Violence Detection Using Dual Stream CNN

Vishnu Prem
Natasha Menon
Shivani Rai
Rahul Rajeev

Abstract

Convolutional neural networks have made huge progress in the field of visual recognition. Action recognition however, is a more challenging machine learning problem as it involves capturing information from multiples frames of a video. This paper discusses two broad convolutional neural network model architectures for action recognition from videos- Single stream method and Dual stream method. Both model architectures have been implemented and compared to determine the better technique. Both models have been trained on a dataset of violent and non-violent videos with the goal of generating a model capable of violence detection from videos

1 Introduction

Violence in videos can negatively influence human behaviour especially when exposed to sensitive audience. Filtering the violent content or blurring it is an effective way of dealing with this. Detecting physical violence in public areas is also important towards maintaining social harmony. To proceed with this, we need to identify the videos that have violence. This project aims to classify the videos into 2 different categories: Violence and Non-violence. We will be implementing 2 different approaches for action recognition and we will be comparing them in the results section. The goal is to find the best technique to classify videos into violence and non-violence or classify any action in general.

In this project we focus on action recognition using single stream and dual stream convolutional neural networks. The dataset consists of equal number of samples in both categories: Violence and Non-violence, and this will be used to train our dual and single stream neural networks. We are contrasting the use of spatiotemporal features with the use of only spatial features used in our binary neural network classifier. This is because the optical flow that is used in the dual stream network computer vision technique which can extract the motion based on consecutive frames in a video. We will be comparing both results that are obtained from both CNNs and also test thier performance using hard negative cases.

2 Related Work

Research in video classification, through action recognition, has increased over the recent years – with many new papers experimenting on various forms of convolutional neural networks. The most basic form implemented so far is to stack consecutive video frames and extend 2D ConvNets into time. On the other hand research like[1] . Advances in Neural Information Processing Systems." use two stream convolutional neural networks approach. This research is closely related to us – the method first decomposes video into spatial and temporal components by using RGB and optical flow fames. These components are fed into separate deep ConvNet architectures, to learn spatial as well as temporal information about the appearance and movement of the objects in a scene. Each stream is performing video recognition on its own and for final classification, they fuse the two networks, at the last convolutional layer (after ReLU) into the spatial stream to convert it into a spatiotemporal stream by using 3D Conv fusion followed by 3D pooling (see Fig. 4, left). After fusion, they let the 3D pooling operate on T spatial feature maps that are frames apart. In addition, the losses of both streams are used for training and during testing they average the predictions of the two streams.
Another research that employs an architecture similar to us is [2] This research uses a multi-response Convolution neural network. The difference of this model architecture with respect to the above is that both streams are processed by identical network as the full frame models. The activations from both streams are concatenated and fed into the first fully connected layer with dense connections.
Since action recognition is relying heavily on the nature of the motion which is extracted from the temporal data, it may be easy for a model to misclassify actions which may seem like physical violence but is infact nonviolent. We will explore such cases and observe the results using both the proposed models.

3 Methodology

In order to get uniformity in the dataset, we first pre-processed each video of both categories in the same way. The processing pipeline performed the followings:

1. Frame Extraction: Ten frames were extracted from the middle second of the video. This is based on the assumption that the middle second has the relevant action

2. Image Cropping: Each frame was converted to an aspect ratio of 1:1 so that any background irrelevant information will not be captured. This was done assuming that the main action would be shown in the center of the frame.

3. Resolution Reduction: Each frame was resized to 100×100, so that the inputs have uniform resolution.

4. Extracting the Optical Flow: The ten frames that were extracted from each video were used to build the optical flow i.e. the temporal information of the video. The direction of action is being stored hence capturing and analyzing the motion features of the data. The optical flow stack is only used with the dual stream neural network.

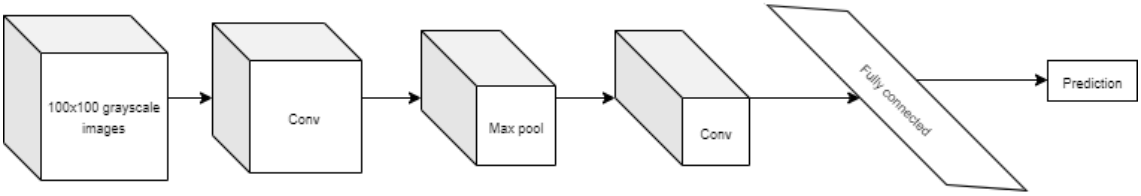

Figure 1: Single Stream Architecture

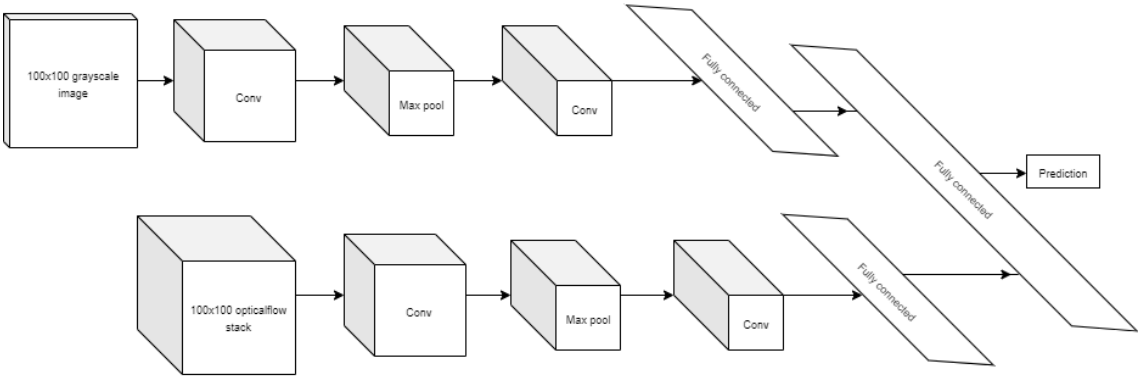

Figure 2: Dual Stream Architecture

After the input data was pre-processed, the data was divided into 3 parts: training(70), validation(15) and test(15) data. The data was used to train two different convolutional neural networks:

