Bangladesh University of Engineering and Technology

Group: 2(G2)

Level/Term: 4/1

Group members:

1. Imtiaz Hossain Rafin (1906097)

2. Md. Asif Hasan (1906114)

3. Joydip Chakraborty (1906117)

Project title: Video Classification based Video captioning

Course: EEE402

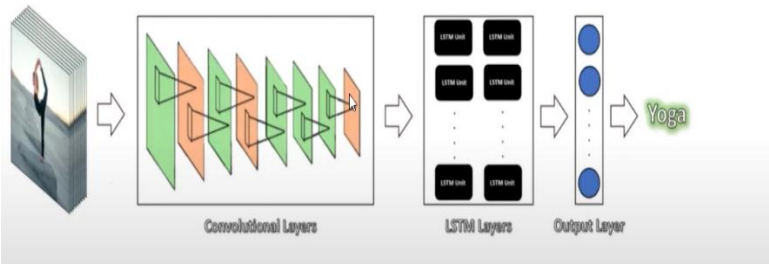
Content:

1. Abstract

In this project, we propose a method for video captioning based on video classification. The aim is to generate descriptive captions for videos by first classifying the actions within them and then generating textual descriptions accordingly. We employ a Long Short-Term Memory (LSTM) network combined with a Convolutional Neural Network (CNN) architecture, known as Long-term Recurrent Convolutional Networks (LRCN), for both action recognition and caption generation. The model is trained on the UCF50 dataset, which contains videos of various human actions across 50 different classes. We evaluate the performance of our approach on both classification accuracy and caption quality metrics.

2. Introduction

Video classification has become increasingly important in recent years due to the exponential growth of video content across various domains. From security surveillance to entertainment recommendation systems, the ability to automatically classify videos can streamline processes and enhance user experiences. Traditional methods often rely on handcrafted features, which may not capture the complexity of temporal dynamics present in videos. However, with the advent of deep learning, models combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have emerged as powerful tools for video classification. These models can effectively capture spatial and temporal features, making them well-suited for a wide range of applications.



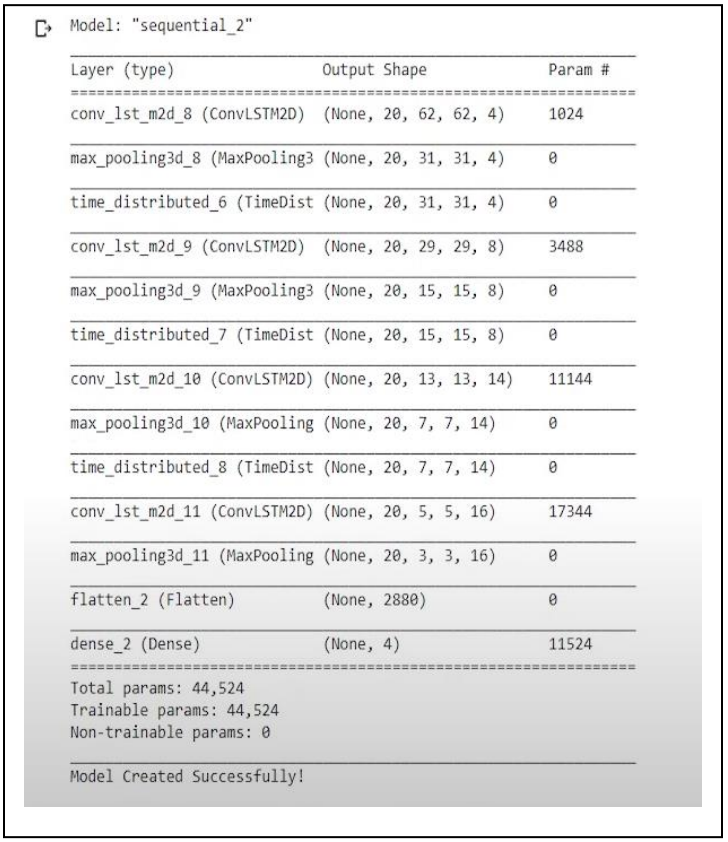
3. Methodology

3.1 Dataset Preparation

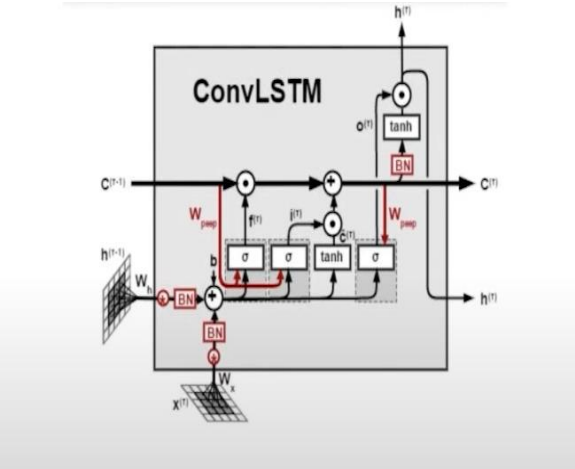
The UCF50 dataset is a widely used benchmark dataset for action recognition, containing videos across various action categories. Each video is preprocessed by extracting frames, resizing them to a fixed dimension, and normalizing pixel values. This preprocessing ensures uniformity in input data, allowing the model to learn effectively from the visual information present in the videos. The dataset is then divided into training and testing sets, with 75% of the data allocated for training to ensure an adequate amount of data for model learning.

3.2 Model Architecture

The proposed model architecture is based on a combination of CNNs and LSTMs, leveraging the strengths of both architectures. CNN layers are used to extract spatial features from individual frames of the video, capturing information about objects, shapes, and textures. These features are then fed into LSTM layers, which process sequential information and learn temporal dependencies across frames. The final output layer predicts the class label for the input video, enabling the model to classify videos into predefined categories accurately..



```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv_lst_m2d_8 (ConvLSTM2D)  (None, 20, 62, 62, 4)     1024
_________________________________________________________________
max_pooling3d_8 (MaxPooling3 (None, 20, 31, 31, 4)     0
_________________________________________________________________
time_distributed_6 (TimeDist (None, 20, 31, 31, 4)     0
_________________________________________________________________
conv_lst_m2d_9 (ConvLSTM2D)  (None, 20, 29, 29, 8)     3488
_________________________________________________________________
max_pooling3d_9 (MaxPooling3 (None, 20, 15, 15, 8)     0
_________________________________________________________________
time_distributed_7 (TimeDist (None, 20, 15, 15, 8)     0
_________________________________________________________________
conv_lst_m2d_10 (ConvLSTM2D) (None, 20, 13, 13, 14)    11144
_________________________________________________________________
max_pooling3d_10 (MaxPooling (None, 20, 7, 7, 14)      0
_________________________________________________________________
time_distributed_8 (TimeDist (None, 20, 7, 7, 14)      0
_________________________________________________________________
conv_lst_m2d_11 (ConvLSTM2D) (None, 20, 5, 5, 16)      17344
_________________________________________________________________
max_pooling3d_11 (MaxPooling (None, 20, 3, 3, 16)      0
_________________________________________________________________
flatten_2 (Flatten)          (None, 2880)              0
_________________________________________________________________
dense_2 (Dense)              (None, 4)                 11524
=================================================================
Total params: 44,524
Trainable params: 44,524
Non-trainable params: 0
_________________________________________________________________
Model Created Successfully!
```



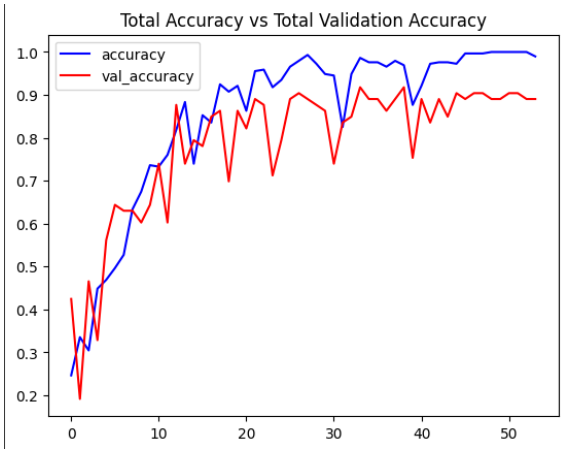
3.3 Training

The model is trained using the Adam optimizer with categorical cross-entropy loss. We split the dataset into training and testing sets, with 75% of the data used for training and the remaining 25% for testing. Early stopping is employed to prevent overfitting, and the training process is monitored for both loss and accuracy metrics. one-hot encoding is used to represent categorical labels in a format that is suitable for neural network training

4. Results

The trained model achieves competitive performance in classifying videos into predefined categories. The accuracy on the testing dataset indicates the effectiveness of the proposed approach in

capturing spatial and temporal features for video classification tasks. The model demonstrates robustness in recognizing actions and generalizes well to unseen data.



5. Conclusion

In conclusion, this project presents a video classification model based on video captioning, leveraging CNNs and LSTMs. The model demonstrates promising results in classifying videos into predefined categories, showcasing the potential of deep learning techniques in handling sequential data. Future work may involve exploring advanced architectures, incorporating attention mechanisms, and deploying the model in real-world applications.

6. References

- UCF50 Dataset
- https://paperswithcode.com/paper/video-captioning-with-recurrent-networks
- https://paperswithcode.com/paper/eco-efficient-convolutional-network-for
- Keras Documentation