method proposed in this paper adopts an end-to-end CNN approach, learning the entire driving pipeline—from raw camera inputs to steering commands—within a single network architecture.

By removing the need for manually crafted rules, the system automatically captures critical features and adapts to diverse driving conditions.

03 Overview of the DAVE-2 system

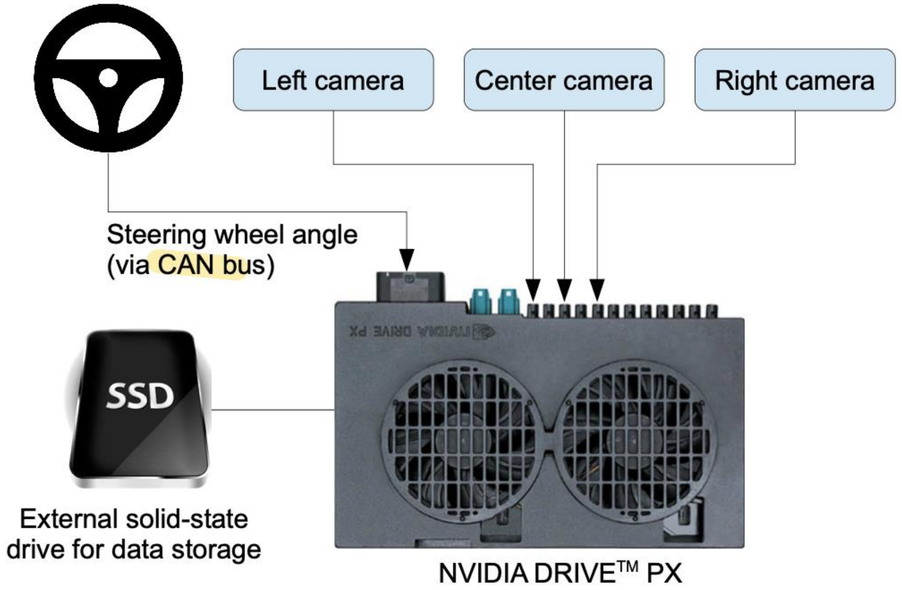

Figure 1: High-level view of the data collection system.

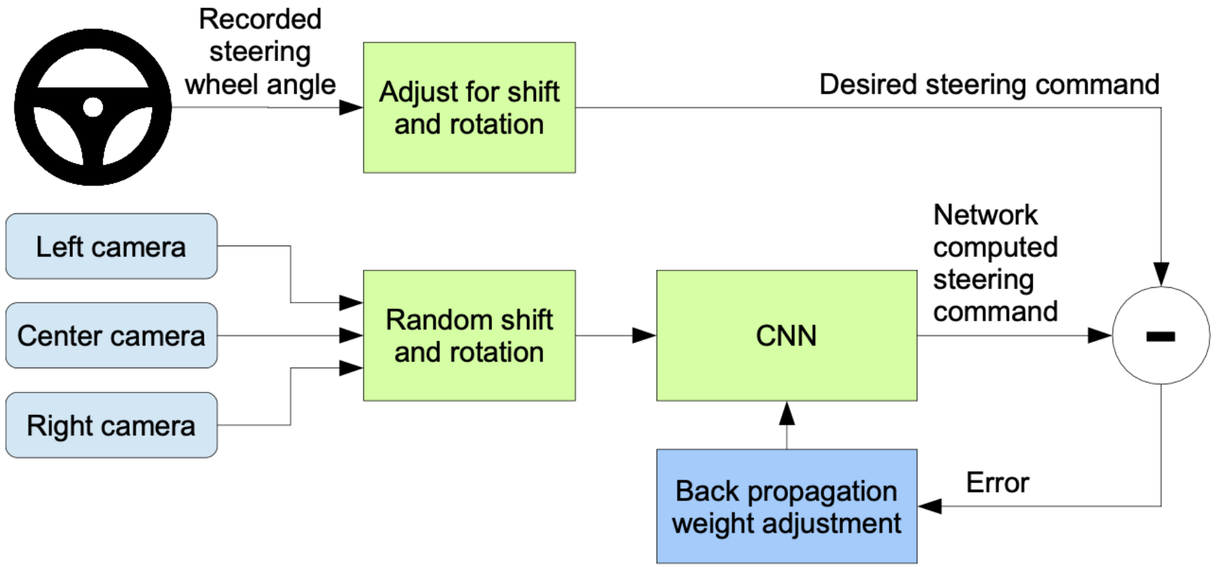

Figure 2: Training the neural network.

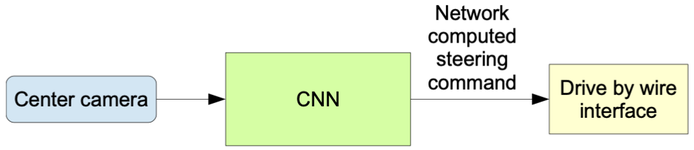

Figure 3: The trained network is used to generate steering commands from a single front-facing center camera.

04 Data Collection

The training data was collected in various road, lighting, and weather conditions, mostly in the United States. It includes snow-covered scenes, day/night driving, and even camera distortions from sunlight. In total, 72 hours of driving data were gathered by different drivers using various car models.

05 Network Architecture

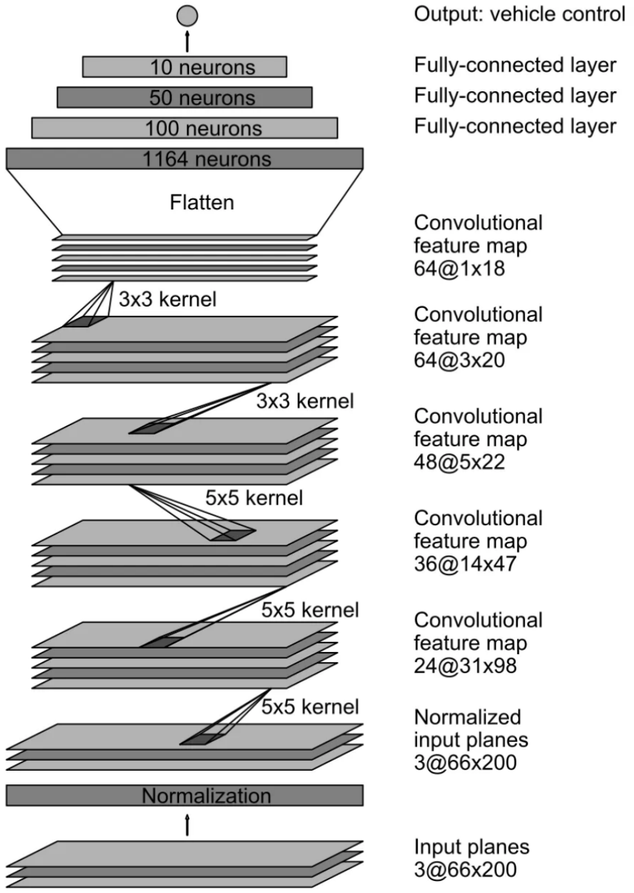

Figure 4: CNN architecture. The network has about 27 million connections and 250 thousand parameters.

The architecture in this paper has nine layers: one normalization layer, five convolutional layers, and three fully connected layers. Images are converted to YUV, which is commonly used by camera sensors and offers better stability for brightness and color separation.

The first three convolutional layers use a stride of (2, 2) and a kernel size of (5, 5) to capture overall features, while the last two use a stride of (1, 1) and a kernel size of (3, 3) for finer details.

Finally, the fully connected layers predict the inverse turning radius (1/r). Because it is an end-to-end model, there is no clear distinction between the parts of the network responsible for feature extraction and final steering prediction.

06 Training Detail

Data Selection

To train the CNN to follow the lane, only frames in which the vehicle is on the lane were used. They chose 10 FPS because higher frame rates produce too many similar images. To reduce bias toward straight roads, they included more frames of curved roads.

Augmentation

They also augmented the data by artificially shifting images and adding frames where the vehicle leaves the lane. These frames were randomly selected based on a normal distribution. Because excessive shifting can result in unrealistic images, the augmentation range must be carefully controlled.

