Violence Detection in Videos

Mustafa Berk Yaldiz, Krushnat More, Jayant Prakash, Pradyumna Mukunda

College of Computing, Georgia Institute of Technology, Atlanta, GA

INTRODUCTION

With widespread internet, we have seen exponential increase in multimedia contents i.e. images, audios, videos. These contents are being created/updated and shared across causing information explosion. We have a problem of understanding these videos for different purposes like search, recommendation, ranking etc. Problems like action recognition, abnormal event detection, and activity understanding of video analysis have been attempted by computer vision community for decades. Considerable contributions has been made in each of these problems by employing specific solutions. However, we still lack a generic video descriptor or classifier solution and we feel need for it, which could help us in solving large-scale video tasks.

In last few years, because of advent of deep learning in the image/video territory we have seen rapid progress have been made in feature learning, various pre-trained convolutional network models are made available for extracting image features. However, image based deep features or training weights doesn't directly work for videos because of lack of motion modelling or information.

We use a simple, yet effective approach for spatiotemporal feature learning using deep 3-dimensional convolutional networks (3D ConvNets) trained on a large scale supervised video dataset. [1]

The action recognition community has focused mostly on detecting simple actions like clapping, walking or jogging, whereas the detection of fights or in general aggressive behaviors has been comparatively less studied. Such capability may be extremely useful in some video surveillance scenarios like in prisons, psychiatric or elderly centers or even in camera phones. [5]

APPROACH

The paper we picked up proposes learning of spatio-temporal features using deep 3D ConvNet. [1] We empirically show that these learned features with a simple linear classifier can yield good performance on various video analysis tasks.

Although 3D ConvNets were proposed before, to our knowledge this work exploits 3D ConvNets in the context of large-scale supervised training datasets and modern deep architectures to achieve the best performance on different types of video analysis tasks. The features from these 3D ConvNets encapsulate information related to objects, scenes and actions in a video, making them useful for various tasks without requiring to finetune the model for each task. C3D has the properties that a good descriptor should have: it is generic, compact, simple and efficient.

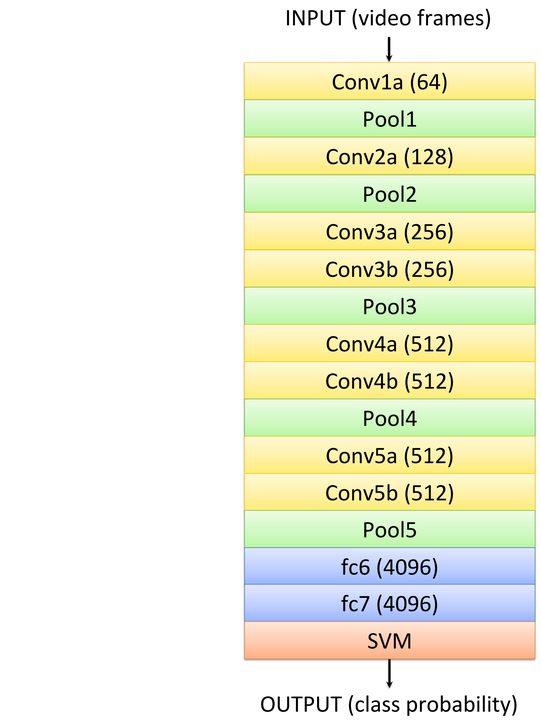
We use different datasets at each level. Like we use "Sports-1M" for initial training process. Then, we do transfer learning. We take learned weights and fine tune those with dataset "UCF 101", to enhance the model accuracy.

We are using SVM classifier at the end, replacing the approach mentioned in the research paper. In our experiments, we saw a significant boost in accuracy with SVM classifier so we decided to use it to be successful in video classification project.

To measure success of a model, we will use test accuracy on unseen dataset like "Fight Detection". We will also check precision, because precision expresses the proportion of the data points our model says was relevant actually were relevant.

The project was implemented using TensorFlow and executed on the Google Cloud Platform.

MODEL ARCHITECTURE



DATA AND PREPROCESSING

We used a multitude of datasets for implementing the violence detection task:

Sports-1M:

Sports-1M dataset collected by Stanford University contains 1,133,158 video URLs which have been annotated automatically with 487 sports labels using the YouTube Topics API. [3]

UCF-101:

An action recognition data set of realistic action videos, collected from YouTube by the University of Central Florida. The dataset is of size 6.5 GB and consists of 13320 videos from 101 action categories. [4]

Violence Detection in Video using Computer Vision Techniques:

This collection is composed of two separate datasets, a Hockey Fight dataset and a Movies dataset. It consists of 1000 video sequences divided in two groups: fights and non-fights. [5]

Violent-Flows Database:

A database of real-world, video footage of crowd violence. The data set contains 246 clips downloaded from YouTube, each of length 1-6 seconds. [6]

The videos are in compressed formats such as mp4, mpg and avi. The individual frames are extracted using ffmpeg and then converted to numpy arrays. This preprocessing is done on-the-fly while training.

The size and diversity of the data was augmented using transformations like mirroring, cropping, off-centering and subtracting mean on the original videos.

DISCUSSION AND RESULT

The model is pre-trained on Sports-1M dataset and then trained the last two fully connected layer with UCF-101 dataset and it should boost the overall accuracy .

Initial C3D paper tried to predict the videos among 10 classes. However, this paper tried to distinguish the video between violent and non violent videos, so the baseline accuracy of this paper is greater than the accuracy of initial C3D paper. Initial training on small UCF dataset is done and getting the training accuracy of around 99% but we still need to train on the whole UCF dataset. This paper expects to have the test accuracy of more than 90% after training on the whole dataset.

This paper used the SVM layer instead of Softmax to predict the violence in the videos . The movie dataset is used to train the last SVM layer. So finally, this paper used Sports-1M dataset to train the convolution and pooling layers, last two fully connected layers are trained with UCF-101 and the last SVM layer is trained with movie dataset.

FUTURE WORK

This paper used one stream model for capturing spatiotemporal information. The two stream network one focusing on spatial information and another on temporal information can also be used in the future. In the future, batch normalization could also be added along with this model to boost the accuracy.

REFERENCES

[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, Learning spatiotemporal features with 3D convolutional networks, arXiv:1412.0767, 2015

[2] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. Arzani, R. Yousefzadeh and L. Van Gool, Temporal 3D ConvNets: New architecture and transfer learning for video classification, arXiv:1711.08200, 2017

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, Large-scale video classification with convolutional neural networks, CVPR, 2014

[4] K. Soomro, A. R. Zamir and M. Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, CRCV-TR-12-01, 2012

[5] E. Bermejo, O. Deniz, G. Bueno and R. Sukthankar, Violence Detection in Video using Computer Vision Techniques, Computer Analysis of Images and Patterns, 2011

[6] T. Hassner, Y. Itcher, and O. Kliper-Gross, Violent Flows: Real-Time Detection of Violent Crowd Behavior, CVPR, 2012