Action Recognition on UCF 101 Dataset using C3D

Ryan Chui
Rochester Institute of Technology
One Lomb Memorial Dr,
Rochester, NY 14623-5603
Rc4307@rit.edu

Vigneshwar Jayakumar
Rochester Institute of Technology
One Lomb Memorial Dr,
Rochester, NY 14623-5603
Vj1181@rit.edu

Abstract

Image processing and pattern recognition in images are some of the most interesting and researched topics in Deep learning. Integrating these two concepts for action recognition into video understanding is useful in a variety of applications, including identification of different actions from video where an action from the video may or may not be performed throughout the entire duration. To get an understanding of the presented UCF-101 action recognition dataset which consists of 101 action classes and 13320+ videos, it became the largest and diversified dataset to study in terms of actions and variations from camera motion, object appearance, pose, and scale. We extracted video features using the proposed Convolutional 3D Network (C3D) to train, test, and fine-tune 3D convolution network against the UCF101 dataset.. We also examined performance-related causes and identify heatmaps for future potential research usage at various frames.

1. Introduction

When human characterize objects, we cannot perfectly identify objects, and monitor every condition 24/7/365 days. Using computer vision and AI system, the attributes of causing human error due to fatigue, subjective decision-making and distractions can be minimized. In many applications, computer vision can be used to help monitor safety condition with better quality control. Additionally, human cannot label an action perfectly within every second. A standard categorial label with attributes that describe objects and the state in which and when they appear can be carried via categorical object recognition and segmentation. In this paper, we will explore the C3D action recognition paper and we will use the UCF-101 video recognition dataset that is composed of realistic action videos from YouTube. The goal of this paper is to understand the C3D network and in general using 3D convolution for action recognition and video semantic analysis. The videos contain changes in continuous background noises, and as a result, detecting real moving objects and motions can become extremely difficult. As most of the available action recognition data sets are not realistic, UCF101 aims to encourage further research into action recognition by learning and exploring new realistic action categories.

2. Related Work

The image recognition is being studied for at least 2 decades now in the computer vision field. But analyzing videos and recognizing actions based on videos is relatively new field. Space-time Interest Points [01] this paper explored an idea of bringing the special interest points used in early computer vision to temporal domain. As the deep neural network architecture progressed with increasing GPU power [02] papers explore using Deep neural networks to video classification. But these papers used the 2D convolutional networks at large except first few layers. This resulted in loss of temporal data early in the network. The above discussed papers are the related work for C3D paper.

Here in this paper, we try to understand the C3D convolutional layers. Visualizing and Understanding Convolutional Networks [03] discuss various ways to visualize the convolutional network with heatmap to understand the segmentation. We explore the use of heat maps to understand how each frame in 3D Convolutional neural network is analyzed.

In recent years, several attribute datasets are used and studied by researchers for different computer vision applications. Below are some of the most relevant ones.

- UCF-101 which we referred and used for the findings.
- HMDB51 dataset is an action recognition dataset which contain videos from web videos and movies. This dataset is composed of 51 action categories (“chew”, “dribble”, “hug” etc.) with 6849 video clips in total.

3. Methodology

3.1. Base Dataset

The UCF 101 dataset composes of 101 categories and is classified into 5 types (Body motion, Human-human interactions, Human-object interactions, Playing musical instruments and Sports). The objects are without editing or post-processing and are not annotated with classification labels or bounding boxes in advance. The objects and actions categories are filmed from realistic videos taken by YouTube. The success of localizing the predicted actions from each video will provide a more general understanding, which will have greater applications in surveillance and human-computer interaction.

3.2. 3D Convolutional Neural Network

After downloading the dataset and split the data into training, testing, and validation, we implemented a convolutional 3D Network (C3D) to extract spatial information and temporal features using the input 16 RGB frames. In our work, we extend the forementioned 3D-CNN works in the following aspects.

First, the network architecture has 5 pooling layers, 2 fully connected layers, and a SoftMax loss layer. The channel size for each convolution block from the first block to the fifth block was 64, 128, 256, 512, and 512. The network employs an 3x3x3 input as the first convolutional layer, followed by a pooling layer of kernel size 1x2x2. The stride size of 3 is reduced to size 2 for the second to fifth pooling layers, because this can preserve from using only a single level spatial feature abstraction and can reduce memory footprint and allow bigger batches to detect smaller objects in motion. The last pooling layer will have highest activation after deconvolution. We then utilize the Relu activation function and normalize it through a SoftMax function, and the fully connected layers of size 4096 dimensions will follow by a SoftMax classifier to obtain the prediction result. Using C3D method on this project can significantly focus on appearance and motion at different occasions of a video clip.

3.3. How C3D Learns

To understand how the c3D network learns, we use a heatmap that is created with the values at the end of the CONV5D convolutional layer. This heatmap is seen at different frames to see what is being segmented. The C3D networks starts focusing on the appearance at the initial frames to find the key features and segment according to them. Once this is done in the next frames this network starts detecting movement patterns and maps the movement patterns accordingly.

The figure 1 and figure 2 gives a clear picture of how the movement is tracked. This is the temporal information advantage compared to 2D CNN. From this we could understand that C3D network uses both the movement and appearance to bring a better conclusion on the recognition tasks.

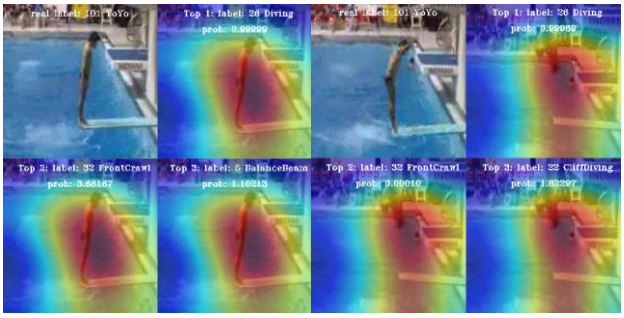

Figure1: Initial frame of diving video and a middle frame from jumping with original image showing one frame behind to get the idea of what is being tracked.

