COMPARISON OF TEMPORAL SEGMENT NETWORK METHOD AND SLOWFAST METHOD FOR VIDEO ACTION RECOGNITION.

Sai Srinath Alla

Dept of Electrical and Computer Engineering
The University of New Mexico
Albuquerque, NM 87131-0001, USA
ssrinathalla@unm.edu

Abstract:

Human activity recognition has gained a lot of importance in recent years due to its application in various areas such as Surveillance, Health, Security, etc. In this project, I will address the action recognition classification using the Temporal Segment Network method and Slow fast method and compare their performance on the test videos. In the TSN Method we use the segment-based sampling of a single video and in the end, we do the aggregation of segments and In the SlowFast Method we use slow pathway and fast pathway to extract the spatial semantics, temporal information respectively. In this project I will be using the UCF101-24 dataset for training and testing both the models. The implementation of the method had been done using the mmaction2.

Keywords—SlowFast, Temporal Segment Network, Temporal Dimension.

i. Motivation

In this project, I want to work on a video activity recognition problem because, In the Image processing class the previous semester, I learned the basics in Deep learning while working on images for a classification problem and application of the Ensemble technique so, this project will help me learn a lot of advanced concepts in computer vision.

ii. Introduction

Video activity recognition because of its uses in various fields and the success of deep neural networks on the image recognition had got a lot of attention form the scientific community. The important aspects in the action recognition problem are the consideration of temporal dimension and the extract of images or frames from the videos which is relevant to the activity. Temporal segment network method has the inbuilt feature to consider the temporal dimension by having a deep neural network which specifically learns from the flow images. In the case of slowfast method, the work based on the temporal dimension is done by the fast pathway of the network. In this project I used mmaction2 to implement these methods on UCF101-24 dataset.

iii. Background

Temporal Segment Network

Temporal Segment Network operates on the Sequence of short snippets Sampled (k snippets from each video) from the Video instead of Single frame or short frame stack. Each snippet from this video produces its own snippet-level prediction of the action classes, and a following function is asked to aggregate these snippet-level predictions into the video level scores. This video-level score is more reliable and informative than the original snippet-level prediction since it captures the long-range information over the entire video. In this TSN method the function chosen to predict the video

level score should have the ability to effectively aggregate the snippet scores into video level score. From those snippets RGB and flow frames will be extracted, and different deep networks will be trained on those frames. The aggregation process also happens separately for both the cases initially and in the end again there will an aggregation for class score fusion of both rgb consensus and flow consensus (both the streams).

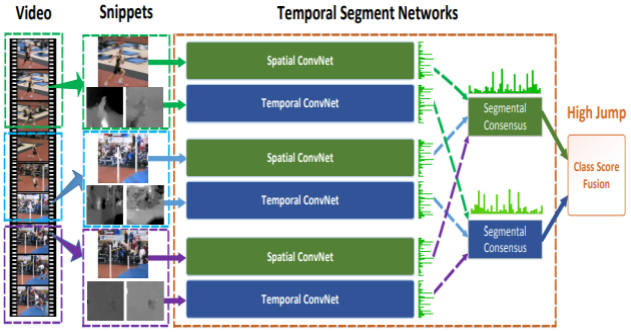

Figure 1. Temporal segment network

SlowFast Method

the slowfast method can be described as a single stream method that operates with different frame rates and lateral connections between the slow frame network and fast frame network. This slow frame network and fast frame network are called slow pathway and fast pathway respectively. The major difference between both networks is the number of frames as input, number of channels. The number of frames in the fast pathway is α times more frames than the number of frames in the slow pathway where $\alpha > 1$. The number of channels in the fast pathway is β times the number of frames in the slow pathway where $\beta < 1$. The high frame rate for the fast pathway helps the network in learning temporal information. Low channel capacity for the fast pathway helps the network in attaining good accuracy and the fast pathway is lightweight. The lateral connection between both the pathways is connected in such a way that the temporal information extracted in the fast pathway will be fused in the slow pathway after undergoing the necessary dimensionality transformation. In the end, a global average pooling is performed on both pathways' output. Then two pooled feature vectors are concatenated as the input to the fully connected classifier layer.

