Modelling a Deep Learning Framework for Recognition of Human Actions on Video

Flávio Santos[1][[0000−0003−2378−5376], Dalila Durães[1][0000−0002−8313−7023], Francisco Marcondes[1][0000−0002−2221−2261], Marco Gomes[1][0000−0001−6370−9955], Filipe Gonçalves[2][0000−0002−8769−4257], Joaquim Fonseca[2][0000−0002−2056−1206], Jochen Wingbermuehle[2], José Machado[1][0000−0003−4121−6169], Paulo Novais[1] [0000−0002−3549−0754]

[1] Centre Algoritmi, University of Minho, Braga, P-4710-057, Portugal
{flavio.santos, dalila.duraes, francisco.marcondes]@algoritmi.uminho.pt, {marcogomes, jmac, pjon}@di.uminho.pt
[2] Bosch Car Multimedia, Braga, P-4705-820, Portugal
{filipe.goncalves,joaquim.fonseca2, jochen.wingbermuehle}@pt.bosch.com

Abstract. In Human action recognition, the identification of actions is a system that can detect human activities. The types of human activity are classified into four different categories, depending on the complexity of the steps and the number of body parts involved in the action, namely gestures, actions, interactions, and activities [1]. It is challenging for video Human action recognition to capture useful and discriminative features because of the human body's variations. To obtain Intelligent Solutions for action recognition, it is necessary to training models to recognize which action is performed by a person. This paper conducted an experience on Human action recognition compare several deep learning models with a small dataset. The main goal is to obtain the same or better results than the literature, which apply a bigger dataset with the necessity of high-performance hardware. Our analysis provides a roadmap to reach the training, classification, and validation of each model.

Keywords: Action Recognition, Deep Learning Models, Video Intelligent Solutions.

1 Introduction

Intelligent solutions of action recognition have been studied, with different perspectives, for several disciplines, including psychology, biomechanics, and computer vision [1-2]. However, in recent years there has been a rapid growth in production and consumption of a wide variety of video data due to the popularization of high quality and relatively low-price video devices [3]. Smartphones and digital cameras contributed a lot to this factor. Simultaneously, on YouTube, there are about 300 hours of video data updates every minute [4]. New technologies such as video captioning, answering video surveys, and video-based activity/event detection are emerging every day along with the growing production of video data [5-6]. From the video input data, human activity detection indicates which activity is contained in the video and locates the regions in the video where the action occurs [7]. Also, from the computer vision community point of view, we can use visual tracking for the process of locating, identifying, and determining the dynamic configuration of one or many moving (possibly deformable), objects (or parts) in each frame of one or several cameras [8].

This paper conducted an experience of action recognition, comparing several deep learning models and obtaining better results. Our analysis provides a roadmap to reach the training, classification, and validation of each model with a dataset with a fewer class. The organization of this paper was: firstly, section 2 introduces the concepts with state of the art, namely models and dataset; then, section 3 presents materials and methods, with training data and validation data; next, section 4 presents result and discussion; and finally, section 5 concludes by performing a global conclusion and some future work.

2 State of Art

Human action recognition used several deep learning models [3, 4, 9]. However, our goal is to develop models that cover a multisensory integration process. In this stage, we will focus on optimizing the video signal learning process and afterwards expanding the architectures for efficient human action recognition by applying audiovisual information. The reason for choosing this path is twofold the different "learning dynamics" between the visual and audio information – audio generally train much faster than visual ones, which can lead to generalization issues during joint audiovisual training [9]. There are several architectures for human action recognition [3, 4]. However, the most used are C2D-Resnet 50, SlowFast, and I3D [8]. Furthermore, these architectures allow them to be combined with audio so, we will focus on these architectures.

2.1 C2D – Resnet 50

C2D is a standard 2D convolution network. A convolution network is a neural network that uses convolution in place of a fully connected matrix multiplication in at least one layer [10].

The Residual Network (ResNet) was conceived to explore a neural network depth [11-12]. It aims to handle the vanishing/exploding gradient problem that worsens according to the number of the layers raises because of a network difficulty on learning identity functions [13]. The numeral 50 denotes the network depth, i.e., the number of layers.

