MAGNETO: Edge AI for Human Activity Recognition - Privacy and Personalization

Jingwei Zuo
Technology Innovation Institute
Abu Dhabi, UAE
jingwei.zuo@tii.ae

George Arvanitakis
Technology Innovation Institute
Abu Dhabi, UAE
george.arvanitakis@tii.ae

Mthandazo Ndhlovu
Technology Innovation Institute
Abu Dhabi, UAE
mthandazo.ndhlovu@tii.ae

Hakim Hacid
Technology Innovation Institute
Abu Dhabi, UAE
hakim.hacid@tii.ae

ABSTRACT

Human activity recognition (HAR) is a well-established field, significantly advanced by modern machine learning (ML) techniques. While companies have successfully integrated HAR into consumer products, they typically rely on a predefined activity set, which limits personalizations at the user level (edge devices). Despite advancements in Incremental Learning for updating models with new data, this often occurs on the Cloud, necessitating regular data transfers between cloud and edge devices, thus leading to data privacy issues. In this paper, we propose MAGNETO, an Edge AI platform that pushes HAR tasks from the Cloud to the Edge. MAGNETO allows incremental human activity learning directly on the Edge devices, without any data exchange with the Cloud. This enables strong privacy guarantees, low processing latency, and a high degree of personalization for users. In particular, we demonstrate MAGNETO in an Android device, validating the whole pipeline from data collection to result visualization.

1 INTRODUCTION

Human Activity Recognition (HAR) tasks has been largely investigated in the past decades. Numerous researches have studied the HAR problem from different aspects, from data collection, learning models, to post-processing and result interpretation [1].

As shown in Figure 1, traditional approaches [1, 5, 11] for HAR tasks are mainly Cloud-based: a classifier is trained on a predefined set of activities in a centralized Cloud environment. User's activity data collected on the Edge device is then sent to the Cloud for inference. However, this centralized, Cloud-based approach raises three main issues: (i) *high latency*, due to the User-Cloud communication, (ii) *lack of flexibility and personalization* to individual user's needs, and (iii) *lower privacy control*, due to the data transfer to the Cloud. In contrast with the conventional Cloud-based approach, Edge AI [13] shifts core processing tasks (e.g., model's training, inference, etc.) to the edge devices and intends to adapt AI technologies to the edge environment. In a specific use case, edge devices can collect user activity data directly and use it to update an initial model, catering well to user demands for *personalization*. This approach facilitates the deployment of optimized models and services directly onto the user devices, thus ensuring rapid real-time response (*low latency*) and enhanced *privacy guarantees*.

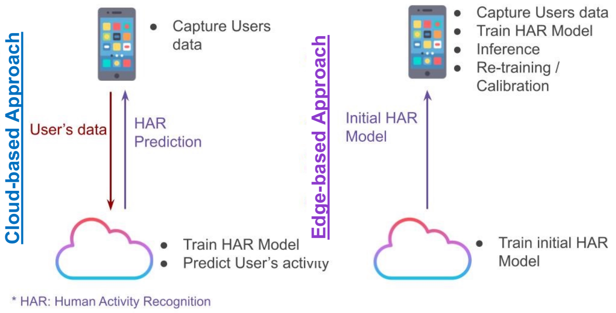


Figure 1: left) Human Activity Recognition (HAR) protocols: left) Cloud-based Approach, constant communication between Cloud and Edge, right) Edge-based Approach, only data transfer from Cloud to Edge is allowed, showing stronger privacy guarantees.

However, shifting both inference and learning process to Edge devices presents unique challenges due to the intrinsic constraints of Edge devices, including (i) Model size, which should be small enough to fit within the Edge but also to operate efficiently, (ii) Data size, which should be very limited due to the low storage capabilities within the Edge, and (iii) Energy consumption, constraining the training process to be very efficient without excessive power consumption. Moreover, in the context of personalized HAR, since models are required to be updated continuously (i.e., *personalization*), *Catastrophic Forgetting* [14] is a practical challenge when learning from the dynamic data streams. This becomes even more daunting with the limited Edge resources.

