NATIONAL INSTITUTE OF TECHNOLOGY,
TIRUCHIRAPPALLI

SUMMER INTERNSHIP REPORT

MAY 30 2023 – JULY 31 2023

HUMAN ACTIVITY RECOGNITION SYSTEM

Submitted by: *Institution:*

RAJAAMANI.P *NIT Trichy*

112120048

Metallurgical and Materials Engineering

DECLARATION

I, **Rajaamani.P**, Roll Number **112120048** , hereby declare that this internship entitled “**Human Activity Recognition System**” represents my original work carried out as a bachelors student of NIT Tiruchirappalli at National Institute of Technology ,Tiruchirappalli under Prof. Bibin Francis Professor, Department of Electronics and Communication Engineering.

Any contribution made to this research by others, with whom I have worked at NIT Trichy or elsewhere, is explicitly acknowledged in the dissertation. Works of other authors cited in this thesis have been duly acknowledged under the section "References".

ACKNOWLEDGEMENT

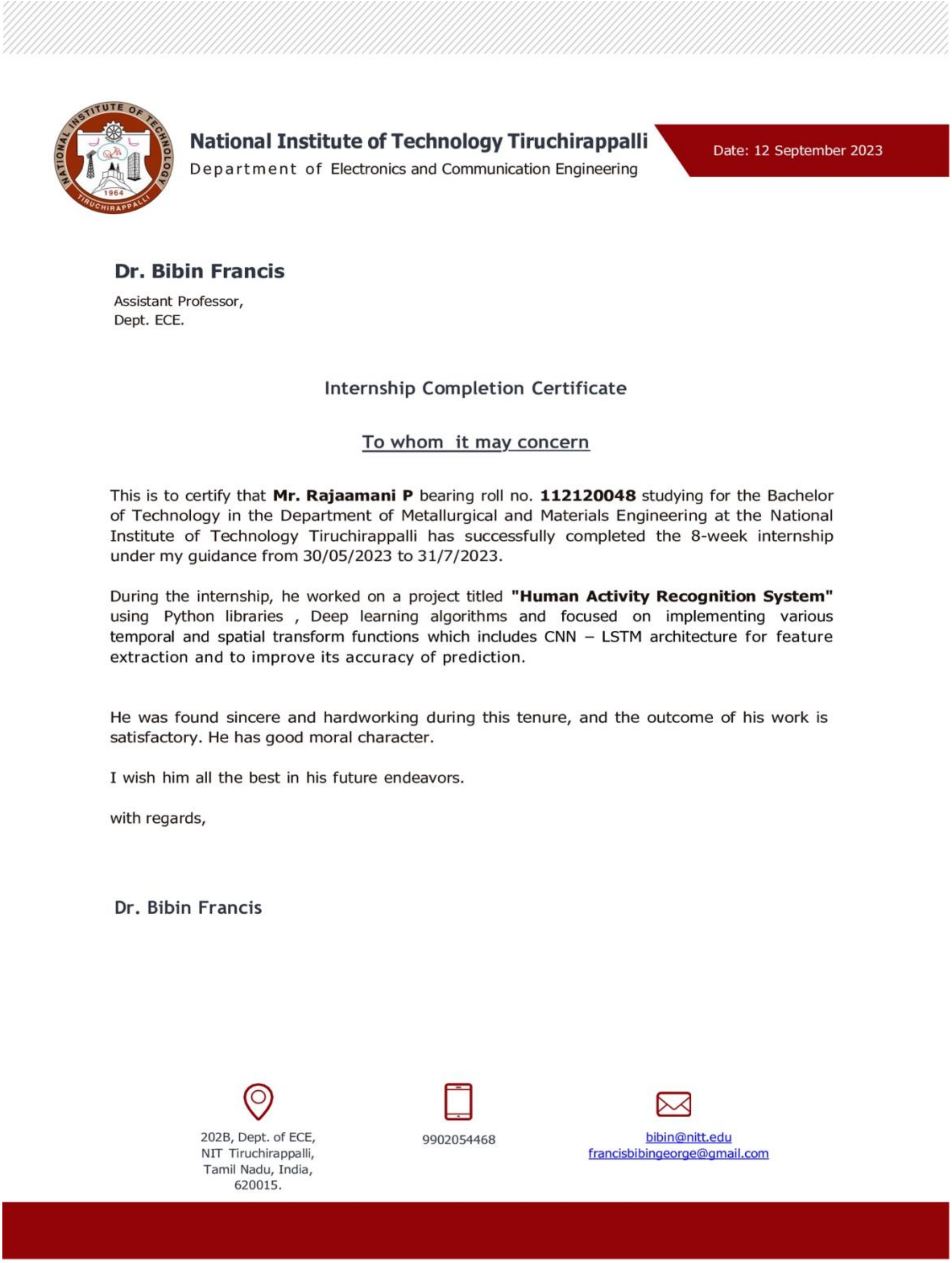
I take this opportunity to express my profound gratitude and deep regards to everyone who guided me in the right path and helped in the successful completion of this course.

This acknowledgement transcends the reality of formality when we would like to express deep gratitude and respect to all those people behind the screen who guided, inspired and helped me for the completion of our project work. This project would add as an asset to my academic profile. I would like to express my thankfulness to Dr. Bibin Francis for their constant motivation and valuable help throughout the development of this project by providing us with required information without whose guidance, cooperation and encouragement, this project couldn’t have been materialized.

And lastly, I express my deepest gratitude to everyone who helped me both directly and indirectly during the entire duration of my internship.

RAJAAMANI.P

112120048



National Institute of Technology Tiruchirappalli

Department of Electronics and Communication Engineering

Date: 12 September 2023

Dr. Bibin Francis

Assistant Professor,
Dept. ECE.

Internship Completion Certificate

To whom it may concern

This is to certify that **Mr. Rajaamani P** bearing roll no. **112120048** studying for the Bachelor of Technology in the Department of Metallurgical and Materials Engineering at the National Institute of Technology Tiruchirappalli has successfully completed the 8-week internship under my guidance from 30/05/2023 to 31/7/2023.

During the internship, he worked on a project titled **"Human Activity Recognition System"** using Python libraries , Deep learning algorithms and focused on implementing various temporal and spatial transform functions which includes CNN – LSTM architecture for feature extraction and to improve its accuracy of prediction.

He was found sincere and hardworking during this tenure, and the outcome of his work is satisfactory. He has good moral character.

I wish him all the best in his future endeavors.

with regards,

Dr. Bibin Francis

202B, Dept. of ECE,
NIT Tiruchirappalli,
Tamil Nadu, India,
620015.

