Data Generation Process Modeling for Activity Recognition

Abstract. The dynamics of body movements are often driven by large and intricate low-level interactions involving various body parts. These dynamics are part of an underlying data generation process. Incorporating the data generation process into data-driven activity recognition systems has the potential to enhance their robustness and data-efficiency. In this paper, we propose to model the underlying data generation process and use it to constrain training of simpler learning models via sample selection. As deriving such models using human expertise is hard, we propose to frame this task as a large-scale exploration of architectures in charge of relating sensory information coming from the data sources. We report on experiments conducted on the Sussex-Huawei locomotion dataset featuring a sensor-rich environment in real-life settings. The derived model is found to be consistent with existing domain knowledge. Compared to the basic setting, our approach achieves up to 17.84% improvement, by simultaneously reducing the number of required data sources by one-half. Promising results open perspectives for deploying more robust and data-efficient learning models. [1]

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 offer a broad range of perspectives allowing to perform robust activity recognition [33]. Indeed, positioned in different places and featuring various sensing modalities, these sources of information generate a lot of data which, if exploited rightfully, could provide many advantages like improved signal-to-noise ratio, reduced ambiguity, and enhanced reliability [22].

Learning tasks 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 [14] 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. 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 largely the activity in question [11,25,35,5].

[1] Code and experiments are available at `https://github.com/sensor-rich/shl-nas`

The dynamics of body movements are part of an underlying data generation process (DGP) and a long line of research, e.g. [30,31,8], proposed to incorporate this kind of prior knowledge into activity recognition models. Specifically, authors in [31] derive 3D body skeleton-based representations while other works encode prior domain knowledge using ontology-based representations [29,36]. These representations are then used to constrain training of activity recognition models. While incorporation of prior knowledge about the dynamics of body movements into learning systems improves performances and is appealing in terms of interpretability, relying solely on human expertise to derive models for these dynamics is hard [39]. Indeed, these dynamics are often driven by large and intricate low-level interactions involving various body parts [21].

In this paper, we propose a novel approach to derive and incorporate DGP into activity recognition models. Our approach enhances the performance of activity recognition models through two major steps. It first constructs a model of the DGP via a large-scale exploration of a neural architecture space. Then, it selects highly confident data sources for inclusion in the final training set using a variance-based importance estimation algorithm.

Our contributions can be summarized as follows. (1) We frame the derivation of the data generation process as an exploration of the neural architecture space; (2) We propose to estimate the relative importance of data sources and their interactions using a variance-based method; (3) Extensive experiments show the effectiveness of combining the data generation process through selection of highly confident data sources. In particular, we achieve improvement of recognition performances of up to 17.84% over the baseline, which is accompanied by a substantial reduction of required data; (4) We perform a comprehensive comparative analysis using different instantiations of the proposed approach (8 exploration strategies) on 4 different representative related datasets.

The rest of the paper is organized as follows. In Section 2, we define the problem of data source selection based on the DGP and Section 3 presents the details of our approach. In Section 4, we detail the empirical evaluation of the proposed approach. We provide a related work in Section 5 and finally, Section 6 concludes this paper with a summary and future works.

2 Problem Statement

This section defines the problem of modeling the data generation process in the context of activity recognition from sensor-rich environments.

2.1 Preliminaries

We consider settings where a collection $\mathcal{S}$ 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 $\mathbf{x}^i = (x_1^i, x_2^i, \ldots)$ of observations of a certain modality. Furthermore, each observation is composed of channels, e.g. the x, y, and z axes of an accelerometer.

In our work, we exploit mainly body-motion modalities that are often used in human activity recognition applications. The goal is to recognize a set $\mathcal{Y}$ of activities, like running or biking, performed in sensor-rich environments.

Definition 1 (Modality). *A modality is a form of perception that conveys a particular perspective about a given phenomenon. E.g. acceleration, gyroscopic and magnetometric observations are different modalities each describing, in a particular way, the motions of the body.*

Definition 2 (Data source). *a given data source (or sensor), denoted s, is characterized by two main attributes: the first is the* **modality** *being produced by the sensor and the second one is the* **position** *where the data source is located on the body. A data source is then uniquely defined with these two attributes.*

2.2 Problem Definition

Human activities are largely determined by the dynamics of the gestures. Indeed, each activity is characterized by a different set of gestures which in turn involve specific body parts. In the case of wearable technologies, where these body parts are equipped with data sources, often, focusing on these specific data sources, allows recognizing a given activity precisely. Therefore, our approach attempts to select subsets of data sources that are highly confident and informative with regards to these dynamics, to create a curated training set for model training. In this work, we focus on two different notions that encode these dynamics: importance of a single data source and degree of interaction among a set of data sources.

