

# Improving Deep Learning for HAR with shallow LSTMs

Marius Bock

University of Siegen  
Siegen, North Rhine-Westphalia, Germany  
marius.bock@uni-siegen.de

Michael Moeller

University of Siegen  
Siegen, North Rhine-Westphalia, Germany  
michael.moeller@uni-siegen.de

Alexander Hölzemann

University of Siegen  
Siegen, North Rhine-Westphalia, Germany  
alexander.hoelzemann@uni-siegen.de

Kristof Van Laerhoven

University of Siegen  
Siegen, North Rhine-Westphalia, Germany  
kvl@eti.uni-siegen.de

## ABSTRACT

Recent studies in Human Activity Recognition (HAR) have shown that Deep Learning methods are able to outperform classical Machine Learning algorithms. One popular Deep Learning architecture in HAR is the DeepConvLSTM. In this paper we propose to alter the DeepConvLSTM architecture to employ a 1-layered instead of a 2-layered LSTM. We validate our architecture change on 5 publicly available HAR datasets by comparing the predictive performance with and without the change employing varying hidden units within the LSTM layer(s). Results show that across all datasets, our architecture consistently improves on the original one: Recognition performance increases up to 11.7% for the F1-score, and our architecture significantly decreases the amount of learnable parameters. This improvement over DeepConvLSTM decreases training time by as much as 48%. Our results stand in contrast to the belief that one needs at least a 2-layered LSTM when dealing with sequential data. Based on our results we argue that said claim might not be applicable to sensor-based HAR.

## CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**.

## KEYWORDS

Human Activity Recognition, Deep Learning, CNN-LSTMs

### ACM Reference Format:

Marius Bock, Alexander Hölzemann, Michael Moeller, and Kristof Van Laerhoven. 2021. Improving Deep Learning for HAR with shallow LSTMs. In *2021 International Symposium on Wearable Computers (ISWC '21)*, September 21–26, 2021, Virtual, USA. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3460421.3480419>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
*ISWC '21, September 21–26, 2021, Virtual, USA*

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-8462-9/21/09...\$15.00  
<https://doi.org/10.1145/3460421.3480419>

## 1 INTRODUCTION

Physical activities play a crucial role in the way we structure our lives. Which activity, and how it is performed, can reveal a person's intention, habit, fitness, and state of mind; it is therefore not surprising that a range of research fields, from cognitive science to healthcare, display a growing interest in the machine recognition of human activities, also known as Human Activity Recognition (HAR) [2]. Deep Learning methods have in the past decade shown to outperform classical Machine Learning algorithms (e.g., [5, 19, 35]) and, as a product of this success, have led to studies investigating the effectiveness of Deep Learning in HAR (e.g., [13, 26]).

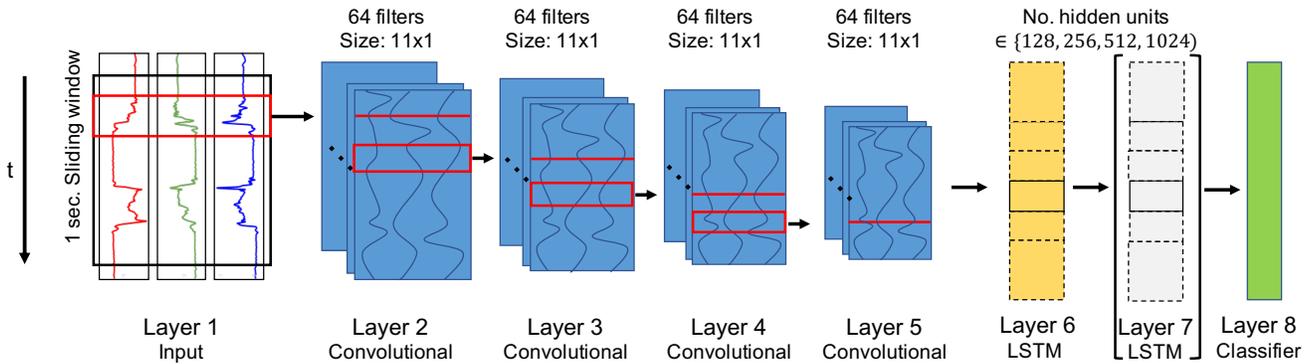
One of the most popular Deep Learning architectures for HAR is the DeepConvLSTM which was proposed by Ordonez et al. in [26]. The suggested architecture combines both recurrent and convolutional layers and received state-of-the-art results on both the Opportunity [29] and the Skoda Mini Checkpoint [43] dataset. The original DeepConvLSTM architecture employs a 2-layered LSTM with 128 hidden units. A common belief, as e.g. stated by Chen et al. [3], based on the findings of Karpathy et al. [17], is that one requires at least a 2-layered LSTM when dealing with sequential data. With this paper, we aim at challenging this belief and suggest that re-examining the architecture of the DeepConvLSTM by employing a one-layered LSTM, has considerable benefits.