In this paper, considering the aforementioned challenges, we introduce MAGNETO, *sMArt sensinG for humaN activity rEcogniTiOn*, an Edge AI platform that shows the whole pipeline of HAR tasks, covering real-time data collection, data preprocessing, model adaptation/re-training/calibration, model inference and result visualization. Importantly, all the operations are taking place on the Edge, and should not have any data exchange with the Cloud, considering data privacy issues. Moreover, MAGNETO is equipped with the Incremental Learning ability for new activities, without *forgetting* previously learned activities, i.e., catastrophic forgetting. To demonstrate the effectiveness of MAGNETO, a user-friendly Android application has been developed. This offers users an interactive and practical way to experience the platform's capabilities on a smartphone. We believe that the HAR on the Edge with incremental learning capability can enable a new area in the health care, fitness or assistant applications.

The rest of this paper is organized as follows: Section 2 presents the most related work to our proposal. Section 3 introduces the

Published in Proceedings of the 27th International Conference on Extending Database Technology (EDBT), 25th March-28th March, 2024, ISBN 978-3-89318-095-0 on OpenProceedings.org.

Series ISSN: 2367-2005

10.48786/edbt.2024.69

Edge AI platform MAGNETO. Section 4 discusses the demonstration scenarios. We conclude and show future work in Section 5.

2 RELATED WORK

In this section, we show the most related work of our Edge AI platform, covering Edge Machine Learning and HAR tasks.

2.1 Edge Machine Learning (Edge ML)

Edge Machine Learning (Edge ML) shows advantages on low latency and strong privacy guarantee. Edge ML can be divided into two categories: Inference and Training on the Edge. For model inference on the Edge, past studies focus on optimizing model scale and quantizing weights to reduce resource costs, employing methods like parameter pruning [6], low-rank factorization [4], and knowledge distillation [8]. Training on the Edge, which has higher resource costs, either uses tiny models or employs distributed/federated learning [12]. The paper investigates models that can efficiently operate on Edge devices with limited resources. The objective of MAGNETO is to build a complete pipeline for HAR tasks, we should note that more sophisticated techniques can be integrated into the platform incrementally.

2.2 Human Activity Recognition (HAR)

Human Activity Recognition (HAR) is a well-established field. Building HAR model is basically considered as a classification task. Depending on data resources and targeted applications, the HAR model can be designed differently. One can use handcrafted features to feed any general ML models for downstream tasks, which is easy-to-deploy and requires linear processing time. More advanced work has been proposed in the Time Series Classification domain, where researchers aim to build general ML models covering various application domains [16], including HAR tasks. For instance, Shapelet features [15] with a kNN classifier, end-to-end models [16] with automatic feature extraction and selection. We should remark that the HAR models have been integrated into various commercial service or products, such as Google platform [11], Samsung health activity trackers [7], and Apple developer kit [5].

However, the existing work either rely on centralized learning or inference on the Cloud [15, 16], or a pre-defined activity sets lacking support for personalization [5, 7, 11]. This fall short in simultaneously addressing the crucial aspects of the Edge Learning on HAR tasks: *Privacy* and *Personalization.*

3 PROPOSAL: MAGNETO

To answer the aforementioned challenges, we propose MAGNETO, *sMArt sensinG for humaN activity rEcogniTiOn*, which is an Edge AI platform primarily designed for HAR tasks. Before describing its system design, we highlight two main definitions:

Definition 1. (Privacy). Given a Cloud server and Edge device, no user data is allowed to be transferred from Edge to Cloud. However, it is less restrict to pull data from Cloud to Edge.

Definition 2. (Personalization). Given a model Θ_o trained on $\mathcal{D}_o$, user's personal data $\mathcal{D}_n = (X_s, ..., X_t)$ with classes $(s, ..., t)$ will enrich Θ_o incrementally, leading to a new model Θ_n.

