[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: dealing 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 thus 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 into 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 are indeed capable of identifying and learning 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. 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 for 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 these reasons, 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 ways of performing activity detection and classification have been proposed: in the literature there's no shortage of models. The trend has been to expand the power of networks, adding more and more layers: this resulted in more accurate models, that had to face though an increasing computational complexity. However, specifically when dealing with real time applications, computational power is limited and the possibility of using too complex models is far from being realizable. Moreover, as pointed out in [2] and [3], despite the proliferation of models to perform activity detection and classification, the lack of common data to perform a baseline evaluation and of structured and fixed implementation details prevented a fair comparison between different solutions.

Considering than that the activity recognition problem has been already widely addressed by many authors, we decided to present a systematic comparison between two different commonly proposed types of pipeline. The two differ on how inactivity is handled: the first one tries to learn a representation of the signals where no action is performed, adding a null class to the other activities; the second one instead splits the classification into two tasks, first deploying an activity detector that filters out inactivity signals and then classifying the remaining activities. Our study is meant to provide a baseline for future work, giving an idea on which system could be more appealing. In order to assess the efficiency of our models, that have been designed against the trend trying to minimize the number of trainable parameters, 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 those solutions;
- comparison of two different approaches.

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 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.

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

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 presented it as a *benchmark dataset*: as a demonstration, they provided the results obtained with four classification techniques (*k-nearest neighbours, nearest centroid, linear discriminant analysis, quadratic discriminant analysis*) and they compared 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 elaborations 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 (jointly with 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.

Again, in our work

III. Processing Pipeline

We start off our analysis by preprocessing the collected signals within the MATLAB environment: we chose that framework because we find it is easier to operate with matrices. In this first step we import the data collected by sensors, which are given as .dat files, then we select the signals from on-body sensors and discard the others, so we replace the missing values by means of interpolation and, at last, we store them as .mat files.

Secondly, we import the preprocessed data in a *Jupyter Notebook* and make the dataset suitable for the classification task: this consists for example of segmenting data into windows, scaling and normalizing raw signals.

Then, after the last step of preprocessing, we define and train a suitable learning model. This is respectively 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: in that case, the first one has the purpose of

detecting activity while the second one classifies the type of movement, if detected. Figure 1 shows a schematic depiction of the pipeline, which we are going to elaborate on the following sections.

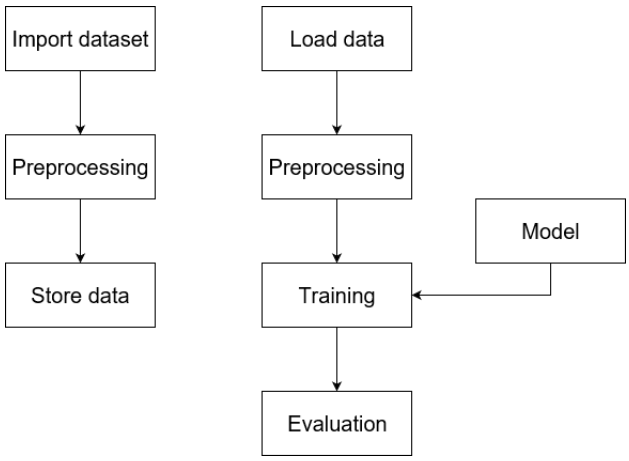

Fig. 1: Framework Pipeline

IV. Signals and Features

The **OPPORTUNITY** dataset, succinctly introduced in section II, has been collected from four subjects accomplishing different Activities of Daily Life (ADLs). As highlighted before, both the subjects and the environment they moved in were meticulously monitored. The process of acquisition consisted in 5 consecutive runs (named ADL1 to ADL5) that followed a predetermined script, plus a sixth run consisting of 20 repetitions of each of the distinct discrete activity present in the script. Then each vector of samples corresponding to a single time-step is labelled; in the following we'll refer to Task A when we consider an high-level modes of locomotion (*Standing, Walking, Sitting, Lying*) while we'll refer to Task B2 for more specific arm gestures, 17 in total. To these tasks we need of course to add the *Null Class* that we mentioned earlier in this paper: specifically, this label represents the state where the participant does nothing (or something that is not classified among the listed classes).

The wireless sensors worn by the subjects (IMU - Inertial Measurement Unit) provided acceleration among the three-axes, rate of turn, magnetic field and orientation information; in addition, 12 accelerometers were placed on the subjects' parts of the body sensible to movements (arms, back, hips and feet). All these sensors for a total of 145 distinct acquired channels. Aggiungere immagine ometti con sensori?

