http://dmirlab.com

Improved Conditional VRNNs for Video Prediction

Lluis Castrejon*
Mila, Université de Montréal
lluis.castrejon@gmail.com

Nicolas Ballas
Facebook AI Research
nballas@fb.com

Aaron Courville
CIFAR, Mila, Université de Montréal
aaron.courville@umontreal.ca

Hao Zhang

2020/02/25

Outline

- Short-tem prediction
- Long-term prediction

Improved VRNN

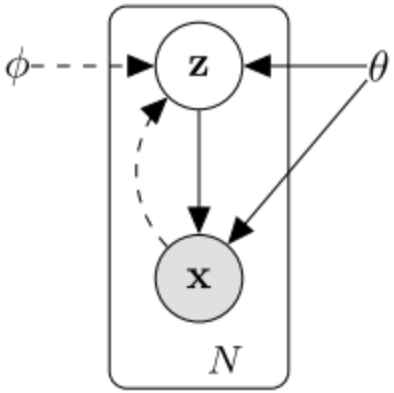
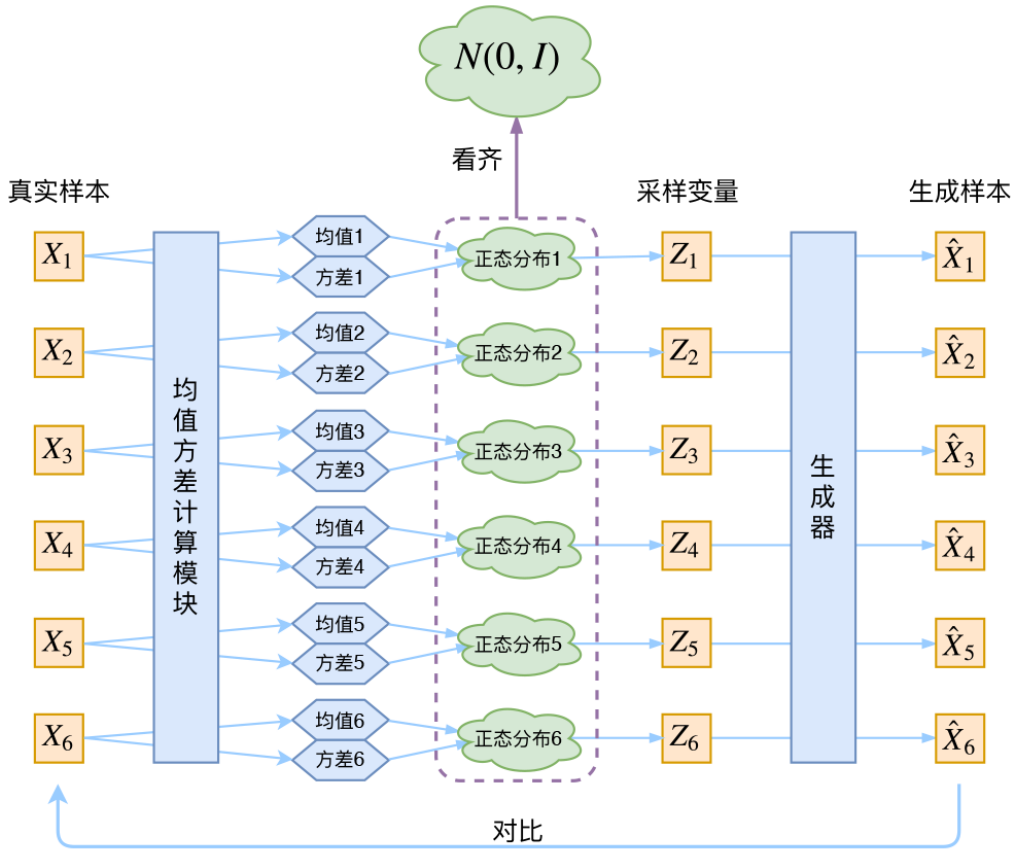


Figure 1: The type of directed graphical model under consideration. Solid lines denote the generative model $p_{\boldsymbol{\theta}}(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, dashed lines denote the variational approximation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to the intractable posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. The variational parameters $\boldsymbol{\phi}$ are learned jointly with the generative model parameters $\boldsymbol{\theta}$.

