Anilata AB Assignment

Question 1: Action Classification System

Introduction

In this document, I will outline several approaches to creating an action classification system. Action classification involves identifying specific human actions from video data, which is a challenging task due to the complexity and variability in human movements. Below, I will describe the different approaches I considered, along with their strengths and limitations.

1. **Approach 1: CNN with Image Frames**

   Concept

   The first approach involved using a Convolutional Neural Network (CNN) architecture to classify actions based on individual image frames extracted from the video. I utilized the VGG-16 architecture, which is pre-trained on large image datasets and fine-tuned it using the 'Human Action Recognition' dataset from Kaggle.

   Methodology

   1. **Frame Extraction:** The video is decomposed into individual frames.
   2. **Preprocessing:** Each frame undergoes normalization and resizing to fit the VGG-16 input dimensions.
   3. **CNN Classification:** The pre-processed frames are passed through the VGG-16 network to extract features, followed by classification using a softmax layer.

   Results and Limitations

   - **Training Time:** This approach required significant training time due to the large volume of image data.
   - **Accuracy:** The model showed limited accuracy, especially in scenarios where temporal dependencies were crucial for recognizing actions.
   - **Temporal Data:** This method was unable to capture the temporal dynamics between frames, leading to suboptimal performance.

2. **Approach 2: OpenPose and MediaPipe for Keypoint Extraction**

   Concept

   The second approach involved using open-source libraries like OpenPose and MediaPipe to extract spatial keypoints from the videos. These keypoints represent the positions of human body parts and can be used to recognize actions. Dataset used is UCF50, which is a consolidated dataset of small videos from youtube.

Methodology

1. **Keypoint Extraction:** OpenPose or MediaPipe are used to extract spatial keypoints from the video frames.
2. **Preprocessing:** The keypoints are normalized and fed into a CNN classifier.
3. **Classification:** The CNN classifier processes the keypoints to classify the action.

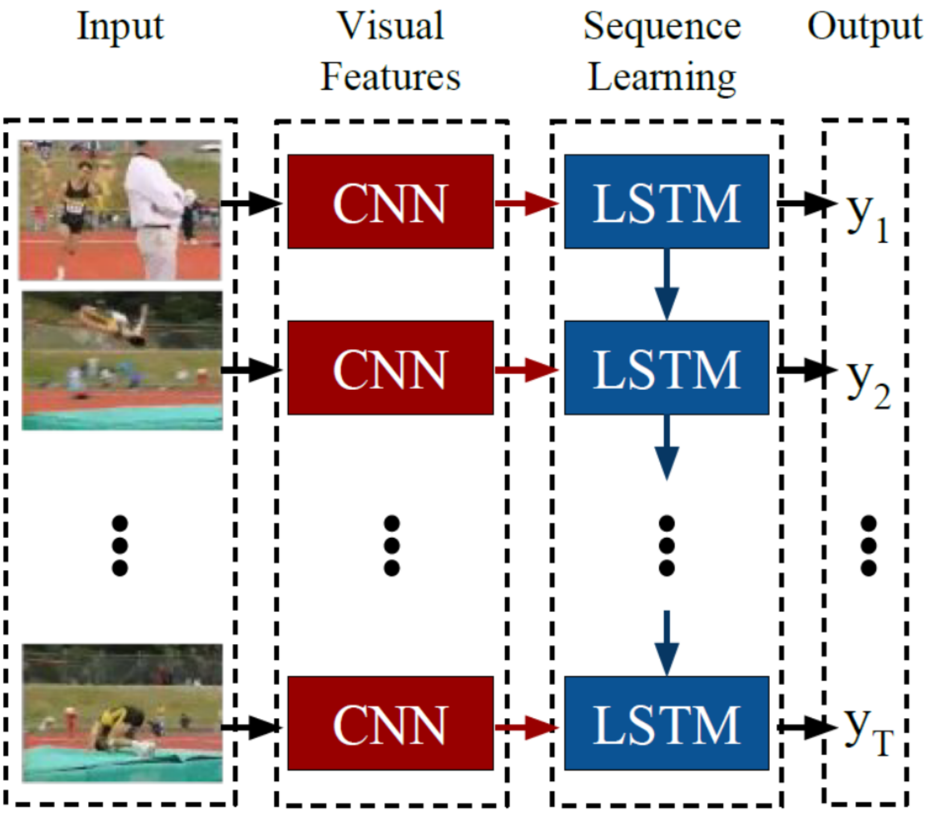
Results and Limitations

- **Improvement over Approach 1:** This method showed better performance in recognizing actions since it focused on key movement points.
- **Temporal Data:** Similar to Approach 1, it lacked the ability to capture temporal dependencies, limiting its effectiveness for complex actions.

3. Approach 3: LRCNN (LSTM-CNN) Architecture

Concept

The third approach combined the strengths of CNNs for spatial feature extraction and Long Short-Term Memory (LSTM) networks for capturing temporal dependencies. This hybrid model is often referred to as an LRCNN architecture.



Methodology

1. **Spatial Feature Extraction:** A CNN, such as ResNet or another robust model, extracts spatial features from the video frames.
2. **Temporal Feature Extraction:** The spatial features are passed to an LSTM network, which processes the sequence of features over time.
3. **Classification:** The LSTM outputs are used to classify the actions.

Results and Benefits

- **Training Efficiency:** The model was faster to train compared to the previous approaches.
- **Performance:** The inclusion of temporal data led to significant improvements in action recognition accuracy.
- **Complexity:** This approach provides a balanced trade-off between model complexity and performance.

4. Approach 4: Object Detection with YOLOv8 and Temporal Models

Concept

The fourth approach involved using advanced object detection techniques like YOLOv8 to classify actions in real-time. YOLOv8 was combined with time-dependent models to capture both spatial and temporal aspects of the actions.

Methodology

1. **Object Detection:** YOLOv8 is used to detect objects and their movements in the video.
2. **Temporal Analysis:** The detected objects are then passed through time-dependent models (such as LSTM or GRU) to analyze the sequence of movements.
3. **Classification:** The combined spatial and temporal data is used to classify the actions.

Results and Benefits

- **Real-Time Application:** This approach is well-suited for real-time action classification in live video feeds.
- **Accuracy:** It achieves high accuracy by effectively combining spatial and temporal information.
- **Complexity:** The method is more complex and computationally intensive but provides superior results.