Our paper's contributions are threefold:

- (1) We show that an altered DeepConvLSTM architecture with a one-layered LSTM overall outperforms architectures employing a two-layered LSTM, by validating our claim using the Opportunity dataset [29] as seen in [26], as well as 4 other popular HAR datasets [28, 32, 33, 36].
- (2) Using our suggested architecture change, we reduce the number of learnable parameters within the DeepConvLSTM and thus are able to decrease training time significantly.
- (3) We provide our PyTorch-based architecture, experiment scripts and log files in a GitHub repository for others to replicate these findings and continue such analyses.

## 2 RELATED WORK

*Deep Learning in HAR.* The predictive performance of classical Machine Learning approaches highly relies on sophisticated, hand-crafted features [27]. In the last decade, Deep Learning has shown to outperform classical Machine Learning algorithms in many areas, e.g. image recognition [7, 35, 37], speech recognition [14, 19, 23, 31] and Natural Language Processing [1, 5, 15, 34]. Much of this success



**Figure 1: Illustration of the suggested change to the DeepConvLSTM [26] architecture. The change involves removing the second LSTM layer (Layer 7). During experiments we either include (original architecture) or exclude Layer 7 (our suggestion) and vary the amount of hidden units within the LSTM layer(s) (i.e. Layers 6 and 7) to be either 128, 256, 512 or 1024.**

can be accredited to the fact that Deep Learning does not require manual feature engineering, but is able to automatically extract discriminative features from raw data input [25]. The advantage of being able to apply algorithms on raw data and not being dependent on handcrafted features has led to studies investigating the effectiveness of Deep Learning in HAR, which e.g. suggested different architectures [4, 10–12, 22, 26, 30, 40–42], evaluated the generality of architectures [13] and assessed the applicability in real-world scenarios [9, 12, 16, 21].

*DeepConvLSTM.* One popular HAR Deep Learning architecture is the DeepConvLSTM which combines both convolutional and Long-Short-Term-Memory (LSTM) layers [26]. By combining both types of layers the network is able to automatically extract discriminative features and model temporal dependencies. The idea of Ordonez et al. [26] to combine recurrent and convolutional layers served as the basis for several further improvements in subsequent works. For example, Murahari and Plötz [24] improved results by appending attention layers to the original architecture. Xi et al. [38] improved the performance of the original DeepConvLSTM architecture by adding dilated convolution layers in addition to normal convolution layers. Demonstrating a different way of combining recurrent and convolutional layers, Xu et al. [40] proposed *InnoHAR* which combines Inception modules based on GoogLeNet [35] and Gated Recurrent Unit (GRU) layers. Xia et al. [39] suggested to first apply a 2-layered LSTM whose output is then fed to two convolution layers and extended the architecture with a max pooling, global average pooling and batch normalization layer. Kim and Cho [18] make use of the architecture proposed by [44] which, similar to our approach, employs a one-layered LSTM. To our knowledge the work of Kim and Cho [18] is the only work which applied a one-layered LSTM within a variation of the DeepConvLSTM architecture.

