Multimodal Deep Learning for Activity and Context Recognition

Valentin Radu[†], Catherine Tong[§], Sourav Bhattacharya[‡], Nicholas D. Lane[‡§], Cecilia Mascolo[*], Mahesh K. Marina[†], and Fahim Kawsar[‡△]

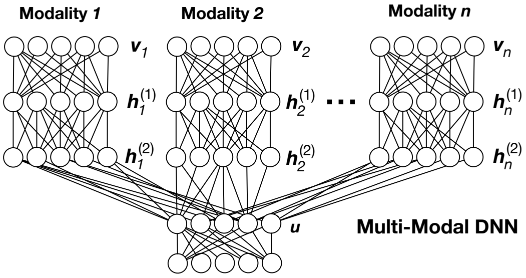
[†]University of Edinburgh, [§]University of Oxford, [‡]Nokia Bell Labs, [*]University of Cambridge, [△]TU Delft

The popularity of smart mobile and wearable devices has given rise to a growing interest in complex sensing tasks such as activity and context recognition. These tasks typically rely on data from a multitude of modalities, captured by low-energy small form-factor sensors such as light detectors, magnetometer, accelerometer and barometer. A successful combination of information across multi-modal sensor streams dictates the fidelity at which they can track user behavior and context.

In this study, we consider the benefits of adopting *deep learning* algorithms for activity and context recognition as captured by multi-sensor systems. Specifically, we use fully-connected Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) and compare two multimodal architectures for each type of neural network. One architecture, *Feature Concatenation* (FC), is a commonly employed approach for multimodal data fusion, which simply concatenate raw sensor streams at the input layer. We compare this to a novel architecture, *Modality-Specific Architecture* (MA); In this architecture (Fig. 1), separate neural networks are built per modality, before their generated concepts are unified through representations which bridge across all sensors. MA is based on the architecture proposed in [Ngiam et al., 2011], although our formation and experiments is the first time that this architecture has been tested on mobile data [Radu et al., 2018].

We use 4 publicly available datasets for evaluation, covering recognition of human activity, gait, sleep stages, as well as indoor-outdoor detection. Our experiments show that these generic multimodal neural network models compete well with shallow methods and task-specific modelling pipelines, across a wide range of sensor types and inference tasks. Although the training and inference overhead of these multimodal deep approaches is in some cases appreciable, we also demonstrate the feasibility of on-device mobile and wearable execution is not a barrier to adoption. This study is carefully constructed to focus on multimodal aspects of wearable data modeling for deep learning by proving a wide range of empirical observations, which we expect to have considerable value in the community. We summarize our observations into a series of practitioner rules-of-thumb and lessons learned that can guide the usage of multimodal deep learning for activity and context detection.

Figure 1: Modality-Specific Architecutre (MA) with Deep Neural Networks. Separate branches exist for each of n modalities, which are joined in unifying cross-sensor layers.



References

[Ngiam et al., 2011] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 689–696. Omnipress.

[Radu et al., 2018] Radu, V., Tong, C., Bhattacharya, S., Lane, N. D., Mascolo, C., Marina, M. K., and Kawsar, F. (2018). Multimodal deep learning for activity and context recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4):157:1–157:27.