DEEP LEARNING DS3040

Project Report

Human Activity Recognition

Name: ADITYA DANDRIYAL **||** 142202003 **||** MTech Data Science

Group Member: Argyhadeep Ghosh (142202017)

Introduction:

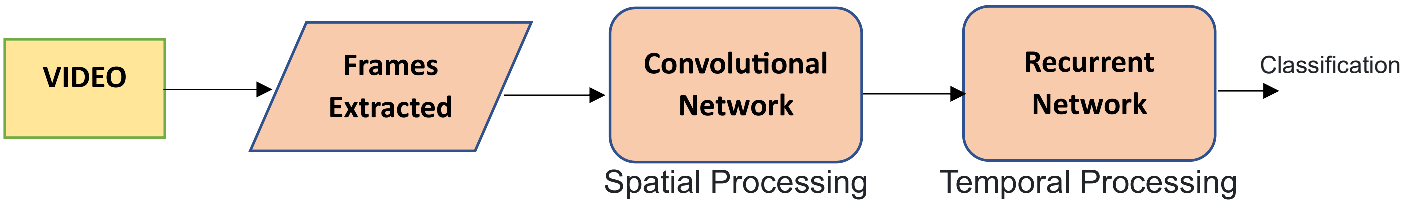
In this project we worked on a subset of UCF-101 dataset in which we considered only 21 classes of human activity/actions instead of the original 101.

A video consists of an ordered sequence of frames. Each frame contains **spatial information**, and the sequence of those frames contains **temporal information**. To model both of these aspects, we use a hybrid architecture that consists of **convolutions** (for spatial processing) as well as **recurrent layers** (for temporal processing).

Approach:

- Since a video is an ordered sequence of frames, we extracted the frames from the video at a fixed interval until a maximum frame count (predefined) was reached.
- Firstly, from the folder of the UCF-101 dataset we extracted the video paths and the corresponding labels (which we further one-hot encoded) of each video contained in the subset relevant to the project requirements.
- Then we applied the prespecified train-test split instead of a random one, because if the videos belonging to the same group are present in both the training and the testing datasets, then this would give a false high performance of our models.
- For each of the videos we extracted 24 frames and also resized the frames to 224x224.
- Also while fitting the model we used data loader to load the data in batches.
- For all the considered models, loss=>‘**Categorical Cross-Entropy’** , optimizer=> ‘**Adam**’, metrics=> ‘**Accuracy**’.
- Then we considered cases with and without transfer learning, and also considered different pre-trained models for transfer learning. Along with this we considered cases using different types of recurrent layers (RNN, LSTM, GRU).
- The training and validation accuracies for the considered cases is tabulated as follows:

Basic Pipeline:



CASE 1: [CNN-RNN architecture]

MODEL	TRAIN ACCURACY	VALIDATION ACCURACY
3DConv + LSTM	0.30	0.29
ConvLSTM	0.94	0.48
LRCN	0.31	0.30
CNN + GRU	0.30	0.29
CNN + RNN	0.30	0.29

CASE 2: [Transfer Learning: VGG16]

We made use of VGG16 Pretrained Model to extract meaningful features from the video frames.

MODEL	TRAIN ACCURACY	VALIDATION ACCURACY
LSTM	0.97	0.81
GRU	0.93	0.77
SimpleRNN	0.90	0.70

CASE 3: [Transfer Learning: Resnet-50]

We made use of Resnet-50 Pretrained Model to extract meaningful features from the video frames.

MODEL	TRAIN ACCURACY	VALIDATION ACCURACY
LSTM	0.96	0.81
GRU	0.85	0.75
SimpleRNN	0.86	0.75

CASE 4: [Transfer Learning: Efficient Net-B7]

We made use of Efficient Net-B7 Pretrained Model to extract meaningful features from the video frames.

MODEL	TRAIN ACCURACY	VALIDATION ACCURACY
LSTM	0.95	0.79
GRU	0.81	0.72
SimpleRNN	0.81	0.65

Inferences drawn:

- Using Pre-trained model for extracting meaningful features from the video frames gives a very high accuracy score as compared to not using transfer learning.
- Greatest validation accuracy is established when using either VGG16 or Resnet-50 for feature extraction along with LSTM model.

Individual Contribution:

Along with the implementation of the general structure of the project the analysis for cases 1 and 2 has been done by me.

References:

- https://keras.io/examples/vision/video_classification/
- https://www.bleedai.com/human-activity-recognition-using-tensorflow-cnn-lstm/
- https://www.tensorflow.org/tutorials/images/transfer_learning