Latent HyperNet: Exploring the Layers of Convolutional Neural Networks

Artur Jordao, Ricardo Kloss, William Robson Schwartz
Smart Surveillance Interest Group, Computer Science Department
Universidade Federal de Minas Gerais, Brazil
Email: {arturjordao, rbk, william}@dcc.ufmg.br

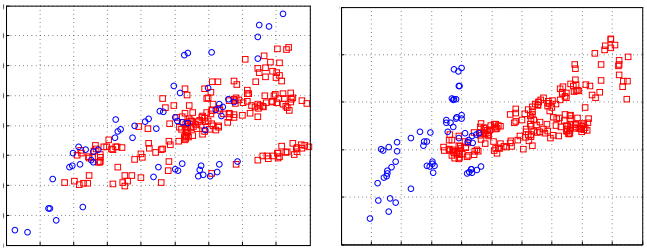
***Abstract*—Since Convolutional Neural Networks (ConvNets) are able to simultaneously learn features and classifiers to discriminate different categories of activities, recent works have employed ConvNets approaches to perform human activity recognition (HAR) based on wearable sensors, allowing the removal of expensive human work and expert knowledge. However, these approaches have their power of discrimination limited mainly by the large number of parameters that compose the network and the reduced number of samples available for training. Inspired by this, we propose an accurate and robust approach, referred to as *Latent HyperNet* (LHN). The LHN uses feature maps from early layers (hyper) and projects them, individually, onto a low dimensionality (latent) space. Then, these latent features are concatenated and presented to a classifier. To demonstrate the robustness and accuracy of the LHN, we evaluate it using four different network architectures in five publicly available HAR datasets based on wearable sensors, which vary in the sampling rate and number of activities. We experimentally demonstrate that the proposed LHN is able to capture rich information, improving the results regarding the original ConvNets. Furthermore, the method outperforms existing state-of-the-art methods, on average, by 5.1 percentage points.**

I. Introduction

Human activity recognition (HAR) has received great attention in the past decade since it is fundamental to healthcare, homeland security and smart environments applications. In particular, human activity recognition based on wearable sensors has attracted the attention of the research community mainly due to easy acquisition and processing of the data [1], [2], [3].

Recent technological advances have allowed this task to migrate from dedicated wearable sensors to sophisticated devices such as smartphones and smartwatches. Besides, these advances also have enabled the use of different sensors (e.g., accelerometer, gyroscope and barometer), which allow performing an improved activity recognition. However, HAR based on wearable sensors faces a large number of challenges, for instance, treatment of noise and definition of discriminative features able to distinguish the different categories of activities [4], [5].

Many works have demonstrated that the feature extraction process is the most important step regarding HAR based on wearable sensors, since finding adequate features significantly improves the activity recognition rate [6], [4], [3]. In the past decade, handcrafted features, such as average, standard deviation and Fourier-based descriptors, were employed to extract higher level descriptions from raw signal data to be presented to a classifier. However, this paradigm requires expert knowledge and expensive human work.



(a) Projection using the feature maps at the last layer.

(b) Projection using the features maps from all layers.

Fig. 1. Projection of two activities onto the two first components of the Partial Least Squares using our ConvNet1 (best viewed in color). The feature space is better separated when features from early and the last layers are combined. This happens due to multi-scale information provided by the low-level information (shallow layers) with the refined information (deep layers).

Recent works employ Convolutional Neural Network (ConvNets) to learn features and the classifier simultaneously. These approaches have achieved better results than works based on handcrafted features and 1D convolutions [7], [8], [9], [10]. On the other hand, ConvNets have some delicate points, such as the need of a large number of training samples, the sensibility to unbalanced data and the large number of parameters to be estimated. As a consequence, the accuracy achieved by ConvNets might be compromised.

To explore the advantages while facing the drawbacks of the aforementioned issues, in this work we propose an accurate and robust approach, referred to as *Latent HyperNet* (LHN). The LHN relies on the hypothesis that early layers composing a ConvNet provide strong, or complementary, clues to better discriminate the categories of activities. In other words, the combination of low-level information, i.e., shallow layers, with the refined information, i.e., deep layers, might help to better distinguish the activities. Figure 1 illustrates our hypothesis, showing that the feature space is better separated when features from early layers are combined with features from the last layer, as seen in Figure 1(b).

