Brain4Cars: Sensory-Fusion Recurrent Neural Models for Driver Activity Anticipation

Ashesh Jain, Shane Soh, Bharad Raghavan, Avi Singh, Hema S Koppula, and Ashutosh Saxena
Department of Computer Science, Stanford University.
Department of Computer Science, Cornell University.
{ashesh,shanesoh,bharadr,hema,asaxena}@cs.stanford.edu

Abstract

Advanced Driver Assistance Systems (ADAS) have made driving safer over the last decade. They prepare vehicles for unsafe road conditions and alert drivers if they perform a dangerous maneuver. However, many accidents are unavoidable because by the time drivers are alerted, it is already too late. Anticipating maneuvers beforehand can alert drivers before they perform the maneuver and also give ADAS more time to avoid or prepare for the danger.

In this work we anticipate driving maneuvers several seconds before they occur [2]. For this purpose we build a vehicular-sensory platform with multiple cameras, tactile sensors, GPS, and a computing device. Our sensors capture the driving context from both inside and outside of the car. We introduce a sensory-fusion Recurrent Neural Network (RNN) architecture which jointly learns to anticipate and fuse information from multiple sensory streams. Our architecture use Long-Short Term Memory (LSTM) units to capture long temporal dependencies. We evaluate our approach on a diverse data set with **1180 miles** of natural free-way and city driving and show that we can anticipate maneuvers **3.5 seconds** before they occur with over 80% F1-score in real-time.[1]

1 Introduction

Over the last decade cars have been equipped with various assistive technologies in order to provide a safe driving experience. Technologies such as lane keeping, blind spot check, pre-crash systems etc., are successful in alerting drivers whenever they commit a dangerous maneuver. Still in the US alone more than 33,000 people die in road accidents every year, the majority of which are due to inappropriate maneuvers [1]. We need mechanisms that can alert drivers *before* they perform a dangerous maneuver in order to avert many such accidents [3]. In this work we address this problem of anticipating maneuvers that a driver is likely to perform in the next few seconds (Figure 1).

In order to anticipate maneuvers, we reason with the contextual information from the surrounding events, which we refer to as the *driving context*. We obtain this driving context from our vehicular-sensory platform which includes multiple cameras, tactile sensors (on steering wheel and brakes), global positioning system (GPS), the vehicle's dynamics, and street maps. The challenge lies in correctly fusing these multiple sensory-streams, and handling long temporal dependencies in driving.

We propose a Recurrent Neural Network (RNN) based architecture which learns rich representations for anticipation. Our architecture learns how to optimally fuse information from different sensors and uses Long Short-Term Memory (LSTM) units to capture temporal dependencies. We train our architecture in a sequence-to-sequence prediction manner such that it implicitly learns to anticipate given only partial context, and introduce a novel loss layer which helps anticipation by preventing over-fitting. We evaluate our approach on a driving data set with 1180 miles of natural free-way and city driving collected across two states – from 10 drivers and with different kinds of driving maneuvers. We demonstrate that our approach anticipates maneuvers 3.5 seconds before they occur with 80% precision and recall. We believe that our work creates scope for new ADAS features to make roads safer. Our code and data set are available here: http://www.brain4cars.com

[1] ashesh@cs.stanford.edu is the corresponding author.

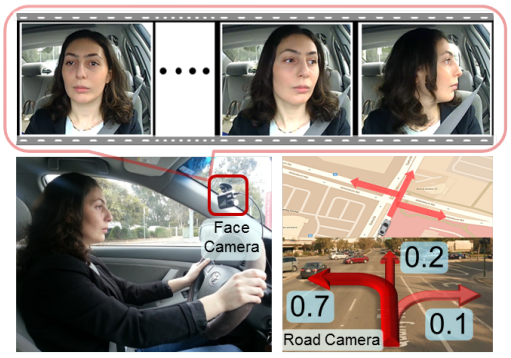

Figure 1: **Anticipating maneuvers.** Our algorithm anticipates driving maneuvers performed a few seconds in the future. It uses information from multiple sources including videos, vehicle dynamics, GPS, tactile sensors, and street maps to anticipate the probability of different future maneuvers.

