ECCV'20
ONLINE
23-28 AUGUST 2020

16TH EUROPEAN CONFERENCE ON

COMPUTER VISION

WWW.ECCV2020.EU

Disentangling Multiple Features in Video Sequences using Gaussian Processes in Variational Autoencoders

Sarthak Bhagat[1], Shagun Uppal[1], Zhuyun Yin[2], Nengli Lim[3]

[1]IIIT Delhi
[2]Bioinformatics Institute, A*STAR, Singapore
[3]Singapore University of Technology and Design

Table of Contents

Introduction

Problem Statement: To disentangle multiple factors of variation simultaneously from video sequences.

- We propose MGP-VAE (Multi-disentangled-features Gaussian Processes Variational AutoEncoder), for the unsupervised learning of disentangled representations for video sequences.
- It utilizes a latent prior distribution that consists of multiple channels of fractional Brownian motions and Brownian bridges.

Variational Autoencoders

Variational autoencoders [9] are powerful generative models which reformulate autoencoders in the framework of variational inference.

- Given latent variables $z \in R^M$, the decoder, typically a neural network, models the generative distribution $p_\theta(x \mid z)$, where $x \in R^N$ denotes the data.
- Due to the intractability of computing the posterior distribution $p(z \mid x)$, an approximation $q_\varphi(z \mid x)$, again parameterized by another neural network called the encoder, is used.

Gaussian Processes

Given an index set $T = \{X_t; t \in T\}$ is a Gaussian Process [5, 14] if for any finite set of indices $\{t_1, \ldots, t_n\}$ of T, $(X_{t1}, \ldots, X_{tn})$ is a multivariate normal random variable.

- We are concerned primarily in the case where T indexes time, and the Gaussian Process $\{X_t; t \in T\}$ can be uniquely characterised by its mean and covariance functions

$$\mu(t) := E[X_t] \quad (3)$$

$$R(s,t) := E[X_t X_s], \quad \forall\, s,t \in T. \quad (4)$$

- The prior distributions employed in MGP-VAE are the appropriately discretized versions of two frequently encountered Gaussian processes in stochastic models, e.g. in financial modeling [1, 3], namely Fractional Brownian Motion (fBM) and Brownian Bridge (BB).

Fractional Brownian Motion (fBMs)

fBMs [10] $\{B_t^H; t \in T\}$ are Gaussian processes parameterized by a Hurst parameter $H \in (0, 1)$, with mean and covariance functions given by

$$\mu(t) = 0, \tag{5}$$

$$R(s,t) = \frac{1}{2}\left(s^{2H} + t^{2H} - |t-s|^{2H}\right), \quad \forall\, s,t \in T. \tag{6}$$

- When $H = 1/2$, $W_t = B_t^{1/2}$ is standard Brownian motion [5] with independent increments.
- Most notably, when $H \neq 1/2$, the process is not Markovian.
  - when $H > 1/2$, the disjoint increments of the process are positively correlated,
  - whereas when $H < 1/2$, they are negatively correlated.

Brownian Bridges (BBs)

The Brownian bridge [3, 8] from $a \in$ R to $b \in$ R on the domain $[0, T]$ is the Gaussian process defined as

$$X_t = a\left(1 - \frac{t}{T}\right) + b\left(\frac{t}{T}\right) + W_t + \frac{t}{T}W_T. \qquad (7)$$

- Its mean function is identically zero and its covariance function is given by

$$R(s,t) = \min(s,t) - \frac{st}{T}, \quad \forall\, s,t \in T. \qquad (8)$$

- From (7), its defining characteristic is that it is pinned at the start and the end such that $X_0 = a$ and $X_T = b$ almost surely.

