1 Topic and goal

In this project, we will focus on self-supervised video frame prediction problem by explicitly modeling the foreground objects, using video with less ambiguity on what the foreground and background objects are.

2 Related work and motivation

There are two main approaches to predict future frames. The first approach adopts a feedforward encoder-decoder architecture, encoding one or more frames (usually concatenated along the color channel) and decoding from latent space the future frame. For example, (Vukotic et al., 2017) uses a convolutional encoder-decoder network to predict the frame in the next time step given the frame of current time step.

The second approach involves recurrent network to capture temporal dependency. Recent years, researchers tend to use LSTM or GRU, as they shew satisfactory performance in capturing long-term dependency. (Srivastava et al., 2015) uses an LSTM encoder to encode the frames into fixed-length hidden states, which are then fed to LSTM decoder for video frame prediction. (Lotter et al., 2016) uses stacked convolutional LSTM, trained by predictive coding hypothesis in cognitive neural science, to perform future frame prediction.

However, other than predicting the frame pixel-wise, it could be possible to explicitly model higher level image elements, thus exploiting the structure of the latent space more efficiently. Video frame prediction with object (or object parts) awareness thus becomes an motivation of this project.

3 Dataset to use

Considering the self-supervised nature, we decide not to use videos that are too cluttered with objects or too ambiguous of telling the foreground from background. We haven’t explored widely which exactly dataset to use yet. UCF-101 and related video dataset could be a helpful choice. However, apparently we need to manually choose which videos could be a good fit.
References

