A Novel Adversarial Inference Framework for Video Prediction with Action Control

Zhihang Hu & Jason T.L. Wang
New Jersey Institute of Technology
ashvegance@gmail.com

Abstract

The ability of predicting future frames in video sequences, known as video prediction, is an appealing yet challenging task in computer vision. This task requires an in-depth representation of video sequences and a deep understanding of real-word causal rules. Existing approaches often result in blur predictions and lack the ability of action control. To tackle these problems, we propose a framework, called VPGAN, which employs an adversarial inference model and a cycle-consistency loss function to empower the framework to obtain more accurate predictions. In addition, we incorporate a conformal mapping network structure into VPGAN to enable action control for generating desirable future frames. In this way, VPGAN is able to produce fake videos of an object moving along a specific direction. Experimental results show that a combination of VPGAN with some pretrained image segmentation models outperforms existing stochastic video prediction methods.

Introduction

Acquiring an in-depth understanding of videos has been a cornerstone problem in computer vision. This problem has been studied by various researchers from different perspectives, among which video prediction has attracted much attention.

However, despite its appealing prospects, accurate video prediction remains an open problem. The major challenge is the inherent uncertainty in the dynamics of the world. A typical example is that the future trajectory of a ball hitting the ground is inherently random. The main contributions of our work include the following:

- We introduce a new adversarial inference model designed for stochastic video prediction and incorporate a novel cycle-consistency loss into the model.
- We incorporate a conformal mapping network structure into our VPGAN framework to enable action control for generating desirable future frames.
- We combine pre-trained image segmentation models with our VPGAN framework to exploit their effectiveness in image understanding. Having more semantic understanding of the frames in video sequences would enable VPGAN to generate more accurate predictions.

The combination of our VPGAN framework with the pre-trained image segmentation models outperforms existing stochastic video prediction methods as shown in our experimental results reported in the paper.

Methods

The task of stochastic video prediction can be formalized as learning a multivalued function $f : R^{N\times M\times T} \mapsto R^{N\times M}$ from a collection of T context frames $X_0, \ldots, X_{T-1}$, each of which is a matrix of N rows and M columns of pixels, to some possible future frames $\{X_T\}$.

It is natural to think that the transformation from frame X_{t-1} to frame X_t is caused by some variation Z_t.

Adversarial Learning

Let X represent the frames and let Z represent the variations under consideration. Let $p_{data}(X)$ represent the true distribution of X. We wish to construct a joint distribution $q(X, Z)$ such that $q(X, Z)$ is a good approximation of $p_{data}(X)$.

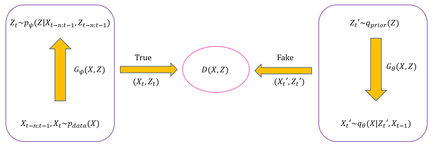

Figure 1: Illustration of our VPGAN learning process. Both G_ψ and G_θ are generators. Discriminator $D(X, Z)$ tries to discriminate between true pair (X, Z) and fake pair (X', Z').

Figure1 illustrates our VPGAN learning process during training. VPGAN employs two generators: $p_\psi = G_\psi(X, Z)$ and $q_\theta = G_\theta(X, Z)$. Let $X_{t-n:t-1}$ denote the frames $X_{t-n}, \ldots, X_{t-1}$ and let $Z_{t-n:t-1}$ denote the variations $Z_{t-n}, \ldots, Z_{t-1}$. When the training is completed, the two joint distributions $q(X, Z)$ and $p(X, Z)$ match with each other.

Denote $p_\psi(Z|X_{t-n:t-1}, Z_{t-n:t-1})$ by $G_\psi(X_{t-n:t-1}, Z_{t-n:t-1})$ and $q_\theta(X|Z'_t, X_{t-1})$ by $G_\theta(Z'_t, X_{t-1})$. The adversarial loss function used in the training is calculated as:

$$\begin{aligned} L_{adv} = {} & E_{X_t\sim p_{data}(X)}[\\ & \log D(X_t, G_\psi(X_{t-n:t-1}, Z_{t-n:t-1}))] \\ + {} & E_{Z'_t\sim q_{prior}(Z)}[\\ & 1 - \log D(G_\theta(Z'_t, X_{t-1}), Z'_t)] \end{aligned} \tag{1}$$

To generate or predict the next frame X_t based on the past frames $X_{t-n:t-1}$, the past frames $X_{t-n:t-1}$ and past encoded vectors $Z_{t-n:t-1}$ are sent to the encoder p_ψ, which generates the next encoded vector (variation) Z_t. Then the decoder q_θ takes X_{t-1} and Z_t together, and predicts the next frame X_t. Depending on different variations (latent variables) Z_t, q_θ can predict multiple possible next (future) frames $\{X_T\}$.