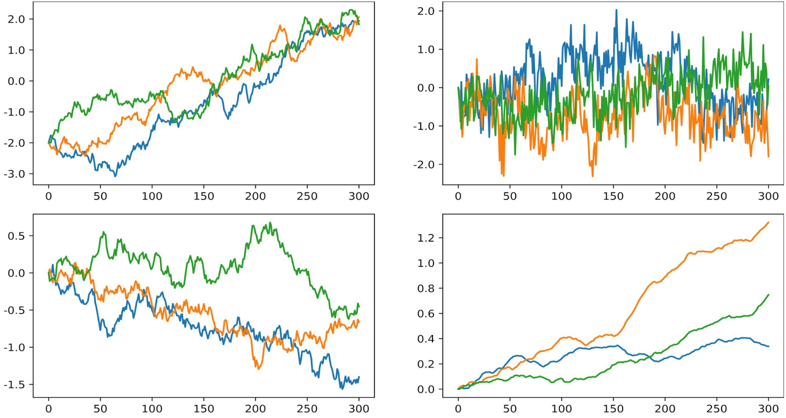


Figure: Sample paths for various Gaussian processes. Top-left: Brownian bridge from -2 to 2; top-right: fBM with H = 0.1; bottom-left: standard Brownian motion; bottom-right: fBM with H = 0.9

MGP-VAE

- For VAEs in the unsupervised learning of static images, the latent distribution $p(z)$ is typically a simple Gaussian distribution, i.e. $z \sim N(0, \sigma^2 I_d)$.
- For a video sequence input $(x_1, \ldots x_n)$ with n frames, we model the corresponding latent code as

$$z = (z_1, z_2, \ldots, z_n) \sim \mathcal{N}(\mu_0, \Sigma_0), \quad z_i \in \mathbb{R}^d, \tag{9}$$

$$\mu_0 = \left[\mu_0^{(1)}, \ldots, \mu_0^{(d)}\right] \in \mathbb{R}^{n \times d}, \tag{10}$$

$$\Sigma_0 = \left[\Sigma_0^{(1)}, \ldots, \Sigma_0^{(d)}\right] \in \mathbb{R}^{n \times n \times d} \tag{11}$$

- Here d denotes the number of channels, where one channel corresponds to one sampled Gaussian path, and for each channel, $\mu_0^{(i)}$, $\Sigma_0^{(i)}$ are the mean and covariance of

$$V + \sigma B_t^H, \quad t = \{1, \ldots, n\}, \tag{12}$$

in the case of fBM or

$$A\left(1 - \frac{t}{n}\right) + B\left(\frac{t}{n}\right) + \sigma\left(W_t + \frac{t}{n} W_n\right) \tag{13}$$

in the case of Brownian bridge.

- V , A are initial distributions, and B is the terminal distribution for Brownian bridge. They are set to be standard normal, and we experiment with different values for σ.

- The covariances can be computed using (6) and (8) and are not necessarily diagonal, which enables us to model more complex inter-frame correlations.

- The output of the encoder is a mean vector μ_1 and a symmetric positive-definite matrix Σ_1, i.e.

$$q(z \mid x) \sim \mathcal{N}(\mu_1, \Sigma_1), \tag{14}$$

- To compute the KL divergence term for the variational autoencoder loss, we use the formula

$$D_{KL}\left[q \mid p\right] = \frac{1}{2}\left[\operatorname{tr}\left(\Sigma_0^{-1}\Sigma_1\right) + \langle \mu_1 - \mu_0, \Sigma_0^{-1}(\mu_1 - \mu_0)\rangle - k + \log\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)\right]. \tag{15}$$

- Following [6], we add a β factor to the KL divergence term to improve disentanglement.

