LATENT HYPERNET: EXPLORING ALL LAYERS FROM CONVOLUTIONAL NEURAL NETWORKS

Artur Jordao, Ricardo Kloss and William Robson Schwartz
Smart Surveillance Interest Group, Computer Science Department
Universidade Federal de Minas Gerais, Minas Gerais, Brazil
{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 space (latent). 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 networks architectures in five publicly available HAR datasets based on wearable sensors, which vary in the sampling rate and number of activities. Our experiments demonstrate that the proposed LHN is able to produce rich information, improving the results regarding the original ConvNets. Furthermore, the method outperforms existing state-of-the-art methods.

1. INTRODUCTION
Human activity recognition (HAR) has received 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 [1,3].

Recent technological advances allowed this task to migrate from dedicated wearable sensors to sophisticated devices (smartphones and smartwatches). Besides, these advances also enabled the use of different sensors (e.g., accelerometer, gyroscope and barometer), which provide 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 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 [3,4,6]. 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.

Recent works employ Convolutional Neural Network (ConvNets) to learn features and the classifier simultaneously. For instance, Chen and Xue [7] proposed a sophisticated ConvNet to classify the different categories of activities using 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. Once the signal image is generated, a discrete Fourier transform is applied to it, producing a new input matrix which is presented to the ConvNet. In contrast to [7,8], Ha et al. [9] proposed a multi-modal ConvNet consisting of convolutional filters and max-pooling of 3 × 3 and 5 × 5, respectively. The filters at the first layer are learned apart from each heterogeneous modality (e.g., accelerometer and gyroscope). For this purpose, the authors introduced zero-padding between the different modalities. Following [9], Ha and Choi [10] introduced zero-padding at the second layer to separate the filters of each modality in all layers from ConvNet.

The ConvNet-based approaches mentioned above have achieved better results than works based on handcrafted features [4,8,11] and 1D convolutions [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, helps 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. [12] 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 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 LHN explores, iteratively, all the layers (in this work the max-pooling layers) that compose a ConvNet in an efficient way, extracting richer information to improve recognition rates for HAR based on wearable sensors.

The proposed method consists in extracting and projecting features of each layer, individually, onto a latent space using Partial Least Squares [13, 14]. 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 ConvNet 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 these layers.

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 produce rich information which improves the results regarding the original ConvNet. Furthermore, the method outperforms existing state-of-the-art methods.

2. PROPOSED METHOD

In this section, we start by describing, briefly, the Partial Least Squares (PLS). Afterwards, we introduce the pipeline to generate our Latent Hypernet approach.

Let $X \subset R^m$ be a matrix representing the samples (activities) in $m$-dimensional space (originated by the layer from a ConvNet). Let $y$ be the matrix with 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 [13]. The NIPALS algorithm, at each iteration, computes a column of $W$, which represents the maximum covariance between $X$ and $y$. Due to lack of space, we recommend [13, 14] for a detailed mathematical definition.

Note that other dimensionality reduction methods could
be used to find the projections \( W_{ith} \), e.g., principal component analysis, however, this work employs PLS since many studies showed that it provides good results to unbalanced and multiclass problems [13, 15, 16].

Now, we explain the generation of LHN. First, we present all the training samples to the network and after each max-pooling 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. Moreover, observe that we could concatenate all the features provided by all max-pooling layers and then project the concatenated features to the latent space, hence, we would have only one PLS model. However, the memory consumption increases significantly since the result of this concatenation is an high dimensional space and wearable devices have limited memory. Therefore, we perform the projection iteratively, layer by layer. Finally, we concatenate and present to a classifier all the latent features (LF), produced by each max-pooling layer \( ith \).

Algorithm 1: Latent HyperNet Algorithm.

| Input: ConvNet, \( LF = \{ \} \) |
| Output: Concatenated Latent Features (LF) |
| foreach max-pooling layer \( \in \text{ConvNet} \) do |
| \( X_{ith} = \text{features maps from max-pooling}_{ith} \) |
| \( \text{Find } W_{ith} \text{ using PLS via NIPALS algorithm} \) |
| \( LF = LF \cup X_{ith}W_{ith} \) |
| end |

3. EXPERIMENTAL RESULTS

In this section, we first introduce the details according to our proposed ConvNets. Then, we show the improvements achieved by LHN and compare it with the state-of-the-art.

**Experimental Setup.** Throughout the experiments, we adopt the 10-fold cross-validation protocol, which is a standard protocol applied to HAR based on sensors [4, 17, 18]. To report the results, we use the recall metric [19], which we also refer to as recognition rate. In addition, following [20, 21], we segment the raw signal using temporal sliding window of 5 seconds (due to space issues we recommend [8, 21] for a detailed discussion regarding this procedure). 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 deep 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.

**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 1 shows the architectures of our proposed ConvNets. The first column in this table shows the depth of the network (Figure 2 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 \( 2 \times 1 \)), the first column also indicates the number of max-pooling layers, i.e., the number the 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 [24], where \( c \) equals to 19 yielded the best result. We use this 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. Our LHN method was able to increase the activity recognition for all ConvNets. In particular, LHN was able to improve up to 16.70 percentage points (p.p.) for the activity recognition, representing a significant improvement since many efforts have been done to achieve just minor improvements [8, 10].

