Self-Attention Map Knowledge Distillation 을 이용한 음성 인식 모델 경량화에 대한 연구

강주연, 이현승, 윤지원, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소 휴먼인터페이스 연구실
{jykang, hslee, jwyoon}@hi.snu.ac.kr, nkim@snu.ac.kr

A Study on the Lightweight Speech Recognition Model using Self-Attention Map Knowledge Distillation

Ju Yeon Kang, Hyeon Seung Lee, Ji Won Yoon, Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ.

요 약

본 논문은 Self-Attention Map Knowledge distillation 을 이용한 음성 인식 모델 경량화 기법을 제시하였다. 일반적으로 음성 인식 모델은 과도한 계산량과 메모리 사용으로 인해 real-world application 에 한계가 존재한다. 따라서 다양한 knowledge distillation 기법을 통해 모델의 경량화를 진행하고 있다. 본 논문에서는 Self-Attention Map 지식을 이용하여 Knowledge distillation 을 진행하였고 실험을 통해 경량화 효과와 성능을 확인하였다.

Ⅰ. 서 론

최근 모델의 사이즈가 매우 큰 deep neural model 들의 우수한 성능이 입증되었다. 일반적으로 large scale 모델들은 계산 복잡도가 높고 큰 메모리가 요구되기 때문에 real time application 에 불리한 특성을 가지고 있다. 따라서 model compression 기법 중 하나인 knowledge distillation 을 이용한 다양한 경량화 기법이 등장하였다.

음성 인식 모델 또한 일반적으로 모델의 사이즈가 클수록 성능이 우수하다고 알려져 있다. 따라서 음성 인식 모델을 on-device 로 사용하기 위해 다양한 model compression 기법이 존재한다. 본 논문에서는 self-attention map 을 이용한 knowledge distillation 기법을 적용하여 음성 인식 모델의 경량화를 진행하였다. 실험을 통해 self-attention map knowledge distillation 의 경량화 효과와 성능의 우수성을 확인하였다.

Ⅱ. 본론

Knowledge distillation 이란 model compression 기법 중 하나로 모델의 사이즈가 크고 성능이 우수한 teacher 모델에서 모델의 사이즈가 작고 상대적으로 성능이 낮은 student 모델로 지식을 전달해주어 student 모델의 성능을 높이는 것을 말한다. Knowledge distillation 은 사용되는 knowledge 에 따라 두가지 기법으로 구분할 수 있다. 첫번째로, Softmax-level knowledge distillation 은 teacher 의 마지막 layer 의 output 인 logit 을 student 에게 전달한다. 일반적으로 teacher 와 student 의 logit, 즉 softmax 간에 cross entropy loss 를 적용한다. 두번째로, feature level knowledge distillation 은 모델의 중간 layer 들의 intermediate feature 들을 이용하고 teacher 와 student 의 feature 들 간에 L2 loss 를 이용하는 것이 일반적이다.

Real world application 에서 모델의 사이즈가 크고 성능이 우수한 음성 인식 모델을 이용하는 것은 어렵다. 따라서 Knowledge distillation 을 통해 음성 인식 모델을 경량화한 다양한 기법들이 존재한다. [1] [2]

본 논문에서는 end-to-end 음성 인식 모델 중 하나인 CTC [3] 모델에 knowledge distillation 을 적용하였다. 한편, CTC 모델의 특성으로 인해 softmax level 에서 cross entropy loss 를 통해 distillation 을 진행한 기존의 방식은 오히려 student 모델의 성능을 낮춘다. CTC 모델의 posterior distribution 은 매우 sparse 하고 spiky 한 특성을 지닌다. 따라서 CTC 모델의 softmax output 은 one hot vector 와 유사한 형태를 지니고 대부분의 frame 에서 blank label 이 나온다. 또한 CTC 모델의 학습에는 speech frame 과 text 사이의 명확한 alignment 가 주어지지 않기 때문에 서로 다른 CTC 모델의 spike timing 이 다른 특성을 가지고 있다. 이로 인해 cross entropy loss 를 이용한 softmax-level knowledge distillation 을 적용하는 것은 어렵다. 또한 경험적으로 feature level distillation 을 진행해 주었을

때 collapse 되는 현상이 종종 발견되었다. 직접적으로 teacher 와 student 사이의 feature matching 을 해주게 되면 학습 과정 상에서 어느 정도 이상부터 유의미한 정보를 학습하지 못한 채 teacher 의 feature 값만 따라가게 되어 collapse 되었다.