Network Architecture

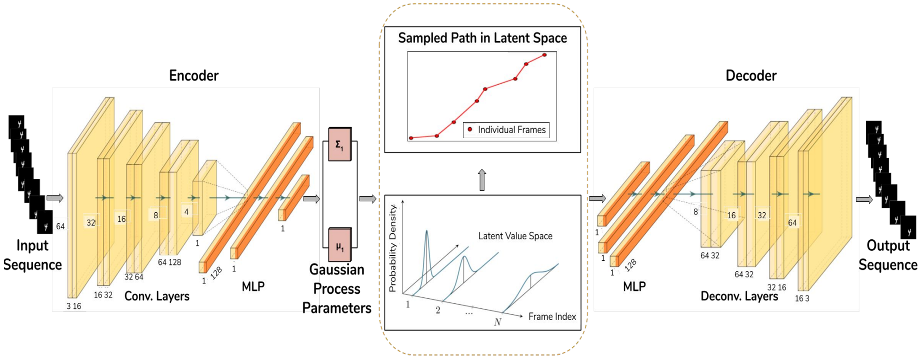


Figure: Network illustration of MGP-VAE

Video Frame Prediction

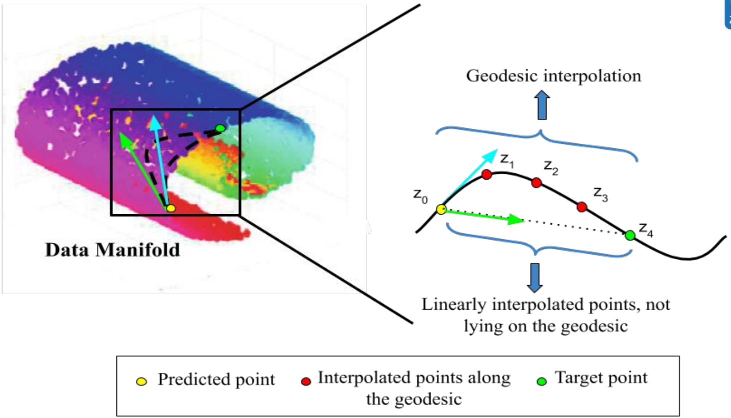


Figure: Using the geodesic loss function as compared to squared-distance loss for prediction.

Geodesic Loss

We use the following algorithm from [11] to compute the geodesic distance.

Algorithm 1: Geodesic Interpolation

Input: Two points, $z_0, z_T \in Z$; α, the learning rate

Output: Discrete geodesic path, $z_0, z_1, \ldots, z_T \in Z$ Initialize z_i as the linear interpolation between z_0 and z_T

while $\Delta E_{zt} > \varepsilon$ **do**

  for $i \in \{1, 2, \ldots, T-1\}$ **do**

    Compute gradient using (17)

    $z_i \leftarrow z_i - \alpha \nabla_{zt} E_{zt}$

  end for

end while

$$E_{z_t} = \frac{1}{2} \sum_{i=0}^{T} \frac{1}{\delta t} |g(z_{i+1}) - g(z_i)|^2 \quad (16)$$

by computing its gradient

$$\nabla_{z_t} E_{z_t} = -(\nabla g(z_i))^T [g(z_{i+1}) - 2g(z_i) + g(z_{i-1})] \quad (17)$$

Experiments

- Disentanglement
  - Attribute Transfer

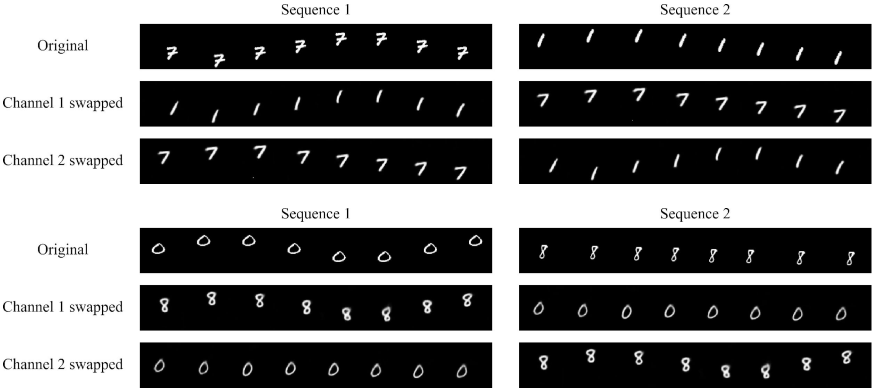
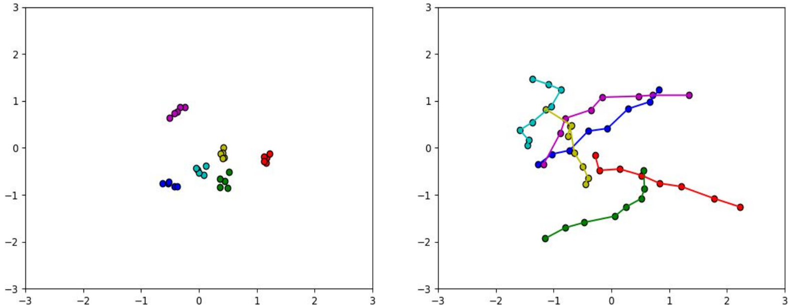


Figure: Results from swapping latent channels in Moving MNIST; channel 1 (fBM($H = 0.1$)) captures digit identity; channel 2 (fBM($H = 0.9$)) captures motion.

