Video Classification Using MoViNet

Mike Bermero[1], and Jean Camacho[1]
and Leo Ramos[1,2]

[1]Yachay Tech University, San Miguel de Urcuquí, Ecuador
[2]Deep Learning for Autonomous Driving, Robotics, and Computer Vision Research Group

Comp. Vision Task

Video classification using TensorFlow lite models over an android application that works in real-time:

- Video classification using MoViNet-A0, MoViNet-A1, and MoViNet-A2
- Models trained with the UCF101 dataset, this dataset is centered on human actions compilation of multiple nature.

Introduction

Video classification is a computer vision task that involves the machine's capacity to extract features, process, and recognize information from a video source [1]. It has many relevant applications, such as self-driving, augmented reality, robotics, and movement recognition [2].

Initially, video recognition was achieved by applying classical image-processing techniques [3] (i.e., object representation with points, geometric shapes, skeletal model, probability density, templates, etc.). The rise of neural networks has facilitated the development of more efficient and effective systems that address this task. Based on this, we re-trained the MoViNet video recognition architecture using transfer learning to fine-tune multiple versions of this architecture in the UCF-101 dataset.

Architecture

The architecture we use in this study is MoViNet. The architecture's most essential property is having streaming models with 3D causal convolutions and temporal ensembles, standing out from most state-of-the-art architectures.

Also, the evaluation of the videos is made frame by frame, in contrast to the traditional multi-clip evaluation approaches, which helps to improve speed and accuracy [2, 4].

Likewise, its architectural composition reduces memory usage, performs less redundant computations, and boosts efficiency.

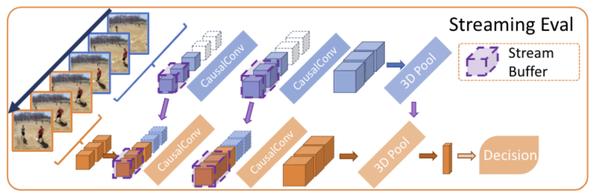

Figure 1:MoViNet structure using streaming evaluation on it architecture

Data set Description

For model retraining, we used the UCF101-action recognition data set. This data set consists of 13,320 video clips collected from YouTube. The data set contains 101 categories, which can be classified into five types: body motion, human-human interaction, human-object interaction, playing musical instruments, and sports. In total, the videos are over 27 hours long and have a fixed frame rate of 25 FPS with a resolution of 320 × 240.

Figure 2:Sample images taken from the data set used

Implementation and Training

We decided to train three versions of the MoViNet architecture. The details of these are shown in Table 1. The models were trained under TensorFlow version 2.12 and using Google Colab. Also, to deploy the models and use them on cell phones, we used Android Studio in the Dolphin version. Likewise, the loss function and optimizer used are cross entropy and RMSprop, respectively. The training was carried out during 5 epochs.

Table 1:MoViNet versions used in this study.

Model	Input shape	Params
MoViNet-A0	50x172x172	3.1M
MoViNet-A1	50x172x172	4.6M
MoViNet-A2	50x224x224	4.8M

Results

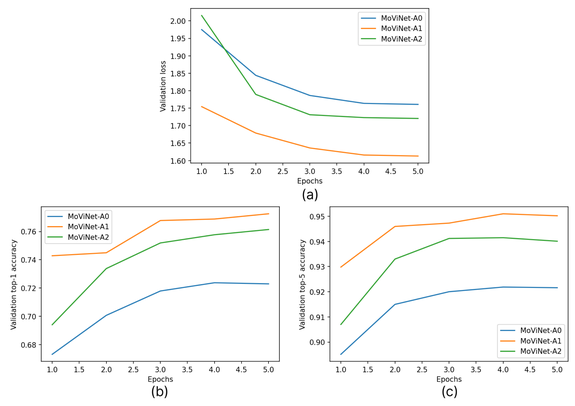

Figure 3:Results using the validation data set. (a): Loss, (b) Top-1 accuracy, (c): Top-5 accuracy

Table 2:Average results using the validation data set

Model	Loss	Top-1	Top-5
MoViNet-A0	1.8260	0.7077	0.9147
MoViNet-A1	1.6596	0.7594	0.9449
MoViNet-A2	1.7959	0.7398	0.9326

Conclusion

In this work, we fine-tuned a MoViNet architecture on the UCF101 data set. With this, we found that the current neural network-based approaches provide excellent solutions to address the video classification task. This was verified with the numerical results since we obtained a top-1 accuracy of 0.76 and a top-5 accuracy of 0.95 with our best model. Furthermore, our practical tests reflected the efficiency and effectiveness of our models when exposed to real-time recognition, as they delivered fast and accurate responses.

References

[1] Sen Jia, Shuguo Jiang, Zhijie Lin, Nanying Li, Meng Xu, and Shiqi Yu.
A survey: Deep learning for hyperspectral image classification with few labeled samples.
Neurocomputing, 448:179–204, August 2021.

[2] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam.
Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[3] Shipra Ojha and Sachin Sakhare.
Image processing techniques for object tracking in video surveillance- a survey.
In *2015 International Conference on Pervasive Computing (ICPC)*. IEEE, January 2015.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean.
Distilling the knowledge in a neural network, 2015.

Contact Information

- Repo: https://github.com/Leo-Thomas/Video-Classification-Computer-Vision-Midterm
- Email:{mike.bermeo, jean.camacho, leo.ramos}@yachaytech.edu.ec