We should note that the *Personalization* is strongly linked to the *Incremental Learning* paradigm [3]. With extra user data, the model can be personalized in two ways: (i) re-calibrate an activity to be more accurate for her/his personal style or (ii) re-train the model to learn new custom activities according to user's habits.

3.1 Architecture

As mentioned before, the goal of MAGNETO is to provide user activity estimation directly on the Edge device and with the capability of adding new activities in the model, without transferring any of the user's data to the Cloud. Figure 2 shows the architecture and the components of MAGNETO, discussed below. To achieve its objectives, MAGNETO proceeds in two steps: (i) an offline step (*Cloud Initialization*) and (ii) an online step (*Edge Inference and Learning*). After learning a class-separable embedding space, a nearest class mean (NCM) classifier [14] can be built to do the Edge Inference.

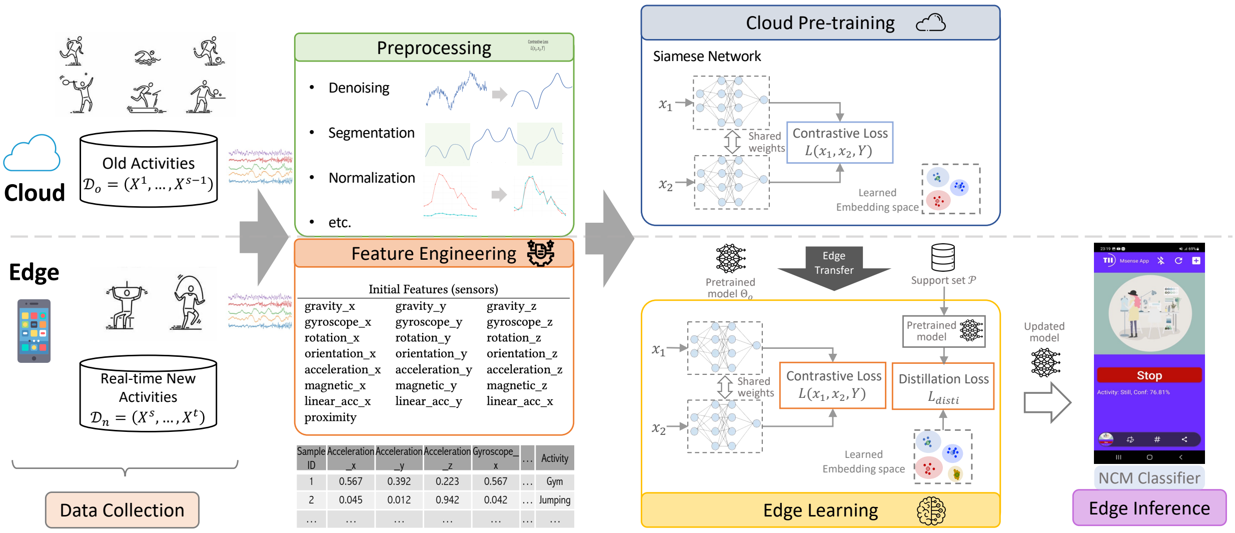
The **offline step** in this context is carried out on the Cloud for two main reasons: (a) resource limitations on the Edge, inviting us to leverage the scalable resources of the Cloud but without compromising user privacy. This task is actually similar to a transfer learning approach where a pre-trained model is used for, e.g., personalization purposes. It's essential for initializing the system, addressing the cold-start problem by providing a pre-trained model as a foundational knowledge base. This base is crucial for subsequent learning on the Edge, as it is not efficient to start from scratch. It's important to note that in the pre-training process, no user data is transferred from Edge to Cloud. The initial model is pre-trained exclusively using open-source data that is readily available, ensuring user data privacy and security.

During the **online step**, MAGNETO performs real-time inference of the user's activities based on their data, while incrementally learning from new data. This feature enhances personalization to the user's specific needs. Notably, learning from new data occurs directly on the device, eliminating the need to transmit data to the cloud. Finally, from the inference standpoint, this approach ensures minimal latency since all operations are executed directly on the Edge. More technical details of the online step (e.g., handling *Catastrophic Forgetting issues* and *Incremental Learning behaviors*) can be found in our previous papers [2, 14].