- Latent Space Visualisation



(a) fBM, H = 0.1

(b) fBM, H = 0.9

Figure: Latent space visualization of fBM channels for 6 videos. Each point represents one frame of a video. The more tightly clustered points in (a) capture digit identity whereas the scattered points in (b) capture motion.

- Attribute Transfer

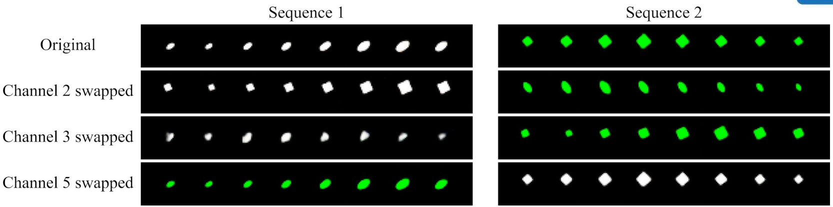


Figure: Results from swapping latent channels in Coloured dSprites; channel 2 captures shape, channel 3 captures scale, channel 4 captures orientation and position, and channel 5 captures color.

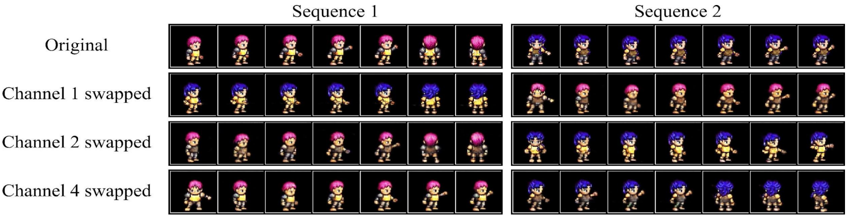


Figure: Results from swapping latent channels in Sprites; channel 1 captures hair type, channel 2 captures armor type, channel 3 captures weapon type, and channel 4 captures body orientation.

- Qualitative results of MGP-VAE and baselines in the video prediction task. Predicted frames are marked in red, and the first row depicts the original video sequence.

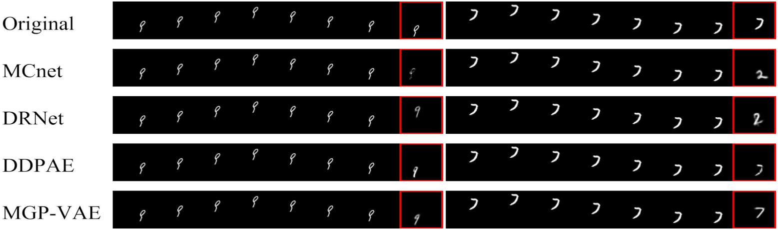


Figure: Moving MNIST

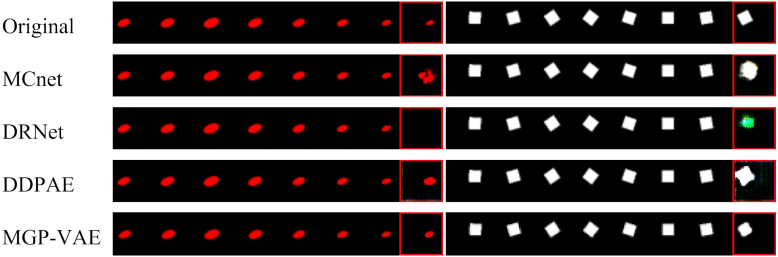


Figure: Colored dSprites

Disentanglement

Table: mAP values (%) for Coloured dSprites

Model	Coloured dSprites					
	Shape	Scale	Colour	x-Pos	y-Pos	Avg.
MCnet [13]	95.6	69.2	94.0	69.7	70.2	79.7
DRNet [2]	95.7	69.6	94.8	72.4	70.6	80.6
DDPAE [7]	95.6	70.3	94.2	71.6	72.4	80.8
MGP-VAE	96.2	77.9	94.0	76.4	72.8	**83.4**

