임베디드 시스템에서 음성 기반 실시간 감성 인식

Paul Angelo Oroceo, 김정인, Alexander Pascual, 임완수
금오공과대학교 전자공학과

oroceopaul@kumoh.ac.kr, wjddls2208@kumoh.ac.kr, alexander.pscl@kumoh.ac.kr

Real-time Speech Emotion Recognition on Embedded system

Paul Angelo Oroceo, Jeongin Kim, Alexander Pascual, Wansu Lim
Kumoh National Institute of Technology

Abstract

Recognizing emotion is an important component of communication in human-to-human type of interaction. This emotional awareness has also been developed in machines using state-of-the-art techniques such as machine learning (ML) and deep learning (DL) methods. As the trend in the human-to-machine ML domain continues to be applied in different areas such as the Internet of things (IOT), the memory footprint of ML architecture is reduced for embedded system implementation. In this paper, we proposed an implementation of embedded speech emotion recognition (SER) in a microcontroller unit (MCU). Our proposed model recognizes the emotion of speech features categorized as: positive, negative, and neutral in real-time by learning and storing the model using a convolutional neural network (CNN). The CNN architecture for SER is mainly divided into (1) input layer, (2) hidden layer, and (3) output layer. SER is developed using TensorFlow lite which enables the on-device machine learning.
Keywords— speech emotion recognition, embedded system, microcontroller, internet of things, machine learning

I . INTRODUCTION

Emotion is one of many that a human considers as an experience in response to an event or situation. Understanding emotion is one key component in a conversation with respect to human-to-human encounter. In the past decade, advancement in the field of artificial intelligence paved the way in the development of emotion recognition algorithm, enabling computer or devices to mimic human behavior. Emotion recognition is becoming prominent in the domain of human-computer interaction [1]. Devices such as computers are now capable of detecting human emotions through various sources such as facial features, audio samples, psychological data and so on. In dealing with audio samples, speech emotion recognition (SER) algorithm is commonly employed for this task. SER aims to recognize emotions in an audio sample or speech sample reflected using CNN [2].

Recent studies achieved state-of-the-art results on SER by using feature representation of extracted features such as Mel-frequency cepstral coefficients (MFCC) [3, 4]. Generally, MFCC feature extraction includes windowing the signal. The idea behind MFCC is to convert an audio sample in time domain into frequency domain so that more features can be learned. MFCC is usually paired with CNN as its model structure that uses this feature extraction method to output different emotions such as neutral, happy, sad, angry, disgust, fear and so on. Earlier implementation of SER using CNN proposed the use of time frequency analysis, transforming speech signal into 2D representation using short time Fourier transform (STFT) [4] then passed through CNNs and long short-term memory (LSTM) architectures.

With the rise of Internet of Things (IOT), embedded systems took the spotlight in the implementation of machine learning where model inference is done locally. Machine learning, SER included, involves complex tasks which require ample processing power and memory for inferencing. Many researchers introduce optimization techniques such as pruning, quantization and Huffman coding to effectively reduce the ML model size and improve its execution in terms of time and power efficiency. However, even with these techniques, implementation of ML on embedded systems still requires further optimization or model compression.

In [5], they introduce the Tensorflow lite Micro (TFLM), an open-source ML inference framework for running deep-learning models on limited-resource embedded systems. TFLM grants a machine learning model to reduce power consumption or memory size since TFLM applies both quantization and weight pruning. Thus, TFLM framework supports direct

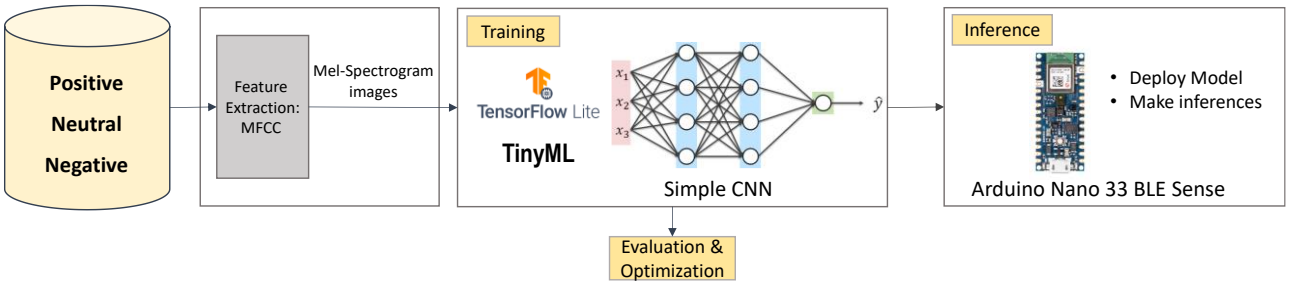

Figure 1. Overall flow of proposed real-time embedded system for speech emotion recognition.

software deployment on resource-limited embedded hardware.

In this paper, we propose an implementation of speech emotion recognition on embedded systems, utilizing TensorFlow lite (TFLite) library as shown in figure 1. To account for memory constraint, we also proposed an algorithm which have a FIFO (First in First Out) structure, where the system receives an audio input, cut it into smaller samples, output prediction in terms of an integer (1 = Positive, 2 = Neutral, 3 = Negative) then dispose the oldest sample. These proposed systems are implemented on a low-powered device.

