Human Activity Recognition using Deep Learning

Akhilesh Nandwal
Department of Computer Science & Engineering
National Institute of Technology Delhi, Delhi, India
181210006@nitdelhi.ac.in

Amrit Raj
Department of Computer Science & Engineering
National Institute of Technology Delhi, Delhi, India
181210008@nitdelhi.ac.in

Ankit Rouniyar
Department of Computer Science & Engineering
National Institute of Technology Delhi, Delhi, India
181210011@nitdelhi.ac.in

Samyak Prajapati
Department of Computer Science & Engineering
National Institute of Technology Delhi, Delhi, India
181210046@nitdelhi.ac.in

ABSTRACT

The uses of technology can be over a wide range as exhibited by the boom of atomic energy, which gave rise to nuclear fission bombs and clean energy generation techniques. The coupling of real-time video feeds with convolutional neural networks would give rise to multiple interdisciplinary applications ranging from general statistics to security-based surveillance.

In this work, we have trained CNNs like ResNet50, ResNet101, InceptionV3 and InceptionResNetV2 on human activity recognition datasets and achieved a mean top-1 accuracy of 78.64% on the imagery and a mean top-1 accuracy of 48.33% on video data. An end-to-end application was also created as a proof of concept for the viability of this technology.

1. INTRODUCTION

In the current age, the products of the 4th Industrial Revolution are establishing their prevalence in our daily lives and technology has advanced to such a level that going "off-grid" is no longer a viable option. The boom in technology is directly correlated with the boom in the economical position of a nation, and while it has proven apt in ameliorating the quality of life, the general trend is leading us to an over-reliance on technology. This dependence has several pros and cons associated with it, where it all depends on us humans, on how we decide to make use of it.

Mobile phones and laptops have now become commonplace items that are at arm's reach for most of us. Data from such sources can prove valuable in establishing a security-critical surveillance system as proven in the 2013 Boston Marathon Bombings [1] where videos recordings from mobile phones used by citizens aided the investigators in determining the cause of the explosion.

With the given abundance of CCTV cameras in nearly every public location, a system designed for activity recognition could prove invaluable in circumventing illegal activities. Such systems could be used for recognizing abnormal and suspicious activities at crowded public locations and aid the on-ground personnel in flagging an individual as needed.

2. RELATED WORKS

The works of Mohammadi et al. [2] built their results on CNNs which were pre-trained on "ImageNet" [3] weights and made use of attention mechanism with an average classification accuracy of 76.83% across 8 models. They were also involved in the creation of ensemble models with 4 models that yielded the highest accuracies and achieved an action classification accuracy of 92.67%.

3. DATA SOURCE

The Stanford 40 Action Classification Dataset [4] was used in this work for training the images. It contains 9532 images across 40 action classes with around 180-300 images dedicated for each action class. Subsequently, a custom dataset [5] was created which embodies three YouTube URLs for each action class present in the Stanford 40 dataset. Each URL is a copy-right free "stock" video, with the video length ranging from 15-30 seconds.

4. METHODOLOGY

The images were first augmented with random rotations between 0 and 359 degrees followed by resizing them to 256x256 pixels. The augmented images were then used to train four CNNs, namely ResNet50, ResNet101, InceptionV3 and InceptionResNetV2 using Keras. The models were initialized with "ImageNet" weights and Stochastic Gradient Descent (SGD) was chosen as the optimizer. The metrics were further improved by different combinations of regularization layers, dropout layers and by hyperparameter tuning.

To introduce the modality of classification by the use of videos, the trained models were tested by decomposing the videos into individual frames, and then each frame was tested by each model and the predicted class with the highest frequency was chosen as the class exhibited in the video.

A browser-based end-to-end deployment was also created using Streamlit in order to have a visualizable experience for the end-user of the product. The deployment features the ability to choose multiple models for the purpose of detection of action classes.

5. PERFORMANCE METRICS

The metrics chosen for model evaluation were chosen as Top 1 Accuracy, Precision, Recall and AUROC (Area under ROC Curve).

$$accuracy = \frac{True\ Positive + False\ Negatives}{Total\ Samples}$$

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

6. RESULTS

The metrics achieved after training and testing on Stanford-40 imagery and corresponding videos are tabulated in Table 1 and Table 2 respectively. The accuracy mentioned henceforth refers to the Top-1 accuracy.

Table 1: Metrics on Stanford-40 Imagery

Model	Accuracy	Precision	Recall	AUC
ResNet50	77.55%	0.81	0.75	0.96
ResNet101	80.41%	0.84	0.78	0.97
Inception V3	79.16%	0.82	0.77	0.96
Inception ResNetV2	77.46%	0.85	0.71	0.98

Table 2: Metrics on Stanford-40 Videos

Model	Accuracy	Precision	Recall	AUC
ResNet50	47.50%	0.47	0.47	0.73
ResNet101	54.16%	0.54	0.54	0.76
Inception V3	42.50%	0.42	0.42	0.70
Inception ResNetV2	49.16%	0.49	0.49	0.73

7. DISCUSSIONS

With the availability of computational equipment which enables us to perform such computations in real-time, the possibility of such systems are endless. Keeping the current COVID-19 pandemic in mind, computer vision is a field that has progressed incredibly in the span of a few months. Our activity recognition model could easily be trained to differentiate between “Wearing a mask properly”, “Wearing a mask improperly” and “Not wearing a mask” and can then be used to flag down violators. This technology, coupled with some hardware could also be used to create an access control system where only specific categories could be allowed access, such as in a construction site, where many workers tend to skimp off on wearing necessary protective gear.

The models could further be improved upon by training further with optimized hyperparameters and by using multiple datasets, such as the Sports-1M Dataset [6], which consists of almost one million videos for around 487 sporting activities and UCF101 Dataset [7], which consists of 13,320 videos for various common actions.

A weighted ensembled model could also be created with multiple other models for improving overall accuracy in the classification of action categories.

8. REFERENCES

1. Hunt for Boston bomber in iPhone era. (2013, April 18). Financial Times. https://www.ft.com/content/48adc938-a781-11e2-bfcd-00144feabdc0
2. S. Mohammadi, S. G. Majelan and S. B. Shokouhi, "Ensembles of Deep Neural Networks for Action Recognition in Still Images," 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 2019, pp. 315-318, doi: 10.1109/ICCKE48569.2019.8965014.
3. J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.
4. B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. Internation Conference on Computer Vision (ICCV), Barcelona, Spain. November 6-13, 2011.
5. Prajapati, S. (2021, April). djsamyak/DM-Stanford40. GitHub. https://github.com/djsamyak/DM-Stanford40
6. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, & Li Fei-Fei (2014). Large-scale Video Classification with Convolutional Neural Networks. In CVPR.
7. Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, CRCV-TR-12-01, November, 2012.