Hand Gesture Recognition on Arduino Using Time Series Classification

Matthew Lipski
Supervisors: Mingkun Yang , Ran Zhu
EEMCS, Delft University of Technology, The Netherlands
m.s.lipski@student.tudelft.nl, m.yang-3@tudelft.nl, r.zhu-1@tudelft.nl

1 Introduction

Machine learning and artificial intelligence have traditionally been restricted to the realms of high-performance and in turn, high power devices. Unfortunately, this means that it has been previously unfeasible to use these technologies with embedded hardware as it lacks the performance to run machine learning models locally, and lacks the dedicated power to be able to transmit sensor data to a remote processor. However, recent advances into machine learning model compression and optimization have changed this, allowing deep neural networks to be run on devices even powered by coin batteries, meaning that low-power microcontrollers can make sense of sensor data in much more sophisticated ways than previously possible. For example, a machine learning enhanced microcontroller may be able to tell when a car is about to break down based solely on the sound its engine produces.

This paper will specifically investigate the usage of machine learning models on embedded hardware to classify hand gestures, with the use of OPT101 photo diodes connected to an Arduino Nano 33 BLE microcontroller. To do this, we must first consider the full processing pipeline, is includes each function from reading the raw photo diode sensor data to outputting a gesture label at the end. This pipeline can be split into the following tasks:

1. Optimizing the number and placement of photo diodes.
2. Reading, processing, and sanitizing data from photo diodes.
3. Creating an appropriate dataset for the ML classifier to recognize gestures.
4. Training an appropriate ML model on the created dataset and ensuring it can run in real-time on an Arduino Nano 33 BLE.

It is important to note that this paper only investigates the final step of this pipeline, i.e. ”Training an appropriate ML model on the created dataset and ensuring it can run in real-time on an Arduino Nano 33 BLE”. Although the remaining steps are being investigated by other group members as part of the CSE3000 Research Project, they are beyond the scope of this paper. Therefore, the research for this paper was guided by the following research question:

”Which machine learning model yields the lowest error rate for real-time gesture recognition on an Arduino Nano 33 BLE, using 3D-formatted data from photo diodes?”

This question can be segmented into the following sub-questions:

1. What does ”3D-formatting” mean in the context of photo diode data?
2. What does ”error rate” mean in the context of machine learning models?
3. What does ”real-time” mean in the context of gesture recognition?
4. How can we ensure that a given machine learning model will run in real-time on an Arduino Nano 33 BLE?
5. How can we determine which machine learning model yields the lowest error rate for gesture recognition?
6. How can we pre-process the 3D-formatted photo diode data to minimize machine learning model error rate in the context of gesture recognition?
7. Why it be useful for embedded devices to be able to process data using machine learning models?

The first of these sub-questions can be answered immediately as ”3D-formatting” is a term specific to this paper, and is best explained with the comparison to ”2D-formatting”. 2D-formatted data can simply be thought of as an image, with some horizontal resolution x and vertical resolution y. In the context of this project, the data we are concerned with is the readings from our photo diodes over some period of time during which the gesture is performed. This means that we can format said data as a 2D image in which x is the number of photo diodes we use and y is the number of samples we receive from the photo diodes in the aforementioned period of time, while the value of each ”pixel” in the image is a reading from a single photo diode at a single point in time. 3D-formatted data can meanwhile be thought of as a video, which splits this 2D-image into a sequence of n frames. 3Dformatting is generally more appropriate when the data is sensitive to time, i.e. when data points should be considered in a specific sequence. Therefore, by 3D-formatting the photo diode data, lower error rates should be achievable as the temporal information from the photo diodes isn't lost.

The remaining sub-questions will be answered in later sections of this paper, while the research question as a whole will be answered in the final section, using conclusions drawn from each individual sub-question.

The planned use case for this technology is hands-free navigation of menus which has especially gained relevance due to the restrictions imposed during the COVID-19 pandemic, which has showed that gestures can be an appealing alternative to physical buttons in social settings. The most important existing literature for this research comes from Pete Warden and Daniel Situnayake, who are pioneers in the field of embedded AI and authors of the book "TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers"[**?**], as well as Qing Wang and Marco Zuniga, who have laid the groundwork for embedded AI specifically in the context of hand gesture recognition[**?**]. However, there is still room for advancement regarding gesture recognition in particular - error rates, required processing power, and sensor counts, are all likely to be reduced due to this field being so new. The research conducted for this paper expands on existing work in this exact way, by using a lower power microcontroller and significantly fewer photo diodes than in the solution created by Wang and Zuniga, while maintaining or improving the classification accuracy of hand gestures.