Violence Detection with 3D Convolutional Networks

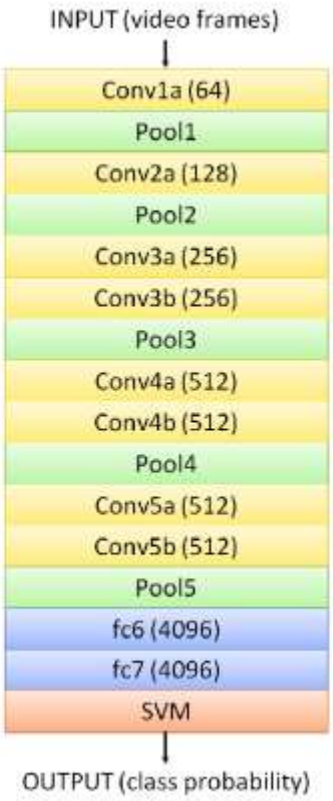
Mustafa Yaldiz (myaldiz3@gatech.edu) Jayant Prakash (jayant2205@gatech.edu) Krushnat More (kmore3@gatech.edu) Pradyumna (pmukunda3@gatech.edu)

Spring 2019 CS 4803 / 7643 Deep Learning: Class Project
Georgia Tech

Abstract

We are trying to classify the content of video whether it is violent or not .We implemented a simple, yet effective approach for spatiotemporal feature learning using deep 3-dimensional convolution networks (3D ConvNets) trained on a large scale supervised video dataset.We used the pre trained weights of sports 1M,then fine tuned the model using UCF101 dataset and again fine tuned and tested the model on Violence detection dataset mainly the Hockey Fight and Movies Dataset . We have achieved 96.5% and 98.6% testing accuracy on Hockey fight and movies dataset respectively on hockey dataset .

C3D architecture



Introduction / Background / Motivation

We are trying to solve the violence detection in videos . Now,people share videos which have violence in it and the company needs to automatically detect the videos and remove it . This project helps in automatically identifying the violence in the video and finally remove them. Now the people use two stream network and this architecture has two separate networks - one for spatial context and one for motion context . However,We have read about T3D ,one of the variant of two stream network, have complicated architecture and require a lot of processing time to train the network .Since , we are training the videos , it takes a lot of time and gpu power to come up with good video detection model. Also, these models doesn't have pre trained weights, so these models need to be trained from scratch and fine tuning is not possible. In this C3D model,we don't need to train the model from scratch and we can finetune the model to adapt to various problems.So, this model can save lot of training time and processing power and also relatively less complex to implement .

Approach

We have implemented C3D models and have weights of Sports 1M trained on this C3D architecture .We have then finetuned the fully connected layer 6 and 7 with UCF101 dataset . Finally, we have trained the final layer SVM with hockey dataset and movies dataset separately and recorded the training and testing accuracy. We have tuned the parameter when we are fine tuning the F6 and F67 layer . It should be successful as it doesn't require training from scratch as we have pre trained weights of sports 1M and thus it requires less training time

, processing power and giving very high accuracy.Earlier,C3D paper has just trained the UCF101 and tested on it only .However, we have gone one step further and trained on violence dataset and tried to find how it works on the violence dataset . We have mainly anticipated the issue related to size of videos and its pre processing can take a lot of time.We have mainly encountered the issue of dataset size ,when we have converted the video dataset in frames, it took around 80 GB of memory,that is our main bottleneck .Also, to pre process the videos of such huge size take a lot of training time .Initial pre processing took a lot of time to convert the videos to frames , then we used ffmpeg to convert videos to frame which relatively speeds up the pre processing time .

Experiments and Results

We have tested the result based on the testing accuracy of two datasets Hockey fight dataset and movie dataset.We have mainly used four datasets Sports 1M, UCF101, Hockey Fight and Movies Dataset .We have experimented with two fully connected layer by training the last two layers with and without pretrained sports 1M weights .The training accuracy and testing accuracy of UCF101 with last softmax layer is 99.27 and 78.53% respectively. Now, before training the violence detection dataset such as movie and hockey dataset , we have removed the softmax and used the SVM instead. The various accuracies of different experiments is mentioned below. We are getting very high testing accuracy of 96.5 and 98.6 % on hockey fight and movies dataset respectively while the training accuracy is 100% for both the datasets .

