An Empirical Study on AI-Powered Edge Computing Architectures for Real-Time IoT Applications

Anne H. H. Ngu, Awatif Yasmin
Department of Computer Science, Texas State University

Motivation

- Edge computing is indispensable for IoT applications, handling data from billions of devices and expected to surpass 41.6 billion installations by 2023.
- It facilitates swift decision-making at the device level
- ,It conserves network bandwidth by processing data locally, making it suitable for resource-constrained or costly networks.
- Bolsters privacy and security by storing data locally, particularly crucial for applications that involves processing personal data.

Methodology

- Conduct a Comparative Analysis of Three different software architectures for running AI-Powered Real-Time IoT Applications, where data is collected in wearables.
- We used SmartFall, a Fall Detection IoT Application [1] for the study
- The study focus on the following key challenges in deploying AI-Powered IoT Applications on Edge devices:
  - Battery Life: Participants wore the watch for an hour to measure battery consumption during daily activities.
  - Data Latency: We recorded the time between accelerometer data sensing and prediction generation for each architecture.
  - Data Loss: Participants wore the watch while doing chores, and we checked for potential data loss by comparing collected data points.
  - Model Accuracy: Participants performed falls and ADL activities in a lab setting to assess fall detection accuracy, calculating precision, recall, and accuracy.
- Machine Learning model:
  - Utilizing LSTM, a popular RNN architecture, for fall detection.
  - Input: Accelerometer data in a 3x128 format, with 3 representing x, y, and z values, and 128 denoting the window size.
  - TFLite Version: Adapted for mobile and edge devices, facilitating widespread deployment.

Watch and Phone-based Architecture (WPA)

- The smartwatch collects data from sensors.
- Data is sent to the smartphone.
- The smartphone uses machine learning to make predictions.
- Predictions are sent back to the smartwatch.

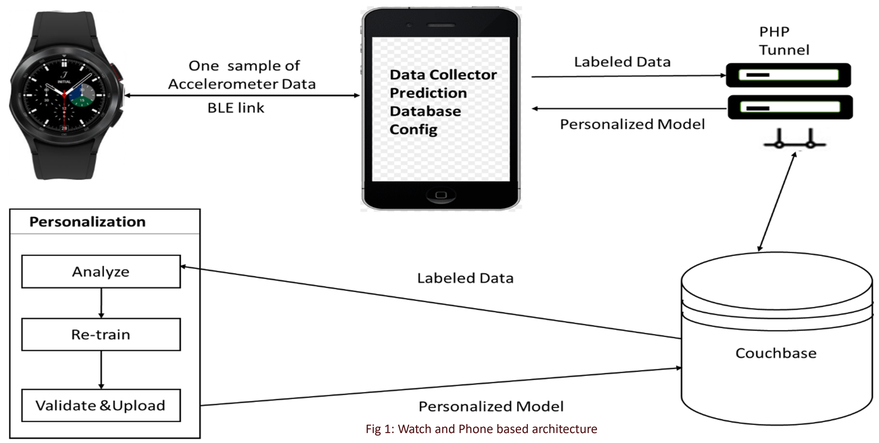

Fig 1: Watch and Phone based architecture

Watch Based Architecture (WA)

- The smartwatch collects data and makes predictions using its own system.
- Every so often, this data and feedback get sent to a cloud server for storage and analysis.
- Mainly use the smartphone to set up user profile at the start.