9902054468

bibin@nitt.edu
francisbibingeorge@gmail.com

ABSTRACT

Human Activity Recognition (HAR) represents a captivating frontier in computer science, merging computer vision, machine learning, and signal processing. HAR involves classifying human actions using sensor data, traditionally requiring meticulous feature engineering. However, modern deep learning techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have revolutionized HAR by automatically extracting meaningful features from raw sensor data. This project delves into HAR, leveraging deep learning to create an efficient system. It encompasses sensor data acquisition, pre-processing, feature extraction, machine learning, deep learning, activity classification, real-time recognition, and explores challenges and applications. By seamlessly integrating CNNs, LSTMs, and ResNet-101 architecture, it aims to enhance accuracy and unlock innovative real-world possibilities across various industries.

The journey begins with sensor data acquisition, the lifeblood of HAR. Subsequently, meticulous data pre-processing ensues, ensuring data quality through cleaning, missing value handling, noise reduction, and normalization. Deep learning models, notably CNNs for spatial feature extraction and LSTMs for temporal dependencies, are at the forefront of the project's methodology. The core innovation lies in the integration of these models within the robust ResNet-101 architecture, facilitating hierarchical spatial feature extraction and precise temporal modelling. Activity classification is the project's heartbeat, with a focus on optimizing metrics like precision, recall, and F1 score.

Moreover, the developed HAR system isn't confined to offline analysis; it excels in real-time recognition. Beyond the technical aspects, the project dives into challenges like computational complexity and overfitting while illuminating the diverse real-world applications of HAR. These applications span healthcare monitoring, sports performance analysis, security surveillance, and smart environments, underscoring the far-reaching impact of this research in enhancing human-technology interactions. In conclusion, this project epitomizes the transformative potential of deep learning in HAR. By seamlessly integrating CNNs, LSTMs, and ResNet-101 architecture, it endeavours to elevate accuracy, resilience, and real-time capabilities, ushering in an era of innovation and utility across diverse industries.

TABLE OF CONTENTS

INTRODUCTION

Human Activity Recognition (HAR) is a field in computer science and AI that focuses on creating systems and algorithms to automatically identify and understand human actions from data, like sensors, cameras, or wearables. The main aim is to enable machines to comprehend and respond to human behaviours, with applications in various domains. In HAR using ResNet-101, Convolutional Neural Networks (CNNs), and Long Short-Term Memory networks (LSTMs), the goal is to use deep learning to accurately recognize and categorize human activities from a dataset. ResNet-101 is a powerful neural network architecture, particularly for image classification tasks, thanks to its deep layers and residual learning.

The dataset used likely contains sequences of sensor data or image frames capturing different human activities, each labelled with an activity like walking or running. This combination of ResNet-101, CNNs, and LSTMs aims to capture both spatial and temporal features in the data, improving the model's ability to grasp complex activity patterns. CNNs are responsible for extracting spatial features from the input data, such as identifying key image features relevant to different activities. ResNet-101 excels at capturing hierarchical features in images due to its depth and residual connections. LSTMs come into play to model temporal dependencies in the data, crucial for recognizing activities involving sequences of motions. LSTMs are adept at remembering patterns over time, making them suitable for such tasks.

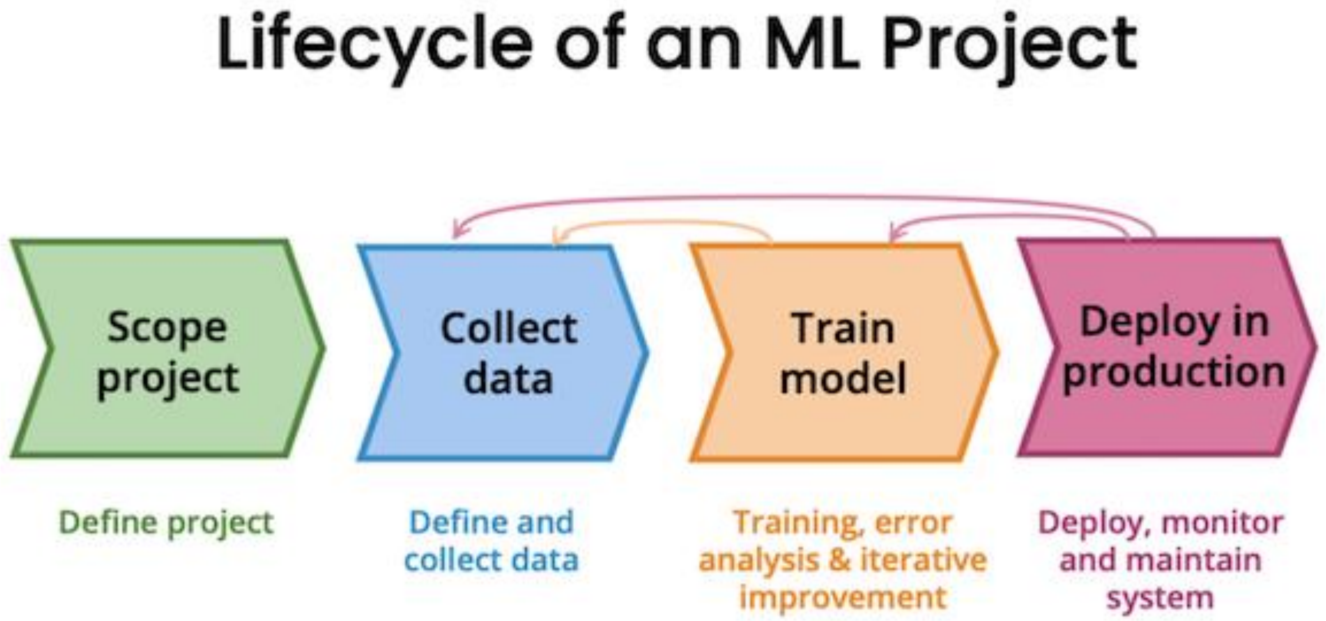
The overall architecture likely involves passing sequences of data (e.g., video frames or sensor readings) through the CNN to extract spatial features, followed by feeding these features into the LSTM to capture temporal dependencies. The final output is the classification of the human activity associated with the input sequence. This approach combines the strengths of CNNs and LSTMs, integrating spatial and temporal information for more accurate HAR. The inclusion of ResNet-101 enhances the model's capacity to learn complex hierarchical features. Success depends on dataset quality, diversity, and effective training strategies for optimal predictions.

OBJECTIVES:

The primary objective of this project is to develop a sophisticated and efficient system for recognizing human activities using advanced deep learning techniques. By leveraging the power of CNNs and LSTMs within the ResNet-101 framework, the goal is to achieve high accuracy in identifying a diverse range of human activities from input sensor data. The system aims to excel in real-world scenarios, providing reliable recognition across various environments, users, and activity types.

