RNN-T MODELS FAIL TO GENERALIZE TO OUT-OF-DOMAIN AUDIO: CAUSES AND SOLUTIONS

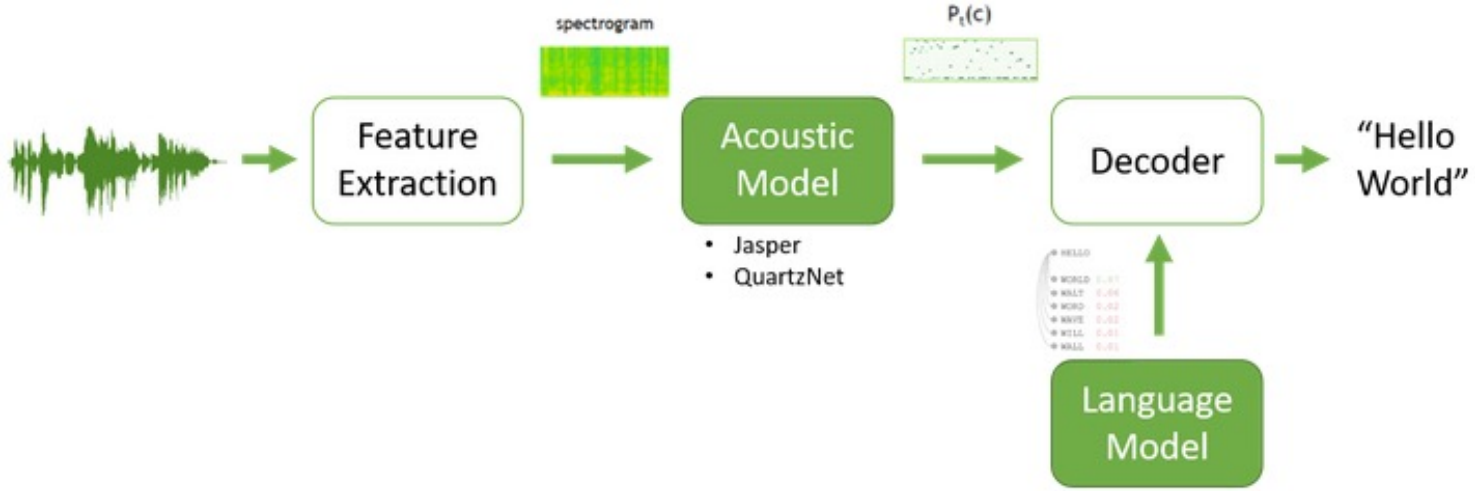
Chung-Cheng Chiu, Arun Narayanan, Wei Han, Rohit Prabhavalkar, Yu Zhang, Navdeep Jaitly , Ruoming Pang, Tara N. Sainath, Patrick Nguyen, Liangliang Cao, Yonghui Wu

Google Inc, D. E. Shaw Group

End-to-End ASR

최근의 End-to-End ASR 모델은 빠르게 발전하고 SOTA를 달성해나가고 있음

- CTC (Connectionist Temporal Classification)
- Attention based Encoder-Decoder
- Transformer