Definition 3 (Importance). *Given a data source s_i that is attached to a given body part and an activity y, the importance of s_i with regards to activity y, denoted $\mu_i^y \in [0, 1)$, is defined as a quantity that represents the relative involvement of that body part in the dynamics of the gestures pertaining to that activity.*

Definition 4 (Interaction). *An interaction involves two or more data sources and is defined as their degree of dependence regarding the relative involvement of the body parts, they are attached to, in the dynamics of the gestures. The greater the degree, the more interacting the data sources. Given a set of interacting data sources, $S \subset \mathcal{S}$, their degree of interaction is denoted by $\mu_S^y \in [0, 1)$. Specifically, in the case of two interacting data sources, s_i and s_j, it is denoted μ_{ij}^y.*

Problem 1 (Data source selection based on DGP). Let $DGP : \wp(\mathcal{S}) \times \mathcal{Y} \to [0, 1)$ be the data generation process, which gives, for each activity $y \in \mathcal{Y}$ the influence of a set of data sources $S \subset \mathcal{S}$. The goal is to use DGP as an indicator function to select data sources (or samples) that are highly confident and informative to be included in the final training set of activity recognition models. Let $\tau_{imp} \in [0, 1)$ and $\tau_{int} \in [0, 1)$ be two parameters that determine the thresholds above which a given set of data sources $S \subset \mathcal{S}$ can be selected. It follows that the subsets of interacting data sources pertaining to activity $y \in \mathcal{Y}$, denoted $\mathcal{S}_y$, is defined as $\mathcal{S}_y := \{s_i \in \mathcal{S} | DGP(\{s_i\}, y) = \mu_{s_i}^y \geq \tau_{imp}\} \cup \{S \subsetneq \mathcal{S} | DGP(S, y) = \mu_S^y \geq \tau_{int}\}$

Learning using curated sources of information is widely used in machine learning [38,17]. The DGP in the Problem 1 presents a natural solution for selecting such sources in the context of activity recognition from sensor-rich environments.

3 Approach

Our approach enhances the performance of activity recognition models through two major steps: (1) construct a model of the DGP as described in the Problem 1 using an architecture space as a surrogate model (proxy), and (2) select highly confident and informative data sources for inclusion in the final training set using a variance-based importance estimation algorithm. These two steps are described in the following and Algorithm 1 outlines the complete process.

3.1 Architecture Space as Proxy for the DGP

We use the space defined by multimodal analysis architectures as a proxy for the dynamics of the body movements. The exploration of this architecture space serves, then, to derive the DGP as defined in Problem 1.

An architecture is defined as a set of architectural components responsible for extracting valuable insights, in the form of features, from the observations and efficiently fusing different data sources carrying different modalities and various spatial perspectives. We distinguish four types of architectural components: *feature extraction* (FE), *feature fusion* (FF), *decision fusion* (DF), and *analysis unit* (AU) as defined in [2]. These are illustrated in Fig. 1 (left). An architectural component takes as inputs either raw data, features, or decisions and outputs either a feature or a decision. The way a given component processes each individual input is controlled by a hyperparameter.

It is convenient to represent an architecture as a directed acyclic graph where the architectural components are connected together using valued edges. We associate a value (hyperparameter) h_u^v with every edge in the directed graph that connects two components C_u and C_v. These values control how architectural components process each individual input and by the same occasion their influence on the overall architecture performance. We refer to the set of all hyperparameters of a given architecture by $\mathcal{H}$.

We focus, particularly, on the insights that stem from tuning and adapting these architectures, through their hyperparameters and specifically those controlling the influence of the data sources. At each layer of a given architecture, setting the right combination of hyperparameters is critical. In particular, choosing the right instantiation for the features learning and sensor fusion components can lead to an architecture capable of building, from the various data sources, an original set of features which is suitable for recognizing a given activity. We take into account the following assumption: let $\mathcal{H}_s \subsetneq \mathcal{H}$ be the set of hyperparameters controlling the impact of a given data source s. The global impact of $\mathcal{H}_s$ on the recognition performances represents the impact of the data source s.

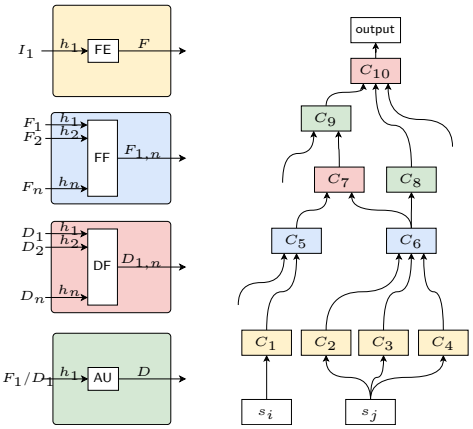

Fig. 1: (Left) feature extraction and multimodal fusion components defined in [2]. Feature extraction (FE), feature fusion (FF), decision fusion (DF), and analysis unit (AU). These building blocks can be combined in order to form feature-level, decision-level, and hybrid multimodal analysis. Additionally, the hyperparameters h_i controlling the effects of each individual input are depicted. (Right) An illustration of an architecture where each node corresponds to a component. An edge from component C_u to component C_v denotes that C_v receives the output of C_u as input.

