Name - Chandan Kumar Roy
Email - chandankumar.r18@iiits.in

REPORT

HUMAN ACTION RECOGNITION

Abstract

Human action recognition is a standard Computer Vision problem and has been well studied. The fundamental goal is to analyze a video to identify the actions taking place in the video. Essentially a video has a spatial aspect to it ie. the individual frames and a temporal aspect ie. the ordering of the frames. Some actions (eg. standing, running, playing, skateboarding etc.) can probably be identified by using just a single frame but for more complex actions(eg. walking vs running, bending vs falling) might require more than 1 frame's information to identify it correctly. Local temporal information plays an important role in differentiating between such actions. Moreover, for some use cases, local temporal information isn't sufficient and you might need long duration temporal information to correctly identify the action or classify the video.

Dataset

Kinetics dataset was first introduced in 2017 primarily for human action classification.t was developed by the researchers: Will Kay, Joao Carreira, Chloe Hillier and Andrew Zisserman at Deepmind. The dataset contains 400 human activity classes, within any event 400 video cuts for each activity. It has 306,245 recordings and is separated into three parts, one for preparing to have 250–1000 recordings for each class, one for approval with 50 recordings per class and one for testing with 100 recordings for every class. Each clip endures around 10s.
Kinetics dataset are taken from Youtube recordings. The activities are human focussed and cover a wide scope of classes including human-object communications, for example mowing lawn, washing dishes, humans Actions e.g. drawing, drinking, laughing, pumping fist; human-human interactions, e.g. hugging, kissing, shaking hands. Since the dataset is huge and downloading each clip would be a waste of time given that we already have pre-trained models by the original author. It would be smarter to work on the pre-trained model than to train and tune it separately.

Model

I have used ResNet_34 3D model for human action recognition. The difference between the normal resnet model and Resnet3D model is that it uses 3d convolution layers instead of 2D convolution layer. More details about the 3D ResNe_34 model can be found here. For more information about the training, please go through this link. I have used opencv and pytorch to make the inference.

Demo

I have provided a python script to run the human action recognition task.

You can simply run the script using below command

“Python har.py -l <link:optional>
Example :

```
(pytorch_x86) chandanroy@Chandans-MacBook-Air HAR % python har.py -l "https://www.youtube.com/watch?v=668nUCeBHyY"
```

I have also provided a default video and the command line argument “-l” is completely optional. In case you want to try the default video, please use the following command.

```
(pytorch_x86) chandanroy@Chandans-MacBook-Air HAR % python har.py
```

References

https://openaccess.thecvf.com/content_cvpr_2018/html/Hara_Can_Spatiotemporal_3D_CVPR_2018_paper.html