Enabling Speech to Text on Embedded Systems

Undergraduate Student Research: Team Project

Mohan Dodda
Georgia Tech
mdodda3@gatech.edu

Taejoon Park
Georgia Tech
tpark71@gatech.edu

Sayuj Shajith
Georgia Tech
sshajith3@gatech.edu

Ramyad Hadidi
Georgia Tech
rhadidi@gatech.edu

Hyesoon Kim
Georgia Tech
hyesoon@gatech.edu

ABSTRACT

Data obtained by sensors on IoT and embedded devices are private, but understanding these data with machine learning requires significant computational power. Since embedded devices cannot handle such computations efficiently, data is sent to remote servers for processing, which raises privacy concerns (e.g., recordings of users in their home). Therefore, it is important to be able to process the data in real-time on the embedded devices without remote servers or connections. An important facet is how users communicate with embedded devices. In this work, we focus on speech-to-text technology. We study the feasibility and the performance of current state-of-the-art machine learning methods for speech-to-text on embedded devices. Our experiments are done on RaspberryPi3 [Raspberry PI Foundation 2017].

Current speech-to-text models are computationally expensive to efficiently run on embedded devices. The first method we try is Mozilla DeepSpeech speech-to-text [Mozilla 2019]. The implementation is based on the Baidu deep speech model [Hannun et al. 2014] architecture which uses TensorFlow recurrent neural networks to interpret speech to text. We compiled a trained model of DeepSpeech model and execute it on RaspberryPi3. In our experiments, for the word "hello", the model performs in 2.8 seconds on a regular machine and in 75.6 seconds on RaspberryPi3, that is 27x slower.

The second method we implement is the updated version of Carnegie Mellon PocketSphinx framework [Huggins-Daines et al. 2006]. PocketSphinx is a lightweight model that could easily be run on embedded devices. This tool breaks speech down to the smallest discernible unit of speech called "Senones" and then uses Hidden Markov Models to determine which word was uttered by using the surrounding Senone. Hidden Markov Models can determine with a certain probability based on the sequence of states, what word was uttered. When we run the "hello" experiment with the vanilla version of PocketSphinx, decoding the word "hello" takes 10.3 seconds on RaspberryPi3. By adding the following extra optimizations that are specific for embedded devices, we can reduce the decoding time significantly: (1) making our dictionary to reduce the number of words in our dictionary, and (2) using our language model to simplify the model as well. These optimizations give us a 10x improvement in the time it took to say hello with a time of 1.6 seconds. Since the embedded devices are usually used for certain tasks, we can optimize the dictionary to only understand words that are relevant to the task that needs to be performed. Such targeted optimization for machine learning-based models, such as DeepSpeech, requires redesigning the model and retraining it.

In our poster and presentation, we will discuss all of our experiments on measuring the execution performance of speech-to-text models on RaspberryPi3, a server, and a regular machine. We will discuss the current optimizations that are possible to reduce their execution overhead on embedded devices. Finally, we will analyze why and how such optimizations help the execution performance on embedded devices (e.g., reducing the model size, relieving pressure on computing units).

© 2020 ACM. ISBN 978-1-4503-2138-9…$15.00
DOI: 10.1145/1235

REFERENCES

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* (2014).

David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alexander I Rudnicky. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. IEEE, I–I.

Mozilla. 2019. GitHub Mozilla/DeepSpeech. github.com/mozilla/DeepSpeech. (2019). [Online; accessed 7/13/19].

Raspberry PI Foundation. 2017. Raspberry Pi 3B. raspberrypi.org/products /raspberry-pi-3-model-b/. (2017). [Online; accessed 7/13/19].