Similar ideas were employed by Kong et al. [11] in the context of object detection, where the authors incorporated features from early layers and employed them jointly to learn a classifier, instead of using only the last convolutional

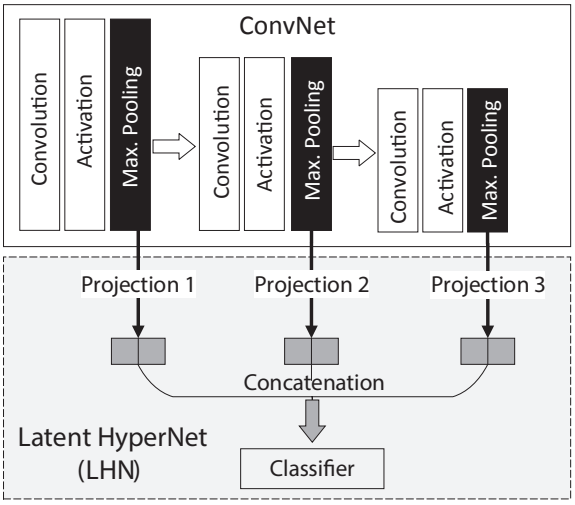


Fig. 2. Process to build the Latent HyperNet. After each max-pooling layer, we apply a dimensionality reduction technique to project the features onto a low dimensional space. For this purpose, we use the Partial Least Squares. Then, we concatenate and present the features to a classifier.

layer (which is the traditional approach). However, because features from earlier layers have a high dimensional space, the computational cost and the number of parameters increased significantly, making its use impracticable in wearable systems due to memory constraints. On the other hand, our proposed latent hypernet explores, iteratively, all the layers (in this work the max-pooling layers) that compose a ConvNet in an efficient way, enabling us to extract richer information to improve recognition rates for HAR associated with wearable sensors.

The proposed method consists of extracting and projecting features of each layer, individually, onto a latent space using Partial Least Squares [12], [13], a dimensionality reduction and regression technique widely employed in social sciences and chemometrics to predict a set of dependent variables from a (large) set of independent variables [13]. Then, we concatenate and present these features (in the latent space) to a classifier. Figure 2 illustrates the LHN approach. It is important to emphasize that we neither modify the design nor the learned weights of the ConvNet during the process to build the LHN. This enables the method to be easily adaptable to any network.

The development of this work presents the following contributions: (1) three accurate ConvNets architectures that explore the signal content in different ways, outperforming previous ConvNet architectures specific to HAR based on wearable sensors and serving as insight to future works, with the intention to build ConvNets architectures; (2) evidences that early layers that compose a ConvNet provide discriminative information that can increase the recognition rate when properly combined; (3) a novel method, the Latent HyperNet (LHN), that effectively combines the network layers, without the requirement of re-training the network.

To validate the robustness of the proposed method regarding the employed ConvNet, we evaluate it on three ConvNets architectures proposed in this work and in a ConvNet proposed by Chen and Xue [7] to HAR based on wearable sensors. We evaluate the recognition rate achieved by the LHN method using five publicly available HAR datasets and compare it with state-of-the-art methods, which are ConvNet architectures built for tackling the problem of wearable data.

Our experiments show that the proposed LHN is able to capture rich information which improves the results regarding the original ConvNet without compromising the computational cost at the prediction stage. To the best of our knowledge, the LHN achieves notable enhancements, since many efforts have been done to achieve smaller improvements in human activity recognition based on wearable sensor data [8], [9]. Furthermore, the proposed approach outperforms existing state-of-the-art methods. On the datasets USCHAD and WISDM, our method achieves a recognition rate of 83.8% and 88.0%, respectively. Additionally, on the datasets UTD-MHAD1 and UTD-MHAD2, our method is able to obtain a recognition rate of 50.1%, 75.3%, respectively. These results outperform existing state-of-the-art methods in 5.1 percentage points, on average.

The remainder of this paper is structured as follows. Section II reviews the works that combine early layers from ConvNets and the state-of-the-art methods in HAR associated with wearable data. Next, Section III explains our proposed LHN method. Finally, Section IV and V present our experimental results and concluding remarks, respectively.

