Paper_ID: 1047 in **IEEE ICASSP 2021**

Fusion Neural Network for Vehicle Trajectory Prediction in Autonomous Driving

Jue Wang[1,2], Ping Wang[1], Chao Zhang[1], Kuifeng Su[2], Jun Li[3]

[1]Peking University, Beijing, China,
[2]Tencent Technology (Beijing) Company Limited, China,
[3]University of Chinese Academy of Sciences, Beijing, China

Introduction

- Predicting future interactive traffic behaviors becomes the bottleneck of safety and comfort solving problems for self-driving vehicles. A vehicle trajectory prediction model designed for the usage of mobile platform on autonomous driving cars is actually required.
- In this paper, we propose a robust and efficient combined architecture with recurrent neural networks and convolutional neural networks named F-Net to deal with vehicle trajectory prediction for autonomous driving application.

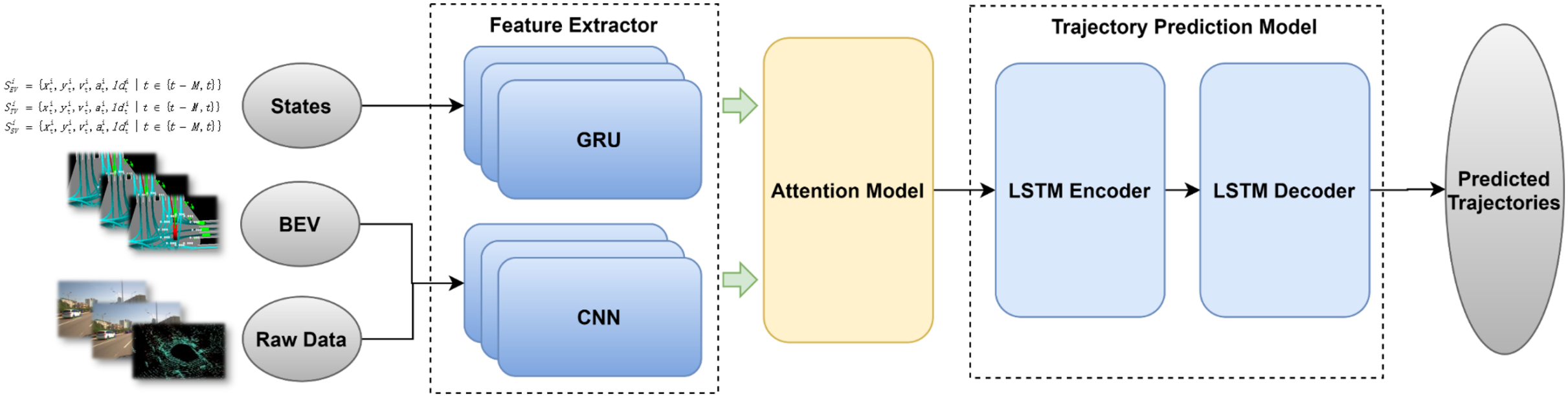
Problem Formulation

- The prediction for the future trajectories of multiple vehicles in driving scenarios could be formulated as a probabilistic optimization problem to compute the conditional distribution $P(O \mid I)$, where the future trajectories of multiple vehicles are $O = \{O_1, O_2, \dots, O_N\}$, their input are $I = \{I_1, I_2, \dots, I_N\}$, and N is the number of predicted vehicles. The future trajectory of predicted vehicle i is defined as $O_i = \{O_{t+1}^i, O_{t+2}^i, \dots, O_{t+L}^i\}$
- Here we extend more sufficient input information I . For predicted vehicle i , the corresponding input I_i includes the following different information, as shown in $I_i = \{S_{TV}^i, S_{SV}^i, E_{BEV}^i, R_i\}$. S_{TV}^i is for the history and current states of target predicted vehicles (TV), S_{SV}^i is for the history and current states of surrounding vehicles (SV), and E_{BEV}^i is for Bird's Eye View (BEV) of environmental information, R_i is for raw sensor data.

Proposed Model

- High uncertainty of traffic behavior and large number of situations that a self-driving car may encounter on roads, makes it very difficult to predict future traffic behaviors.
- To solve the existing prediction problems, our proposed F-Net architecture consists of three components: Feature Extractor with GRUs and CNNs, Attention Mechanism and LSTM-based Trajectory Prediction Model, as illustrated in the Figure Architecture.

Architecture



Feature Extractor

- Our feature extractor with GRUs and CNNs extracts both temporal and spatial features from the static and dynamic objects states, BEV and raw sensor data of the entire environmental and interaction scene.
- Because of the temporal feature extracting power of RNNs, we use GRU encoder to learn the history and current states of predicted target vehicles and surrounding vehicles, and we have the temporal features as shown in $F^{\mathrm{i}}_{Temporal} = GRU(S^i_{TV}, S^i_{SV})$.
- Because of the spatial feature extracting ability of CNNs, we use CNNs to learn BEV of environmental information and raw sensor data, and we have the spatial features as shown in $F^{\mathrm{i}}_{Spatial} = CNN(E^i_{BEV}, R_i)$.
- In this paper, we use MobileNetV3 as $CNN(.)$ with the initial training weights on ImageNet and fine-tuning on the nuScenes dataset.

Attention Mechanism

- Motivated by that people always pay more attention to nearby static and dynamic obstacles, upcoming road structure and traffic signals when driving, we use two attention mechanisms to focus on the salient and more relevant parts in the progress of both GRU and CNN feature extraction, as shown in $F^i_{Att} = ATT(F^i_{Temporal}, F^i_{Spatial})$.
- The inputs to the two attention mechanisms are separately the hidden states of the GRU which has the information for predicting the vehicles' future trajectories interacted with each other, and the spatial visual features extracted from the CNN which includes the rich scene texture information of sensor raw data.

LSTM-based Trajectory Prediction Model

- We use an LSTM encoder-decoder network that takes the output of attention mechanisms to generate the future trajectories.
- Our LSTM encoder is a two-layer LSTM, the input of the encoder has 512 channels that are the output dimension of attention mechanisms at each time step which have extracted more relevant features.
- Our LSTM decoder is a two-layer LSTM, and it is fed by the hidden feature of the encoder LSTM, concatenated with coordinates of objects at the previous and current time step, to predict the future trajectories.

Experiments Results

- We compare its performance with different methods on the nuScenes dataset. The qualitative results could show the effectiveness of our proposed F-Net.

Table 1 The Average Displacement Error (ADE) comparison results of our proposed F-Net with the following baseline and some existing state-of-art methods.

Method	1s	2s	3s	4s	5s
CV	0.71	1.79	3.09	4.75	6.65
V-LSTM	0.68	1.35	2.16	3.51	4.75
CS-LSTM	0.63	1.28	2.07	3.2	4.36
ST-LSTM	**0.55**	1.21	1.94	2.81	3.78
F-Net(Proposed)	0.67	**1.16**	**1.7**	**2.59**	**3.57**

Table 2 The computation frequency (FPS) comparison results of our proposed F-Net comparing with the state-of-art CS-LSTM method.

Method	Predicted#	FPS 128 batch	FPS 1 batch
CS-LSTM	1000	3.5	0.03
F-Net(Proposed)	1000	**15**	**0.2**

Thanks