[TO BE DEFINED]

Matteo Drago, Riccardo Lincetto†

***Abstract*—With the increasing interest in deep learning techniques and its applications, also Human Activity Recognition (HAR) saw significant improvements; before neural networks were put into practice, most of the research activities on the field relied on hand-crafted features which, however, couldn't represent nor distinguish well enough complex and articulated movements. Moreover, the use of smart devices and wearable sensors brought the challenge to another level: having to deal with high-dimensional and noisy time series while assuring optimal performances, requires a detailed study and most of all a considerable computational effort.**

In our paper we present the design of an HAR architecture which implements convolutional layers in order to extract significant features from windows of samples, along with Long-Short Term Memory (LSTM) layers, suitable to exploit time dependencies among consecutive samples. For our study, we designed the system in order to minimize the collection of layers per network and so the amount of parameters to train, which could be of great advantage in real time application. In addition, we also decided to study how performances change if we split the process in two distinct phases: the first one that performs *activity detection* while the last one activity classification. The dataset that we used to assess the efficiency of our architecture is the OPPORTUNITY dataset.

***Index Terms*—Human Activity Recognition, Machine Learning, Neural Networks, Motion Detection.**

I. Introduction

During the past decade, time series classification has captured growing interest thanks to the introduction of deep learning mechanisms, such as neural networks. These tools indeed are capable of identify and learn signal features, which are then exploited for classification, without the need of human domain-knowledge: this is a huge step forward considering that features were traditionally hand-crafted.[non trovo nessuna reference per questo] Human Activity Recognition (HAR) in particular has been fostered by the spread of powerful, efficient and affordable sensors, which nowadays are commonly found in mobile phones and wearable devices, with multiple applications, ranging from health care to gaming and virtual reality. [1] Wearable sensors allow us to collect and process a huge amount of signals, which are essential for deep neural networks (DNN) to work properly: in fact, in order for them to learn and being accurate enough to be preferred over standard machine learning approaches, we need the input training set to be heterogeneous, meaningful and representative of the problem.

For this reason, HAR is not an easy classification problem: when dealing with on-body sensors, system performances heavily depends on human behaviour, which is a source of high variability; moreover, data collected from sensors is typically high-dimensional, multi-modal and subjected to noise, making the problem even more difficult from a machine learning perspective. In the recent years, several models to perform activity detection and classification have been proposed, but as pointed out in [2] and [3], the lack of a baseline evaluation and of structured and fixed implementation details prevented a fair comparison between different solutions.

Considering that many authors in the field of machine learning and activity recognition tried to solve these problems, after an accurate study of the state of the art we decided to focus on recent works and to start from them in order to study and design improvements to the framework. Our aim was to find an architecture which gives comparable (and possibly better) performances while minimizing the number of trainable parameters and the assortment of layers in the network. The reason why we decided to go down this path is that in the literature they usually tend to expand the power of the network via increasing the computational complexity, adding layers over layers with the hope that the more number of layers, the more accurate the model. However, specifically when dealing with real time application, computational power is limited and the possibility of using difficult models is far from being realizable. [forse andrebbe messa qualche citazione, ma non ne trovo nessuna di specifica] As a first step towards the exemplification of the architecture, we decided to design two distinct networks: one dedicated to detect movements, consisting of da aggiungere descrizione del modello che decideremo; the other instead created with the purpose of classifying the movement, when detected, in this case built as da aggiungere descrizione del modello che decideremo. Then, we compare the performances of this cascade-model with a classification system that comprehends also the **Null Class** which in our case represents the state of **no activity**. With our study we aim also to provide a baseline for future works, exploiting this two-steps technique for reducing computational complexity. In order to assess the efficiency of our models, we used the **OPPORTUNITY** dataset [2], [4] which will be described in details in the following sections.

In conclusion, the contributions of this paper are:

- overview of the latest progresses of the state of the art
- implementation of these solution for comparison purpose
- the design of two separated pipelines for reducing complexity.

