CTC Ratio Scheduling을 이용한 Joint CTC/Attention 한국어 음성인식

문영기[1,2,O], 조용래[2], 조원익[3], 조근식[1]

인하대학교[1], 보이스루[2], 서울대학교[3],
ykmoon0814@gmail.com, yongrae.jo@voithru.com, wicho@hi.snu.ac.kr, gsjo@inha.ac.kr

Joint CTC/Attention Korean ASR with CTC Ratio Scheduling

YoungKi Moon[1,2,O], YongRae Jo[2], WonIk Cho[3], GeunSik Jo[1],
Inha University[1], Voithru Co., Ltd.[2], Seoul National University[3]

요 약

본 논문에서는 Joint CTC/Attention 모델에 CTC ratio scheduling을 이용한 end-to-end 한국어 음성인식을 연구하였다. Joint CTC/Attention은 CTC와 attention의 장점을 결합한 모델로서 attention, CTC 단일 모델보다 좋은 성능을 보여주지만, 학습이 진행될수록 CTC가 attention의 학습을 저해하는 요인이 된다. 본 논문에서는 이러한 문제를 해결하기 위해, 학습 진행에 따라 CTC의 비율(ratio)를 줄여나가는 CTC ratio scheduling 방법을 제안한다. CTC ratio scheduling를 이용하여 학습한 결과물은 기존 Joint CTC/Attention, 단일 attention 모델 대비 좋은 성능을 보여주는 것을 확인하였다.

주제어: End-to-End Automatic speech recognition, Multitask learning, CTC, Attention

1. 서론

음성인식(Automatic Speech Recognition)은 사람의 음성을 입력으로 받아 텍스트로 변환, 출력해주는 기술을 말한다. 실제 음성인식 기술은 가전제품부터 시작해서 스마트폰, 스마트 홈 시스템 등 여러 방면에서 사용되고 있으며, 기계와 사람의 소통을 도와주는 중요한 기술 중 하나로 자리매김하고 있다.

통상적으로, 아날로그 신호인 음성은 디지털 신호로 변환된 후 기 학습된 음성인식 모델을 거쳐 텍스트가 되는데, 이 때 쓰이는 음성인식 모델로는 HMM (Hidden Markov Model) 기반 모델[1,2]과 딥러닝 기반 end-to-end 모델[3,4,5,6,7]이 주로 사용되고 있다. HMM 기반 모델은 음향 모델, 언어 모델과 같은 모듈들을 따로 구축하여, 결합하는 방식으로 만들어진다. 그렇기 때문에 각 모듈을 구축함에 있어서 각 모듈에 맞는 전문 지식이 필요하고, 이는 음성인식 모델을 만듦에 있어 큰 부담으로 작용한다.

딥러닝 기반 end-to-end 모델의 경우, 여러 개의 모듈을 결합하는 방식이 아닌, 하나의 모듈을 통해 음성인식을 학습하고 추론한다. 이러한 방법은 HMM 기반의 방법과 비교했을 때, 전문지식 없이도 만들 수 있고, 출력 결과물에 대해 사후처리를 해줄 필요도 없으며, 구현과 활용도 HMM 기반 모델에 비해 더 간단하다는 장점이 있다. 이 배경을 살펴보자면, Alexnet[8]의 Imagenet Challenge 우승 이후 딥러닝에 대한 관심도는 높아졌으며, 딥러닝의 발달은 HMM 기반 모델에도 영향을 주어서, GMM-HMM 모델을 DNN-HMM [2] 모델로 확장시켰다. 계속되는 딥러닝의 발전 속에서 자연스럽게 딥러닝 기반의 end-to-end 음성인식 모델이 제안되게 되었다.

초기 제시된 CTC (Connectionist Temporal Classification) [9] 기반의 end-to-end 모델은 HMM 기반의 음성인식 모델의 성능을 뛰어넘지는 못했다 [3]. 하지만, [6]에서 Encoder-Decoder 구조의 Seq2Seq와 attention을 결합한 모델을 제안한 이후로 HMM 기반 모델과의 성능 격차는 점점 좁혀지게 되었다. 최근에는 SpecAugment 기법을 이용한 data augmentation[10], [11]에서 제안한 Transformer 구조를 이용한 [12]와 같이 다양한 방법들이 HMM 기반 모델의 성능을 웃도는 결과를 보여주고 있다.

