임베디드 시스템을 위한
심층신경망 기반 스트리밍 음성 인식 모델

안성환, 우범준, 이동준, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실

{shahn, bjwoo, djlee}@hi.snu.ac.kr, nkim@snu.ac.kr

Streaming Automatic Speech Recognition System Based On Deep Neural Network For Embedded Systems

Sung Hwan Ahn, Beom Jun Woo, Dongjune Lee, Nam Soo Kim
Human Interface Laboratory,
Department of Electrical and Computer Engineering and INMC,
Seoul National University

요 약

본 논문은 경량화된 심층신경망 기반 실시간 음성인식 모델을 제안한다. 1 차원 합성곱 레이어 및 제한된 비선형 함수만으로 심층신경망을 구성하여 자원이 제한된 임베디드 시스템에서도 동작하며 음성인식 성능이 우수하다.

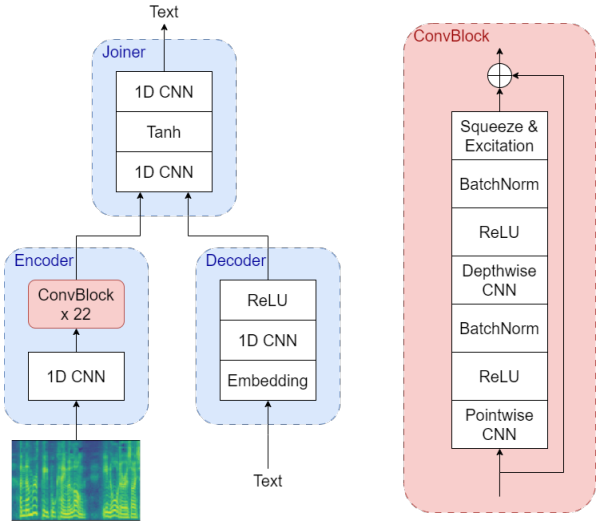
Ⅰ. 서 론

Streaming 음성인식은 음성 입력이 미리 지정한 chunk 크기로 들어올 때 그 음성에 해당하는 문자로 실시간으로 변환해주는 시스템이다. 최근 심층신경망 모델의 발전과 함께 streaming 음성인식 시스템의 성능도 크게 증가하였다. 그중 가장 성능이 좋은 transformer 기반의 모델[1]은 softmax 또는 sigmoid 등의 비선형 연산이 필수적으로 들어가며, 상당한 연산량 및 메모리를 필요로 하기 때문에 on-device 및 임베디드 시스템에서 사용되기에는 한계가 있다. 자원이 부족한 상황에 적합한 convolutional neural network (CNN) 기반의 모델 연구도 이루어지고 있지만, 실시간 동작이 불가능하거나[2] 1 차원 CNN 외의 다른 레이어도 사용된다[3]. 임베디드 시스템에 심층신경망 기반의 음성인식 모델을 탑재하기 위해서는 메모리 사용량과 연산량이 적을 뿐만 아니라, 해당 시스템의 하드웨어가 빠르게 처리할 수 있는 연산만으로 이루어진 모델을 설계하는 것이 중요하다.

본 연구는 1 차원 CNN, 병렬 덧셈 및 곱셈, ReLU 와 hyperbolic tangent 연산을 빠르게 처리할 수 있는 가속기가 탑재되었으며, 다른 연산에 대해서는 속도가 느린 CPU 로 처리해야 하는 임베디드 시스템을 타겟으로 하는 음성인식 시스템을 제안한다. LibriSpeech [4] 데이터셋으로 학습 및 검증한 결과 기존의 경량화된 실시간 음성인식 모델 대비 우수한 성능을 보였다.

Ⅱ. 본론

제안하는 모델은 [그림 1]과 같다. Mel spectrogram 입력이 640ms chunk 단위로 순차적으로 들어오면



[그림 1] (왼쪽) 음성인식 모델 구조. (오른쪽) ConvBlock 구조. Tanh 는 hyperbolic tangent 를 의미.