The problem of modeling the DGP becomes, then, an exploration of the architecture (hyperparameter) space. This exploration is determined by three aspects: (1) a search space which defines the architectural components and the type of branching that is allowed for the architectures (e.g. convolutional layers); (2) a search strategy which decides how the exploration of the space should be carried (e.g. Bayesian optimization of the hyperparameters); and (3) a performance estimation strategy (e.g. sequence classification problem) [9].

In the case of convolutional layers, for example, architectures can be constructed by stacking a series of Conv1d/ReLU/MaxPool blocks followed by Fully-Connected/ReLU layers. Denote by ν_k the validation loss of a particular instantiation k of the set of hyperparameters. The exploration strategy tries to find an architecture k^* that minimizes the validation loss $\nu_k^*(w^*)$. The weights w associated with the architecture are obtained by optimizing the weights of the components using, for example, a gradient descent algorithm over a predefined class of functions.

Given an exploration budget B, the exploration strategy yields a series of validation losses $\nu_1, \ldots, \nu_B$ including partial validation losses pertaining to individual activities. The task of modeling the DGP, therefore, reduces to find a link between these validation losses and the impact of each individual data source.

3.2 Variance-Based Importance Estimation

Let $\mathcal{V}$ be a set of validation losses where each validation loss ν_k represents the estimated performance of a particular instantiation of the hyperparameters. To estimate the importance of each individual data source, we decompose the non-linear relation f described by $\mathcal{V}$ as follows

$$f(\mathcal{S}, y) = \mu_0^y + \sum_{i=1}^{M} \mu_i^y(s_i) + \sum_{i \neq j} \mu_{ij}^y(s_i, s_j) + \cdots + \mu_{1\ldots M}^y(s_i, \ldots, s_M) \quad (1)$$

a constant mean μ_0^y plus first-order effects (μ_i^y), plus second-order effects (μ_{ij}^y) and so on. The lower the variance induced by a data source, the higher its influence of on the non-linear relation f. This formulation corresponds to an additive expansion and the variance of each term can be estimated using the functional analysis of variance (fANOVA) [18]. It can be quantified using the efficient implementation proposed in [19] which is based on a linear-time algorithm for computing marginals of random forest predictions.

As we have access to the set of validation losses indexed by the hyperparameters instantiation, in order to estimate the decomposition, we have to determine, first, the correspondence between each individual data source and the set of hyperparameters that controls their influence.

Data source/hyperparameters correspondence. Given an architecture A, we determine a correspondence, $Corr_A : \mathcal{S} \rightarrow \wp(\mathcal{H} \times \mathbb{R})$, between each individual data source and the hyperparameters that influence their effects, as follows:

$$Corr_A(s) = \bigcup_{(u,v) \in s \rightarrow^* t} < h_u^v, w > \quad (2)$$

where $s \rightarrow^* t$ denotes all paths in the architecture that have s as a source and t as sink, h_u^v the hyperparameter associated with edge (u, v), and w corresponds to a weight computed as $w = \frac{\omega_1 \cdot dist(s,v) + \omega_2 \cdot \delta^-(v)}{\omega_1 + \omega_2}$ which ponders the correspondence of a given hyperparameter h_u^v depending on its distance ($dist(s, v)$) to the input s and the number of incoming edges to the component v ($\delta^-(v)$). The weight parameters $\omega_1, \omega_2 \in [0, 1)$ are both set to $\frac{1}{2}$. In the case an edge is shared by many different paths, we sum the weights assigned to the corresponding hyperparameter following each path.

4 Experiments and Results

In this section, we perform empirical evaluation of the proposed approach. We first derive a model of the data generation process from the SHL dataset using different space exploration strategies. We, then, demonstrate the effectiveness of incorporating the derived model into four different activity recognition datasets (including SHL). Code to reproduce the experiments is publicly made available [2].

[2] Software package and code to reproduce empirical results are publicly available at: `https://github.com/sensor-rich/shl-nas`

Algorithm 1: DGP-based Data Sources Selection

Input : (i) $\{\mathbf{x}^i\}_{i=1}^M$ streams of annotated observations generated by the data sources, (ii) B exploration budget, (iii) τ_{imp}, (iv) τ_{int}, (v) E exploration strategy, (vi) O maximal order of interaction effects

