Synthesizing Patterns for Videos Based on the Adversarial ConvNet

Wei Tian¹ and Jingyuan Lv²
¹ School of Electrical Engineering, University of Jinan, Jinan, China
² Information and Communications Division, Shandong Provincial Public Security Department, Jinan, China
Email: cse_tianw@ujn.edu.cn; 85123991@163.com

Abstract. In this paper, we proposed the application of the adversarial ConvNet mode with back-propagation algorithm and weighted method for synthesizing dynamic patterns. A multi-layer spatiotemporal filter is designed in the adversarial ConvNet to capture the spatiotemporal information that may be lost in the convolution phase. The model defines probability distributions. The problem of pattern design is to obtain the best synthesis by training these probability models and maximizing the objective likelihood function. In the training phase, multiple layers parameters of filters are updated through the matching between the synthesized videos generated based on the adversarial trained ConvNets and the observed training videos. The maximum likelihood learning algorithm is used to realize the synthesis of real natural image patterns. The effectiveness of this method is evaluated on public DynTex datasets and compared with other technology through the experimental results.

1. Introduction
In social media analysis and processing, synthesizing patterns for videos is an important part. Given some input sample video, synthesize some new video completely different from the given sample video. How to reduce more effectively the error has been an interesting problem for synthesizing and analyzing such dynamic patterns.

Based on convnet [1-4], various successful dynamic texture modeling and synthesis have been proposed in the field of graphics and vision in recent years. In fact, the research of dynamic pattern synthesis based on convnet is rare. For instance, Synthesizing patterns for videos are generated through the cooperative training of spatiotemporal descriptor network and spatiotemporal generation network [5-6]. The general antagonistic video network with adversarial convolution structure separates the foreground from the background [7]. Ref. [8] used a spatial-temporal discriminative ConvNet for analyzing video data. A nonlinear dynamic texture modeling method based on Bayes is proposed in Ref. [9]. The model can automatically construct a nonlinear kernel function suitable for dynamic modeling, so it can fit various types of dynamics [10]. A vector linear autoregressive model with Gaussian innovations based on single valued decomposition is proposed in Ref. [11]. Ref. [12] extended the spatially generated ConvNet to capture spatiotemporal patterns at different filter scales by increasing the time dimension.

Our work is inspired by the generative ConvNet model proposed in Ref. [13], which is a very effective discriminant learning machine or predictive learning method, and has a broad application prospect. The generative ConvNet model is a random field or energy-based, which appears in the exponential slant form of reference distribution. More commonly used are Gaussian white noise
distribution or uniform distribution [14, 15]. The main work of this paper is to propose a synthesizing patterns technology using generative ConvNet, aiming to provide a model for synthesizing patterns for videos. The design of the network structure has a confrontational ConvNet network, which is composed of multi-layer filters to obtain spatiotemporal parameters of different scales and synthesize video sequences from the learning model. The conventional part can translate potential informations into signals using the difference between the synthetic video sequence and the initial training image, the model filter parameters are updated. The maximum likelihood learning algorithm is used to realize the synthesis of real natural video modes.

Our model is an energy model based on image space, and we expect to be able to capture complex spatiotemporal patterns more flexibly by using multi-layer networks. Compared with the recursive temporal model, our model does not need a start frame, and is easier to capture time patterns. In our method, the synthesis of dynamic texture patterns has obvious advantages. Firstly, when the training sample is small, the model and unsupervised learning are very effective. Secondly, the existing network maximum likelihood learning algorithm is used in the learning process to capture spatiotemporal patterns through multi-layer spatiotemporal filters of different scales. Finally, the experimental results show that the training adversarial ConvNet method can synthesize real patterns for videos.

The rest is organized as follows. In the next section, we first review the principles of the framework model. A training algorithm is given to obtain the adversarial ConvNet weight parameters. The third section introduces the experiment, evaluates the effectiveness of this method compared with other method, and summarizes the content of this paper in the fourth section.

