Video Prediction

Independent Work Report (MAE 340, Spring 2021)

Matthew Coleman

April 26, 2022

This project represents my own work, in accordance with the University regulations.

/s/Matthew Coleman

Contents

1 Introduction

While humans cannot perfectly predict the future, they are indeed capable of predicting near events to some extent, and this knowledge greatly aids them in planning out their actions, such as which movements to take to reach a goal. To some extent, this ability to forecast the future is a direct result of an understanding of causality that is learned through observation and interaction [1].

A great amount of human predictions are, of course, entirely erroneous and result in some consequences, but even the humans least adept at predicting likely outcomes are still masters of predicting the future in some respects. For example, humans have a good sense for where a car will move in the street, or which direction a pedestrian may continue walking. Even a young child can predict where to toss a football to a moving receiver, and even this small knowledge reveals an infinite wisdom compared to the most advanced video prediction methods.

The task of video prediction is comprised of several open challenges in computer vision; it uses some of the most recent model architectures that have been developed and it even deals directly with an impossible task altogether, which is to predict the future. Although it is a particularly confusing task, however, it also has the potential for immense impact and immediate practical applications, such as in autonomous driving [3], video interpolation [4] and most interesting in the context of this report, robotic control systems [2].

2 Prediction Model Families

In addition to convolution, modern video prediction models tend to adopt a few canonical architectures, which are each very well-researched and developed. Although they are typically implemented slightly differently in practice, the overarching themes are still in effect, and many of these model paradigms are combined or adapted for alternative tasks. The best video prediction models use them in conjunction with each other to create a larger network of interleaving models.

Of these are recurrent models, which are trained on sequences of data, generative models, which are trained to approximate the conditional probability $p(xy)$ in order to generate images in the same distribution as ground-truth images.

2.1 Recurrent Models

2.2 Generative Models

3 Convolutional LSTM

4 FutureGAN

5

6

7

8

References

[1] A. Cleeremans and J. Mcclelland. Learning the structure of event sequences. *Journal of experimental psychology. General*, 120:235–53, 10 1991. doi: 10.1037//0096-3445.120.3.235.

[2] C. Finn and S. Levine. Deep visual foresight for planning robot motion, 2016. URL https://arxiv.org/abs/1610.00696.

[3] A. Hu, F. Cotter, N. Mohan, C. Gurau, and A. Kendall. Probabilistic future prediction for video scene understanding, 2020. URL https://arxiv.org/abs/2003.06409.

[4] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala. Video frame synthesis using deep voxel flow, 2017. URL https://arxiv.org/abs/1702.02463.