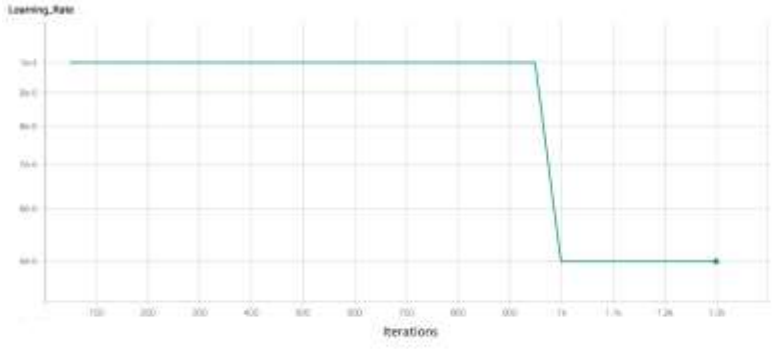
Result

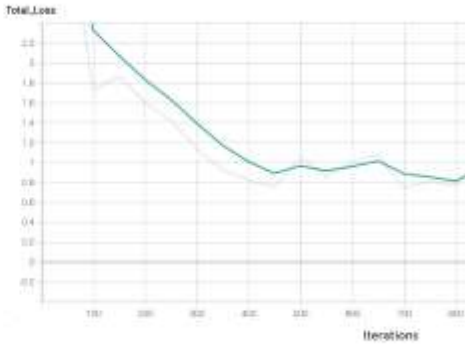
UCF101 Finetuning Results	Accuracy
Training Accuracy - C3D + Softmax	99.27%
Testing Accuracy - C3D + Softmax	78.53%
Paper Testing Accuracy - C3D + Linear SVM	82.30%
Below trained on aggregated whole UCF101 data and showing testing accuracy	
Hockey Fight Dataset	
UCF101 Finetuned, FC Random Initialize + SVM	95.80%
UCF101 Finetuned, FC Sports1M Initialize + SVM	96.50%
Movies Dataset	
UCF101 Finetuned, FC Random Initialize + SVM	98.70%
UCF101 Finetuned, FC Sports1M Initialize + SVM	98.90%
Below trained on aggregated whole UCF101 data and showing training accuracy	
Hockey Fight Dataset	
UCF101 Finetuned, FC Random Initialize + SVM	100.00%
UCF101 Finetuned, FC Sports1M Initialize + SVM	100.00%
Movies Dataset	
UCF101 Finetuned, FC Random Initialize + SVM	100.00%
UCF101 Finetuned, FC Sports1M Initialize + SVM	100.00%
Train Hockey Test Movies	
UCF101 Finetuned, FC Random Initialize + SVM	77.50%
UCF101 Finetuned, FC Sports1M Initialize + SVM	81.00%
Train Movies Test Hockey	
UCF101 Finetuned, FC Random Initialize + SVM	52.50%
UCF101 Finetuned, FC Sports1M Initialize + SVM	57.50%

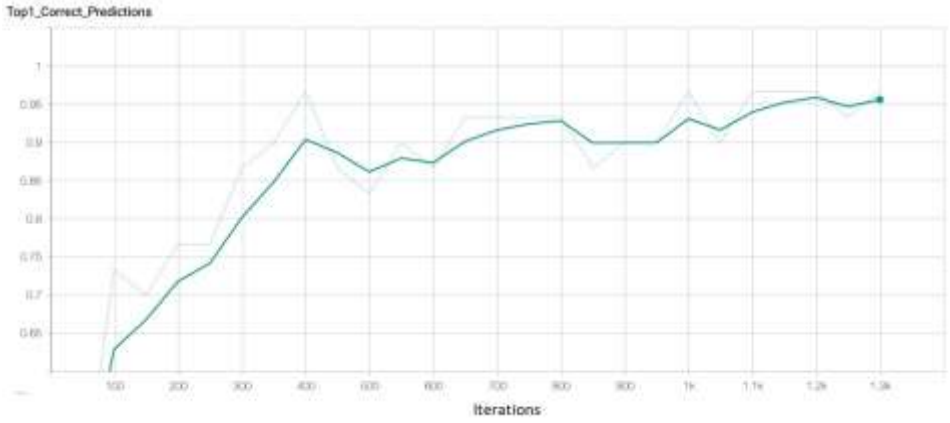
© Mustafa Yaldiz (myaldiz3@gatech.edu) Jayant Prakash (jayant2205@gatech.edu) Krushnat More (kmore3@gatech.edu) Pradhyumna (pmukunda3@gatech.edu)

