Spatial Attention Adapted to a LSTM Architecture with Frame Selection for Human Action Recognition in Videos

Carlos Ismael Orozco [1] **María Elena Buemi** [2] **Julio Jacobo Berlles** [2]

Abstract

Action recognition in videos is currently a topic of interest in the area of computer vision, due to potential applications such as: multimedia indexing, surveillance in public spaces, among others. In this work we propose an attention mechanism adapted to a CNN–LSTM base architecture. To carry out the training and testing phases, we used the HMDB-51 and UCF-101 datasets. We evaluate the performance of our system using accuracy as the evaluation metric, obtaining 57.3% and 90.4% for HMDB-51 and UCF-101 respectively.

1. Introduction

Human Action Recognition (HAR) is a topic of great interest in the field of pattern recognition and computer vision since the automatic identification of the action executed in a video can be a valuable tool for many applications, such as: surveillance video analysis, automatic monitoring and recognition of daily activities, video summarization, human-computer interaction, behavioral biometrics, etc. For a comprehensive guide to the current challenges of this problem, read the work by (Jegham et al., 2020).

For this, different solution strategies are proposed. Classical approaches eg. (Liu et al., 2013) propose a framework for the detection and recognition of human actions. To achieve a robust estimate of the region of interest, they use a combination of optical flow in conjunction with a Harris $3D$ edge detector to obtain space-time information from the video. Then, with the calculation of the local characteristics SIFT and STIP, they train a universal model background (UBM) for the task at hand. (Wang et al., 2011) propose a dense trajectory approach. They take dense points in each frame of the video and track them according to the displacement information of the optical flow.

Attention Mechanisms have become a very important concept within deep learning (Vaswani et al., 2017; Wang et al., 2016; Du et al., 2017; Li et al., 2018; Meng et al., 2019), its operation tries to imitate the visual capacity of the people to focus the attention on relevant parts of a scene to extract important information. These mechanisms make it possible to capture the spatial information of the scene, that is, static information such as objects, contexts, entities, etc.

(Sharma et al., 2015) first propose a soft attention based LSTM recurrent neural network for action recognition. At each time step, an attention map is learned to weight the convolutional characteristics.

(Wang et al., 2016) proposes a hierarchical attention structure to model the temporal transitions between frames and video segments. It effectively incorporates short-term movement information and long-term temporary structures.

(Li et al., 2018) replaces full connections in the LSTM with convolutional connections. It is capable of generating a 2D attention map directly for the grouping of spatial characteristics.

(Meng et al., 2019) proposes an interpretable and easy-to-connect spatio-temporal attention mechanism. Learn a featured mask to focus on the salient features in the spatial domain and employ a convolutional LSTM-based attention mechanism to identify the most relevant frames in the time domain.

The objective of this work is to implement a video action recognition system. For this we propose: (1) working on a CNN–LSTM architecture, that is, a convolutional neural network extracts the features of the video, while an LSTM neural network classifies the video in a certain category. (2) include a attention mechanism on this base proposal and (3) perform pre-processing on all videos to extract homogeneous frames by calculating the Bhattacharyya distance between consecutive frames. The work is organized as follows: in the section 2, the attention mechanism is presented. In the section 3 the databases used, the evaluation method, the experiments carried out and the results obtained

[1]DI. FCE. Universidad Nacional de Salta, Argentina. [2]DC. FCEyN. Universidad de Buenos Aires, Argentina. Correspondence to: Carlos Ismael Orozco <ciorozco.unsa@gmail.com>, María Elena Buemi <mebuemi@dc.uba.ar>, Julio Jacobo Berlles <jacobo@dc.uba.ar>.

Proceedings of the LatinX in AI (LXAI) Research workshop at ICML 2021.

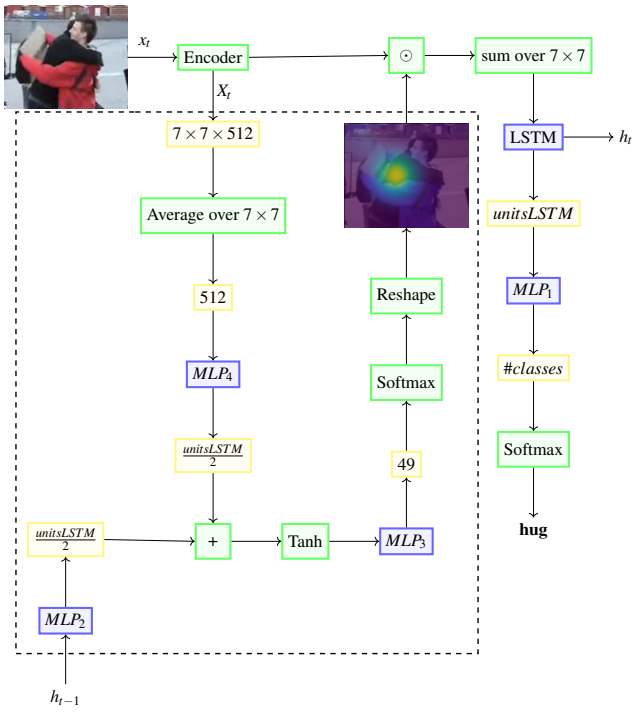

Figure 1. Architecture with attention mechanism.

