음성 인식 모델 경량화를 위한 Teacher Decoder 재사용 기반 지식 증류 기법

강주연, 김세민, 이동준, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실

{jykang, smkim21, djlee}@hi.snu.ac.kr, nkim@snu.ac.kr

Knowledge Distillation Based on Teacher Decoder Reuse for Lightweight Speech Recognition Models

Ju Yeon Kang, Semin Kim, Dongjune Lee, Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ.

요 약

딥러닝 기반의 음성 인식 모델은 높은 계산 비용과 메모리 요구량으로 인해 자원이 제한된 환경에서 이용의 어려움을 겪는다. 이러한 문제를 해결하기 위해, 음성 인식 모델의 경량화를 위한 다양한 지식 증류 (Knowledge Distillation) 기법이 제안되었다. 본 논문에서는 Teacher Decoder 재사용 기반의 지식 증류 기법을 제안하고 실험을 통해 제안된 기법이 음성 인식 모델의 성능을 효과적으로 개선함을 확인하였다.

Ⅰ. 서 론

딥러닝의 발전으로 End-to-End 음성 인식 모델들은 큰 성능 향상을 이루었다. 그 중 Connectionist Temporal Classification (CTC) [1] 기반의 모델은 단순한 구조와 빠른 추론 속도로 인해 주목받고 있다. CTC 모델은 acoustic feature 를 모델링하는 encoder 와 분류를 수행하는 linear layer 인 decoder 로 이루어져 있다. 또한 non-autoregressive 방식으로 출력 토큰을 예측해 추론 속도가 빠르다. 이는 RNN-T [2] 혹은 Attnetion 기반 모델 [3]과 같은 autoregressive 모델과 비교해 큰 장점을 지닌다.

그러나 CTC 기반 모델은 높은 계산 비용과 메모리 요구량 때문에 자원이 제한된 환경에서 활용이 어려운 단점을 가지고 있다. 이러한 문제를 해결하기 위한 주요 접근법 중 하나는 모델의 크기를 줄여 경량화를 이루는 것이다. 이를 위해, 다양한 지식 증류 (Knowledge Distillation) 기법들이 등장하고 있다. 지식 증류는 모델의 사이즈가 크고 성능이 좋은 Teacher 모델이 학습한 지식을 사이즈가 작은 Student 모델에게 전달함으로써, Student 모델의 성능을 효율적으로 향상시키는 방법이다. CTC 기반 모델에서는 일반적으로 student 모델이 ground truth label 뿐만 아니라 Teacher 모델이 제공하는 soft label 을 활용하여 학습을 진행한다. 이러한 지식 증류 기법은 Student 모델이 Teacher 모델의 풍부한 정보를 학습할 수 있도록 돕는다.

본 논문에서는 Teacher 모델의 decoder 를 재사용함으로서 Teacher 모델의 분류 지식을 직접적으로 이용하는 지식 증류 기법을 제안한다. 제안된 기법은 feature-level distillation (1st stage)과 softmax-level distillation (2nd stage)을 통한 2-stage 학습 방식을 통해 Student 모델이 Teacher Decoder 의 분류 지식을 효과적으로 이용할 수 있도록 한다. 실험을 통해 Student 모델의 성능 향상과 제안 기법의 효과를 확인하였다.

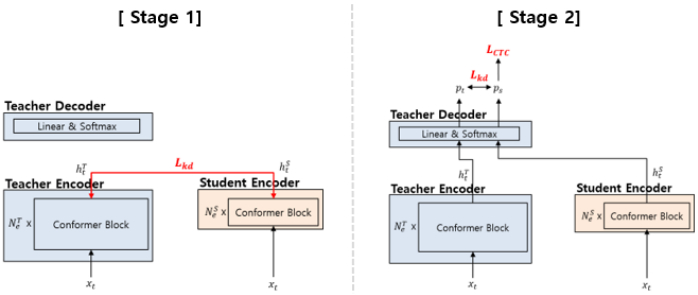

그림 1. 제안 기법의 2-stage 학습 방식

Ⅱ. 본론

본 논문에서는 2-stage 학습 방식을 이용한 Teacher Decoder 재사용 지식 증류 기법을 제안한다. CTC 기반의 모델들은 acoustic feature 를 모델링하는 encoder 와 classifier 역할을 하는 decoder 로 구성되어 있다. 제안 기법은 Teacher 의 decoder 를 재사용함으로써 Teacher 의 분류 능력을 이용한다.

