ANOMALY DETECTION IN VIDEOS USING SPATIO-TEMPORAL AUTOENCODERS

Computer Vision Project Report

PRESENTED BY

Hadia Eman Mahdi

1. INTRODUCTION

Video anomaly detection is a crucial application in computer vision, with use cases ranging from surveillance systems to robotics. This project implements an approach for detecting anomalies in video sequences using a Spatio-Temporal Autoencoder. By reconstructing input video sequences and analyzing reconstruction errors, the model identifies patterns that deviate from expected behavior.

2. SPATIO-TEMPORAL AUTOENCODERS

2.1 CONCEPT OVERVIEW

A spatio-temporal autoencoder is a neural network architecture designed to learn compact representations of spatio-temporal data, such as videos. It comprises two main components:

1. Encoder: Compresses high-dimensional input data (video sequences) into a lower-dimensional latent space.
2. Decoder: Reconstructs the original input from the compressed latent representation.

This reconstruction is learned by minimizing the difference between the input and the reconstructed output. The architecture is particularly suited for detecting anomalies in videos because:

- Normal patterns (e.g., pedestrian motion) are reconstructed accurately.
- Abnormal patterns (e.g., sudden accidents) result in higher reconstruction errors.

2.2 PROJECT WORKING

- Input Representation:
- The model processes sequences of video frames (spatio-temporal data) as input.
- Each sequence consists of 10 consecutive frames resized to a fixed resolution (e.g., 64x64).
- Encoder:
- 3D Convolutions: Capture spatial and temporal features simultaneously.
- Layer-wise downsampling reduces spatial resolution while extracting hierarchical features.
- Latent Space:
- The encoder outputs a compact latent representation that encodes the most critical information from the video sequence.
- Decoder:
- 3D Transposed Convolutions: Gradually reconstruct the original input from the latent space.
- The reconstruction mimics the normal input sequence.
- Reconstruction Error:
- Reconstruction quality is quantified using Mean Squared Error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$$

- Higher errors indicate potential anomalies.

3. IMPLEMENTATION DETAILS

3.1 DATASET AND FEATURES

The Avenue Dataset is used, which contains surveillance video sequences of normal and anomalous events. The dataset features:

- Training Data: Video sequences with only normal events.
- Test Data: Video sequences containing normal and anomalous events.
- Key Features:
  - Videos are split into frame sequences.
  - Each frame is resized to 64×6464 \times 6464×64 pixels for uniform processing.
  - Frames are normalized to [0, 1] for better model performance.

3.2 DATA PREPROCESSING

The preprocessing pipeline:

- Frame Extraction: Frames are extracted from each video.
- Normalization: Pixel values are scaled between 0 and 1.
- Sequence Formation: Consecutive frames are grouped into fixed-length sequences (10 frames per sequence).

3.3 SPATIO-TEMPORAL AUTOENCODER ARCHITECTURE

The network comprises:

- Encoder:
  - Layer 1: 3×3×3 \times convolution, 64 filters, stride 1,2,2
  - Layer 2: 3×3×3 \times convolution, 128 filters, stride 1,2,2
- Decoder:
  - Layer 1: 3×3×3 \times transposed convolution, 128 filters, stride 1,2,2.
  - Layer 2: 3×3×3 \times transposed convolution, 64 filters, stride 1,2,2.

Output: 3×3×3 \times transposed convolution, 3 filters, stride 1,1,1.

3.4 TRAINING

- Loss Function: Mean Squared Error (MSE).
- Epochs: 2 (adjusted for computational resources).
- Validation Split: 20% of the training data is used for validation.

3.5 ANOMALY DETECTION

- Compute reconstruction errors for test sequences.
- Set an anomaly threshold

$$\text{Threshold} = \mu + 2\sigma$$

where μ is the mean and σ is the standard deviation of reconstruction errors on normal data.

- Detect anomalies as sequences with errors exceeding the threshold.