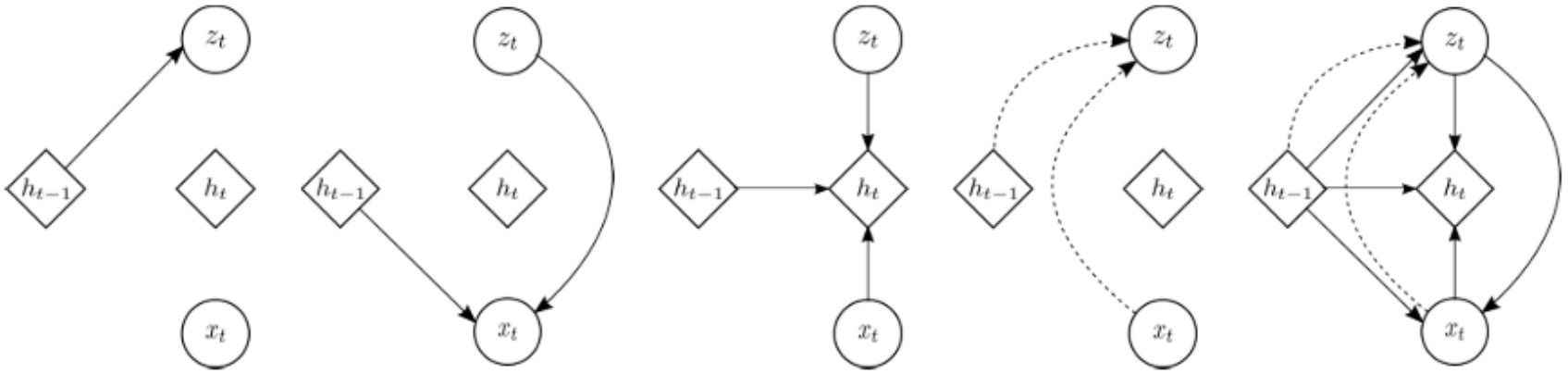
Improved Conditional VRNNs for Video Prediction (Lluis et al, ICCV2019)

Improved VRNN



Improved Conditional VRNNs for Video Prediction (Lluis et al, ICCV2019)

Improved VRNN



(a) Prior (b) Generation (c) Recurrence (d) Inference (e) Overall

Figure 1: Graphical illustrations of each operation of the VRNN: (a) computing the conditional prior using Eq. (5); (b) generating function using Eq. (6); (c) updating the RNN hidden state using Eq. (7); (d) inference of the approximate posterior using Eq. (9); (e) overall computational paths of the VRNN.

Improved Conditional VRNNs for Video Prediction (Lluis et al, ICCV2019)

Improved VRNN

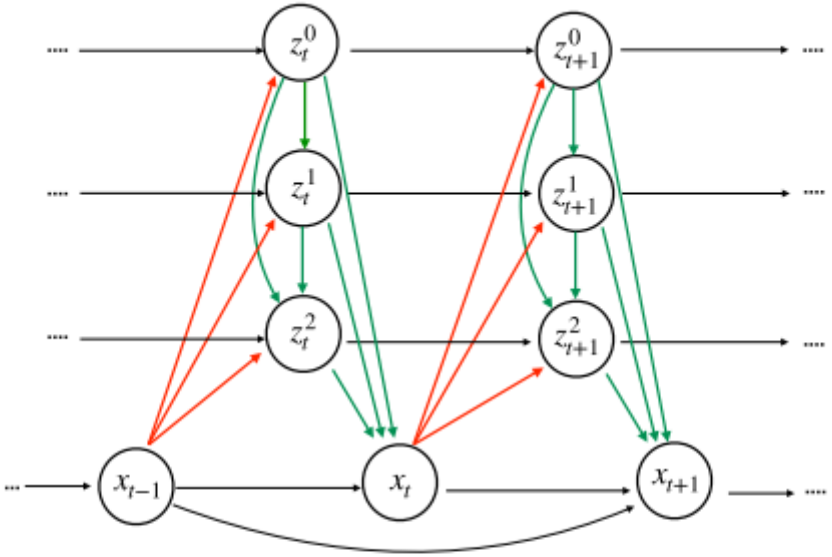


Figure 2: **Graphical model for the learned prior with the dense latent connectivity pattern**. Arrows in red show the connections from the input at the previous timestep to current latent variables. Arrows in green highlight skip connections between latent variables and connections to outputs. Arrows in **black** indicate recurrent temporal connections. We empirically observe that this dense-connectivity pattern eases the training of latent hierarchies.

