Mutual Information based Method for Unsupervised Disentanglement of Video Representation

International Conference on Pattern Recognition 2020

P Aditya Sreekar, Ujjwal Tiwari and Anoop Namboodiri

Center for Visual Information Technology
International Institute of Information Technology, Hyderabad

January 15, 2021

What is Video Prediction?

Video Prediction is a challenging but interesting task of predicting future frames from a given set context frames that belong to a video sequence.

Motivation and Perspective Applications

- Vision based reinforcement learning, Simulation
- Mobile robotics, Autonomous Navigation and Maneuver Planning
- Frame Reconstruction and Image Denoising

Key Challenges and Concurrent Work

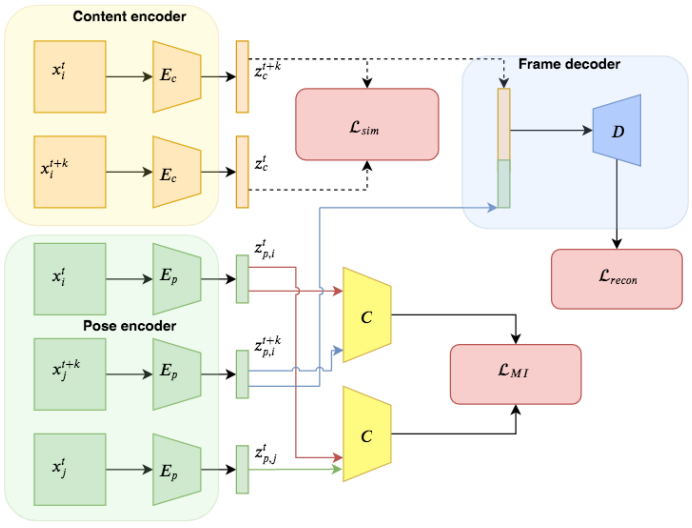
One of the major challenges in future frame generation is due to the high dimensional nature of visual data

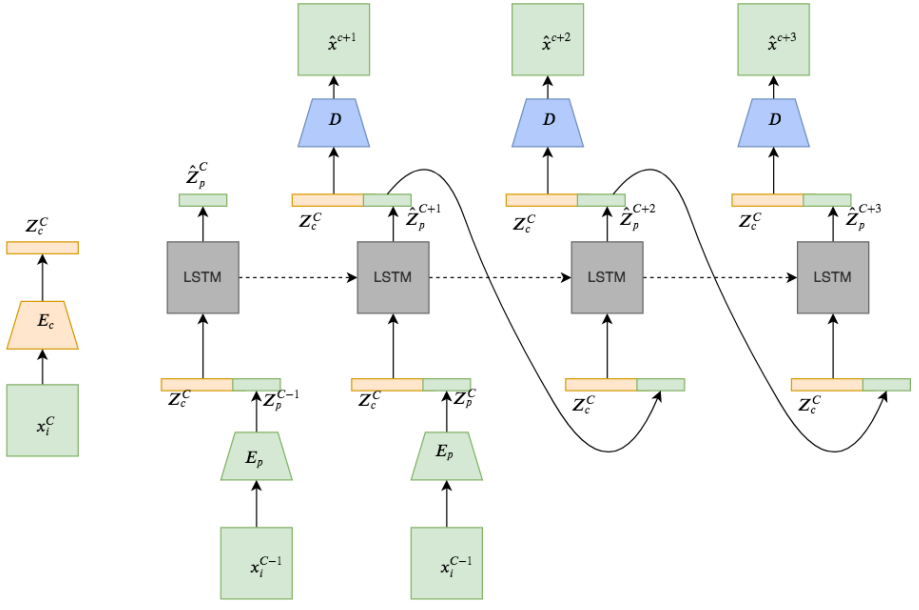
- Concurrent video prediction methods overcome this challenge by factorising video representations into a low dimensional temporally varying component and another temporally invariant component.
- Tulyakov *et al.*(Tulyakov et al. 2018) factorised video representations into time dependent pose and time independent content representations
- Vondrick-Pirsiavash *et al.*(Vondrick, Pirsiavash, and Torralba 2016) decomposed video into salient (foreground) and non-salient (background) regions
- DRNET (Denton et al. 2017) disentangled video representation into time dependent pose and time independent content latent representation