첫번째 stage 에서는 Teacher encoder output 과 Student encoder output 간의 feature-level distillation 을 진행한다. L2 손실 함수를 통해 feature matching 을 함으로써 Teacher 모델과 Student 모델의 feature 가 align 될 수 있도록 한다.

두번째 stage 에서는 Teacher 모델의 softmax output 과 Student 모델의 softmax output 간의 softmax-level distillation 을 진행한다. Student softmax output 의 경우, Student encoder 와 Teacher decoder 를 이용하여 출력한다. 학습 과정에서 Student encoder 는 Teacher decoder 의 gradient 를 통해 학습함으로서 Teacher decoder 의 분류 지식을 활용할 수 있다. Softmax-level distillation 은 CTC 모델의 특성 [4]에 따라 L2 손실 함수를 이용한다.

Inference 시에는 Student encoder 와 Teacher decoder 를 이용하여 최종 예측을 수행한다.

	KD scheme	WER
CTC w/o distillation	None	4.70
SKD [4]	Softmax-level	4.19
Proposed	Stage 1: Feature-level	4.93
	Stage 2: Softmax-level	**3.54**

표 1. 베이스라인과 제안 기법의 WER

KD scheme	Teacher Decoder Reusing	WER
Stage 2: Softmax-level	X	3.632
	O	**3.540**

표 2. Ablation Study

Ⅲ. 실험 및 결과

본 실험에서는 학습 데이터로 Librispeech train-clean-100, train-other-360, train-other-500 를 이용하였고, 테스트 데이터로 Librispeech dev-clean 을 이용하였다. Teacher 모델과 Student 모델은 Conformer 모델을 encoder 로 활용하였다. Teacher 모델은 121M, Student 모델은 13M 의 parameter 사이즈를 이용하였다. Word Error Rate (WER)를 이용하여 성능을 측정하였다.

표 1.을 통해 distillation 을 적용하지 않은 CTC 모델에 비해, 제안된 기법이 효과적으로 성능을 향상시켰음을 확인할 수 있다. 표 2.를 통해 Teacher decoder 를 재사용하지 않은 경우에 비해 Teacher decoder 를 재사용한 경우의 성능이 더 좋음을 알 수 있다. Teacher decoder 를 재사용하는 것이 Student 모델의 성능 개선에 더 효과적임을 확인하였다.

Ⅳ. 결론

본논문에서는 Teacher Decoder 를 재사용하는 2-stage 지식 증류 기법을 제안하였다. 제안된 기법은 Feature-Level Distillation 과 Softmax-Level Distillation 으로 구성된 2 단계 학습 방식을 통해, Student 모델이 Teacher 모델의 분류 지식을 효과적으로 학습할 수 있도록 설계되었다. 실험을 통해 제안 기법의 효과와 성능을 입증하였다.

ACKNOWLEDGMENT

이 논문은 2024 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의해 지원되었음

참 고 문 헌

[1] Graves, Alex, et al., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *Proceedings of the 23rd international conference on Machine learning*. 2006.

[2] Graves, A. (2012). Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711.*

[3] Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015). Listen, attend and spell. *arXiv preprint arXiv:1508.01211.*

[4] Yoon, J. W., Lee, H., Kim, H. Y., Cho, W. I., & Kim, N. S. (2021). TutorNet: Towards flexible knowledge distillation for end-to-end speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1626-1638.*