Towards Multimodal Deep Learning for Activity Recognition on Mobile Devices

Valentin Radu
University of Edinburgh

Nicholas D. Lane
University College London and
Nokia Bell Labs

Sourav Bhattacharya
Nokia Bell Labs

Cecilia Mascolo
University of Cambridge

Mahesh K. Marina
University of Edinburgh

Fahim Kawsar
Nokia Bell Labs

Ubicomp/ISWC'16 Adjunct, September 12-16, 2016, Heidelberg, Germany
ACM 978-1-4503-4462-3/16/09.
http://dx.doi.org/10.1145/2968219.2971461

Abstract
Current smartphones and smartwatches come equipped with a variety of sensors, from light sensor and inertial sensors to radio interfaces, enabling applications running on these devices to make sense of their surrounding environment. Rather than using sensors independently, combining their sensing capabilities facilitates more interesting and complex applications to emerge (e.g., user activity recognition). But differences between sensors ranging from sampling rate to data generation model (event triggered or continuous sampling) make integration of sensor streams challenging. Here we investigate the opportunity to use deep learning to perform this integration of sensor data from multiple sensors. The intuition is that neural networks can identify nonintuitive features largely from cross-sensor correlations which can result in a more accurate estimation. Initial results with a variant of a Restricted Boltzmann Machine (RBM), show better performance with this new approach compared to classic solutions.

Author Keywords
multimodal sensing; deep learning; context detection; activity recognition; mobile sensing

ACM Classification Keywords
I.2.6 [Learning]: Parameter learning; I.5.1 [Models]: Neural nets; I.5.4 [Applications]: Signal processing

Introduction

With the advancement of electronics and processor miniaturization, a new generation of smart devices is emerging for personal monitoring and private data processing. A common characteristic across these devices is a rich set of embedded sensors allowing them to run interesting applications to benefit their users. These simple, numerous sensors provide the opportunity to help with more complex inference tasks by combining capabilities across complementary modalities; for example, as seen in [8]. But due to their intrinsic nature and sensing characteristics (e.g., sampling rate and statistical properties) integrating sensor streams is often very challenging. Extracting relevant features and finding correlations between these features across sensing modalities to improve inference accuracy is therefor a pressing problem of immediate interest.

The focus of this work is to investigate the ability for deep-learning to advance the state-of-the-art in multimodal sensing on mobile and embedded devices.

Deep-learning [3] is an area of machine learning that is revolutionizing several domains from computer vision to speech recognition and many others. This fast growing area of research has the potential to influence key topics like sensor data fusion, with study of this learning paradigm applied to mobile devices only recently begun [6, 4].

One attractive characteristic of deep learning is the ability to transform close to raw sensor data into a dense representation of features through different activation patterns of artificial neurons (i.e. units) within a deep neural network. This network is used to perform inferences (e.g., estimating the activity class) directed by the activation pattern of neurons in the network, and often achieves higher accuracy than classic modeling methods.

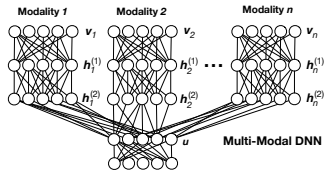

Figure 1: RBM-specific deep multimodal learning architecture.

With evidence from deep architectures on dual modalities like text mixed with images [11] and audio linked with video [9, 7], similar impressive gains should be attainable with other combinations of modalities; for example, in our case multiple cheap sensors present on mobile and wearable devices. The goal of this work is to provide the initial answers to whether these algorithms can increase the accuracy of ubiquitous tasks (e.g., user activity recognition) using sensor data from wearable devices, which is not well explored in the literature. We start this exploration by using a multimodal RBM architecture (a promising deep-learning algorithm) and initial results seem promising, while resource requirements make this architecture viable to resource constrained computation units like wearable devices.

Deep Learning for Multimodal Sensing

The multi-layer feature representation of deep-learning architectures allows them to extract more complex information than readily used shallow methods. Common shallow methods require a selection of hand crafted features to be extracted out of sensor data as a pre-processing stage, with performance being affected by the quality of these features.