Result: $\mathcal{S}_y \in \wp(\mathcal{S})|_{y\in\mathcal{Y}}$, the sets of most important and interacting data sources for each individual human activity

begin
  $\mathcal{V} \leftarrow \varnothing$, $\mathcal{S}_y \leftarrow \varnothing$
  $(X, Y) \leftarrow$ segmentation$(\{\mathbf{x}^i\}_{i=1}^M)$; % *preprocess for sequence classif. pblm*
  $\mathcal{V} \leftarrow E(X, Y, B)$; % *architecture space exploration*
  foreach $s \in \mathcal{S}$ **do**
    $\{(h, w)|h \in \mathcal{H}, w \in \mathbb{R}\}_s \leftarrow Corr_A(s)$; % *DS/HPs correspondence*
  end
  $\{\mu_S^y | S \subsetneq \mathcal{S}\} \leftarrow$ QuantifyImportance$(\mathcal{V}, \{\{(h, w)\}_s\}, O)$; % *Section 3.2*
  foreach *activity* $y' \in \mathcal{Y}$ **do**
    foreach $\mu_S^y \in \{\mu_S^y | S \subsetneq \mathcal{S}, y = y'\}$ **do**
      if $\mu_S^y > \tau_{int}$ **then**
        ; % *use* τ_{imp} *if* $S = s$
        $\mathcal{S}_y \leftarrow \mathcal{S}_y \cup S$
      end
    end
  end
  return $\{\mathcal{S}_y\}_{y\in\mathcal{Y}}$
end

4.1 Datasets

We use the SHL dataset primarily to derive the data generation model. The derived model is then incorporated into the SHL dataset itself and three other datasets including (1) *USC-HAD* [42] containing body-motion modalities of 12 daily activities collected from 14 subjects (7 male,7 female) using MotionNode, a 6-DOF inertial measurement unit, that integrates a 3-axis accelerometer, 3-axis gyroscope, and a 3-axis magnetometer; (2) *HTC-TMD* [41] containing accelerometer, gyroscope, and magnetometer data all sampled at 30Hz from smartphone built-in sensors in the context of energy footprint reduction; and (3) *US-TMD* [6] featuring motion data collected from 13 subjects (9 male, 4 female) using smartphone built-in sensors.

SHL dataset. The SHL dataset [15] is a highly versatile and precisely annotated dataset dedicated to mobility-related human activity recognition. In contrast to related representative datasets like [43,42,41,6], the SHL dataset provides, simultaneously, multimodal and multilocation 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. Data collection was performed by each participant using four smartphones simultaneously placed in different body locations: *Hand*, *Torso*,

Hips, and *Bag*. These four positions define the topology that allows us to model and leverage the dynamics of body movements for activity recognition models. 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.

4.2 Training Details

We use Tensorflow [1] for building the neural architectures. In this work, we construct neural architectures by stacking Conv1d/ReLU/MaxPool blocks. These blocks are followed by a Fully Connected/ReLU layers. Architecture performance estimation is based on the validation loss and is framed as a sequence classification problem. As a preprocessing step, annotated input streams from the SHL dataset are segmented into sequences of 6000 samples which correspond to a duration of 1 min. given a sampling rate of 100 Hz. For weight optimization, we use stochastic gradient descent with Nesterov momentum of 0.9 and a learning-rate of 0.1 for a minimum of 12 epochs (we stop training if there is not improvement). Weight decay is set to 0.0001. Furthermore, to make the neural networks more stable, we use batch normalization on top of each convolutional layer [20].

Different exploration strategies will lead to different sets of hyperparameter instantiations. In our experiments, we instantiate our approach with various exploration strategies. We use the Microsoft-NNI toolkit [3] which provides a comprehensive list of exploration strategies, in particular, those based on hyperparameter tuning, including (1) exhaustive search (random search [3], and grid search); (2) heuristic search (naïve evolution [34], anneal [4], and hyperband [23]); and (3) sequential model-based optimization (Bayesian optimization hyperband [10], tree-structured Parzen Estimator [4], and Gaussian process tuner [4]).

We quantify the influence of data sources using the efficient implementation of fANOVA proposed in [19], which is based on a linear-time algorithm for computing marginal predictions in random forests. Interaction structure of the data sources is estimated using fanova-graph [26].

4.3 Performance Evaluation

In our experiments, each architecture is 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 [16]. This technique relies on a modified partitioning procedure that alleviates the neighborhood bias, which results from the high probability that adjacent (moreover, overlapping) segments fall into training and test-set at the same time.

We use the f1-score in order to assess performances of the architectures. We compute this metric following the method recommended in [12] to alleviate bias

[3] `https://github.com/microsoft/nni`

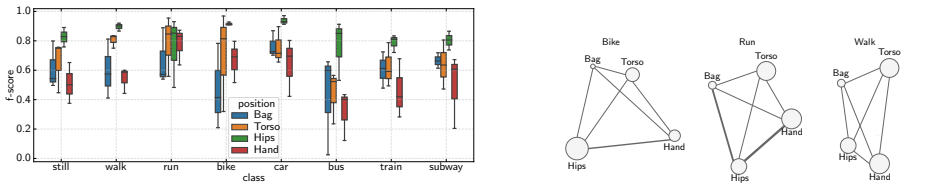

Fig. 2: (a) Contribution of the data sources to the overall recognition performances of each human activity. (b) Estimated interaction structure (fANOVA graph [26]) of the data sources for 3 different activities (bike, run, and walk). Data sources are grouped by their respective positions. The circumference of the circles represents main effects (importance), the thickness of the edges represents total interaction effects.

