D22-069

應用於智慧終端裝置之高能效語音辨識加速器晶片

An Energy-Efficient Speech Recognition Accelerator IC for Intelligent Edge Devices

隊伍名稱　正港臺中人
Taichung Real Man
隊　　長　蔡宇軒 / 臺灣大學電子工程學研究所

指導教授

楊家驤　臺灣大學電機工程學系

美國加州大學洛杉磯分校電機博士，現為臺灣大學電機工程學系教授。實驗室致力於開發低功耗之客製化晶片以提升資料處理速度與能量效率。

研究領域

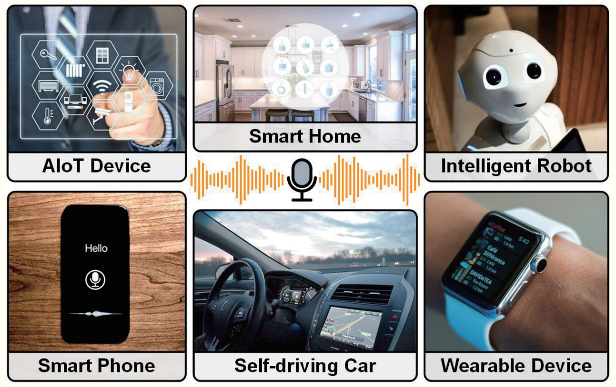
AI 晶片設計、基頻通訊積體電路、生醫訊號處理晶片設計

作品摘要

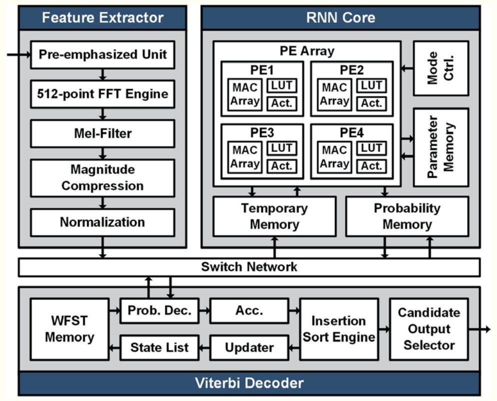
語音辨識是一個將語音訊號轉換成文字的一項技術。它在人工智慧相關領域有廣泛應用，包括穿戴式裝置、智慧型手機、智能家庭、智慧機器人與自駕車等。由於機器學習技術的快速演進，近年來的辨識準確度有著顯著成長。然而語音辨識系統中的高運算複雜度，使得要在邊緣裝置進行運算，會有延遲與能耗的設計挑戰。因此需要一個專門的語音辨識加速器來解決這些問題。

這次參賽的設計展示了一個專門處理基於遞歸神經網路語音辨識系統的高能效加速器。整個語音辨識流程由三個模組組成，包括特徵萃取、遞歸神經網路和集束搜索。輸入的語音訊號先經過濾波器組進行特徵萃取，其中的運算包括快速傅立葉轉換與梅爾濾波器濾波。遞歸神經網路則根據萃取的特徵來計算觀察機率。本設計採用輕量門控遞迴單元，在使用更少的參數與運算量的同時，達成更好的辨識結果。在集束搜索部分，採用維特比解碼進行實現，根據由聲學模型、辭典與語言模型三個模型構成的加權有限狀態轉換器來找出最有可能的輸出結果。

經過網路壓縮流程後，遞歸神經網路大小有顯著壓縮，可以放在晶片上不需要外部記憶體存取。透過演算法與架構層級優化，使得功耗與面積最小化。透過28-nm製程下線，本晶片包含9.42 M個邏輯閘。本晶片實現即時的語音辨識並只消耗少於2 mW的功耗。相比於過去的最佳設計，本設計達到37.5倍短的最低延遲，並達到6.5倍低的正規化能耗。本晶片同時達到最低的音素錯誤率。不像之前的作品，本設計所有的參數都存在晶片上，節省從外部記憶體存取所造成的能耗與延遲。



圖一 語音辨識應用



圖二 本設計之系統架構

Abstract

Speech recognition is a technique that translates the spoken languages into text. It can be applied to various AI applications, including wearable devices, smartphones, smart homes, intelligent robots, and self-driving cars. Thanks to the development of deep learning, the accuracy of speech recognition has rapidly improved in recent years. However, the high computational complexity poses design challenges to edge devices in latency and energy consumption. Hence, a dedicated speech recognition accelerator is required to address these issues.

This work presents an energy-efficient dedicated accelerator for recurrent neural network (RNN)-based hybrid speech recognition. The overall system is comprised of three blocks: feature extraction block, RNN, and beam search block. The input speech signal is first processed by a filter bank (FBANK) to extract features through fast Fourier transform (FFT) and Mel-filter. The RNN is used to compute the observation probability based on the extracted features. The light gated recurrent unit (LiGRU) is adopted to achieve a high accuracy with less computations and fewest parameters. For beam search, Viterbi decoding is applied on the weighted finite-state transducer (WFST), in which three models (acoustic model, lexicon model, and language model) are included, to find the most-likely sequence.

By applying network compression, the RNN model is compressed significantly so that it can be included on the chip without the need for an external memory. Power and area are minimized through optimization across the algorithm and architecture layers. The chip integrates 9.42M logic gates in in a 28-nm CMOS technology. The chip delivers real-time speech recognition and dissipates less than 2 mW. Compared to prior arts, this work achieves a 37.5× lower attainable latency with 6.5× lower normalized energy. The chip also achieves the lowest phone error rate. Unlike prior works, all the model weights are stored on the chip, saving more energy and latency when external memory access is considered.

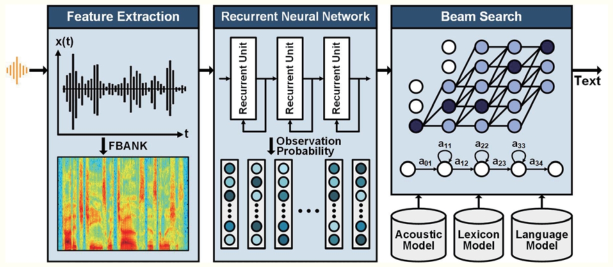


Fig. 3 RNN-based hybrid speech recognition system

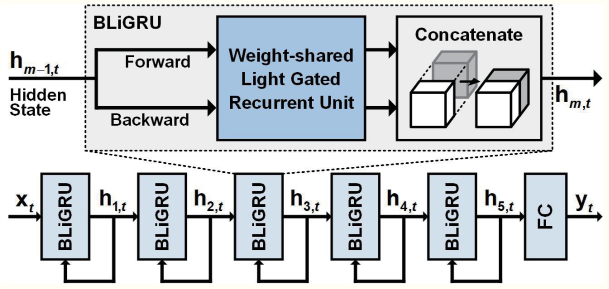


Fig. 4 The RNN model in this work