RNN-T 기반 음성 인식에서 최적화 속도 향상을 위한 가중치 선택 기법

한진모, 김석민, 정명훈, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실
{jmhan, smkim, mhjeong}@hi.snu.ac.kr, nkim@snu.ac.kr

Weight Selection for Faster Optimization in RNN-T Speech Recognition

Jinmo Han, Seokmin Kim, Myeonghun Jeong, Nam Soo Kim
Human Interface Laboratory,
Department of Electrical and Computer Engineering and INMC,
Seoul National University

요 약

본 연구에서는 RNN-T 기반 음성 인식 시스템에서 가중치 선택 초기화가 최적화 속도에 미치는 영향을 탐색한다. 이 기법은 사전 학습된 교사 모델의 가중치를 활용하여 학습자 모델의 초기값을 설정함으로써 모델 성능의 최적화 속도를 가속화한다. 실험 결과는 제안 기법이 기존의 초기화 기법보다 빠른 수렴을 유도함을 입증한다. 이 연구는 가중치 선택 기법이 자원 제약이 있는 실시간 음성 인식 시스템의 효율적인 성능 최적화를 달성하는 방법이 있음을 시사한다.

Ⅰ. 서 론

실시간 음성 인식 시스템에서는 계산 효율성을 위해 작은 모델을 효과적으로 훈련하는 것이 요구된다. 본 연구에서는 RNN-T(순환 신경망 변환기) 기반 음성 인식 시스템의 훈련 과정을 가속화하기 위한 가중치 선택 초기화 기법을 제안한다. 이 기법은 사전 학습된 교사 모델의 가중치를 활용하여 학습자 모델의 초기화 성능을 향상시키며, 학습 속도를 가속화한다.

본 연구는 교사 모델의 그래디언트가 이미 0 에 가깝다는 가정을 바탕으로 가중치 선택 기법이 최적화를 촉진하는 과정에 관한 직관을 제공한다. 또한, 실험을 통해 가중치 선택 기법이 실제로 훈련을 가속함을 입증한다. Librispeech 데이터셋에서 수행된 실험 결과는 가중치 선택 기법이 초기 학습 단계에서 수렴을 상당히 가속화함을 보여준다. 특히 초기 10 에폭에서 전통적인 무작위 초기화보다 더 빠른 Word Error Rate(WER) 감소를 보인다.

이 연구는 가중치 선택 기법이 자원 제약이 있는 실시간 음성 인식 시스템에서 최적화 효율성을 개선할 수 있음을 강조한다.

Ⅱ. 본론

1. 관련 연구

가중치 초기화는 모델의 훈련 과정에서 중요한 요소로, 수렴 속도와 학습 안정성에 직접적인 영향을 미친다. 전통적인 가중치 초기화 기법은 그래디언트 흐름을 유지하고, 그래디언트 폭발 또는 소멸 문제를 방지하는 것을 목표로 한다. 주요 기법으로는 레이어 간 신호 분산을 균형 있게 유지하여 그래디언트 문제를 완화하는 Xavier 초기화[1], ReLU 활성화 함수를 사용할 때 적절한 가중치 스케일링을 통해 수렴 속도를 최적화하는 He 초기화[2] 등이 있다.

이러한 기법들은 내부 모델 구조를 활용하여 최적화를 수행하지만, 교사 모델의 사전 학습된 가중치와 같은 외부 정보를 활용하지 않는다. 반면, 가중치 선택(Weight Selection)은 교사 모델의 특정 가중치 부분을 선택하여 학습자 모델을 초기화함으로써, 대규모 모델에서 소규모 모델로 지식을 전이한다[3]. 기존 연구에서는 교사 모델에서 학습자 모델에게 필요한 층을 선택하는 층 선택(Layer Selection) 및 각 층의 가중치 요소를 선택하여 학습자 모델에 맞게 조정하는 요소 선택(Element Selection)과 같은 기법이 훈련 시간 단축과 정확도 향상에 기여할 수 있음을 입증하였다.

2. 연구 방법론

가중치 선택(Weight Selection)은 사전 학습된 교사 모델의 가중치를 활용하여 학습자 모델의 가중치 값을 초기화하는 방법이다. 교사 모델의 가중치를 선택하는 규칙을 미리 정의한 뒤 이에 따라 교사 모델의 일부 가중치를 선택하거나 변환하여 학습자 모델을 초기화하는 방식으로 이루어진다. 초기화된 학습자 모델은 더 안정적이고 효율적인 최적화 경로를 따를 수 있다.