that could stem from unbalanced class distribution. Given the usual definition of precision $\mathrm{Pr}^{(i)}$ and recall $\mathrm{Re}^{(i)}$ for the ith fold, we compute the f1-score by averaging its different components obtained for each fold as $\mathrm{F}_{\mathtt{avg}} = \frac{1}{k} \cdot \sum_{i=1}^{k} \mathrm{F}^{(i)}$ where $\mathrm{F}^{(i)} = 2 \cdot \frac{\mathrm{Pr}^{(i)} \cdot \mathrm{Re}^{(i)}}{\mathrm{Pr}^{(i)} + \mathrm{Re}^{(i)}}$, if both $\mathrm{Pr}^{(i)}$ and $\mathrm{Re}^{(i)}$ are defined. The i-superscripted measures correspond to measures obtained when the ith fold is used as the test set.

4.4 Evaluation of the Data Generation Model

Here we evaluate the data generation model that is derived using our proposed approach. We specifically assess the plausibility of the derived important data sources and their interactions based on a comprehensive set of studies around activity recognition. These studies are compiled into a data generation model that we refer to as human expertise-based data generation model *HExp*. Furthermore, we instantiate our proposed approach using different space exploration strategies and compare the derived knowledge using each strategy.

Fig. 2 shows how data sources grouped by their respective positions contribute to the overall recognition of each human activity. Fig. 3 and Table 1 summarize results of the variance-based importance estimation conducted using the fANOVA framework (Sect. 3.2). The estimated first and second order effects of the hyperparameters controlling the importance of each considered modality are illustrated, respectively.

Data source location. Results in Fig. 2 show that the contributions of data sources for recognizing bus, train, and subway related activities are equivalent. More variability appears in the case of the bus activity. Data sources located on the hips, for their part, yield overall the smallest variability. This variability is to some extent more important in the case of bus and run activities but stays in fairly acceptable terms. In the case of car-related activities, relying on the data sources located on the hips seems to be sufficient, this position yielding the best models overall (90%-95% f1-score). The same observation can also be made

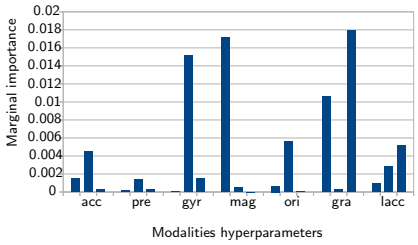

Fig. 3: Individual marginal importance of the kernel size (convolution) hyperparameters controlling the impact of each modality. 3 layers per modality are shown.

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 interactions of the kernel size (ks) hyperparameter. Hyperparameters are grouped by the corresponding modalities that they control.

regarding bike and walk activities where Hips data sources 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 [6].

Data source modality. 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 [40,35]. Indeed, the respective individual marginal importance of the accelerometer-related data lies approximately around 0.004 and does not exceed 0.006, while those of gravity, gyroscope, as well as magnetometer, reach 0.01 and almost 0.02 (See Fig. 3). This observation is further confirmed when we analyze the pairwise marginals of the hyperparameters controlling the set of three modalities mentioned above.

Impact of the space exploration strategies. Here, we compare the data generation models obtained using different space exploration strategies. Specifically, we compare the derived subsets of data sources in terms of their level of agreement with those aggregated in the human expertise-based data generation model (HExp). We use for this, Cohen's kappa coefficient [7] which measures the agreement between two raters. We also compare the average recognition performance ν_k of the explored architectures which can indicate many aspects concerning the exploration strategy, like the concentration of important sets of data sources in regions of the architecture space.

Results in Table 2 show that the sequential model-based exploration strategies are indeed better than heuristic search-based ones. Exhaustive search-based strategies are far behind with an agreement that does not exceed 3. It is worth mentioning that even with a larger exploration budget allowed to exhaustive search, using these kinds of strategies does not allow to derive a valuable data generation process. This could be explained by the fact that important sets of

Exploration strategy	Agreement	ν_k on avg.
Exhaustive search		
Random Search	0.156± 0.04	67.12%
Grid Search	**0.251±0.05**	66.78%
Heuristic search		
Naïve evolution	0.347±0.12	73.35%
Anneal	**0.481±0.05**	75.47%
Hyperband	0.395±0.08	74.2%
Sequential Model-Based		
BOHB	0.734±0.03	84.25%
TPE	0.645±0.1	83.87%
GP Tuner	**0.865±0.02**	84.95%

Table 2: Degree of agreement with human expertise and average cardinality of the derived sets of data sources obtained using different space exploration strategies.

data sources are concentrated in very specific regions the grid search, for example, can not capture. As the GP tuner yields the highest agreement with HExp, in the following, we will, first, use the data generation process derived using this strategy to assess the effectiveness of incorporating such knowledge into activity recognition models.

