Pedestrian Intent Understanding

Understanding the behaviors and intentions of humans are one of the main challenges autonomous ground vehicles still faced with. More specifically, when it comes to complex environments such as urban traffic scenes, inferring the intentions and actions of vulnerable road users such as pedestrians become even harder. Following is a review of different approaches developed to tackle this problem.

Datasets For Pedestrian intent understanding
Pedestrian Attribute Dataset
PPSS Dataset (Large human parsing benchmark dataset)
ETH Dataset
UCY Dataset
Pedestrian direction recognition dataset http://www.rovit.ua.es/dataset/pedirecog/.
TUD-det
CVC
USC-A, USC-B, USC-C
Daimler-DB
INRIA
NICTA
TUD-Brussels

Models

- **Architecture involving Dense neural network**
  A dense neural network has been used for a fixed number of time-steps and features to directly classify a pedestrian's intention to cross the street at a given cross walk. The input data which is the feature vector leads to a fully connected layer. In the activation layer, rectified linear function is used as activation function to attain non-linearity and training stability. Then, a dropout layer is used for regularization. The combination is repeated a few times. The last fully-connected layer has a single output neuron for classification which a sigmoid function transforms to values between -1.0 for not crossing the street with a very high probability and 1.0 for crossing with very high probability.
  The dataset used contains car and pedestrian tracks recorded with a Velodyne laser scanner. Every track is associated with a precise digital map, which describes road boundaries, crosswalks etc. Around 2000 trajectories with 10,000 data points have been used.
  The best performing dense neural network has three layers having 32, 64, 128 neurons per hidden layer., each with rectified linear functions with a dropout of 50%.

  Performance Evaluation: This network has achieved an average cross-validation accuracy of 96.21%.

  The advantage of the dense network is that it has simultaneous access to all currently relevant time steps and can make its decision based on all of those at the same time.

- **State refinement for LSTM**

  This model uses information from human-human interaction to predict pedestrian path trajectory. V-LSTM is a model that infers all pedestrians independently without considering interactions among them. In a V-LSTM, the location of the i th pedestrian at time t is embedded as a vector which is the input to a LSTM. With the hidden states extracted from LSTM, coordinates at time step t+1 are predicted. The same process is used by SR-LSTM to extract features from the trajectory of each pedestrian separately. In addition to this, state refinement module is used to refine cell states by passing message among pedestrians. Thus for the i th pedestrian, hidden states from neighbouring pedestrians are integrated through a message passing function and then combined with cell state of pedestrian i to obtain refined cell state. In order to focus on the most useful neighboring information and guide the message passing, the message is passed with a social-aware information selection mechanism. A motion gate is also used which acts on each hidden state for feature selection based on motion of pedestrian i and j and their relative spatial location. Pedestrian wise attention is used to emphasize important neighbors and control the amount of neighborhood message.

  Performance evaluation: The model is evaluated on ETH and UCY datasets (ETH-univ,ETH-hotel,UCY-zara01,UCY-zara02,UCY-univ). Performance of trajectory prediction is evaluated using Mean Average Distance error and Final Average Distance error.
  Model evaluation is done for the following 2 variations in SR-LSTM:

  SR-LSTM1: Contains motion gate, pedestrian attention layer, neighbourhood size-4x4, 1 refinement layer using current state

  SR-LSTM2: Contains motion gate, pedestrian attention layer, neighbourhood size-20x20, 2 refinement layersusing current state

  Average performance of SR-LSTM1 on the five datasets is 0.46/0.97
  Average performance of SR-LSTM2 on the five datasets is 0.45/0.94

- **Convolutional Neural Network**
  The network architecture has 5 convolutional layers and three fully connected layers. The last layer has a softmax function which classifies the pedestrian trajectory. All layers have their neurons activate using ReLU.
  Three different classes are defined: pedestrian moving to the left, to the right and to the front, considering each direction of pedestrians motion.

The dataset used comprised of 7416 images for training and 1752 for validation. Images were classified according to three different categories- right, left and front. 2907 images were assigned to Right, 3099 images to Left and 1410 to Front.

