UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

Anomaly Detection in Surveillance Videos

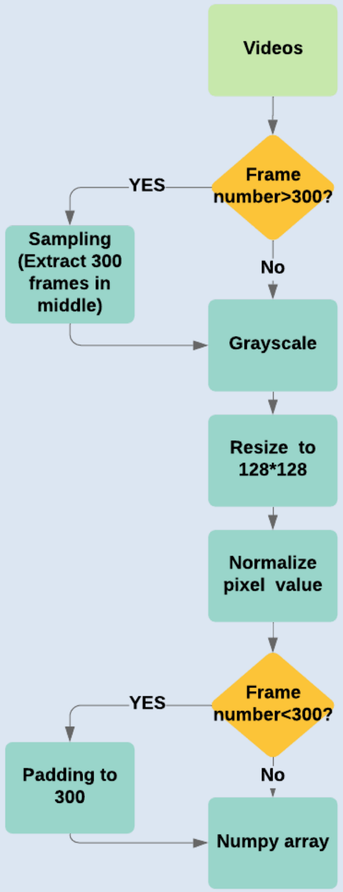
Aishwarya Phanse, Tariq Haque, Liya Zhang

INTRODUCTION

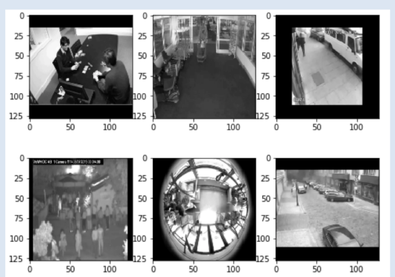
- Video surveillance gains an increasing demand to improve public safety, reduce crime rate, help catch criminals and curb illegal activities.
- The goal of our project is to identify anomalies in surveillance videos to enhance monitoring capabilities of law enforcement agencies to improve public safety.
- Some recent researches include:

- Detection of 13 anomalous activities from 128 hours, 1900 long and untrimmed surveillance videos by Waqas Sultani et al. using a deep Multiple Instance Learning ranking model with Conv 3D and TCNN.
- Deploying AnomalyNet consisting of motion fusion block and feature transfer block for anomaly detection using sparse long short-term memory (SLSTM) by researchers Joey Tianyi Zhou et al..

DATA PROCESS



376 videos



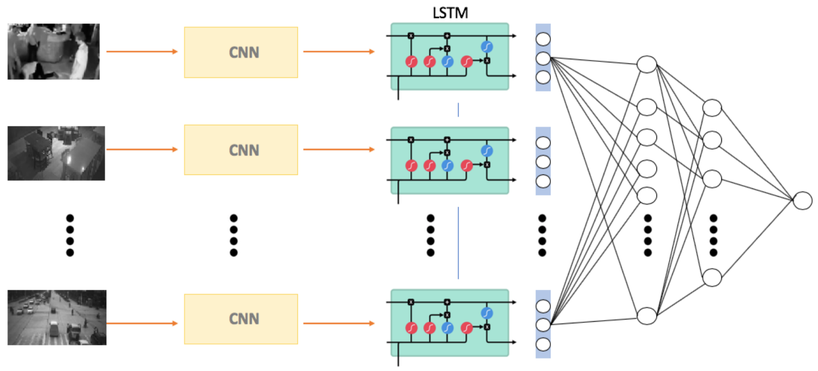
The shape of the training data is **(376, 300, 128, 128, 1)**

This indicates:

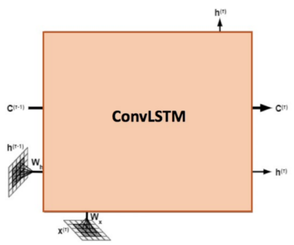
1. There are **376 videos**.
2. Each video has **300 frames**.
3. The dimension of **each frame is 128x128 pixels**.
4. **Only 1 channel** for each frame, i.e., all videos are gray scale.

METHODOLOGY

MODEL 1: CNN-LSTM combined CNN and LSTM layers where Conv2D (Keras) layers learn features in frames of a video followed by LSTM. To make sure that CNN is applied to each image of a video and passed to LSTM as a single time step, CNN layers are wrapped in a TimeDistributed Layer.