PROJECT METHODOLOGY :

The methodology for a Human Activity Recognition (HAR) project involves several key steps, from data collection and pre-processing to model training, testing and evaluation. A general outline of the project methodology:


Lifecycle of an ML Project
Scope project
Collect data
Train model
Deploy in production
Define project
Define and collect data
Training, error analysis & iterative improvement
Deploy, monitor and maintain system

SCOPE OF THE PROJECT :

The scope of this project is to enhance the accuracy and effectiveness of a Human Activity Recognition (HAR) model through a comprehensive approach that spans multiple stages of development. This includes data collection and preprocessing, the design and implementation of a specialized model based on the ResNet-101 architecture, rigorous training, thorough evaluation, continuous optimization, considerations for real-time recognition , validation through user studies or domain expert feedback, and meticulous documentation with a strong emphasis on ethical compliance.

The project's primary objective is to improve the model's ability to accurately identify and classify a diverse range of human activities based on sensor data. Data collection involves obtaining the UCF101 dataset and ensuring its labels are accurate. Subsequently, data preprocessing steps are carried out to clean the data, handle missing values, remove noise, and normalize it to ensure high-quality input for the model.The core of the project centers on the development of a ResNet-101-based model, tailored to excel in HAR. The model's architecture, including layers, connections, and parameters, is carefully configured to optimize performance. Training and evaluation are crucial phases, with metrics such as accuracy, precision, recall, and F1 score used to assess and refine the model's capabilities.

DATA COLLECTION:

UCF 101 DATASET

The UCF101 dataset is a comprehensive and widely recognized resource for human activity recognition. It consists of a total of 13,320 video clips, encompassing 101 distinct human actions or activities. These activities cover a broad spectrum of daily life and sports-related motions, including but not limited to activities such as jogging, eating, basketball, yoga, and more. Each video clip is carefully labeled, providing precise annotations for the corresponding activity, which is invaluable for supervised machine learning tasks like Human Activity Recognition (HAR).

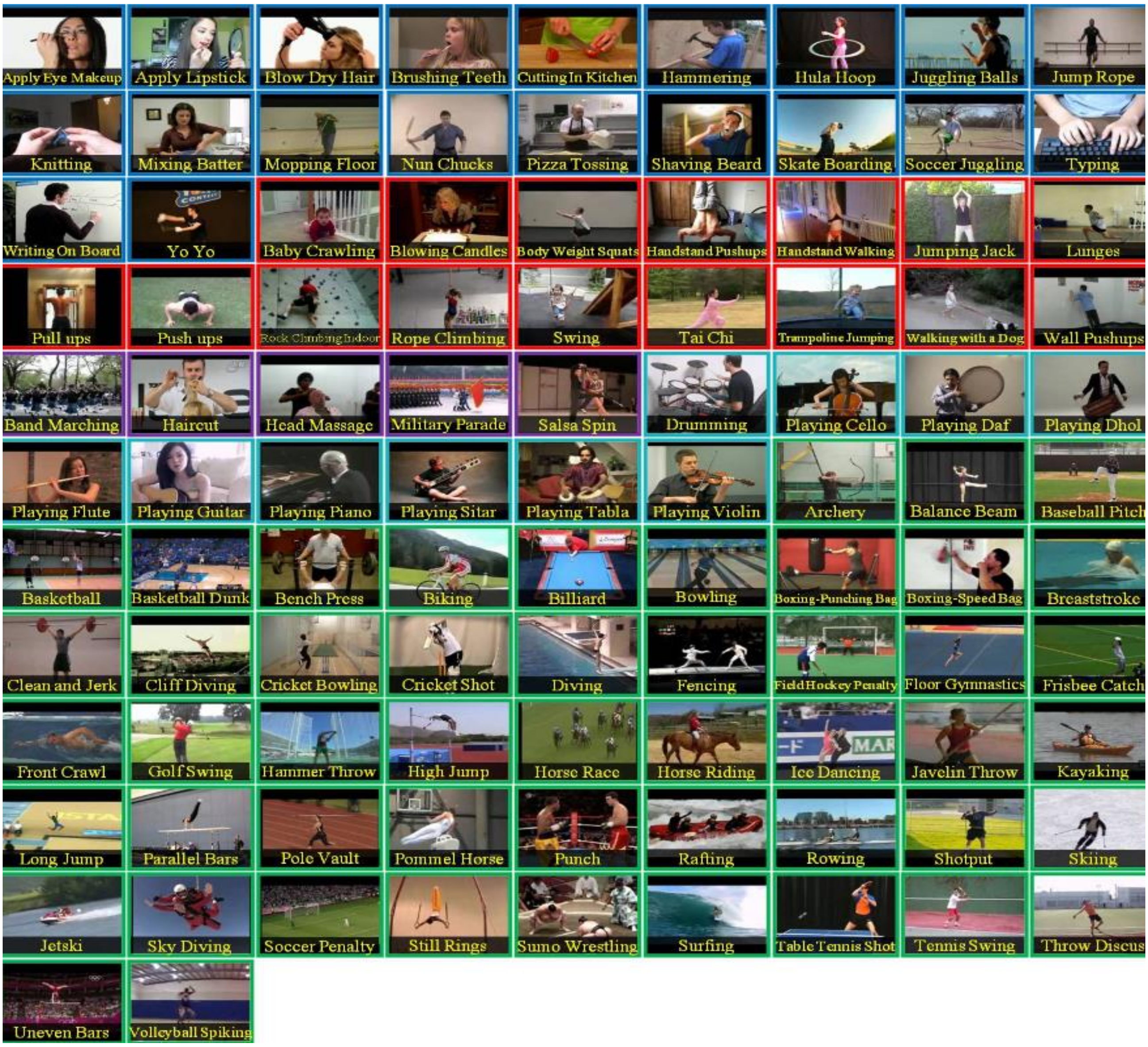
The dataset is divided into three main splits: training, validation, and testing sets. The training set typically comprises approximately 9,500 video clips, while the validation set contains around 1,000 clips, and the testing set includes roughly 3,800 clips. This division allows for proper model training, validation of model performance during development, and robust testing to assess the model's ability to generalize to unseen data.

The UCF101 dataset's richness in activities, large number of video clips, and well-defined labeling make it an ideal choice for HAR model development. It provides a realistic representation of human activities, enabling the model to learn and recognize a wide range of actions, making it suitable for various real-world applications where accurate activity recognition is crucial.