Our Framework and Assumption

- **Assumption:** Two stage video generation process.
- We propose Mutual Information Predictive Auto-Encoder (MIPAE) framework for predicting future frames of a video sequence. MIPAE framework reduces the task of predicting high dimensional video frames by factorising video representations into content and low dimensional pose latent variables.
- Content and the predicted pose representations then decoded to generate future frames.
- We also propose a Mutual Information Gap (MIG) metric to quantitatively access and compare the effectiveness in disentanglement of latent representation.

Automated Spectral Kernel Learning





How do we achieve proper disentanglement?

1. We leverage the temporal structure in latent generative factors by applying the following three loss functions in video prediction architecture shown in fig. 5:
   - **Similarity Loss** $\mathcal{L}_{sim}$ between the content latent representations z_c of different frames from a given sequence.
   - **Mutual Information Loss** $\mathcal{L}_{MI}$ is minimized between the pose latent representations z_p^t across time.
   - **Reconstruction Loss** $\mathcal{L}_{recon}$, which is l_2 reconstruction error $\mathcal{L}_{recon}$ is minimised between the ground truth and decoded frame to ensure proper reconstruction.

Mathematical Formulation of Objective Functions

- **Similarity Loss :** Time invariance of content representation is enforced by penalizing change in content representation between two different frames from the same video sequence that are separated by random offset $k \in [0, K]$ time steps:

$$\mathcal{L}_{sim} = \mathbb{E}_{P(x^t, x^{t+k})} \left[\| E_c(x^t) - E_c(x^{t+k}) \|_2^2 \right] \quad (1)$$

- Reconstruction Loss : Pixel-wise l_2 loss is minimized between decoded frame $D(E_c(x^t), E_p(x^t))$ and the ground truth frame x^t:

$$\mathcal{L}_{recon} = \mathbb{E}_{P(x^t)} \left[\| D(E_c(x^t), E_p(x^t)) - x^t \|_2^2 \right] \quad (2)$$

Mathematical Formulation of Objective Functions

- Mutual Information Loss : For estimating the mutual information between z_p^t and z_p^{t+k}, we train a critic C to classify whether z_p^t and z_p^{t+k} are sampled from joint distribution $P(z_p^t, z_p^{t+k})$ or the product of marginal distributions $P(z_p^t)P(z_p^{t+k})$ by using the standard GAN discriminator objective, which is maximized for the optimal critic:

$$\begin{aligned}\mathcal{L}_C = \mathbb{E}_{P(x^t, x^{t+k})}\left[\sigma(C(E_p(x^t), E_p(x^{t+k})))\right] \\ +\mathbb{E}_{P(x^t)P(x^{t+k})}\left[1-\sigma(C(E_p(x^t), E_p(x^{t+k})))\right]\end{aligned} \tag{3}$$

- We use a variational lower bound estimates of MI to enforce mutual information loss,

$$\begin{aligned}\mathcal{L}_{MI} = \mathbb{E}_{P(z_p^t, z_p^{t+k})}\left[C(z_p^t, z_p^{t+k})\right] \\ -\mathbb{E}_{P(z_p^t)P(z_p^{t+k})}\left[\exp(C(z_p^t, z_p^{t+k}))\right]\end{aligned} \tag{4}$$

- Minimizing this MI estimate, restricts E_p from encoding any content information.

