HAR Data Generation Model to Highly Reduce Learning Space

Massinissa Hamidi
Laboratoire LIPN-UMR CNRS 7030
PRES Sorbonne Paris Cité
hamidi@lipn.univ-paris13.fr

Aomar Osmani
Laboratoire LIPN-UMR CNRS 7030
PRES Sorbonne Paris Cité
osmani@lipn.univ-paris13.fr

Abstract

In this paper, we consider leveraging the intrinsic structure that characterizes recognition of human activities (HAR) in the context of sensor-rich environments such as wearables. We propose to (1) model the underlying data generation process and (2) use it to constrain training of simpler learning models. We formulate this task as a large-scale exploration of the architectures responsible of relating sensory information. We propose an approach based on neural architecture search (NAS) techniques in order to perform the exploration and derive the global effects of the data sources. A large-scale experimental setting is used to evaluate the ability of the proposed approach to derive an effective model of the data generation process. We report on experiments conducted on the Sussex-Huawei locomotion dataset collected in real-life settings. The derived interactions and individual importances are found to be consistent with empirical results in the literature. We then demonstrate the effectiveness of the obtained model via a setting that exploits the subsets of interacting data sources in order to constrain the learning process of a neural network. Promising results open perspectives for deploying learning systems which are more robust and efficient in terms of data sampling.

1 Introduction

Proliferation of internet of things technologies allows the emergence of sensor-rich environments where sensing-enabled devices constitute sources of diverse forms of information describing their surrounding. These sources have the ability to offer a broad range of perspectives for a given situation to be interpreted [2]. Indeed, positioned in different places and generating various sensing modalities, these sources of information generate a lot of data which, if exploited rightfully, could potentially bring the robustness needed by learning processes.

Learning problematics that emerge in these sensor-rich environments are profoundly structured. This is the case of wearable technologies with the considered Sussex-Huawei locomotion-transportation (SHL) dataset [8] studied in this paper. Our work focuses on recognizing mobility-related human activities from data sources materialized by on-body sensors placed at different locations of the body following a pre-defined and fixed topology.

Indeed, the main assumption is that the recognition of human activity from wearables is a profoundly structured issue. For example, it has been observed that for a given activity, there is the emergence of dynamics that involve very specific positions of the body parts for which a set of specific modalities can provide complementary information. Primarily, what characterizes these dynamics is the fact that they define precisely the activity in question [6, 15, 19, 4]. In this paper, we propose to model the data generation processes underlying these sensor-rich wearable environments, in particular, the interactions that emerge between data sources and their respective importance. The goal is to improve the performances of human activity recognition (HAR) systems.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Currently, wearable technologies tend to develop quickly with deployments involving, for example, smart clothes, or more closely related to the considered dataset, the combination of smartphones and smartwatches, which, depending especially on the smartphone, can derive completely different topologies. Knowing, or better, deriving a model, describing the process of generating data has real potential for these use cases, where learning systems can rely on well-defined data sources, thus facilitating their adaptation to these topology evolutions.

Another benefit, which would stem as a side effect from having an explicit model of the interactions, is a smarter management of data quantities being used for recognizing human activities, by sampling solely from the data sources defined in the model being derived. This contrasts with developed HAR learning systems which are often based on fixed configurations and which assume the availability of all data sources when deployed in real-world sensor-rich environments.

2 Problem statement

In this section we start by introducing the notation, and formulate the problem of exhibiting a model for the data generation process in sensor-rich environments. We provide some challenges that led to the definition of our approach.

2.1 Preliminaries

We consider settings where a collection of M sensors (also called data generators or data sources), denoted $\{s_1, \ldots, s_M\}$, are carried by the user during daily activities and capture the body movements. Each sensor s_i generates a stream (or sequence) $\mathbf{x}^i = (x_1^i, x_2^i, \ldots)$ of observations of a given body-motion modality. Each observation is made-up of a given number of channels, e.g. x, y, and z axes of an accelerometer.