Problem of End-to-End ASR

하지만 비교적 짧고(15초 이내) 같은 도메인을 가지는 데이터 셋으로 Train되고 Test 되었던 결과

End-to-End ASR 모델은 크게 두 가지 문제가 발생

1. 도메인 불일치 상황에 매우 민감함
2. 문장 길이에 민감함

이전 연구들에서는 해당 문제에 대한 원인을 분석하지 않았으므로 이 논문에서는 실험을 통해 원인을 분석해보았음

RNN - Transducer

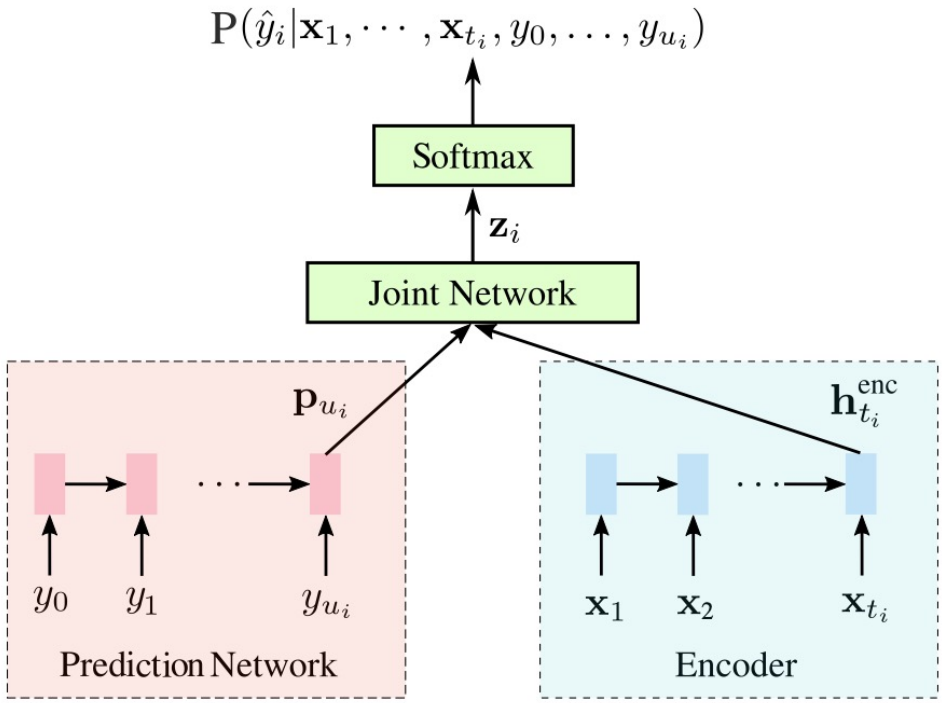


Fig. 1: Block diagram of an RNN-T model [22, 13].

CTC의 단점을 극복하기 위해 Alex Graves에 의해 제안(2012, 2013)

Encoder, Prediction, Joint Network의 세 부분으로 구성

상태가 조건부 독립이라 가정했던 CTC와는 다르게 이전 출력이 prediction network와 joint network를 통해 새로운 출력에 관여하여 종속적이게 해줌

RNN - Transducer

이 논문에서는 RNN-T를 사용하여 실험을 진행함

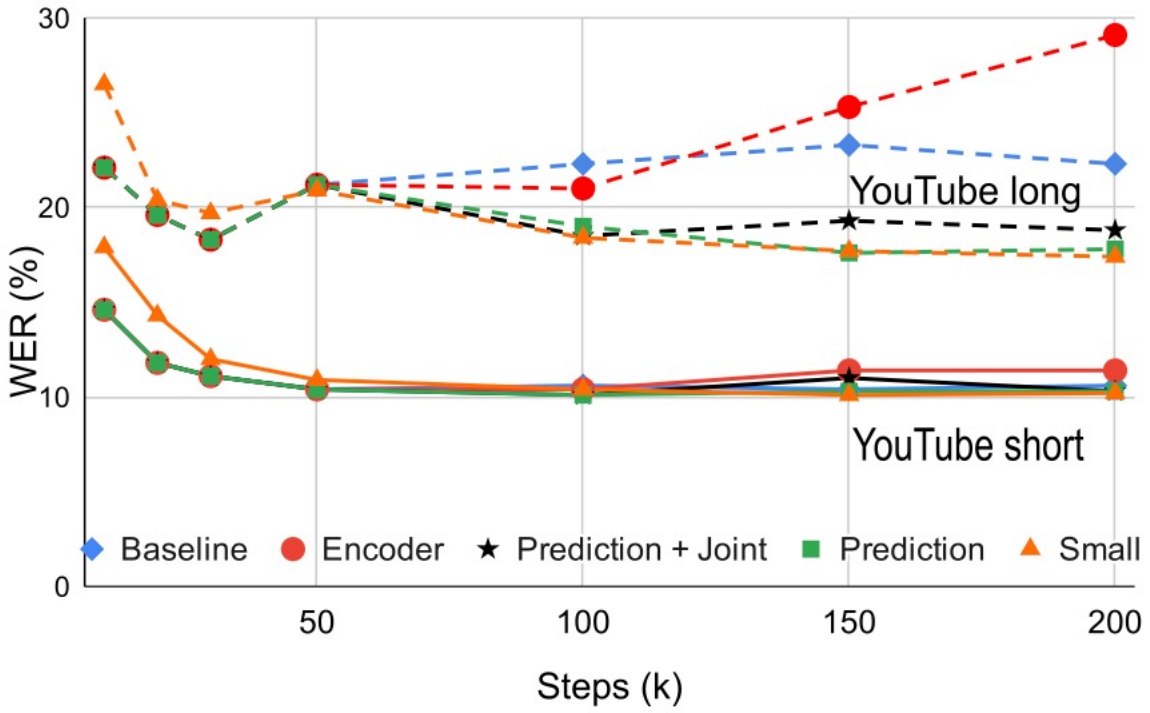
1. Youtube에서 추출한 짧은 음성으로 학습한 모델
2. 짧은 Search data(In-house data)로 학습한 모델
3. LibriSpeech Corpus로 학습한 모델

