Deep Speech : Scaling up end-to-endspeech recognition

https://arxiv.org/pdf/1412.5567

0. Introduction

- 음성 인식은 기존에 특징 추출, 음향 모델, 언어 모델 등 여러 모듈이 필요해 복잡한 구조를 가졌음
- 대규모 데이터와 GPU 연산이 가능해지면서 end to end 방식의 단순한 모델이 필요해졌음
- 논문은 단일 RNN 기반 end to end 모델 Deep Speech 를 제안
- 대규모 음성 데이터, 합성 데이터, GPU 병렬화를 활용해 기존 시스템을 뛰어넘는 성능을 보였음

1. Overview

- 하나의 신경망이 스펙트로그램 입력을 받아 문자 확률을 직접 출력하는 구조
- 모델은 비순환 레이어 3개, 양방향 순환 레이어 1개, 비순환 레이어 1개로 구성됨
- 활성화 함수는 clipped ReLU
- 학습 과정은 CTC 손실을 기반으로 하며 정렬 정보 없이 음성을 문자 시퀀스로 변환함
- 다양한 환경의 음성 데이터에서 동작하도록 설계됨

2. Challenges

- 정렬 정보가 없기 때문에 시간 길이가 긴 음성에서 문자를 직접 예측하기 어려움
- 양방향 순환 레이어는 병렬화가 어려워 연산 부담이 큼
- 실사용 환경은 잡음, 발화 속도, 억양 변화가 크므로 일반화 어려움
- 대규모 모델 학습 시 과적합을 방지하기 위한 증강과 정규화 전략이 필요함

3. Method

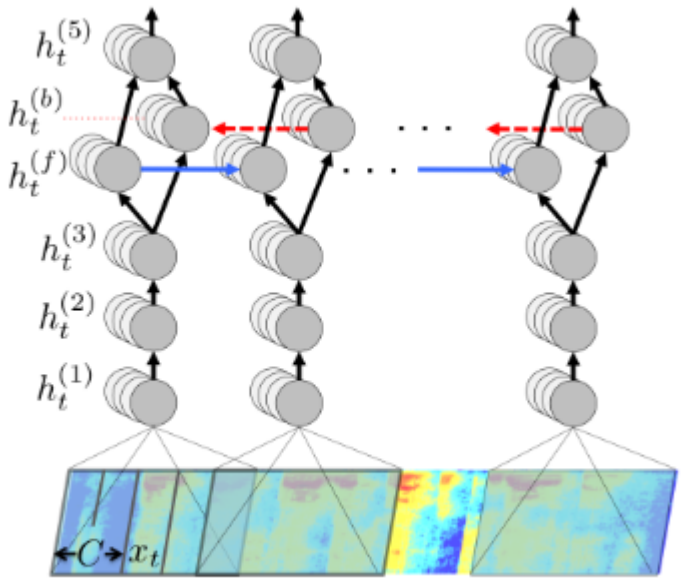

Figure 1: Structure of our RNN model and notation.

- 입력은 스펙트로그램 기반
- 문자 집합은 알파벳, 공백, 아포스트로피, blank 로 구성
- 모델은 총 5개의 레이어로 구성되며 4 번째 레이어가 양방향 순환 구조
- clipped ReLU 로 그라디언트 폭주를 방지
- CTC 손실과 Nesterov momentum 최적화 적용
- GPU 분산 학습을 위해 모델 파티셔닝과 멀티 GPU 병렬 학습 활용
- 드롭아웃을 비순환 레이어에 적용
- 시간 축 시프트 기반 테스트 증강 사용

4. Experiments

Dataset	Type	Hours	Speakers
WSJ	read	80	280
Switchboard	conversational	300	4000
Fisher	conversational	2000	23000
Baidu	read	5000	9600

- 공개 데이터와 자체 수집 데이터를 포함해 대규모 음성 데이터를 사용
- 합성 잡음, 발화 속도 변형, 다양한 환경을 반영한 데이터 증강 활용
- 평가 지표는 WER
- 모델 단일, 앙상블, 테스트 시 시프트 적용 등 다양한 설정을 실험
- 비교 대상은 기존 상용 시스템 및 연구 모델
- 데이터 전체 규모는 일부 구간에서만 정확한 수치가 제공되어 부분적으로 모호함 (확실하지 않음)

5. Results

Model	SWB	CH	Full
Vesely et al. (GMM-HMM BMMI) [44]	18.6	33.0	25.8
Vesely et al. (DNN-HMM sMBR) [44]	12.6	24.1	18.4
Maas et al. (DNN-HMM SWB) [28]	14.6	26.3	20.5
Maas et al. (DNN-HMM FSH) [28]	16.0	23.7	19.9
Seide et al. (CD-DNN) [39]	16.1	n/a	n/a
Kingsbury et al. (DNN-HMM sMBR HF) [22]	13.3	n/a	n/a
Sainath et al. (CNN-HMM) [36]	11.5	n/a	n/a
Soltau et al. (MLP/CNN+I-Vector) [40]	**10.4**	n/a	n/a
Deep Speech SWB	20.0	31.8	25.9
Deep Speech SWB + FSH	12.6	**19.3**	**16.0**

System	Clean (94)	Noisy (82)	Combined (176)
Apple Dictation	14.24	43.76	26.73
Bing Speech	11.73	36.12	22.05
Google API	6.64	30.47	16.72
wit.ai	7.94	35.06	19.41
Deep Speech	**6.56**	**19.06**	**11.85**

- Switchboard Hub5 00 에서 WER 16.0 기록
- 잡음 환경 데이터셋에서는 WER 19.1 로 기존 상용 시스템보다 월등히 우세
- 합성 데이터, 모델 앙상블, GPU 병렬 학습이 성능을 크게 높였음
- Ablation 형식의 실험을 통해 각 요소의 기여도를 확인함

6. Insight

- end to end 모델이 기존 파이프라인을 대체할 실질적 가능성을 보여줌
- 거대한 데이터와 계산 자원이 성능 향상의 핵심 요인이었음
- 다양한 환경에서 성능 확보를 위해 합성 데이터가 매우 효과적임
- 계산 비용이 높아 저자원 환경 적용에는 한계가 존재함
- 후속 연구는 모델 효율화, 실시간 처리, 합성 데이터 품질 개선 등이 필요함