Modality. A modality is a form of perception that convey a particular perspective about a given phenomenon. In our work, we exploit mainly body-motion modalities and are often used in human activity recognition applications. E.g. acceleration, gyroscopic and magnetometric observations are different modalities each describing, in a particular way, the motions of the body.

Data source. In the present work, we consider sensor-rich environments where a given data source (or sensor), denoted s, is characterized by two main attributes: the first is the *modality* being produced by the sensor which includes for example accelerometric, gyroscopic, and magnetometric observations. The second attribute is the *position* where the data source is located on the body. A data source is then, uniquely defined with these two attributes.

(Neural) architecture. An architecture is defined as a set of components, e.g. analysis unit, feature fusion unit, decision fusion unit, etc. [3], responsible of extracting valuable insights, in the form of features, from the observations and efficiently fusing different data sources carrying different modalities and various spatial perspectives. Architectures are parameterized by a set of N hyperparameters, $h_1, \ldots, h_N$, controlling the effects of the various components, and eventually impacting the architecture performance ν.

Configuration. A configuration is an instantiation of the hyperparameters, $H = \{h_1, \ldots, h_N\}$, controlling the effect of architecture components. It defines by the same occasion an architecture. We will refer to configuration and architecture interchangeably in the rest of this paper. A configuration is refered to as θ_i.

2.2 Modeling of the data generation process

In this work, we are interested in deriving a model of the data generation process in order to exploit this knowledge to build more robust and data efficient HAR learning systems.

The main assumption is that the recognition of human activity from wearables is a deeply structured issue. For example, it has been observed that for a given activity, there is the emergence of dynamics that involve very specific positions of the body parts for which a set of specific modalities can provide complementary information. Primarily, what characterizes these dynamics is the fact that they define

precisely the activity in question. From this observation follows the concept of subset of data sources, the main characteristic of which is that they allow a robust recognition of human activities.

Regarding this point in particular, several results in the HAR literature agree on the existence of such insights. Indeed, numerous works such as [6, 15, 19], and more recently [4], where authors consider multi-modal and multi-location data sources, provide evidence of the large influence of subsets of data sources and specific interactions that emerge towards the recognition of particular human activities.

2.3 Exploration of the architecture space

As mentionned above, the goal is to find subsets of data sources and specific interactions that emerge towards the recognition of particular human activities. Precisely, we are seeking subsets $X \subsetneq S$ such that the individual and global impact of the data sources belonging to X, on the recognition of a particular human activity, are substantial (relatively to the rest of data sources).

This requires modeling the dynamics, that involve the subsets of data sources, directly from the body movements of each human activity. This is a rather complex task which requires, beyond the huge amount of heavy experiments that have to be conducted in real settings, human expertise which is scarce and limited in terms of the insights it may convey. Thus, rather than relying on a direct modeling of these dynamics, we recast this problem into an analysis of the behavior of the architectures which are responsible of modeling these dynamics through feature extraction and sensor fusion schemes. In other words, the problem is reframed as an exploration of the architecture space and for which adequate tools are available to solve it.

In particular, we focus on the potential insights that could stem from tuning and adapting these architectures, through their hyperparameters and those controlling specifically the influence of the data sources. Indeed, in order to obtain such subsets, we have to be able to leverage information contained *within* and *across* data sources using efficient features extraction and sensor fusion schemes: the information contained *within* a data source determine its individual impact, while information contained *across* data sources determine, this time, the notion of interaction or global impact of interacting data sources.

Challenges. This problem formulation leads us to exhibit three main challenges. First, the various sensing modalities featured by sensor-rich environments carry information with various perspectives. However, building appropriate features extraction and sensor fusion schemes, is not trivial. So, how can we leverage information contained within and across data sources when we know limits of human expertise in terms of features extraction and sensor fusion schemes? Second, modeling of the data generation processes being recasted into and exploration of an induced architecture space, how can one explore such, potentially, very large spaces. Framed differently, the exploration of the whole space being infeasible, how can we get a fairly precise picture of it with a constrained exploration budget? And third, after doing all this, how can we quantify the individual and global impact of the data sources, based on the exploration results, in order to form subsets that satisfy the hypothesis we formulated previously?