The UCF101 dataset was meticulously created through a multi-step process involving the collection of video clips from diverse sources, precise annotation of human activities, data cleaning to remove extraneous content, organization into distinct activity classes, and division into training, validation, and testing subsets. This dataset, known for its quality and diversity, serves as a fundamental resource for the development and evaluation of Human Activity Recognition (HAR) models, enabling researchers and practitioners to advance the field by training models to recognize a wide range of real-world human activities accurately.

Certainly, here are some of the actions included in the UCF101 dataset:

- Walking
- Sitting
- Standing
- High Jump
- Jumping
- Swimming
- Boxing
- Weightlifting
- Diving



UCF 101 DATASET - 13,320 VIDEOS - 101 ACTION CATEGORIES

DATA PREPROCESSING:

Data pre-processing refers to the set of techniques and procedures applied to raw data to transform it into a format that is suitable for analysis, modelling, and machine learning. It is a crucial step in the data pipeline because the quality and suitability of the data directly impact the results and performance of any data analysis or machine learning task. It involves various steps:

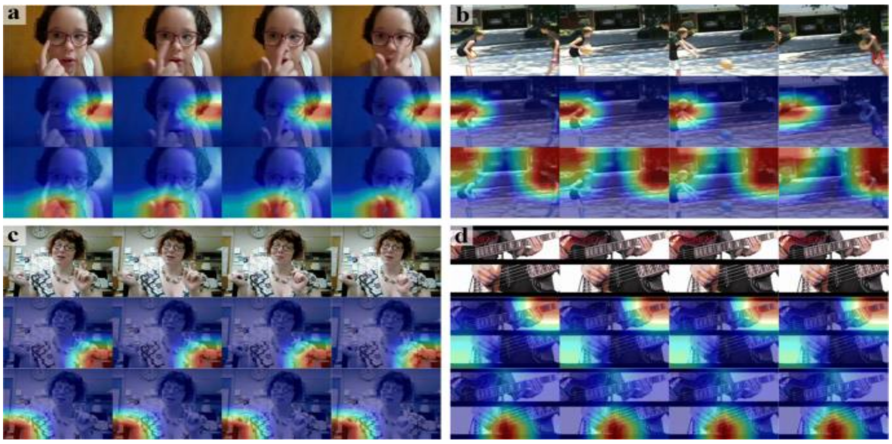
- **Temporal Transformations**
- **Spatial Transformations**
- **Target Transformations**

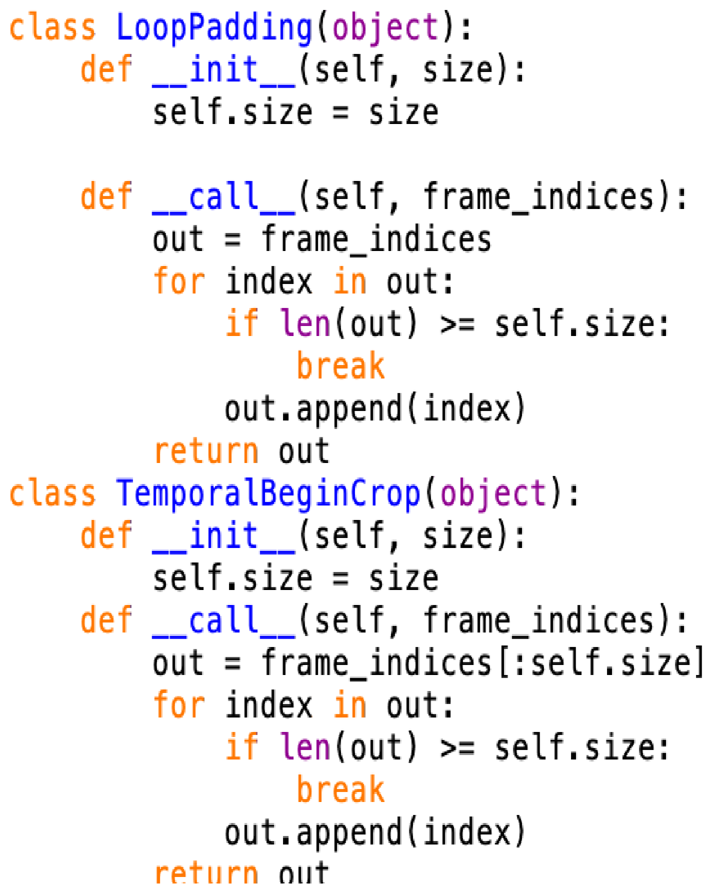
Temporal Transformations:

Temporal transformations are essential pre-processing operations for video data used in machine learning models, particularly for tasks like action recognition. Here's a concise overview of key temporal transformations:

1. **Loop Padding:**
   - **Purpose:** Ensures all video clips have the same number of frames for model input.
   - **How it works:** Adds extra frames to shorter videos from the beginning.
   - **Importance:** Ensures consistent frame counts for efficient model training.
2. **Temporal Begin Crop:**
   - **Purpose:** Focuses on the beginning part of each video sequence for analysis.
   - **How it works:** Selects the first 'n' frames, where 'n' is the desired sequence length.
   - **Usefulness:** Valuable when essential information is concentrated at the start.
3. **Temporal Center Crop:**
   - **Purpose:** Selects the central and most informative segment of videos.
   - **How it works:** Extracts a segment from the video's center, preserving key frames.
   - **Benefit:** Useful when the primary action or information is in the middle.
4. **Temporal Random Crop:**
   - **Purpose:** Introduces data augmentation for model robustness.
   - **How it works:** Randomly selects a segment from each video, exposing the model to different action parts.
   - **Value:** Enhances adaptability to various scenarios and improves generalization.

These temporal transformations are akin to preparing ingredients before cooking, where each operation has a specific role. Loop Padding ensures uniformity, TemporalBeginCrop focuses on the start, Temporal Center Crop selects the central segment, and Temporal Random Crop adds variability to boost the model's adaptability.





```
class LoopPadding(object):
    def __init__(self, size):
        self.size = size

    def __call__(self, frame_indices):
        out = frame_indices
        for index in out:
            if len(out) >= self.size:
                break
            out.append(index)
        return out
class TemporalBeginCrop(object):
    def __init__(self, size):
        self.size = size
    def __call__(self, frame_indices):
        out = frame_indices[:self.size]
        for index in out:
            if len(out) >= self.size:
                break
            out.append(index)
        return out
```

Spatial Transformations:

The Spatial Transform Functions in spatial_transform.py are a crucial component of video data pre-processing for deep learning tasks. These functions prepare individual video frames to be effectively utilized by neural networks, enhancing the model's ability to generalize across different spatial conditions within the frames. Let's delve into the key spatial transformations:

1. **Compose**: Compose is a versatile wrapper that allows you to chain multiple spatial transformations together. It facilitates the creation of a transformation pipeline, making it easy to apply a consistent set of operations to each video frame.