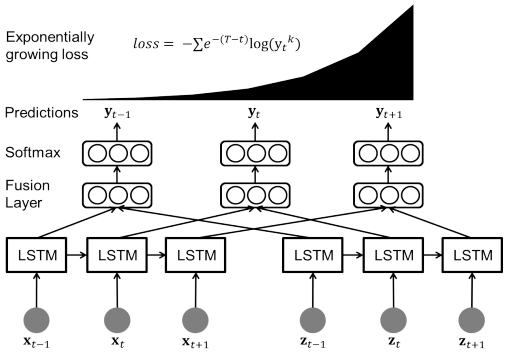

Figure 2: **Sensory fusion RNN for anticipation.** (**Bottom**) In Fusion-RNN each sensory stream is passed through their independent RNN. (**Middle**) High-level representations from RNNs are then combined through a fusion layer. (**Top**) In order to prevent over-fitting early in time the loss exponentially increases with time.

1.1 Sensory-Fusion RNN with LSTM units for anticipation

In order to anticipate, an algorithm should learn to predict the future only based on partial temporal context. At training time, temporal sequences of context $\mathbf{x}$ and events y $\{(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)_j, y_j\}_{j=1}^N$ are provided. At test time, however, the goal is to predict the future event as early as possible, i.e. by observing only a partial sequence of context $\{(\mathbf{x}_1, ..., \mathbf{x}_t)|t < T\}$. We train RNN for anticipation by mapping the sequence of context $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$ to the sequence of events $(y_1, ..., y_T)$ such that $y_t = y, \forall t$. Our RNN trained in this manner attempts to map all sequences of partial context $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t)$ $\forall t \leq T$ to the future event y. This way our model implicitly learns to anticipate.

An obvious way to allow sensory fusion in RNN is by concatenating the streams. However, we found that this sort of simple concatenation performs poorly. We instead learn a sensory fusion layer which combines the high-level representations of sensor data (Figure 2). Our proposed architecture first passes the sensory streams independently through separate RNNs. The high level representations from RNNs are then concatenated at each time step, and passed through a fully connected layer which fuses the representations. This recurrent operations is performed from $t = 1$ to T.

Since the model is trained for anticipation it should be encouraged to anticipate early. We therefore propose a new scheme under which the architecture suffers a loss of $-e^{-(T-t)}\log(y_t)$ at each t. This loss penalizes the architecture exponentially more for the mistakes it makes as it sees more context. This encourages the model to fix mistakes as early as it can in time. It also penalizes the network less on mistakes made early in time when there is not enough context available. This way it acts like a regularizer and reduces the risk to over-fit very early in time.

1.2 Experiments

Our data set consists of 1180 miles of free-way and city driving from 10 drivers. We anticipate maneuvers every 0.8 seconds where the algorithm processes the recent context and assigns a probability to each of the five maneuvers: {*left lane change, right lane change, left turn, right turn, driving straight*}. On the holdout set our sensory-fusion RNN anticipates manuevers 3.58 seconds before they happen with 84.5% precision and 77.1% recall. While a random guess will only give 20% precision and recall. More details are available here: `http://www.brain4cars.com`

References

[1] 2012 motor vehicle crashes: overview. *N. Highway Traffic Safety Administration, Washington, D.C., Tech. Rep.*, 2013.

[2] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *ICCV (To appear)*, 2015.

[3] T. Rueda-Domingo, P. Lardelli-Claret, J. Luna del Castillo, J. Jimenez-Moleon, M. Garcia-Martin, and A. Bueno-Cavanillas. The influence of passengers on the risk of the driver causing a car collision in spain: Analysis of collisions from 1990 to 1999. *Accident Analysis & Prevention*, 2004.