Video Action Recognition

Runxin Gao
rgao35@wisc.edu

Kenny Jin
jjin59@wisc.edu

Tong Li
tli287@wisc.edu

Nan Yang
nyang49@wisc.edu

1. Introduction

Image classification is an important and popular research topic of computer vision. Compared to image classification, the task of video classification has raised fewer attention by researchers. Computer vision tasks regarding videos are more complicated than those of images in that each video consists of multiple images and motions of objects are involved. For this project, we plan to use deep learning methods to build a program that can classify videos. Specifically, we will work with video datasets that are about human daily actions, and the program we build will be able to classify these videos into different categories of human action when we pursue a high classification accuracy.

1.1. Related Works

Previous works regarding image classification usually utilize Convolutional Neural Networks (CNN). Since a video clip can be viewed as a stack of images, using CNNs for video classification seems to be a viable approach. Many researchers have tried 2D CNNs for video classification. For example, Karpathy et al.[2] improved and evaluated their 2D CNN models on the Sports-1M dataset, which consists of 1 million YouTube videos annotated with 487 classes. However, it is difficult for 2D CNN models to capture temporal structures of videos. Many other researchers have used 3D CNNs for video classification since 3D CNNs can both deal with temporal and spatial features simultaneously. For example, Tran et al.[4] used 3D CNNs to learn spatiotemporal features of videos and achieved good prediction accuracy on UCF-101 dataset. Training and using 3D CNNs, however, is computationally expensive. Thus, later efforts regarding this topic involve both lowering the computational costs and improving the classification accuracy. Recently, Lin et al.[3] proposed Temporal Shift Module (TSM) for video recognition that is efficient and accurate. The TSM model achieved state of the art performance on the Something-something video dataset. In addition, Feichtenhofer et al.[1] proposed SlowFast Networks, which also achieved state of the art performance in other video recognition benchmarks.

2. Motivation

Recently, there has been significant development in image classification with the state-of-the-art neural networks such as CNNs and RNNs. Most of the neural networks focus on categorizing still objects. However, the physical world we are living in is not static. Moving actions cannot be captured by a single image. Therefore, to better understand the dynamic world, we want to explore models that are able to perform video classification and recognize human activities, and compare performances of different algorithms.

Successful (high prediction accuracy) classifiers on human actions can be used for enhancing public security practices. They can be integrated into surveillance camera systems to effectively detect whether hazardous activities are happening, and can provide timely alerts for police agencies to react. Such automatic detection is likely to perform better than humans' in that it can get a broader view through the cameras than human eyes can, and thus provide warnings even on blind spots. In addition, it can also be used for home security by adapting this classifier in home security cameras, and thus alert homeowners something is happening around their properties.

Moreover, this classifier can be utilized for studying infant behavioral development. It can automatically detect the infants' behavior through cameras on volunteering participants, which thus diminishes the need for human observers and offers them more spare time to do something else. Likewise, it could also result in better performance because it is able to detect infant behaviors 247 uninterruptedly for a long period, which is impossible for human observers. Therefore, such a classifier can help us better understand behavioral development of the initial stage of human beings.

3. Evaluation

Although this is a human action classification task, the key remains the same as that of all classification problems: prediction accuracy. Therefore, the most straightforward way to measure the performance of the model is to divide the number of correctly categorized items over the total number of items, which can be constructed as follows:

$$\text{ACC1} = \frac{\text{Total \# Correctly Predicted (Exactly Match)}}{\text{Total Number of Videos}}$$

However, such evaluation scheme seems harsh, given that some categories can be confusingly similar, and different people may reach different results for the same video. Therefore, we may also add another accuracy scheme that only counts the classification as correct if the top 5 predictions contain the true label. The equation is shown as follows:

$$\text{ACC2} = \frac{\text{Total \# Correctly Predicted (Top 5 Predictions)}}{\text{Total Number of Videos}}$$

4. Resources

The data set comes from 20 Billion Neurons site [1]. It has 220,847 total videos, 174 labels, and 168,913, 24,777, 27,157 videos in training, validation, test set, respectively. The videos are in Webm format with VP9 as encoding, have a resolution of 240px, and are named with number 1 to 220847. In each video, a human is conducting a pre-defined action with common projects such as pouring water into a glass. All videos are collected by crowd workers. The labels describe the general action instead of the detailed action, for example, it labels the action “pouring water into a glass” as “pouring something into something”. The data set is not balanced, with the largest number of labels (4081) for “putting something on a surface” , and the least (115) for “poking a hole into some substance.” Since we will primarily use our personal computers for the task, we plan to only use 8G video data and at most 50 labels. We may also resort to cloud computing resources such as Google Cloud if need arises.

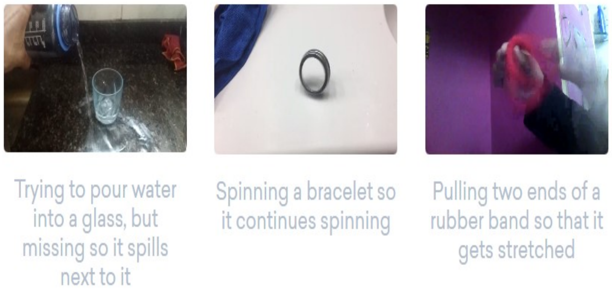
Trying to pour water into a glass, but missing so it spills next to it
Spinning a bracelet so it continues spinning
Pulling two ends of a rubber band so that it gets stretched

Figure 1. A snippet of data with labels

5. Contributions

Each group member contributed evenly to this proposal. For experiments each group member is responsible to train a kind of classification model and then all group members will come up ways to optimize these models. For the final report and presentation each member will contribute evenly.

References

[1] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition, 2018. arXiv:1812.03982.

[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[3] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding, 2018. arXiv:1811.08383.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[1] https://20bn.com/datasets/something-something/v2