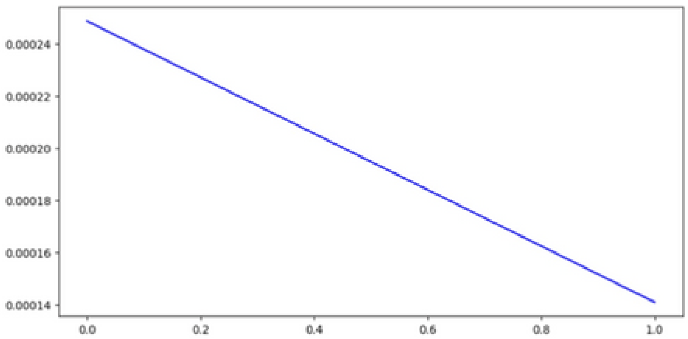
4. RESULTS AND VISUALIZATIONS

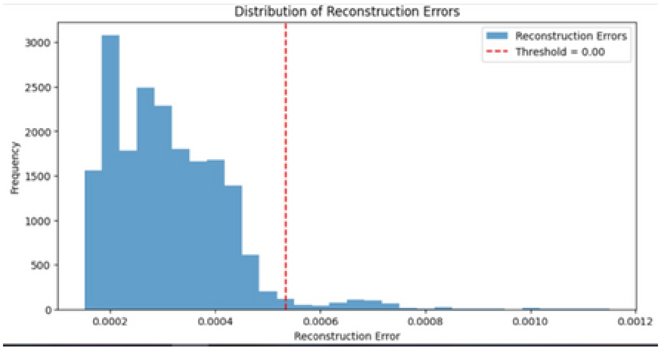
4.1 PERFORMANCE METRICS

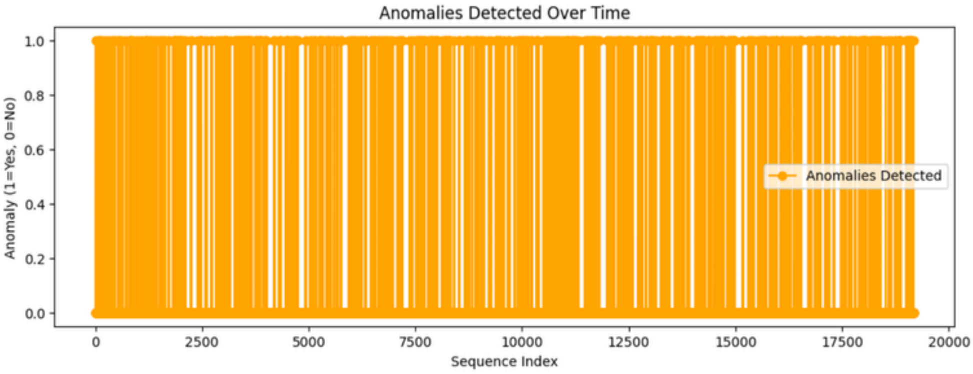
- Reconstruction Errors: Normal sequences exhibit low errors, while anomalies have significantly higher errors.
- Anomalies Detected: The number of anomalies identified is consistent with expectations from the dataset.