For the purposes of our work, however, we based the analysis only on on-body sensor signals using just a subset of the available sensors: in this way, we ended up with a total of 113 channels. Another important point is that in the preprocessing phase we performed spline interpolation (which uses a cubic polynomial) in channels that manifested missing data (equivalent to a NaN vale); however, this type of interpolation ends up meaningless if more than the 30% of data is missing. For this reason we had to discard all the three columns corresponding to one of the physical devices: finally, this led us to work with 110 channels. Since we noticed that the head and tail of the measurement sessions correspond to a transient where most of the sensors are turned-off, we decided to discard them. In this way we ensure that the interpolation phase provides consistent results; the subsequent step consists in normalizing each column with respect to mean and variance. After that, in order to instruct the framework on one subject, we stack sessions from ADL1 to 3 and Drill as a first step to create our training set; then, we assemble also ADL4 and ADL5 to build the test set.

To conclude the preprocessing phase, finally we decided to follow the same procedure as in several works we cited earlier: we apply in fact the *sliding window* technique on the datasets, obtaining a tensor of windows constituted of ... samples (... ms), using a stride of length .... In our case, we decided to assign to each window the most frequent label: 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 for being representative of a movement.

We found in fact that the choice of window size and stride is fundamental in order to obtain good results: we observed that a window too short can't represent a single gesture with fair precision and, on the contrary, if the window is too large we could include two distinct movements (or modes of locomotion), leading to misplaced classes. The choice of the size of the displacement separating two consecutive windows can guarantee a good trade-off between windows diversity and dataset population.

In the first phase of our project we also tried to create a reduced dataset of features: since each accelerometer returns the acceleration value of in the direction of all the three axes, we tried to substitute these three values with a unique value representing the *mean acceleration* measured by the sensors. In our first experiments, however, this led to poor results: with respect to the simulations where all features were fed to the model, the configuration using the reduced dataset was affected by almost a 5% loss in terms of accuracy performance.

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

As in most of the works mentioned in section II, besides accuracy, we used F_1 measure to estimate the goodness of our

models. Defining precision and recall as:

$$p = \frac{TP}{TP + FP} \qquad r = \frac{TP}{TP + FN} \tag{1}$$

the F_1 measure is the harmonic average between the two (in the previous formula we have *TP = true positive*, *FP = false positive*, *FN = false negative*). In particular, since we deal with a multi-class problem we need to add a measure of weights to the F_1 equation:

$$F_1 = \sum_i 2w_i \frac{p_i \cdot r_i}{p_i + r_i} \tag{2}$$

where the weights are defined as the number of samples of a particular class divided by the total number of samples $w_i = \frac{n_i}{N}$. The weighted measure can help also with the class imbalance problem that we outlined in the previous section; however we must highlight that our models are still trained on an imbalanced dataset, so in our opinion this could provide only a minor improvement.

In Figures 2 and 3 we present the results regarding task A (modes of locomotion, high level movements) for each participant of the experiment: in those configurations we wanted to see if there was one architecture that evidently outperform the others. As we can see, unfortunately that is not the case since among all the frameworks that we tested the differences in terms of weighted F_1 measure is negligible. The variation that we can clearly see is among distinct subjects, since they are all studied separately: S_1 performs better in both the configurations, while S_2 is the worst. Considering that we consistently performed the same procedure independently from the subject, in this case this discrepancy could be due to a problem during data collection.

Fig. 2: Task A : One-Shot Classification

With an F_1 value of ~ 0.91 for the *one-shot* scenario and ~ 0.94 in the *two-steps* (if we consider S_1), we can definitely say that the results presented in [2] are outperformed by these more powerful deep learning models; with respect to the other subjects, we can say that the result that we obtained are state of the art. In particular it's interesting to see that, when we consider the *Null Class*, the best model with S_2 is represented by the neural network with one convolutional layer (Figure 2); when the *Null Class* is ignored, instead, the hybrid model that combines convolutional layers and LSTM provides the best results (Figure 3). In our opinion the reason is that in certain cases the *Null Class* has to be intended as noise, so when the windows assigned to that class are removed the LSTM can exploit and reveal the time correlation among different sample more clearly. In addition, we tried to train our architectures on S_2 and S_3 jointly, using ADL_4 and ADL_5 of both subjects as test set: as is shown on the last column on the right in Fig 2 and 3, the results are not brilliant. In fact, we obtained a kind of average performance between the results of S_2 and S_3, in line again with what presented in [2].

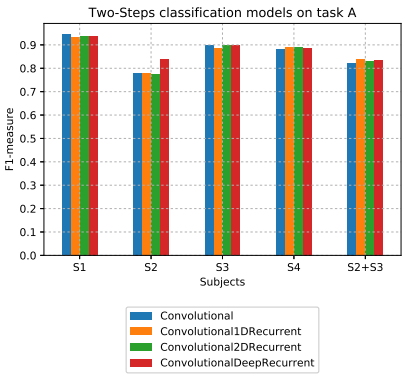

Fig. 3: Task A : Two Steps Classification

Finally, as we can see, for task A we can say that using the *two-step* approach definitely helped performances.

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.