Handling Real-time Unlabeled Data in Activity Recognition using Deep Bayesian Active Learning and Data Programming

Gautham Krishna Gudur[†], Prahalathan Sundaramoorthy[‡], Venkatesh Umaashankar[†]
[†]Ericsson, India, [‡]University of Southern California, USA

The ubiquitous proliferation of low-cost mobile and wearable devices has spawned growing research in extracting contextual information from sensor data, particularly for Human Activity Recognition (HAR) owing to its applications in fall detection, fitness tracking, health monitoring, etc. Contemporary Deep Learning models are engineered to alleviate the difficulties posed by conventional Machine Learning algorithms which require extensive domain knowledge to obtain heuristic hand-crafted features that have questionable generalizability. The pervasive deployment of such models necessitates us to primarily perform inference, and also progress towards minimalistic training on the edge with continuous incoming stream/flow of data from different users. In an on-device *Incremental Learning* [1] setting, to facilitate continuous updation of incoming streams of unseen test user data in real-world across various users (*User Adaptability*), a small portion of data is included to update the weights of a pre-trained stocked deep learning model, thereby eliminating the need to retrain the model from scratch.

However, acquiring labeled data in real-time has always been a major challenge, particularly in HAR wherein, conventional systems assume that the incoming activities performed by users are already labeled, thereby necessitating ground truthing techniques like *Active Learning* (AL). In this study, we propose a unified *Deep Bayesian Active Learning* framework for HAR which supports high-dimensional data, by combining recent advancements in Bayesian deep learning (Bayesian Neural Networks) with AL [2]. This solves the need for labeled data by querying only a handful of data points from the oracle, and also mitigates the problem of representing uncertainty in deep learning without sacrificing either computational complexity or test accuracy. By utilizing the stochastic regularization technique - *Dropout* which is considered as an approximate Bayesian inference in deep Gaussian processes by Yarin et al. [3], we sample from the approximate posterior using multiple stochastic forward passes to arrive at the most uncertain data points from incoming pool/test data. We experiment with multiple AL acquisition functions like BALD (Bayesian Active Learning by Disagreement), Max Entropy, Variation Ratios & Random Sampling across 3 different publicly available mobile- and wearable- based HAR & Fall Detection datasets [4][5]. Preliminary results show considerable increase in accuracies and f1-scores on a Leave-One-User-Out (LOOCV) Incremental Learning setting, with just $\sim$40% of data points on the *HARNet* architecture [1].

Data Programming, being an automated data-labeling paradigm, enables us to provide certain initial heuristics called *Labeling Functions* (LFs) to automate the process of ground truthing [6]. The users can programmatically create labels using weak supervision strategies or crowd-sourced labels, which naturally tend to be noisy and conflicting. The entire labeling process using LFs is then represented as a Generative model, and further tuned using a noise-aware Discriminative model. In our future works, we aim to leverage such data programming techniques to establish ground truths for unknown incoming test data by finding minimalistic patterns that govern LFs and also extracting dependencies between different classes, thereby facilitating label generation on-the-fly. This effectively reduces any oracle involvement during real-time and enables non-experts to easily automate ground-truthing. Data Programming can also be used to label training data from scratch when scarce or no labeled data is available.

References

[1] Prahalathan Sundaramoorthy et al. (2018) HARNet: Towards On-Device Incremental Learning using Deep Ensembles on Constrained Devices. In *Proceedings of the 2nd International Workshop on Embedded and Mobile Deep Learning (EMDL'18)*.

[2] Yarin Gal et al. (2017). Deep Bayesian Active Learning with Image Data. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*.

[3] Yarin Gal et al. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML'16)*.

[4] Allan Stisen et al. (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 (SenSys '15)*.

[5] Taylor R. Mauldin et al. (2018). SmartFall: A Smartwatch-Based Fall Detection System Using Deep Learning. In *Sensors 2018*.

[6] Alexander J. Ratner et al. (2016). Data Programming: Creating Large Training Sets, Quickly. In *Advances in Neural Information Processing Systems (NIPS 2016)*.