Pastor Wang's Bible Teachings in Mandarin

Speech to Text System (Speech Recognition)

Stanford University

Introduction

- Aim to ease a Bible study group everyday life.
- Transcribe Bible teachings in audio form recorded by Pastor Wang in Chinese mandarin to searchable text.

Unique in several ways:

- Focus on biblical context in mandarin
- Limit scope to a single speaker, Pastor Wang

Generalization challenges:

- Overlooked accented mandarin speaking
- Missed Bible terminology
- Went astray from coherency of Bible teachings

Take on the challenges:

- Leverage end-to-end deep learning model, DeepSpeech2, to investigate Automatic Speech Recognition (ASR) system
- Use Single Character Error Rate (CER) metric
- Compare different model architecture
- Apply larger dataset for model training
- Tune hyperparameter, alpha and beta
- Accelerate model training time

Project Goal: Develop a practical ASR system which can transcribe Pastor Wang's Bible teachings in Chinese mandarin. The CER is aimed for 5% and below.

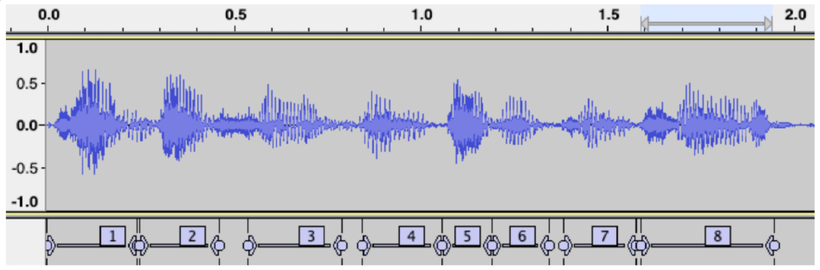
Results: We successfully structured and applied strategy of deep learning to drop CER from 50.39% to 24.97%. Although the final CER does not meet initial goal of 5%, we have identified several areas for future to further improve the CER.

Dataset and Feature

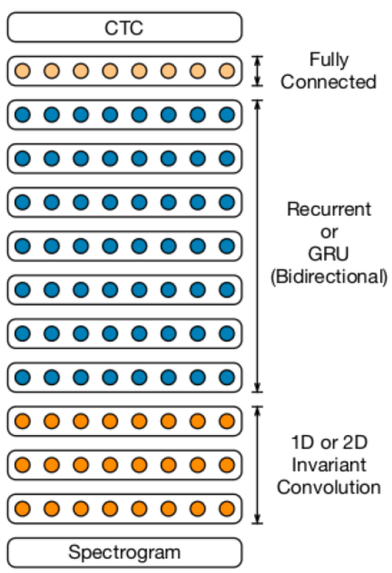
- Two dataset (aishell) and (aishell2) are used for model training.
- Pastor Wang's teachings are used as Test Set.
  - Dataset is divided according to its size.

	Aishell (aishell)	Aishell (aishell2)	Pastor Wang (Teachings)
Dataset Size (utterance)	141,925	1,0009,222	226
Dataset Size (hour)	151	1,001	0.3
Training Set (%)	84 .8	99.0	0
Training-dev Set (%)	10.1	0.5	0
Training-test Set (%)	5.1	0.5	0
Test Set (%)	0	0	100

- Illustration of an audio clip in format of waveform
  - Encoding eight Chinese mandarin characters



Method And & System Model



- DeepSpee2 may include Conv layers from 1 to 3, and GRU from 1 to 7
- We explored 2-Conv, 3-GRU layers and the maximum capacity 3-Conv, 7-GRU
- Using multiple GPUs is not accelerating training in our use case
- To avoid running out GPU memory
  - GRU size uses 1024
  - Mini-batch size is 16

Experiments Results Discussion

	DeepSpeech2 (aishell)	DeepSpeech2 (baiducn1.2k)	AWS Transcribe (Unknown)
Language Model	zhidao_giga	zhidao_giga	Unknown
Dataset Available	Yes	No	No
Dataset Size (utterance)	141,925	Unknown	Unknown
Dataset Size (hour)	151	1,204	Unknown
Wang Test Set CER (%)	50.39	54.30	3.83

- Poor initial CER on DeepSpeech2
- State-of-the-art benchmark CER on AWS

Error Analysis:

- Output length is not matched with target
- Transcribed character is not accurate

Target Transcription: 讲到神伟大的主权
Output Transcription: 将到人轨道的主持
Character Error Rate [CER]: 62.50%

CER Improvement:

- Larger dataset (aishell2) for model training
- Hyperparameter (alpha, beta) tuning
  - Dropped more than 20% CER
- Larger network helps not much

	DeepSpeech2			
Dataset	(aishell2)		(aishell)	
Dataset Size (utterance)	999,077		141,925	
Dataset Size (hour)	990		151	
Conv Layers	2	3	2	2
RNN Layers (GRU)	3	7	3	3
Training Epoch	50	50	50	50
Training Time (hour)	103.3	107.8	15.1	15 .1
Batch Size	16	16	16	16
Training Loss	0.188	0.069	0.009	0.007
Training-Dev Loss	5.284	5.011	10.900	13.112
alpha tuned best	2.6	2.2	4.2	2.6
bata tuned best	5.3	4.4	10.0	5.0
Wang Testing CER (%)	28.98	28.08	42.87	50.39

Conclusion

- An elaborate engineering task to apply deep learning
- Data preprocessing is mandatory for audio to fit model
- Large training dataset matters the most
- Tuning hyperparameter and network architecture matter
- Future work to train efficiently and further lower CER

Yen Peng (Karl) Kao (karlkao@stanford.edu)

Presentation Video Record URL (last letter is capital i): https://youtu.be/Eg4t52cphYk