2. **ToTensor**: The ToTensor transformation is responsible for converting various image formats, such as NumPy arrays or PIL images, into PyTorch tensors. It also performs pixel value normalization, ensuring that the pixel values fall within a suitable range. This is crucial because neural networks typically expect input data in tensor format.

3. **Normalize**: Normalize adjusts pixel values based on mean and standard deviation, a common pre-processing step that centers the data around zero and scales it appropriately. This step aids in faster model convergence during training.

4. **Scale**: The Scale transformation is used for resizing video frames while preserving their aspect ratio. It can either resize frames to a fixed size or scale them proportionally. Maintaining consistent frame dimensions is essential for neural networks.

5. **Center Crop**: Center Crop focuses on cropping the central portion of an image. This is valuable when resizing isn't desirable, ensuring that the core spatial information is retained.

6. **CornerCrop**: CornerCrop offers the flexibility to crop a corner of an image based on the specified position, such as top-left, top-right, bottom-left, or bottom-right. It enables the selection of regions of interest within frames.
7. **RandomHorizontalFlip**: This transformation introduces diversity during training by randomly flipping video frames horizontally (left to right) with a 50% chance. It helps models become robust to horizontal variations.
8. **MultiScaleCornerCrop**: MultiScaleCornerCrop combines random corner selection with scaling to a specified size. This transformation provides both spatial diversity and consistent frame sizes, enhancing the model's ability to capture various spatial conditions.
9. **MultiScaleRandomCrop**: Similar to MultiScaleCornerCrop, MultiScaleRandomCrop selects a random region within an image and scales it to a specified size. This introduces spatial diversity while maintaining consistent frame dimensions.

In summary, these spatial transformations play a critical role in preparing video frames for deep learning models. They ensure that each frame is appropriately processed, allowing models to generalize effectively to different spatial conditions within video frames. This, in turn, contributes to improved performance in tasks like video classification and object recognition, where spatial information is vital.

Target Transformations:

The "Target Transformations" are custom transformations designed to process and extract specific information from the target data associated with each sample in a dataset. Here's a summary of the purpose and usage of each transformation class:

1. **Compose Transformation**:
   - Purpose: Combines multiple target transformations into a sequence, allowing you to apply a series of transformations to the target data.
   - Usage: It takes a list of transformation objects and applies each of them to the target data. The result is a list containing the outcomes of all applied transformations.
2. **Class Label Transformation**:
   - Purpose: Extracts the 'label' field from the target data, which typically represents the class or category associated with a sample.
   - Usage: It retrieves the 'label' value from the input dictionary, making it convenient for class-related tasks.
3. **Video ID Transformation:**
   - Purpose: Extracts the 'video_id' field from the target data, which can be used to uniquely identify a video sample.
   - Usage: It retrieves the 'video_id' value from the input dictionary, enabling you to access video-specific information.

These target transformations are beneficial when dealing with datasets that have complex target data structures. They allow you to extract specific components of the target data, such as class labels and video IDs, for use during the training or evaluation of machine learning models. Typically, these transformations are applied sequentially using the Compose transformation to process and extract the necessary target information effectively.

MODEL ARCHITECTURE:

CONVOLUTIONAL NEURAL NETWORK(CNN):

Convolutional Neural Networks (CNNs) are a class of deep neural networks designed for processing structured grid data, such as images. They are particularly effective in computer vision tasks, such as image classification, object detection, and segmentation. A general overview of the architecture of a typical CNN are :

1.Input Layer: The raw input data, usually an image in the UCF 101 Data Set like Knitting ,Skiing, Surfing, Apply eye makeup.

2.Convolutional Layer: Convolutional layers are the core building blocks of CNNs. They apply convolutional operations to the input data using filters or kernels. These filters slide over the input image, performing element-wise multiplication and aggregation to capture local patterns and features. The result is a set of feature maps that highlight different aspects of the input.

3.Activation Function: After the convolution operation, an activation function (commonly ReLU - Rectified Linear Unit) is applied element-wise to introduce non-linearity into the network. The non-linear activation helps the network learn complex patterns and relationships in the data.

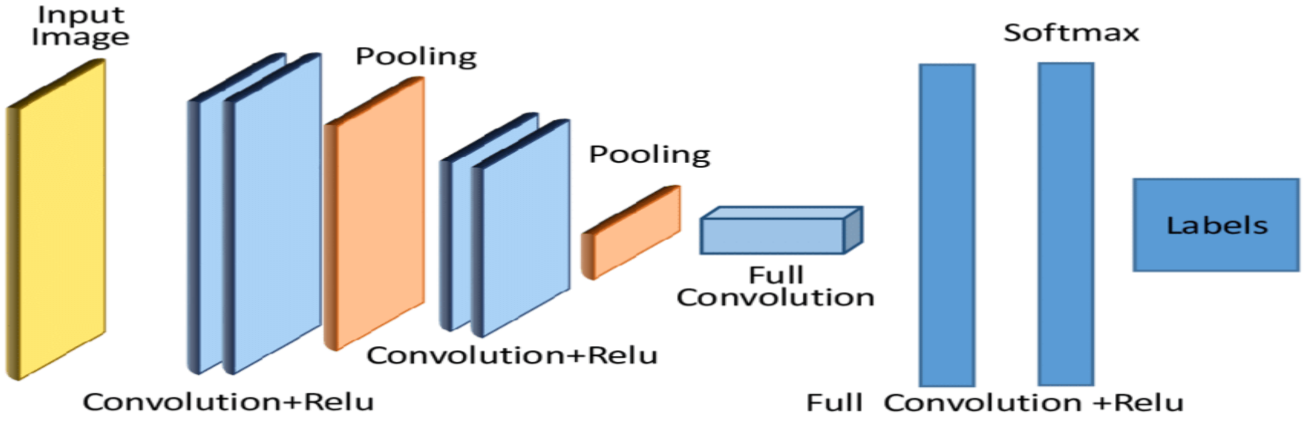
4.Pooling Layer: Pooling layers are used to reduce the spatial dimensions of the feature maps while retaining important information. Common pooling operations include max pooling (retaining the maximum value in a local region) or average pooling (taking the average value).

5.Flattening:The output from the convolutional and pooling layers is flattened into a vector. This step prepares the data for the fully connected layers.

6.Fully Connected (Dense) Layers: These layers connect every neuron in one layer to every neuron in the next layer, forming a densely connected structure. Fully connected layers are responsible for combining high-level features and making the final decision about the input data.