3 Approach

In this section, we describe an original approach for exhibiting the data generation process underlying a sensor-rich environment. The main idea is to use architectures based on neural networks in order to overcome aforementioned limits of human expertise and to come-up with genuine features extraction and sensor fusion schemes. Exploration of the induced space is then recasted as a problem that can be treated with neural architecture search NAS techniques.

We describe the convolutional modes and the resulting neural architectures space in Sec. 3.1. We discuss how we explore the neural architectures space in Sec. 3.2 and how we come-up with the global impact of data sources using the functional analysis of variance in Sec. 3.3.

3.1 Features learning and fusion strategies

Neural networks hold important properties which are advantageous to multimodal recognition tasks. They are able to construct, or learn, hierarchies of abstract features and relate efficiently modalities between them. In our work, to replicate such capabilities within our architectures, we use convolutional neural networks.

Beyond the frequent application of convolutional neural networks for the recognition of human activities, which show, by the way, good performances, e.g. [10, 16, 18, 4], these kinds of networks are being adopted, primarily, for their ability to efficiently aggregate heterogeneous data from different sources. In [10] for example, authors proposed various convolutional architectures featuring an explicit mechanism for partial and full weight sharing, by placing separate convolution kernels on each modality and in the upper layers responsible for aggregating features maps.

In this work, we construct neural architectures by stacking Conv/ReLU/MaxPool blocks. These blocks are followed by a Fully Connected/ReLU layers. In order to allow for the emergence of cross-modal relationships at both low and high-levels of abstractions, we define three *convolutional modes* of the input sequences with each set of filters of the first layer (input layer):

- whole modalities grouped and convolved, which we refer to as *grouped modalities*.
- each modality convolved apart, which is designated by *split modalities*.
- each channel convolved apart, referred to as *split channels*.

These various convolutional modes can be considered as different levels of sensor fusion. From this perspective, *split channels* would correspond to a late fusion and, on the contrary, *grouped modalities* would correspond to an early fusion scheme.

In order to train a given architecture, we frame recognizing human activities as a sequence classification problem, where the goal is to learn a function $\mathcal{F} : X \to Y$ mapping inputs to outputs. In order to simplify the given sequences of observations $\mathbf{x}^i|_{i\in\{1\dots M\}}$ (defined above), we thus introduce a segmentation step, which will be responsible of decomposing the stream of observations, $\mathbf{x}^i$, into sequences of fixed size, and optionally overlapping. Each sequence of observations (indexed by j), $\mathbf{x}^i_j \in X$, is assigned with a label, $y_j \in Y = \{1, \dots, A\}$, corresponding to the human activity being performed. As in the traditional classification setting, performance of the neural architecture is quantified with a loss function $\ell : X \times Y \to \mathbb{R}$, and a mapping is found via

$$f^* = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(f(\mathbf{x}^i_j), y_j) \tag{1}$$

which can be optimized using a Gradient descent algorithm over a pre-defined class of functions $\mathcal{F}$. In our case, $\mathcal{F}$ will be convolutional neural networks parametrized by their weights and the loss function will be $\ell(f(\mathbf{x}_i), y_i) = \mathbb{1}\{f(\mathbf{x}_i) \neq y_i\}$. For a fixed architecture, i.e. a particular instantiation of the hyperparameters, the optimization process will tune the weights of the network and, by the same occasion, the subsequent uni-modal and multi-modal features that are extracted from the input signals.

But beyond the optimization of the parameters of an architecture, one of the most important aspects, which defines to an extremely great extent its potential performances, is the set of hyperparameters which determines and forges the shape of the architecture. In the case of convolutional neural networks, the set of hyperparameters include the filters of the different layers, both in terms of size and number, the strides, and so on.

