Violence Detection Project

Team Members -

Mustafa Yaldiz (myaldiz3@gatech.edu)
Jayant Prakash (jayant2205@gatech.edu)
Krushnat More (kmore3@gatech.edu)
Pradyumna (pmukunda3@gatech.edu)

Introduction / Background / Motivation -

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.

There are four properties for an effective video descriptor or classifier:

1) We need it to be generic, so that it can help us represent different video types well while being discriminative as well. For. e.g. we have food, pets, movies, tv shows, sports clips are some of its types.
2) We need it to be compact as well, we have billions of videos and they are getting generated or spread across very fast, we need to operate at a large scale, so we need a compact descriptor which could help us in processing, storing and retrieving videos at scale.
3) We need it to be efficient. As thousands of videos are expected to be processed every minute in real world situation, so we need computationally efficient system.
4) System can be simple to implement. Instead of using complicated feature encoding methods and classifiers, a good descriptor should work well even with a simple model (e.g. linear classifier).

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.

These features we get as activations of the CNN’s last fully connected layers, which we can use as it is and perform a transfer learning which works very well. However, image based deep features or training weights doesn’t directly work for videos because of lack of motion modelling or information.

The paper we picked up proposes learning of spatio-temporal features using deep 3D ConvNet. 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.

Approach -

We propose 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.
Our findings are three-fold:
1) 3D ConvNets are more suitable for spatiotemporal feature learning compared to 2D ConvNets;
2) A homogeneous architecture with small 3 × 3 × 3 convolution kernels in all layers is among the best performing architectures for 3D ConvNets;
3) Our learned features, namely C3D (Convolutional 3D), with a simple linear classifier outperform state-of-the-art methods on 4 different benchmarks and are comparable with current best methods on the other 2 benchmarks.

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.

We used 3d convolutional neural networks to detect violence through transfer learning, we use learned weights and train those with dataset “UCF 101”, to enhance the model accuracy.

Experimental Plan -

We use different datasets at each level, like we use “Sports 1M” for initial training process. Then, we use trained weights and do transfer learning and use dataset “UCF 101” for further training.
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.

Dataset links -
https://docs.opencv.org/3.0-beta/modules/datasets/doc/datasets/ar_sports.html
https://www.crcv.ucf.edu/data/UCF101/UCF101.rar
https://cs.stanford.edu/people/karpathy/deepvideo/
https://www.openu.ac.il/home/hassner/data/violentflows/
http://visilab.etsii.uclm.es/personas/oscar/FightDetection/index.html

---

Current Status -

We have finalized our approach and model, we’ve prepared setup as well to work with the model. We have partial but working implementation ready. Currently, we are working on improvising model as discussed in aforementioned sections. Our task distribution is as follows,

Mustafa Yaldiz -
1) Found the base paper and worked on it, designed end to end flow for the model based on the paper.

Paper - https://arxiv.org/pdf/1412.0767.pdf

2) Checking whether current preprocessing method, which will work with our setup.
3) Checking violence detection papers and comparing their accuracies.
4) Responsible for partial working implementation of the model.

Jayant Prakash -
Reading different action recognition papers and finding out the feasible architecture.
LRCN, Conv3D, Conv3D & Attention, Two Stream Fusion, TSN, ActionVlad, HiddenTwoStream, I3D, T3D
Will be writing the code to build the different layers of model, then train the model on sports 1M and violence datasets and finally test the implementation of video classification task.

Krushnat More -
1) Reading different action recognition papers, finding respective code, datasets and suitability to the problem at hand, based on model accuracy -
Paper - https://arxiv.org/pdf/1609.08675v1.pdf
Paper - https://arxiv.org/pdf/1706.07960v2.pdf
Paper - https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6165309
2) Will be writing the code to build the different layers of model, then train the model on sports 1M and violence datasets and finally test the implementation of video classification task.
3) Prepared setup - We needed a shared cloud setup for the video classification project, which could hold and process min. 100 GB of data. Using clouderizer and GCP compute at the backend I created the setup for group wide exploration and usage. We are using paid google drive storage to store ~ 100 GB of video data for the training process.

Setup Details - Status (Done)
Clouderizer setup details - https://console.clouderizer.com
Associated GCP/Google Account - deeplearning7643@gmail.com
Project Name - “dl-vcp-1”

Pradumna -
1) Worked and explored aforementioned datasets.
2) Working on data preprocessing also bringing efficiency to the training process.

Help -

We suspect that it might take longer time to read and process data, causing slowness in model training. Currently, we are experimenting efficiently read the data and preprocess it, working with local and clouderizer resources. Depending on compute resources of the server, we will decide how to proceed.

We will not be needing any help as of now, we are thankful to Instructors or TA’s for providing clouderizer and GCP free credits for the purpose of this class.