Design Document: Video Anomaly Detection and Summarization Pipeline

Table of Contents

1. Overview

This pipeline is designed to perform anomaly detection on video files and generate a summarization of video content. The core of this pipeline combines a **Diffusion Model** to extract reconstruction errors as features and a **CNN+LSTM** model to classify videos based on these features.

Objective

- **Anomaly Detection**: Capture reconstruction errors to detect anomalies in the video frames.
- **Video Classification**: Use a CNN+LSTM to analyze frame sequences and classify videos as "anomalous" or "normal."
- **Video Summarization**: Extract descriptions of key frames and create a summarized narrative of the video content.

2. Architecture and Component Details

System Architecture

1. **Video Preprocessing**: Converts videos into frames and resizes each frame to match model requirements.
2. **Diffusion Model for Feature Extraction**: Uses a Vision Transformer (ViT) to capture reconstruction errors for each frame, acting as features for classification.
3. **CNN+LSTM Model for Classification**: Processes the extracted features and classifies videos as normal or anomalous.
4. **Evaluation Metrics**: Evaluates the model's performance.
5. **Content Description and Summarization**: Uses a captioning model to describe frames and a summarization model to create an overview of the video.

Component Details

1. Diffusion Model for Feature Extraction

- **Purpose**: Identify reconstruction errors in frames to highlight anomalies.
- **Architecture**: Vision Transformer (ViT) based image classification model.
- **Process**:
    1. Original frames are passed through the ViT model to obtain baseline features.
    2. Slightly perturbed versions of frames are processed to obtain reconstructed features.
    3. The absolute difference between baseline and reconstructed features forms the reconstruction error, which serves as an anomaly indicator.

2. CNN+LSTM Model for Classification

- **Purpose**: Classify video sequences based on the reconstruction error features.
- **Architecture**:
    1. **CNN Layers**: Extract spatial features from each frame.
    2. **LSTM Layer**: Captures temporal dependencies across frames.
    3. **Classifier**: Binary classification to indicate normal or anomalous video.
- **Process**:
    1. Each frame's reconstruction error is fed into CNN layers to extract feature maps.
    2. These features are processed by an LSTM layer to learn temporal patterns.
    3. The final LSTM output is passed to a fully connected layer for classification.

3. Evaluation Metrics

The pipeline uses metrics for binary classification:

- **Accuracy**
- **Precision**
- **Recall**
- **F1 Score**

These metrics are essential to evaluate model performance in anomaly detection.

4. Video Content Description and Summarization

- **Image Captioning**: Describes frames using a pretrained captioning model (BLIP).
- **Text Summarization**: Combines individual captions using a summarization model (BART).
- **Process**:
    1. Key frames are sampled and described.
    2. All descriptions are combined and summarized to provide a narrative for the video.

---

3. Pipeline Workflow

1. **Video Upload**: The user uploads a video.
2. **Preprocessing**: The video is split into frames, resized, and normalized.
3. **Feature Extraction and Reconstruction Error Calculation**:
    - Original and perturbed frames are passed through the diffusion model.
    - Reconstruction errors are calculated.
4. **Classification with CNN+LSTM**: The CNN+LSTM model classifies the video as "anomalous" or "normal."
5. **Evaluation Metrics**: Model performance metrics are calculated.
6. **Content Summarization**: Key frames are described, and their descriptions are summarized to provide a narrative.
7. **Results Display**: The classification, metrics, and summary are displayed.

---

4. Justification of Model Choices

1. **Diffusion Model for Feature Extraction**:
    - Diffusion models and their variants (like Vision Transformers) are well-suited for identifying fine-grained reconstruction errors.
    - Reconstruction error detection is particularly useful in anomaly detection, where discrepancies in visual features can indicate anomalies.
2. **CNN+LSTM for Temporal Classification**:
    - CNN layers are effective for spatial feature extraction.
    - LSTM layers capture temporal dependencies, allowing the model to analyze frame sequences effectively.
    - This combination is ideal for video classification tasks with a time series nature.
3. **Image Captioning and Summarization**:

    - The BLIP model for captioning and BART for summarization provides interpretable content descriptions, making the output more understandable.

---

5. Configurations and Assumptions

Configurations

- **Frame Size**: 224x224 (suitable for most image classification models).
- **Number of Frames**: 32 frames per video to ensure temporal consistency.
- **CNN Feature Dimensionality**: 512.
- **LSTM Hidden Size**: 256.
- **Classification**: Binary output (1 for anomaly, 0 for normal).

Assumptions

- Video input is in `.mp4` format.
- The pipeline processes a single video file at a time.
- The CNN+LSTM structure is appropriate for the temporal classification task.

---

6. Training and Inference Process

Training Process

The pipeline does not include a training process, as per project constraints. However, if training were required:

1. **Data Preparation**: Collect a labeled dataset of normal and anomalous videos.
2. **Feature Extraction**: Generate reconstruction error features for each frame.
3. **Training the CNN+LSTM**: Use extracted features to train the CNN+LSTM model with a binary cross-entropy loss.

Inference Process (Current Pipeline)

1. **Upload Video**.
2. **Preprocessing**: Video is split into frames and normalized.
3. **Feature Extraction**: Compute reconstruction errors.
4. **Classification**: CNN+LSTM model processes features and outputs a binary classification.
5. **Metrics Calculation**: Accuracy, precision, recall, and F1 score are computed.

6. **Summarization**: Key frame descriptions are generated and summarized.

---

7. Evaluation Metrics

Metrics Used

- **Accuracy**: Measures overall correctness.
- **Precision**: Measures the model's ability to identify true positives among predicted positives.
- **Recall**: Measures the model's ability to identify true positives among actual positives.
- **F1 Score**: Harmonic mean of precision and recall, offering a balanced performance metric.

Evaluation Strategy

1. **Single Video Inference**: Given the single-video constraint, each video is evaluated independently.
2. **Interpretation of Metrics**:
    - High accuracy with low precision/recall would indicate an imbalance in classification.
    - Balanced high values across metrics indicate good model performance.

---

8. References and Research Benchmarks

1. **Diffusion Models for Anomaly Detection**:
    - Research papers on Vision Transformers (ViT) and reconstruction error detection serve as the foundation for the diffusion model.
2. **CNN+LSTM for Video Classification**:
    - CNN+LSTM architectures have shown effectiveness in video anomaly detection tasks due to their spatiotemporal modeling capabilities.
3. **Image Captioning and Summarization**:
    - BLIP and BART are well-researched models for image captioning and summarization, respectively, and are benchmarks in their fields.