At each layer of a given architecture, setting the right combination of hyperparameters is critical. In particular, setting the right instantiation for the features learning and sensor fusion stages can lead to an architecture capable of building an original set of features from the various data sources which is suitable for recognizing a given activity. This is particularly appealing in the sense that by varying (particular sets of) hyperparameters (independently or not), we can discover some interesting insights regarding the interactions and phenomenon that emerge in these high dimensional spaces. Moreover, recently, the field of neural networks has seen a resurgence of approaches focusing on this aspect, i.e. architecture engineering, hyperparameters tuning, where the performance of a model is largely determined by the architecture. These approaches are grouped under the term of neural architecture search (NAS).

Clearly, the previously defined convolutional modes provide an additional value to the configuration that these form with the hyperparameters tuning, especially in terms of sensor fusion strategies via weight sharing. The combination of the large-scale optimization of the hyperparameters, provided by NAS techniques, and the previously introduced convolutional modes defines what we call the *neural architecture space*. In the following, we present how we explore this space using Bayesian optimization (BO) which offers a good trade-off between exploration and exploitation.

3.2 Optimal exploration of the neural architectures space

Exploring the hyperparameters space implies that we use an efficient method that can save us from exploring a lot of configurations, that do not provide with useful insights, and concentrate rather on regions of high improvement potential. In this sense, recent advances in neural architecture search demonstrated noticeable successes in many fields leading in certain cases to state-of-the-art architectures using Bayesian optimization in particular [20]. We tune hyperparameters via BO based on Gaussian process as a surrogate model in which the generalization performance of a given configuration θ is modeled as a sample from a Gaussian process, referred to as the response surface. Expected improvement $EI(\theta) = -\mathbb{E}[f(\theta) - f(\theta + t)]$ is used as an acquisition function in order to direct sampling, at step t, of areas of the neural architecture space where an improvement of the performances is likely to happen. The Gaussian process from which hyperparameter instances are sampled is updated at each step with the recognition performance achieved by the induced neural architecture.

In this work, the performance of an architecture is denoted ν_k and is obtained using a performance metric such as the f1-score or accuracy. We also exploit the partial performances obtained during the training of the architectures for each of the human activities considered. These are denoted ν_k^a, where k refers to the configuration and a to the activity. The partial performances ν_k^a will give us an idea about, precisely, the behavior of the architectures towards the dynamics of particular activities.

3.3 Global effects of data sources

In the previous section, we performed the exploration of the neural architecture space and obtained a response surface which ensures a good coverage. Precisely, at this point, we have the performances (and the partial performances), $\nu_1, \ldots, \nu_R$, which correspond to R runs of the BO. These performances are indexed by the corresponding architecture (or configuration).

The efficient implementation of fANOVA proposed in [13], is used to quantify the individual and global impact of data sources. Their implementation is based on a linear-time algorithm for computing marginal predictions in random forests. It is used to build a predictive model of the explored neural architecture space as a function of the configurations, that have been evaluated.

Before diving into the details of the computations, we first formalize the notions of *importance* and *interaction*, which have been mentioned, in the problem formulation (Sec. 2.3), as being the information contained *within* and *across* data sources, respectively.

Importance. Importance of a data source is defined as a quantity that corresponds to the level of its influence on the performances of a (neural) architecture. We denote the importance of a given data source m_i by $\mu_{m_i} \in [0, 1[$, or simply μ_i.

Interaction. An interaction, in the other hand, involves two or more data sources and is defined as their degree of dependence regarding their influence on the performance of a (neural) architecture. Given a set of interacting data sources, I, their degree of interaction is denoted by μ_I, and specifically, for two interacting data sources, m_i and m_j, it is denoted by $\mu_{<m_i,m_j>}$, or simply $\mu_{i,j}$.

In addition, we define two parameters, τ_{imp} and τ_{int}, which correspond to the importance and interaction thresholds, respectively. These two parameters determine limits above which a given data source, with a given importance, or a set of interacting data sources, could be included in a subset of data sources. It follows that the subsets of interacting data sources, denoted X, is defined as

$$X := \{m_i \in S | \mu_{m_i} \geq \tau_{imp}\} \cup \{I \subsetneq S | \mu_I \geq \tau_{int}\} \tag{2}$$

In the following, we will detail how the importance and interaction of the data sources are computed. Functional analysis of variance (functional ANOVA) is a prominent data analysis method [12]. Broadly, ANOVA partitions the observed variation of an architecture performance ν into components due to each of its hyperparameters h_i. In the following, we will use the formalism proposed in [13].