The paper is organized as follow: section II provides a summary of the latest and more important works related to our studies; in section III we start delving into the details of how we organized our HAR architecture, step by step; section IV is

†Department of Information Engineering, email: {matteo.drago,riccardo.lincetto}@studenti.unipd.it

dedicated to the description of the dataset and to the decisions we made in the preprocessing phase; finally in section V we are ready to describe meticulously the learning framework, while sections VI and VII are for discussion of results and for drawing our conclusions.

II. Related Work

The **OPPORTUNITY** activity recognition dataset has been introduced in [4] to overcome the lack of an evaluation setup, to compare different classification systems and to provide a more exhaustive dataset compared to the others, which "are not sufficiently rich to investigate opportunistic activity recognition, where a high number of sensors is required on the body, in objects and in the environment, with a high number of activity instances". As pointed out in [2] in fact, previously, several datasets were related to the activities which were to be classified: this is due to researchers acquiring signals only from sensors located in specific locations, according to the task of interest. To overcome this drawback, the **OPPORTUNITY** dataset has been gathered from a monitored, sensor rich environment aggiungere img dell'ambiente? : objects from the scene were connected to acquisition sensors, while people participating to the session were equipped with on-body sensors; signals collected from different sensors will be described in section IV. This particular dataset has been fundamental over the past years, it provided indeed an heterogeneous and complete set of time series, perfectly suitable for different studies in the **HAR** domain. In [2] they present it as a *benchmark dataset*; as a demonstration, they provide the results obtained with four classification techniques (*k-nearest neighbours, nearest centroid, linear discriminant analysis, quadratic discriminant analysis*) and they compare them with other works that used the same dataset. inseriamo anche i valori che ottengono nel paper per confronto?

Given that we had to perform our elaboration on this dataset, in order to make the discussion consistent in the following we introduce a collection of interesting works that use the same dataset (and others, when available) for their evaluation.

The authors in [5] proposed an exhaustive framework which, besides the standard preprocessing on the activity data sequence (filling of the gaps via interpolation and data normalization), presents also a solution for the well-known class imbalance problem [6]. Moreover, they also include a post-processing procedure after classification consisting of a smoothing operation along the temporal axis (i dati non vengono finestrati e quindi loro li filtrano) and of a strategic fusion procedure to integrate prediction sequences from different classifiers, in order to reduce the risk of making an erroneous classification. The classifiers used in this work consisted in a 1-layer neural network (1NN) and a Support Vector Machine (SVM, complete overview of this tool in [7]).

We can see an example of CNN applied to HAR in [8] where, in order to evaluate their model, they use also the Hand Gesture dataset [9]. In this configuration the authors designed a network with three consecutive convolutional blocks; the first two are constituted by a convolutional layer, a rectified linear unit layer (ReLU) and a max pooling layer. The latter instead is constituted of a convolutional layer followed by a ReLU and a normalization layer. The reason behind this collection of layers is that, while the first ones identify **basic** movements in human activity, higher layers characterize the combination of these basic movements. At the end of these core blocks, two fully-connected layers are added in order to complete the classification structure. It's important to notice also that here a sliding window strategy has been adopted to segment the time series: in this way the prediction is not focused on a single sample but is associated to a temporal matrix; the corresponding label is determined by the most-frequent label among the matrix of samples.

As a form of regularization, in [10] the authors added to the CNN also a dropout layer (more on this technique in [11]); moreover, they also put together a recurrent neural network in order to exploit time dependencies between different windows. In particular, they built two flavours of LSTM networks: one that contains multiple layers of recurrent units connected forward in time, and another which exploit dependencies either backward and forward with respect to the time-step of interest. These last configuration in particular gave the best results when applied to the **OPPORTUNITY** dataset. (dovremmo scrivere anche "in termini di f1-measure"? più che altro non l'abbiamo ancora introdotta)

In conclusion, the recent work in [12] provided a complete comparison among different features extraction and classification techniques; first, they demonstrated how the method of hand-crafted features give poor results with respect to deep learning mechanism. Then, they made a step forward with respect to the work in [10] as they created an hybrid architecture comprehensive of both a convolutional and a recurrent LSTM layer. In this way they exploited correlation among samples of a single window (as in [8]), searching for significant features; also, with LSTM they exploit correlation in time among independent windows. This combination results in a slight improvement with respect to the configurations when CNN and LSTM layers are not implemented jointly.

