Textual Video Prediction

Center for Research in Computer Vision
UNIVERSITY OF CENTRAL FLORIDA

Emily Cosgrove
ecosgrov@aum.edu
Auburn University at Montgomery

Amir Mazaheri
amirmazaheri@knights.ucf.edu
University of Central Florida

Dr. Murabak Shah
shah@crcv.ucf.edu
University of Central Florida

I. Problem

- **Video Prediction**
  - **Generate a sequence of frames**
- **Predicting Movie Clips**
  - **More Challenging**
    - **Camera motion**
    - **Many possible scene events**
  - **Audio Description**
    - **Can help to have better prediction**

II. Dataset

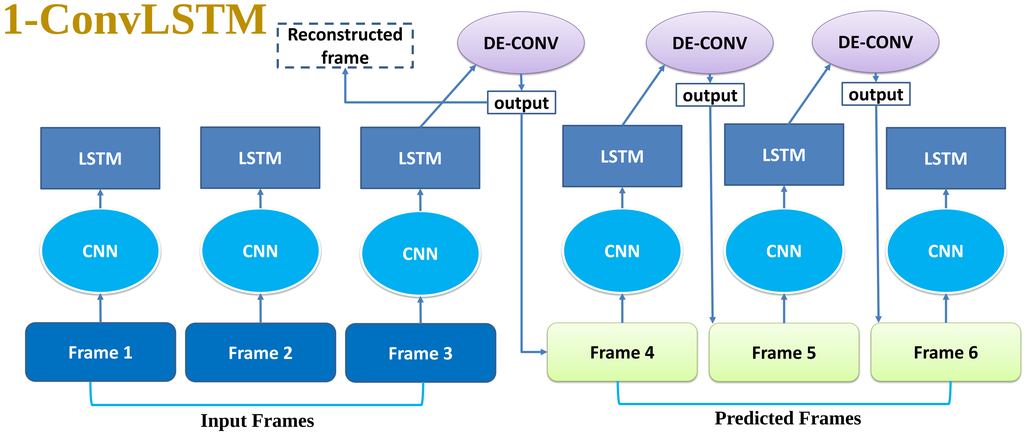
- **Large Scale Movie Description Challenge dataset[1]**
  - **128,000 video (2- 20 seconds)**
- **Each video is clipped into 1 second shots**
- **Shots with abrupt scene change are removed**
- **159,000 shots**
  - **90% for training and 5% for each of test and validation**
- **Each video comes with a description**
- **64 × 64 pixel frames**



Someone defensively grabs a picture frame and presses her back to a wall by a white-trim doorway.

III. Model/Methods

1-ConvLSTM

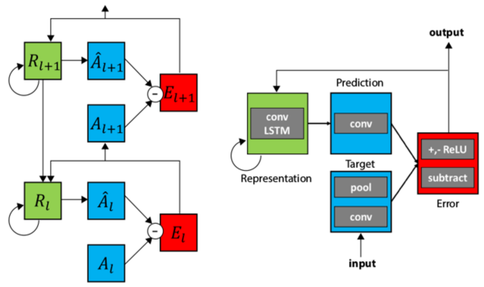


2-Spatial Transformer Network

1. **Predict the future motion map for each pixel**
2. **Copy pixel values from previous frame to new locations**

$$T_{x,y}^{t+1} = F\left(T_{x',y'}^{t}\right)$$
$$x' = x + \boldsymbol{t_x}$$
$$y' = y + \boldsymbol{t_y}$$

We predict t_x and t_y for each pixel using ConvLSTM.



3-PredNet [2]

1. **Based on predictive coding**
2. **Minimizes the reconstruction loss**

IV. Loss and Measurement

Loss:

$$\mathrm{MSE}(I, \tilde{I}) + \mathrm{MSE}(E_s(I) - E_s(\tilde{I}))$$

E_s: Sobel edge detector

I: Ground Truth **$\tilde{I}$: Prediction**

- **Preserve intensities**
- **Preserve edges**

Measurement:

Peak signal-to-noise ratio

$$\mathrm{PSNR} = 10 * log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

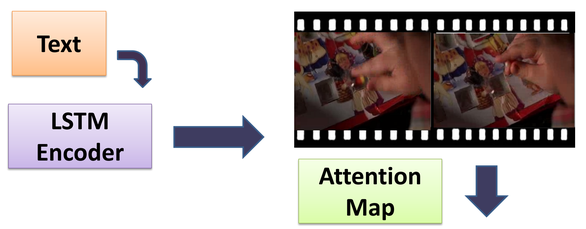
V. Quantitative Results

Method Name	PSNR
PREDNET	17.3
PREDNET	N/A
ConvLSTM	21.3
STN	**31.235**
ConvLSTM + Text	N/A

V. Qualitative Results



VI. Next Step



- **Separated Motions**
  - ***Global Motion*** **prediction (background)**
  - ***Foreground Motions***
    - **Computed per pixel**



VII. Conclusion

In this project, we tackle the video prediction problem. We gathered a more challenging dataset of movie clips. We showed that the baselines can fail due to the variety of events that can happen in our dataset. Finally, we introduced our method based on ConvLSTM and per pixel motions. Also, we will extend this work to use text to separate foreground and background motions.

VIII. References

[1] Maharaj, Tegan, et al. "A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering." ***arXiv preprint arXiv:1611.07810*** **(2016).**

[2] Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104 (2016)

IX. Acknowledgements

This work has been sponsored by the NSF grant number 100461121