Given subsets, U, of the hyperparameters H, functional ANOVA [13] decomposes the performance of the neural architecture space $\hat{y} : \theta_1 \times \cdots \times \theta_n \rightarrow \mathbb{R}$ into the sum of its marginal components corresponding to the various hyperparameter instantiations that only depend on subsets of the hyperparameters N:

$$\hat{y}(\theta) = \sum_{U \subsetneq H} \hat{f}_U(\theta_U) \tag{3}$$

The components $\hat{f}_U(\theta_U)$ are defined as follows:

$$\hat{f}_U(\theta_U) = \begin{cases} \hat{f}_\emptyset & \text{if } U = \emptyset \\ \hat{a}_U(\theta_U) - \sum_{W \subsetneq H} \hat{f}_W(\theta_W) & \text{otherwise} \end{cases} \tag{4}$$

where the constant $\hat{f}_\emptyset$ is the mean value of the function over its domain, and $\hat{a}_U(\theta_U)$ is defined as the average performance of all complete instantiations θ that agree with θ_U. For each hyperparameter h_i, the function $\hat{f}_{\{h_i\}}(\theta_{\{h_i\}})$ captures the effect of varying it on the recognition performances of the neural architectures, and this, by averaging across all possible values of all other hyperparameters.

By definition, the variance of $\hat{y}$ across its domain θ is

$$\mathbb{V} = \frac{1}{||\theta||} \int (\hat{y}(\theta) - \hat{f}_\emptyset)^2 d\theta \tag{5}$$

Given the individual components, functional ANOVA decomposes the variance $\mathbb{V}$ of $\hat{y}$ into the contributions by all subsets of hyperparameters $\mathbb{V}_U$:

$$\mathbb{V} = \sum_{U \subset H} \mathbb{V}_U, \text{with } \mathbb{V}_U = \frac{1}{||\theta_U||} \int \hat{f}_U(\theta_U)^2 d\theta_U, \tag{6}$$

where $\frac{1}{||\theta_U||}$ is the probability density of the uniform distribution across θ_U. The importance of each main and interaction effect $\hat{f}_U$ can thus be quantified by the fraction of variance it explains: $\mathbb{F}_U = \mathbb{V}_U / \mathbb{V}$. It follows that the individual impact of a data source m_i is determined

$$\mu_i = \frac{1}{|h_i|} \sum_{h_i} \sum_{U \subset H} \mathbb{F}_U \tag{7}$$

This final step produces for each activity, subsets of the most relevant and interacting data sources.

4 Experiments

In this section, we perform empirical evaluations of the proposed approach on a subset of the SHL dataset. Code to reproduce the experiments is publicly made available [1].

4.1 Dataset description

The SHL dataset [9] [2] is a highly versatile and precisely annotated dataset dedicated to mobility-related human activity recognition and aiming to overcome the lack of such datasets. The total amount of collected data corresponds to 2812 h of labeled data and 17 562 km of traveled distance. In contrast to related representative datasets like [23] which includes solely global positioning system information, the SHL dataset contains multi-modal locomotion data recorded in real-life settings.

There are in total 16 modalities including accelerometer, gyroscope, cellular networks, WiFi networks, audio, etc. making it suitable for a wide range of applications and in particular the task we are

[1] `https://github.com/alphaequivalence/shl-nas`

[2] The preview of the SHL data set can be downloaded from: `http://www.shl-dataset.org/download/`.

interested in. Indeed, there are 8 primary categories of activities that we are interested in: *Still*, *Walk*, *Run*, *Bike*, *Car*, *Bus*, *Train*, *Subway (Tube)*. The locomotion data was recorded by three participants, referred to in the dataset as *User1*, *User2*, and *User3*, involved full-time. Data collection was performed by each participant using four smartphones simultaneously placed in different body locations where people are commonly carrying phones, *Hand*, *Torso*, *Hips*, and *Bag*.