1. Single Stream CNN: The frames extracted from the videos and the respective labels are passed into the network as training data, hence using only the spatial information of the dataset. This model is composed of 2 convolutional layers, max pool and fully connected layers.

2. Dual Stream CNN: This network consists of two separate recognition streams: Temporal and Spatial, where the first image of each video is passed into the spatial stream whereas the optical flow stack will be passed into the temporal stream.Both streams have convolutional layers which are later fused

together and passed through a fully connected layer to obtain the final binary classification.

Typically, violence involves rapid, haphazard movement, 2 or more people, a big group,etc. but there are certain other activities that also involve similar settings or movements. For example, when it comes to sports or a dance performance, it may seem to look like violence due to the swift movements and close proximity between people, hence such cases are considered as hard negative cases.

In order to validate how effective our classifier is, we also created a dataset composed only of hard negative cases i.e. 6 second videos of close interaction of the figures but are not violent. This is to determine whether our classifier is able to contrast the violent action from non-violent action, as compared to finding a difference in close proximity figures and figures which that are spaced out in the video. For this we have used a set of 200 videos including a compilation of people dancing (salsa, hip-hop and break dance), teams playing a sport (soccer, football and boxing) and public crowded areas.

4 Results

We have used the 'accuracy' metric in order to compare and formulate our results. The results obtained support our hypothesis that dual stream neural networks fed with optical flow, along with the frames, help recognize certain actions in the videos better than a single stream neural network fed with only frames from the videos.

The below results show the validation accuracies that we obtained from both models, as well as the accuracies obtained from unseen hard negative cases:

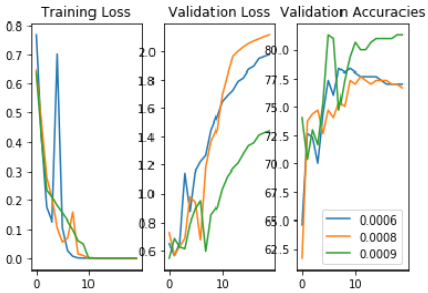

Fig. 1. Single Stream CNN

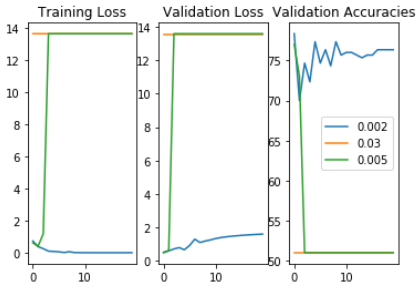

Fig. 2. Dual Stream CNN

Testing Accuracies (Percentage)			
Single Stream Neural Network	Dual Stream Neural Network	Single Stream Neural Network (Hard Negative Cases)	Dual Stream Neural Network (Hard Negative Cases)
79.667	81.333	58.455	62.888

As observed from the results table, the accuracy of the dual stream network outperforms that of the single stream network. This is due to the fact that the

dual stream network has the temporal information that has been extracted which is vital for the action recognition problem. Hence our results are consistent with the literature.
Moreover, the experiment with the hard negative cases resulted in a drop in the accuracy of the model. Our understanding is that the model has learnt to predict videos with rapid motion with people very close by as violent. The inability of the model to identify a context from the video can be pointed out as the reason for the drop in performance.

5 Future Work

This project can be extended by identifying the amount of violence present in a video. It should be able to understand the percentage of violence in an entire video so as to determine the criticality of the violent content. Furthurmore methods to understand the context and setting of a scene could be more useful for identifying violence. Hard negative cases could be classified more accurately if the poses of the people in the video can be extracted and included as a feature to the model. In these ways the performance of the model can be improved considerable, making it more robust for real life applications.

6 Conclusion

The work condcuted clearly cements Dual Stream Neural Network with optical flow inputs as a better classifier for action recogntion problem. Optical flow data helps understand the temporal information that is a vital part of any video and thus the model that learnt from optical flow data, in addition to the frame, gave better results. As suggested in the future work section, improvements can be made but the final model generated at this stage is suitable for adequate violence detection applications in the real world.

7 References

1. Dataset: https://www.kaggle.com/mohamedmustafa/real-life-violence-situations-dataset

2. Karpathy, Andrej Toderici, George Shetty, Sanketh Leung, Thomas Sukthankar, Rahul Li, Fei Fei. (2014). Large-Scale Video Classification with Convolutional Neural Networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

3. Simonyan, Karen Zisserman, Andrew. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. Advances in Neural Information Processing Systems. 1.

4. M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In Human Behavior Understanding, pages 29–39. Springer, 2011

5. T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In Proc. ECCV, pages 25–36, 2004.