High Fidelity Video Prediction with Large Stochastic RNN

Summary

1. Large-scale empirical study &
2. demonstrate SOTA result for video prediction
3. Build upon SVG by Emily Denton

Evaluation Metrics

Frame-wise eval

1. Peak Signal-to-Noise Ratio (PSNR) } pixel-wise comparison
2. Structural Similarity (SSIM) }
3. VGG Cosine similarity & "perceptual" comparison at feature level

Dynamic-based eval

1. Fréchet Video Distance (FVD) : a 3D CNN trained for video classification is used to extract a single feature vector from a video.
2. Human Evaluations

Device details

1. They use 32 Google TPUv3 pods for each experiment and a batch size of 32.
2. There is a single batch element in each TPU.

Approach

1. They use the same stochastic component as SVG, with a few changes:
   - Shallower encoder-decoder that only have convolutional layers to enable more detailed image reconstruction.
   - A conv LSTM instead of fully-connected LSTM to fit the shallow encoder-decoder.
   - Optimize for l_1 instead of l_2 for img reconstruction loss.

2. To scale up the parameter, they use factor:
   - M for LSTM
   - K for the encoder & decoder

Ablations

1. Removing stochastic component, only deterministic component
2. Removing the LSTM, leaving behind the encoder-decoder architecture.
   The encoder observes the same no. of initial history as the LSTM

Experiments

1. Obj interactions with Ebert.et al. dataset, conditioned on action.
2. Structured motion with Human 3.6M dataset
3. Partial observability with KITTI driving dataset.