II. Related Work

The use of early layers to improve the data representation has been explored in the context of object detection. However, these approaches demand a high computational cost[1] since the combination of the layers is performed by a deconvolution or 1×1 convolution layer, which makes it usage unfeasible in wearable systems

Bell et al. [14] proposed a method to extract contextual and multi-scale information. For this purpose, their method, called Inside-Outside Net, combines previous layers from ConvNet. This combination is performed by normalizing and presenting the features maps to a 1×1 convolution, a trick employed to avoid the high dimensionality. The authors noticed that this combination aids, mainly, in the detection of small objects, which require higher spatial resolution produced by lower-level layers. Instead of combining the layers using a 1×1 convolution, as in [14], Kong et al. [11] used max-pooling and deconvolution layers to re-scale all feature maps to the same size. Then, the authors presented these re-scaled features maps to a fully-connected layer.

The idea behind using 1×1 convolution and deconvolution in [14] and [11] is to reduce the data dimensionality and compress the feature maps into a uniform space (to properly combine them), respectively. However, the number of parameters of the network increases considerably, since these new

[1]The computation cost increases since a larger number of parameters must be learned by the ConvNet due to addition of 1×1 convolution and deconvolution layers.

layers must be learned. In our proposed method, on the other hand, the number of parameters of the network is not modified (since we do not change the network) and the problem of the high dimensional of the data is easily handled by the PLS.

Previous works on human activity recognition based on wearable sensors have showed that ConvNets approaches are able to provide significant improvements. Chen and Xue [7] proposed a sophisticated ConvNet to classify the different categories of activities from raw signal. Their proposed ConvNet consists of three convolutional layers, where each layer is followed by a 2×1 max-pooling layer. Similar to [7], Jiang and Yin [8] proposed a shallow ConvNet with only two layers. However, to improve the activity representation, Jiang and Yin [8] suggested a method, called signal image, in which once the signal image is generated, a discrete Fourier transform is applied to it, producing a novel matrix which is presented as the ConvNet input.

Different from [7] and [8], Ha et al. [15] proposed a multimodal ConvNet consisting of convolutional filters and max-pooling of sizes 3×3 and 5×5, respectively. The filters at the first layer are learned separately to each heterogeneous modality (e.g., accelerometer and gyroscope). For this purpose the authors introduced zero-padding between the different modalities, in this way, the modalities are not merged during the convolution process. Following [15], Ha and Choi [9] introduced zero-padding at the second layer to separate the filters of each modality in all layers from ConvNet. An interesting aspect of their work is that the authors demonstrated that ConvNets (2D convolutions) are more suitable to HAR based on wearable sensors than 1D convolutions. They showed that the models generated by ConvNets are smaller and have, nearly, 4.7 times fewer parameters than the 1D convolutions, which is an important issue since mobile devices have limited memory and computational power.

A drawback in [15], [9], however, is that due to their proposed ConvNet architecture (design of the filters), it is possible to execute these ConvNets only in datasets where multiple devices are used to capture the data (e.g., devices placed on different body parts to acquire accelerometer and gyroscope). This restriction happens because the convolution process produces feature maps smaller than the input provided to it and its dimension can reach invalid values (i.e., zero) in deep ConvNets. Hence, in data provided by single devices the input samples are small, which contributes to the above-mentioned limitation.

In contrast to the aforementioned studies, our work focuses on enhancing the data representation generated from ConvNets, which enables improvements independently of the sensor data and ConvNet architecture. This data representation improvement is achieved by multi-scale information (obtained through the combination of shallow with deep layers from ConvNet), which is enhanced by our LHN method.

III. Proposed Approach

In this section, we start by briefly describing the Partial Least Squares (PLS) used for projecting the features maps in our method. Afterwards, we introduce the approach to generate our proposed Latent Hypernet approach.