*LSTMs and sequential data.* Upon the experiments conducted by Karpathy et al. [17], Chen et al. claim within their recent survey paper that “the depth of an effective LSTM-based RNN needs to be at least two when processing sequential data” [3]. Within this paper we investigated the effect of employing a 1-layered instead of a 2-layered LSTM within the DeepConvLSTM architecture. As Karpathy et al. [17] obtained their results using character-level

language models, i.e. text data, our paper aims at challenging the belief that their claim is applicable to sensor-based HAR.

### 3 METHODOLOGY

Contrary to the belief that one needs at least a two-layered LSTM when dealing with sequential data [17], we propose to change the DeepConvLSTM to have a one-layered LSTM. Figure 1 illustrates the suggested architecture change. We further investigate the varying amount of hidden units in the LSTM layers during experiments.

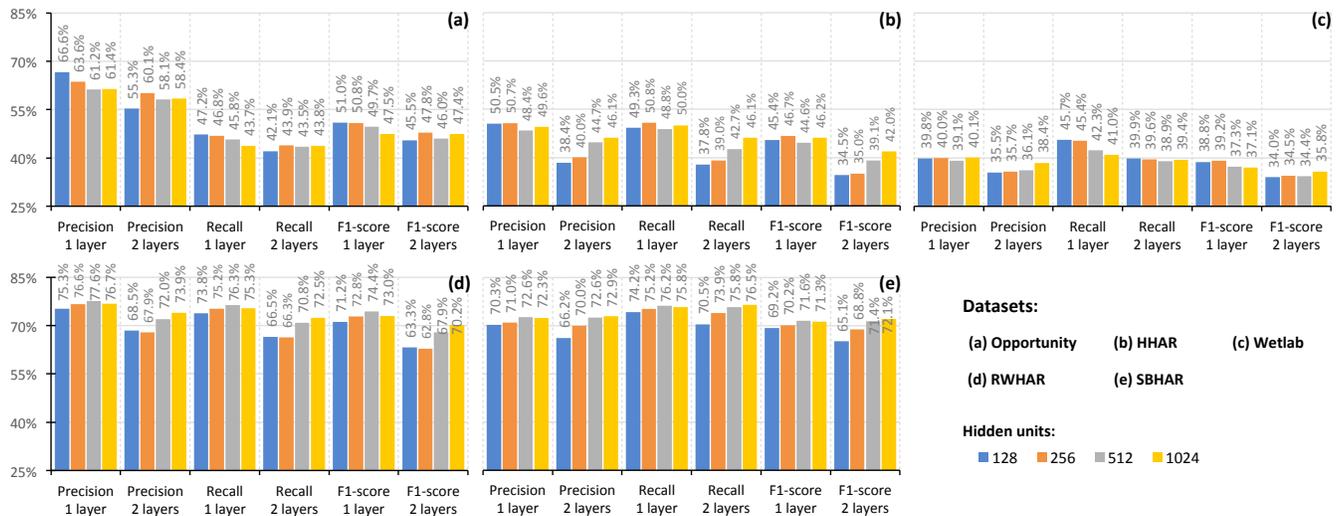
#### 3.1 Datasets

To validate our architecture change, we chose to use the preprocessed version of the Opportunity dataset [29] as used by Ordonez et al. [26], as well as four popular HAR datasets, namely the Wetlab [32], RealWorld HAR (RWHAR) [36], Smartphone-Based Recognition of Human Activities and Postural Transitions (SBHAR) [28] and the Heterogeneity Activity Recognition (HHAR) [33] dataset.

*Opportunity.* The Opportunity datasets consists of 4 individuals performing a set of activities of daily living [29]. For the gesture recognition challenge of the dataset, there are 18 classes which are to be predicted (*open/close door 1 and 2, fridge, dishwasher and drawer 1, 2 and 3, clean table, drink from cup and toggle switch*) as well as a *null* class. During the experiments, we applied the same preprocessing as well as train-test split as suggested by Ordonez et al. [26]. The resulting preprocessed dataset consists of in total 113 feature channels each representing an individual sensor axis from body-worn accelerometers and inertial measurement units (combining 3D accelerometers, gyroscopes and magnetometers).

