Action Recognition in Videos using Two-Stream Convolutional Networks

Saurabh Kumar 2015088 | Kaustav Vats 2016048 | Manish Mahalwal 2016054

Group 15

Abstract

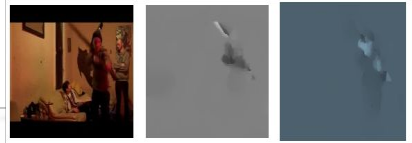
Action recognition task involves the identification of different actions from videos where the action may or may not be performed throughout the entire duration of the video.
Classifying actions (jumping, picking, playing etc) in still images has been solved with a high accuracy using Deep Convolutional Networks. However, for action recognition in videos, there exists only limited work with poor accuracy. Action recognition from videos requires capturing context from the whole video rather than just capturing information from each frame as compared to an image.

Literature Review

We aim to extend the existing model of Deep Convolutional Neural Network used for classifying action in still images. Relevant work already exists for the same which uses stacked frames for input to CNN, but the results are poor. So, to improve upon this architecture, we implement a new architecture for action recognition in videos which is based on the human visual system. This architecture is based on the human visual cortex system which uses two-streams to detect and recognise motion:

1. **Ventral Stream**: used for performing object recognition
2. **Dorsal Stream:** used for recognising motion

The author's architecture involves two separate recognition streams - **spatial and temporal**. Spatial Stream performs action recognition in still frames which extracted from the video. Temporal Stream recognises action from motion using dense optical flow.



(a) Video frame (b) Vertical component of optical flow (c) Horizontal component of optical flow

Architecture

We can decompose a video into two components i.e. spatial and temporal. Each of the two streams are implemented using deep CNN:

1) Spatial Stream ConvNet: Spatial Stream carries information about the objects and scene in the video-in RGB format. This network operates on individual frames extracted from the video at some sampling rate.

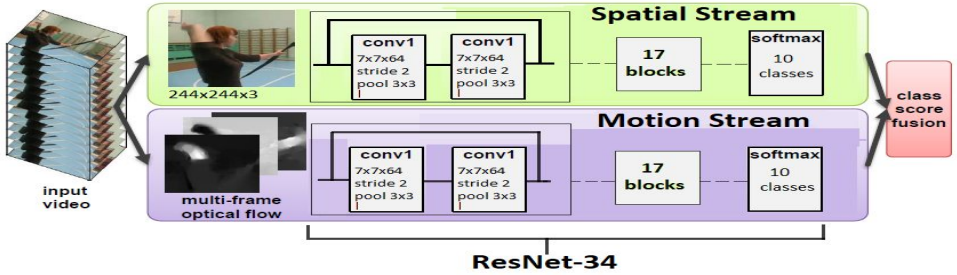
2) Temporal Stream ConvNet: Temporal Stream carries information about the movement of the of the object(or the camera) in the scene (refer Fig. 2) i.e. the direction of the object. The input fed to this model, is formed by stacking multiple optical flow displacement images between consecutive video frames.

Dataset

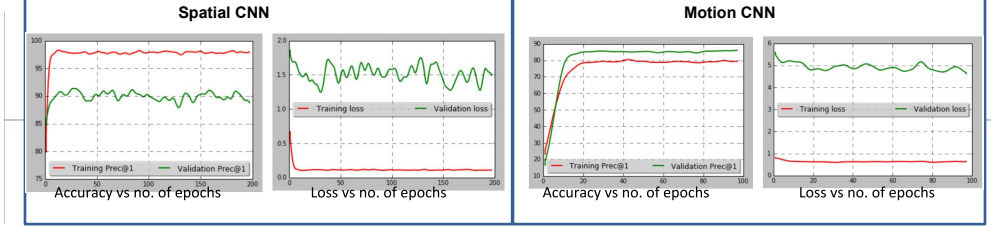
We are using the famous dataset called UCF101 for our project. UCF101 is a dataset of human actions.We extracted 1. UCF101
have 101 classes which are divided into five types:

1) Human-Object Interaction
2) Body-Motion Only
3) Human-Human Interaction
4) Playing Musical Instruments
5) Sports

The dataset consists of realistic user-uploaded videos containing cluttered background and camera motion. It gives the largest diversity in terms of actions and with the presence of large variations in camera motion, cluttered background, illumination conditions.



Results



Approach

The motivation is to capture both, the spatial feature of the video as well as Temporal ones. For example, To classify either the action being performed by a person in the video is archery or playing basketball, Our spatial network captures the still-frame information about the action being performed. Basically it's doing classification task on frame of video. Temporal network tries to distinguish the action using motion in the video. For this motion we are using segmentation based optical flow in both, horizontal and vertical directions. Both the models uses ResNet34 as the underlying network.

Results and Conclusions

Table 1. Accuracies on training and validation set.

	Spatial CNN		Motion CNN		After Fusion
	Training	Validation	Training	Validation	
Loss	0.10	1.20	0.63	4.67	---
Top-1 Accuracy	98.47	88.94	79.28	86.37	92.80
Top-5 Accuracy	99.89	98.71	97.96	97.42	98.97

As mentioned in the paper, our results also demonstrate that simple classification work is also quite impressive when in comes to action detection in videos. But some temporal information can't be caught from still Images and hence concatenating results from both the networks is giving us higher accuracies over the dataset.

References

Two-Stream Convolutional Networks for Action Recognition in Videos : Karen Simonyan, Andrew Zisserman
https://github.com/feichtenhofer/twostreamfusion
https://github.com/jeffreyhuang1/two-stream-action-recognition
https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py