Ⅱ. METHODOLOGY

2.1. Software

For training the speech emotion recognition model, SER technology first needs to establish a speech emotion database. Now the database we often use is Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) emotion database. RAVDESS is a multimodal library of songs and speeches with strong emotions. 24 professional actors with a neutral North American accent, representing a gender-balanced representation of the population, perform lexically matched sentences in the database. Both speech and music may display a range of emotions, including neutral, calm, happy, sad, angry, fearful, disgust and surprise. For simplified SER implementation for embedded system, we categorized the emotions into 3 classes: positive, neutral and negative. It is necessary to select appropriate speech feature parameters that effectively contain human emotion information. We extracted MFCC as a single feature from the input. Our proposed model recognizes the emotion of speech features in real time by learning and storing the model using a CNN. The CNN architecture for SER is mainly divided into input layer, hidden layer, and output layer. The input layer is responsible for the multi-dimensional feature data in tensor form. In the hidden layer, CNN learns the variety of MFCC features over time, which are important for emotion classification. The output layer is responsible for related classification tasks, and its upper layer network is usually composed of a fully connected layer. The machine learning model is developed using TensorFlow lite.

TensorFlow Lite is a collection of tools that allow developers to run their trained models on mobile, embedded, and IoT devices and computers, enabling on-device machine learning. It supports platforms such as embedded Linux, Android, iOS, and MCU. TensorFlow lite can generate a small TensorFlow model that can fit your target device such as Arduino Nano 33 BLE Sense, STM32, and ESP32.

2.2. Hardware

For the hardware, a microcontroller unit (MCU) is used to run the SER model. In our implementation, the Arduino nano 33 BLE Sense is a well-suited device, based on the Arm® MbedTM OS and the nRF52840 microcontroller. The Nano 33 BLE Sense is packed with sensors that can detect color, proximity, motion, temperature, humidity, and audio, in addition it can connect through Bluetooth® Low Energy. Given the nature of embeddable systems, its always-on configuration can benefit with the low-power consumption trait of MCU.

2.3. Expected results

Real-time speech emotion recognition on edge devices typically requires more processing power, which can be provided by cloud services or a much more powerful local device. However, utilizing external devices will result in a dependent speech emotion recognition system, wherein the microcontroller's ability to recognize emotion will be compromised in the moment that it loses contact with the cloud or another device. To overcome this issue and achieve an effective deep learning-based infrastructure on embedded systems, we need to use the appropriate tools in developing models efficient for edge devices. Many researchers have leveraged the use of frameworks such as TFLite and PyTorch Mobile for an efficient model training. For instance, TFLite improves the latency, model size, and deployment in a small, resource-limited embedded device. In this paper, we propose our own CNN architecture, reducing the output labels from 7 emotions into 3 emotion classifications (positive, neutral, negative) and will be trained using the TFLite framework. In this way, we can ensure the feasibility of embedding complex algorithms in relation to SER and achieve an efficient edge-based deployment in real-time applications.

Ⅲ. CONCLUSION

Implementing speech emotion recognition in embedded systems can require complex customization and optimization, and most of the time, not enough for a deployable system. In his study, the use of TinyML and TensorFlow lite is proposed for implementing machine learning models on low-powered devices such as mobile devices and microcontrollers. Optimization and quantization are applied into the model to reduce size and latency, leaving little to no loss in model accuracy. Also, an algorithm is used for SER inference where, using continuous inferencing, smaller sampling buffers (slices) are used and passed to the inference process. In the inferencing process the buffers are time sequentially placed in a FIFO (First in First Out) buffer. After each iteration, the oldest slice is removed at the end of the buffer and a new slice is inserted at the beginning, thus enabling low memory device for inferencing.

ACKNOWLEDGMENT

This work was supported by the Ministry of SMEs and Start-ups, S. Korea (S3010704), and by the National Research Foundation of Korea (2020R1A4A101777511, 2021R1I1A3056900).

REFERENCES

[1] D. Joshi, A. Dhok, A. Khandelwal, S. Kulkarni and S. Mangrulkar, "Real Time Emotion Analysis (RTEA)," 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), pp. 1-5, Jan. 2022.

[2] C. Jie, "Speech emotion recognition based on convolutional neural network," 2021 International Conference on Networking, Communications, and Information Technology (NetCIT), 2021, pp. 106-109

[3] Araño, K.A., Gloor, P., Orsenigo, C. et al., “When Old Meets New: Emotion Recognition from Speech Signals,” Cogn. Comput., vol. 13, pp. 771-783, Apr. 2021.

[4] N. Kumar, R. Kaushal, S. Agarwal and Y. B. Singh, "CNN based approach for Speech Emotion Recognition Using MFCC, Croma and STFT Hand-crafted features," 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 981-985, Dec. 2021

[5] R. David, J. Duke, A. Jain, V. J. Reddi, N. Jeffries, J. Li, N. Kreeger, I. Nappier, M. Natraj, S. Regev, R. Rhodes, T. Wang, and P. Warden, ‘‘TensorFlow lite micro: Embedded machine learning on TinyML systems,’’ CoRR, vol. abs/2010.08678, pp. 1-12, Oct. 2020.