2. Frame Models

2.1. The Adversarial ConvNet

Let \( I = (I(x,t), x \in D, t \in T) \) be an image matrix, where \( x \) indexes the coordinates \((x_i, x_j)\) of pixels, and \( t \) indexes the frames in the video sequence.

The adversarial convnet filter can capture spatiotemporal patterns on multiple scales, which are composed of multilayer filetering, the recursive formula is:

\[
[F^{(l)}_k * I](x,t) = h\left(\sum_{i=1}^{N_l} \sum_{(s_1,s_2) \in S_l} \alpha^{(l,k)}_{i,y,s} \times [F^{(l-1)}_i * I](x + s_1, t + s_2) + b_{i,k}\right)
\]  

(1)

\(F^{(l)}_k\) and \(F^{(l-1)}_i\) represent the linear filterings, \(i\) and \(k\) represent the filters index, \(N_l\) and \(N_{l-1}\) represent the numbers of the filters, \(l\) and \(l-1\) represent the filter layer, \((s_1,s_2)\) is the filter’s range. \([F^{(l-1)}_i * I]\) is a linear filtering with \(\alpha^{(l,k)}_{i,y,s}\) weight parameters. \(h(\bullet)\) is ReLU operation.

The design of adversarial learning ConvNet structure defines the probability distribution of video sequences, which is a Markov random field model, as expressed by the following formula:

\[
P(I;W_D) = \frac{1}{Z(\omega)} \exp[ f(I;\omega)] q(I)
\]  

(2)

where \(Z(\omega)\) is the normalization constant, \(q(I)\) is the Gaussian white noise model. \(f(I;\omega)\) is the scoring function, which is:

\[
f(I;\omega) = \sum_{k=1}^{K} \sum_{x \in D} \sum_{t \in T} [F^{(L)}_k * I](x,t)
\]  

(3)

The filter response is aggregated over all filters, locations, and times at the \(L\) level. Space and time pools can identify non-Gaussian spatiotemporal features or patterns.
2.2. Learning by Maximum Likelihood

The stochastic gradient algorithm is used to train the model using the training video sequences. Firstly, the video sequences are synthesized from the learned model. Then the difference between the synthetic and the training image is used to update the filter parameters.

The logarithmic probability $L(\omega)$ is defined by an adversarial ConvNet to capture spatiotemporal patterns of different scales.

$$L(\omega) = \frac{1}{M} \sum_{m=1}^{M} \log P(I; \omega)$$  \hspace{1cm} (4)

The ConvNets parameters $\omega$ can be trained by maximum likelihood. The logarithmic probability gradient is defined by:

$$\dot{L}(\omega) = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial \omega} f(I_m; \omega) - E_{\omega}[\frac{\partial}{\partial \omega} f(I; \omega)]$$  \hspace{1cm} (5)

The adversarial ConvNet can be sampled through Langevin dynamics. The expected value is difficult to deal with in analysis, so it must be calculated approximately. The iterative steps are as follows:

$$I_{\tau+1} = I_\tau - \frac{\epsilon^2}{2}[I_\tau - B_{\omega,\delta(I,\omega)}] + \epsilon U_t$$  \hspace{1cm} (6)

Where $\epsilon$ represents the training step length, $U_t$ represents the Gaussian white noise in the model, and $\tau$ is the time steps index. The $\omega$ can be updated with the computation $\frac{\partial f(I_m; \omega)}{\partial \omega}$ and $\frac{\partial f(I_m; \omega)}{\partial \omega}$ using the following approximation:

$$\frac{\partial L(\omega)}{\partial \omega} \approx \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial \omega} f(I_m; \omega) - \frac{1}{M} \sum_{m=1}^{M} \frac{\partial}{\partial \omega} f(I_m; \omega)$$  \hspace{1cm} (7)

where synthesized video sequences $\{I_m, m = 1, \cdots, M\}$ are generated based on the Langevin dynamics which is a stochastic relaxation process. $\omega^{(t+1)}$ Update With step $\eta_t$:

$$\omega^{(t+1)} = \omega^{(t)} + \eta_t \dot{L}(\omega)$$  \hspace{1cm} (8)

3. Experiments

In our experiments, Matlab code of MatConvNet of [16] and MexConv3D of [17] are used. We run on a Intel (R) Core(TM)i7-8850H CPU@2.60GHz with 32 GB RAM. For comparison, we use public DynTex dataset and run 800 iterations to train the model. Experiments on texture images show that the method can learn and generate real natural patterns. The ConvNet parameters set up just liking [7].