*Wetlab.* The Wetlab dataset consists of 22 participants performing two DNA extraction experiment within a wetlab environment [32]. During experiments subjects were equipped with a wrist-worn sensing unit capturing 3D acceleration data with a sampling rate of 50Hz. Using recorded video footage, Scholl et al. identified 9 different actions (*cutting, inverting, peeling, pestling, pipetting, pouring, stirring, transfer*) which, along the *null* class, are the target labels to be predicted within this dataset.

*RealWorld (HAR).* The RWHAR dataset contains data of 15 participants performing 8 different activities (*walking upstairs, walking*



**Figure 2: Results obtained on the (a) Opportunity, (b) HHAR, (c) Wetlab, (d) RWHAR and (e) SBHAR datasets using the 8 variations of the DeepConvLSTM [26] architecture, grouped by 1- and 2-layered variations and color-coded by how many hidden units were employed in the LSTM layer(s). Results are given as the average precision, recall and F1-score across 5 runs using a set of 5 varying seeds. The 1-layered architecture variants continuously outperform the 2-layered ones for the HHAR, Wetlab and RWHAR dataset; The worst performing 1-layered variant even still outperforms the best performing 2-layered variant. For the Opportunity dataset, the 1-layered variant outperforms the 2-layered for 128, 256 and 512 hidden units and is on par for 1024 hidden units. For the SBHAR dataset, the 1-layered variant only outperforms for 128 and 256 hidden units and is on par with the 2-layered version when using 512 and 1024 hidden units.**

*downstairs, jumping, lying, standing, sitting, running, walking*) as well as a *null* class [36]. In order to make results more comparable to the Wetlab dataset, we chose to only use 3D acceleration data captured by a wrist-worn sensor which samples the data at 50Hz.

**SBHAR.** The SBHAR dataset consists of 30 participants performing activities of daily living (*standing, sitting, lying, walking, walking downstairs, walking upstairs*) [28]. In addition to the 6 activities, class labels also include 6 postural transitions (*stand-to-sit, sit-to-stand, sit-to-lie, lie-to-sit, stand-to-lie, lie-to-stand*) as well as a *null* class. For the same reason as mentioned with the RWHAR dataset, we only used the raw 3D acceleration sensor data, which was sampled at 50Hz.

**HHAR.** Similar to the RWHAR dataset, the HHAR dataset contains data of 9 human participants performing activities of daily living. There are 6 activities (*biking, sitting, standing, walking, walking upstairs and downstairs*) and a *null* class which are to be predicted [33]. As for the previous datasets, we only use the 3D-acceleration data obtained from a wrist-worn sensor unit, which was sampled at 100Hz.

### 3.2 Training

To justify our architecture and illustrate the effectiveness of said change to it, we compare two variations of the DeepConvLSTM architecture with each other: One equipped with a single-layer LSTM, and one equipped with a two-layer LSTM. We further vary the number of hidden units, more specifically using in both variations either 128, 256, 512 or 1024 hidden units per LSTM layer. This leaves us with a total of 8 variations that were evaluated, for all five

datasets. It is important to note that only the LSTM was altered. Specifications of other layers, such as the number of convolution layers (4), the number of convolution filters (64) and the dropout rate (0.5), were left unchanged from the original DeepConvLSTM architecture.

To train our networks we employ a sliding window approach similar to the one as seen in [26]. To obtain suitable set of hyperparameters we evaluated multiple settings based on results obtained from using the Wetlab dataset as input. We identified a sliding window of 1 second with a 60% overlap to be most suitable. Unlike Ordonez et al. [26] we use the *Adam* optimizer with a smaller weight decay ( $1e^{-6}$ ) and learning rate ( $1e^{-4}$ ) and initialize the network weights using the Glorot initialization [8]. During training we computed the loss using a weighted cross-entropy loss to enable the networks to learn also imbalanced datasets (e.g. Wetlab). Since the focus of our paper lies on evaluating the change to the architecture, we did not perform any hypertuning on the RWHAR, SBHAR and HHAR dataset and thus kept hyperparameters consistent across datasets. Nevertheless, unlike the other 3 datasets, the HHAR was sampled at 100Hz. We thus increased the convolutional filter size by a factor of two to be 21, to maintain the relation between convolutional filter and sliding window size and capture the same amount of information with each filter across all datasets. We used the same set of hyperparameters for the Opportunity dataset as was used for the other datasets to allow a comparison across the different types of activity recognition scenarios, and changed the sliding window size and overlap to be identical with the one employed in [26], i.e. 0.5 seconds with a 50% overlap.