In short, the ResNet aims to handle the gradient descent problem caused by identity function by skipping the layers expected to compute these functions [11, 14]. Notice that ResNet cannot be directly applied to C3D. This is because the search for temporal data significantly increases the resources consumption.

2.2 SlowFast Network

The generic architecture of a SlowFast network can be described as a single stream architecture that operates at two different temporal rates (Slow pathway and Fast pathway), which are fused by lateral connections. The underlying idea is to model two tracks separately, working at low and high temporal resolutions. One is designed to capture fast-changing motion but fewer spatial details (fast pathway) and the other as a lightweight version more focused on the spatial domain and semantics (slow path) [15].

As presented in [15], the fast pathway data is fed into the slow pathway via lateral connections throughout the network, allowing the slow path to becoming aware of the fast pathway results. To do it, it requires a match to the sizes of features before fusing. At the end of each pathway, SlowFast performs global average pooling, a standard operation intended to reduce dimensionality. It then concatenates the results of the two tracks. It inserts the concatenated result into a fully connected classification layer, which uses Softmax to classify which action is taking place in the image [15].

2.3 Inflated 3D ConvNet (I3D)

By adding one dimension into a C2D (e.g. k×k) it becomes a C3D (e.g. t×k×k) [16]. Inflating is not a plain C3D but a C2D, often pre-trained, whose kernels are extended into a 3D shape. Growing is as simple as including an additional, usually temporal, dimension [12]. The I3D stands for two-stream inflated 3D convolution network [16]. Therefore, I3D is a composition of an inflated C2D with optical flow information [12, 16].

2.4 Dataset

Kinetics 700 Dataset.

The Kinetics dataset is a project that provides a large scale of video clips for human action classification, covering a varied range of human actions. This dataset contains real-world applications with video clips having a duration of around 10 seconds. The dataset's primary goal is to represent a diverse range of human actions, which can be used for human action classification and temporal localization. Another characteristic is that clips also contain audio so that the dataset can be used for multi-modal analysis. The fourth version, created in 2019, was the Kinetics-700 dataset with 700 classes, each with 700 video clips [17].

This dataset is essentially focused on human actions, where the list of action classes includes three types of actions: person actions, person-person actions, and person-object actions. The person-actions are a singular human action and include drawing, drinking, laughing, and pumping first. The person-person actions cover human actions like kissing, hugging, and shaking hands. Finally, the person-object actions contain actions like opening a present and washing dishes. Furthermore, some actions required more emphasis on the object to be distinguished, such as playing different wind instruments. Other actions required temporal reasoning to distinguish, for example, different types of swimming [17].

AVA Kinetics Dataset.

The AVA Kinetics dataset creates a crossover of the two datasets. The AVA-Kinetics dataset builds upon the AVA and Kinetics-700 datasets by providing AVA-style human action and localization annotations for many f the Kinetics videos.

The AVA-Kinetics dataset extends the Kinetics dataset with AVA style bounding boxes and atomic actions. A single frame is annotated for each Kinetics video, using a frame selection procedure described below. The AVA annotation process is applied to a subset of the training data and all video clips in the validation and testing sets from the Kinetics-700 dataset. The procedure to annotate bounding boxes for each Kinetics video clip was as follows: person detection, key-frame selection, missing box annotation, human action annotation, and human action verification [18].

3 Materials and Methods

As mentioned in section 1, the idea was to classify activities in video. The first step was to download the AVA-Kinetics datasets and cross between AVA Actions and Kinetics datasets. Downloading files from YouTube was relatively slow since the program itself blocks excess downloads. During the download IP some problems have occurred like "this video is no longer available because the YouTube account associated with this video has been terminated", the owner of this video has granted you access, please sign in: "This video is private", and this video is no longer available because the uploader has closed their YouTube account. On the second step, we evaluate the top-60 most frequent classes, and our dataset has 283 videos of the 430 videos from AVA v2.2 and 100 classes from Kinetics-700 datasets, where each class has between 650 and 1000 videos.