Generalization Problem을 확인하기위해 다른 데이터셋 사용하여 평가

1. Youtube-Short(2분 ~ 10분 길이)
2. Youtube-Long(41.8초 ~ 30분 길이)

Non-Streaming 방식과 Streaming 방식 모두 분석했음

Generalization Problem



테스트 진행해보니
짧은 문장으로 train 후 각
데이터셋에 대해 inference 했고
50k step 이후에 각 모듈들을
freeze 해보면서 더 학습시켰더니
낮은 성능을 보였음

Generalization Problem

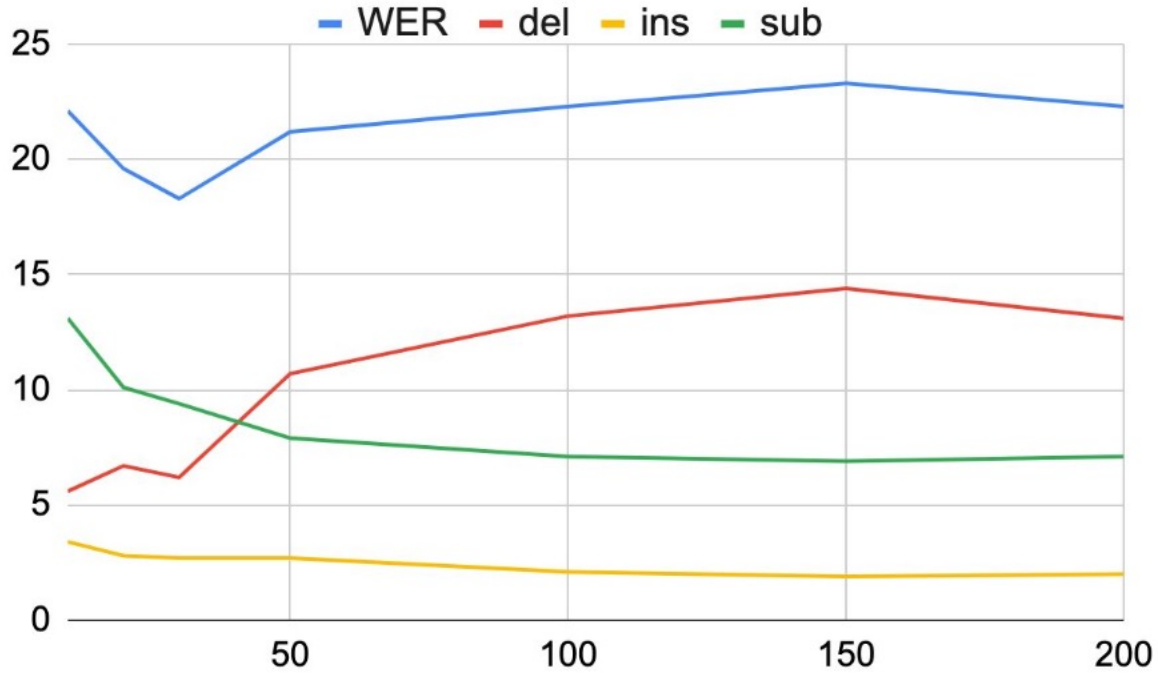
