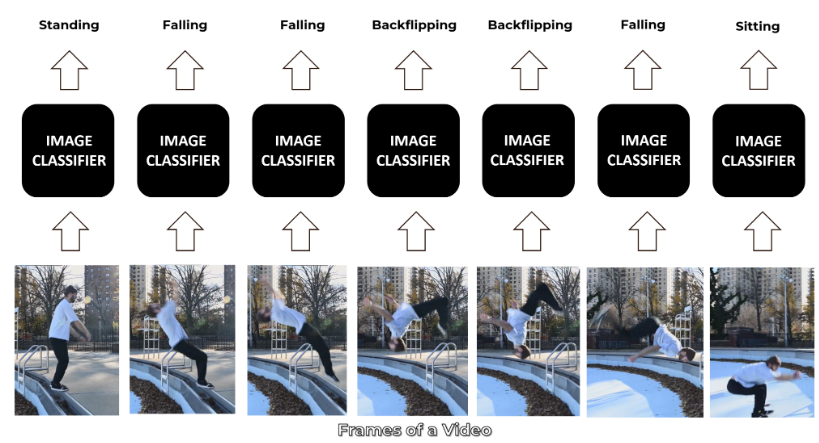
Methodology

**Approach 1: Single-Frame Classification**

The simplest and most basic way of classifying actions in a video can be using an image classifier on each frame of the video and classify action in each frame independently. So if we implement this approach for a video of a person doing a backflip, we will get the following results.

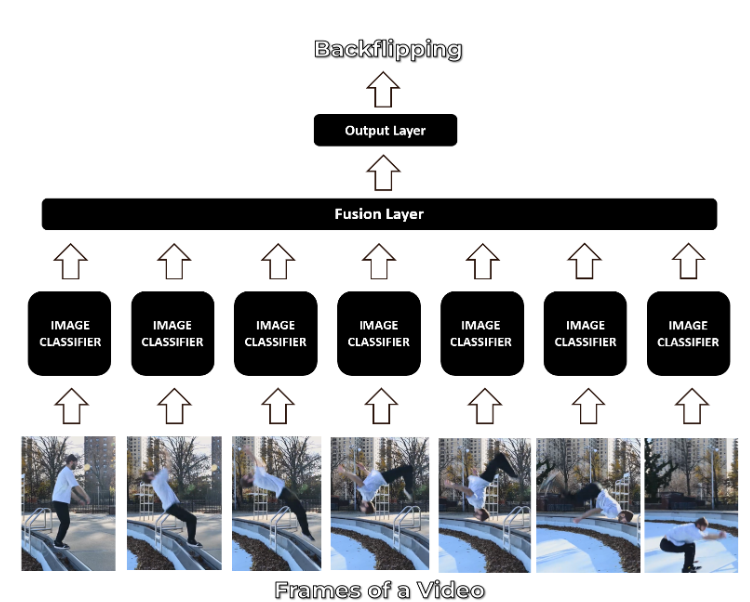


The classifier predicts Falling in some frames instead of Backflipping because this approach ignores the temporal relation of the frames sequence. And even if a person looks at those frames independently he may think that the person is Falling.

Now a simple way to get a final prediction for the video is to consider the most frequent one which can work in simple scenarios but is Falling in our case and is not correct. So another way to go about this is to take an average of the probabilities of predictions and get a more robust final prediction.

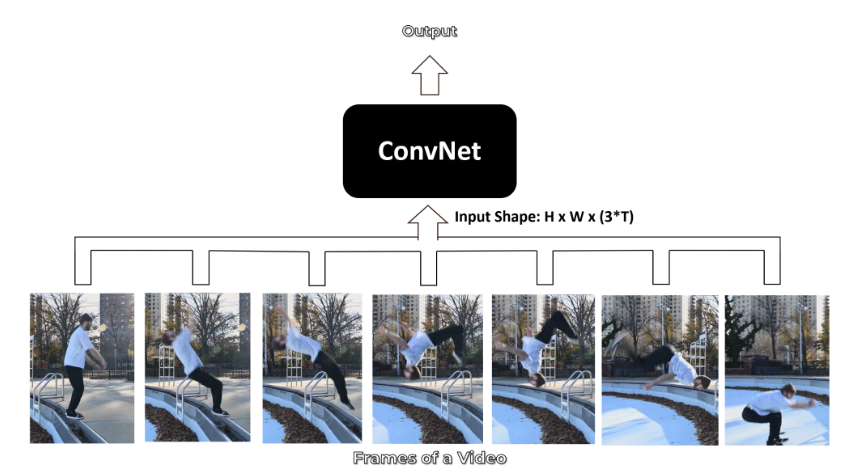
**Approach 2: Late Fusion**

Another slightly different approach is late fusion, in which after performing predictions on each frame independently, the classification results are passed to a fusion layer that merges all the information and makes the prediction. This approach also leverages the temporal information of the data.