The PLS is a dimensionality reduction technique which projects the high dimensional space onto a latent space, where the covariance between the feature and its label is maximized. The PLS technique works as follows. Let $X \subset R^m$ be a matrix of independent variables representing the samples (activities) in m-dimensional space (originated by the layer from a ConvNet). Let y be the matrix dependent variables denoting the class label in a k-dimensional space, where k represents the number of categories of activities. The PLS projects X onto a new c-dimensional space (where c is the single parameter of the method), $X' \subset R^c$, in terms of $X' = XW$, where W is a weight matrix and can be computed, iteratively, using the NIPALS algorithm [12], Algorithm 1.

The NIPALS algorithm computes a column of W at each iteration (step 3 in Algorithm 1), which represents the maximum covariance between X and y. It should be mentioned that in Algorithm 1, the convergence step is achieved when, from an iteration to another, no change occurs in w_a (defined in the algorithm). Additionally, we can use a fixed number of steps (i.e., 20) to ensure the method stop in step 7. Finally, before presenting the matrices X and y for the NIPALS algorithm they are normalized to operate in the same scale (transformed into Z-scores). Thereby, we ensure that the output of different layers works in a common scale.

Algorithm 1: NIPALS Algorithm.

Input : m-dimensional data X, label matrix y
Input : Number of components c
Ouput: Weight matrix W

1 **for** $a = 1$ ***to*** c **do**
2 randomly initialize $u \in \mathbb{R}^{n \times 1}$
3 $w_a = \frac{X^T u}{\|X^T u\|}$, where $w_a \in W$
4 $t_a = X w_a$
5 $q_a = \frac{y^T t_a}{\|y^T t_a\|}$
6 $u = y q_a$
7 Repeat steps $3 - 6$ until convergence
8 $p_a = X^T t_a$
9 $X = X - t_a p_a^T$
10 $y = y - t_a q_a^T$
11 **end**

Note that other dimensionality reduction methods, such as principal component analysis or linear discriminant analysis, could be used to find the projection matrix W, however, this work employs PLS since many studies showed that it robust to unbalanced and multiclass problems [16], [17], [18].

The generation of the LHN works as follows. First, we present all the training samples to the network and after each

max-pooling[2] layer i, we use its feature maps to learn a PLS model. Clearly, we can notice that each max-pooling layer will have a PLS model associated with it. Even though we could have concatenated the features provided by all max-pooling layers and then project the concatenated features to the latent space, generating a single PLS model, the memory consumption would increase significantly since the result of this concatenation is a high dimensional space and wearable devices have limited memory. Therefore, we perform the projection iteratively, layer by layer. In addition, iteratively projecting the layer enables the method to be efficient in deeper networks. Finally, we concatenate and present to a classifier all the latent features (LF) produced by each max-pooling layer i-th. Algorithm 2 presents the steps of the described process.

Algorithm 2: Latent HyperNet Algorithm.

Input : ConvNet, $LF = \{\}$
Output: Concatenated Latent Features (LF)

foreach *max-pooling layer* $\in ConvNet$ **do**
  $X_i =$ features maps from max-pooling$_i$
  Find W_i using Algorithm 1
  $X' = X_i W_i$ (Projection step in Figure 2)
  $LF = LF \cup X'$ (Concatenation step in Figure 2)
end

IV. Experimental Results

In this section, we first present the datasets employed to validate our proposed LHN approach. Then, we describe the experimental setup and our proposed ConvNets, respectively. Finally, we show the improvements achieved by LHN, the importance of the stage of dimensionality reduction, the computational cost of the method and compare our approach with other state-of-the-art methods.

Datasets. Instead of describing each dataset individually, we summarize the main features of them in Table I. From this table, it is possible to observe that the datasets vary in terms of the sampling rate (Freq.), number of available sensors and activities. In this way, we can examine the robustness of the methods regarding the high variation in the essence of the data.

Experimental Setup. Throughout the experiments, we adopt the 10-fold cross-validation protocol, which is a standard protocol applied to HAR based on sensors [23], [4], [24]. To report the results, we use the recall metric [25], which we also refer to as recognition rate.