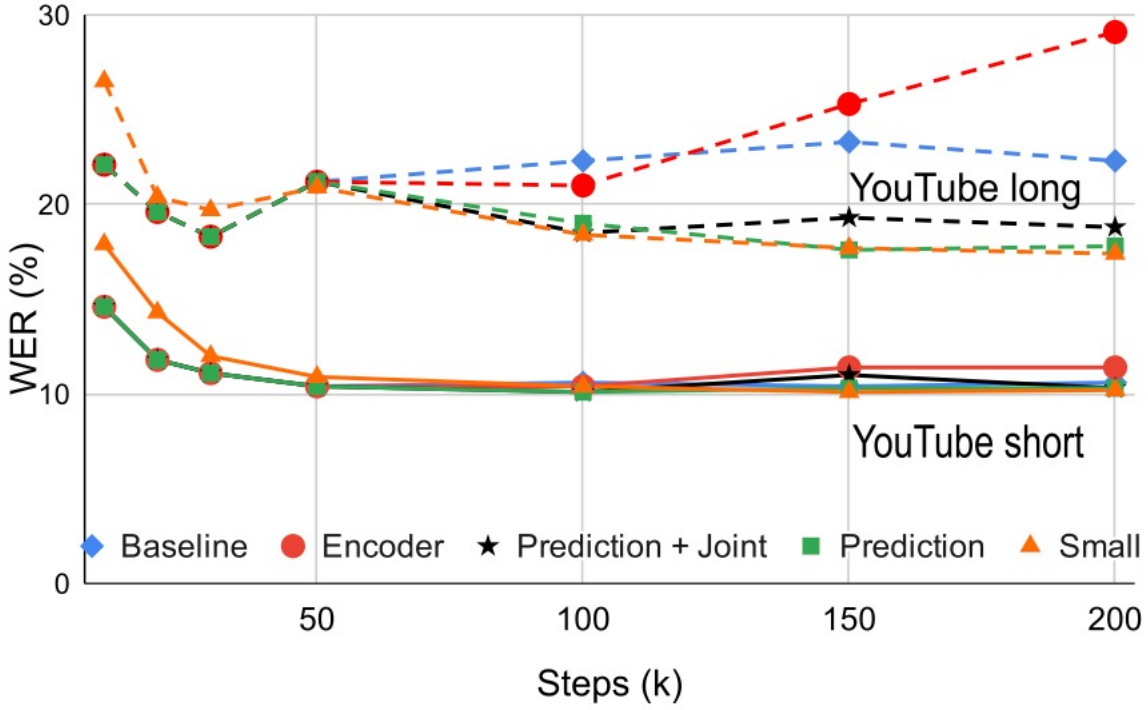


Fig. 3: WERs and the respective deletion, insertion, and substitution errors for non-streaming model on *YT-long* as a function of training steps.

높은 WER의 원인을 분석해보니
Deletion error가 주된 요인 이었음

Generalization Problem



실험은 Encoder, Prediction, Joint Network들을 각각 freeze 해보면서 성능을 살펴봤는데 주된 요인은 Encoder Network의 overfitting임을 추측할 수 있음