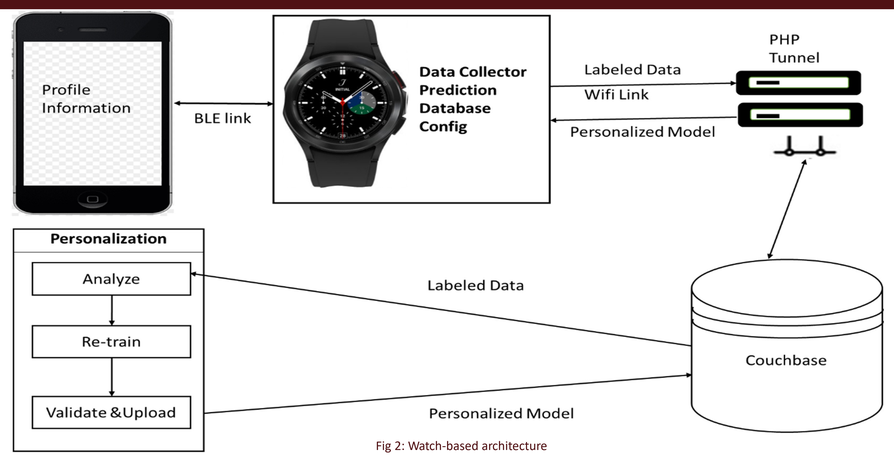

Fig 2: Watch-based architecture

Server Based Architecture (SA)

- The smartwatch sends data to a cloud server.
- Predictions are made using machine learning on the server.
- The results are sent back to the watch for users to see or interact with. The smartphone is used for setting up user profiles.

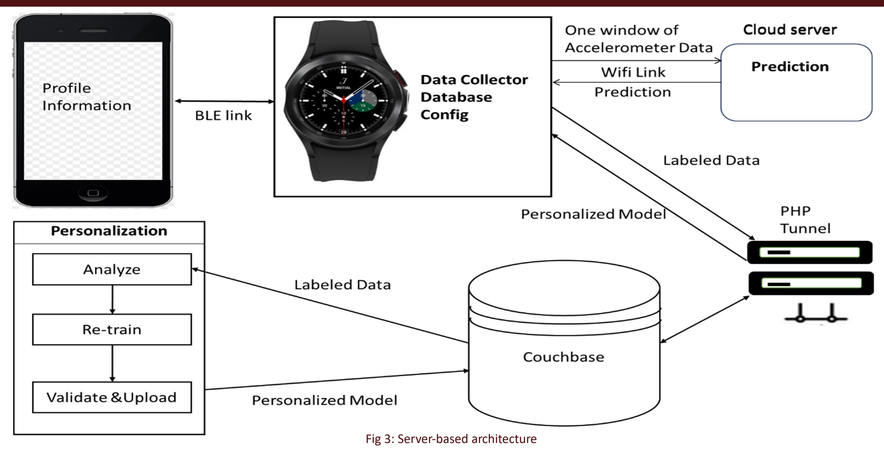

Fig 3: Server-based architecture

Results

Three participants used three configuration of SmartFall App, wearing the watch on the left wrist and carrying the phone nearby.

- Battery Optimization:
  - Phone : For WPA, phone drains the most battery and for others it is almost (<1%).
  - Watch: For WA and SA it drains almost 17% in an hour.
- Latency:
  - Highest latency in SA (101 ms);
  - WA lowest (73 ms).
- Data Loss
  - Most data loss in WPA, highest (8,838 points).
  - Least loss in SA (4,443 points).
- Model Accuracy:
  - SA Matches training F1 score (0.79, 0.76, 0.77).

Discussion and Conclusion

- Server-based architecture exhibited the best model prediction accuracy but incurred higher in data latency.
- Server-based architecture requires stable data transmissions
- There is quite a significant decrease in model accuracy for WPA and WA . This lower F1 score is due to hardware (no GPU available on edge device) and the use of light weight version of deep learning framework called TFLite.

Acknowledgement

- We thank the National Science Foundation for funding the research under the Smart and Connected Health Program (NSF-SCH-21223749) at Texas State University.

References

1. A. H. Ngu, V. Metsis, S. Coyne, P. Srinivas, T. Salad, U. Mahmud, and K. H. Chee, "Personalized watch-based fall detection using a collaborative edge-cloud framework," International journal of neural systems, vol. 32, no. 12, p. 2250048, 2022.

TEXAS STATE UNIVERSITY

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM