Predicting a vehicles velocity using dashcam footage

Florian Wolf
Department of Mathematics and Statistics, University Konstanz
Konstanz, Germany
florian.2.wolf@uni-konstanz.de

Franz Herbst
Department of Physics, University Konstanz
Konstanz, Germany
franz.herbst@uni-konstanz.de

***Abstract*—In this report bla bla bla**

***Index Terms*—deep learning, computer vision, visual odometry, dense optical flow, siamese network**

I. Introduction

Here are some motivational words needed

A. Aim of the project

Here we need to clarify the aim of the project and the levels of measurement

ASSUMPTIONS ON MEASUREMENT OF THE RESULTS

ASSUMPTION TRAINING ERROR UNDER 3 IS EXTREMLY GOOD

II. Data collection, analysis and preprocessing

For our data set, we used the comma ai speedchallenge[1] data base. This data set provides two dashcam videos: a training video, (20400 frames, shoot at 20 frames per second) including ground truths and a testing video (10798 frames, shoot at 20 frames per second) without labels, which they use for applications to check how well a submitted model is able to generalize. As we only have access to the labels of the test video frames, we decided to split the provided train data by the 80/20 principle into training and testing subsets. Here we did not shuffle the data randomly, as we needed to always have two consecutive frames to calculate the optical flow. We initially used a hard cut off after 80% of the frames, as we wanted to test our model on unseen data, to measure how good our model is able to generalize. We will analyse this naive approach later, when we take a closer look at the results.

A. Data analysis

To analyse the velocity distribution in the two subsets, we plotted the velocity per frame curve in Fig. 1.

The first have of the video mostly represents highway scenarios and the second half only consists of city driving scenes. Therefore we did not expect our model to perform really well using the initial splitting, as the models is mostly trained in highway road traffic scenarios.

[1]https://github.com/commaai/speedchallenge

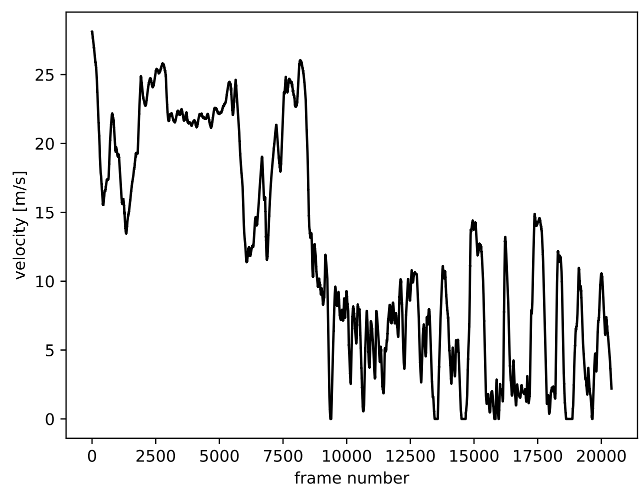

Fig. 1: Distribution of the velocities of all frames in the training video.

B. Preprocessing

Each of the provided frames has a size of $(640, 480, 3)$ pixels. Due to computational limitations, we decided to cut off the last 60 pixels from the lower border, to remove a black frame inside the car, which did not have any effect on the optical flow. Furthermore, we cut the frame size in half and calculated the optical flow using the Farneback pyramid [1] method with the following parameters

$$\begin{aligned} \text{pyramid levels} &:= 3 \\ \text{pyramid scaling} &:= 0.5 \\ \text{window size} &:= 6 \\ \text{pixel neighborhood size} &:= 5 \\ \text{SD of the gaussian filter} &:= 1.1 \end{aligned}$$

We choose three pyramid levels, because we wanted the calculations to be more accurate. To decrease the training duration, we halved the size of the optical flow frames again, resulting in a resolution of $(160, 105, 3)$ pixels per frame. As we used a window size of six pixels, a comparison between the original optical flow and the down sampled one lead to the result, that we do not loose a lot of information.

The optical flow calculation returns the magnitude and the angle of the flow vectors, which we transformed into polar coordinates. To get an RGB image representing the optical flow of two consecutive frames, we normalized the magnitudes and put them into the third channel of the frame. The values of the second channel were all set to the value 255. We then multiplied the angle with the factor $180/(2\pi)$ and set this value for the first channel.

We wanted to see, if the model performs better using the dashcam frames as additional material. Therefore, we did the same down sampling with the frames.

III. Method selection and architecture

The prediction of the vehicles speed is a non-linear regression task , so the choice of a neural network is reasonable. Recent architectures we discussed in the lecture (ResNet, GoogLeNet, etc.) have shown that using multiple stacked convolution layers combined with stacked dense layers, perform well on image classification tasks. Therefore the choice of a convolutional neural network is justified.

As a performance measure, we choose mean squared errors (MSE).

A. *Initial Network*

As an initial architecture we decided to give the model of https://arxiv.org/pdf/1604.07316v1.pdf a try. As the group used it for self-driving cars, the model has enough complexity to handle a task like ours and with the amount of layers, we had a lot of possibilities to fine-tune and improve the model.

HERE SHOULD BE AN IMAGE OF THE NETWORK

In a first approach, we tried the raw model gaining an MSE of around 18-20 on testing and under 3 on training. One can clearly see, that the model really overfits the train dataset.

IV. Tuning of the model

To speed up the training we used batch normalization layers https://arxiv.org/pdf/1502.03167.pdf and we tested different activation functions, to improve the performance of the model.

As proposed in the lecture, we used

$$\mathrm{ReLu}: \mathbb{R} \to \mathbb{R}_0^+, x \mapsto \max\{0, x\}$$

as the initial activation function. Using the ReLU function and 15 epochs for training, we achieved a MSE of around 15 on the testing set. We ran the code multiple times, to ensure this result holds. This result was not really promising, so we wanted to decrease the error by modifying the model even more.

As we still had a lot of issues with overfits, we decided to include a dropout layer, according to https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf. We tried different numbers and positions for dropout layers, but using one layer with a dropout probability of $p = 0.5$ after the third concolutional layer seemed to work best.

To solve the problem of dead neurons[2] of the ReLU function, we tried the leakyReLU function

$$\mathrm{leakyReLU}: \mathbb{R} \to \mathbb{R}, x \mapsto \begin{cases} x, x \geq 0 \\ c \cdot x, x < 0 \end{cases}$$

with a parameter $c = 0.01$. Using the leakyReLU function, we achieved a MSE of around 11 on the testing set and under 3 on the training set. All of the models were trained with only eight epochs, as we still had problems with overfits.

We identified three possible reasons for our poor results:

(i) Too complex model, as the paper used it for autonomous driving or we put too little information into the model.
(ii) Problems with different brightnesses in the frames (lack of generality), which leads to unstable calculations in the optical flow, as the optical flow is quite sensitive to brightness. (Explain how in the train part the sky is quite dark and in the city (end) the sky is bright)
(iii) Too naive/ambigous splitting of the data into train and testing set, as both datasets seem to represent totally different scenarios in the road traffic.

We came up with the following approaches to solve these problems

(i) Simplify the model by using Pooling (we will try average and max pooling), to get more compression and we tried on the other hand to feed in more information into the model, by using a linear combination of the optical flow and the raw frames itself, and we tried using a siamese network, to put simultaneously the of and raw frames into the convolutional layers.
(ii) We wanted to try adding some additional noise into the frames before calculating the optical flow, to make the calculation more robust against brightness changes. As intentionally adding noise to a frame is quite atypical in computer vision, this idea looked quite interesting.
(iii) Use another splitting. To get a better ratio between highway and city driving scenarios, we decided to split the data into blocks of 100 frames and take the first 80 for training and the last 20 for testing. Therefore our model should have seen some city driving.

A. *Pooling layers with initial splitting*

We added two generic pooling layers to the network. One after the second convolutional layer and the second one before the fully connected layers start. We tested maximum and average pooling with the following parameters

$$\begin{aligned} \text{kernel size} &:= 2 \times 2 \\ \text{stride} &:= 2 \\ \text{padding} &:= 1 \\ \text{dilatation} &:= \text{None} \end{aligned}$$

The implementation of pooling layers helped a lot, as now the loss on the train and test data seem to decrease nearly equally.

[2]As one can clearly see in the defintion of the ReLU function, neurons with a value below zero cannot participate in the learning process.

Initial splitting, dropout layer	ReLU		leakyReLU	
	Train	Test	Train	Test
No pooling	2.85	12.08	2.45	10.75
Max pooling	5.62	11.82	5.52	10.29
Max pooling (15 epochs)	-	-	**3.22**	**9.63**
Average pooling	7.70	11.40	6.08	13.09

TABLE I: MSE results of the network using pooling and dropout layers, different activation functions and the initial splitting.

Our results on the initial splitting are shown in Table I. We gave the network with max pooling a try with 15 epochs, as the loss on the train and test set was decreasing in a pretty stable manner. Using this network, we achieved a MSE of under 10 on the test set, which is according to the companies guidelines a pretty good result. Especially, as we trained the model mostly on highway scenes and tested it only in city driving scenarios.

B. New splitting and siamese network

C. Augmented brightness

V. New Approach using siamese network for two consecutive frames

VI. Error analysis and results

VII. Further work

Use MUCH MUCH MUCH MUCH MUCH more data!

Create Validation Data that is not connected to training data and still covers all situations

Acknowledgment

We like to thank bla bla bla

References

[1] Gunnar Farnebäck. “Two-Frame Motion Estimation Based on Polynomial Expansion”. In: *Scandinavian Conference on Image Analysis* (2003), pp. 363–370.