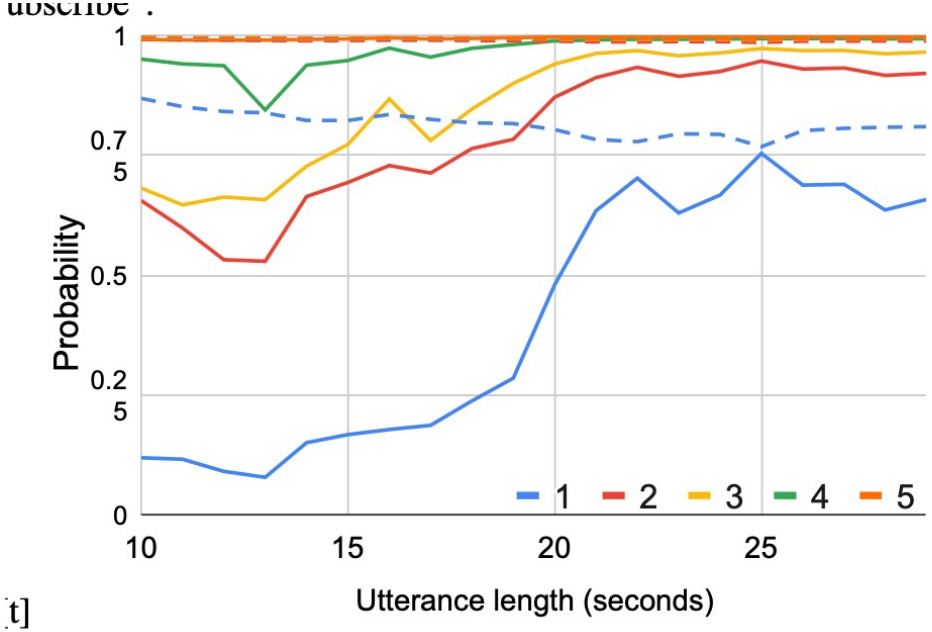
Generalization Problem

Beam search step 200	Beam search step 280	Beam search step 299
hey guys i don't know a	*<empty>*	*<empty>*
<empty>	*hey guys i don't know a lot of work yes oh you're*	*subscribe*
oh hey guys i don't know a	*hey guys i don't know a lot of work yes oh*	*yeah*
hey hey guys i don't know a	*hey guys i don't know a lot of work yes*	*h*
hey guys i don't know	*subscribe*	*this*
uh hey guys i don't know a	*hey guys i don't know a lot of work yes please*	*uh*
oh hey guys i don't know a	*hey guys i don't know a lot of work yes subscribe*	*hmm*
hey guys i don't know a a	*hey guys i don't know a lot of work yes for*	*is*

Beamsearch의 진행을 살펴봤더니<Blank> 토큰의 불안정한 예측으로 인해서 Beamsearch 가 진행될수록 더 심해져 결국 토큰이 없는 prefix가 남아 높은 deletion error를 유발했다고 함

Generalization Problem

논문 그림 오류...



요약하자면...
학습이 진행될 수록 긴 문장에서 Blank 토큰의 예측 확률이 증가했고, 이에 따라 빔서치는 pruning 시 높은 확률로 빈 시퀀스를 남기게 되어 결과적으로 deletion error가 증가하여 높은 WER을 유발하게 됨