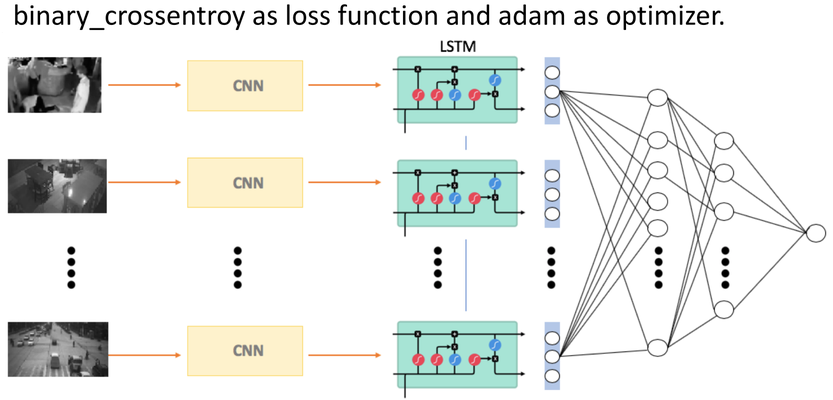


MODEL 2: ConvLSTM is a integration of CNN and LSTM where instead of doing matrix multiplication, convolution operation is conducted at each gate of a LSTM cell thereby, capturing the underlying spatial features by convolution operations in multiple-dimensional data.

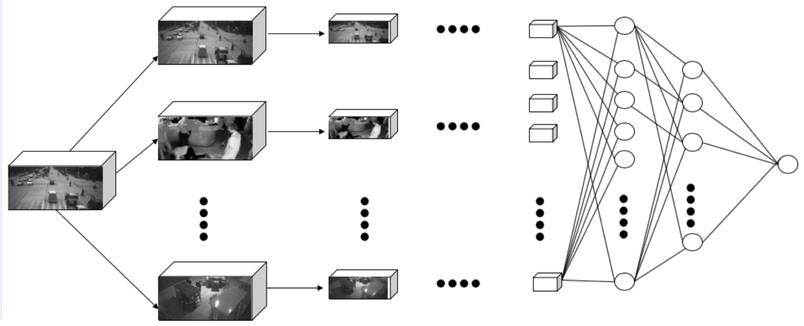


Layers (Type)	Output Shape	Number of Parameters
ConvLSTM2D	None, 300, 128,128,40	59,200
Batchnormalization	None, 300, 128,128,40	160
ConvLSTM2D	None, 300, 128,128,40	115,360
Batchnormalization	None, 300, 128,128,40	160
ConvLSTM2D	None, 300, 128,128,40	115,360
Batchnormalization	None, 300, 128,128,40	160
Flatten	None, 19660800	0
Dense	None, 1	19,660,801

MODEL 3: VGG-16 LSTM VGG-16 has 16 layers. We connected the last layer of the VGG-16 network with LSTM, then a flatten and output layer with sigmoid as activation function, binary_crossentroy as loss function and adam as optimizer.



MODEL 4: **Conv3D** 3D CNN can capture both temporal and spatial information and hence, are useful to build function to model videos that have both temporal and spatial information. In 3D convolution neural network, kernels are also 3 dimensional. Due to lack of memory the model may be disadvantaged in learning a long sequence.



RESULTS

Model	Model Input Shape	Validation Accuracy
CNN LSTM	(376, 300, 128, 128, 1)	53.98%
ConvLSTM	(376, 300, 128, 128, 1)	54.19%
VGG-16 LSTM	(376, 300, 128, 128, 1)	53.98%
Conv3D	(376, 300, 128, 128, 1)	53.98%

CONCLUSION

- We have achieved a significant progress in achieving a maximum accuracy of 54.19 % in prediction of anomalous videos.
- However, we believe that there is a significant need to carry this research forward especially by changing model architecture to improve prediction accuracy.
- In addition, of having used only 376 of the 1900 available videos due to computational restrictions; we believe that prediction can further improve if complete dataset is used.
- We would also like to highlight that we have sampled videos based on 300 frames in the middle of video. While, this is a regularly adopted method, other methods for sampling should also be explored in the pursuit of prediction accuracy.