4.5 Effectiveness of the Data Generation Model

In this second experiment, we incorporate the derived data generation model into activity recognition models via sample selection. We select highly informative data sources to form training sets. During the training phase, activity recognition models are encouraged to concentrate on the provided subsets of data sources to learn the corresponding human activities. We refer to this setting as *w-DGP*, which stands for, with data generation process.

For this, we construct activity recognition models based on neural networks, similar to the architectures used to derive the data generation model, but restricted to 3 Conv1d/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 remaining inputs. We provide training examples to the neural network according to the given subsets of interacting data sources that we extract from the derived model. Furthermore, we experiment with different values of τ_{int} and τ_{imp} to extract the subsets of data sources.

For comparison, we train the activity recognition models on the whole data sources of each dataset, i.e., without incorporation of the derived data generation

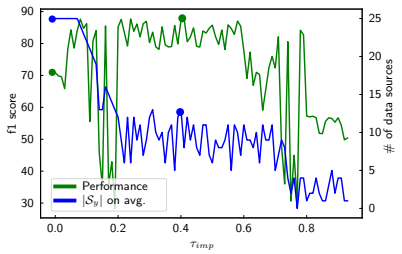

Fig. 4: Recognition performances as a function of the data source importance threshold τ_{imp}. In parallel, the cardinality on average of the subsets $|\mathcal{S}_y|$ used to train the models is shown. The left-most points correspond to a configuration where all data sources are used, i.e., no DGP.

Dataset	Performances		
	wo-DGP	w-HExp	w-DGP
USC-HAD	72.1%	75.38%	89.33%
HTC-TMD	74.4%	77.16%	78.9%
US-TMD	71.32%	80.28%	83.64%
SHL	70.86 %	77.18%	88.7%

Table 3: Comparison of different DGP incorporation settings in terms of recognition performances. Scores of column w-DGP correspond to top-performing models selected while varying the data source importance threshold τ_{imp}.

model. These models constitute our baselines and we refer to this setting as *wo-DGP*. In addition, we also incorporate the data generation model based on human expertise (HExp). We refer to this setting as *w-HExp*. Table 3 compares recognition performances obtained, on each dataset, using these settings. Overall, we obtain substantial improvements for all datasets when incorporating a data generation process (either w-HExp or w-DGP). It is to note, though, that for HTC-TMD, we get a smaller improvement compared to the other datasets. This could be related to the limited number of modalities and unavailability of the precise location of the data sources.

Fig. 4 shows the evolution of the obtained recognition performances 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, for example, τ_{imp} and τ_{int} are set 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 which rely on a higher number of data sources. Noticeably, we get a recognition performance of 88.7%$\pm$0.6, measured by the f1-score, using subsets containing on average 12 data sources. Thus, an improvement over the baseline of 17.84% in terms of recognition performances and a reduction of one-half concerning the required quantities of data. 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 5 and 12. It is also worthy to note that in some configurations where $|\mathcal{S}_y| = 13$, the trained model performs badly (less than 40%$\pm$0.16 f1-score). In the contrary, for smaller subsets ($|\mathcal{S}_y| \leq 5$), trained models get high recognition performances (more than 80%$\pm$0.05 f1-score). A Deeper inspection of these configurations reveals that the location of selected data sources plays an important role, in particular, the latter subsets are mainly composed of hips data sources.

Dataset	**Exhaustive search**		**Heuristic search**			**Sequential model-based**		
	Random	Grid	Naïve	Anneal	HB	BOHB	TPE	GP Tun.
USC	79.28%	79.58%	80.76%	83.56%	85.27%	86.66%	82.37%	**89.33%**
HTC-TMD	76.34%	75.17%	74.98%	73.18%	77.45%	75.86%	**80.13%**	78.9%
US-TMD	74.14%	72.21%	79.71%	81.13%	80.80%	79.17%	**84.39%**	83.64%
SHL	72.2%	71.32%	79.46%	84.16%	82.33%	84.22%	86.7%	**88.7%**

Table 4: Recognition performances of activity recognition models while incorporating the data generation models derived using different space exploration strategies.

4.6 Alternative Exploration Strategies

In the previous experiment, we constrain training of activity recognition models using data generation model derived using the Gaussian process tuner as it had the highest degree of agreement with HExp. Since the exploration strategies tend to favor different regions of the architecture space, we hypothesize that the derived models will be characterized by variety in terms of combinations of data sources but will still hold the same property, which is being highly informative with regards to the dynamics of body movements. Here we evaluate the effectiveness of the data generation models derived using the other exploration strategies. Table 4 presents the results obtained for this setting on each individual dataset.

Note that TPE outperforms GP tuner in the case of HTC-TMD and US-TMD datasets. It is also interesting to note that even though exhaustive search strategies have a low degree of agreement with HExp, incorporation of their corresponding data generation models is competitive for both HTC-TMD and USC-HAD, which can be explained by the ability of our approach to derive knowledge that is hardly captured by the sole human expertise.

