Video Prediction and Learning Causality

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 inferring a great deal of information about near events in the future, and this knowledge greatly aids them in planning out their actions, such as which movements to take to reach a goal. 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, erroneous in major respects, but even the humans least adept at inferring far-off outcomes and consequences still are masters of learning very near-term ones. 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, 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].

This project will examine the current state-of-the-art in video prediction models, report on several experiments carried out by implementing and testing such a model on various existing datasets, and attempt to make meaningful conclusions about video prediction and learning causality.

2 The Task of Video Prediction

The task of video prediction is to construct an approximation for the completion of a sequence of frames, given only the initial sequence of frames. Formally, given an ordered set of n image frames $\boldsymbol{X} = (X_1, X_2, X_3 \cdots X_n)$, the task is to predict the latter m frames of the sequence $\boldsymbol{Y} = (Y_1, Y_2, Y_3 \cdots Y_m)$, each frame of which having the same dimensions, for example with c channels, height h, and width w. The model predictions $\hat{\boldsymbol{Y}} = \left(\hat{Y}_1, \hat{Y}_2, \hat{Y}_3 \cdots \hat{Y}_n\right)$ are conditioned on the input sequence $\boldsymbol{X}$, and the model weights are updated typically by the gradient of a loss function computed between the predictions and ground truth sequence $\boldsymbol{Y}$ directly. Critically, since there is no human intervention or labeling required for the model to do this, video prediction is a self-supervised task [5].

3 Families of Prediction Models

Modern video prediction models tend to adopt canonical architectures.

3.1 Recurrent Models

Recurrent neural networks (RNNs) consist of a network of nodes that stretches over some sequential information, typically in the form of a time sequence (Which is the case in video prediction). Each node will perform another learning technique on the data in sequence (This could be a convolution operation, a linear layer, or a combination of several, for example) and output its own activation. Commonly, this is implemented not with a network of individual nodes but rather in the form of a feedback loop over a single node, which passes some data embedding forward through the network to itself in a loop (a hidden state, or variable), only taking in new data from the original input sequence at each step. In this way, an RNN equipped with convolution is capable of learning from time-varying information while preserving spatio-temporal relations (That is, relationships in the data that exist over space, such as the shape of a person's leg and hip, as well as relationships in the data that exist over time, such as the motion of a person walking, will be preserved in the final activations of the network).

The outputs of each node are then used for other purposes, depending on the task, and the result is then backpropagated against a loss function. In machine learning parlance, this is referred to as backpropagation through time (BPTT), since a gradient must be computed in the input sequence's reverse order, i.e., backwards through time, and in the case of an RNN implemented with feedback, this gradient must be computed with respect to the input data at each time step and then added to the weights of the single node in sum.

3.2 Generative Models

Generative neural networks (for example, GANs) consist of mostly the same architecture as discriminative neural nets, such as the ones which are classically used for image classification. While discriminative nets seek to learn the conditional probability $p(y \mid x)$ of an input x belonging to a particular class y, generative nets seek to learn the opposite conditional probability $p(x \mid y)$ of an input data given the output. In practice...

4 Convolutional LSTM

5 FutureGAN

6

7

8

9

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.

[5] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020. ISSN 1939-3539. doi: 10.1109/tpami.2020.3045007. URL http://dx.doi.org/10.1109/TPAMI.2020.3045007.