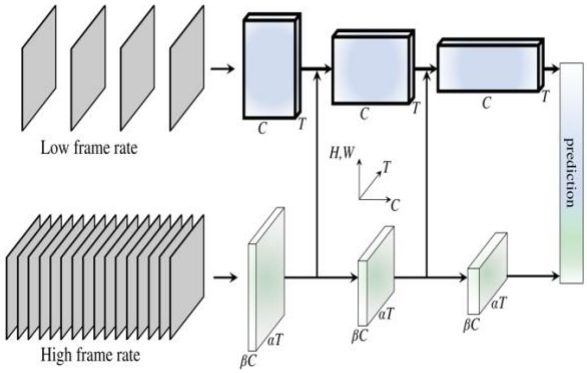

Figure 2. Slowfast Network

iv. Method

Temporal segment network

While working on using the TSN method number of snippets I worked on for each video k=1(i.e., I didn't segment the videos at all). I used TSN from mmaction2 to train and test the data. Using the mmaction2(Dense flow) I extracted RGB images and flow images from all the videos and following their documentation I perfectly created text files for both training and test data samples. I changed the config file for the training of the RGB images (RGB stream) from one of the preexisting configs so that it will suit my job objectives like the Number of classes, learning rate, the path for the dataset. In this method, the main backbone is based on the pre-trained Resnet50. For the second stream, I again changed another config file relating to TSN and flow.

SlowFast Method

Slowfast method only requires RGB images for training, testing, and validation. As we know that this is a single stream method, I just changed the config file in mmaction2 as per the requirements of my problem like learning rate, the number of classes, etc. various factors affect the training process like channel_ratio, speed_ratio in the

config file which I just set them for default values. the main backbone is based on the 3D Resnet50.

v. Dataset and results

Dataset

UCF101-24 is an action recognition data set with 24 action categories, consisting of realistic videos taken from YouTube.

No of categories:

This Dataset consists of 24 Categories relating to various actions. Each category consists of around 133 videos for training and each video is of length around 5 seconds.

- **133** Average Videos per Action Category
- **199** Average Number of Frames per Video
- **320** Average Frames Width per Video
- **240** Average Frames Height per Video
- **26** Average Frames Per Seconds per Video

Performance of TSN-RGB and SlowFast:

The training for the RBG stream of TSN is done using a single GPU so the learning rate used is 0.0016 as per the documentation from the mmaction2. I ran the model for 75 epochs by the end of the training there is a clear improvement in the performance of the model in training accuracy. By the end of the 75th epoch, the top1 training accuracy is 97.66% and the top5 accuracy is 99.83%. By testing the final model on the testing data, the predictions are as following Top1 accuracy – 88.13%, Top5 accuracy – 97.47 % and the mean class accuracy is 88.06 %

The training for slowfast is done using a single GPU so the learning rate used is 0.0125 as per the documentation from the mmaction2. I ran the model for 120 epochs by the end of the training there is a clear improvement in the performance of the model in training accuracy. By the end of the 120th epoch, the top1 training accuracy is 97.66% and the top5 accuracy is 99.83%. By testing the final model on the testing data, the predictions are as following Top1 accuracy – 75.82%, Top5 accuracy – 92.97 %, and the mean class accuracy is 75.50.

Method	Top1 acc	Top5 acc	Class acc
TSN(RBG)	88.13	97.47	88.06
slowfast	75.82	92.47	75.50

Table 1. Accuracy of both the models on the testing data.

vi. Conclusion

On the Ucf101-24 dataset, both the slowfast and the RGB stream of TSN worked well in Activity recognition. The performance of TSN would have been even better if the flow stream of TSN worked without any error, even though the RGB stream of TSN outperformed the slowfast network regarding the accuracy on the same testing set. The performance would have been better if I used more hyperparameter tuning. This is a perfect example problem to run on multi-GPUs for training.

References

1. SlowFast Networks for Video Recognition https://arxiv.org/pdf/1812.03982.pdf

2. Temporal Segment Networks for Action Recognition in Videos https://arxiv.org/pdf/1705.02953v1.pdf

3. Large-scale Video Classification with Convolutional Neural Networks https://www.cv-foundation.org/openaccess/content_cvpr_2014/html/Karpathy_Large-scale_Video_Classification_2014_CVPR_paper.html