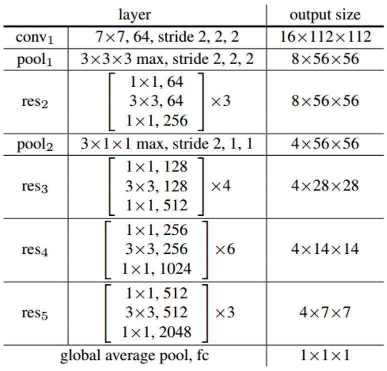
The annotation format presented was the video_id, middle_frame_timestamp, person_box, action_id, and score. The video_id is a YouTube identifier. The middle_frame_timestamp is measured in seconds from the start of the video. The person_box is normalized at upper left (x1, y1) and lower right (x2, y2) about the frame size, where (0.0, 0.0) corresponds to the upper left corner and (1.0, 1.0) corresponds the bottom right corner. The action_id is a whole identifier of an action class, from ava_action_list_v2.2_for_activitynet_2019.pbtxt. Moreover, finally, the score is a float indicating the score for that labelled box.

3.1 Architectures Networks

C2D – Resnet 50

Initially, we began training with the C2D-ResNet 50 architecture. All characteristic of this architecture is presented in Table 1.

Table 1. Global average pool of the architecture C2D-ResNet 50.



layer		output size
conv$_1$	7×7, 64, stride 2, 2, 2	16×112×112
pool$_1$	3×3×3 max, stride 2, 2, 2	8×56×56
res$_2$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	8×56×56
pool$_2$	3×1×1 max, stride 2, 1, 1	4×56×56
res$_3$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	4×28×28
res$_4$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	4×14×14
res$_5$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	4×7×7
global average pool, fc		1×1×1

The batch size was 12, LR: 0.1, the optimizer: SGD with 85 epochs and Cross-Entropy loss for this architecture.

SlowFast

Figure 1 presents a slowfast network's generic architecture, which can be described as a single stream architecture that operates at two different temporal rates (Slow pathway and Fast pathway), which are fused by lateral connections.

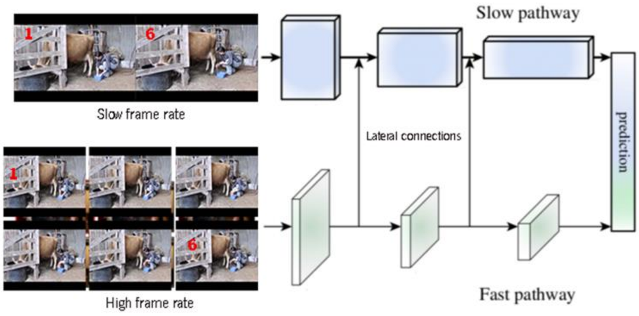


Fig. 1. A slowfast network.

Inflated 3D ConvNet (I3D)

Figure 2 presents the approach for I3D architecture. This approach begins with a 2D architecture and inflates all the filters and pooling kernels, adding a dimension layer (time).

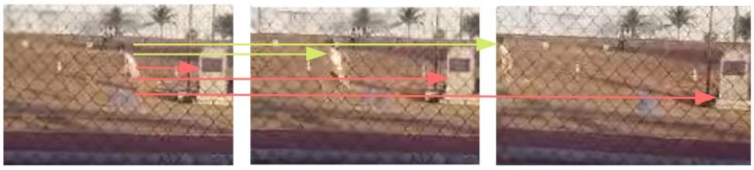


Fig. 2. Non-local action recognition example.

3.2 Training Data

Hence, we began the training only with the classes we had a download. However, for these 100 classes, we did not have the 650 - 1000 videos for each class. Because some videos are no longer available, or the owner has changed the video to private, or de video is no longer available on YouTube. Thus, the following data visualization shows the difference between the downloaded videos and full dataset. Table 2 compares the total of videos for the complete training dataset and the video download training dataset.

Table 2. Comparison between the videos of the complete training dataset and the download training dataset.

	Completed	Download
q1	510.5	454
q3	888.5	812.5
Max	997	972
Min	393	127
median	683	602

3.3 Validation Data

Regarding the validation of the videos, the script had a full validation of 50 videos. Table 3 present a comparison for the total of videos of complete validation dataset and the videos download validation dataset.

Table 3. Comparison between the videos for the complete validation dataset and the download validation dataset.

	Completed	Download
q1	48	46
q3	50	48
max	50	50
min	44	40
Median	49	47

Remember that the accuracy is obtained with the number of correct predictions, based on the total number of predictions.

4 Results and Discussion

This section presents the results and discuss the data presented in section 3, based on state-of-art, illustrated in section 2.