### 3.3 Results

For the Wetlab, RWHAR, HHAR, SBHAR and HHAR dataset, results were obtained using Leave-One-Subject-Out (LOSO) cross-validation. This means that each subject was being used as the validation set exactly one time while all other subjects were used as training data. The final validation results are then the average across all subjects. Using LOSO cross-validation ensures that obtained results were not a product of overfitting on subject-specific traits. As we are using the same preprocessed version of the Opportunity dataset as introduced in the original work [26], we also employed the same train/ test split and trained the networks using said inputs.

Figure 2 shows results for all five datasets using varying hidden units and either a 1- or 2-layered LSTM within the DeepConvLSTM architecture. Results were obtained by averaging across 5 runs with a set of 5 varying random seeds. We report standard evaluation metrics, namely precision, recall and F1-score.

As we can see in Figure 2 architectures which employ a 1-layered LSTM overall outperforms architectures which employ a 2-layered LSTM across four (Opportunity, Wetlab, HHAR and SBHAR) out of five datasets. Only for the case of 1024 employed hidden units and using the Opportunity dataset as input, our architecture is on par with the original DeepConvLSTM architecture. For the Wetlab, RWHAR and HHAR dataset we can even see that the best performing 2-layered LSTM architecture variation performs worse, as far as precision, recall and F1-score, than the worst performing 1-layered LSTM architecture variation. Only for the SBHAR dataset our architecture is marginally worse than the original one when using 1024 hidden units within the LSTM layers.

By removing the second LSTM layer we are able to get performance increases up to 11.7% in F1-score (HHAR, 256 hidden units). Looking at Figure 2 one can see that the performance increase is the largest for variations employing 128 and 256 units and steadily decreases for larger amounts of hidden units. While one can see that the performance difference was not as significant for the SBHAR dataset when employing 512 hidden units and even slightly negative for the 1024 hidden units architecture, we nevertheless argue that both architecture variations are at least on par due to the fact that one has to consider statistical variance. Over the five runs, we witness an average standard deviation of 1.67% (1.53%) for precision, 1.69% (1.44%) for recall and 1.65% (1.38%) for the F1-score, using the 1-layered (2-layered) version of the architecture, respectively. The exact averages per architecture variation can be found in the repository (see <https://github.com/mariusbock/dl-for-har>).

As HAR datasets are relatively small in size compared to other popular Deep Learning datasets, e.g. [6], larger networks are more likely to be prone to overfitting. Though we witnessed this trend for the HHAR and Opportunity dataset, by monitoring the generalization gap (i.e. the difference in loss, accuracy, precision, recall and F1-score between the train and validation set), we did not see any indication that the 2-layered architecture variations are more likely to overfit and saw both architectures equally suffering from this phenomenon.

With the suggested removal of one LSTM layer we are decreasing the complexity of the DeepConvLSTM architecture. Assuming both layers in the original LSTM have the same number of hidden units

$h$ , the number of learnable parameters  $p_2$  within the LSTM of DeepConvLSTM is  $p_2 = 4sh + 8h + 12h^2$  with  $s$  denoting the size of the sliding window. By removing the second layer, the number of LSTM parameters in our architecture reduces to  $p_1 = 4sh + 4h + 4h^2$ , which shows that the increase in parameters with an increasing number of hidden units is dominated by  $4h^2$  for  $p_1$  opposed to  $12h^2$  for  $p_2$ . In our experiments the removal of one LSTM layer roughly equated in 63% fewer learnable LSTM parameters across all experiments. This decrease in complexity can be also seen in decreased runtimes using our architecture compared to the original DeepConvLSTM architecture. On average we are seeing a decrease in runtime of 30% across all experiments with the difference increasing (going up to as much as 48%) the more hidden units are employed and the larger the input dataset is.