Considering all datasets evaluated in Figure 2 the proposed LHN was able to improve the ConvNets 1-3, on average, 5.68, 6.16, 6.40 p.p., respectively. Moreover, the use of LHN enhanced the recognition rate in 6.20 p.p. when employed to the architecture proposed by [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
Table 1. Configurations of our proposed ConvNets.

| ConvNet | #Conv. Layers (Depth) | # Filters per Layer | Kernel Shape (height \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 |

Table 2. Comparison with state-of-the-art methods. Values in bold denote the top 2 best methods for each dataset. Cells with the symbol – denote that it is not possible to execute the ConvNet on the respective dataset, due to its architecture.

| Method | MHEALTH [23] | USCHAD [24] | UTD-MHAD1 [22] | UTD-MHAD2 [22] | WISDM [25] |
|--------|--------------|-------------|-----------------|-----------------|-------------|
| 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. [9]  | 67.9 | - | - | - | - |
| Ha and Choi [10] | 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 |

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 [10] improved the works of Kwapisz et al. [26] and Ha et al. [9] 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 [24]. By removing the dimensionality reduction, the recognition rate decreased, on average, 30 p.p. 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.

**Comparison with the State-of-the-art.** As we mentioned before, the current state-of-the-art results in HAR based on wearable sensors are achieved with methods based on ConvNets [10-11]. 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).

According to Table 2, in the most of 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][9] use kernels of $3 \times 3$ and $5 \times 5$, 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.

Finally, 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 2 shows that our LHN outperforms state-of-the-art methods in activity recognition based on wearable data. 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.

4. 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 and outperforms existing state-of-the-art methods. 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.
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).

5. REFERENCES

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

[2] Xizhe Yin, Gary Shen, Xianbin Wang, and Weiming 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] Muhammad Shoaib, Stephan Bosch, Ozlem Incel, Hans Scholten, and Paul Havinga, “A Survey of Online Activity Recognition using Mobile Phones,” in Sensors, 2015.

[5] Barbara Bruno, Fulvio Mastrogiovanni, and Antonio Sgorbissa, “Wearable Inertial Sensors: Applications, Challenges, and Public Test Benches,” in IEEE Robot. Automat. Mag., 2015.

[6] Cagatay Catal, Selin Tufekci, Elif Pirmit, and Guner Kocabag, “On the use of ensemble of classifiers for accelerometer-based activity recognition,” in Applied Soft Computing, 2015.

[7] Yuqing Chen and Yang Xue, “A Deep Learning Approach to Human Activity Recognition Based on Single Accelerometer,” in SMC, 2015.

[8] Wenchao Jiang and Zhaozheng Yin, “Human activity recognition using wearable sensors by deep convolutional neural networks,” in ACM Multimedia Conference, 2015.

[9] Sojeong Ha, Jeong-Min Yun, and Seungjin Choi, “Multi-modal convolutional neural networks for activity recognition,” in SMC, 2015.

[10] Sojeong Ha and Seungjin Choi, “Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors,” in IJCNN, 2016.

[11] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu, “Deep learning for sensor-based activity recognition: A survey,” CoRR, vol. abs/1707.03502, 2017.

[12] Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun, “Hypernet: Towards accurate region proposal generation and joint object detection,” in CVPR, 2016.

[13] H. Wold, “Partial Least Squares,” in Encyclopedia of Statistical Sciences, vol. 6, pp. 581–591. Wiley, New York, NY, USA, 1985.

[14] Herv 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, 2010.

[15] William Robson Schwartz, Aniruddha Kembhavi, David Harwood, and Larry S. Davis, “Human detection using partial least squares analysis,” in ICCV, 2009.

[16] Cassio Elias dos Santos, Ewa Kijak, Guillaume Gravier, and William Robson Schwartz, “Learning to hash faces using large feature vectors,” in CBMI, 2015.

[17] Ming Zeng, Le T. Nguyen, Bo Yu, Ole J. Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang, “Convolutional neural networks for human activity recognition using mobile sensors,” in MobisCASE, 2014.

[18] Juan C. Quiroz, Min Hooi Yong, and Elena Geangu, “Emotion-recognition using smart watch accelerometer data: preliminary findings,” in UbiComp/ISWC, 2017.

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

[20] Dan Morris, T. Scott Saponas, Andrew Guillory, and Ilya Kelner, “Recofil: using a wearable sensor to find, recognize, and count repetitive exercises,” in CHI, 2014.

[21] Huan Song, Jayaraman J. Thiagarajan, Prasanna Sat tigeri, Karthikeyan Natesan Ramamurthy, and Andreas Spanias, “A deep learning approach to multiple kernel fusion,” in ICASSP, 2017.

[22] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz, “UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor,” in ICIP, 2015.

[23] Oresti Baños, Rafael García, Juan A. Holgado-Terriza, Miguel Damas, Héctor Pomares, Ignacio Rojas Ruiz, Alejandro Saez, and Claudia Villalonga, “mhealthdroid: A novel framework for agile development of mobile health applications,” in IWAAL, 2014.

[24] Mi Zhang and Alexander A. Sawchuk, “Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors,” in UbiComp, 2012.
[25] Jeffrey W. Lockhart, Gary M. Weiss, Jack C. Xue, Shaun T. Gallagher, Andrew B. Grosner, and Tony T. Pulickal, “Design considerations for the wisdm smart phone-based sensor mining architecture,” in SIGKDD, 2011.

[26] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel Moore, “Activity recognition using cell phone accelerometers,” SIGKDD Explorations, vol. 12, no. 2, pp. 74–82, 2010.