레벤슈타인 트랜스포머 기반 빠른 추론이 가능한 음성 인식 모델에 관한 연구

이동준, 강주연, 한진모, 홍지민, 김남수
서울대학교

{djlee, jykang, jmhan, jmhong}@hi.snu.ac.kr, nkim@snu.ac.kr

Towards Fast Decoding in Speech Recognition via Levenshtein Transformer

Dongjune Lee, Ju Yeon Kang, Jinmo Han, Ji Min Hong and Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ.

요 약

본 논문은 Levenshtein Transformer 를 기반으로 한 새로운 음성 인식 모델을 제안한다. 본 논문에서는 자기 회귀 방식으로 동작하는 일반적인 음성 인식 모델과 대비하여 비자기 회귀(non-autoregressive) 방식을 적용한 음성 인식 모델을 제안한다. 제안하는 모델은 추론 시간을 대폭 줄일 수 있으며, 자기 회귀 방식 기반으로 삽입(insertion) 및 삭제(deletion) 기능이 추가되어 초기 예측이 부정확할 경우 이를 수정할 수 있는 유연성을 제공한다. 이러한 기능은 모델이 초기 오류로부터 빠르게 회복하도록 도와주며, 전체적인 정확도를 향상시키는 데 기여한다. 본 논문에서 제안하는 모델은 기존 음성 인식 모델들이 가진 추론 속도와 정확도의 한계를 극복할 수 있는 새로운 대안을 제시한다.

Ⅰ. 서 론

본 논문은 Levenshtein Transformer 를 기반으로 한 새로운 음성 인식 모델을 제안한다. 기존의 음성 인식 시스템은 일반적으로 자기 회귀(autoregressive) 방식을 채택하고 있으며, 각 토큰을 순차적으로 생성해야 하므로 추론 속도가 상대적으로 느린 단점이 있다. 또한, 이 방식은 추론 과정에서 이전 예측이 잘못되면 그 오류가 누적되어 전체 예측 결과의 정확도를 저하시키는 문제를 가지고 있다. 이에 대한 대안으로, 본 논문에서는 비자기 회귀(non-autoregressive) 방식을 적용한 음성 인식 모델을 제안한다. 이 모델은 추론 시간을 대폭 줄일 수 있으며, 자기 회귀 방식 기반으로 삽입(insertion) 및 삭제(deletion) 기능이 추가되어 초기 예측이 부정확할 경우 이를 수정할 수 있는 유연성을 제공한다. 이러한 기능은 모델이 초기 오류로부터 빠르게 회복하도록 도와주며, 전체적인 정확도를 향상시키는 데 기여한다. 본 논문에서 제안하는 모델은 기존 음성 인식 모델들이 가진 추론 속도와 정확도의 한계를 극복할 수 있는 새로운 대안을 제시한다.

Ⅱ. 본론

본 논문에서는 Levenshtein Transformer decoder 를 기반으로 새로운 음성 인식 학습 및 추론 방식을 채택한 음성 인식 모델을 제안한다. 일반적인 Transformer 기반의 음성 인식 모델[1]의 학습 및 추론 방식은 자기 회귀(autoregressive) 방식이며, 이는 이전까지의 출력 토큰을 기반으로 다음 토큰을 예측하는 방식으로 이루어진다. 이에 추론 과정 시 디코딩 (decoding) 속도가 느리며, 과거의 출력에 오류가 존재할 경우, 해당 오류가 누적되어 잘못된 음성 인식 예측 결과를 출력한다는 단점이 존재한다.

이에 본 논문에서는 Levenshtein Transformer[2]를 이용하여 기존 음성 인식 모델의 decoder 를 변경하여, 기존과는 달리 더 빠른 decoding 속도를 보이며, 예측 출력 결과에 오류가 존재할 시 이를 수정할 수 있는 음성 인식 모델(Levenshtein ASR)을 제안한다.

