Foresee: Attentive Future Projections of Chaotic Road Environments

Extended Abstract

Anil Sharma
Indraprastha Institute of Information Technology, Delhi
Delhi, India
anils@iiitd.ac.in

Arun Balaji Buduru
Indraprastha Institute of Information Technology, Delhi
Delhi, India
arunb@iiitd.ac.in

ABSTRACT

In this paper, we train a recurrent neural network to learn dynamics of a chaotic road environment and to project the future of the environment on an image. Future projection can be used to anticipate an unseen environment for autonomous driving. Road environment is highly dynamic and complex due to the interaction among traffic participants such as vehicles and pedestrians. Even in such a complex environment, a human driver can easily anticipate the environment and is efficacious to drive safely on the chaotic roads. Proliferation in deep learning research has shown the efficacy of neural networks in learning this kind of human behavior. In the same direction, we investigate recurrent neural networks to understand the road environment. We propose *Foresee*, a unidirectional gated recurrent units (GRUs) network with attention to project future of the environment in the form of images. We have collected several videos on Delhi roads consisting of various traffic participants, background and infrastructure differences (like 3D pedestrian crossing) at various times on various days. We show that our proposed model performs better than state of the art methods (prednet [9], Enc. Dec. LSTM [15]).

KEYWORDS

Future projection; prediction; attention; chaotic environments.

ACM Reference Format:
Anil Sharma and Arun Balaji Buduru. 2018. Foresee: Attentive Future Projections of Chaotic Road Environments. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 INTRODUCTION

Environment anticipation is an important task for situation awareness and decision making. There is a recent progress in anticipation of road environments [5, 7] for safe driving. However, anticipation becomes difficult in uncertain and dynamic environments [1]. The road environment is highly dynamic and stochastic due to the presence of a diverse set of human drivers and pedestrians. We define chaotic environment as that environment where the traffic participants follow no rule and move randomly. The same case is seen on road in developing countries like India. In such environments, the road space is shared by pedestrians, vehicles (cars, trucks, buses, motor-bikes etc.), and sometimes animals as well. Even when the environment is complex, its behavior can be modeled [7]. Modeling such an environment requires detection, tracking, and understanding of the dynamics of the traffic participants. Anticipating behavior of the environment is essential in various applications such as autonomous driving [13], driving assistance [14] etc. However, humans are very good at anticipating such an environment. For example, they drive very successfully by anticipating maneuvers even in a very crowded and chaotic shared space such as markets, street roads etc. We explore ways to achieve that anticipation power in machines using neural networks by exploiting the predictive power of recurrent neural network to capture this human behavior. In this paper, we propose a deep learning architecture to generate future projections in terms of the camera frames few frames in advance.

In this work, we propose *Foresee*, a deep learning architecture for future projections of the chaotic road environment directly from the raw camera images. The network is composed of two layers of GRUs (Gated Recurrent Units [4]) to encode the dynamics of the environment into a small representation in the hidden layers. We train the network in an unsupervised way to achieve the desired performance. We formulate the above problem as a sequence generation task, where a sequence is the collection of images that are contiguous in time.

Related works have also explored future projections from various viewpoints. One common approach is Bayesian filtering to predict next state as in Kalman filter [8]. Authors in Ning et al. [10], Redmon and Farhadi [12] have used CNN and LSTMs to find the future trajectory of an object using current camera location. The above approaches are supervised and requires a labeled dataset to predict target locations. [2] has modeled the interaction among pedestrians using a LSTM [6] network. Authors in Ondrúška and Posner [11] have looked one step ahead for object tracking in partially observable environment from simulated data. The papers Lotter et al. [9], Srivastava et al. [15] are very similar to our work. They evaluated their proposal on simple environment like fewer vehicles and only vehicles in the scene. We show that *Foresee* performs better than above two approaches for chaotic environments.

The subsequent sections are structured as follows. Section 2 describe the proposed method for environment anticipation. Section 3 describe the experimental setup and results. section 4 concludes the paper.

2 METHODOLOGY

The future projection is carried out in various steps. The input image is first normalized in range between 0 and 1 and then gamma

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

correction is applied to enhance illumination. The normalized corrected image is then re-sized to shape $32*32*3$. An image sequence is then created by concatenating the last 10 normalized images. The input image sequence is the sequence of images starting from current frame to 9 frames in past and the output image sequence is the sequence of images in the future. The prepared image sequence is then passed to the recurrent network to encode the temporal sequence for future projections. The output sequence is reconstructed from the encoded representations using a fully connected layer with hidden units equal to the number of pixels in the output image ($32 * 32 * 3$). The proposed network consists of GRUCells which has a hidden state corresponding to each time step. The figure 1 diagrams *Foresee* for future projections few frames in advance.

The input x_t at time t is fed into the network and the necessary information to encode the temporal sequence till time t is stored in the hidden state of the GRU cell. In GRU networks, the reconstruction quality degrades with the longer sequence as it cannot stuff all the information into its hidden layer (see [3]). To resolve this problem, attention methods were employed. Since road environment is not markov i.e., it does not depend only on previous frame, the attention method helps the network to attend to past frames as compared to only the previous frame. The attention mechanism takes outputs from all previous time steps and makes a context vector. The context vector is then multiplied with the new hidden state and then the output is reconstructed using fully connected layer (FC block in the figure). The output of the network is y_t.