Featuring a broad range of spatial perspectives for the human activities constitutes another strength that is credited to the SHL dataset in contrast to related datasets like [22]. These four positions define the topology of the underlying infrastructure that we study in this paper. Among the 16 modalities of the original dataset, we select the body-motion modalities to be included in our experiments, namely: accelerometer, gyroscope, magnetometer, linear acceleration, orientation, gravity, and in addition, ambient pressure.

Performance evaluation. In our experiments, each architecture was evaluated with a 10-fold meta-segmented cross-validation to avoid the problem of overestimation of the quality of results induced by standard cross-validation procedure [11]. This technique relies on a modified partitioning procedure that alleviates the neighborhood bias, which results from the high probability that adjacent (moreover, overlapping) sequences fall into training and test-set at the same time.

Performance metrics. Concerning the performance metrics, we use the f1-score in order to assess models' recognition performances. This metric has the advantage of giving more realistic insights about the models' performances but presents some weaknesses when it is confronted with strongly unbalanced datasets [7] which introduce bias depending on the way this metric is calculated, *i.e.* averaging f1-score in each fold, averaging over precision and recall in each fold, etc. We compute the f1-score as in [7] by averaging its different components obtained for each fold. Given the usual definition of precision $\mathrm{Pr}^{(i)}$ and recall $\mathrm{Re}^{(i)}$ for the ith fold, we compute the f1-score as in [7] by averaging its different components obtained for each fold.

4.2 Qualitative evaluation of the neural architecture space exploration

In our first experiment, we evaluate the ability of the proposed approach to derive an effective model of the data generation process that underlies the considered sensor-rich environment. We thus perform a large-scale analysis of interactions through the proposed approach as well as an extended literature review around the recognition of human activities from various modalities and spatial perspectives.

We work on the SHL dataset and consider the provided body-motion sensing modalities located on the four body locations. We assume that a certain structure underly each considered human activity and seek the subsets of interacting data sources and confront them to related empirical evidences.

Training details. Models development and BO process are based on off-the-shelf implementations, all of which are free software. In particular, we use TensorFlow [1] for building the neural architectures, and the scikit-optimize library [17] specialized in optimizing cost functions. We make use of an efficient implementation of the fanova framework proposed by [13] which is based on random forests. Inputs are segmented into sequences of 6000 samples which correspond to a duration of 1 min. given a sampling rate of 1 Hz. We use batch normalization on top of each convolutional layer. This procedure makes the neural networks more stable by normalizing the inputs of the different layers [14].

Each hyperparameter instantiation is trained for a maximum of 52 epochs and a minimum of 12 epochs. After 13 epochs, if there is no improvement in the recognition performances over the best score for 8 subsequent epochs, we consider that the model has reached a local minimum so we stop training.

Evaluation. More than $6k$ architectures were evaluated during our experiments which resulted in more than $54k$ partial performances. Figure. 1 shows how data sources grouped by their respective positions contribute to the overall recognition performances of each human activity. Figure. 2 and Table 1 summarize results of the analysis conducted via the fANOVA framework and show respectively the individual and the pairwise marginal importance of the hyperparameters controlling features learning and fusion stages.

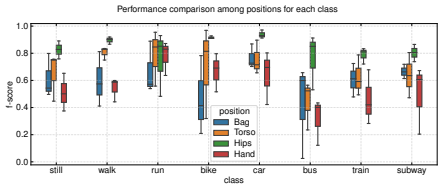

Figure 1: Contribution of the data sources to the overall recognition performances of each human activity. Data sources are grouped by their respective positions.

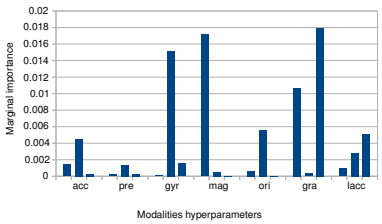

Figure 2: Individual marginal importance of the kernel size hyperparameters controlling the impact of each modality.