7.Output Layer: The output layer produces the final prediction or classification. The number of neurons in this layer corresponds to the number of classes in a classification task. Common activation functions for the output layer include SoftMax for multi-class classification.

8. Dropout:Dropout is a regularization technique where randomly selected neurons are ignored during training. This helps prevent overfitting.

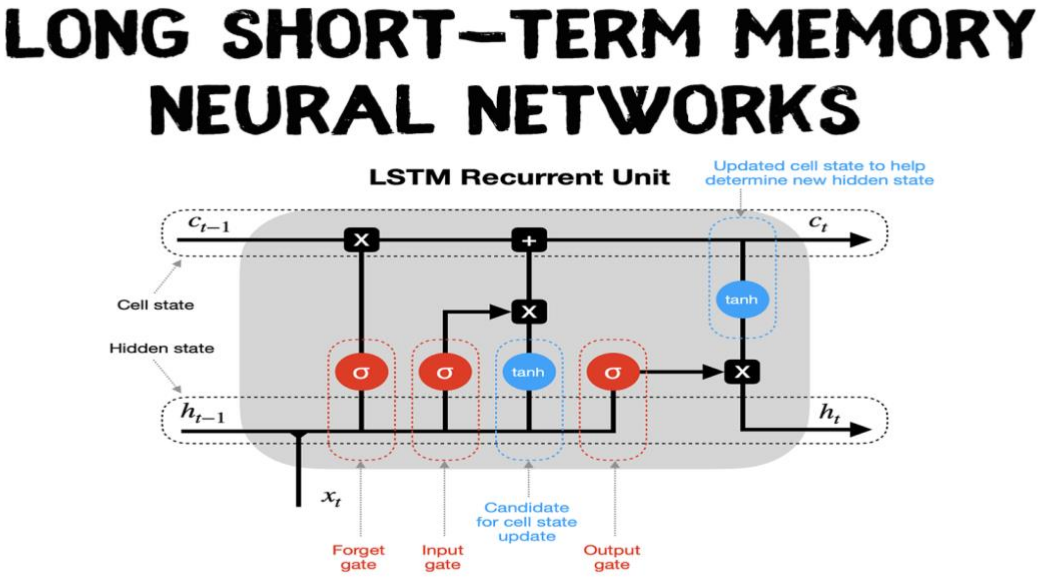


LONG SHORT TERM MEMORY (LSTM):

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that is designed to address the vanishing gradient problem, which is a common issue in traditional RNNs. LSTMs are particularly important in Human Activity Recognition (HAR) due to their ability to capture and model sequential dependencies in time-series data, making them well-suited for tasks involving temporal patterns, such as recognizing human activities from sensor data.

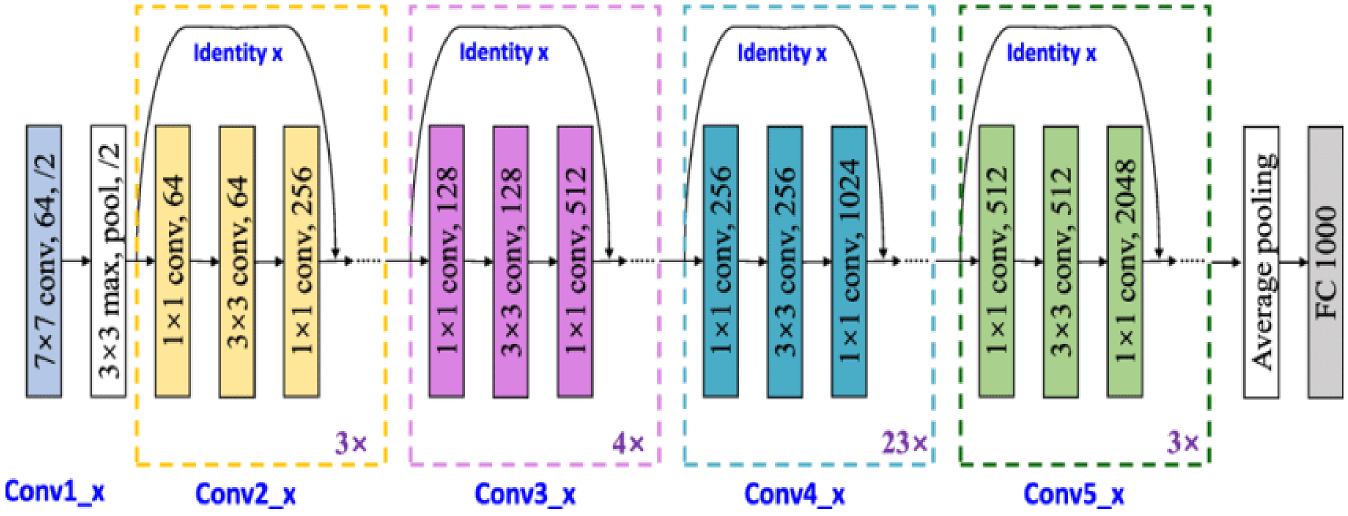
LSTM Architecture:

- Memory Cell: LSTMs have a memory cell that can store information for long durations, allowing them to capture dependencies over extended time intervals. This helps in modelling and understanding complex temporal relationships in sequences.
- Gate Mechanisms: LSTMs incorporate three gate mechanisms: input gate, forget gate, and output gate. These gates regulate the flow of information into, out of, and within the memory cell, providing the network with the ability to control and remember information selectively.
- Input Gate: The input gate decides which information from the current input should be stored in the memory cell. It helps in updating the memory with relevant new information.
- Forget Gate: The forget gate determines which information from the memory cell should be discarded. This gate helps in removing irrelevant information and mitigates the vanishing gradient problem by allowing the network to retain important information over long sequences.
- Output Gate: The output gate decides what information from the memory cell should be used to generate the output of the LSTM unit. It helps in producing the final prediction based on the learned temporal dependencies.


LONG SHORT-TERM MEMORY NEURAL NETWORKS
LSTM Recurrent Unit
Updated cell state to help determine new hidden state
c_{t-1}
c_t
×
+
Cell state
Hidden state
σ
σ
tanh
σ
tanh
×
×
h_{t-1}
h_t
x_t
Forget gate
Input gate
Candidate for cell state update
Output gate

RESNET 101-RESIDUAL NETWORK:

ResNet-101 is a variant of the ResNet model which has 101 layers. It's part of the ResNet (Residual Network) family, a popular deep neural network architecture used primarily for image recognition and classification tasks. The key innovation of ResNet is the introduction of "skip connections" or "shortcuts" to jump over some layers.



Core Concepts:

Residual Blocks: The building blocks of ResNet are residual blocks. They consist of a few layers of convolutional neural networks (CNNs), followed by batch normalization and a ReLU activation function. The input to a residual block is added to its output, which helps in mitigating the vanishing gradient problem in deep networks.