The input samples to the ConvNets are temporal windows generated from raw signal. These windows are produced by dividing the signal into subparts (windows) and considering each subpart as an entire activity. Formally, we can define a temporal window in terms of

$$w = [s_{k-t}, ..., s_{k-2}, s_{k-1}, s_k]^\top, \quad (1)$$

where k denotes the current sample captured by the sensor and t denotes the temporal sliding window size. The windows that do not fit within the temporal window are dropped. For a detailed discussion regarding this procedure we recommend [8], [26]. In addition, following [27], [26], we segment the raw signal using temporal sliding window of 5 seconds (value t in Equation 1). However, since the activities of UTD-MHAD [22] dataset have the duration of 2-3 seconds, to this dataset, we use a window of 1 second, which limits the use of deeper architectures and large convolutional kernels. This happens because the convolution process generates feature maps smaller than the input provided to it and its size can reach zero in deep ConvNets.

TABLE I
Main features of each dataset. The available sensors in each dataset vary from Accelerometer (Acc) to Gyroscope (Gyro), and Magnetometer (Mag).

Dataset	Freq. (Hz)	#Sensors	#Activities
USCHAD [19]	100	2 (Acc and Gyro)	12
WISDM [20]	20	1 (Acc)	7
MHEALTH [21]	50	3 (Acc, Gyro, Mag)	12
UTD-MHAD1 [22]	50	2 (Acc, Gyro)	21
UTD-MHAD2 [22]	50	2 (Acc, Gyro)	5

Convolutional Neural Networks. As mentioned in the previous sections, to evaluate the LHN robustness regarding the ConvNets, we propose three different ConvNets (*ConvNet1*, *ConvNet2* and *ConvNet3*), which vary in the number of filters, kernel dimensions (shape) and depth of the network. Table II shows the architectures of our proposed ConvNets. The first column in this table shows the depth of the network (Figure 2

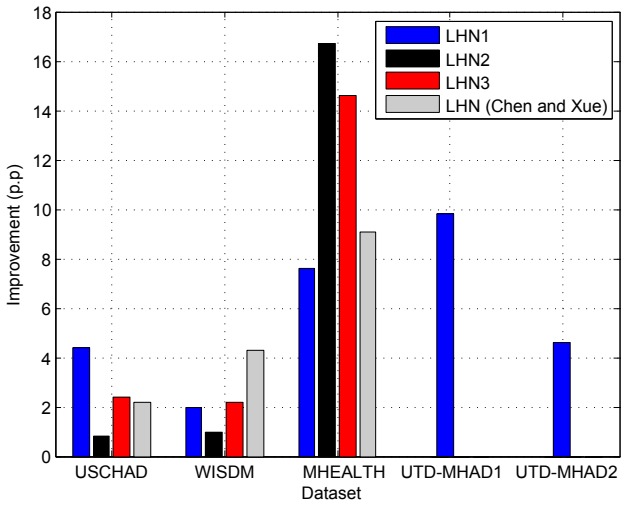


Fig. 3. Improvements achieved by our LHN method compared to the original ConvNet (best viewed in color).

[2]We have selected the max-pooling layer since it is robust to spatial shift.

TABLE II
CONFIGURATIONS OF OURS PROPOSED CONVNETS.

	#Conv. Layers (Depth)	# Filters per Layer	Kernel Shape (heigh $\times$ width) per Layer
ConvNet1	2	$24, 32$	$12 \times 2, 12 \times 2$
ConvNet2	3	$24, 32, 40$	$6 \times 1, 8 \times 1, 10 \times 1$
ConvNet3	4	$24, 32, 40, 48$	$12 \times 1, 12 \times 1, 6 \times 1, 2 \times 1$

illustrates a ConvNet of depth 3, i.e., three convolutional layers). In particular, since each convolutional layers is followed by a max-pooling layer (with kernel of 2×1), the first column also indicates the number of max-pooling layers, i.e., the number of PLS models that compose the LHN.

Latent HyperNet. Since the essence of the LHN is the dimensionality reduction step, we need to find the best number of components, c, to the PLS. For this purpose, we range c from 1 to 20 and evaluate the results achieved using the USCHAD dataset [19], where c equals to 19 yielded the best result. We use the same value on the other datasets. In addition, to render a fair comparison and show the improvement obtained by the LHN, we use the same classifier employed by the original ConvNet, which is a fully connected layer followed by a SoftMax classifier. In this way, our LHN is not biased by the classifier.

