Peeking into the Future: Predicting Future Person Activities and Locations in Videos

Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander Hauptmann, Li Fei-Fei

Seminar Computer Vision
10th of July 2019
Naya Baslan
Supervisor: Osama Makansi
Outline

- Motivation
- Related Work (for trajectory and activity prediction)
- Approach: Next Model
  - Person Behavior Module
  - Person Interaction Module
  - Trajectory Generator with Focal Attention
  - Activity Prediction
- Experiments: ActEV/VIRAT and ETH & UCY
- Conclusion
Motivation

Main Contribution of this Paper:
Path Prediction  Activity Prediction

joint task

Results in better prediction of future path and activity

Figure 1: Introduction
Outline

- Introduction

- Related Work (for trajectory and activity prediction)
  - Approach: Next Model
    - Person Behavior Module
    - Person Interaction Module
    - Trajectory Generator with Focal Attention
    - Activity Prediction
  - Experiments: ActEV/VIRAT and ETH & UCY
  - Conclusion
Related Work for Trajectory Prediction

**Person-person models:**

- Human Interactions and Behaviour
- Trajectory Prediction

examples:
Social-LSTM and Social-GAN

- Allows sharing of hidden states between different LSTMs

Source: Social LSTM Human Trajectory
Related Work for Trajectory Prediction

Person-scene models:

- Models the interaction between the person and the surrounding

Examples:
- Scene-LSTM, Car-Net and SoPhie

Source: Scene-LSTM: A Model for Human Trajectory Prediction

• Models the interaction between the person and the surrounding
• Using LSTMs to recognize human activities
• In this paper: both activity and trajectories are jointly predicted using rich visual features and focal attention

Source: Learning Activity Progression in LSTMs for Activity Detection and Early Detection
Outline

- Introduction
- Related Work (for trajectory and activity prediction)
- Approach: Next Model
  - Person Behavior Module
  - Person Interaction Module
  - Trajectory Generator with Focal Attention
  - Activity Prediction
- Experiments: ActEV/VIRAT and ETH & UCY
- Conclusion
Approach:

- Next: Multi-task learning model used to predict path and activity.

Problem Formulation:

<table>
<thead>
<tr>
<th>1 to $T_{obs}$</th>
<th>$T_{obs+1}$ to $T_{pred}$</th>
<th>$T_{pred}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>make observations</td>
<td>predict positions of person</td>
<td>predict activity</td>
</tr>
</tbody>
</table>

- Trajectory prediction from $T_{obs+1}$ to $T_{pred}$.
- Activity prediction at $T_{pred}$.
Next Model: Framework

**Figure 2: Model Overview**

- Person Behavior Module
- Person Interaction Module
- Trajectory Generator
- Activity Prediction
This module has 2 sub-modules:
- Person Appearance Encoder (uses a pre-trained object detection model)
  Source: J. Liang et al.
- Person Keypoint Encoder (uses a pre-trained person keypoint model)
  Source: H.-S. Fang et al.

(both related to visual information about every person in the scene)
- Model person-scene interactions
- Model person-object interactions
Person Interaction Module: Person-Object Encoder

 observes person
 \((x_b, y_b, w_b, h_b)\)

 observes object
 \((x_k, y_k, w_k, h_k)\)

\(k \in [1, K]\)

- compute the geometric relation \(G\) where

\[
G_k = \left[ \log\left( \frac{|x_b - x_k|}{w_b} \right), \log\left( \frac{|y_b - y_k|}{h_b} \right), \log\left( \frac{w_k}{w_b} \right), \log\left( \frac{h_k}{h_b} \right) \right]
\]

- geometric distance
- fraction box size
Visual Features

Person Appearance Features
Person Keypoint Features
Person-Object Features
Person-Scene Features

Tensor Q contains the hidden states of all LSTM encoders

5 x T_{obs} x d

Visual Feature Tensor Q

Trajectory embedding: 
\[ e_{t-1} = \tanh \{ W_e [x_{t-1}, y_{t-1}] \} + b_e \in \mathbb{R}^d \]
(fully connected layer with activation function $\tanh$)