Video Frame Prediction

Table: Prediction results on Moving MNIST

Model	$k = 1$		$k = 2$	
	MSE	BCE	MSE	BCE
MCnet [13]	50.1	248.2	91.1	595.5
DRNet [2]	45.2	236.7	86.3	586.7
DDPAE [7]	35.2	201.6	75.6	556.2
Grathwohl, Wilson [4]	59.3	291.2	112.3	657.2
MGP-VAE	25.4	198.4	72.2	554.2
MGP-VAE (with geodesic loss)	**18.5**	**185.1**	**69.2**	**531.4**

Figure: Without geodesic loss

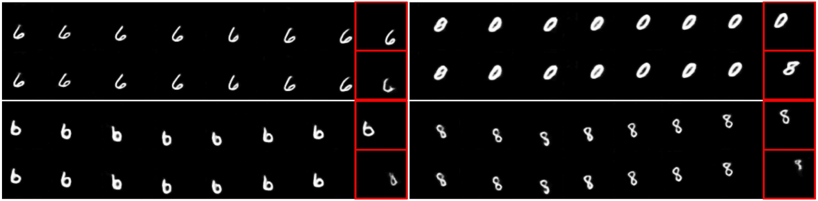


Figure: With geodesic loss

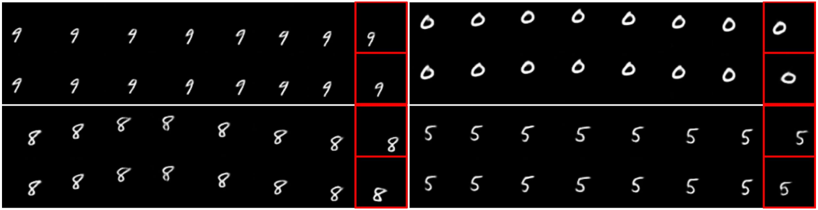


Figure: Comparison between predictions with and without using the geodesic loss function for Moving MNIST.

- For more details, please check our paper: https://arxiv.org/pdf/2001.02408.pdf
- The code for the paper is available at: https://github.com/SUTDBrainLab/MGP-VAE

Thank You!

References I

1 C. Bayer, P. Friz, and J. Gatheral. Pricing under rough volatility. *Quantitative Finance*, 16(6):887–904, 2016.

2 E. L. Denton and V. Birodkar. Unsupervised Learning of Disentangled Representations from Video. In *Conference on Neural Information Processing Systems (NIPS)*, 2017.

3 P. Glasserman. *Monte-Carlo Methods in Financial Engineering*. Springer-Verlag, NY, 2003.

4 W. Grathwohl and A. Wilson. Disentangling Space and Time in Video with Hierarchical Variational Auto-encoders. *ArXiv*, abs/1612.04440, 2016.

5 T. Hida and M. Hitsuda. *Gaussian Processes*. American Mathematical Society, 2008.

6 I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations (ICLR)*, 2017.

7 J.-T. Hsieh, B. Liu, D.-A. Huang, F.-F. Li, and J. C. Niebles. Learning to Decompose and Disentangle Representations for Video Prediction. *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

References II

8 I. Karatzas and S. E. Shreve.
Brownian Motion and Stochastic Calculus. Springer-Verlag, NY, 1998.

9 D. P. Kingma and M. Welling. Auto-Encoding Variational
Bayes.
In *International Conference on Learning Representations (ICLR)*, 2013.

10 B. B. Mandelbrot and J. W. Van Ness.
Fractional brownian motions, fractional noises and applications.
SIAM Review, 10(4):422–437, 1968.

11 H. Shao, A. Kumar, and P. T. Fletcher.
The Riemannian Geometry of Deep Generative Models.
In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

12 J. Stu¨hmer, R. E. Turner, and S. Nowozin.
Independent Subspace Analysis for Unsupervised Learning of Disentangled Representations.
In *AISTATS*, 2020.

13 R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing Motion and Content for
Natural Video Sequence Prediction.
International Conference on Learning Representations (ICLR), 2017.

14 C. K. Williams and C. E. Rasmussen.
Gaussian Processes for Machine Learning, volume 2. MIT press Cambridge, MA, 2006.