Figure 3 shows the improvements (difference between the recall achieved by the ConvNet using our LHN and the one without using the LHN) achieved by the LHN method regarding the employed ConvNet, where the ith LHN represents the LHN using the ith ConvNet. The proposed LHN method was able to increase the activity recognition for all ConvNets. In particular, the LHN was able to improve up to 16.70 percentage points (p.p.) the activity recognition, representing a significant improvement since many efforts have been done to achieve just minor improvements [8], [9].

Considering all datasets evaluated in Figure 3, the proposed LHN was able to improve the ConvNets 1-3, on average, 5.68 p.p., 6.16 p.p. and 6.40 p.p., respectively. Moreover, the use of the LHN enhanced the recognition rate in 6.20 p.p. when employed to the architecture proposed by Chen and Xue [7]. These results reinforce our hypothesis that shallower layers, when properly combined with deeper layers, are able to enhance the discrimination of activities, allowing a better activity recognition.

Although the achieved improvements seem small, many efforts have been done to achieve smaller improvements in HAR based on wearable sensor data. For instance, Catal et al. [6] and Ha and Choi [9] improved the works of Kwapisz et al. [28] and Ha et al. [15] in 2.81 p.p. and 2.19 p.p., respectively. Therefore, our LHN achieves notable enhancements.

Importance of the dimensionality reduction. In this experiment, we show the importance of the dimensionality reduction step in our LHN method. To this end, we measure the results of the LHN without the dimensionality reduction step on the USC-HAD dataset [19].

By removing the dimensionality reduction, the recognition rate decreased 30 p.p. on average. This occurs due to the high dimensionality generated from the concatenation of the feature maps, rendering the learning stage more complex since the network needs to learn a larger number of parameters. On the contrary, by using the dimensionality reduction we generate a low-dimensional feature space, where there are fewer parameters to be learned, which aids the learning stage.

Time issues. Since our method performs a projection after each max-pooling layer (as explained in Section III), it introduces an extra cost to predict the samples. In this experiment, we show that this cost is irrelevant, which enables the LHN method to be computationally efficient compared to the traditional ConvNet. To demonstrate that, we perform a statistical evaluation [29], on the prediction time. In this evaluation, we computed the confidence interval by using a confidence of 95% and estimate the average prediction time by considering 30 executions.

Figure 4 shows the average and the confidence interval of the original ConvNets (red bars) and the ones using the proposed LHN method (gray bars). According to this figure, it is possible to observe that the confidence intervals present overlap, thereby, the methods might be statistically equivalent. To validate this claim, as suggested by Jain [29], we perform a *unpaired t-test* between the methods. On this test, our method has been shown to be equivalent to original ConvNet, which makes the prediction time of the LHN statistically equivalent

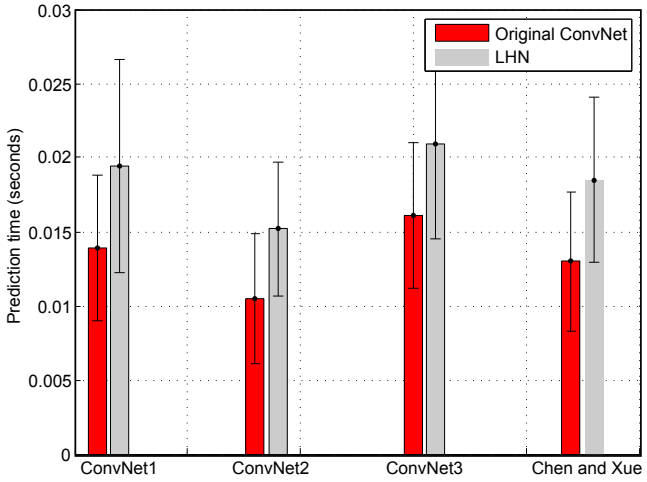


Fig. 4. Average prediction time, lower values are better (best viewed in color). It is possible to note that, our proposed method does not compromise the prediction time, since the time of the LHN is statistically equal to the original ConvNet time.