Improved Conditional VRNNs for Video Prediction (Lluis et al, ICCV2019)

Improved VRNN

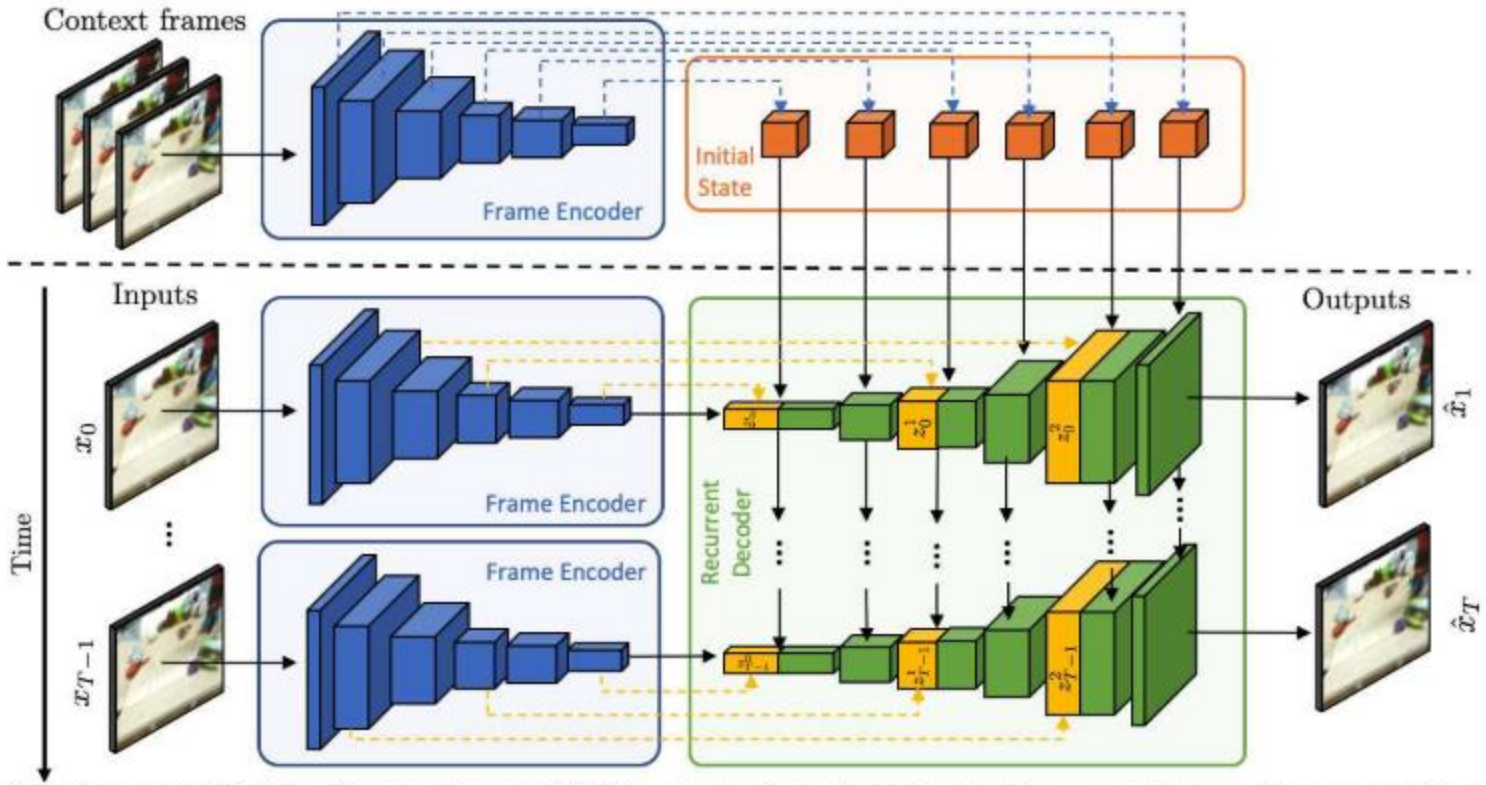


Figure 3: **Model Parametrization.** Our model uses a CNN to encode frames individually. The representation of the context frames is used to initialize the states of the prior, posterior and likelihood networks, all of which use recurrent networks. At each timestep, the decoder receives an encoding of the previous frame, a set of latent variables (either from the prior or the posterior) and its previous hidden state and predicts the next frame in the sequence.

Improved Conditional VRNNs for Video Prediction (Lluis et al, ICCV2019)

Thank you !

Q&A