Group Proposal

Deyu Kong
Aihan Liu

1. TOPIC

   With the development of the Internet, more and more videos are shared online. The computer vision community is also developing research on video, such as behavior recognition, abnormal event detection, activity understanding, etc. Considerable progress has been made on these individual problems by employing different specific solutions. However, a broad set of video feature representation learning methods is still needed for solving large-scale video tasks.

   The topic we chose is video classification, also called video recognition. It is probably the most straightforward computer vision task related to the video. Unlike video object detection, it is easier to understand, and we have enough time to understand its knowledge.

2. DATASET

   The data set we use is UCF101. It consists of 101 categories and 13,320 videos recorded in unconstrained environments and uploaded to YouTube, featuring camera motion, various lighting conditions, partial occlusion, low-quality frames, etc.

   We focus on the playing Musical Instruments group. They have 1268 clips in total, and each of them contains 4-7 videos, and the videos from the same group have some similar characteristics, such as background, characters, etc. This dataset is large enough to train a deep network.

3. NETWORK

   The network we used is based on CNN3D. The 3D convolution kernel is more efficient for spatiotemporal feature learning. Tran et. (2015) proposes a network called C3D, and we will customize it with different kernel size or neural size to learn this architecture more clearly.

4. FRAMEWORK

   Since we want to customize the network ourselves, PyTorch is the best option for debugging and changing network architectures.

5. MATERIALS

   We will look for the published papers related to the video classification tasks, and the PyTorch help documents will help us understand the framework.

6. PERFORMANCE

   We will use the metrics in Exam2 to judge the network's performance and accuracy - hlm. Also, because this is a multiclassification task, we will look at the F1 score.

7. SCHEDULE
   - Late March and early April: looking for the questions and dataset we are interested in (done)
   - The first week of April: complete the data loader and run the C3D successfully.
   - The second week of April: complete a customized model.
   - The third week of April: complete demo, presentation slides, and report
   - April 26th: presentation

Reference

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489-4497).