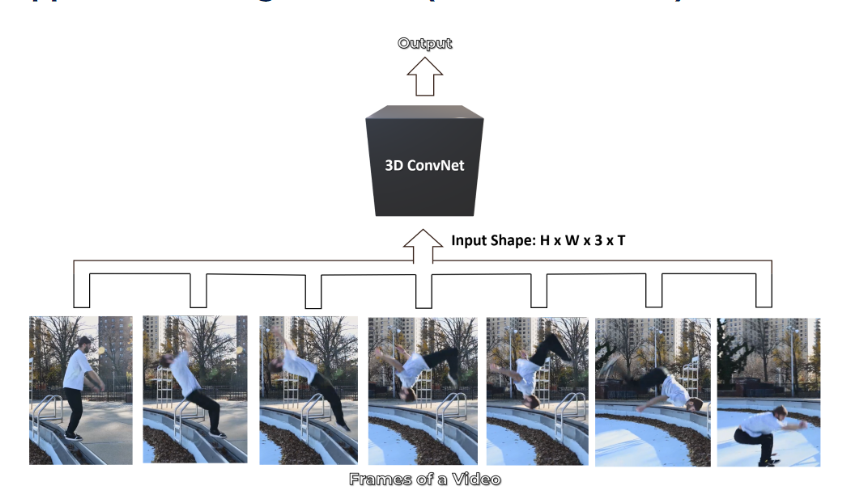
This approach does give decent results but is still not powerful enough.

**Approach 3: Early Fusion**

****

Another approach of video classification is early fusion, in which all the information is merged at the beginning of the network, unlike late fusion which merges the information in the end. This is a powerful approach but still has its own limitations.

**Approach 4: Using 3D CNN’s (aka. Slow Fusion)**

****

Another option is to use a 3D Convolutional Network, where the temporal and spatial information are merged slowly throughout the whole network that is why it’s called Slow Fusion. But a disadvantage of this approach is that it is computationally really expensive so it is pretty slow.

**Approach 5: Using Pose Detection and LSTM**



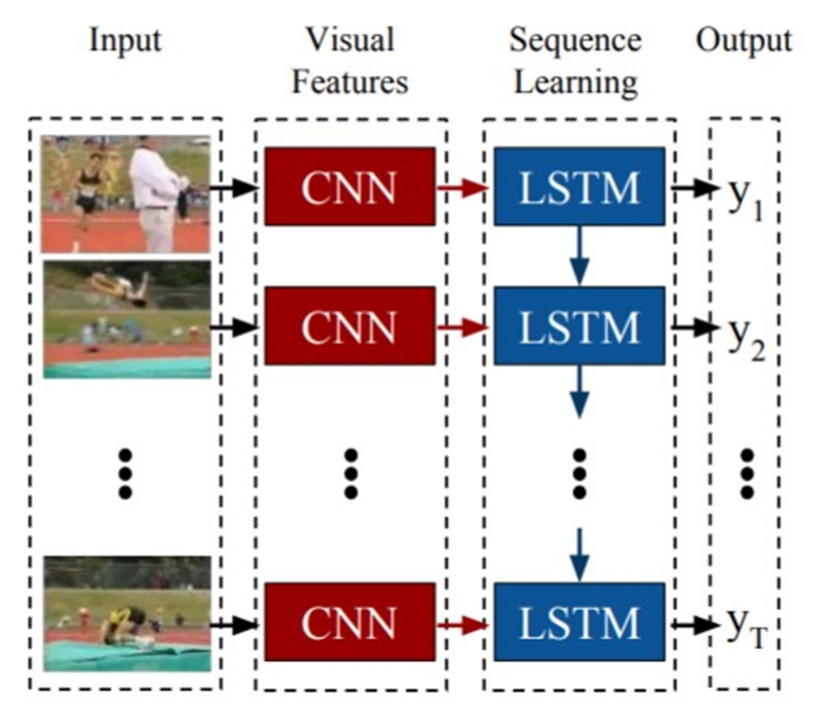
Another method is to use a pose detection network on the video to get the landmark coordinates of the person for each frame in the video. And then feed the landmarks to an LSTM Network to predict the activity of the person.

There are already a lot of efficient pose detectors out there that can be used for this approach. But a disadvantage of using this approach is that you discard all the information other than the landmarks, like the environment information can be very useful, for example for playing football action category the stadium and uniform info can help the model a lot in predicting the action accurately.

Before going to the approach that we will implement.

**Approach 6: CNN + LSTM**

We will be using a CNN to extract spatial features at a given time step in the input sequence (video) and then an LSTM to identify temporal relations between frames.



The two architectures that we will be using to use CNN along with LSTM are:

1. **ConvLSTM**
2. **LRCN**

Both of these approaches can be used using TensorFlow.