5 Related Work

In our work, we proposed to derive and incorporate the DGP into activity recognition models. Incorporating domain knowledge into activity recognition models is particularly appealing and attracted lots of research.

A long line of research, e.g. [39,21,30,31,8], proposed to incorporate the 3D body skeleton-based representation into activity recognition models. Specifically, authors in [31] estimate centroids for upper, middle and lower body and use slopes of the segments delimited by these centroids in order to represent the posture in terms of the overall orientation of the upper and lower body. In [30], authors introduced a representation based on the calculation of spherical angles between selected joints and the respective angular velocities. They used their system for real-time tracking of human activities. Other works encode prior domain knowledge using ontology-based representations [29,36] which are then

used to constrain training of activity recognition models. While incorporation of domain knowledge into learning processes is beneficial, the way it is done differs substantially from one approach to another. In [21], the obtained ontology serves as a basis for constructing a network of Bayesian inference while in [31], the constructed representations help the neural networks to self-organize.

Beyond activity recognition, many other applications leverage domain models to enforce certain conditions or equations, which are part of prior knowledge, within machine learning models. In [37,27], authors propose to incorporate domain knowledge, like known laws of physics, by constraining neural networks via regularization. Their settings introduce new challenges for encoding knowledge into appropriate loss functions and avoiding trivial solutions in the constraint space. In the same vein, authors in [28] propose to make use of a more experimented model, a proxy, that is responsible for selecting samples in order to train new generations of models in the context of industrial monitoring. New paradigms, like Vapnik's learning using privileged information [38] and Hinton's distilled knowledge [17], propose to incorporate high capacity models, similar to proxy's, called "intelligent teachers" into machine learning models.

A growing volume of work proposes to exploit domain knowledge to improve the performances of machine learning models. Our experiments encourage an even broader range of future applications, where larger and more experienced models like the proposed neural architecture space, form surrogates for prior domain knowledge and provide guidance to simpler models via sample selection.

6 Summary and Future Work

We presented in this paper a novel approach for deriving a model of the data generation process underlying sensor-rich environments. We framed this task as an exploration of the neural architecture space and proposed a variance-based method to estimate the relative importance of data sources and their interactions. Incorporating the derived data generation model into activity recognition models allows us to obtain consistent improvement in recognition performances using a reduced number of data sources. We performed a comprehensive comparative study on various representative datasets using different instantiations of the space exploration strategy. Obtained promising results open perspectives for the development of more robust and data-efficient learning systems pertaining to the internet of things.

In this work, we used exploration strategies based on hyperparameters tuning. An alternative way is to have fine-grained control on the architectural components that make up the neural architectures allowing for more specialized architectures. Recent approaches in neural architecture search, such as ENAS [32] and DARTS [24], enable this kind of granularity. Furthermore, recent advances in weight-agnostic neural architectures [13] and the possibility of building architectures that are completely specialized in a given task and requiring no further weight adjustments open perspectives for these kinds of approaches. As part of our future work, we plan to derive more precise data generation processes using these fine-grained control mechanisms.

References

1. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI. vol. 16, pp. 265–283 (2016)
2. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia systems **16**(6), 345–379 (2010)
3. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. JMLR **13**, 281–305 (2012)
4. Bergstra, J.S., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. In: NIPS. pp. 2546–2554 (2011)
5. Bevilacqua, A., MacDonald, K., Rangarej, A., Widjaya, V., Caulfield, B., Kechadi, T.: Human activity recognition with convolutional neural networks. In: ECML-PKDD. pp. 541–552. Springer (2018)
6. Carpineti, C., Lomonaco, V., Bedogni, L., Di Felice, M., Bononi, L.: Custom dual transportation mode detection by smartphone devices exploiting sensor diversity. In: Int. Conf. on Perv. Comp. and Comm. Workshops. pp. 367–372. IEEE (2018)
7. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
8. Dhiman, C., Vishwakarma, D.K., Aggarwal, P.: Skeleton based activity recognition by fusing part-wise spatio-temporal and attention driven residues. arXiv preprint arXiv:1912.00576 (2019)
9. Elsken, T., Metzen, J.H., Hutter, F.: Neural architecture search: A survey. JMLR **20**(55), 1–21 (2019)
10. Falkner, S., Klein, A., Hutter, F.: Bohb: Robust and efficient hyperparameter optimization at scale. In: ICML. pp. 1437–1446 (2018)
11. Foerster, F., Smeja, M., Fahrenberg, J.: Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring. Computers in Human Behavior **15**(5), 571–583 (1999)
12. Forman, G., Scholz, M.: Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. SIGKDD Explor. Newslet. **12**(1), 49–57 (2010)
13. Gaier, A., Ha, D.: Weight agnostic neural networks. In: NeurIPS (2019)
14. Gjoreski, H., Ciliberto, M., Morales, F.J.O., Roggen, D., Mekki, S., Valentin, S.: A versatile annotated dataset for multimodal locomotion analytics with mobile devices. In: Conf. on Embed. Net. Sens. Sys. p. 61. ACM (2017)
15. Gjoreski, H., Ciliberto, M., Wang, L., Ordonez Morales, F.J., Mekki, S., Valentin, S., Roggen, D.: The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. IEEE Access (2018)
16. Hammerla, N.Y., Plötz, T.: Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In: UbiComp. pp. 1041–1051. ACM (2015)
17. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
18. Hoeffding, W.: A non-parametric test of independence. The annals of mathematical statistics pp. 546–557 (1948)
19. Hoos, H., Leyton-Brown, K.: An efficient approach for assessing hyperparameter importance. In: International Conference on Machine Learning. pp. 754–762 (2014)
20. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. vol. 37, pp. 448–456. PMLR (2015)
21. Kovalenko, M., Antoshchuk, S., Sieck, J.: Real-time hand tracking and gesture recognition using semantic-probabilistic network. In: Int. Conf. on Computer Modelling and Simulation. pp. 269–274. IEEE (2014)