Seq2Seq with attention 모델의 경우, attention을 이용하여 입력값 frame과 출력값 label 사이의 alignment를 학습함으로써 좋은 성능을 보장해주었다. 하지만, 이러한 모델은 attention의 alignment 초기 학습이 잘 되지 못할 경우 큰 성능의 저하가 나타났다. 이러한 문제를 해결하기 위해, [7]에서는 attention 기반으로 end-to-end 모델을 학습할 때, CTC알고리즘을 결합하는 방식으로, 초기 alignment를 안정적으로 학습할 수 있도록 하는 방법을 제안하였다. 그러나, 이 알고리즘 또한, attention 모델의 alignment 학습이 어느 정도 되었을 경우, CTC 알고리즘이 학습을 저해한다는 문제점이 있었다.

이 논문에서는 이러한 문제를 해결하기 위해 모델이 학습되어감에 따라 CTC 결합 비율을 조정하는 CTC ratio scheduling이라는 방법을 제안한다. 우리는 CTC ratio scheduling를 통해서 모델이 기존 Joint CTC/Attention 방식 대비 더 높은 성능과 안정적인 수렴성을 보임을 확인할 수 있었다.

2. 관련 연구

일반적으로 음성인식은 입력 값인 acoustic frame의 크기가 정답값인 label의 크기보다 아주 크다. 이러한 점은 입력값과 출력값 사이의 alignment를 모델링하는 데에 문제를 유발한다. 이는 기존 기계학습뿐만 아니라 딥러닝 기반의 end-to-end 음성인식에서도 중요한 문제이다. 딥러닝 기반의 end-to-end 음성인식 방법은 alignment를 모델링하는 방법에 따라 Hard alignment, Soft alignment 의 두 가지로 나눌 수 있다.

Hard alignment 중 대표적인 방법으로는 CTC 알고리즘이 있다. CTC 알고리즘은 [9]에서 처음 제안된 방법으로서, 대표적인 CTC 기반 end-to-end 모델로는 Deepspeech2[4]가 있다. CTC 알고리즘은 학습시에 RNN과 같은 모델을 거쳐 나온 출력 값과 정답 label 사이에 가능한 모든 decoding path를 고려하여 alignment를 맞춰가는 방식이다. 하지만 CTC는 loss 계산시 각 sequence 사이에 조건부 독립을 가정하고 계산하기 때문에 좋은 성능을 보장하지 못하고, 실제로 단일 딥러닝 네트워크에 CTC를 결합하여 모델을 만들 경우 HMM 기반 모델의 성능보다 훨씬 떨어지는 성능을 보여준다 [3]. 또한, 입력 acoustic frame의 갯수보다 label의 갯수가 많은 경우에는 계산하지 못한다는 단점이 있다.

Soft alignment는 encoder-decoder 기반의 Seq2Seq 구조를 그 예시로 들 수 있다. Encoder-decoder 구조는 입력값이 encoder에서 압축되고, decoder를 통해서 다시 decoding되기 때문에 CTC처럼 조건부 독립을 가정하지 않고, differentiable하게 alignment를 학습할 수 있게 되었다. [5]는 이 구조에 attention을 적용하여, 더욱 성능을 발전시켰다. Soft alignment 방법은 가능한 모든 hard alignment를 계산하지 않고, 입력 데이터와 출력 레이블 사이의 soft alignment를 직접 계산한다. 또한 Seq2Seq 방식을 사용하기 때문에 다양한 길이의 음성 입력을 처리할 수 있다. 하지만, alignment의 초기 학습이 잘되지 못할 경우, 음성인식 성능이 떨어진다는 단점이 있었다.

Joint CTC/Attention은 attention 기반의 Seq2Seq 모델의 단점을 개선하기 위해 제안된 모델이다. 이 모델은 CTC 알고리즘에서 계산하는 Hard alignment를 Seq2Seq with attention의 soft alignment 계산에 더해서, soft alignment의 초기 학습이 잘 이루어지도록 하였다. 하지만 앞서 말했듯, alignment의 학습이 어느정도 이루어졌을 때에는 CTC의 hard alignment가 모델 성능을 저해한다는 단점이 있다. 우리는 이러한 점에 착안하여 Joint CTC/Attention 모델의 학습이 진행될 수록 CTC의 ratio를 조절하는 CTC Ratio Scheduling을 고안했다.

3. 모델

3.1 Joint CTC/Attention Model

Joint CTC/Attention Model은 Seq2Seq with attention의 loss 계산시에 Seq2Seq with attention의 encoder로부터 계산된 CTC Loss를 가중합하는 방식으로 구성되는 음성인식 모델이다. 모델의 구조는 그림 1과 같다. 모델의 입력은 $x = (x_1, x_2, \dots, x_T)$인 acoustic frame으로 정의되며, 이에 대응되는 출력값은 $y = (y_1, y_2, \dots, y_U)$로 정의된다.

