Deep Learning on Video Classification

Jingxi Liu, Yuetong Liu, Yingkun Wang, Jiaqian Zhong
jliu238, yliu390, ywang601, jzhong16

1 Introduction

Video classification becomes a hot trend in current technological field. With deep learning methods, we could analyze human actions, gestures, and do AR interactions in videos. We learnt and practiced many concepts about image classification in class, and we further think videos are a collection of a set of images arranged in a specific order. As a result, we want to do a video related project to both solid what we've learnt in this semester and to have some innovative ideas go beyond. In this project, we are going to choose the sports videos in UCF101 dataset to perform motion classification task. We will first pre-process videos to get important features and representations. Then we will choose appropriate models to train and validate the data. Finally, we will use the test data to find out the most accurate model. Our ultimate goal is to accurately classify different kind of sports.

2 Dataset and Features

UCF101 is an action recognition dataset of realistic action videos, collected from YouTube, having 101 action categories. The action categories can be divided into five types: 1)Human-Object Interaction 2) Body-Motion Only 3) Human-Human Interaction 4) Playing Musical Instruments 5) Sports. In this project, we will focus on Sports. There are 50 action categories in Sports, and each action category is grouped into 25 groups, where each group can consist of 4-7 videos of an action.

The videos from the same group may share some common features, such as similar background, similar viewpoint, etc. Therefore, we will sample 1 video per group. Then there are 50 classes (action categories) in our data set, and each class has 25 data entries (videos).

3 Methods

For the data pre-processing part, we would like to use OpenCV. Specifically, we would like to select one or a few more frames each second from the video and resize the images to standardize the input. After data pre-processing, the data will be split into training, validation and testing set with a ratio of 3:1:1.

For the baseline model, we would like to try CNN as our classification method, because a video is just a series of frames. We want to loop over all selected frames in a video, and use CNN to classify each frame individually. Then we choose the label which has the largest probability as our predicted output.

To further improve our model, we will augment the dataset by flipping, rotation and cropping, etc as well as including more frames from the selected videos. In terms of the model architecture, we would like to apply a 3D convolution network called C3D as our improved model because it has been widely implemented in video classification and has achieved 0.85 accuracy on the UCF101 dataset. In addition, we would like to add layers that are able to capture the relation among consecutive frames. For example, one state-of-art network called ARTNet-ResNet18 improve C3D by stacking multiple generic building blocks to simultaneously model appearance and relation from RGB input.

4 Evaluation

Computing accuracy score is our major way to evaluate models. Also, similar to our homework, we would like to plot graphs between accuracy and loss to visualize the model performance.

References

[1] Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015. APA

[2] Wang, Limin, et al. "Appearance-and-relation networks for video classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[3] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild., CRCV-TR-12-01, November, 2012.