3.2 Cloud Initialization

To empower MAGNETO with the best possible initial model, instead of building an initial dataset from openly available data source, we launched data collection campaigns collecting an initial sensory dataset of more than 100GB, which was stored and processed on the cloud. A neural network is built from the pre-processed data, targeting the prediction of existing activities, embedded in the system as an initialization step. At the end of this step, the following items are transferred into the Edge device:

(1) **The pre-processing function**: We do popular pre-processing operations on raw sensor data, including denoising, segmentation, normalization, as shown in Figure 2. Moreover, we adopt a primary feature extractor that relies on handcrafted statistic features [14], requiring linear processing time. Nevertheless, more advanced feature extractors can be explored and integrated into our framework, by considering the Edge constraints. This is orthogonal to our work.
(2) **The Initial ML Model**: Being efficient and memory friendly, a Siamese Network-based model [10] with contrastive loss [9] is designed, which learns a class-separable embedding space. The backbone model is a simple Fully Connected (FC) neural network with dimensions $[1024 \times 512 \times 128 \times 64 \times 128]$, which can be replaced by any other advanced networks. The lightweight model can be efficiently deployed on Edge devices for retraining, even though the training data is very limited [14].



Sample ID	Acceleration _x	Acceleration _y	Acceleration _z	Gyroscope_ x	…	Activity
1	0.567	0.392	0.223	0.567	…	Gym
2	0.045	0.012	0.942	0.042	…	Jumping
…	…	…	…	…	…	…

Figure 2: Global system architecture of MAGNETO

(3) **The support set**: As MAGNETO supports learning new activity data on the fly, it is necessary to keep a minimal dataset to update the learning model (mainly for handling *Catastrophic Forgetting* issues [14]). The support set, containing a limited amount of data samples which are representative for each class, can greatly reduce the resource cost for Edge devices. For instance, 200 observations per class cost roughly 0.5MB in 32-bit precision. This support set has a twofold mission: (i) serving to calculating the class prototypes for building the *NCM* classifier, (ii) updating the model by combining with the new activity data as training set.

3.3 Edge Inference and new activities learning

With the aforementioned transferred items from the cloud initialization stage, the Edge device is capable of performing the inference on the fly by reading its sensors and passing the captured measurements sequentially from the pre-processing function to the pre-trained model. Again, this is performed without sharing any of the user's raw information with the Cloud.

For learning new data on the Edge, either on existing or new activities, we adopt the same base model as the *Cloud Initialization*, i.e., Siamese Network [10] with contrastive loss [9]. This few-shot learning approach allows updating the model with minimal data. To handle the *Catastrophic Forgetting* issue [14], we jointly optimize the model with contrastive loss [9] and distillation loss [8]. Here are the main steps happened on Edge devices:

(1) **Samples recording**: Users capture new activity samples that does not exist in the initial dataset, while annotating the activity, e.g., roughly 20-30 seconds of recording, that will be fed into the pre-processing function.
(2) **Support set update**: To keep track of the activities of users and ensure an incremental learning process, MAGNETO adds the freshly captured data into the support set, which serves to further re-training of the model.
(3) **Model re-training**: The initial model will be updated by integrating the patterns of the newly captured data. As mentioned previously, the cost function is a combination of Contrastive and Distillation Loss, optimized on the updated support set.

Following the re-training steps, the inference resumes as previously described. It's important to note two key aspects: first, the learning process can be repeated to accommodate the addition of multiple activities as per the user's requirements; second, calibrating an activity to more closely align with the user's behavior is a focal point of interest. This calibration mirrors the re-training process, with the distinction that the data for the targeted activity within the support set is replaced with newly acquired data.

4 ABOUT THE DEMONSTRATION

4.1 Demonstration settings

4.1.1 Demonstration environment. MAGNETO is developed and integrated into an Android Application. The core models are implemented in PyTorch 1.6.0. The proposal will be demonstrated in an Android smartphone, freely accessible to audiences.