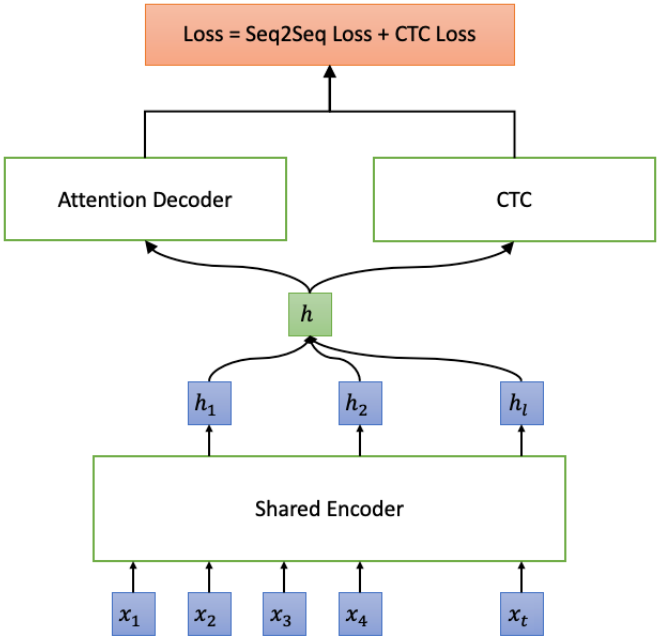

그림 1 Joint CTC/Attention model 구조, Shared Encoder는 acoustic frame($x_1, x_2, x_3, \dots, x_t$)를 입력으로 받아 hidden vector $h(h_1, h_2, \dots, h_t)$를 출력한다. 이 hidden vector는 각각 CTC와 attention Decoder의 입력값으로 들어가고, 이 출력값을 토대로 각각 Loss가 계산된다. 이렇게 계산된 Loss들의 가중합으로 Joint CTC/Attention model의 Loss가 계산된다.

3.1.1 CTC

[9]에서 제안된 CTC 알고리즘은 각 입력 acoustic frame(x)이 하나의 label(y)에 대응되는 방식이다. 이때에 output label sequence 내의 중복을 막기 위해 blank label를 사용한다. 또한, 학습시 정답 label과 blank label을 참조하여 label sequence $\phi(y')$를 만들고, 이를 토대로 학습을 진행한다. CTC의 학습 목표(Training objective/ $P(y|x)$) 는 (1)과 같이 정의된다.

$$P(y|x) = \sum_{\pi \in \phi(y')} P(\pi|x) \quad (1)$$

CTC는 encoder network의 결과물 위에 적용되며, $P(\pi|x)$는 조건부 독립 가정하에 network 결과물(q_t)의 곱으로 계산된다. ((2) 참조)

$$P(\pi|x) \approx \prod_{t=1}^{T} P(\pi_t|x) = \prod_{t=1}^{T} q_t(\pi_t) \quad (2)$$

CTC Loss는 이렇게 정의된 $P(y|x)$를 토대로 정답 label인 y^*의 확률 $P(y^*|x)$의 negative loglikelihood로 정의되며 (3), 이 loss를 최소화 시키는 방향으로 학습이 진행된다.

$$Loss_{CTC} \triangleq -\log P(y^*|x) \quad (3)$$

3.1.2 Seq2Seq with attention

Seq2Seq with attention 모델도 CTC와 같이 입력 acoustic frame(x)에 대해 label(y)를 예측하는 모델이다. 하지만 CTC와 달리 조건부 독립 가정이 없는

encoder decoder 구조를 사용하여서 더 좋은 성능을 보여준다.

$$P(y|x) = \prod_u P(y_u|x, y_{1:u-1}) \quad (4)$$

$$h = Encoder(x) \quad (5)$$

$$y_u = AttentionDecoder(h, y_{1:u-1}, c_u) \quad (6)$$

(4-6)은 Seq2Seq with attention 구조를 개괄적으로 보여준다. (6) 에서 c_u는 attention mechanism으로 생성된 context vector로 (7-9)와 같은 방법으로 계산된다. 여기서 ω, W, V, b는 모델을 통해 학습되는 모수(parameter)이다.

$$e_{u,l} = \omega^T tanh(Ws_{u-1} + Vh_l + b) \quad (7)$$

$$a_{u,l} = \frac{exp(e_{u,l})}{\sum_l exp(e_{u,l})} \quad (8)$$

$$c_u = \sum_l a_{u,l} h_l \quad (9)$$

위와 같은 과정을 거쳐 Seq2Seq loss는 (10)과 같이 정의된다.

$$Loss_{seq2seq} \triangleq -logP(y^*|x) = -\sum_u logP(y_u^*|x, y_{1:u-1}^*) \quad (10)$$

3.1.3 Joint CTC/Attention Loss

Joint CTC/Attention Model은 3.1.1과 3.1.2에서 정의했던 CTC Loss와 Seq2SeqLoss를 가중합 형태로 계산해서 최종 Loss를 계산함으로써 모델을 학습시킨다. 이 때에 α는 학습 전에 설정하는 초모수(hyperparameter)이다.

$$Loss = \alpha Loss_{CTC} + (1-\alpha) Loss_{seq2seq} \quad (11)$$

3.2 CTC ratio scheduling

Joint CTC/Attention Model 모델의 경우, Seq2Seq with attention의 초기 alignment 학습이 잘 안됨으로써 생기는 문제를 CTC를 결합하여 해결하였다. 이렇게 함으로써 CTC 단일 모델이나 Seq2Seq with attention 단일모델보다 좋은 성능을 보여주었다. 하지만 해당 모델은 CTC Loss와 Seq2Seq Loss를 가중합하는 모수인 α가 고정되었기 때문에 Seq2Seq with attention의 alignment 학습이 어느 정도 된 이후에는 CTC가 오히려 학습을 저해한다는 문제가 있었다. 이러한 문제를 해결하기 위해 이 논문에서는 CTC ratio scheduling 기법을 제안하며, 그 방법은 Algorithm 1과 같다.

Algorithm 1: CTC ratio scheduling

```
Result: ctc_ratio
joint_ctc_ratio, final_ctc_ratio,
  current_epoch,freezing_epochs,scheduling_epochs;
if current_epoch < freezing_epochs then
    ctc_ratio = joint_ctc_ratio ;
else if current_epoch < scheduling_epochs + freezing_epochs then
    gradient = (joint_ctc_ratio - final_ctc_ratio) / scheduling_epochs ;
    ctc_ratio = joint_ctc_ratio - (current_epoch - freezing_epochs) ×
      gradient ;
else
    ctc_ratio = final_ctc_ratio ;
```

우리는 제안하는 방법을 이용하여, Joint CTC/Attention Model의 한국어 음성인식에서 기존보다 더 좋은 성능을 얻을 수 있었다.

4. 실험

4.1 Dataset 및 준비

우리는 Dataset으로 KSS(Korean Single Speaker Speech Dataset)[1]와 ClovaCall[13]을 사용하였다. KSS는 전문 여성 성우가 낭독한 책 구절을 기반으로 만들어진 데이터이며, 12시간 이상의 분량으로 이루어져 있다. ClovaCall은 전화상의 식당 예약 대화를 기반으로 만들어진 데이터이며, 11,000명 이상의 사람들이 참여하였고, 100 시간 이상 분량으로 이루어져 있다. 모든 음성 파일들은 8000khz로 sampling하였다. 또한, 전체 데이터의 5%를 검증용 데이터로 추출하였고, 나머지 데이터를 학습용 데이터로 사용하였다. Label 텍스트는 전처리하지 않고 음절 단위로 토크나이징하였다.

4.2 실험 환경 및 세팅

음성 파일들은 window size 25ms, hop size 10ms로 처리하였고, delta, delta-delta 정보를 모두 사용하였으며, acoustic feature로는 log mel-spectogram을 80 dimension으로 하여 acoustic frame으로 변환하였다.

학습 및 추론은 모두 음절 단위로 진행하였다. 학습시 KSS 데이터는 GRU 기반의 Joint CTC/Attention 모델을 사용하였고, ClovaCall 데이터는 GRU 및 Transformer 기반의 Joint CTC/Attention 모델을 사용하여 학습하였다.

모든 모델들은 subsampler로 VGGNet을 사용하였고, batch accumulation은 16, dropout 0.2로 설정하였고, 40000개의 acoustic frame으로 batch를 구성하였다. GRU의 경우 hidden dimension은 256, encoder 및 decoder depth는 각각 3, 1에 double learning rate scheduler로 학습이 진행되었다. 또한, Transformer는 hidden dimension 512, encoder 및 decoder depth는 각각 6, 3으로 설정하였고, noam learning rate scheduler[11]에 warmup steps는 2000으로 하여 학습하였다. CTC ratio scheduling의 초모수는 KSS 데이터의 경우 freezing epochs 70, scheduling epochs 10으로 구성하였고,

[1] https://github.com/Kyubyong/kss

ClovaCall-RNN은 freezing epochs 100, scheduling epochs 10, ClovaCall-Transformer는 freezing epochs 130, scheduling epochs 10으로 구성하였으며, 3개의 실험 모두 ctc ratio 0.4에서 시작하여 ctc ratio 0.0까지 scheduling되도록 학습하였다.

4.3 실험 결과

학습된 모델들을 미리 추출된 데이터를 통해 검증을 하였고, 이 때에 평가 지표로는 character error rate (CER)[14]를 이용하였다. 실험 결과는 표 1과 같다.실험 결과를 보자면, KSS 데이터는 ClovaCall 데이터에 비해 CER이 높은데, 그 이유는 KSS 데이터 자체의 시간이 적어서로 보인다. 그럼에도 불구하고, scheduling를 적용한 결과물이 그렇지 않은 두 방법보다 더 좋은 결과를 보여주고 있다. ClovaCall 데이터에서는 RNN 모델과 Transformer 모델 사이에는 3~4% 정도의 CER 차이를 보이며, 앞의 KSS 데이터와 같이 scheduling를 적용한 결과물이 다른 방법들보다 좋은 결과를 보여줌을 확인할 수 있었다.

표 1. 실험 결과

CER(%)	KSS	ClovaCall-RNN	ClovaCall-Transformer
CTC ratio 0.0	13.5	6.9	2.2
CTC ratio 0.4	12.8	5.1	1.9
CTC ratio scheduling	**12.1**	**4.7**	**1.5**

5. 결론

본 논문은 기존 joint CTC/Attention 음성인식 모델에서 CTC ratio를 점차 줄여나가는 CTC ratio scheduling을 제안하였다. CTC는 학습 초반에는 attention decoder의 attention 학습을 돕는 장점을 지니고 있으나 학습 후반에는 정교한 attention의 학습을 방해한다는 가설을 세웠고, KSS 및 ClovaCall에서 CTC ratio scheduling을 적용한 모델이 기존 모델에 비해 더 낮은 CER을 달성하는 것을 확인하였다.

앞으로 이를 발전시켜 선형적인 CTC ratio scheduling 외에 triangular 또는 sinusoidal 등 비선형적인 scheduling 방법들에 대한 연구가 진행될 수 있다. 또한 scheduling 기법은 새로운 초모수를 필요로 하여 최적의 초모수 탐색을 어렵게 한다는 점을 보완하기 위해 설정해야 하는 초모수의 개수를 줄이는 scheduling 기법에 대한 연구도 가능하다.

참고문헌

[1] H. A. Bourlard and N. Morgan, Connectionist SpeechRecognition: A Hybrid Approach., USA: Kluwer Academic Publishers, 1993.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, Vol. 29, No. 6, pp. 82-97, 2012.

[3] A. Graves and N. Jaitly, "Towards end-to-end speechrecognition with recurrent neural networks," ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., Vol. 32, No. 2, pp. 1764-1772, 22-24, Jun 2014.

[4] D. Amodei, S. Ananthanarayanan, ..., J. Zhan, and Z. Zhu, "Deep speech 2 : End-to-end speech recognition in english and mandarin," ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q.Weinberger, Eds., Vol. 48, pp. 173-182, 20-22 Jun 2016.

[5] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 577-585, 2015.

[6] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," ICASSP, 2016.

[7] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," pp. 4835-4839, 03 2017.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097-1105, 2012.

[9] A. Graves, S. Fernandez , F . Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," Proceedings of the 23rd International Conference on Machine Learning, ser. ICML '06, p. 369-376, 2006.

[10] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple augmentation method for automatic speech recognition," INTERSPEECH, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L . u. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., pp. 5998-6008, 2017.

[12] L. Dong, S. Xu, and B. Xu, “Speech-transformer: A no-recurrence sequence-to-sequence model for speech recog-nition,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5884-5888, 2018.
[13] J. K. S.-W. L. S. Y. H. J. E. K. H. K. S. K. H. A. K.K. D. C. K. L. N. S. S. K. Jung-Woo Ha, Kihyun Nam,“Clovacall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers,” arXiv preprint arXiv:2004.09367, 2020.
[14] D. Jurafsky and J. H. Martin, Speech and Language Processing (2nd Edition). USA: Prentice-Hall, Inc., 2009.