Stochastic Video Generation with a learned prior

Summary

1. Existing approach fails to capture the full distribution of possible future or yields blurry prediction.

2. They learn a prior of the uncertainty in a given environment

3. Future frames are predicted based on combining a samples from the prior and a deterministic prediction of the future.

Approach

1. The loss function is:

$$\mathcal{L}_{\theta,\phi}(x_{1:T}) = \sum_{t=1}^{T} \Big[\mathbb{E}_{q_\phi(z_{1:t}|x_{1:t})} \big[\log p_\theta(x_t | x_{1:t-1}, z_{1:t}) \big] - \beta D_{KL}(q_\phi(z_t | x_{1:t}) \,||\, p(z)) \Big]$$

where:
- p_θ is the prediction module that generates the next frame $\hat{x}_t$ based on previous ones in the sequence $x_{1:t-1}$ and a latent variable z_t.
- q_ϕ is an approximate inference network.
- $p(z)$ is a prior distribution

2. The loss function can also include a learned prior, the KL term in the loss becomes:

$$D_{KL}(q_\phi(z_t | x_{1:t}) \,||\, p_\psi(z_t | \underset{\uparrow}{x_{1:t-1}}))$$

does not depend on x_t, the frame being predicted.

3. The drawback of the fixed prior model is that the samples at each timestep will be drawn randomly, thus ignoring temporal dependencies between frames.

4. At test time, a frame is generated by at time t by:
- first sampling z_t from the prior $z_t \sim \mathcal{N}(0, I)$
  $z_t \sim p_\psi(z_t | x_{1:t-1})$
- a frame is generated by $\hat{x}_t = \mu_\theta(x_{1:t-1}, z_{1:t})$

mean of the prediction normal.

Architectures

1. p_θ, q_ϕ & p_ψ are generic convolutional LSTM
2. For a timestep t during training, the generation is, where the LSTM recurrence is omitted for brevity:

$$\mu_\phi(t), \sigma_\phi(t) = LSTM_\phi(h_t) \qquad h_t = Enc(x_t)$$

$$z_t \sim N(\mu_\phi(t), \sigma_\phi(t))$$

$$g_t = LSTM_\theta(h_{t-1}, z_t) \qquad h_{t-1} = Enc(x_{t-1})$$

$$\mu_\theta(t) = Dec(g_t)$$

the Enc & Dec are feed-forward conv. network.

3. In the learned prior model, the parameters of the prior distribution at time t is:

$$h_{t-1} = Enc(x_{t-1})$$

$$\mu_\psi(t), \sigma_\psi(t) = LSTM_\psi(h_{t-1})$$

Experiments

1. An interesting point is that they shown the variance of the learned prior accurately predicts the collision points in the Stochastic Moving MNIST experiments (figure 6)
2. They also add skip connection from the last real frame to the decoder to facilitate easy reconstruction of the background in the image sequences.