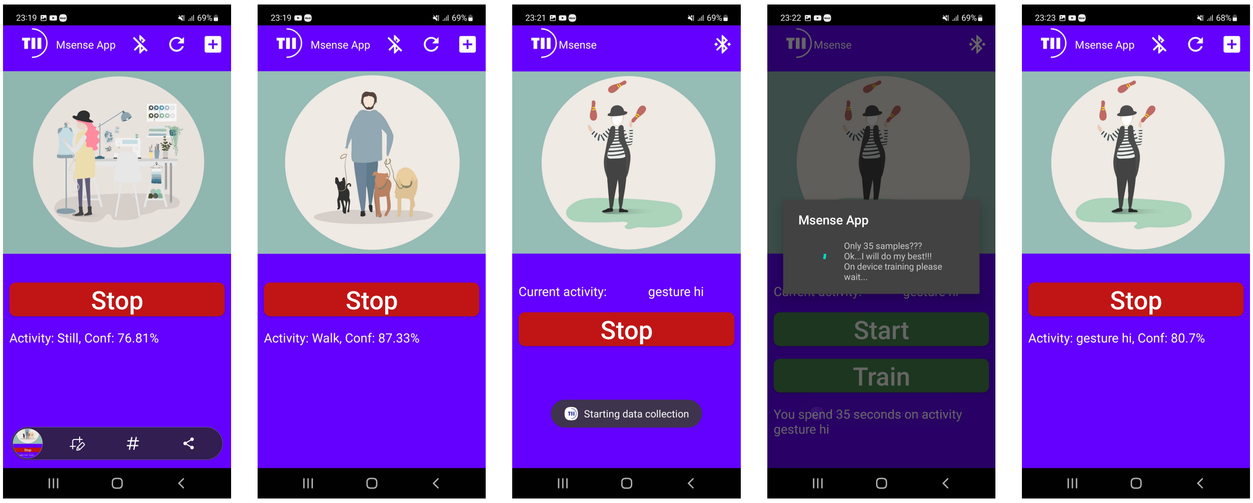
4.1.2 Dataset description. We base our demonstration on real human physical activity data collected on edge devices. We have launched data collection campaigns, capturing an initial dataset of more than 100GB of sensor data. We split the sensory data into a one-second window with roughly 120 sequential measurements from 22 mobile sensors, e.g., accelerometer, gyroscope, and magnetometer. We extract 80 statistical features. After data preprocessing, five activities with $\sim 200k$ records are collected: *Drive*, *E-scooter*, *Run*, *Still*, *Walk*. The data is applied to pre-train a HAR model. On the Edge device, the real-time coming data can be processed instantly, as the preprocessing requires linear time.

4.2 Demonstration scenarios

During the demonstration of MAGNETO [1], we focus on the following aspects: (i) Inference on the Edge, (ii) Incremental Learning of new activities on the fly, with no interaction with the Cloud. The demonstration device(s) will be disconnected from the Internet, ensuring the execution of the processes locally. In order to improve the user's experience, the phone's outputs will be projected on a screen for a real-time visualization.

4.2.1 Real-time Inference. As a first step, participants will be given a smartphone with MAGNETO installed, and try a few

[1] The demo video can be found in https://bit.ly/magneto_edbt2024



Activity: Still, Conf: 76.81%

Activity: Walk, Conf: 87.33%

Current activity: gesture hi

Starting data collection

Only 35 samples???
Ok...I will do my best!!!
On device training please wait...

You spend 35 seconds on activity gesture hi

Activity: gesture hi, Conf: 80.7%

(a) (b) (c) (d) (e)

Figure 3: GUI of the MAGNETO App: (a) Inference on *Still* with the initial model, (b) Inference on *Walk* with the initial model, (c) Collecting new activity data for *Gesture Hi*, (d) Updating the Edge model, (e) Inference on new activity *Gesture Hi*

existing activities in real time, i.e., *Drive*, *E-scooter*, *Run*, *Still*, *Walk*. Figure 3(a-b) show GUI examples for the real-time activity prediction. The participants will gain a clear understanding of the imperceptible prediction latency, which is only a few milliseconds.

