### 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 x 64 pixel frames

### III. Model/Methods

#### 1-ConvLSTM
- LSTM
- Input Frames
- Predicted Frames
- DE-CONV

#### 2-Spatial Transformer Network
1. Predict the future motion map for each pixel
2. Copy pixel values from previous frame to new locations

\[
T_{x'y'} = F(T_{x',y'})
\]

\[
x' = x + \delta_x
\]

\[
y' = y + \delta_y
\]

We predict \(\delta_x\) and \(\delta_y\) for each pixel using ConvLSTM.

#### 3-PredNet [2]
1. Based on predictive coding
2. Minimizes the reconstruction loss

\[
\text{MSE}(I, \hat{I}) + \text{MSE}(E_s(I) - E_s(\hat{I}))
\]

- \(I\): Ground Truth
- \(\hat{I}\): Prediction

### IV. Loss and Measurement
- Loss:
- \(E_s\): Sobel edge detector

#### Loss:
- \(\text{MSE}(I, \hat{I}) + \text{MSE}(E_s(I) - E_s(\hat{I}))\)
  - Preserve intensities
  - Preserve edges

#### Measurement:
- Peak signal-to-noise ratio

\[
\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}_I^2}{\text{MSE}}\right)
\]

### V. Quantitative Results

<table>
<thead>
<tr>
<th>Method Name</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDNET</td>
<td>17.3</td>
</tr>
<tr>
<td>PREDNET</td>
<td>N/A</td>
</tr>
<tr>
<td>ConvLSTM</td>
<td>21.3</td>
</tr>
<tr>
<td>STN</td>
<td>31.235</td>
</tr>
<tr>
<td>ConvLSTM + Text</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### 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


### IX. Acknowledgements
This work has been sponsored by the NSF grant number 100461121