Table 1. Detail of MLPs

MLP	Layer	Parameter
MLP_1	Dropout	0.5
	FC	#classes (neurons)
$MLP_{2,3and4}$	FC	128 (neurons)
	Dropout	0.5
MLP_h and **MLP_c**	FC	256 (neurons)
	Dropout	0.5

are explained. Finally, section 4 presents the conclusions and future work.

2. CNN–LSTM Approach with Attention

The Fig. 1 shows the general architecture scheme. To generate an attention map we compress the cuboid to a vector shape and together with the context vector we create a weighting vector.

For the initialization of h_0 and c_0 Xu et al. (Xu et al., 2015), compress all the information of the video achieving a faster convergence, this is calculated as:

$$h_0 = mlp_h\left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{F^2}\sum_{i=1}^{F^2}X_{t,i}\right) \quad (1)$$

$$c_0 = mlp_c\left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{F^2}\sum_{i=1}^{F^2}X_{t,i}\right) \quad (2)$$

where T is the number of frames of the videos and F is dimension of the VGG16 feature map. In our approach we adopted $T = 40$ and $F = 7$. Table 1 shows the MLP settings for initialization.

To select the T frames that will feed the architecture, we measure the dissimilarity d between the consecutive frames of the video. In particular, the dissimilarity measure is the Bhattacharya distance (d_B) between consecutive histograms h_1, and h_2 and is given by the equation 3.

$$d_B = \sqrt{1 - \sum_{i=1}^{n_b}\sqrt{h_{1,i} * h_{2,i}}}, \quad (3)$$

where h_1 and h_2 are consecutive histograms of n_b bins.

3. Experiments y Results

3.1. Databases

- HMDB-51 Human Motion dataset proposed by (Kuehne et al., 2011) has 6766 videos that belong to one of the following 51 classes: clap, drink, hug, jump, somersault, etc. It also provides three training test divisions, each of which consists of 5100 videos, 3570 for training and 1530 for testing, that is, a ratio of 70/30 per class. We evaluate the average precision over these three divisions.

- UCF-101 dataset proposed by (Soomro et al., 2012). These 101 categories can be classified into 5 types (human-object interaction, body movement only, human-human interaction, playing musical instruments and sports). The total duration of these videos is over 27 hours. All videos are collected from YouTube and are rated at 25 FPS with a resolution of 320×240. For the division of the training and test sets, we follow the configuration proposed in the original article (Soomro et al., 2012).

3.2. Results

Our system was implemented in Python using Tensorflow (Abadi et al., 2016) library on an Intel CORE i7-6700HQ computer with 16GB of DDR3 memory and Ubuntu 16.04 operating system. The experiments were carried out on an NVIDIA Titan Xp GPU mounted on a server with similar characteristics. The network parameters are optimized by minimizing the cross-entropy loss function using stochastic gradient descent with the RMSProp update rule (Dauphin et al., 2015).

Table 2 summarizes the results obtained by our system: (a) **LSTM**: Approach base see in section, (b) **A_LSTM**: with attention mechanism included, and (c) **AB_LSTM**: with attention and pre-processing of frames selection. We also include a comparison with other approaches cited in the literature.

Table 2. Video classification results

Approach	Dataset	
	HMDB-51	**UCF-101**
(Kuehne et al., 2011)	23.0%	-
(Kliper-Gross et al., 2012)	29.2%	-
(Jiang et al., 2012)	40.7%	-
(Sharma et al., 2015)	41.3%	-
(Li et al., 2018)	63.0%	-
(Wang et al., 2016)	64.3%	-
(Ye & Tian, 2016)	-	85.4%
(Zhang et al., 2018)	-	86.4%
(Zhu et al., 2017)	-	97.1%
LSTM Approach	40.7%	75.8%
A_LSTM Approach	51.2%	87.2%
AB_LSTM Approach	57.3%	90.4%

The Fig. 2 shows examples of our system's output for the HMDB-51 (left) and UCF-101 (rigth) datasets respectively. Each example is accompanied by the following information: (a) Label: action tagged for the video (ground truth). (b) Prediction: Output of our system corresponding to the class with the highest score, that is, the most probable class. (c)Attention map overlay. The region in yellow is where the system is facing and the brightness indicates the weighting.