Cycle-Consistency

Cycle consistency is based on the idea of using transitivity as a way to regularize structured data. Here we propose a new cycle consistency loss function for video prediction. With the same generator in (1), we generate the frame at time $t-1$, X_{t-1}, conditioned on the opposite of Z_t and X_t. That is, we generate $\bar{X}_{t-1}$ conditioned on $-Z_t$ and X_t where $\bar{X}_{t-1}$ is approximately equal to X_{t-1} as expressed in (2):

$$X_{t-1} \approx \bar{X}_{t-1} \sim q_\theta(X| - Z_t, X_t) \tag{2}$$

Mathematically, denote $q_\theta(X|Z_t, X_{t-1})$ by $G_\theta(Z_t, X_{t-1})$ and $q_\theta(X| - Z_t, X_t)$ by $G_\theta(-Z_t, X_t)$. Our cycle consistency loss is calculated as

$$\begin{aligned} L^1_{cycle} = {} & E_{X_t,X_{t-1}\sim p_{data}(X)}\{ \\ & \parallel X_t - G_\theta(Z_t, G_\theta(-Z_t, X_t)) \parallel_1 \\ + {} & \parallel X_{t-1} - G_\theta(-Z_t, G_\theta(Z_t, X_{t-1})) \parallel_1\} \end{aligned} \tag{3}$$

Here, we utilize L_1 loss as the reconstruction loss. The loss function L_{cycle} in (3) only considers one-step cycle consistency. We can generalize the formula in (3) to take into account cycle consistency of multiple steps (more precisely, k steps) for video prediction. Combining the multi-cycle loss and multi-reconstruction loss in our paper, our overall loss, denoted as L_{loss}, is calculated as follows,

$$L_{loss} = \alpha L_{adv} + \beta L^k_{cycle} + \lambda L^k_{recon} \tag{4}$$

Action Control

We wish to manipulate the latent variable space $\mathcal{Z}$ so as to generate desirable moving directions, through preserving 'orthogonality,' or more precisely, through preserving angles between the encoding space $\mathcal{Z}$ and the moving directions of an object. This orthogonality property can be preserved by first enforcing the latent variable space $\mathcal{Z}$ to be R^2. Then, we added a conformal sub network, denoted τ, which maps a latent variable from the latent variable space $\mathcal{Z}$ to the moving direction space $\mathcal{D}$.

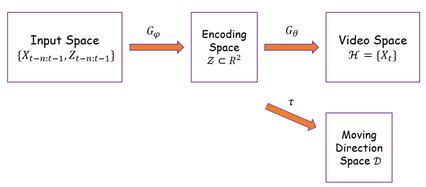

Figure 2: Illustration of our modified model for action control..

A mapping f is conformal iff it is homomorphic and its derivative is nowhere zero. In our VPGAN framework, the mapping τ is implemented using a 3-layer affine transform. It is very easy to prove that such an affine transform enforces τ to be conformal, therefore it preserves angles between any two vectors through $'0'$. In this way, if we know a latent variable Z moving toward a specific direction, we can control the generated moving direction by manipulating the latent variable Z (through rotating with some angle since the angle is preserved between the latent variable space $\mathcal{Z}$ and the moving direction space $\mathcal{D}$).

Results

The BAIR robot pushing dataset involves a series of videos generated by a Sawyer robotic arm pushing a variety of objects. All of the videos have relatively similar surroundings (table settings) with a static background. Each video also collected actions taken by the robotic arm corresponding to the commanded gripper pose.

KTH action dataset contains various types of videos collected in real world cameras including human subject doing one of six activities(walking, jogging, running, boxing, hand waving, and hand clapping). For first three activities, the human enters and leaves the frame multiple times, leaving the frame empty with a mostly static background for multiple frames at a time.

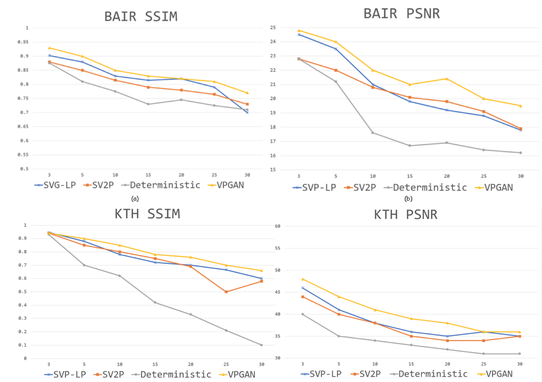

Figure 3: Performance results of our approach on BAIR and KTH datasets compaired with SVG-LP, SV2P and deterministic models.

Conclusion

In this paper, we proposed a novel framework for video prediction which reaches the state-of-the-art in several dataset and introduced a novel idea of action control.