DESIRE

Chandraker

Summarized: 24 October 2018

1 Introduction

- DESIRE uses RNN Encoder Decoder Framework for future prediction. It first predicts a wide range of future predictions using a CVAE (Conditional Variational Autoencoder). The evaluation is done on KITTI and Stanford Drone Dataset

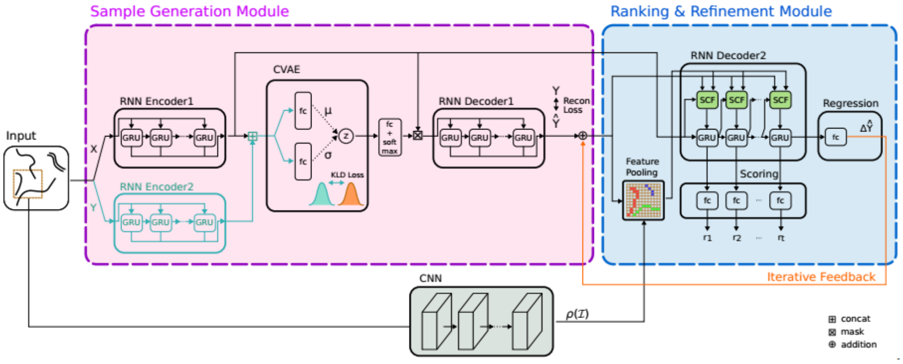


Figure 1: Network Design

2 Main Points

- Scene consists of semantic elements and dynamic participants like cars / pedestrians.
  - Diverse Sample Generation is done using the CVAE. The CVAE uses the past trajectories and predicts future hypothesis. The CVAE is combined with a RNN for this purpose
  - There is an IOC (Inverse Optimal Control) based ranking and Refinement. The most likely trajectory is ranked on the basis of IOC for better generalization.
  - The Scene Context Fusion aggregates the interaction bw agents and scene using CNN. This output is then passed to the RNN module to produce the rankings.

- The CVAE introduces a latent variable z_i which basically encodes the predictions Y_i given input X_i. The loss terms used are Reconstruction Loss and the KLD loss. The RNN's use GRU instead of LSTMs. Check cs231n for Variational autoencoders

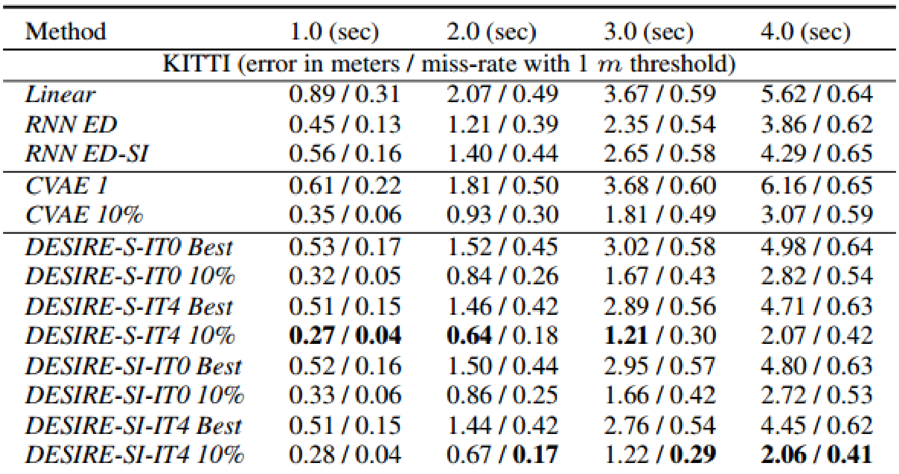
- IOC framework is used to train the trajectory ranking objective by measuring the long term future rewards. It uses iterative feedback and the losses used are cross entropy loss and regression loss. *Further details later*

- Scene Context Fusion: *Further details later*

- **KITTI Data**
  - Semantic Segmentation of images is done and velodyne points are used to get the labelled 3D points. Feature maps of size $H \times W \times C$ (C is no. of scene elements, $H, W = 80m$). Moving objects are removed during registration and there are approx 2500 train samples.
  - The major metrics used are: 1) L2 distance bw ground truth and prediction, 2) Miss rate with threshold of L2 distance and 3) Max L2 distance over entire time frames
  - Several baselines were used:
    - Linear Regression using MSE Loss
    - RNN encoder decoder model that regresses on the prediction
    - RNN ED-SI model which also consists of the scene context fusion model
    - DESIRE which is actual method proposed in the paper

- Training was done with adam optimizer using $lr = 0.0004$ which decreases after every quarter of total epochs. Gradient clipping was used and each model observes maximum of 2 seconds of past trajectories and predicts upto 4 seconds in the future.

3 Results



Method	1.0 (sec)	2.0 (sec)	3.0 (sec)	4.0 (sec)
	KITTI (error in meters / miss-rate with 1 m threshold)			
Linear	0.89 / 0.31	2.07 / 0.49	3.67 / 0.59	5.62 / 0.64
RNN ED	0.45 / 0.13	1.21 / 0.39	2.35 / 0.54	3.86 / 0.62
RNN ED-SI	0.56 / 0.16	1.40 / 0.44	2.65 / 0.58	4.29 / 0.65
CVAE 1	0.61 / 0.22	1.81 / 0.50	3.68 / 0.60	6.16 / 0.65
CVAE 10%	0.35 / 0.06	0.93 / 0.30	1.81 / 0.49	3.07 / 0.59
DESIRE-S-IT0 Best	0.53 / 0.17	1.52 / 0.45	3.02 / 0.58	4.98 / 0.64
DESIRE-S-IT0 10%	0.32 / 0.05	0.84 / 0.26	1.67 / 0.43	2.82 / 0.54
DESIRE-S-IT4 Best	0.51 / 0.15	1.46 / 0.42	2.89 / 0.56	4.71 / 0.63
DESIRE-S-IT4 10%	**0.27** / **0.04**	**0.64** / 0.18	**1.21** / 0.30	2.07 / 0.42
DESIRE-SI-IT0 Best	0.52 / 0.16	1.50 / 0.44	2.95 / 0.57	4.80 / 0.63
DESIRE-SI-IT0 10%	0.33 / 0.06	0.86 / 0.25	1.66 / 0.42	2.72 / 0.53
DESIRE-SI-IT4 Best	0.51 / 0.15	1.44 / 0.42	2.76 / 0.54	4.45 / 0.62
DESIRE-SI-IT4 10%	0.28 / 0.04	0.67 / **0.17**	1.22 / **0.29**	**2.06** / **0.41**

- The RNN encoder decoder performed much better than the linear regression. Their model RNN ED-SI performed worse than the simple encoder decoder network for Kitti dataset as it was very small. However, the CVAE based model performed much better due to the multi-modal nature of the problem.