4. Conclusions and Future Work

In this work we implement a video action recognition system, using a CNN–LSTM neural network. First, a VGG-16 extracts the features from the video. An LSTM neural network then classifies the scene into the class to which it belongs. We include a soft attention mechanism adapted for the base architecture and perform pre-processing on all videos to extract homogeneous frames by calculating the bhattacharyya distance between consecutive frames. The architecture was implemented in Python using the Tensorflow library, it was trained and tested using the databases HMDB-51 (Kuehne et al., 2011) and UCF-101 (Soomro et al., 2012) it was performed on an NVIDIA GPU Titan Xp.

We evaluate the performance of the architecture following the standard evaluation metrics for the databases used, we have obtained 40.7% (base), 51.2% (attention) and 57.3% (attention and pre-processing) for HMDB-51, 75.8% (base), 87.2% (attention) and 90.4% (attention and pre-processing) for UCF-101. We want to emphasize that the results obtained are competitive with respect to those published in the literature, taking into account the simplicity of the architecture.

For future work, we will consider using other databases, such as Hollywood2 (Marszalek et al., 2009) and UCF-50 (Reddy & Shah, 2013) to make the system more robust and delve into techniques to avoid overfitting.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

Dauphin, Y., de Vries, H., and Bengio, Y. Rmsprop and equilibrated adaptive learning rates for non-convex optimization. In *NIPS*, 2015.

Du, W., Wang, Y., and Qiao, Y. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3725–3734, 2017.

Jegham, I., Khalifa, A. B., Alouani, I., and Mahjoub, M. A. Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32:200901, 2020. ISSN 2666-2817. doi: https://doi.org/10.1016/j.fsidi.2019.200901. URL http://www.sciencedirect.com/science/article/pii/S174228761930283X.

Jiang, Y.-G., Dai, Q., Xue, X., Liu, W., and Ngo, C.-W. Trajectory-based modeling of human actions with motion reference points. In *European Conference on Computer Vision*, pp. 425–438. Springer, 2012.

Kliper-Gross, O., Gurovich, Y., Hassner, T., and Wolf, L. Motion interchange patterns for action recognition in unconstrained videos. In *European Conference on Computer Vision (ECCV)*, Oct. 2012. URL https://osnathassner.github.io/talhassner/projects/MIP.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., and Snoek, C. G. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018.

Figure 2. Examples of our architecture output. HMDB-51(left) and UCF-101 (rigth), along with the highest scoring class.

Liu, D., Shyu, M., and Zhao, G. Spatial-temporal motion information integration for action detection and recognition in non-static background. In *2013 IEEE 14th International Conference on Information Reuse Integration (IRI)*, pp. 626–633, Aug 2013. doi: 10.1109/IRI.2013.6642527.

Marszalek, M., Laptev, I., and Schmid, C. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2929–2936, June 2009. doi: 10.1109/CVPR.2009.5206557.

Meng, L., Zhao, B., Chang, B., Huang, G., Sun, W., Tung, F., and Sigal, L. Interpretable spatio-temporal attention for video action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 0–0, 2019.

Reddy, K. K. and Shah, M. Recognizing 50 human action categories of web videos. *Mach. Vision Appl.*, 24(5):971–981, July 2013. ISSN 0932-8092. doi: 10.1007/s00138-012-0450-4. URL http://dx.doi.org/10.1007/s00138-012-0450-4.

Sharma, S., Kiros, R., and Salakhutdinov, R. Action recognition using visual attention. *CoRR*, abs/1511.04119, 2015. URL http://arxiv.org/abs/1511.04119.

Soomro, K., Zamir, A. R., and Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Wang, H., Kläser, A., Schmid, C., and Liu, C. Action recognition by dense trajectories. In *CVPR 2011*, pp. 3169–3176, June 2011. doi: 10.1109/CVPR.2011.5995407.

Wang, Y., Wang, S., Tang, J., O'Hare, N., Chang, Y., and Li, B. Hierarchical attention network for action recognition in videos. *arXiv preprint arXiv:1607.06416*, 2016.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.

Ye, Y. and Tian, Y. Embedding sequential information into spatiotemporal features for action recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1110–1118, 2016.

Zhang, B., Wang, L., Wang, Z., Qiao, Y., and Wang, H. Real-time action recognition with deeply transferred motion vector cnns. *IEEE Transactions on Image Processing*, 27(5):2326–2339, 2018.

Zhu, Y., Lan, Z., Newsam, S. D., and Hauptmann, A. G. Hidden two-stream convolutional networks for action recognition. *CoRR*, abs/1704.00389, 2017. URL http://arxiv.org/abs/1704.00389.