이러한 특성들로 인해 본 논문에서는 2-stage 로 knowledge distillation 을 적용해주었다. 첫번째 단계에서는 feature-level distillation 을 진행하였고 이를 이용하여 student 모델을 초기화하였다. 이때 직접적인 feature matching 의 collapse 되는 현상을 막기 위해 feature 대신 self-attention map 을 통해 frame 들이 자체적으로 가지고 있는 관계 정보를 전달해주었다. 그림 1.과 같이 CTC 모델에 사용된 conformer encoder 의 마지막 layer 의 self-attention map 에 L2 loss 를 이용하여 feature-level distillation 을 진행하였다.

$$L_{AT} = \sum_{a=1}^{A_h} \sum_{t=1}^{|x|} ||A^T_{L,a,t} - A^S_{M,a,t}||_2^2$$

두번째 단계에서는 feature-level distillation 을 통해 초기화한 student 모델과 teacher 모델 간에 softmax-level distillation 을 진행하였다. cross entropy loss 를 이용하는 것이 어렵기 때문에 [1]의 방법을 이용하였다. 또한 기존 CTC 모델의 loss 인 CTC loss 를 추가적으로 이용하여 학습을 진행하였다. 전체적인 과정은 그림 1.과 같다.

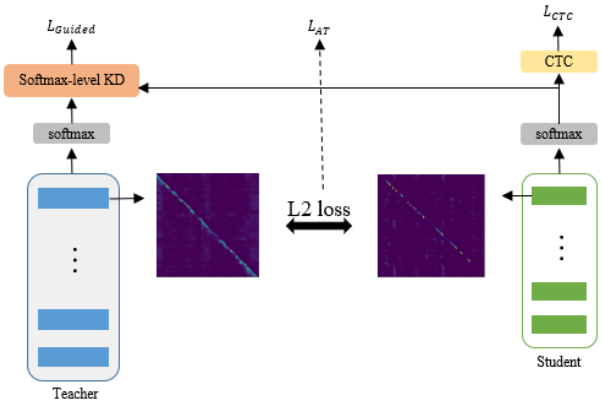

그림 1. 제시한 knowledge distillation 기법

실험은 Librispeech-960h 데이터를 이용하였다. Teacher 모델과 student 모델은 각각 121M, 5M 사이즈의 parameter 를 가지는 Conformer encoder 를 이용하였다. 베이스라인으로 distillation 을 진행하지 않은 student CTC 모델과 Guided CTC [1]을 적용한 student CTC 모델을 사용하였다. Word Error Rate 를 이용하여 성능을 측정하였고 제시한 방법의 성능은 표 1.과 같다. Distillation 을 진행하지 않은 student CTC 모델과 비교하여 [1]과 제안한 모델 모두 성능 향상을 보였다. 또한 [1]과 비교하여 feature-level distillation 과 softmax-level distillation 을 진행한 제안 기법이 성능 향상을 보였다.

KD scheme	Training Method	WER(%)
None	CTC w/o distillation	7.26
Softmax-level	Guided CTC	5.37
(1) Feature-level -> (2) Softmax-level	**Proposed**	***5.19**

표 1. 베이스라인과 제안 기법의 WER

Ⅲ. 결론

본논문에서는 self-attention map 을 이용한 knowledge distillation 기법을 제안하였고 이를 음성 인식 모델에 적용하였다. 실험 결과 기존의 방법과 비교하여 우수한 성능을 보였다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] Kurata, G., and Audhkhasi, K., “Guiding CTC posterior spike timings for improved posterior fusion and knowledge distillation,” *In Annual Conference of the International Speech Communication Association.* 2019

[2] Panchapagesan, Sankaran, et al., "Efficient knowledge distillation for rnn-transducer models," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2021.

[3] 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.