06 simulation

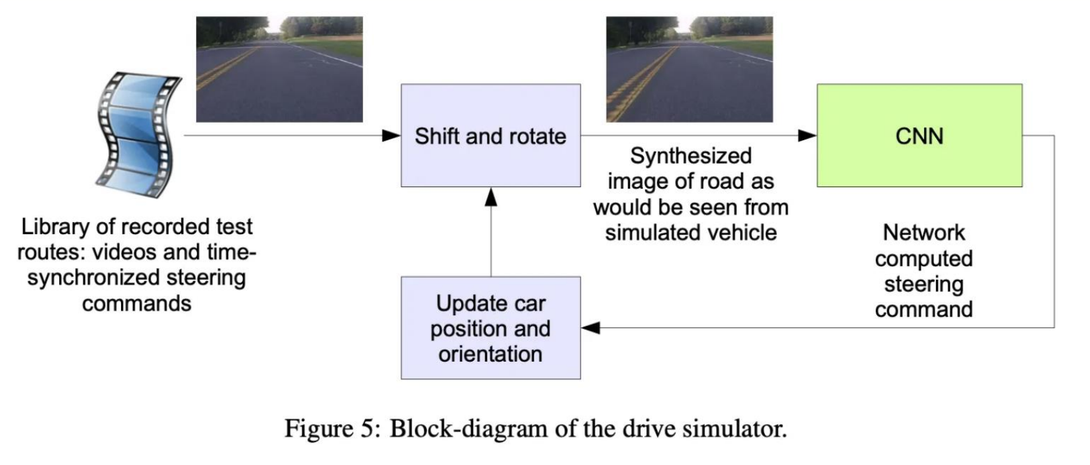

Figure 5: Block-diagram of the drive simulator.

Before testing the trained CNN on a real road, its performance was evaluated in a simulation. A video feed from the front camera (used for data collection) was processed so that, for autonomous driving, the CNN would receive the corresponding “would-be” frames.

Because the driver did not always stay perfectly centered in the lane, all images were transformed to align the vehicle with the lane center.

The transformed image was then fed into the CNN to predict the next frame’s steering. If the CNN’s predicted path deviated by more than 1 meter from the ground truth, a “virtual human” intervened, and the correct image was reintroduced to the system.

This process was repeated continuously.

06 Evaluation

$$\text{autonomy} = (1 - \frac{(\text{number of interventions}) \cdot 6 \text{ seconds}}{\text{elapsed time [seconds]}}) \cdot 100$$

$$(1 - \frac{10 \cdot 6}{600}) \cdot 100 = 90\%$$

Using these metrics, the percentage of self-driving can be calculated. For example, if there are 10 interventions over 600 seconds of driving, the system achieves 90% autonomous driving.