4.2.2 Incremental Learning. With the same mobile phone, participants can collect new activity data with their personal behaviors, e.g., a greeting gesture as shown in Figure 3(c) for recording a few seconds[2], and request MAGNETO to learn and integrate the activity into the existing model, as shown in Figure 3(d). Participants can then check the new model's inference capability and check if it has managed to learn the new activity, see Figure 3(e).

We should remark that the entire data size that the demonstration needs on the Edge device (including support set, preprocessing, and the model) does not exceed $5MB$, that is lower than two high definition pictures a smartphone can take, further motivating and highlighting the value of such a mechanism.

5 DISCUSSIONS AND CONCLUSION

There are multiple motivations for pushing the Human Activity Recognition (HAR) tasks to the Edge. As the sensing part, the Edge devices seamlessly integrate the data collection and processing tasks in the same device, providing extremely low latency. The isolated Edge environment naturally unblocks the strong privacy guarantees and personalization possibilities. However, Edge devices are extremely limited in terms of computational resources. This necessitates a careful design of machine learning models which can be efficient executed on the Edge devices.

Conclusion In this paper, we present MAGNETO, an Edge AI platform designed for Human Activity Recognition (HAR) tasks, which provides rapid real-time response (low latency), enhanced privacy guarantees and personalization capability. Importantly, different from existing work or products in the market, MAGNETO allows adding new activities on the fly without the need for retraining the whole model. With a developed Android application, audiences can check the functionalities of the system in a user-friendly way. Even though MAGNETO is primarily designed for HAR tasks, its potential extends far beyond this application. Leveraging incremental learning, the system can adapt to diverse data types, such as time series, text, and voice. By adjusting its feature extractor or backbone model, the system offers a versatile solution adaptable to a wide range of scenarios.

REFERENCES

[1] Anindya Das Antar, Masud Ahmed, and Md Atiqur Rahman Ahad. 2021. Recognition of human locomotion on various transportations fusing smartphone sensors. *Pattern Recognition Letters* 148 (2021).

[2] George Arvanitakis, Jingwei Zuo, Mthandazo Ndhlovu, and Hakim Hacid. 2023. Practical Insights on Incremental Learning of New Human Physical Activity on the Edge. In *DSAA*. IEEE.

[3] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence* 44, 7 (2021), 3366–3385.

[4] Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. 2014. Exploiting linear structure within convolutional networks for efficient evaluation. *Advances in neural information processing systems* 27 (2014).

[5] Apple developers. 2018. https://developer.apple.com/documentation/coremotion/cmmotionactivity.

[6] Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. *NeurIPS* 28 (2015).

[7] Samsung health. 2023. https://www.samsung.com/global/galaxy/apps/samsung-health/.

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *stat* 1050 (2015), 9.

[9] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *NeurIPS*, Vol. 33.

[10] Gregory R. Koch. 2015. Siamese Neural Networks for One-Shot Image Recognition. In *ICML*.

[11] Google Platform. 2018. https://developers.google.com/location-context/activity-recognition.

[12] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM TIST* 10, 2 (2019), 1–19.

[13] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang. 2019. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proc. IEEE* 107, 8 (2019), 1738–1762.

[14] Jingwei Zuo, George Arvanitakis, and Hakim Hacid. 2023. On Handling Catastrophic Forgetting for Incremental Learning of Human Physical Activity on the Edge. In *EDBT 2023*. 792–798.

[15] Jingwei Zuo, Karine Zeitouni, and Yehia Taher. 2019. Exploring Interpretable Features for Large Time Series with SE4TeC. In *EDBT*. 606–609.

[16] Jingwei Zuo, Karine Zeitouni, and Yehia Taher. 2021. SMATE: Semi-Supervised Spatio-Temporal Representation Learning on Multivariate Time Series. In *ICDM*. IEEE, 1565–1570.

[2] A variety of activities can be used here. The listed ones are for illustration only