## 4 CONCLUSION AND FUTURE DIRECTIONS

We proposed to change the DeepConvLSTM architecture as frequently used in activity recognition to employ a 1-layered instead of a 2-layered LSTM. We validated our architecture change with experiments using 5 publicly available datasets [28, 29, 32, 33, 36] with varying numbers of hidden units within the LSTM.

Results show that for 4 out of 5 datasets, one LSTM layer consistently outperforms the original DeepConvLSTM architecture [26] in terms of precision, recall and F1-score. For [28], it outperforms the original 2-layer architecture only for LSTMs with smaller number of hidden units, and is on par for larger LSTMs. With our suggested removal of the second LSTM layer, we are able to decrease the number of learnable parameters of the LSTM within the DeepConvLSTM [26] by an average factor of 62% and are able to decrease training time on average by a factor of 38%, going up to as much as 48% for larger networks (with more hidden units).

These findings contradict with the belief that one needs at least a 2-layered LSTM when dealing with sequential data [17]. As this belief finds its origins in the area of text analytics, we suggest that it might not be applicable to sensor-based HAR. Our choice of datasets to use for our experiments was driven by testing our hypotheses against datasets containing complex (Wetlab and Opportunity), transitional (SBHAR) and simpler/ periodical activities (SBHAR, HHAR and RWHAR). Looking at the results one can see that the hypothesis holds for all three types of activities. Furthermore, with the Opportunity dataset and its 113 feature dimensions we showed that a performance increase can be also witnessed when using more than one sensor. Therefore we currently cannot clearly identify specific reasons for why 1-layered LSTMs are performing better within the setting of HAR. Next steps within this research include investigating whether the hypothesis holds true for larger HAR datasets to further rule our overfitting being a reason for the performance increase. Using e.g. a feature extraction tool as seen in [20] could be used to create such larger datasets. Further, we will analyse how 1-layered LSTMs learn compared to 2-layered ones within the setting of HAR by replicating a similar analysis as performed by Karparthy et al. [17].

This paper’s architecture and experiment scripts are publicly downloadable via <https://github.com/mariusbock/dl-for-har>, to support the continuation, replication, as well as further analysis of our architecture and experiments.

## REFERENCES

- [1] Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question Answering With Subgraph Embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 615–620. <http://doi.org/10.3115/v1/D14-1067>
- [2] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors. *Comput. Surveys* 46, 3 (2014). <https://doi.org/10.1145/2499621>
- [3] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep Learning for Sensor-Based Human Activity Recognition: Overview, Challenges, and Opportunities. *Comput. Surveys* 54, 4 (2021). <https://doi.org/10.1145/3447744>
- [4] Yuwen Chen, Kunhua Zhong, Ju Zhang, Qilong Sun, and Xueliang Zhao. 2016. LSTM Networks for Mobile Human Activity Recognition. In *Proceedings of the 2016 International Conference on Artificial Intelligence: Technologies and Applications*. Atlantis Press, 50–53. <https://doi.org/10.2991/icaita-16.2016.13>
- [5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) From Scratch. *Journal of Machine Learning Research* 12, 76 (2011), 2493–2537. <http://jmlr.org/papers/v12/collobert11a.html>
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [7] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. 2013. Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1915–1929. <https://doi.org/10.1109/TPAMI.2012.231>
- [8] Xavier Glorot and Yoshua Bengio. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 9)*, Yee Whye Teh and Mike Titterton (Eds.). 249–256. <http://proceedings.mlr.press/v9/glorot10a>
- [9] Yu Guan and Thomas Plötz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition Using Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017). <https://doi.org/10.1145/3090076>
- [10] Sojeong Ha and Seungjin Choi. 2016. Convolutional Neural Networks for Human Activity Recognition Using Multiple Accelerometer and Gyroscope Sensors. In *International Joint Conference on Neural Networks*. 381–388. <https://doi.org/10.1109/IJCNN.2016.7727224>
- [11] Sojeong Ha, Jeong-Min Yun, and Seungjin Choi. 2015. Multi-Modal Convolutional Neural Networks for Activity Recognition. In *IEEE International Conference on Systems, Man, and Cybernetics*. 3017–3022. <https://doi.org/10.1109/SMC.2015.525>
- [12] Nils Hammerla, James Fisher, Peter Andras, Lynn Rochester, Richard Walker, and Thomas Ploetz. 2015. PD Disease State Assessment in Naturalistic Environments Using Deep Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 29, 1 (2015). <https://ojs.aaai.org/index.php/AAAI/article/view/9484>
- [13] Nils Y. Hammerla, Shane Halloran, and Thomas Ploetz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. 1533–1540. <http://www.ijcai.org/Proceedings/16/Papers/220.pdf>
- [14] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* 29, 6 (2012), 82–97. <https://doi.org/10.1109/MSP.2012.2205597>
- [15] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2014. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1–10. <http://doi.org/10.3115/v1/P15-1001>
- [16] WenChao Jiang and Zhaozheng Yin. 2015. Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks. In *Proceedings of the 23rd ACM International Conference on Multimedia*. 1307–1310. <https://doi.org/10.1145/2733373.2806333>
- [17] Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and Understanding Recurrent Networks. *CoRR abs/1506.02078* (2015). <http://arxiv.org/abs/1506.02078>
- [18] Tae-Young Kim and Sung-Bae Cho. 2019. Predicting Residential Energy Consumption Using CNN-LSTM Neural Networks. *Energy* 182 (2019), 72–81. <https://doi.org/10.1016/j.energy.2019.05.230>
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems* 1 (2012), 1097–1105. <https://dl.acm.org/doi/10.5555/2999134.2999257>
- [20] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D. Abowd, Nicholas D. Lane, and Thomas Plötz. 2020. IMUTube: Automatic Extraction of Virtual on-Body Accelerometry From Video for Human Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (Sept. 2020). <https://doi.org/10.1145/3411841>
- [21] Nicholas D. Lane and Petko Georgiev. 2015. Can Deep Learning Revolutionize Mobile Sensing?. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*. 117–122. <https://doi.org/10.1145/2699343.2699349>
- [22] Song-Mi Lee, Sang M. Yoon, and Heeryon Cho. 2017. Human Activity Recognition From Accelerometer Data Using Convolutional Neural Network. In *IEEE International Conference on Big Data and Smart Computing*. 131–134. <https://doi.org/10.1109/BIGCOMP.2017.7881728>
- [23] Tomáš Mikolov, Anoop Deoras, Daniel Povey, Lukáš Burget, and Jan Černocký. 2011. Strategies for Training Large Scale Neural Network Language Models. In *IEEE Workshop on Automatic Speech Recognition & Understanding*. 196–201. <https://doi.org/10.1109/ASRU.2011.6163930>
- [24] Vishvak S. Murahari and Thomas Plötz. 2018. On Attention Models for Human Activity Recognition. In *Proceedings of the ACM International Symposium on Wearable Computers*. 100–103. <https://doi.org/10.1145/3267242.3267287>
- [25] Maryam M. Najafabadi, Flavio Villanustre, Taghi M. Khoshgoftar, Naeem Seliya, Randall Wald, and Edin Muharemagic. 2015. Deep Learning Applications and Challenges in Big Data Analytics. *Journal of Big Data* 2, 1 (2015). <https://doi.org/10.1186/s40537-014-0007-7>
- [26] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (2016). <https://doi.org/10.3390/s16101115>
- [27] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria P. Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S. Iyengar. 2018. A Survey on Deep Learning: Algorithms, Techniques, and Applications. *Comput. Surveys* 51, 5 (2018). <https://doi.org/10.1145/3234150>
- [28] Jorge-L. Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-Aware Human Activity Recognition Using Smartphoneseon-Body Localization of Wearable Devices: An Investigation of Position-Aware Activity Recognition. *Neurocomputing* 171 (2016), 754–767. <https://doi.org/10.1016/j.neucom.2015.07.085>
- [29] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczeck, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkil, Alois Ferscha, Jakob Doppler, Clemens Holzmann, Marc Kurz, Gerald Holl, Ricardo Chavarriaga, Hesam Sagha, Hamidreza Bayati, Marco Creatura, and José del R. Millán. 2010. Collecting Complex Activity Datasets in Highly Rich Networked Sensor Environments. In *7th International Conference on Networked Sensing Systems*. 233–240. <https://doi.org/10.1109/INSS.2010.5573462>
- [30] Charissa A. Ronao and Sung-Bae Cho. 2016. Human Activity Recognition With Smartphone Sensors Using Deep Learning Neural Networks. *Expert Systems with Applications* 59 (2016), 235–244. <https://doi.org/10.1016/j.eswa.2016.04.032>
- [31] Tara N. Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhubana Ramabhadran. 2013. Deep Convolutional Neural Networks for LVCSR. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 8614–8618. <https://doi.org/10.1109/ICASSP.2013.6639347>
- [32] Philipp M. Scholl, Matthias Wille, and Kristof Van Laerhoven. 2015. Wearables in the Wet Lab: A Laboratory System for Capturing and Guiding Experiments. 589–599. <https://doi.org/10.1145/2750858.2807547>
- [33] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor S. Prentow, Mikkel B. Kjærgaard, Anind Dey, Tobias Sonne, and Mads M. Jensen. 2015. Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. 127–140. <https://doi.org/10.1145/2809695.2809718>
- [34] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems, Vol. 2*. 3104–3112. <http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>
- [35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. <https://doi.org/10.1109/CVPR.2015.7298594>
- [36] Timo Szttyler and Heiner Stuckenschmidt. 2016. On-Body Localization of Wearable Devices: An Investigation of Position-Aware Activity Recognition. In *IEEE International Conference on Pervasive Computing and Communications*. 1–9. <https://doi.org/10.1109/PERCOM.2016.7456521>
- [37] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. 1799–1807. <https://dl.acm.org/doi/10.5555/2968826.2969027>
- [38] Rui Xi, Mengshu Hou, Mingsheng Fu, Hong Qu, and Daibo Liu. 2018. Deep Dilated Convolution on Multimodality Time Series for Human Activity Recognition. In