The trajectory embedding function is obtained from the trajectory output at the last time instant and is fed into an LSTM encoder.
The hidden states of all 5 LSTMs are packed into a tensor.
Input feature vectors are all passed into a 3D tensor. Then multi-task prediction is done (simultaneously predicting both the trajectory and the future activity).
Trajectory Generator

Without Focal Attention:

LSTM decoder: \[ h_t = \text{LSTM} \left( h_{t-1}, [e_{t-1}] \right) \]

With Focal Attention:

LSTM decoder: \[ h_t = \text{LSTM} \left( h_{t-1}, [e_{t-1}, \tilde{q}_t] \right) \]

Focal attention identifies what content of the sequential data needs to be focused on to make correct predictions.
In summary: focal attention models the correlation among different features and summarizes them into a lower dimensional feature vector.
Main Task: activity label prediction

Auxiliary Task: activity label prediction

Purpose of auxiliary task:
- to bridge the gap between trajectory generation and activity label prediction
- to minimize the errors in trajectory that have been accumulating over time
Activity Location Prediction:

- predicts the final location of the person's activity

Manhattan Grid

$h \times w$
• **Activity Label Prediction:**
  • Predicts the person's future activity at time $T_{pred}$

\[
\text{cls}_{act} = \text{softmax}(W_a \cdot [Q_{1T_{obs}}, \cdots, Q_{MT_{obs}}])
\]
Training: (3 Losses)

Total Loss: \[ L = L_{xy} + \lambda (L_{grid\_cls} + L_{grid\_reg}) + L_{act} \]

- Trajectory Prediction Loss
- Activity Location Prediction Loss
- Activity Label Prediction Loss
Outline

- Introduction
- Related Work (for trajectory and activity prediction)
- Approach: Next Model
  - Person Behavior Module
  - Person Interaction Module
  - Trajectory Generator with Focal Attention
  - Activity Prediction
- Experiments: ActEV/VIRAT and ETH & UCY
- Conclusion
Public dataset for activity detection released by NIST in 2018

Metrics used for trajectory prediction:

**Average Displacement Error (ADE)**

\[
ADE = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_{pred}} \| \tilde{Y}_t^i - Y_t^i \|_2}{N \times T_{pred}}
\]

**Final Displacement Error (FDE)**

\[
FDE = \frac{\sum_{i=1}^{N} \| \tilde{Y}_{T_{pred}}^i - Y_{T_{pred}}^i \|_2}{N}
\]
### Comparison to Baseline Methods Quantitative Results

<table>
<thead>
<tr>
<th>Method</th>
<th>ADE</th>
<th>FDE</th>
<th>move_ADE</th>
<th>move_FDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>32.19</td>
<td>60.92</td>
<td>42.82</td>
<td>80.18</td>
</tr>
<tr>
<td>LSTM</td>
<td>23.98</td>
<td>44.97</td>
<td>30.55</td>
<td>56.25</td>
</tr>
<tr>
<td>Social LSTM</td>
<td>23.10</td>
<td>44.27</td>
<td>28.59</td>
<td>53.75</td>
</tr>
<tr>
<td>SGAN-PV</td>
<td>30.51</td>
<td>60.90</td>
<td>37.65</td>
<td>73.01</td>
</tr>
<tr>
<td>SGAN-V</td>
<td>30.48</td>
<td>62.17</td>
<td>35.41</td>
<td>68.77</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>17.99</strong></td>
<td><strong>37.24</strong></td>
<td><strong>20.34</strong></td>
<td><strong>42.54</strong></td>
</tr>
<tr>
<td><strong>Ours-Noisy</strong></td>
<td>34.32</td>
<td>57.04</td>
<td>40.33</td>
<td>66.73</td>
</tr>
<tr>
<td>SGAN-PV-20</td>
<td>23.11</td>
<td>41.81</td>
<td>29.80</td>
<td>53.04</td>
</tr>
<tr>
<td>SGAN-V-20</td>
<td>21.16</td>
<td>38.05</td>
<td>26.97</td>
<td>47.57</td>
</tr>
<tr>
<td><strong>Ours-20</strong></td>
<td><strong>16.00</strong></td>
<td><strong>32.99</strong></td>
<td><strong>17.97</strong></td>
<td><strong>37.28</strong></td>
</tr>
</tbody>
</table>