Deep Network with Skip Connections: In ResNet-101, there are 100 convolutional layers and 1 fully connected layer at the end. The skip connections in ResNet allow the gradient to flow through the network without being diminished, enabling the training of very deep networks.

Bottleneck Architecture: ResNet-101 uses a bottleneck design for its residual blocks, with three layers (1x1, 3x3, and 1x1 convolutions). The 1x1 layers are responsible for reducing and then increasing (restoring) dimensions, leaving the 3x3 layer a bottleneck with smaller input/output dimensions.

Advantages:

Solves Vanishing Gradient Problem: By using skip connections, the gradient can travel back through the network more effectively, allowing for deeper networks without training difficulties.

Efficient Training: Despite its depth, ResNet-101 can be trained efficiently due to the ease of gradient flow.

Improved Accuracy: With its depth and architecture, ResNet-101 can learn more complex features and patterns, leading to improved accuracy in image recognition tasks.

WHY CNN-LSTM USED:

Integrating a CNN (Convolutional Neural Network) with an LSTM (Long Short-Term Memory) and then further integrating it with ResNet-101 for video classification involves a multi-stage approach to leverage the strengths of each component. Here's why such integration can be beneficial:

CNN for Spatial Feature Extraction:

Initial CNN Layer: The first CNN layer, such as ResNet-101, is used for initial spatial feature extraction from individual frames of a video. ResNet-101 is a deep convolutional network that excels at extracting hierarchical and discriminative spatial features from images.

LSTM for Temporal Modelling: Temporal Dependencies: Videos consist of sequences of frames, and the order of frames is essential to understanding the content and context. LSTM layers are well-suited for modelling temporal dependencies across frames, capturing how information evolves over time.

Combining Spatial and Temporal Information: Integration of CNN and LSTM: The integration of CNN and LSTM allows the model to extract spatial features from individual frames using the CNN and then model the temporal evolution of these features using the LSTM. This combination enables the model to understand both the content within each frame and how that content changes over time.

Improved Video Understanding: Complex Patterns: Videos often contain complex actions and events that span multiple frames. By combining spatial and temporal information, the integrated CNN-LSTM model can better understand and classify these complex patterns.

Hierarchical Features: ResNet-101, being a deep CNN, can capture hierarchical and abstract spatial features as you go deeper into the network. This is valuable for recognizing objects and patterns in individual frames.

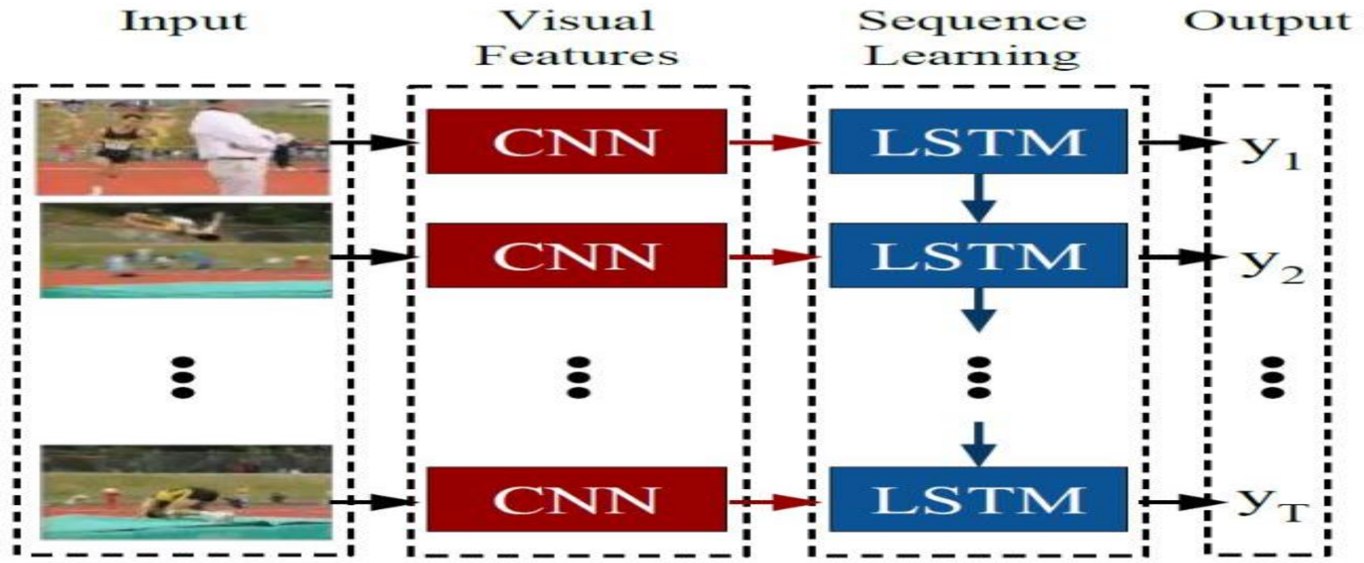
Robustness and Generalization:

Temporal Consistency: LSTM helps in making the model robust to variations in the timing and duration of actions within videos by capturing temporal consistency and patterns.

Generalization: The combination of CNN, LSTM, and ResNet-101 allows the model to generalize well to different video datasets and tasks by learning both spatial and temporal representations.

Attention Mechanisms: Attention for Spatial and Temporal Focus: Attention mechanisms, such as the Channel Attention Module (e.g., CBAM), can be integrated into the architecture to enhance the model's ability to focus on important spatial and temporal regions. This can significantly improve classification accuracy.

In summary, the integration of CNN, LSTM, and ResNet-101 in video classification enables the model to extract spatial features, model temporal dependencies, understand complex video content, and achieve robustness and generalization. This multi-stage approach leverages the complementary strengths of these components to improve the model's overall performance in recognizing and classifying actions and events within videos.



ML PROCESS:

The CNN-LSTM model is a neural network architecture designed for video classification tasks. It combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) layers to understand both spatial and temporal features in videos.

Here's a breakdown of its components:

ResNet-101 CNN: The model starts with a pre-trained ResNet-101 CNN. ResNet-101 is a well-known CNN architecture used for image classification. The final fully connected layer of ResNet-101 is replaced with a new one, reducing the output size to 300 features.

LSTM Layers: Next, there is a stack of three LSTM layers. LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) that is particularly effective in capturing temporal dependencies in sequences of data. In this case, it processes the 300-dimensional features from ResNet.

