Assignment 2: Evaluation Metrics

By: Pradyumna Seethamraju

Evaluation Metrics for Architecture in the Pipeline

The model architecture in the pipeline uses several key evaluation metrics to assess classification and reconstruction accuracy, which help determine both model quality and anomaly detection efficacy. This includes standard metrics like **Accuracy**, **F1 Score**, and **AUC (Area Under the Curve)**, alongside **Reconstruction Error**, which is crucial for monitoring anomaly detection performance in the **Diffusion Model**.

1. Evaluation Metrics

1.1. Accuracy

- **Definition**: Measures the proportion of correct predictions (i.e., frames correctly classified as either normal or anomalous).
- **Use Case**: Given that the pipeline is designed to classify sequences as anomalous or normal, accuracy provides a basic measure of how often the CNN+LSTM model is correct.
- **Calculation**: Accuracy=True Positives+True Negatives/Total Number of Predictions
- **Interpretation**: Higher accuracy indicates that the model is performing well in distinguishing between classes. However, in cases with class imbalance, this metric alone may be insufficient, necessitating additional metrics like F1 and AUC.

1.2. F1 Score

- **Definition**: The harmonic mean of Precision and Recall, particularly useful for binary classification where class distributions may be imbalanced.
- **Use Case**: Anomaly detection often deals with imbalanced data (e.g., fewer anomalies than normal cases). F1 Score is well-suited for this, as it considers both **False Positives** and **False Negatives**.
- **Calculation**: F1 Score=2×Precision×RecallPrecision+Recall
    - **Precision** = True Positives/(True Positives+False Positives)
    - **Recall** = True Positives/(True Positives+False Negatives)
- **Interpretation**: A high F1 score suggests a balanced performance in detecting anomalies (high precision) and minimizing missed anomalies (high recall).

1.3. AUC (Area Under the Curve)

- **Definition**: Represents the area under the Receiver Operating Characteristic (ROC) curve, which plots **True Positive Rate** against **False Positive Rate** at different threshold levels.
- **Use Case**: AUC is ideal for evaluating model performance independent of a specific threshold, especially valuable in evaluating the model's capability to rank anomalous frames higher than normal ones.
- **Calculation**:
    - The ROC curve is generated by plotting the **True Positive Rate** (TPR) against the **False Positive Rate** (FPR) as the classification threshold varies.
- **Interpretation**: AUC ranges from 0 to 1, with 1 representing perfect distinction between classes. In anomaly detection, a high AUC indicates strong separation between normal and anomalous sequences.

1.4. Reconstruction Error (Diffusion Model)

- **Definition**: Measures the difference between the input frame and the reconstructed frame output by the **Diffusion Model**. This is particularly useful in anomaly detection, as anomalies should have a higher reconstruction error.
- **Use Case**: Used as a threshold-based detection mechanism within the Diffusion Model to identify frames that significantly deviate from the learned data distribution.
- **Calculation**: Reconstruction Error= Mean(∣Input Frame−Reconstructed Frame∣)
- **Interpretation**: Higher reconstruction error values may indicate an anomaly, making this metric a primary indicator for the quality of feature extraction. A well-tuned threshold can help classify high-error frames as anomalies.

2. How the Metrics Are Applied

2.1. CNN+LSTM Model Evaluation

- **Accuracy, F1 Score, and AUC** are calculated on a **test dataset** to evaluate the model's binary classification performance.
- These metrics are assessed after each epoch, and the **best model checkpoint** is saved based on the highest F1 score, prioritizing balanced precision and recall.

2.2. Diffusion Model Evaluation

- **Reconstruction Error** is used to evaluate how well the Diffusion Model reconstructs each frame. Frames with a high reconstruction error (above a certain threshold) are flagged as potentially anomalous, allowing this metric to serve as a secondary layer of anomaly detection before classification by the CNN+LSTM.

3. Insights from Each Metric

- **Accuracy** provides an overall view of performance but may mask issues if classes are imbalanced.

- **F1 Score** is critical for balancing precision and recall, which is particularly important for imbalanced datasets common in anomaly detection.
- **AUC** provides a threshold-independent view of the model's ranking ability, crucial when adjusting classification thresholds.
- **Reconstruction Error** specifically assesses the Diffusion Model's reconstruction capabilities, enabling it to act as an initial filter for anomalies based on how well frames fit the learned data distribution.

4. Combining the Metrics for Comprehensive Evaluation

By using these metrics together, the pipeline can be evaluated on both general classification accuracy (Accuracy) and balanced anomaly detection (F1 and AUC). Reconstruction Error serves as a unique measure within the pipeline, providing additional validation of anomalous frames before reaching the CNN+LSTM model. This multi-metric approach provides robust insights into model performance across all stages of anomaly detection and classification in video sequences.