Convolutional Social Pooling for Vehicle Trajectory Prediction

Authors: Nachiket Deo, Mohan M. Trivedi
University of California, San Diego
La Jolla, 92093

Presented by: Hafiza Rimsha Rauf
Objectives

• The purpose of this paper is to predict multi model future trajectories of a vehicle using track histories
Challenges:

- **Multimodal Behavior**
  
  Drivers can take multiple decision under the same traffic circumstances

- **Traffic Context**
  
  Interaction between different vehicles affect the motion

- **Driving style**
  
  Variation of driving styles across different drivers
Problem Formulation

**Maneuvers Classes**
- 3 *Lateral Maneuvers Classes*
  - Right Lane Change
  - Left Lane Change
  - Keep Same Lane
- 2 *Longitudinal Maneuver Classes*
  - Normal Driving
  - Applying Brakes
- Overall six possible combinations of maneuver classes
Problem Formulation

- **Stationary Frame of Reference**
  
  Origin fixed at the predicted vehicle at time $t$
Problem Formulation

- **Inputs**
  The input to our model are track histories over \( t_n \) number of frames

\[
X = [x^{(t-t_n)}, \ldots, x^{(t-1)}, x^{(t)}]
\]

\[
x^{(t)} = [x_0^{(t)}, y_0^{(t)}, x_1^{(t)}, y_1^{(t)}, \ldots, x_n^{(t)}, y_n^{(t)}]
\]

- **Outputs**
  The output of the model is a probability distribution over

\[
Y = [y^{(t+1)}, \ldots, y^{(t+t_f)}]
\]

\[
y^{(t)} = [x_0^{(t)}, y_0^{(t)}]
\]
Problem Formulation

- **Probabilistic Motion Prediction**
  Our model estimates the conditional distribution $P(Y|X)$. In order to have the model produce multi-modal distributions, we expand it in terms of maneuvers $m_i$, giving:

  \[
P(Y|X) = \sum_i P_\Theta(Y|m_i, X)P(m_i|X)
  \]

  where,

  \[
  \Theta = [\Theta^{(t+1)}, ..., \Theta^{(t+t_f)}]
  \]

- **Loss Function**

  \[
  - \log \left( \sum_i P_\Theta(Y|m_i, X)P(m_i|X) \right)
  \]
Methodology

**LSTM Encoder**
Capture motion dynamics of all the vehicles

**Convolutional Social Pooling**
Infers motion dependencies of all vehicles in the scene based on spatial configuration

**LSTM Decoder**
- Probabilistic distributions for future motion corresponding to each maneuver
- Probability of each maneuver class
Model
Model

Social Tensor
- Spatial Grid (13, 3) defined around predicted vehicle populated with LSTM states of surrounding agents based on spatial configuration
- Each column corresponds to one lane
- Each row is separated by 15 feet distance
Model
Implementation Details

- **LSTM Encoder**
  (64 Dimensional states)

- **Convolutional Social Pooling**
  2 3x3 Convolutional Layers, 1 Max pooling Layer

- **Maneuver Based LSTM Decoder**
  128 Dimensional states

- Learning Rate: 0.001
- Activation: Leaky Relu
- Optimization: Adam
Implementation Details

- **Evaluation Metric**

  Root Mean Square Error (RMSE)
  
  i. RMSE between true trajectory and future trajectory (highest probable)
  ii. Skewed in the favor of the models that average the modes

  Negative Log Likelihood
  
  i. For comparison of uni-modal and multi-modal predictive distribution

- **Datasets**

  i. NGSIM US-101
  ii. I-80
Baseline Models

- **Constant Velocity (CV)**
  - No Information about neighboring vehicle

- **Vanilla LSTM (V-LSTM)**
  - No Information about neighboring vehicle

- **LSTM with fully connected social pooling (S-LSTM)**
  - Destroys spatial structure of social tensor

- **LSTM with convolutional social pooling (CS-LSTM)**
  - Uni modal

- **LSTM with convolutional social pooling and maneuvers (CS-LSTM(M))**
  - Multi modal
## Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0.73</td>
<td>0.66</td>
<td>0.69</td>
<td>0.68</td>
<td>0.65</td>
<td><strong>0.61</strong></td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.78</td>
<td>1.56</td>
<td>1.51</td>
<td>1.65</td>
<td>1.31</td>
<td><strong>1.27</strong></td>
<td>1.29</td>
</tr>
<tr>
<td>RMSE (m)</td>
<td>3</td>
<td>3.13</td>
<td>2.75</td>
<td>2.55</td>
<td>2.91</td>
<td>2.16</td>
<td><strong>2.09</strong></td>
<td>2.13</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4.78</td>
<td>4.24</td>
<td>3.65</td>
<td>4.46</td>
<td>3.25</td>
<td><strong>3.10</strong></td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6.68</td>
<td>5.99</td>
<td>4.71</td>
<td>6.27</td>
<td>4.55</td>
<td><strong>4.37</strong></td>
<td>4.52</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3.72</td>
<td>2.02</td>
<td>-</td>
<td>1.17</td>
<td>1.01</td>
<td>0.89</td>
<td><strong>0.58</strong></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5.37</td>
<td>3.63</td>
<td>-</td>
<td>2.85</td>
<td>2.49</td>
<td>2.43</td>
<td><strong>2.14</strong></td>
</tr>
<tr>
<td>NLL</td>
<td>3</td>
<td>6.40</td>
<td>4.62</td>
<td>-</td>
<td>3.80</td>
<td>3.36</td>
<td>3.30</td>
<td><strong>3.03</strong></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7.16</td>
<td>5.35</td>
<td>-</td>
<td>4.48</td>
<td>4.01</td>
<td>3.97</td>
<td><strong>3.68</strong></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>7.76</td>
<td>5.93</td>
<td>-</td>
<td>4.99</td>
<td>4.54</td>
<td>4.51</td>
<td><strong>4.22</strong></td>
</tr>
</tbody>
</table>
Analysis

- Uni-modal vs. multi-modal predictions
Analysis

uni-modal
(tends to learn average)

vs.

multi-modal
(all possible trajectories)
Analysis

- Fully Connected vs. Convolutional Social Pooling
Analysis

- Effect of surrounding vehicles on predictions
Advantages

- Improvement in Prediction Accuracy
  Improvement over the state of the art in terms of negative log-likelihoods of true future trajectories under the model’s predictive distribution

- Multi Modal Future Trajectory Prediction
  Multi-modal predictive distribution for future motion of vehicle based on different maneuvers

- Robustly Models Traffic Context
  Models interactions between different vehicles based on spatial configurations in a robust and generalized way
Disadvantages

- Does not take into account the effect of other agents on the road for trajectory prediction i.e. pedestrian

- High computational complexity for generating parameters of bivariate gaussian distribution corresponding to each maneuver class
Takeaways

- Vehicle dynamics can be learned using **LSTM encoders**

- **Social tensors** are helpful to preserve spatial configuration of the road scene

- Traffic context can be captured using **convolutional layers**
Thank You
Quiz

What is the benefit of using social tensor? Would its configuration be changed for a different road scenario? If yes, how?