Experiment 1: generating texture/object patterns. Experiments show that the method can learn and generate real natural image patterns. Figure 1 displays the training images, and figure 2 shows the synthesized sequence corresponding to the training images. The results show that our method can learn from images effectively and can handle all kinds of textures from random to structured.

Experiment 2: Generating dynamic textures. Figure 3 shows the results. For each category, six frames of the training sequence are showed in the first line, and the corresponding frames generated based on our technology are showed in the second and third lines.
Experiment 3: comparisons with other models. Figure 4 shows the results of the comparison. For each category, first row is the initial sequence, the other two rows are obtained sequences by our algorithm and Ref. [10] respectively. Compared with our method, Ref. [10] appears more blurry.

It can be seen from the experiments that our model has the ability to change its structure and parameters, and make its imagination fix the statistical characteristics involved in the model.
4. Conclusions
The application of adversarial ConvNet pattern in dynamic pattern synthesis is studied in this paper. We define probability distributions in the adversarial ConvNet to grasp spatial-temporal information and obtain the optimal results by maximizing the target likelihood function. Experiments showed that this method can generate patterns with different styles and colors more effectively.

Acknowledgments
This work is supported by the Joint Special Fund of Shandong Provincial Natural Science Foundation (No. ZR2017LF018).

References
[1] Kashyap R L 1981 Analysis and Synthesis of Image patterns by spatial interaction models Pattern Recognition 149-186.
[2] Zahner T, Lochbühler T, Mariethoz G, et al. 2017 Image synthesis with graph cuts: A fast model proposal mechanism in probabilistic inversion Geophysical Journal International 204 1179-90.
[3] Voelz D G, Gonglewski J D and Idell P S 1981 Image synthesis from nonimaged laser-speckle patterns: comparison of theory Computer Simulation, and Laboratory Results 30 3333-44.
[4] Kirchner M and Böhme R 2009 Synthesis of color filter array pattern in digital images The International Society for Optical Engineering.
[5] Xie J, et al. 2018 Cooperative training of descriptor and generator networks IEEE Transactions on Pattern Analysis & Machine Intelligence.
[6] Xie J, Lu Y, Zhu S C, et al. 2016 Cooperative training of descriptor and generator networks IEEE Transactions on Software Engineering.
[7] Vondrick C, Pirsiavash H and Torralba A 2016 Conf. on NIPS p 613.
[8] Ji S, Xu W, Yang M and Yu K 2013 3D convolutional neural networks for human action recognition IEEE Transactions on Pattern Analysis and Machine Intelligence 35 221-231.
[9] Wang Y and Zhu S C 2002 Conf. on ECCV p 583.
[10] Zhu Z, et al. 2016 Dynamic texture modeling and synthesis using multi-kernel Gaussian process dynamic model Signal Processing.
[11] Doretto G, Chiuso A, Wu Y N and Soatto S 2003 Dynamic textures Computer Vision 51 91-109.
[12] Xie J, Zhu S C and Wu Y N 2017 Conf. on Computer Vision and Pattern Recognition (CVPR).
[13] Xie Y L, Zhu S C and Wu Y N 2016 Conf. on ICML.
[14] Yann L, Chopra S, Hadsell R, Ranzato M and F Huang 2006 A tutorial on energy-based learning Predicting Structured Data.
[15] Ngiam J, Chen Z, Koh P W and Ng A Y 2011 Conf. on ICML p 1105.
[16] Vedaldi A and Lenc K 2014 Matconvnet-convolutional neural networks for matlab CoRR
Handbook on abs/1412.4564.

[17] https://github.com/pengsun/ MexConv3D.