TABLE III

COMPARISON WITH STATE-OF-THE-ART METHODS. VALUES IN BOLD DENOTE THE TOP 2 BEST METHODS FOR EACH DATASET. CONVNET1-3 DENOTE THE PROPOSED CONVNETS WITHOUT THE EMPLOYMENT OF THE LHN METHOD. THE LAST ROW INDICATES THE CONVNET3 USING THE LHN METHOD, EXCEPT FOR THE DATASETS UTD-MHAD1-2, WHERE WE USE THE LHN WITH THE CONVNET2. CELLS WITH THE SYMBOL − DENOTE THAT IT IS NOT POSSIBLE TO EXECUTE THE CONVNET ON THE RESPECTIVE DATASET, DUE TO ITS ARCHITECTURE.

	MHEALTH [21]	USCHAD [19]	UTD-MHAD1 [22]	UTD-MHAD2 [22]	WISDM [20]
Jiang and Yin [8]	55.6	76.4	**42.0**	70.0	82.3
Chen and Xue [7]	65.7	78.7	-	-	86.0
Ha et al. [15]	67.9	-	-	-	-
Ha and Choi [9]	**84.8**	-	-	-	-
ConvNet1 (Ours)	68.2	80.7	40.3	**70.7**	86.3
ConvNet2 (Ours)	61.8	79.9	-	-	**87.0**
ConvNet3 (Ours)	63.5	**81.4**	-	-	85.8
LHN (Ours)	**78.1**	**83.8**	**50.1**	**75.3**	**88.0**

to the original ConvNet. Therefore, the employment of the LHN method does not compromise the prediction time.

Comparison with the State-of-the-art. As we mentioned earlier, the current state-of-the art results in HAR based on wearable sensors are achieved with methods based on ConvNets [9], [10]. Therefore, our last experiment compares the LHN with such methods. It is important to note that all the methods used in this experiment are ConvNets dedicated to HAR based on wearable data. Moreover, to provide a fair comparison, we re-train all the methods on the same conditions (e.g., number of epochs and training samples). Finally, we do not compare our method with LSTMs-based approaches since, while they presented good results in natural language processing [30] and speech recognition [31], in the context of HAR based on wearable sensors ConvNets have presented superior results [9].

According to Table III, in most of the cases, our proposed ConvNets (even without using LHN) outperform existing ConvNets in the literature. We believe that our ConvNets provide better results due to the convolutional kernel dimensions. For instance, [8], [15] use kernels of 3×3 and 5×5, respectively, which capture a small temporal pattern besides being sensitive to noise by data acquisition. On the other hand, our kernels are able to capture a large temporal relation of the signal and, hence, to be more robust to noise.

In order to compare our LHN with the state-of-the-art methods, we select the LHN using our proposed ConvNet3. However, since UTD-MHAD dataset does not enable deep architectures (as we argued before), we select the LHN using ConvNet1 for this dataset. Table III shows that our proposed method outperforms state-of-the-art methods in activity recognition based on wearable data. For instance, on the datasets USCHAD, UTD-MHAD2 and WISDM we outperformed the state-the-art in 5.1 p.p., 5.3 p.p. and 2.0 p.p., respectively. Moreover, it is important to notice that the recognition rate was reduced drastically on the UTD-MHAD1 dataset due to the large number of activities contained in this dataset. However, our method outperformed the previous best method in 8.1 p.p., demonstrating that our LHN provides a richer data representation. Finally, as can be noticed in Table III, on the MHEALTH dataset the best recognition rate was achieved by the method of Ha and Choi [9]. Additionally, when we apply the LHN method on this ConvNet, we obtain an improvement of 7.3 p.p., outperforming the state-of-the-art once again.

V. CONCLUSIONS

This work presented a robust and accurate method, referred to as Latent HyperNet (LHN), to improve ConvNets applied to HAR based on wearable sensor data. The method individually projects the features from each layer onto a latent space, where a richer representation of these features is obtained. We evaluate the proposed method using different ConvNet architectures and our experiments demonstrated that our method improves the recognition rate regarding the original ConvNet, without compromising its the computational cost at the prediction stage, besides outperforming existing state-of-the art methods. We highlight that many efforts have been done to achieve small improvements in human activity recognition based on wearable sensor data, which reinforces that the LHN produces notable improvements.

Since LHN does not modify the design of the ConvNet, it can be easily adaptable to any network. Therefore, as future work, we intend to apply it to other applications which employ ConvNets, such as image classification. In addition, we plan to employ other dimensionality reduction techniques, such as Linear Discriminant Analysis and Principal Components Analysis, to build the LHNs.