In the initial frame, As there is little to no movement the C3D network has a higher confidence for balancing beam as well but as the video progresses the heatmap shows us how the C3D network tracks its features and finds it as similar to high jump or cliff diving. This clearly shows the recognition is done by tracking the movement pattern as well.

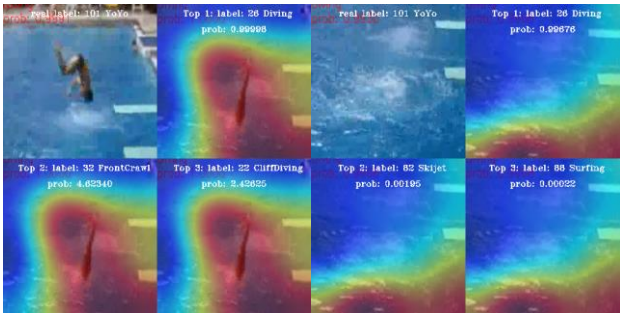

Fig 2. The probability of the diving does not change at all in all three figures

4. Experimental results

We are able to obtain 98.7% accuracy of test loss 4.5% on the epoch of 19. When we trained with multiple epoch values we found that any epoch above 16 performs well on the given dataset.

```
Device being used: cuda:0
F:\project models C3D-ucf101_epoch-19.pth.tar
Initializing weights from: F:\project\models\C3D-ucf101_epoch-19.pth.tar...
Total params: 78.41M
F:\project
True .\ucf101
Number of test videos: 2701

2701it [11:10,  4.03it/s]

[test] Loss: 0.045740264844880994 Acc: 0.9870418363569049
Execution time: 670.4697442999986
```

The performances for all action classes in UCF101, and these confusion matrixes are well diagonalized in figure below. Some categories are easy to recognize, but some

are hard to distinguish. For example, when compared with the data set, we get easily recognizable categories such as “Bench-press” and “Band Marching”, but it is hard to recognize “Apply Lipstick” category ,since videos in “Apply Lipstick” can be confused with “Apply Eye Makeup” and “Brushing Teeth”, as well as “Shaving Beard” as opposed to “Brushing Teeth”. This may be due to the set of classes from the frames contain some commonality in the main subject or machine simply learns small object per second without any furthermore accurate texture after deconvolution methods applied.

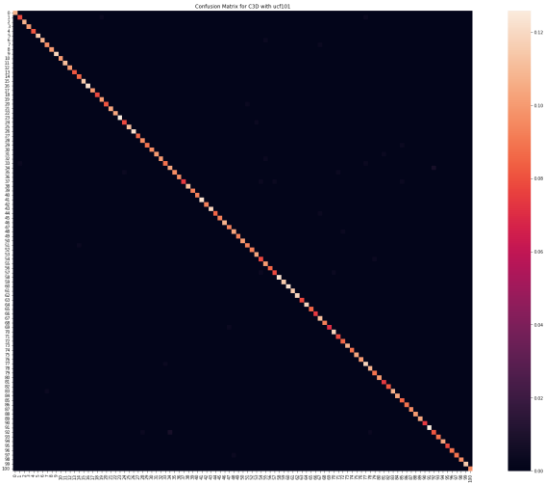

Figure 3: Confusion matrix for the 101 classification not named because of the short space.

Notice that our test set with 2701 videos and 101 classes is considerably small, the estimate performance is going to have high variance. Nonetheless, the proposed C3D model performs quite well in almost all categories.

Figure 5: showing a detection of bike

This last is done for run time analysis. We were getting around 400 fps in a GPU based machine in realtime.

5. Conclusion

From this paper we were able to study about the general structure of 3D convolutional networks and the advantages of using 3D CNN over 2D CNN on action recognition and video in general. We studied how to create Heatmaps for convolutional networks and We also studied how the Convolutional Layers work by viewing the heatmaps at various frames.

6. Appendix

Ryan Chui worked on the understanding of Data loader functions and explaining data loader and Vigneshwar Jayakumar worked on writing the code and visualization of confusion matrix and heatmaps. Both of us worked on the preparation of the report where Vigneshwar Jayakumar worked on the how C3D network learns and some of related work and Ryan Chui worked on introduction, 3D CNN network explanation and experimental results.

References

[1] Ivan Laptev and Tony Lindeberg - Space-time Interest Points 2004 in IEEE. https://www.irisa.fr/vista/Papers/2003_iccv_laptev.pdf

[2] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, George Toderici - Beyond Short Snippets: Deep Networks for Video Classification Mar 2015. arxiv.org/abs/1503.08909

[3] Matthew D. Zeiler and Rob Fergus,Visualizing and Understanding Convolutional Networks

[4] Wu, X.; Ji, Q. TBRNet: Two-Stream BiLSTM Residual Network for Video Action Recognition. *Algorithms* **2020**, *13*, 169. https://doi.org/10.3390/a13070169

[5]. Simonyan, Karen, and Andrew Zisserman. “Two-Stream Convolutional Networks for Action Recognition in Videos.” *ArXiv.org*, 12 Nov. 2014, https://arxiv.org/abs/1406.2199.

[6] Liu, K., Liu, W., Gan, C., Tan, M., & Ma, H. (2018). T-C3D: Temporal Convolutional 3D Network for Real-Time Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/12333