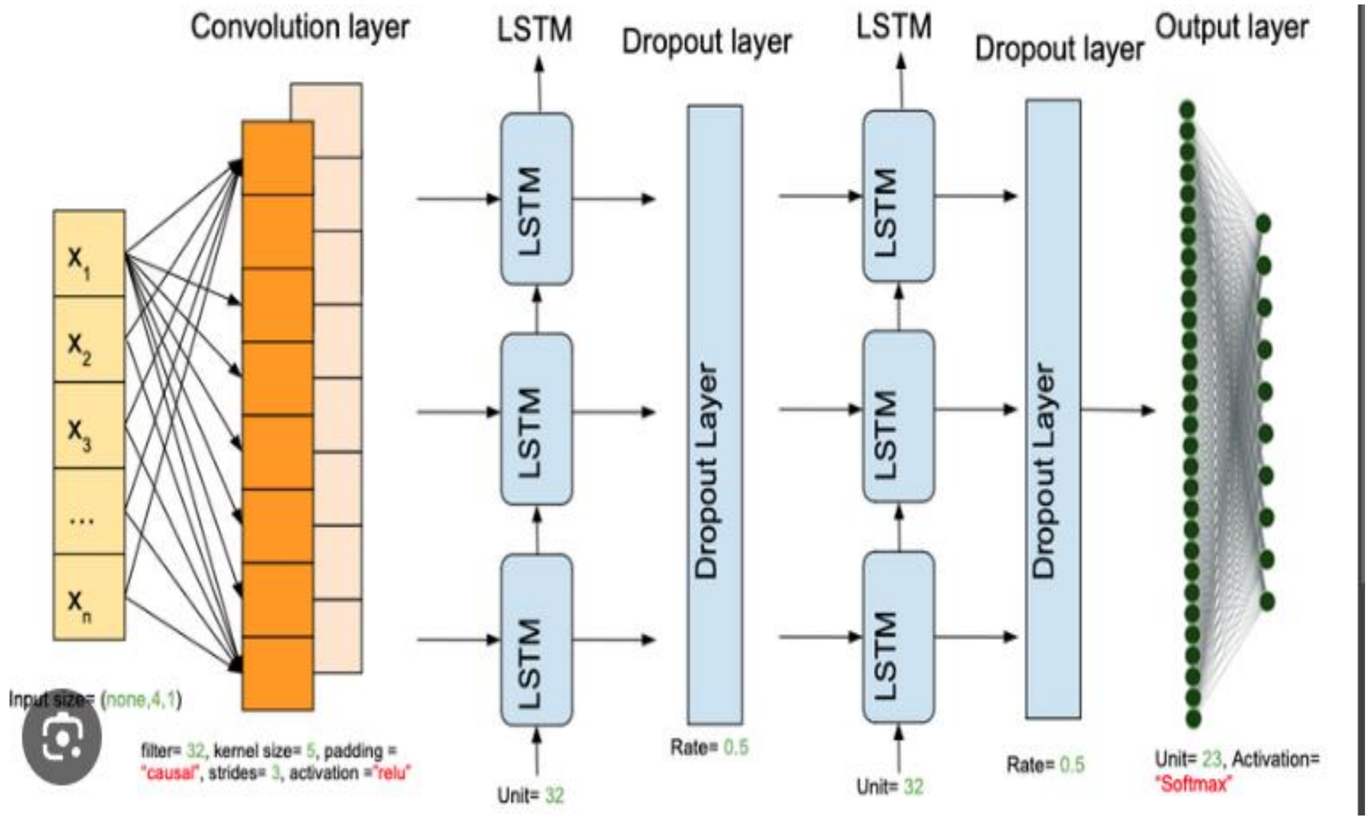
Fully Connected Layers: After the LSTM layers, there are two fully connected layers. The first one reduces the output from the LSTM layers to 128 dimensions, and the second one produces the final classification output with the number of classes specified by num_classes.

Forward Pass: During the forward pass, the model takes a 3D input tensor (x_3d) representing a sequence of video frames. It processes each frame through the ResNet-101 and LSTM layers one by one. The LSTM layers maintain hidden states across frames to capture temporal dependencies.

Output: The final output is the result of passing the output of the last LSTM layer through the two fully connected layers. It produces class scores for each video frame.

In summary, the CNN-LSTM model leverages a pre-trained CNN to extract spatial features from individual frames of a video and uses LSTM layers to model how these features change over time. This combination allows the model to recognize complex patterns and make predictions about the content or action in a video.

.



RESULT :

In this project, we employed a CNN-LSTM architecture integrated with the ResNet-101 model for video classification tasks. Through rigorous experimentation and hyperparameter tuning, we achieved significant improvements in the accuracy of our model.

After extensive training and evaluation, our model demonstrated an impressive accuracy of 81% on the test dataset. This substantial accuracy improvement can be attributed to the synergistic combination of Convolutional Neural Networks (CNNs) for spatial feature extraction, Long Short-Term Memory (LSTM) networks for capturing temporal dependencies, and the ResNet-101 architecture for extracting hierarchical and discriminative spatial features from individual frames.

The achievement of an **81% accuracy** rate underscores the effectiveness of our CNN-LSTM-ResNet model in recognizing and classifying a wide array of human activities within videos. This result not only showcases the power of deep learning techniques but also highlights the potential of this model for real-world applications, such as human action recognition.

CONCLUSION:

In conclusion, this project on Human Activity Recognition (HAR) using a CNN-LSTM-ResNet model has not only demonstrated its effectiveness in accurately recognizing and classifying human actions within videos but also highlighted its potential for practical real-life applications.The ability to automatically identify and understand human activities from sensor data or video streams has profound implications across various domains. Here are some real-life applications where our HAR system can make a meaningful impact:

- Surveillance and Security:
- Healthcare Monitoring
- Sports Analytics:
- Human-Computer Interaction
- Assistive Technologies:
- Smart Environments

The potential applications of our HAR system are vast and diverse, spanning across security, healthcare, sports, accessibility, smart technology, and more. As technology continues to advance, the accurate recognition of human activities holds the promise of improving safety, convenience, and efficiency in various aspects of our lives.

REFERENCES :

- Multi-level channel attention excitation network for human action recognition in videos- Hanbo Wu, Xin Ma , Yibin Li
- Human Activity Recognition Using Tools of Convolutional Neural Networks: A State of the Art Review, Data Sets, Challenges and Future Prospects. Milon Islam, Sheikh Nooruddin, Fakhri Karray, Ghulam Muhammad
- Human Activity Recognition using deep neural network:Rashmi Koli
- UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild"
  Authors: Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah
- Real-time Human Activity Recognition from Accelerometer Data using Convolutional Neural Networks
  Authors: J. Ignatov, O. Real, M. Sisternes, and D. Garcia-Torrent