- This method outperforms all other baseline methods
- Experiments are divided into static and moving scenes
- Best performance is achieved when repeated 20 times
Comparison to Baseline Methods Qualitative Results

- Next model results in better trajectory predictions for all 3 people.
- In addition to that, it also predicts future activities.
Further Qualitative Results - Next Model

- Successful case is where both future trajectory and activity are predicted correctly
- Imperfect case predicts the correct trajectory but incorrect activity
- Failed case predicts both trajectory and activity incorrectly
• Few frames are observed and then predictions are made on both the future activity and trajectory
### Ablation Studies

<table>
<thead>
<tr>
<th>Method</th>
<th>ADE ↓</th>
<th>FDE ↓</th>
<th>Act mAP ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our full model</td>
<td>17.91</td>
<td>37.11</td>
<td>0.192</td>
</tr>
<tr>
<td>No p-behavior</td>
<td>18.99</td>
<td>39.82</td>
<td>0.139</td>
</tr>
<tr>
<td>No p-interaction</td>
<td>18.83</td>
<td>39.35</td>
<td>0.163</td>
</tr>
<tr>
<td>No focal attention</td>
<td>19.93</td>
<td>42.08</td>
<td>0.144</td>
</tr>
<tr>
<td>No act label loss</td>
<td>19.48</td>
<td>41.45</td>
<td>-</td>
</tr>
<tr>
<td>No act location loss</td>
<td>19.07</td>
<td>39.91</td>
<td>0.152</td>
</tr>
<tr>
<td>No multi-task</td>
<td>20.37</td>
<td>42.79</td>
<td>-</td>
</tr>
</tbody>
</table>

- Person behavior features are more essential
- Worse results if focal attention is removed
- Worse results if multi-task learning is removed
- These datasets are for predicting person trajectory

<table>
<thead>
<tr>
<th>Method</th>
<th>ETH</th>
<th>HOTEL</th>
<th>UNIV *</th>
<th>ZARA1</th>
<th>ZARA2</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>1.33/2.94</td>
<td>0.39/0.72</td>
<td>0.82/1.59</td>
<td>0.62/1.21</td>
<td>0.77/1.48</td>
<td>0.79/1.59</td>
</tr>
<tr>
<td>LSTM</td>
<td>1.09/2.41</td>
<td>0.86/1.91</td>
<td><strong>0.61/1.31</strong></td>
<td><strong>0.41/0.88</strong></td>
<td>0.52/1.11</td>
<td>0.70/1.52</td>
</tr>
<tr>
<td>Alahi et al. [1]</td>
<td>1.09/2.35</td>
<td>0.79/1.76</td>
<td>0.67/1.40</td>
<td>0.47/1.00</td>
<td>0.56/1.17</td>
<td>0.72/1.54</td>
</tr>
<tr>
<td>Ours-single-model</td>
<td><strong>0.88/1.98</strong></td>
<td><strong>0.36/0.74</strong></td>
<td>0.62/1.32</td>
<td>0.42/0.90</td>
<td><strong>0.34/0.75</strong></td>
<td><strong>0.52/1.14</strong></td>
</tr>
<tr>
<td>20 Outputs Single Model</td>
<td>0.81/1.52</td>
<td>0.72/1.61</td>
<td>0.60/1.26</td>
<td>0.34/0.69</td>
<td>0.42/0.84</td>
<td>0.58/1.18</td>
</tr>
<tr>
<td>Gupta et al. <a href="V">7</a></td>
<td>0.87/1.62</td>
<td>0.67/1.37</td>
<td>0.76/1.52</td>
<td>0.35/0.68</td>
<td>0.42/0.84</td>
<td>0.61/1.21</td>
</tr>
<tr>
<td>Gupta et al. <a href="PV">7</a></td>
<td><strong>0.70/1.43</strong></td>
<td>0.76/1.67</td>
<td><strong>0.54/1.24</strong></td>
<td><strong>0.30/0.63</strong></td>
<td>0.38/0.78</td>
<td>0.54/1.15</td>
</tr>
<tr>
<td>Sadeghian et al. [26]</td>
<td>0.73/1.65</td>
<td><strong>0.30/0.59</strong></td>
<td>0.60/1.27</td>
<td>0.38/0.81</td>
<td><strong>0.31/0.68</strong></td>
<td><strong>0.46/1.00</strong></td>
</tr>
<tr>
<td>Ours-20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ADE  FDE
DEMO
Peeking into the Future: Predicting Future Person Activities and Locations in Videos