ACKNOWLEDGMENTS

The authors would like to thank the Brazilian National Research Council – CNPq (Grant #311053/2016-5), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project).

REFERENCES

[1] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *in IEEE Sensors Journal*, 2015.

[2] X. Yin, G. Shen, X. Wang, and W. Shen, "Mitigating sensor differences for phone-based human activity recognition," in *SMC*, 2016.

[3] K. Karagiannaki, A. Panousopoulou, and P. Tsakalides, "An online feature selection architecture for human activity recognition," in *ICASSP*, 2017.

[4] M. Shoaib, S. Bosch, O. Incel, H. Scholten, and P. Havinga, "A Survey of Online Activity Recognition using Mobile Phones," *in Sensors*, 2015.

[5] B. Bruno, F. Mastrogiovanni, and A. Sgorbissa, "Wearable Inertial Sensors: Applications, Challenges, and Public Test Benches," *in IEEE Robot. Automat. Mag.*, 2015.

[6] C. Catal, S. Tufekci, E. Pirmit, and G. Kocabag, "On the use of ensemble of classifiers for accelerometer-based activity recognition," *in Applied Soft Computing*, 2015.

[7] Y. Chen and Y. Xue, "A Deep Learning Approach to Human Activity Recognition Based on Single Accelerometer," in *SMC*, 2015.

[8] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *ACM Multimedia Conference*, 2015.

[9] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *IJCNN*, 2016.

[10] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *CoRR*, vol. abs/1707.03502, 2017.

[11] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," in *CVPR*, 2016.

[12] H. Wold, "Partial Least Squares," in *Encyclopedia of Statistical Sciences*. New York, NY, USA: Wiley, 1985, vol. 6, pp. 581–591.

[13] H. Abdi, "Partial least squares regression and projection on latent structure regression (pls regression)," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97–106, 2 2010.

[14] S. Bell, C. L. Zitnick, K. Bala, and R. B. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *CVPR*. IEEE Computer Society, 2016, pp. 2874–2883.

[15] S. Ha, J. Yun, and S. Choi, "Multi-modal convolutional neural networks for activity recognition," in *SMC*, 2015.

[16] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis." in *ICCV*, 2009.

[17] C. E. d. Santos, E. Kijak, G. Gravier, and W. R. Schwartz, "Learning to hash faces using large feature vectors," in *CBMI*, 2015.

[18] R. B. Kloss, A. Jordão, and W. R. Schwartz, "Boosted projection: An ensemble of transformation models," in *CIARP*, 2017.

[19] M. Zhang and A. A. Sawchuk, "Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors," in *UbiComp*, 2012.

[20] J. W. Lockhart, G. M. Weiss, J. C. Xue, S. T. Gallagher, A. B. Grosner, and T. T. Pulickal, "Design considerations for the wisdm smart phone-based sensor mining architecture," in *SIGKDD*, 2011.

[21] O. Baños, R. García, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. R. Ruiz, A. Saez, and C. Villalonga, "mhealthdroid: A novel framework for agile development of mobile health applications," in *IWAAL*, 2014.

[22] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *ICIP*, 2015.

[23] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *MobiCASE*, 2014.

[24] J. C. Quiroz, M. H. Yong, and E. Geangu, "Emotion-recognition using smart watch accelerometer data: preliminary findings," in *UbiComp/ISWC*, 2017.

[25] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *in Journal of Machine Learning Technologies*, 2011.

[26] H. Song, J. J. Thiagarajan, P. Sattigeri, K. N. Ramamurthy, and A. Spanias, "A deep learning approach to multiple kernel fusion," in *ICASSP*, 2017.

[27] D. Morris, T. S. Saponas, A. Guillory, and I. Kelner, "Recofit: using a wearable sensor to find, recognize, and count repetitive exercises," in *CHI*, 2014.

[28] J. R. Kwapisz, G. M. Weiss, and S. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explorations*, vol. 12, no. 2, pp. 74–82, 2010.

[29] R. Jain, *The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling*. John Wiley & Sons, 1990.

[30] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learning Syst.*, 2017.

[31] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.