15740 project proposal: ml-on-IoT

Yifei Yang(yifeiy3@andrew.cmu.edu)
Yang Yang(yangy4@andrew.cmu.edu)

September 2021

1 Project Github

https://youngy0y.github.io/ml-on-IoT/

2 Project Description

Nowadays, machine learning, especially deep learning, is one of the hottest topics in the world. In recent years, the size of the models are growing larger and consequently becoming computationally more expensive (BERT, a widely used pre-trained model for NLP tasks, has more than 300 million parameters). Therefore, lots of current researches focus on optimizing the models or building accelerators to speed up the application but, due to the near-infinite storage space we have in most situations due to cloud computing, the size of these deep learning models are often ignored. However, the issue of model size rises when resources from cloud computing is no longer available, particularly, when running on IoT devices with microcontroller units (MCU) which often has megabytes of storage and kilobytes of memory rather than terabytes and gigabytes we get through cloud computing. Given the limited computing power and memory size of MCU, running these complex ML algorithms directly would be be a very challenging task. On the other hand, being able to run these ML algorithms on such IoT devices are extremely useful, especially in many IoT AI applications such as personalized healthcare and automated retailing where we can then directly perform data analytics near the sensor of the devices.

In this project, we want to explore the possibility of running deep learning models on tiny IoT devices. We would like to first choose a typical MCU architecture (e.g. ARM Cortex-M7 MCU) and implement a reasonable ml task (e.g. image recognition, wake word detection, etc). Then, we will try to optimize code interpretation, memory scheduling and kernel specialization to increase the efficiency and speed. Finally, we will compare the result with using cloud to see if doing ML tasks on MCUs are realistic. We have listed our project goals below.

- 75% Goal: Finish implementing a deep learning task on MCU architecture and comparing our finished model with running task on cloud to check for performance of our model.
- 100% Goal: Finish 75% Goal, additionally, add in code interpretation, memory scheduling and kernel specialization to the model and analyze our performance improvement.
- 125% Goal: Finish 100% Goal, additionally, using the same implementation and optimization on other deep learning models to check for generality of our result.

3 Logistics

3.1 Plan of attack and schedule

Starting from today, we have about 10 weeks to finish our proposed project. As a result, we would like to complete our project through the following steps:

1. Week 1-2: Finding out on whether we are able to virtually simulate a MCU environment using our computer, and what kind of IoT device we should run our implementation on if virtual simulation is not possible.
2. Week 3-4: Learning the APIs needed for simulation or for running our model on physical IoT device and decide on a deep learning network to implement.
3. Week 5-7: Completing our model on the MCU architecture and evaluate our performances. (75% goal)
4. Week 7-10: Research and add on optimizations to finish our 100% goal. Work on implementing and evaluating our model for other deep learning models to reach our 125% goal if there is time.

For the initial weeks, Yang will be focusing on implementing the deep learning model with his research background in ML, and Yifei will be focusing on the IoT aspect of the project with his research background. Since both of us are fairly new in computer architecture topics, we would like to use our knowledge through this class and additional research to work on the optimizations together for the final step of our project.

3.2 Milestone

The milestone on Nov.3 is about 6 weeks away from the time this proposal is written. At that point, we hope to be close to finish our 75% goal. At the least, we would like to finish our model on the MCU architecture and getting ready to evaluate our performances.

3.3 Literature Research

Currently, we have done a limited amount of research in the field, with our idea solely being inspired by the work Mcunet from Lin et.al (`https://arxiv.org/pdf/2007.10319.pdf`) for deep learning on IoT devices. In the immediate future, we would like to research more on running ml programs under MCU environment and identifying good IoT MCU candidates for us to run our project. We would also want to find a deep learning model that is suitable for our project, which is large and complicated enough that it is not possible to implement on MCU with traditional methods, and not too complicated for the scope of our project. These needed literature research, according to our planned schedule, will be done by week 2.

3.4 Resources Needed

Currently, we are still in research on whether it is possible to virtually simulate a MCU running environment for our deep learning model. If such software is not applicable, we will purchase the IoT device we decided from literature research to physically run our model. Later on, we would also need to search for softwares for evaluating the performance of our model, but we are also able to implement such performance checks ourselves if such softwares are unavailable.

3.5 Getting Started

So far, we have not done too much on our proposed project other than reading the Mcunet paper. Currently, we are still researching on softwares for virtually simulate a MCU running environment or finding a physical device if simulation is not possible. If any course staff happens to know a way to virtually simulate a MCU environment, it will be extremely helpful. Thank you for your help in advance!