As we can see, Figure 3 show the training data loss for epoch in the three different architectures. In this case, we can observe that SlowFast and I3D present the worst results, and the best results were obtained for the C2D-Resnet50 architecture.

Figure 4 and Figure 5 present the training data evaluation Top1 and Top5, respectively. The best accuracy for epoch was obtained for C2D-Resnet50 architecture in Top1 and Top5.

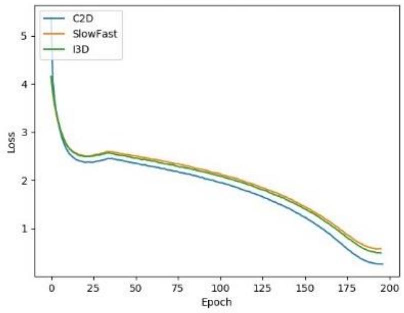


Fig. 3. Training data Loss for epoch.

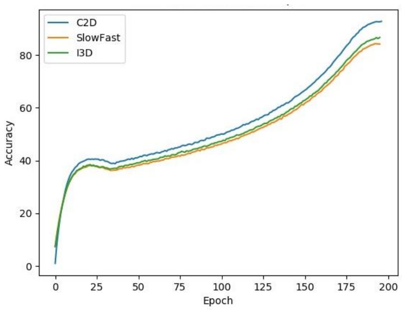


Fig. 4. Training data evaluation Top1.

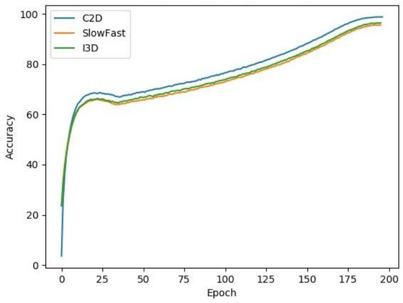


Fig. 5. Training data evaluation Top5.

For the validation data Top 1 and Top5, the results are presented in Figures 6 and 7, respectively. Also, the best accuracy for epoch was obtained for C2D-Resnet50 architecture in Top1 and Top5.

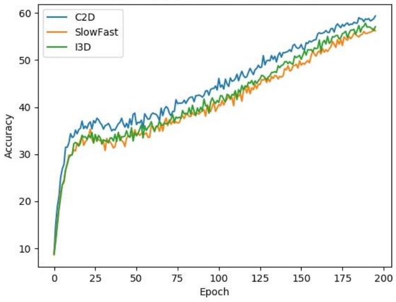


Fig. 6. Validation data – Evaluation Top1.

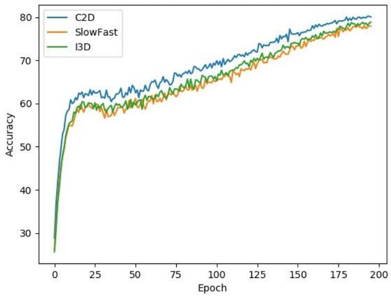


Fig. 7. Validation data – Evaluation Top5.

Table 4 presents a comparison of results for top1 and top 5 training and validation. It is possible to observe that C2D-Resnet50 architecture has better results compared with the other architecture.

Table 4. Comparison of results for top 1 and top 5 training and validation.

Architecture	Train – Top 1	Train – Top 5	Val – Top 1	Val Top 5
C2D-ResNet 50	92.79	98.82	59.36	80.19
SlowFast	84.33	95.64	56.69	78.15
I3D	86.70	96.47	57.83	78.84

Table 4 shows that the training Top1 for C2D - Resnet 50 architecture was the best accuracy of 92.79 versus 86.70 for I3D versus 84.33 for SlowFast. In the case of training Top5, C2D-Resnet50 architecture was also the best accuracy of 98.82 compared with 94.47 for I3D, and 95.64 for SlowFast. Furthermore, the validation Top1 the C2D – Resnet 50 architecture was the best accuracy with 59.36 compared with 57.83 of I3D, and 56.69 of SlowFast. Finally, validation Top5 the C2D-Resnet 50 architecture was also the best accuracy 80.19, compared with 78.84 of I3D, and 78.15 of SlowFast.

5 Conclusions and Future Work

This paper conducted an experience on video Human action recognition, which is to compare several deep learning models and obtain better results with a fewer class dataset. We began to compare three architectures C2D-Resnet 50, SlowFast, and I3D with the same baseline parameters after downloading the dataset. Comparing the results of training and validation, we can observe that C2D-Resnet 50 obtained better accuracy results for the three architectures. Our experiment results are consistent with the present in the literature, and we used a small dataset.

We intend to extend these architectures to work with the synchronized audio information to achieve better results in the next steps. Moreover, we intend to introduce Attention Models to learn which frames are most important in the classification process. Another future intends it is to apply the late fusion for the audio and video models.

Acknowledge

This work is supported by: European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project nº 039334; Funding Reference: POCI-01-0247-FEDER- 039334].

References

1. T. Ko, "KO, Teddy. A survey on behavior analysis in video surveillance for homeland security applications," 37th IEEE Applied Imagery Pattern Recognition Workshop. IEEE, pp. 1-8, 2008.
2. Analide, C., Novais, P., Machado, J., & Neves, J. (2006). Quality of knowledge in virtual entities. In Encyclopedia of Communities of Practice in Information and Knowledge Management (pp. 436-442). IGI Global.

3. Durães, D., Marcondes, F. S., Gonçalves, F., Fonseca, J., Machado, J., & Novais, P. (2020, June). Detection Violent Behaviors: A Survey. In International Symposium on Ambient Intelligence (pp. 106-116). Springer, Cham.
4. Marcondes, F. S., Durães, D., Gonçalves, F., Fonseca, J., Machado, J., & Novais, P. (2020, June). In-Vehicle Violence Detection in Carpooling: A Brief Survey Towards a General Surveillance System. In International Symposium on Distributed Computing and Artificial Intelligence (pp. 211-220). Springer, Cham.
5. Durães, D., Carneiro, D., Jiménez, A., & Novais, P. (2018). Characterizing attentive behavior in intelligent environments. Neurocomputing, 272, 46-54.
6. Costa, R., Neves, J., Novais, P., Machado, J., Lima, L., & Alberto, C. (2007, December). Intelligent mixed reality for the creation of ambient assisted living. In Portuguese Conference on Artificial Intelligence (pp. 323-331). Springer, Berlin, Heidelberg.
7. Y. Zhu, X. Zhao, Y. Fu and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," Asian conference on computer vision. Springer, Berlin, Heidelberg, pp. 660-671, 2010.
8. Jesus, T., Duarte, J., Ferreira, D., Durães, D., Marcondes, F., Santos, F., ... & Machado, J. (2020, November). Review of Trends in Automatic Human Activity Recognition Using Synthetic Audio-Visual Data. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 549-560). Springer, Cham.
9. M. Shokri, A. Harati and K. Taba, "Salient object detection in video using deep non-local neural networks," Journal of Visual Communication and Image Representation, p. 68: 102769, 2020.
10. I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, 2016.
11. K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," Proceeding of the IEEE conference on computer vision and pattern recognition, pp. 770 - 778, 2016.
12. G. Huang, Z. Liu , L. Van Der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," Proceesings of the IEEE conference on computer vision and pattern recognition, pp. 4700 - 4708, 2017.
13. S. Hochreiter, Y. BEngio, P. Fransconi and J. Schmidhuber, Gradient flow in recorrent nets: the difficulty of learning long-terms dependencies, 2001.
14. G. Huang, S. Yu, L. Zhung, S. Daniel and Q. W. Killian, "Deep networks with stochastic depth," in European conference on computer vision, Springer Cham, 2016, pp. 646-661.
15. C. Feichtenhofer, H. Fan, J. Malik and K. He, "Slowfast networks for video recognition," Proceedings of the IEEE international conference on computer vision, pp. 6202-6211, 2019.
16. J. Carreira and Z. Andrew, "Quo vadis, action recognition? a new model and the kinetics dataset," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6299-6308, 2017.
17. J. Carreira, E. Noland, C. Hillier and A. Zisserman, "A short note on the Kinetics-700 human action dataset," arXiv, vol. preprint, no. 1907.06987, 2019.
18. A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov and A. Zisserman, The AVA-Kinetics Localized Human Actions Video Dataset, arXiv preprint 2005.00214, 2020.