06 Evaluation

VISUALIZATION OF INTERNAL CNN STATE

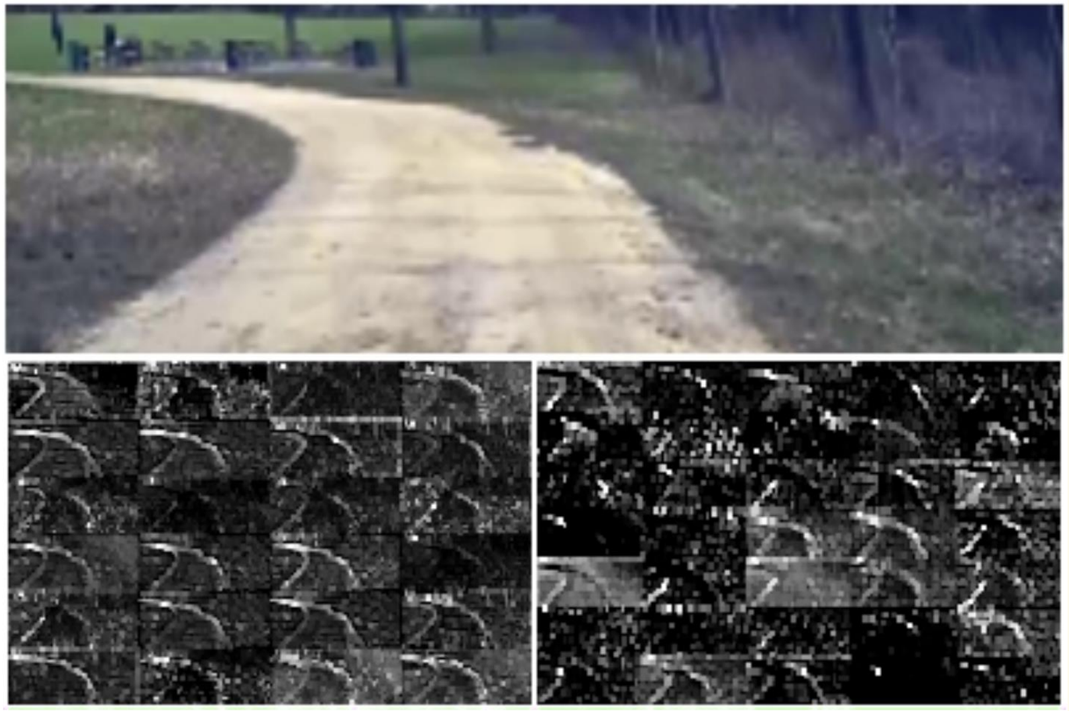

Figure 7: How the CNN "sees" an unpaved road. Top: subset of the camera image sent to the CNN. Bottom left: Activation of the first layer feature maps. Bottom right: Activation of the second layer feature maps. This demonstrates that the CNN learned to detect useful road features on its own, i. e., with only the human steering angle as training signal. We never explicitly trained it to detect the outlines of roads.

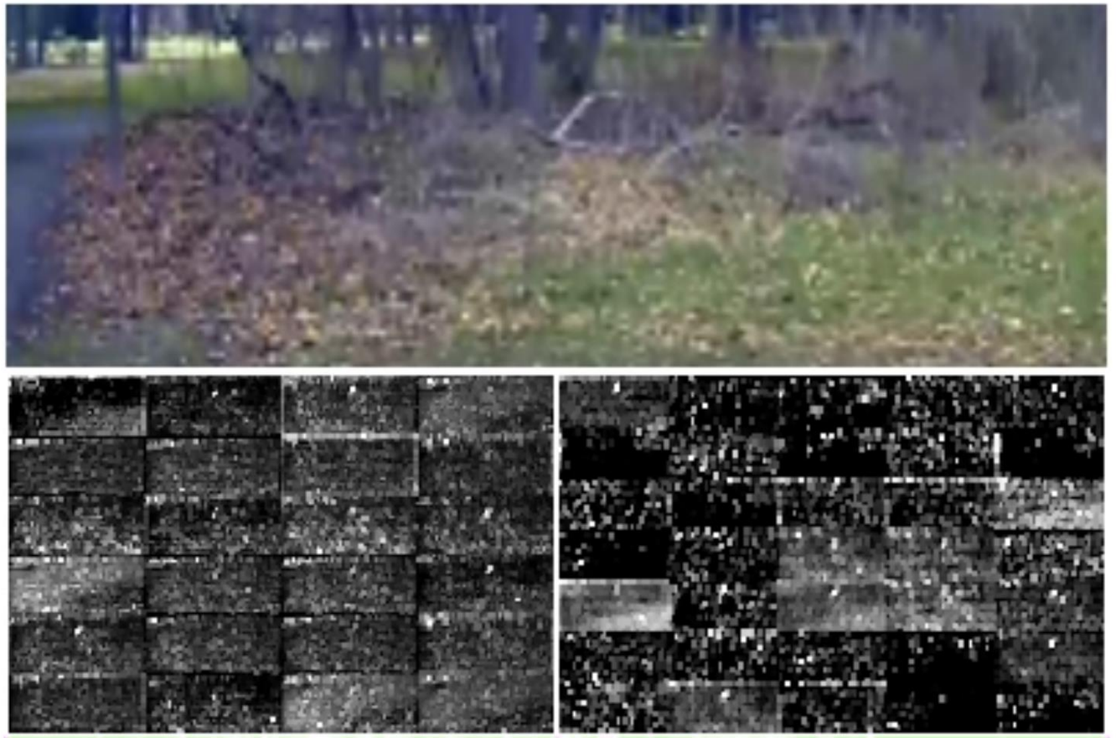

Figure 8: Example image with no road. The activations of the first two feature maps appear to contain mostly noise, i. e., the CNN doesn't recognize any useful features in this image.

07 conclusion

This paper experimentally demonstrated that a CNN can learn the entire task of lane and road driving in an end-to-end manner, without the need for manually decomposing functions like lane marking detection and path planning.

Moreover, it proved that effective training under various road, weather, daytime, and nighttime conditions is possible with less than 100 hours of data.

Thanks