In our work we

III. Processing Pipeline

We start off our analysis by preprocessing the collected signals within the MATLAB framework: we chose that because it makes it simple to deal with matrices. What we do in this first step is then to import the data collected by sensors, which are given as .dat files, select the signals from on-body sensors and discard the others, replacing the missing values by means of interpolation and, at last, store them as .mat files. What we do next is to import the stored data, this time using python, and prepare the matrices for the classification task: this consists of concatenating the data, segmenting it into windows, scaling the signals and other common steps. Once the data is ready to be classified, a model is defined and trained on the available data. This is done for both the locomotion activity and gestures recognition, i.e. with two different sets of labels. This system,

which is forced to learn also the null class together with the actual movements, is then compared to a different system where two models are deployed: the first one has the only purpose of detecting activity, while the second classifies the activity, if present.

IV. Signals and Features

The signals that we use to perform HAR are the ones collected in the OPPORTUNITY activity recognition dataset. The measurement setup is then the one presented in [**?**] and [**?**]. Our analysis though is based only on on-body sensor signals, which means that we kept the signals of only a subset of the available sensors: discarding the other signals then, we ended up with 113 signals. During the preprocessing, we discarded also 3 of them, belonging to the same physical device, because there weren't any measurements in most of the cases. This led us to work on 110 signals. Since we noticed that the almost all the sensors, at the beginning and at the end of the measurement sessions, have sequences where there isn't any sample recorded, we decided to discard the head and the tail of each session, in such a way that we start and stop with all the measurements being registered. This choice was made also to facilitate interpolation. In MATLAB we perform a splines interpolation, which uses a cubic polynomial. The decision of cutting head and tail prevented our code from interpolating a piece of signal which has only one "edge". Then, to perform classification on the data of one subject, we stacked sessions ADL 1 to 3 and Drill to create our training set, and then ADL4 and ADL5 as test set. In some cases, interpolation leaves entire columns to NaN because it isn't provided any data to interpolate those values. We solved the problem by setting to 0 those entire columns. Subsequently we scaled the signals by subtracting their means and dividing by their variance (or sigma?). After this, data is shaped into windows of 15 samples (500 ms) with a stride of 5 samples. The approach that we used to segment the data was then the sliding window introduced above. To perform classification, though we had to assign to each window a unique label, which we decided to be corresponding to the label present with more samples. This doesn't constitute a problem per se, even when changing the size of the sliding window, as long as it is kept short enough and ...

V. Learning Framework

One of the main problems in Human Activity Recognition is handling inactivity.

Thinking of a real recognition system, In this paper we compare two different learning strategies, mimicking a real system. In the first V-A, One Shot Classification, the model is trained to learn a representation of the involved classes together with the null class

A. One Shot Classification

B. Two Steps Classification

VI. Results

VII. Concluding Remarks

References

[1] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys Tutorials*, vol. 15, pp. 1192–1209, Third 2013.

[2] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. del R. Millán, and D. Roggen, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognition Letters*, 2013.

[3] F. Li, K. Shirahama, M. A. Nisar, L. Kping, and M. Grzegorzek, "Comparison of feature learning methods for human activity recognition using wearable sensors," *Sensors*, vol. 18, no. 2, 2018.

[4] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Frster, G. Trster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. d. R. Milln, "Collecting complex activity datasets in highly rich networked sensor environments," in *2010 Seventh International Conference on Networked Sensing Systems (INSS)*, pp. 233–240, June 2010.

[5] H. Cao, M. N. Nguyen, C. Phua, S. Krishnaswamy, and X. Li, "An integrated framework for human activity classification.," in *UbiComp*, pp. 331–340, 2012.

[6] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[7] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[8] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition.," in *Ijcai*, vol. 15, pp. 3995–4001, 2015.

[9] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 33, 2014.

[10] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," *arXiv preprint arXiv:1604.08880*, 2016.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[12] F. Li, K. Shirahama, M. A. Nisar, L. Köping, and M. Grzegorzek, "Comparison of feature learning methods for human activity recognition using wearable sensors," *Sensors*, vol. 18, no. 2, p. 679, 2018.