Unsupervised Keypoint Learning for Guiding Class-Conditional Video Prediction

Abraham Jose

1 Summary

The proposed model is for conditional generation of videos for actions guided by the action label and the extracted keypoint to predict the future frames. The network has 3 modules including keypoint detection, motion generation and keypoint-guided image translation. The method was able to bring the state of the art results in the video-generation for the action videos. The results are realistic and tested on datasets including Penn Action, UvA-NEMO and MGIF dataset.

2 Good points

Author was able to use multiple techniques effectively to reproduce robust and natural looking action videos which when compared to other techniques really stands out in terms of perceptual quality. Also, the method was compared to other techniques using AMT study and Frechet Video Distance. By employing random sampling in the latent space variable, z, they were able to introduce randomness to the actions as well. The background masking ensures that the result is consistent as well.

3 Weak points

The model was trained and tested in a really small dataset which is not capable of showing the capacity of the model and they have deliberately excluded the dataset with occlusion. Since the method relies on keypoint extraction, the performance on multiple objects can be tricky and with lot of errors in localizing keypoints internally. The generated video samples are significantly unique as compared to the
The model is significantly heavy with 3 neural network models involved for keypoint detection, motion generation and keypoint-guided image translation.
Also, the performance of the model depends on image translation network, which handles the dynamic regions based on soft background mask, which requires additional efforts to carefully get the soft masks, either require annotation or the existing masking technologies are not good enough to isolate high level of background noise.

4 Questions

How is FVD different from FID? in terms of how effectively it describes perceptual quality in video clips(of length = 16 or 32) compared to a averaging FID over images.

Even though the method is said to be pseudolabeled and unsupervised, for training the cVAE motion generator, the author uses the ground truth keypoint. It is confined for just testing the model.

5 Ideas

Using the method based on SIFT features rather than the keypoints on videos might help us to use the technique for many other video generation purposes which does not has any keypoints available to extract such a natural scenic videos.