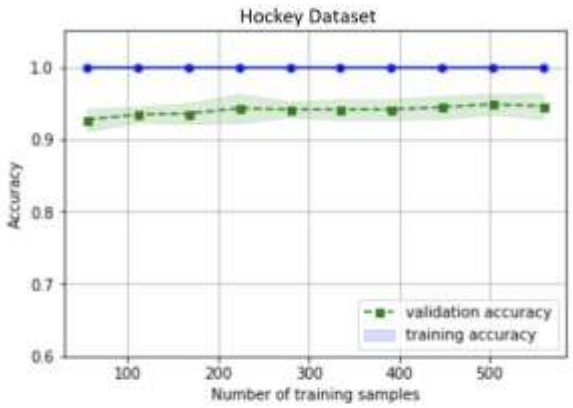
Analysis

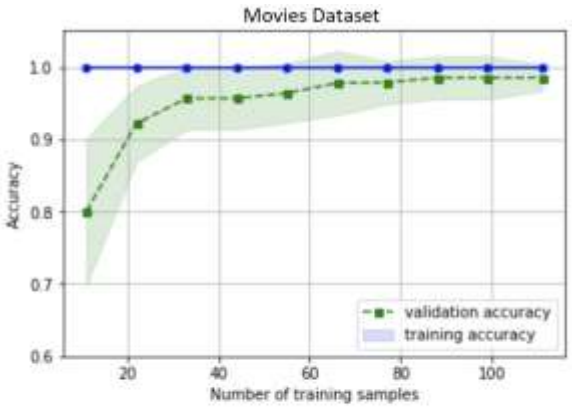
The results make sense and we are getting almost 96.5% and 98.6% testing accuracy on the hockey fight dataset and movies dataset. We have plotted the training loss vs iteration of UCF dataset and training loss decreases with iterations, so the model is getting better and better with iterations. Also,we have plotted iterations with top 1% correct prediction on UCF training dataset and here we have found that training accuracy goes on increasing with iterations which also establishes the claim that model is working perfectly.Finally, we have plotted training and validation accuracy plot with increasing amount of training data for both hockey fight and movies dataset . Both the validation and training accuracy almost converges, so we can say our model is neither overfitting nor under fitting and working perfectly .The original C3D paper have done experiment with UCF + SVM and got 82.3% accuracy but we have tried only UCF101 and SVM and got 78% accuracy which is somewhat less than paper . Also,we have tried training on hockey fight dataset and training on movies dataset, we get testing accuracy of 81% . Also, we have tried testing the hockey dataset trained on hockey dataset and got 57.5% accuracy . So it seems we need to work on the model to generalize for all the dataset.When we are training and testing on same dataset,we are getting testing and training accuracy of about 98.9 and 100% for movies dataset. We think it is not getting generalized on hockey or movie dataset as these are small datasets and we may now need to tune the ucf dataset more to get correct representations.











Team Member Identification

Name	Description of Work
Jayant Prakash	Implemented the C3D models and written code to use the sports 1M pre trained weights. Also read and compared different research paper to come up with the C3D model
Mustafa	Trained the UCF101 dataset on the FC6 and FC7 layer and also hyper tuned the model to have the UCF101 work optimally on the model.Helped in pre processing of videos also.
Krushnat More	Trained the hockey dataset on SVM and also hyper tuned the model to have the hockey dataset work optimally on the model.Also, tested the hockey fight and movies dataset accuracy
Pradyumna	Implemented the pre processing of videos to frames and also applied different techniques like cropping the mean of frames,rotation etc. Also,compared different data sets to be used in the project.

References

Dataset references https://docs.opencv.org/3.0-beta/modules/datasets/doc/datasets/ar_sports.html
https://www.crcv.ucf.edu/data/UCF101/UCF101.rar
http://visilab.etsii.uclm.es/personas/oscar/FightDetection/index.html (Hockey dataset)
http://visilab.etsii.uclm.es/personas/oscar/FightDetection/index.html (Movies dataset).
https://www.dropbox.com/s/zvco2rfufryivqb/conv3d_deepnetA_sport1m_iter_1900000_TF.model?dl=0
(Sports 1M Weights)

[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