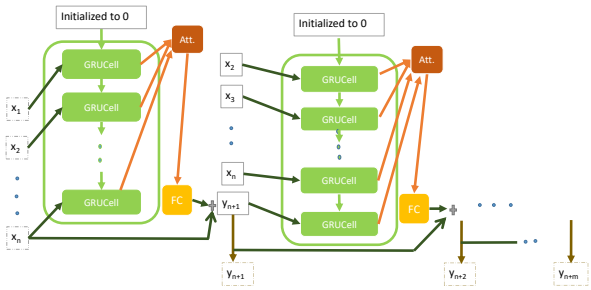

Figure 1: *Foresee* model: attention is applied on the GRUCells. Single bigger block is applied recursively to predict a longer output sequence.

The training loss at time t is mean square error between the target frame T_t and the projected frame y_t. The training loss is mentioned in equation 1.

$$L_{train} = \frac{1}{N}\sum_{i=0}^{N} |T_t(i) - y_t(i)|^2 \quad (1)$$

The above procedure is applied to project the next frame from the input sequence (please note that we are using input sequence of length $N = 10$).

Table 1: Performance comparison on test data.

Approach	MSE	SSIM
Enc. Dec. LSTM	$7.9 * 10^{-4}$	62.92
PredNet	$5.8 * 10^{-2}$	70.04
Foresee	$1.08 * 10^{-5}$	86.40

Figure 2: Image showing next frame projection using *Foresee*. Each image is $32 * 32 * 3$.

3 EXPERIMENT AND RESULTS

In this section, we will show initial results using our proposed framework for future projection.

We found many-to-many sequence prediction networks to be more effective for encoding the environment representations for future projections. Method proposed in [15] is a many-to-many sequence prediction network using LSTMs (Long Short Term Memory [6]). In our work, we first employed the approach proposed by Srivastava et. al. [15] and identified that it is not able to persist the sequence representation even for few frames because the next frame does not depend completely on the current frame. The collected data of chaotic environments is divided into train and test set. Table 1 shows the mean square error and structural similarity index measure of our proposal and proposals in [9, 15] on the test set. In the final network, the hidden state size is 512 and the input sequence length is 10 frames (1 second). Our model has 2 layers of GRU cells. Figure 2 show the future projected images using the proposed method.

4 CONCLUSION

We proposed *Foresee*, a deep learning architecture for future projections using Gated Recurrent Units and attention methods. We showed that attention when applied at all steps of the reconstructed output of the input sequence performs better. We showed that the proposed architecture performs better than the state of the art methods for future projections.

REFERENCES

[1] [n. d.]. thedrive. http://www.thedrive.com/tech/12032/self-driving-cars-are-flummoxed-by-indias-chaotic-roads/

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–971.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014). arXiv:1409.0473 http://arxiv.org/abs/1409.0473

[4] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *CoRR* abs/1406.1078 (2014). arXiv:1406.1078 http://arxiv.org/abs/1406.1078

[5] Frank Havlak and Mark E. Campbell. 2013. Discrete and Continuous, Probabilistic Anticipation for Autonomous Robots in Urban Environments. *CoRR* abs/1309.0766 (2013). arXiv:1309.0766 http://arxiv.org/abs/1309.0766

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] Ashesh Jain, Hema Swetha Koppula, Shane Soh, Bharad Raghavan, Avi Singh, and Ashutosh Saxena. 2016. Brain4Cars: Car That Knows Before You Do via Sensory-Fusion Deep Learning Architecture. *CoRR* abs/1601.00740 (2016). arXiv:1601.00740 http://arxiv.org/abs/1601.00740

[8] Rudolph Emil Kalman et al. 1960. A new approach to linear filtering and prediction problems. *Journal of basic Engineering* 82, 1 (1960), 35–45.

[9] William Lotter, Gabriel Kreiman, and David Cox. 2016. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104* (2016).

[10] Guanghan Ning, Zhi Zhang, Chen Huang, Zhihai He, Xiaobo Ren, and Haohong Wang. 2016. Spatially Supervised Recurrent Convolutional Neural Networks for Visual Object Tracking. *arXiv preprint arXiv:1607.05781* (2016).

[11] Peter Ondrúška and Ingmar Posner. 2016. Deep Tracking: Seeing Beyond Seeing Using Recurrent Neural Networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI Press, 3361–3367. http://dl.acm.org/citation.cfm?id=3016100.3016374

[12] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. *arXiv preprint arXiv:1612.08242* (2016).

[13] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2016. Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. *CoRR* abs/1610.03295 (2016).

[14] S. Sivaraman and M. M. Trivedi. 2014. Dynamic Probabilistic Drivability Maps for Lane Change and Merge Driver Assistance. *IEEE Transactions on Intelligent Transportation Systems* 15, 5 (Oct 2014), 2063–2073. https://doi.org/10.1109/TITS.2014.2309055

[15] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. 2015. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*. 843–852.