Generalization Problem

Streaming 모델은 음성 검색 TASK에 대해 실험해보았음

Streaming을 사용해야되는 환경을 고려하여 더 작은 모델로 구성

Train은 평균 6.3초 길이의 Search data로 학습

Test는 in-domain과 out-domain으로 평가

In-domain은 평균 6초 길이,

Out-domain은 평균 62초의 TTS 음성으로 평가

Generalization Problem

Streaming 모델은 음성 검색 TASK에 대해 실험해보았음

Streaming을 사용해야되는 환경을 고려하여 더 작은 모델로 구성

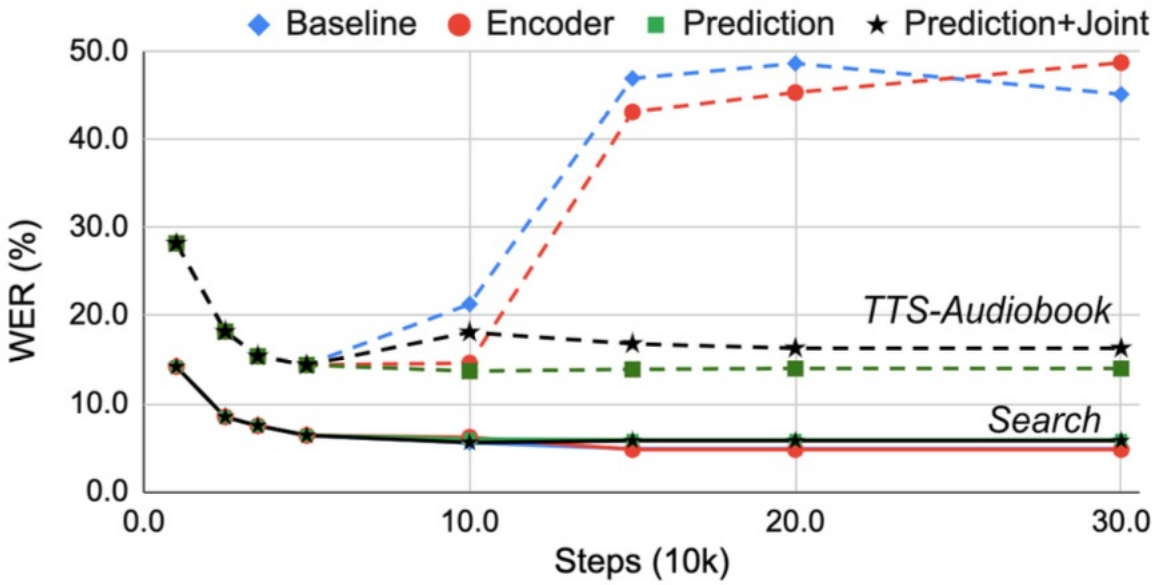
Train은 평균 6.3초 길이의 Search data로 학습

Test는 in-domain과 out-domain으로 평가

In-domain은 평균 6초 길이,

Out-domain은 평균 62초의 TTS 음성으로 평가

Generalization Problem



In-domain의 경우 학습이 진행되도 성능 하락은 없었지만,
Out-domain의 경우 Non-Streaming모델과 마찬가지로 인코더가 성능하락에 많은 기여를 하고 있음을 확인,

Regularization Cocktail

Generalization Problem은 주로 Encoder로 인해 발생하는 것을 확인 했으므로
이는 Regularization 기법을 조합해서 해결할 수 있음

논문에서는 세가지 방법의 Regularization을 사용했음

- Variational Weight Noise
- SpecAugment
- Random state sampling and random state passing