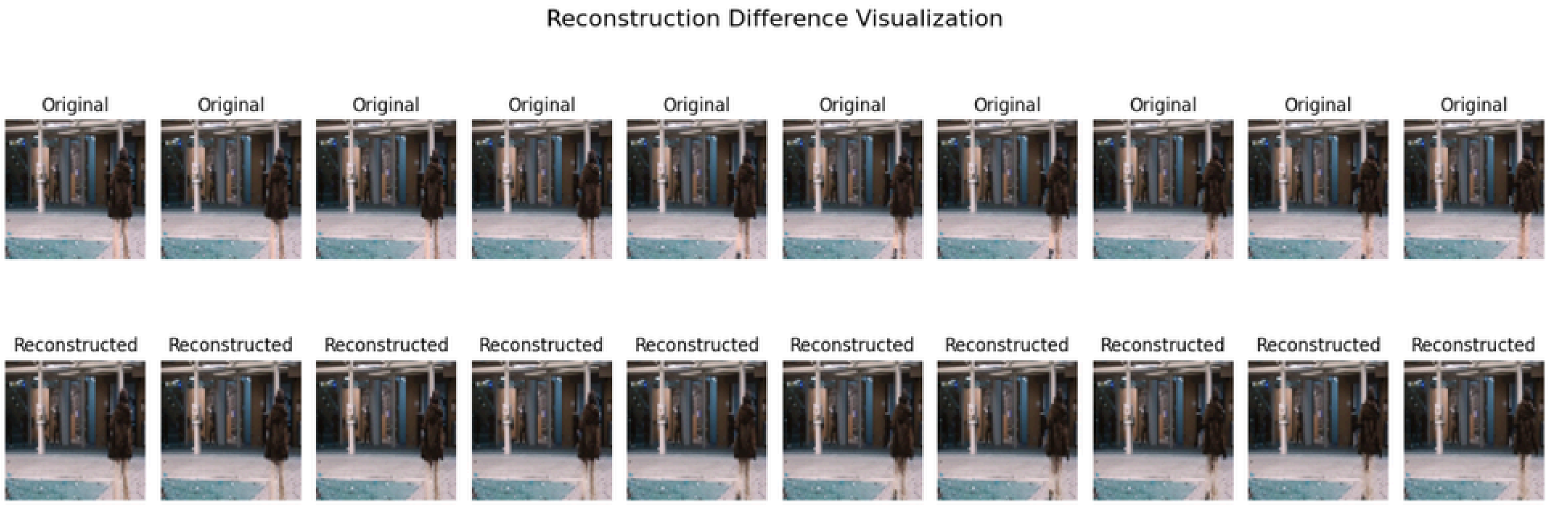
4.2 VISUALIZATIONS

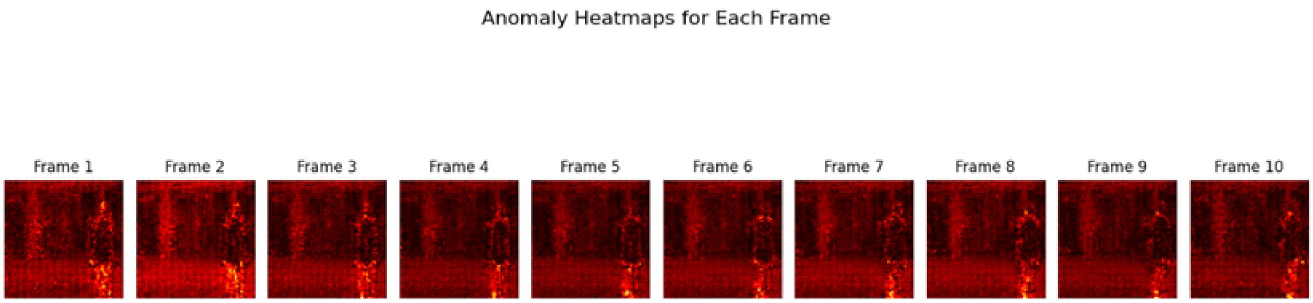
- Reconstruction Difference: Highlights the model's struggles in reconstructing anomalous patterns.
- Heatmaps: Show spatial localization of anomalies within frames.
- Side-by-Side Videos: Allow qualitative analysis of model performance.











Anomaly Heatmaps for Each Frame

Frame 1 Frame 2 Frame 3 Frame 4 Frame 5 Frame 6 Frame 7 Frame 8 Frame 9 Frame 10

5. CHALLENGES AND LIMITATIONS

- Memory Constraints: Large datasets required batch-based processing to avoid memory overflow.
- Threshold Tuning: The anomaly threshold is sensitive to dataset characteristics.

6. CONCLUSION

This project successfully demonstrates the application of spatio-temporal autoencoders for video anomaly detection. The architecture efficiently learns normal patterns in video sequences, and reconstruction errors provide a robust mechanism for detecting anomalies. Future improvements could include advanced architectures like Variational Autoencoders or incorporating attention mechanisms to enhance performance.