CVPR Submission
Paper ID 2601
Outline

- Introduction
- Related Work (for trajectory and activity prediction)
- Approach: Next Model
  - Person Behavior Module
  - Person Interaction Module
  - Trajectory Generator with Focal Attention
  - Activity Prediction
- Experiments: ActEV/VIRAT and ETH & UCY
- Conclusion
Conclusion

• The proposed model in this paper is used to predict future activities of pedestrians as well as future activities

• A person is modeled using visual semantics such as appearance and behavior

• Experiments were done on two common benchmarks and results show that it outperforms state-of-the-art methods

• Applications of this model range from autonomous driving to human assisting robots
References


Thank you!

Questions?
• Focal Attention

\[ S^t = h_{t-1}^T \cdot Q_{ij} \]

\[ A^t = \text{softmax}( \max_{i=1}^{M} S^t_i:) \in \mathbb{R}^M \]

\[ B^t = [\text{softmax}(S^t_{1,:}), \cdots, \text{softmax}(S^t_{M,:})] \in \mathbb{R}^{M \times T_{obs}} \]

\[ \tilde{q}_t = \sum_{j=1}^{M} A^t_j \sum_{k=1}^{T_{obs}} B^t_{jk} Q_{jk} : \in \mathbb{R}^d \]
Object and Activity Classes

<table>
<thead>
<tr>
<th>Class</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object</td>
<td>Bike, Construction_Barrier, Construction_Vehicle, Door, Dumpster, Parking_Meter, Person, Prop, Push_Pulled_Object, Vehicle</td>
</tr>
<tr>
<td>Activity</td>
<td>Carry, Close_Door, Close_Trunk, Crouch, Enter, Exit, Gesture, Interaction, Load, Object_Transfer, Open_Door, Open_Trunk, PickUp, PickUp_Person, Pull, Push, Ride_Bike, Run, SetDown, Sit, Stand, Talk, Talk_phone, Texting, Touch, Transport, Unload, Use_tool, Walk</td>
</tr>
</tbody>
</table>
- Single Feature Ablation and Activity Detection Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>ADE ↓</th>
<th>FDE ↓</th>
<th>Act mAP ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our full model</td>
<td>17.91</td>
<td>37.11</td>
<td>0.192</td>
</tr>
<tr>
<td>No p-object</td>
<td>18.17</td>
<td>37.13</td>
<td>0.198</td>
</tr>
<tr>
<td>No p-scene</td>
<td>18.18</td>
<td>37.75</td>
<td>0.206</td>
</tr>
<tr>
<td>No p-keypoint</td>
<td>18.25</td>
<td>37.96</td>
<td>0.190</td>
</tr>
<tr>
<td>No p-appearance</td>
<td>18.20</td>
<td>37.79</td>
<td>0.154</td>
</tr>
<tr>
<td>Act Detect</td>
<td>18.27</td>
<td>37.68</td>
<td>0.155</td>
</tr>
</tbody>
</table>