Performance Evaluation: An accuracy of 82% was obtained on using Gaussian distribution for weight initialization and stochastic gradient descent.
Shown below is the performance using Nesterov's accelerated gradient descent with variation in different parameters

Learning Rate	Initialized weights manually	Weight initialization method	Accuracy	Loss
0.01	No	Gaussian distribution	82%	0.14
0.01	Convolutional layer1, Fullyconnected layer6, 8	Gaussian distribution(for those not initialized manually)	83.8%	-
0.01	No	Xavier init(Convolutional layer1, Fullyconnected layer6, 8) Gaussian distribution for rest	80%	0.09
0.01	No	Xavier init(Fullyconnected layer6, 8) Gaussian distribution for rest	80%	0.04

The model performs better when weights are initialized manually for Convolutional layer1, Fully connected layers 6, 8.

- **Pedestrian intent prediction by planning**

This model firstly infers a pedestrian's destination from images and positions. Then, trajectory planning is applied towards these destinations for prediction.
The proposed model consists of 3 networks- Destination network, Topology network, Planning network. The Destination Network predicts a mixture of possible destinations from pedestrian images and positions in the form of a grid map. Topology Network generates a planning topology from destination grid maps as well as other environment features. Planning Network then, runs prediction as planning on the topology maps. The final output is a position probability map per predicted time step.
In the destination network, images and position of pedestrians serves as input. The first part consists of Recurrent mixed density network. Image is processed through a CNN and concatenated with position vector and is sent as input to LSTM. The network predicts possible destinations in the form of a probability distribution map. The probability distribution map is sent to the topology network where a fully connected network predicts a map on which planning is executed. Planning applied on topological map gives actual prediction. Output is in the form of a position probability grid per time step. Planning techniques used are Markov Decision process and Forward-Backward algorithm. IT has been found that forward-backward algorithm performs better.

Benchmark

On comparing performance of dense network, LSTM and CNN on a dataset containing car and pedestrian tracks recorded on a Velodyne laser scanner. Dense network gave the best accuracy of 96.21%. The best LSTM network contained 2 layer LSTM with 64 and 128 hidden units and showed 95.77% cross validation accuracy. It was observed that LSTMs outperform when information has to be stored for a longer period of time. For pedestrians crossing the street, information about orientation and velocity from a few time-steps ago is not very useful. Dense network has an advantage that iit has simultaneous access to all currently relevant time steps
and can make its decision based on all of those at the same time
SR-LSTM model takes into account human-human interaction on road for pedestrian intent detection. Thus, this model would give better performance in crowded roads where neighbouring pedestrians would affect trajectory of a given pedestrians.

Model we should use

SR-LSTM would be the best model we could use in crowded roads. Otherwise, dense network could be used.

References

1. A. Dominguez-Sanchez, S. Orts-Escolano, M. Cazorla. "Pedestrian movement direction recognition using convolutional neural networks", IEEE Transactions on Intelligent Transport Systems. In press. 2017
2. Fang Z, Vázquez D, López AM. On-Board Detection of Pedestrian Intentions. *Sensors (Basel)*. 2017;17(10):2193. Published 2017 Sep 23. doi:10.3390/s17102193
3. Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). *Pedestrian Detection: An Evaluation of the State of the Art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4), 743–761.* doi:10.1109/tpami.2011.155

4. Eike Rehder, Florian Wirth. Pedestrian prediction by planning using deep neural networks,2018 IEEE International Conference on Robotics and Automation (ICRA)May 21-25, 2018, Brisbane, Australia
5. Alex Dominguez-Sanchez, Sergio Orts-Escolano, Recognizing Pedestrian Direction Using Convolutional Neural Networks
6. Benjamin Volz, Karsten Behrendt, Holger Mielenz, A data driven approach for pedestrian intent estimation, 2016 IEEE 19th International Conference on Intelligent Transportation System
7. Pu Zhang, Wanli Ouyang, Pengfei Zhang, State refinement for LSTM towards Pedestrian trajectory prediction