실시간 음성인식을 수행한다. Encoder 는 22 개의 CNN Block 으로 구성되어 있다. CNN Block 은 [3]의 global encoder 와 비슷한 구조이다. 모든 CNN Block 의 입/출력 채널은 384 이다. CNN Block 은 pointwise CNN, depthwise CNN, ReLU, batch norm, Squeeze & Excitation (SE)으로 구성되어 있다. 첫 번째 pointwise CNN 은 채널 크기를 1536 으로 늘리고, 마지막 pointwise CNN 은 채널 크기를 384 로 줄인다. 실시간 동작을 위해 모든 CNN 과 SE 는 causal 로 구성했다. SE 에서 채널 크기 감소율은 1/8 으로 설정했으며, gate 함수로는 일반적으로 사용되는 sigmoid 대신 가속기가 지원하는 연산인 hyperbolic tangent 을 사용했다.

	파라미터 개수 (M)	test-clean	test-other
ConvRNN-T	29	5.11	13.82
Proposed	29.2	4.72	12.28

[표 1] LibriSpeech test 데이터셋에서의 WER 성능

Decoder 는 stateless transducer[5]을 사용했으며, embedding layer, causal CNN, ReLU 로 이루어져 있다. Joiner 는 pointwise CNN 과 hyperbolic tangent 로 구성하였다.

본 모델에 사용된 레이어 종류를 종합하면 1D CNN, batch norm, ReLU, hyperbolic tangent 이다. Batch norm 은 학습이 완료된 후에는 병렬 덧셈/곱셈 연산으로 대체하여 batch norm 이후에 나오는 CNN 과 합칠 수 있으므로 모든 레이어는 가속기로 연산이 가능하다. 모델의 총 파라미터 개수는 29.2 Million 개가 되도록 구성하였다.

학습 데이터셋으로는 LibriSpeech 의 train-clean 과 train-other set 을 사용하였다. Optimizer 와 learning rate scheduler 는 [1]에서 제안된 Eve 와 Eden 을 각각 사용하였으며, initial learning rate 은 0.001 으로 설정했다. 3090 GPU 4 대로 총 84 시간동안 학습하였다. 학습 이후 weight averaging 기법을 적용하여 성능을 검증하였다. 성능 검증은 word error rate (WER)을 측정하였다. 비교 모델로는 1 차원 및 2 차원 CNN, LSTM 으로 이루어진 실시간 음성인식 모델인 ConvRNN-T 모델[3]을 사용하였으며, 해당 논문에 보고된 결과를 가져왔다.

실험 결과는 [표 1]과 같다. ConvRNN-T 모델과 거의 동일한 파라미터 개수 하에서 LibriSpeech test-clean 과 test-other 모두에서 더 낮은 WER 을 달성하였다. 또한, ConvRNN-T 모델과는 다르게 오직 1 차원 CNN, ReLU, hyperbolic tangent 함수만을 이용하였기 때문에 임베디드 시스템에서 실시간 동작이 가능하다.

Ⅲ. 결론

본 논문은 성능이 우수할 뿐 아니라, 특정 제한점을 지닌 임베디드 시스템에 탑재할 수 있는 새로운 심층신경망 기반 실시간 음성인식 시스템을 제안했다. 누군가 실제 서비스를 위해 본인의 임베디드 시스템에 음성인식 시스템을 탑재하고자 한다면, 본 모델은 좋은 baseline 으로 활용될 수 있을 것이다.

ACKNOWLEDGMENT

이 논문은 2023 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] Zengwei Yao et al., “Zipformer: A faster and better encoder for automatic speech recognition,” arXiv:2310.11230, 2023

[2] Wei Han et al., “ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context,” Interspeech, 2020

[3] Martin Radfar et al., “ConvRNN-T: Convolutional Augmented Recurrent Neural Network Transducers for Streaming Speech Recognition,” Interspeech, 2022

[4] Vassil Panayotov et al., “Librispeech: An ASR corpus based on public domain audio books,” ICASSP, 2015

[5] Mohammadreza Ghodsi et al., “Rnn-Transducer with Stateless Prediction Network,” ICASSP, 2020