Hyperparam.	Interaction ($\times 10^{-4}$)
(ks^2_{gyr}, ks^2_{gra})	9.2778
(ks^1_{mag}, ks^2_{ori})	7.0166
(ks^2_{gyr}, ks^2_{ori})	5.5122
(ks^1_{acc}, ks^1_{mag})	4.0382
(ks^1_{pre}, ks^3_{gyr})	2.3154
(ks^3_{gyr}, ks^1_{mag})	2.2472

Table 1: Most important pairwise marginals (interactions) of the kernel size hyperparameter. Hyperparameters are grouped by their corresponding modalities.

The contribution of the data generators for the bus, train, and subway are equivalent, with more variability that appears slightly in the case of the bus activity. Models trained on the Hips data generators for their part, show the smallest variability overall. The variability of these model's performances is to some extent more consequent in the case of bus and run activities but stays in fairly acceptable terms. This could be linked to the problems raised above. In the case of car activities, the use of the Hips data generators seems to be sufficient, this position yielding the best models overall (90%-95% f1-score). This same observation is also made on bike and walk activities where Hips seem to discriminate them accurately. This may be explained by the tight link that exists between these activities and the hips position: biking, walking and conducting a car involve specific repetitive patterns that are their hallmark [5].

From modalities perspective, data sources carrying gravity, gyroscope, and magnetometer account for a large part of the variability that is observed on the recognition performances. Surprisingly, another set of modalities emerges from the derived model rather than the accelerometric data which is considered to be one of the most important modalities in representative related work [21, 19]. Indeed, the respective individual marginal importance of the accelerometer-related data lies approximately around 0.4% and does not exceed 0.6%, while those of gravity, gyroscope, as well as magnetometer, reach 1% and almost 2% (See Figure. 2). This observation is further confirmed when we analyze the pairwise marginals of the hyperparameters controlling the set of three modalities mentioned above.

4.3 Effectiveness of the interactions model

In the previous experiments, we evaluated empirically the ability of the proposed approach to make explicit the subsets of interacting data sources that emerge in a sensors deployment, and this, w.r.t. the recognition of each particular human activity. We saw in particular the emergence of some interactions that are found to be consistent with empirical results in the literature. In addition, as we setup the neural architectures to construct abstract features and novel sensor fusion schemes, we see the emergence of some *less* common forms of interactions.

We can leverage all these interactions in order to constrain training of (simpler) architectures. During this *constrained* training phase, these architectures are encouraged to concentrate on the provided subsets of data sources to learn the correspondings human activities. This could substantially improve robustness of these architectures as only "currated" subsets of inputs are provided making it easy to learn patterns. Thus, in this second experiment, we want to evaluate the effectiveness of the obtained

interactions model via a simple setting where the training of our neural architectures is constrained with these interactions. Specifically, our main assumption is that the subsets of interacting data sources, which are obtained using our approach, are able to define precisely a given human activity.

Training details. We construct neural networks, similar to the architectures used to derive the data generation model, with exactly 3 Conv/ReLU/MaxPool stacked blocks. These blocks are followed by a Fully Connected/ReLU layers. The weights of the layers corresponding to all inputs are optimized during training without distinction, the constraining being specified via data augmentation.

Indeed, in this setting, for each subset of interacting data sources, we perform data augmentation by assigning values, drawn from a normal distribution, to the unimportant data sources. The goal is to make the neural network insensitive to the noisy inputs. We provide training examples to the neural network according to the given pairs and triplets of interacting data sources that we extract from the derived model. Training of the neural networks was performed using subsets of data sources parameterized with different values of τ_{int} and τ_{imp}.

Evaluation. For comparison, we also trained a neural network on the whole data sources. This network achieved a performance of 70.86 % measured by the f1-score. We used for this neural network the set of hyperparameters that were obtained for the best configuration during the exploration of the neural architecture space.

