Stochastic Video Generation with a Learned Prior: A Review

Manav Prabhakar, Shiv Nadar University

The paper focusses on devising an efficient approach for generating video frames that accurately predict the future world states. The authors have proposed an unsupervised video generation model that learns a prior model of uncertainty in given environment. Video frames are then generated by drawing samples from this prior and combining them with a deterministic estimate of the future frame.

The approach seems to be built upon the Bayes properties of prior and posterior probabilities. Two variants of the model have been proposed, one that works on using a fixed prior (SVG-FP) and the other with a learned prior (SVG-LP).

The authors have discussed about the importance of a non-deterministic approach when dealing with future frame generation. Most of the deterministic approaches rely on a single possible future which may provide good results on the train set but may perform poorly on unseen data. The reason behind the same being that multiple future frames may be possible. The future is highly uncertain and thus, using a deterministic model particularly those with a deterministic loss reduce that uncertainty leading to deviation from ground truth.

The model has primarily two distinct components: a prediction model that generates the next frame $\hat{x_t}$(t^{th} frame) based on the previous ones in sequence (frames 1 to (t-1)) and a latent variable z_t. The next task is to fix the value of z_t which has been sampled from Gaussian distribution. To prevent overfitting which might occur if z_t copies the value of x_t, the authors have used a KL - divergence term. The model is recurrent in nature, implying that at any time step, the frame predictor has only x_{t-1} and z_t while x_{t-2} and $z_{1:t-1}$ are calculated recursively. The model architecture comprises of a ConvLSTM in which frames are input via a feed forward CNN. The convolutional frame decoder maps the output of the frame predictor's recurrent network back to pixel space.

The use of KL-divergence theorem for preventing overfitting seems to be an effective option. Further, the use of a stochastic approach rather than a deterministic approach is more intuitive and realistic which is evident from the better performance of the SVG when compared to deterministic models.

However, there can be different ways to determine the priors. A drawback is that samples at each time step will be drawn randomly, thus ignoring temporal dependencies present between frames.

Further, the authors talked about multiple possible scenarios and what they are dependent upon. An approach that works towards identifying those factors might prove to be more rewarding than what the authors have currently used. These factors include but are not limited to terrain type and kinematics of the object. These could be used deterministically to determine the future frames. We can provide weights to the possible scenarios based on the kinematics of the object and terrain. This would help us in reducing the possible frames while keeping the model non-deterministic.

The authors have proposed a proficient model which seems to provide competitive if not better results than existing approaches.