Levenshtein ASR 은 기존 음성 인식 모델[3] 과 비교하여 decoder 구조 및 토큰 생성 과정에서 차이점을 보인다. 구체적으로, 구조적으로는 각 Transformer decoder 의 layer 에 Insertion Classifier(Placeholder classifier, Token Classifier), Deletion Classifier 가 존재한다. Insertion Classifier 는 출력된 시퀀스 안에서 추가 되어야할 토큰 및 해당 토큰 위치를 추정하며, Deletion Classifier 는 출력된 시퀀스 안에서 생성된 각 토큰이 필요한지 여부를 판정하는 역할을 한다. 이와 같은 구조를 기반으로, Levenshtein ASR 은 출력 시퀀스 생성 시 자기회귀방식이 아닌 비자기회귀방식으로, 이전까지 생성된 과거 출력 시퀀스를 요구하지 않아, 더 빠른 속도로 음성 인식이 가능하다. 또한, 필요 시 반복(iteration)을 통해 불필요한 토큰 삭제 또는 필요한 토큰 삽입을 통하여 잘못된 예측 시퀀스 출력을 바로잡을 수 있다. 이와 같은 과정은 이전까지 출력된 전체 시퀀스를 기반으로 그 다음 보정된 출력 시퀀스를 출력하는 방식으로 자기 회귀 방식으로 이루어진다.

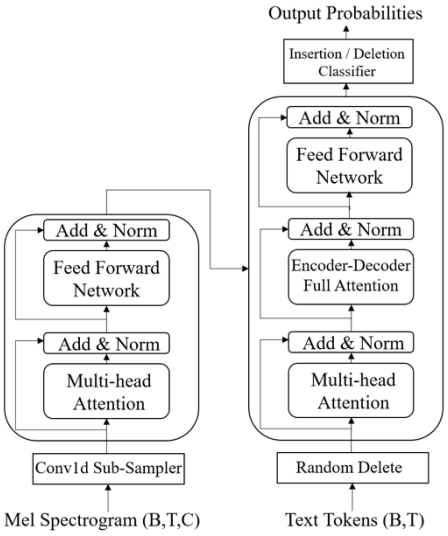

그림 1. 제안하는 모델 구조

III. 실험 및 결과

본 실험에서 기본 베이스라인 모델로 30M 규모의 Conformer 기반의 음성 인식 모델을 사용하였다. 제안하는 모델은 베이스라인과 동일한 Conformer Encoder 와 Levenshtein Transformer decoder 로 구성하였다. 제안하는 모델과 베이스라인 모델의 가장 큰 차이점은 베이스라인 모델은 추론 시 자기 회귀 방식으로 진행된다는 점이다. 본 실험에서는 약 1,000 시간 규모의 LibriSpeech 데이터셋을 학습에 사용하였으며, LibriSpeech Test Clean 데이터셋을 평가 및 비교에 사용하였다. Levenshtein ASR 은 기본적으로 비자기회귀 기반의 모델이므로 자기 회귀 기반 모델인 베이스라인보다 약 4~5 배 빠른 추론 속도를 보였다. 이는 모델 추론 과정이 parallel 하게 진행되어, 자기 회귀 방식의 추론 과정보다 더 빠른 추론이 가능함을 알 수 있었다. 또한, greedy decoding 방식 임에도 삽입(insertion)과 삭제(deletion)가 가능하여 잘못된 토큰이 삽입된 경우, 다음 과정에서 이를 삭제하여 지워내는 과정을 통해 greedy decoding 방식의 추론 과정임에도 베이스라인 모델에 가까운 성능을 낼 수 있었다.

표 1. Results

	Baseline	Levenshtein ASR
Decoding Strategy	Autoregressive	Non-autoregressive (Almost)
Dataset	Librispeech Test Clean	
# Params	30M	30M
WER (%)	**4.4**	6.4
# Iteration	-	2
Inference Speed	70.59 sentences /s 1706.40 tokens /s Total: 37.1s	**333.53 sentences /s** **8641.80 tokens /s** **Total: 7.9s**

IV. 결론

본 논문에서는 Levenshtein Transformer 기반의 음성 인식 모델을 제안한다. 본 제안하는 모델은 기존의 음성 인식 모델과 학습 및 추론 방식에 있어 비자기회귀 방식으로 진행되어 자기 회귀 방식의 기존 음성 인식 모델보다 더 빠른 추론이 가능하며, 자기 회귀 기반의 insertion 및 deletion 기능을 차용하여, 비자기회귀한 방식을 보완 가능하여 추후 음성 인식 모델의 새로운 추론 방식으로 활용할 수 있는 가능성을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2024 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음.

참 고 문 헌

[1] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." International Conference on Machine Learning. PMLR, 2023.

[2] Gu, Jiatao, Changhan Wang, and Junbo Zhao. "Levenshtein transformer." Advances in neural information processing systems 32, 2019.

[3] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R. (2020) Conformer: Convolution-augmented Transformer for Speech Recognition. Proc. Interspeech 2020, 5036-5040.