Regularization Cocktail

- Variational Weight Noise

학습이 어느정도 진행 된 후에 Gaussian Noise를 첨가

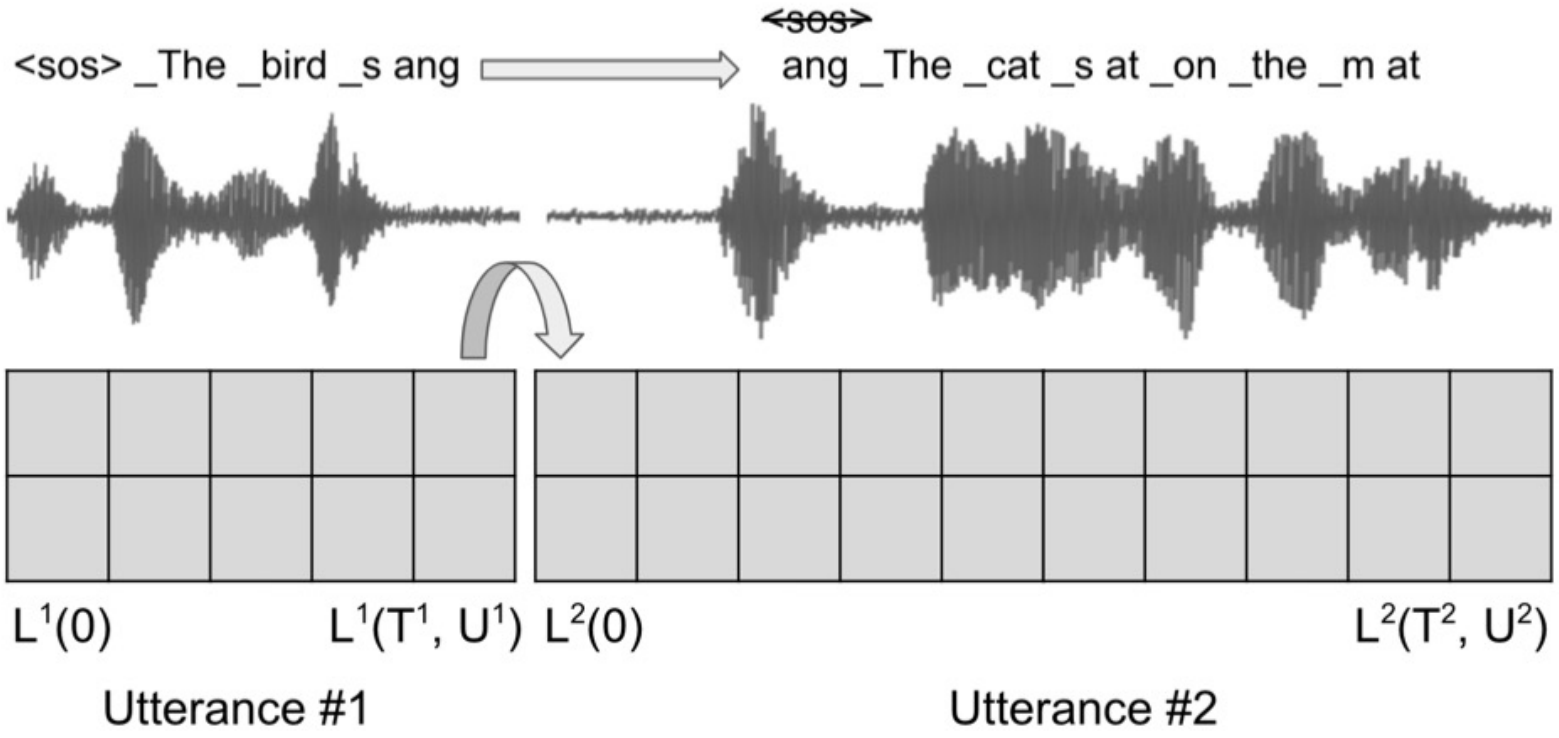
- SpecAugment

Time Warping + Masking, Frequency Masking

- Random state sampling and random state passing

RSS – LSTM의 초기 State 를 0 vector 대신 정규분포로 초기화
RSP – LSTM의 초기 State 및 Token을 이전 미니 배치의 마지막 State 와 Token으로 초기화

Regularization Cocktail



Dynamic Overlapping Inference

Inference 시 긴 문장을 짧은 문장으로 나눠서 Inference 후 결과물을 overlapping 되도록 이어 붙임

Experiment

Models	YT-short		YT-long		Call-center	
	Reg.	DOI	Reg.	DOI	Reg.	DOI
Base	10.6	9.7	22.3	17.0	27.6	22.4
SpecAugment	9.4	9.4	15.9	15.3	21.5	20.3
+ RSS	9.3	9.2	15.6	15.2	19.6	19.5
+ RSS + VN	**9.1**	**9.0**	**14.8**	**14.9**	**19.3**	**19.2**

Models	Search	TTS-Audiobook	YT-short
Baseline	4.9	48.6	67.0
VN	4.7	31.3	59.8
SpecAugment	**4.6**	16.5	52.9
+ RSP	5.1	**11.9**	27.3
+ RSP + VN	5.1	**11.9**	**25.3**

	Reg.	DOI
Test	3.2 (0.2/0.4/2.6)	3.2 (0.2/0.4/2.6)
Test Other	7.8 (0.7/0.8/6.3)	7.8 (0.6/0.9/6.3)
YT-short	99.8 (99.5/0.1/0.2)	33.0 (3.6/7.2/22.2)

Contribution

RNN-T 기반 모델에서 Generalization Problem에 대한 원인을 분석하고 성능 개선 방안을 제시

분석 결과 Encoder의 Overfitting이 Generalization Problem의 주된 요인임을 확인

여러 Regularization 기법을 조합한 Regularization Cocktail과 Dynamic Overlapping 을 통해 Generalization 성능이 크게 향상됨을 확인

Q&A