Figure. 3 shows the evolution of the neural network depending on the parameters τ_{int} and τ_{imp}. In addition, this figure illustrates the average number of data sources, that are included in the subsets, depending on these two thresholds.

In particular, when τ_{imp} and τ_{int} are set, for example to 0, all data sources are included. We find that the neural networks trained with smaller subsets of data sources perform better than the baseline and most of the settings relying to higher number of data sources. Surprisingly, we do not see a lot of bad subsets of interacting data sources for $0.2 \leq \tau_{imp} \leq 0.6$, where the number of data sources per subset is confined between 12 and 5.

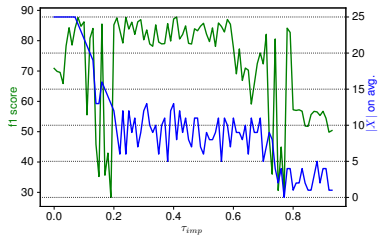

Figure 3: Recognition performance (in green) obtained in the constrained learning setting and average cardinality of the subsets of data sources (in blue) as a function of τ_{imp}.

5 Conclusion

We presented in this paper a novel approach for deriving subsets of interacting data sources. These subsets are found to impact, substantially and in various degrees, the recognition of different human activities. We demonstrate the effectiveness of the obtained interactions model via a setting that exploits the subsets of interacting data sources in order to constrain the learning process of a neural network. Obtained promising results open perspectives for the development of more robust and data-efficient learning systems.

References

[1] Martín Abadi et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Jake K Aggarwal and Quin Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.

[3] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[4] Antonio Bevilacqua, Kyle MacDonald, Aamina Rangarej, Venessa Widjaya, Brian Caulfield, and Tahar Kechadi. Human activity recognition with convolutional neural networks. In *Joint*

European Conference on Machine Learning and Knowledge Discovery in Databases, pages 541–552. Springer, 2018.

[5] Claudia Carpineti, Vincenzo Lomonaco, Luca Bedogni, Marco Di Felice, and Luciano Bononi. Custom dual transportation mode detection by smartphone devices exploiting sensor diversity. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 367–372. IEEE, 2018.

[6] F Foerster, M Smeja, and J Fahrenberg. Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring. *Computers in Human Behavior*, 15(5):571–583, 1999.

[7] George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1):49–57, 2010.

[8] Hristijan Gjoreski, Mathias Ciliberto, Francisco Javier Ordoñez Morales, Daniel Roggen, Sami Mekki, and Stefan Valentin. A versatile annotated dataset for multimodal locomotion analytics with mobile devices. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, page 61. ACM, 2017.

[9] Hristijan Gjoreski, Mathias Ciliberto, Li Wang, Francisco Javier Ordonez Morales, Sami Mekki, Stefan Valentin, and Daniel Roggen. The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. *IEEE Access*, 2018.

[10] Sojeong Ha and Seungjin Choi. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 381–388. IEEE, 2016.

[11] Nils Y Hammerla and Thomas Plötz. Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1041–1051. ACM, 2015.

[12] Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

[13] Holger Hoos and Kevin Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *International Conference on Machine Learning*, pages 754–762, 2014.

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37, pages 448–456. PMLR, 2015.

[15] Jani Mantyjarvi, Johan Himberg, and Tapio Seppanen. Recognizing human motion with multiple acceleration sensors. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 2, pages 747–752. IEEE, 2001.

[16] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[17] Fabian Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[18] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):157, 2018.

[19] Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13, 2010.

[20] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[21] Shuangquan Wang, Canfeng Chen, and Jian Ma. Accelerometer based transportation mode recognition on mobile phones. In *Wearable Computing Systems (APWCS), 2010 Asia-Pacific Conference on*, pages 44–46. IEEE, 2010.

[22] Meng-Chieh Yu, Tong Yu, Shao-Chen Wang, Chih-Jen Lin, and Edward Y Chang. Big data small footprint: the design of a low-power classifier for detecting transportation modes. *Proceedings of the VLDB Endowment*, 7(13):1429–1440, 2014.

[23] Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.