Digital Video: Perception and Algorithms

Human Activity Recognition

Himanshu Singh
himanshusing@iisc.ac.in

Abstract

Human Activity Recognition is an important and well studied topic in the field of image and video processing, with many applications in surveillance (security,sports, etc.), activity detection, video-content-based monitoring, man-machine interaction, and health/disability care. Here, we explored some architectures that have been used to perform action recognition on both still images and videos. For still images, we used several well known architectures to fine tune on the Stanford40 dataset and achieved 83.25% accuracy with the ResNet-50 model. For videos we evaluated a 3D CNN, and a CNN with LSTM on a UCF101 dataset. The 3D CNN on the UCF101 dataset gave us the best results at 57% accuracy. The CNN with LSTM gave better results with an accuracy of 91.35%.

1 Introduction

Action recognition can be defined as the ability to automatically recognize a specific activity in a video stream. With its huge application spectrum, human action recognition (HAR) is proven to be an important part of computer vision research.The advances in both computing power and research in Deep Learning, these applications have made many breakthroughs in recent years. Recently, Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) networks have shown great success in image/video classification and voice recognition. Our work here aims to evaluate a number of neural networks to see how well they can perform HAR on the Stanford40 dataset for images, and the UFC101 dataset for videos. Given an image, or a video, we will evaluate whether the models can correctly classify the activity.

2 Related Work

Classic models are based on handcrafted features and typically use local ones. There are generally two main classical approaches. A number of classic methods use holistic representations, as a global representation of the human body and its movements while some methods use local features. Space-Time Volumes (STVs), which are among the global representation methods.

Deep learning models can extract more complicated concepts from videos by considering frame sequences. Many efforts have been made to involve the time information available in the video into CNN models. Deep-learning action recognition methods can be grouped in two categories: two-stream networks and spacetime networks. Most existing deep-learning HAR methods use two-stream networks, one stream for spatial information and the other for inter-frame motion information. Three dimensional CNNs are examples of space-time networks. Other space-time networks such as Recurrent Neural Networks (RNN), such as LSTM, have also been used to incorporate temporal information into the video.

A number of 2D CNN architectures are available for image classification tasks with weights pretrained on ImageNet. These include the architectures we used: **VGG16**[1],

ResNet-50[2], **MobileNet-v2**[3], **DenseNet-121**[4] and **ResNet-152**[2] which we leverage to fine-tune on the Stanford40 dataset[5] and UCF101 dataset[6] for action classes.

3 Dataset

For still image HAR, the Stanford40 Action dataset[5] was used which contains 40 different human activity classes with 180-300 images per action class with a total of 9532 images. Action categories include "brushing teeth", "playing guitar", "jumping", among others. In dataset train-test split is divided as 100 training images per class and rest belongs to test, but due to relatively small size of the dataset, we decided to go with at 70-30 train-test split per class. Images were rescaled to 224×224 pixels. For data augmentation, horizontal flipping, a rotation range of 10 degrees was applied to training dataset.

For video HAR,the UCF101 dataset[6] was used which is one of the most popular action recognition datasets, with real-world videos. This dataset contains 13320 videos taken from YouTube divided into 101 classes, including "Apply Eye Makeup", "Bench Press", "Trampoline Jumping", among others. The UCF101 dataset contains a wide range of actions from the five main categories: human object interaction, body movement, human-to-human interaction, playing musical instruments, and sports. Each category contains 100 to 200 videos. The shortest video contains 28 frames and the dimensions of each frame is 320×240. We first need to pre-process the videos by extracting the input video frames before inputting them to the proposed model. Libraries like OpenCV or FFmpeg can be used to extract video frames. We directly imported the pre-processed dataset used in [7][8].

4 Methods

4.1 Image HAR

For still images, we used transfer learning on VGG16[1], ResNet-50[2], MobileNet-v2[3], DenseNet-121[4] pre-trained on ImageNet weights, replacing the final layer with a 40-way softmax output layer. We used cross Entropy as our loss function, usually the loss is averaged over the batch, instead we summed over the batch for backpropagation. The models were optimized with Adam optimizer. The learning rate was varying across epochs:

Epochs	Learning Rate
1-100	10^{-4}
101-200	5×10^{-5}
201-300	10^{-5}
301-400	5×10^{-6}
401-500	10^{-6}

4.2 Video HAR

For videos we used two models.

First model is a 3D convolutional network. The shortest video in the entire dataset was of 28 frames, so we used starting first 28 frames of all the videos to predict the output. The network consist of 2 3D conv-layer with batch normalization after which ReLU activation and max-pool is applied. Then finally a full connected network of 256,

256, 101 layers are added. The loss function used was cross entropy, to optimize the model we used Adam optimize with learning rate of 10^{-5}. We also tried using dropout of neuron but it didn't improve the accuracy.