As an alternative, uni-modal deep architectures do not have separate layers per sensor, which prevents the network first learning sensor-specific information before these concepts are unified across all sensors. Previous work has shown this intra-sensor relationship to be much stronger than inter-sensor counterpart [10].

Multimodal RBM Learning

To understand the utility of deep learning for activity recognition under systems with multiple sensors, this work studies the use of a multimodal version of Restricted Boltzmann Machines [7] (RBMs) (presented in Figure 1). Prior constructions of this variety of RBMs have been used to fuse

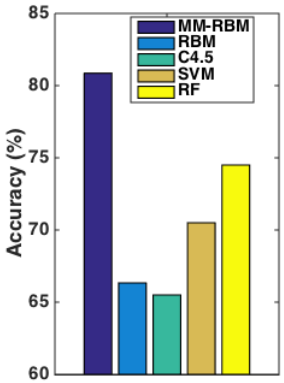

Figure 2: Comparative performance of the proposed deep-learning architecture (**MM-RBM**), a RBM architecture with concatenated sensor streams input and three shallow classifiers (C4.5, SVM and Random Forest). The proposed method outperformes previous solutions for activity recognition tasks on sensor data inputs from multiple sensors.

pairs of text, video and audio data for the purpose of image captioning [10, 11], and speech [7] or emotion recognition [5]. Instead, our objective is to empirically verify if these multimodal RBMs are still suited for new sensing tasks and if these can run on wearable devices.

Formalizing the inference process of RBMs: the state ($\mathcal{A}_i^{L+1}$) of each individual RBM unit (x_i^{L+1}) within a layer ($L+1$) is dependent on the unit weights connecting the j^{th} node in layer L to the i^{th} node in layer $L+1$. In this fully connected approach there is a connection between each (x_i^{L+1}) node to all nodes (x_j^L) on layer L, weighted as w_{ij}^{L+1}. Specifically this relationship is computed as:

$$\mathcal{A}_i^{L+1} = \frac{1}{1+\exp(-\sum_j w_{ij}^{L+1}\, x_j^L)} \tag{1}$$

As shown in Figure 1, separate architectural branches (M_k) exist for each sensing modality (sensor type) without any inter-sensor connections between these initial layers until later unifying independent sensor layers (U_l) in the final layers of the architecture. As an effect, all layers contribute to the learning of a joint representation of all sensor modalities. This aspect is expressed as:

$$P(\mathbf{v},\, \mathbf{h}; \Theta) = \frac{1}{\mathcal{Z}(\Theta)} \exp(-\mathcal{E}(\mathbf{v},\, \mathbf{h}; \Theta)) \tag{2}$$

where $\mathbf{v}$ represents the visible units (input modalities), $\mathbf{h}$ represents the hidden units inside the network, $\mathcal{Z}(\Theta)$ is the normalizing function, $\mathcal{E}$ is the cumulative state of the final layer and $\Theta = \{\mathbf{a}, \mathbf{W}\}$ represent the set of RBM parameters ($\mathbf{a}$ are the biases for the hidden layers).

In essence, feature learning is performed at the level of network parameters, represented by the weights between the nodes and network depth. Training is performed by back-propagation, running several times over a training set and gradually adjusting the network parameters to match the expectation as observed from the training set. We perform a learning method adopted originally from earlier studies of such architectures [7].

This training process can be computationally expensive, though fortunately this can be performed offline (i.e., on cloud servers), only requiring the final weights and parameters to be transferred to the wearable device.

Evaluation

We use a publicly available dataset [12], containing accelerometer and gyroscope signals, collected from 9 participants performing a set of common activities (sitting, standing, walking, climbing stairs, descending stairs, biking) and annotated with ground truth information. Another important characteristic of this dataset is that users perform these activities with 6 different smartphones, which increases the complexity of the inference task.