Overall Training Objective

The overall training objective for E_c, E_p and D is as follows:

$$\min_{E_c,E_p,D} \mathcal{L}_{recon} + \alpha\mathcal{L}_{sim} + \beta\mathcal{L}_{MI} \tag{5}$$

Training object for the critic C is given by:

$$\max_{C} \mathcal{L}_{C} \tag{6}$$

Training Procedure

- The LSTM L is trained separately after training the main network, E_c, E_p and D.
- To predict a future frame $\hat{x}^t$, first, the LSTM L predicts $\hat{z}_p^t$ from previous frame's pose $\tilde{z}_p^{t-1}$ and content representation z_c^C of the last known frame x^C.

$$\hat{z}_p^t = L(z_c^C, \tilde{z}_p^{t-1}) \; where \; \tilde{z}_p^t = \begin{cases} E_p(x^t) & t < C+1 \\ L(z_c^C, \tilde{z}_p^{t-1}) & t \geq C+1 \end{cases} \quad (7)$$

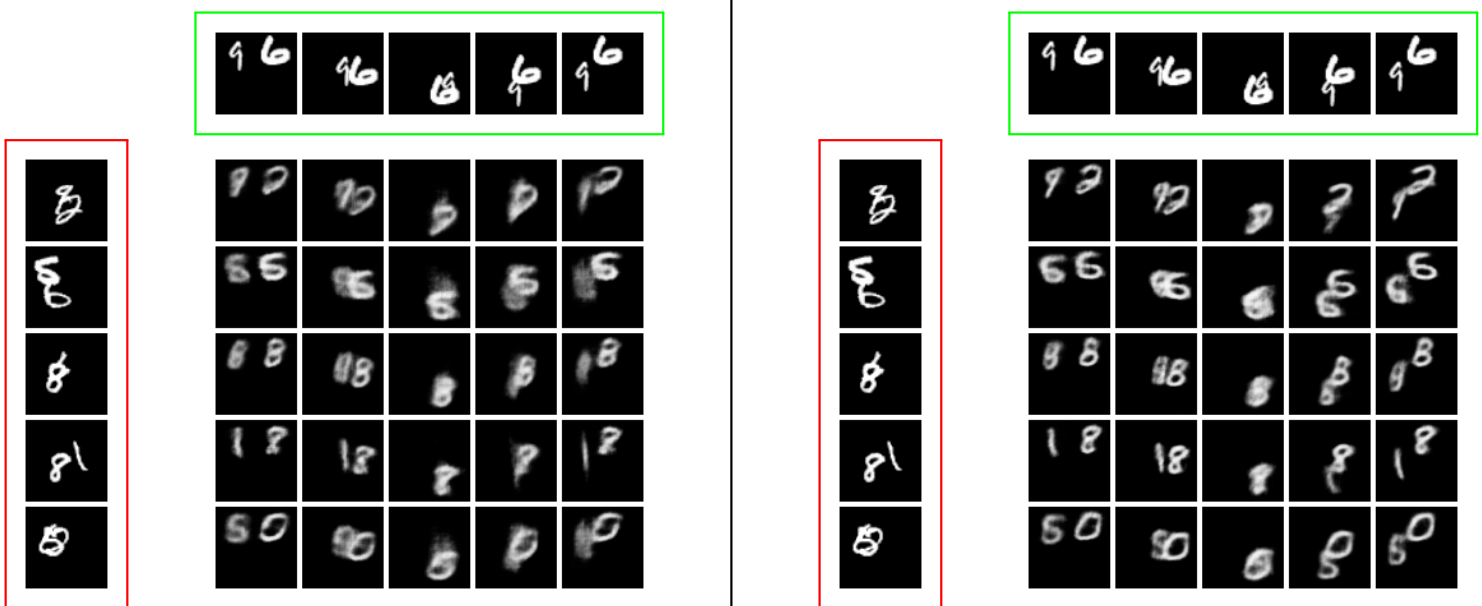
- The training objective for L is to minimize the l_2 loss between predicted poses, $\hat{z}_p^{2:C+T}$, and poses inferred from ground truth frames, $z_p^{2:C+T}$.
- Decoder D is used to generate the future frame $\hat{x}^t$ from the content z_c and the predicted pose representation $\hat{z}_p^t$ of the future frame, such that $\hat{x}^t = D(z_c^C, \hat{z}_p^t)$.