- International Joint Conference on Neural Networks*. 1–8. <https://doi.org/10.1109/IJCNN.2018.8489540>
- [39] Kun Xia, Jianguang Huang, and Hanyu Wang. 2020. LSTM-CNN architecture for human activity recognition. *IEEE Access* 8 (2020), 56855–56866. <https://doi.org/10.1109/ACCESS.2020.2982225>
- [40] Cheng Xu, Duo Chai, Jie He, Xiaotong Zhang, and Shihong Duan. 2019. InnoHAR: A deep neural network for complex human activity recognition. *IEEE Access* 7 (2019), 9893–9902. <https://doi.org/10.1109/ACCESS.2018.2890675>
- [41] Jian B. Yang, Minh N. Nguyen, Phyo P. San, Xiao L. Li, and Shonali Krishnaswamy. 2015. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. In *International Joint Conference on Artificial Intelligence*. 3995–4001. <http://ijcai.org/papers15/Papers/IJCAI15-561.pdf>
- [42] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing. In *Proceedings of the 26th International Conference on World Wide Web*. 351–360. <https://doi.org/10.1145/3038912.3052577>
- [43] Piero Zappi, Clemens Lombriser, Thomas Stiefmeier, Elisabetta Farella, Daniel Roggen, Luca Benini, and Gerhard Tröster. 2008. Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection. In *Wireless Sensor Networks*, Roberto Verdone (Ed.). 17–33. [https://doi.org/10.1007/978-3-540-77690-1\\_2](https://doi.org/10.1007/978-3-540-77690-1_2)
- [44] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A C-LSTM Neural Network for Text Classification. *CoRR* abs/1511.08630 (2015). <http://arxiv.org/abs/1511.08630>