Under this dataset, the proposed deep learning architecture, Multimodal RBM (MM-RBM) is formed by two hidden layers on each of the sensor data streams (acceleration and angular velocity). The extracted data are combined with a hidden layer. Implementation of the deep learning architecture is performed in Torch 7 [2], which interprets the model into C, before being compiled for the platform.

Figure 2 shows the performance of the proposed MM-RBM, along with a simple RBM with concatenated sensor streams input (referred to as RBM in the figure) and the best performing shallow classifiers evaluated on a leave one user out approach (with training on all but one user) [12]. What is important is that this performance is achieved without any hand selection features, skipping a required process in the case of shallow classifiers.

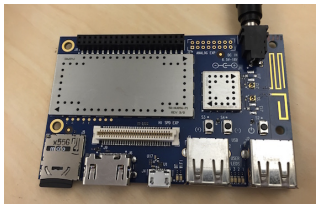

Figure 3: Qualcomm Snapdragon 410c. This development board runs a processor common to many smartwatches. We measure the performance of our algorithm on this processor to replicate the performance on typical wearable device.

Latency (ms)	50
Memory (MB)	2.75
Energy (mJ)	97

Table 1: Resource requirements of the Multimodal RBM. The low resource demands of MM-RBM makes the model feasible for constrained devices.
Time and energy consumption are indicated per inference.

The implications of these results are important because it suggests deep learning methods, such as the one we propose, are able to better extract discriminative information from multimodal sensor data than more commonly used shallow learners. Furthermore, their robustness across users, as shown from training and evaluating on different users, encourages their use in large-scale mobile sensing applications.

Mobile Hardware Feasibility

To test the feasibility of running the proposed multimodal deep architecture on wearable devices, we experiment with the Qualcomm Snapdragon 410c development board. The same processor is found in many smartwatches currently on the market (e.g. LG GWatch R [1]) and includes a quad-core 1.4 GHz CPU and 1 GB of RAM. The key finding is that our multimodal RBM is practical for this platform, and consumes a low enough amount of resources (see Table 1) that is feasible for wearable and mobile use.

Conclusion

Using deep learning to combine different perspectives captured in signals of multimodal data seems very promising. Results indicate this outperforms previous solutions for activity recognition, with resource requirements suited for constrained devices.

Although we show performance under a single dataset, the concepts for activity recognition we propose here can potentially generalize to other activity domains. Furthermore, we are looking to experiment with other types of networks like Convolution Neural Networks.

REFERENCES

1. 2016. LG G Watch R. `https://www.qualcomm.com/products/snapdragon/wearables/lg-g-watch-r`. (2016).
2. 2016. Torch. `http://torch.ch/`. (2016).
3. Y. Bengio, I. J. Goodfellow, and A. Courville. 2015. Deep Learning. (2015).
4. N. D. Lane and P. Georgiev. 2015. Can Deep Learning Revolutionize Mobile Sensing?. In *In Proc. HotMobile*. ACM.
5. W. Liu, W.-L. Zheng, and B.-L. Lu. 2016. Multimodal Emotion Recognition Using Multimodal Deep Learning. *CoRR* (2016).
6. T. Plotz N. Y. Hammerla, S. Halloran. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. In *In Proc. IJCAI*.
7. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. 2011. Multimodal Deep Learning. In *In Proc. ICML*.
8. V. Radu, P. Katsikouli, R. Sarkar, and M. K. Marina. 2014. A Semi-Supervised Learning Approach for Robust Indoor-Outdoor Detection with Smartphones. In *In Proc. SenSys*. ACM.
9. D. S. Sachan, U. Tekwani, and A. Sethi. 2013. Sports Video Classification from Multimodal Information Using Deep Neural Networks. In *AAAI*.
10. K. Sohn, W. Shang, and H. Lee. 2014. Improved Multimodal Deep Learning with Variation of Information. In *In Proc. NIPS*.
11. N. Srivastava and R. R. Salakhutdinov. 2012. Multimodal Learning with Deep Boltzmann Machines. In *In Proc. NIPS*.
12. A. Stisen and et al. 2015. Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition. In *In Proc. SenSys*.