MIG Metric

- Concurrent evaluation method, for example, latent traversal are effective in finding methods that are unable to disentangle the generative factors of data but do not provide any quantitative measure of the effectiveness of disentanglement.
- MIG can be used in scenarios where mutual information can be calculated (i.e where factors of data generation are known a priori).
- In our adaptation of the MIG metric for video prediction, mutual information is calculated between generative factors and the learned pose, content representations:

$$MIG = \frac{0.5}{H(f_c)}\Big(I(f_c, z_c) - I(f_c, z_p)\Big) + \frac{0.5}{H(f_p)}\Big(I(f_p, z_p) - I(f_p, z_c)\Big) \tag{8}$$

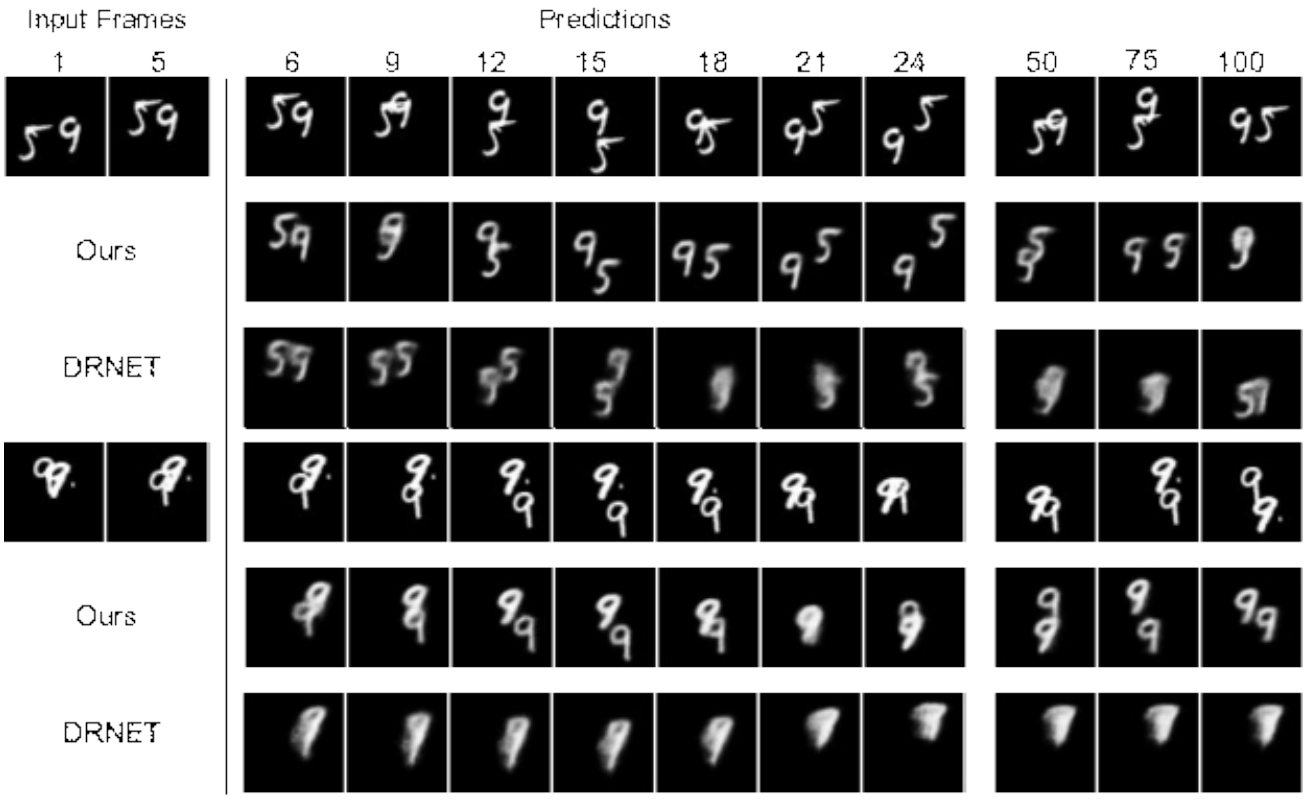
Results



(a) DRNET (b) MIPAE

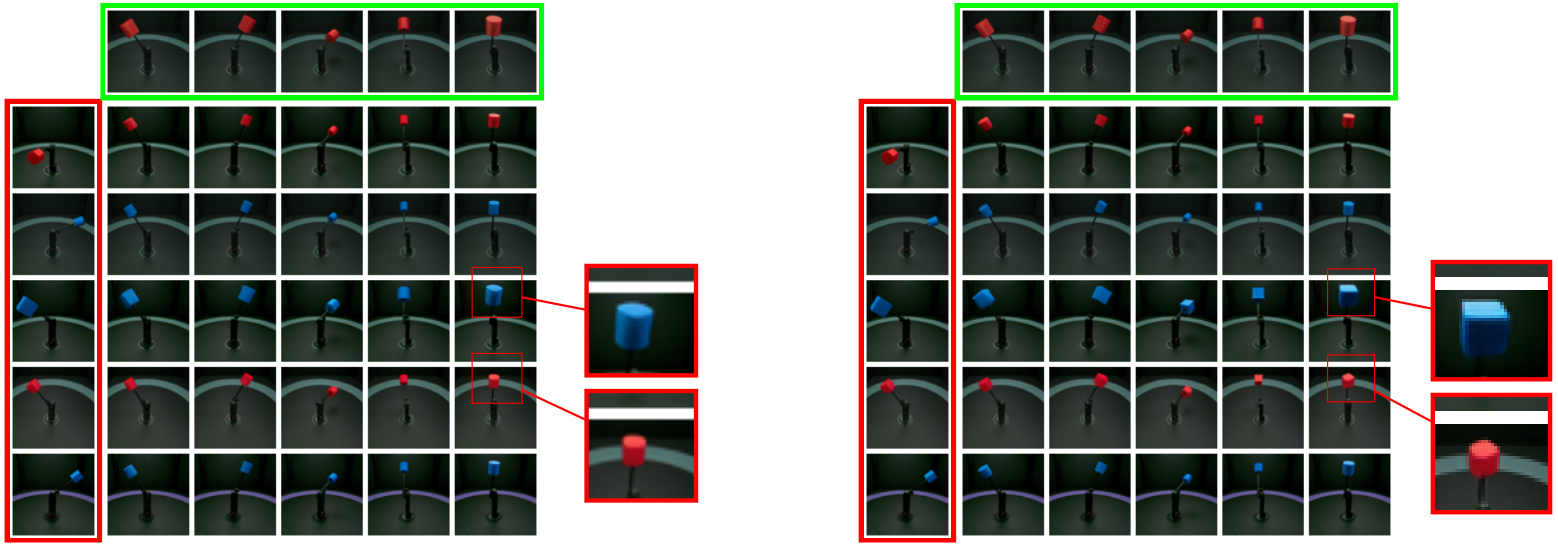
Qualitative comparison on moving MNIST dataset

Results



Qualitative comparison of video prediction on moving MNIST dataset

Results



(a) DRNET (b) MIPAE

Qualitative comparison of disentanglement on MPI 3D Real

Results

Table: MIG Scores

Dataset	Experiment	$I(f_c, z_c)$	$I(f_c, z_p)$	$I(f_p, z_c)$	$I(f_p, z_p)$	MIG
Dsprites	DRNET	5.6476	0.7483	0.0748	6.3434	0.8574
	Ours	5.6992	0.4660	0.725	6.4977	**0.8975**
MPI3D Real	DRNET	8.1353	0.0376	0.0448	6.2029	0.5658
	Ours	8.3866	0.0461	0.0080	7.1034	**0.6126**

Code

- Please checkout our code at https://github.com/blackPython/mipae

Thank You

Literature I

- Denton, Emily L et al. (2017). "Unsupervised learning of disentangled representations from video". In: *Advances in neural information processing systems*, pp. 4414–4423.
- Tulyakov, Sergey et al. (June 2018). "MoCoGAN: Decomposing Motion and Content for Video Generation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba (2016). "Generating Videos with Scene Dynamics". In: *ArXiv* abs/1609.02612.