Pattern Recognition Letters 110 (2018) 58–65

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Deep generative video prediction

Tingzhao Yu[a,b,*], Lingfeng Wang[a,c], Huxiang Gu[a], Shiming Xiang[a], Chunhong Pan[a]

[a] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 101408, China
[c] Hunan Provincial Key Laboratory of Network Investigational Technology, Hunan Police Academy, Changsha 410138, China

ARTICLE INFO

Article history:
Received 25 July 2017
Available online 28 March 2018

MSC:
41A05
41A10
65D05
65D17

Keywords:
Video prediction
Two stream
Adversarial training
Convlstm

ABSTRACT

Video prediction plays a fundamental role in video analysis and pattern recognition. However, the generated future frames are often blurred, which are not sufficient for further research. To overcome this obstacle, this paper proposes a new deep generative video prediction network under the framework of generative adversarial nets. The network consists of three components: a motion encoder, a frame generator and a frame discriminator. The motion encoder receives multiple frame differences (also known as *Eulerian motion*) as input and outputs a global video motion representation. The frame generator is a pseudo-reverse two-stream network to generate the future frame. The frame discriminator is a discriminative 3D convolution network to determine whether the given frame is derived from the true future frame distribution or not. The frame generator and frame discriminator train jointly in an adversarial manner until a Nash equilibrium. Motivated by theories on color filter array, this paper also designs a novel cross channel color gradient (3CG) loss as a guidance of deblurring. Experiments on two state-of-the-art data sets demonstrate that the proposed network is promising.

1. Introduction

Understanding videos is a core problem of pattern recognition and artificial intelligence [35]. It has many applications such as video classification [10], video segmentation [2], video retrieval [6], action recognition [31], crowd analysis [24], event detection [36] and video prediction [18]. Among these applications, video prediction has received growing interests in computer vision and is of great significance for video surveillance [3], video forecasting [32] and autonomous vehicles [13].

As a promising avenue for video understanding, video (or Pixel-level) prediction is of great challenge. This paper addresses the issue of future frame prediction [5,17,20,27,30,33]. Existing methods mainly focus on exploiting the neighbor frame correlation via cross channel or cross frame convolution. For a given video, in order to estimate the discrete joint distribution of the raw pixel values, Video Pixel Network (VPN) [9] constructs a probabilistic model. The model captures the four-dimensional video structure in the temporal dimension of the sequence, in the two spatial dimensions of each frame and in the color channels of a pixel. [20] propose a new spatial-temporal video autoencoder. It consists of a classic spatial image autoencoder and a novel nested temporal autoencoder. At each time step, the network receives a video frame, predicts the optical flow and generates the next frame. Another possible solution for video prediction is generative adversarial networks (GAN) [7], *e.g.* [33] utilize GAN to generate videos from scratch instead of conditioned on the past. Besides, inspired by the concept of predictive coding in neuroscience, [15] propose Predictive Network (PredNet) to predict future frames in a video sequence. Each layer of the network only makes local predictions to subsequent network layers.

This paper proposes a new deep generative architecture for future frame prediction. This work is mainly inspired by Motion and Content Network (MC-net) [32] and Multi Scale Deep Generative network (MSDG) [17]. This paper differs from them in three aspects:

1) Different from MC-net, which is a totally hierarchical network, this paper adopts an adversarial training strategy. This enables the network to automatically generate frames from the original future frame distribution.
2) Different from MSDG, which needs multi-scale RGB information, this paper utilizes single scale frame difference. Using frame difference helps to characterize the motion information more concretely.
3) Different from these two works, this paper employs a new cross channel color gradient loss. This loss function forces the cross

[*] Corresponding author.
E-mail address: tingzhao.yu@nlpr.ia.ac.cn (T. Yu).

https://doi.org/10.1016/j.patrec.2018.03.027
0167-8655/© 2018 Elsevier B.V. All rights reserved.

channel color difference to be consistent, thus it can reduce the blur effect.

In fact, the essence of future frame prediction is to minimize the reconstruction error $\|\hat{\mathbf{x}}_{t+1} - \mathbf{x}_{t+1}\|_{\alpha}^{\alpha}$ between the true $(t+1)$th future frame $\mathbf{x}_{t+1}$ and the predicted $(t+1)$th future frame $\hat{\mathbf{x}}_{t+1}$, where α is an integer greater or equal to 1. To some extent, it can be dealt with an autoencoder. Srivastavaet al. [30] use a LSTM autoencoder to minimize the reverse video sequence reconstruction error and at the same time present the predict future frame. Walker et al. [34] propose to use Conditional Variational AutoEncoder to depict the uncertain future via optimizing a $\mathcal{KL}$-divergence term $\mathcal{KL}[p(\hat{\mathbf{x}}_{t+1}|\mathbf{x}_{1:t})\|q(\hat{\mathbf{x}}_{t+1}|\mathbf{x}_{1:t})]$, where p represents the predicted distribution and q describes a random distribution.