22. Kurle, R., Günnemann, S., van der Smagt, P.: Multi-source neural variational inference. In: AAAI. vol. 33, pp. 4114–4121 (2019)
23. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A.: Hyperband: A novel bandit-based approach to hyperparameter optimization. JMLR **18**(1), 6765–6816 (2017)
24. Liu, H., Simonyan, K., Yang, Y.: Darts: Differentiable architecture search. In: ICLR (2019)
25. Mantyjarvi, J., Himberg, J., Seppanen, T.: Recognizing human motion with multiple acceleration sensors. In: SMC. vol. 2, pp. 747–752. IEEE (2001)
26. Muehlenstaedt, T., Roustant, O., Carraro, L., Kuhnt, S.: Data-driven kriging models based on fanova-decomposition. Stat. and Comput. **22**(3), 723–738 (2012)
27. Nabian, M.A., Meidani, H.: Physics-driven regularization of deep neural networks for enhanced engineering design and analysis. J. of Comput. and Inf. Sci. in Eng. **20**(1) (2020)
28. Osmani, A., Hamidi, M., Bouhouche, S.: Monitoring of a dynamical system based on autoencoders. In: IJCAI (2019)
29. Ousmer, M., Vanderdonckt, J., Buraga, S.: An ontology for reasoning on body-based gestures. In: SIGCHI EICS. pp. 1–6. ACM (2019)
30. Papadopoulos, G.T., Axenopoulos, A., Daras, P.: Real-time skeleton-tracking-based human action recognition using kinect data. In: Int. Conf. on Multimedia Modeling. pp. 473–483. Springer (2014)
31. Parisi, G.I., Tani, J., Weber, C., Wermter, S.: Emergence of multimodal action representations from neural network self-organization. Cogn. Syst. Res. **43**, 208–221 (2017)
32. Pham, H., Guan, M., Zoph, B., Le, Q., Dean, J.: Efficient neural architecture search via parameters sharing. In: ICML. vol. 80, pp. 4095–4104. PMLR (2018)
33. Radu, V., Tong, C., Bhattacharya, S., Lane, N.D., Mascolo, C., Marina, M.K., Kawsar, F.: Multimodal deep learning for activity and context recognition. IMWUT **1**(4), 157 (2018)
34. Real, E., Moore, S., Selle, A., Saxena, S., Suematsu, Y.L., Tan, J., Le, Q.V., Kurakin, A.: Large-scale evolution of image classifiers. In: ICML. pp. 2902–2911 (2017)
35. Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M.: Using mobile phones to determine transportation modes. TOSN **6**(2), 13 (2010)
36. Rodríguez, N.D., Wikström, R., Lilius, J., Cuéllar, M.P., Flores, M.D.C.: Understanding movement and interaction: an ontology for kinect-based 3d depth sensors. In: Int. Conf. on UCAmI, pp. 254–261. Springer (2013)
37. Stewart, R., Ermon, S.: Label-free supervision of neural networks with physics and domain knowledge. In: AAAI. vol. 1, pp. 1–7 (2017)
38. Vapnik, V., Izmailov, R.: Learning using privileged information: similarity control and knowledge transfer. JMLR **16**(2023-2049), 2 (2015)
39. Vatavu, R.D., Pentiuc, S.G.: Multi-level representation of gesture as command for human computer interaction. Computing and Informatics **27**(6), 837–851 (2012)
40. Wang, S., Chen, C., Ma, J.: Accelerometer based transportation mode recognition on mobile phones. In: APWCS. pp. 44–46. IEEE (2010)
41. Yu, M.C., Yu, T., Wang, S.C., Lin, C.J., Chang, E.Y.: 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)
42. Zhang, M., Sawchuk, A.A.: Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In: UbiComp. pp. 1036–1043 (2012)
43. Zheng, Y., Xie, X., Ma, W.Y.: Geolife: A collaborative social networking service among user, location and trajectory. IEEE Data Eng. Bull. **33**(2), 32–39 (2010)