Second model consists of two parts, which is based on [9]. In the first part, we use a CNN network for categorizing images and videos. The desired features are extracted from each video frame by the CNN network. We use a pre-trained CNN network named ResNet-152. Every frame of the video was converted to a 512 dimension feature vector. After extracting feature vector from every frame of video, we used a Recuurent neural network for detecting complex and sequential patterns hidden in video frames across temporal dimension. For this we use a 3-layered LSTM. We use the last output of the LSTM as out final feaure vector. This feature vector is fed to a Fully connected network with 256, 256, 101 neurons. Here also we used cross entropy as loss function with Adam optimizer at learning rate of 10^{-4}.

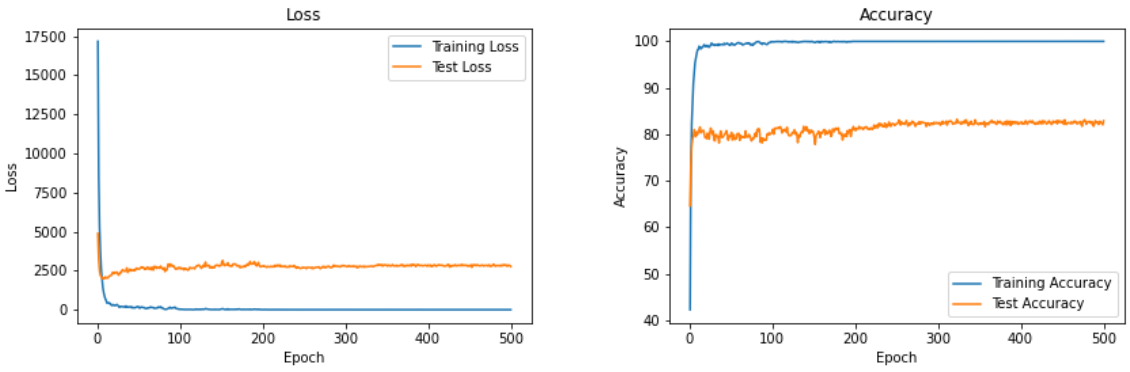
5 Results

For Image HAR our finding are as follows:

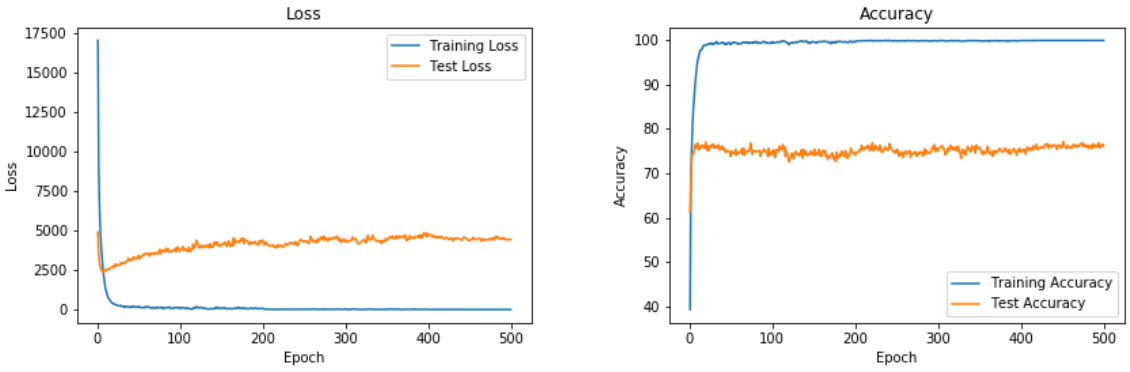
Methods	Accuracy (%)
MobleNet-V2	77.04
VGG-16	77.54
ResNet-50	82.32
DenseNet-50	83.25

From our observations, the MobileNet-v2 model gave the best bang for the buck which has significantly faster training times and a decent accuracy. The loss and accuracy per epoch for the models are as follows:

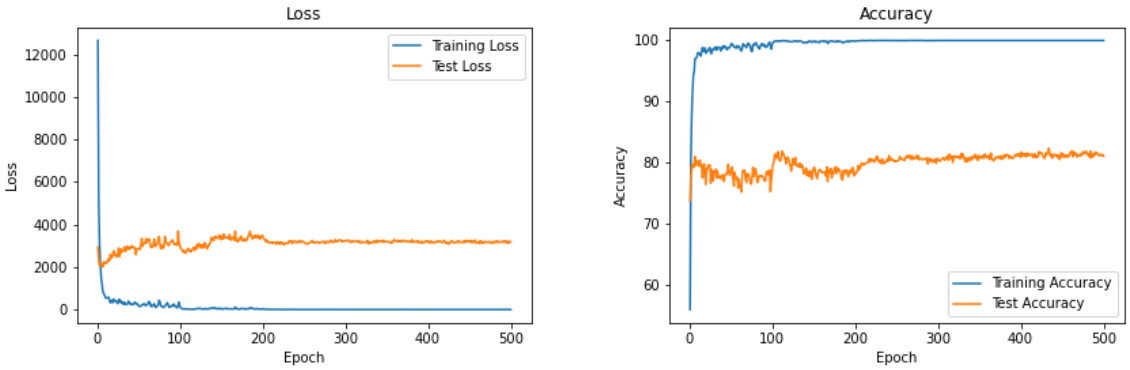
DenseNet-121



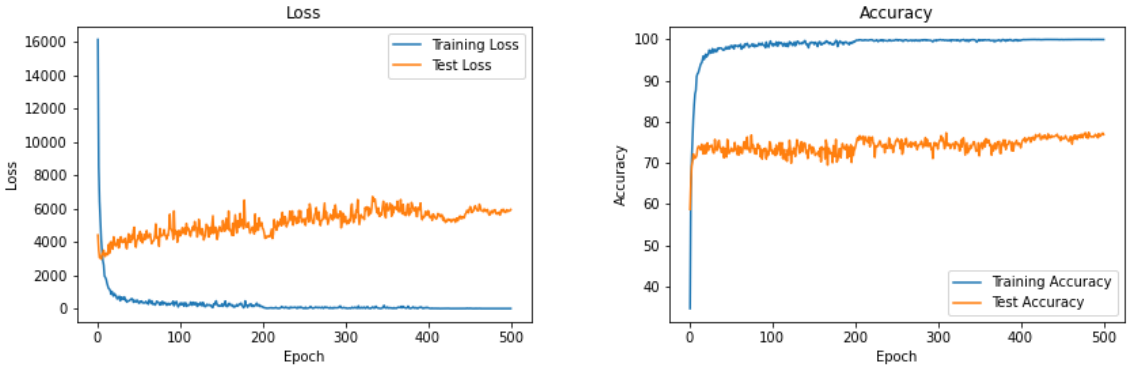
MobileNet-V2



ResNet-50



VGG-16

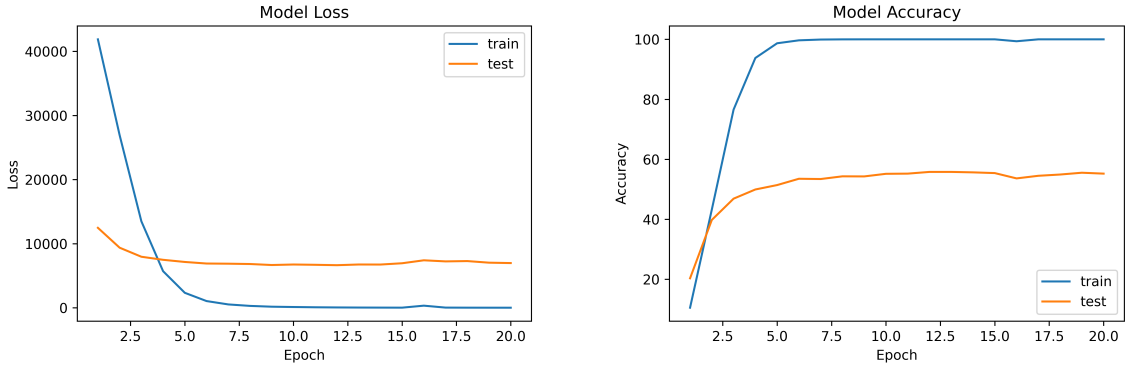


For Video HAR, our finding are as follows:

Methods	Accuracy (%)
3D-CNN	56.36
ResNet-LSTM	91.35

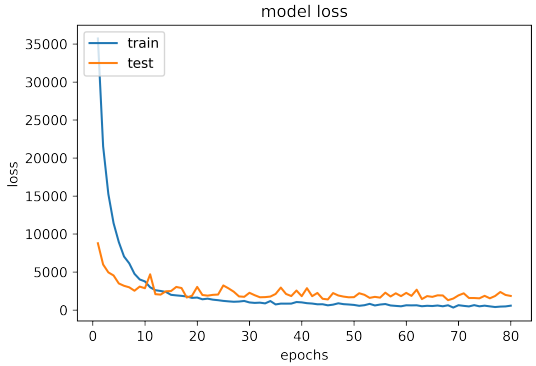
The Resnet-152 with 3-layered LSTM does a good job with relating spatial features with temporal features and achieving a great accuracy, but at an expense of epoch taking 15 min to train. The loss accuracy for Video HAR models are as follows:

3D-CNN



Since the model approached near to 100% accuracy on train dataset, so we trained for only 20 epochs

ResNet-LSTM



References

[1] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, jun 2016, pp. 770–778, IEEE Computer Society.

[3] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2018, cite arxiv:1801.04381.

[4] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[5] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lin, Leonidas Guibas, and Fei Fei Li, "Human action recognition by learning bases of action attributes and parts," 11 2011, pp. 1331–1338.

[6] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[7] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Christoph Feichtenhofer, "twostreamfusion," `https://github.com/feichtenhofer/twostreamfusion`, 2016.

[9] A. Mihanpour, M. J. Rashti, and S. E. Alavi, "Human action recognition in video using db-lstm and resnet," in *2020 6th International Conference on Web Research (ICWR)*, 2020, pp. 133–138.