The contributions of this paper are summarized as follows:

1) A new deep generative network is proposed for video prediction. Within this framework, a generative model generates the future frame of a given video utilizing frame differences. And a discriminative model estimates the probability of a given image being the true future frame. Benefiting from the net architecture, this network can be trained end-to-end via back propagation.
2) A pseudo-reverse two-stream frame generator is proposed for future frame generation. A dynamic stream is designed to encode the motion representation and predict the future motion representation. A static stream is designed to preserve the static content. The future frame is obtained by integrating the outputs of these two streams.
3) A new cross channel color gradient loss is designed to improve the generated frame quality. The motivation is to preserve the consistency among color channels. Theories on color filter array demonstrate that color gradients are consistent for real world images. This can help to highlight the edge areas and reduce blur effect.

Even though the ideas of utilizing difference images Xue et al. [37] and GAN Vondrick et al. [33] have been exploited, the proposed framework is quite different from these works in four aspects.

1) Given a sequence of static frames, Xue et al. [37] take difference images as convolution kernels. However, we take difference images as the input data. Employing difference images as input is essential, because it simplifies the issue of frame prediction to motion representation.
2) For getting better results, Xue et al. [37] need multi-scale frames as input. On the contrary, we only require frames within a single resolution. Thus the proposed network is more efficient.
3) Vondrick et al. [33] utilize GAN for video generation given a random noise. However, we employ GAN for video prediction given a series of known frames. Thus the two problems are quite different from each other.
4) Vondrick et al. [33] apply a two-stream network for both background and foreground generation. Nonetheless, we adopt a pseudo two-stream network, in which only the foreground is predicted and the background is given by the former frame. This in turn makes our network more accurate and more efficient.

2. Related work

The main consideration of future frame prediction is to minimize the reconstruction error between the true future frame and the generated future frame. The related works to this paper are video prediction, video synopsis and two-stream networks.

Video prediction. Given a short video clip, Ranzato et al. [22] propose a baseline of video prediction based on theories about language model. They construct multiple quantized patch dictionaries and apply a recurrent neural network to classify whether an image patch is the future frame. Yet, the future frame is indeterminate, Xue et al. [37] propose to characterize the future frame in a probabilistic manner. This is implemented via cross convolutional. They regard image and motion as feature maps and convolutional kernels, respectively. A Conditional Variational Autoencoder [12] form loss function, which makes synthesizing many possible future frames possible, is designed to model the probability of the future frame. Specifically, Oh et al. [19] firstly encode the frame level information through CNN and then depict the motion information via LSTM. Deconvolution is applied to decode the predicted frames from the transformed encoding. Nevertheless, the main drawback of these methods is the blur effect caused by the minimization of reconstruction error. Mathieu et al. [17] propose a MSDG network to deal with the inherently blurry predictions. Besides, they design a new image gradient loss function to address the problem of lack of sharpness.

Video synopsis. The aim of video synopsis [11,21,28] is to select a sparse subset of video frames that can optimally represent the input video. On the contrary, in this paper, we aim to obtain an optimal motion representation of the input video. Zhang et al. [39] use two LSTMs, one along the time sequence and the other in reverse from the video's end, to select key video frames. To ensure diversity of the selected frames, the network is trained via minimize the cross-entropy loss on annotated ground-truth key frames with an additional loss based on determinantal point process. Mahasseni et al. [16] learn a deep synopsis network to minimize the distance between the training videos and the distribution of their summarizations in an unsupervised way. They utilize a LSTM summarizer to select video frames and a LSTM discriminator to distinguish their similarity.

Two-stream network. Motivated by researches on physiology, Simonyan and Zisserman [26] first propose two-stream network for video-based human action recognition. In this paper, we employ a pseudo-reverse two-stream network for future frame generation. Within the framework of two-stream, a spatial network is designed to detect the moving object, and a temporal network is employed for motion recognition [4]. Saito and Matsumoto [23] first implement a two-stream adversarial network, named Temporal Generative Adversarial Network (TGAN), for generating future frames. Different from MSDG, TGAN consists of two generators, a temporal generator and a frame generator. The temporal generator corresponds to motion transformation and the frame generator handles object generation. Besides, Villegas et al. [32] also propose a two-stream network, called MC-net, to decompose the motion and content in videos. The network is built both upon autoencoder and LSTM. Thus it can capture the spatial layout of and temporal dynamics independently.

3. Architecture

This section formulates the task of future frame prediction and presents the details of the proposed network. Let $\mathbf{x}_{1:t} \in \mathbb{R}^{t\times w\times h\times c}$ represents the first t frames of a given video $\mathbf{x}$, where *t, w, h* and *c* denote the *temporal length, spatial width, spatial height* and *channel numbers*, respectively. The aim of future frame prediction is to predict the following future frame $\hat{\mathbf{x}}_{t+1}$ conditioned on the given input video frames $\mathbf{x}_{1:t}$. This equals to maximize the conditional distribution $p_\theta(\hat{\mathbf{x}}_{t+1}|\mathbf{x}_{1:t})$ [37], where p_θ describes the frame distribution.

Nevertheless, a straightforward modeling of $p_\theta(\hat{\mathbf{x}}_{t+1}|\mathbf{x}_{1:t})$ is difficult due to the complex backgrounds in real world videos. Considering the high correlation among neighbor frames, the frame