가중치 선택이 모델 학습을 가속화하는 이론적 직관은 교사 모델의 안정적인 그래디언트 특성을 학습자 모델에 전이하는 데서 찾을 수 있다. 교사 모델은 이미 최적화된

상태에서 안정적인 그래디언트를 가진다. 가중치 선택 기법은 이러한 안정적인 특성을 유지하도록 초기화되기 때문에 학습자 모델의 초기 최적화 단계에서 그래디언트의 크기 변화가 최소화된다. 이로 인해 학습자 모델은 불필요한 탐색을 줄이고, 안정적인 경로를 따라 수렴하게 된다.

3. 실험

제안된 기법의 효과를 평가하기 위해, RNN-T(순환 신경망 변환기) 기반 음성 인식 모델을 대상으로 실험을 수행하였다. 교사 모델(76M 파라미터)과 학습자 모델(9M 파라미터) 모두 RNN-T 구조로 구성하였다. 가중치 선택(Weight Selection, WS)의 구체적인 규칙은 층 선택(Layer Selection)의 경우 앞에서부터 순서대로 연속된 층을 선택하도록 하였다. 각 층 내의 요소 선택(Element Selection)의 경우 균등한 간격으로 선택하는 방식으로 정의하였다.

WS 초기화와 무작위 초기화(Random Initialization)를 학습 에폭별 단어 오류율(WER)을 기준으로 비교하였다. 모델의 성능 평가는 단어 오류율(WER)을 기준으로 수행하였으며, 이 값이 낮을수록 모델의 성능이 좋음을 의미한다.

데이터셋은 공개 도서 오디오북으로 구성된 1,000 시간 분량의 Librispeech 데이터셋을 사용하였다[4]. 이 데이터셋은 잡음이나 왜곡이 적은 깨끗한 음성 데이터(clean)과, 잡음과 왜곡이 포함된 어려운 음성 데이터(other)로 구분된다.

WS 초기화와 무작위 초기화 후 테스트셋에서 비교한 에포크별 WER 값은 다음과 같다.

Epoch	무작위 초기화	WS 초기화
5	8.15	4.86
10	5.56	4.30
20	4.15	3.96
30	3.73	3.72
40	3.58	3.60
50	3.52	3.53

표 1: Clean 데이터셋에서의 WER 값

Epoch	무작위 초기화	WS 초기화
5	17.68	12.07
10	12.84	10.60
20	10.19	9.82
30	9.42	9.22
40	8.83	8.92
50	8.70	8.89

표 2: Other 데이터셋에서의 WER 값

WS 초기화는 초기 에폭에서 무작위 초기화보다 현저히 낮은 WER 을 달성하였다(예: 5 에폭, Test-Clean: 4.86%, Test-Other: 12.07%). 실험 결과는 가중치 선택 초기화가 초기 수렴 속도와 최종 성능을 모두 향상시킴을 입증한다. 결과적으로, 가중치 선택 초기화는 최적화 가속과 효율성 향상을 위한 효과적인 전략임을 확인하였다.

Ⅲ. 결론

본 연구는 가중치 선택 초기화(Weight Selection Initialization) 기법이 학습 초기화 성능을 향상시키는 데 효과적임을 실험적으로 입증하였다. Librispeech 데이터셋과 RNN-T 모델을 활용한 결과, 제안된 방법은 무작위 초기화와 비교하여 더 빠른 단어 오류율(WER) 감소를 달성하였다. 이러한 결과는 특히 자원 제약이 있는 음성 인식 시스템에서 교사 모델의 가중치를 활용하여 모델의 학습 속도를 개선할 수 있음을 입증한다.

한편, 본 연구는 RNN-T 모델에 초점을 맞췄지만, 다른 아키텍처에 대한 추가 검증을 통해 가중치 선택 초기화의 일반화 가능성을 확인하는 연구가 이뤄질 것으로 기대한다. 한편, 최적화 경로에 대한 가중치 선택 초기화의 영향을 더 깊이 탐구하고, 수렴 속도를 예측할 수 있는 수학적 모델을 탐색하는 것 또한 후속연구로 제안한다.

ACKNOWLEDGMENT

이 논문은 2024 년도 BK21 FOUR 정보기술 미래인재

교육연구단에 의해 지원되었음.

참 고 문 헌

[1] Glorot, X., & Bengio, Y., "Understanding the difficulty of training deep feedforward neural networks," Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249- 256, 2010.

[2] He, K., Zhang, X., Ren, S., & Sun, J., "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1026- 1034, 2015.

[3] Xu, Z., Chen, Y., Vishniakov, K., Yin, Y., Shen, Z., Darrell, T., Liu, L., & Liu, Z., "Initializing models with larger ones," The Twelfth International Conference on Learning Representations, 2023.

[4] Panayotov, V., Chen, G., Povey, D., & Khudanpur, S., "LibriSpeech: An ASR corpus based on public domain audio books," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206- 5210, 2015.85.