Neural Speech Recognizer

Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition

Soltau, H.
Liao, H.
Sak, H.

Google

arXiv:1610.09975

Background

- Previous Study (Sak et al. 2015)
  - LSTM RNN + CTC
  - Learn an alignment between acoustic input and label sequences
  - Can recognize **whole words**
  - Vocabulary of 90k words
  - Fast and accurate, without decoding, but still far from the sub-word phone-based models
- This Paper
  - Applied the techniques on a **larger** dataset
  - Data sparsity can be alleviated

Background

Labels	Initialization			+sMBR	
	Method	Uni	Bi	Uni	Bi
CD state	CE	15.6	14.0	14.0	12.9
CI phone	CTC	15.5	14.1	14.2	12.7
CD phone	CTC	14.3	13.6	12.9	12.2

Table 2: *WERs (%) for sequence-trained LSTM RNN models.*

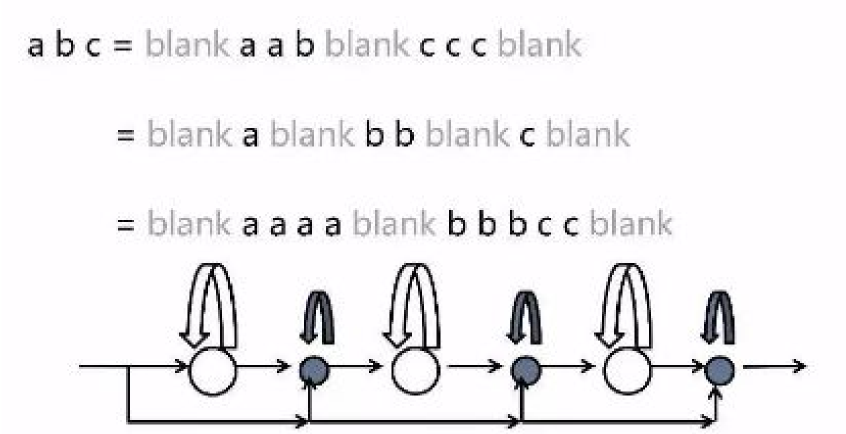
Vocabulary	OOV	WER (%)	In vocab. WER (%)
25k Word	4.8	19.5	14.5
7k Word	13	26.8	11.8

Table 3: *LSTM RNN CTC word acoustic models. The WERs and out of vocabulary (OOV) rates for word models are on held-out data with no decoding or language model. WERs in the last column are computed ignoring utterances containing OOVs.*

Sak, H., Senior, A., Rao, K., & Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv:1507.06947.

CTC

- Connectionist Temporal Classification
  - A sequence alignment/labeling technique
  - An additional unit for the **blank** label used to represent outputting no label at a given time



- Relieves the network from having to label each frame by introducing the blank label, enables the use of **longer duration modeling units**

CTC

- Loss Function of CTC

$$\mathcal{L}_{CTC} = -\sum_{(\boldsymbol{x},\boldsymbol{l})} \ln p(\boldsymbol{z}^{\boldsymbol{l}}|\boldsymbol{x}) = -\sum_{(\boldsymbol{x},\boldsymbol{l})} \mathcal{L}(\boldsymbol{x},\boldsymbol{z}^{\boldsymbol{l}})$$

  - $\boldsymbol{x}$: input sequence of acoustic frames
  - $\boldsymbol{l}$: input label sequence
  - $\boldsymbol{z}^{l}$: lattice encoding all possible alignments of $\boldsymbol{x}$ with $\boldsymbol{l}$
  - $\boldsymbol{p(z^{l}|x)}$: probability for correct labelings

CTC

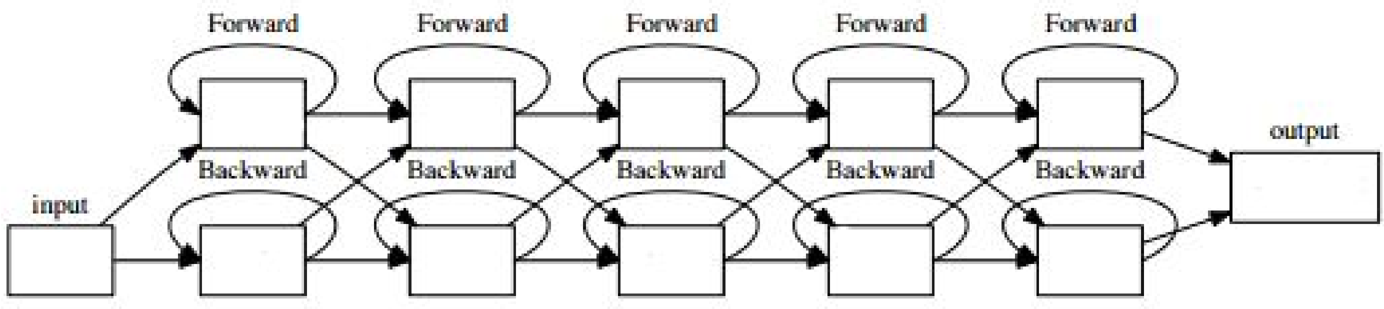
- Gradient of CTC Loss

$$\frac{\partial \mathcal{L}(x, z^l)}{\partial a_l^t} = y_l^t - \frac{1}{p(z^l|x)} \sum_{u \in \{u : z_u^l = l\}} \alpha_{x,z^l}(t,u)\beta_{x,z^l}(t,u)$$

  - $\boldsymbol{y_l^t}$: softmax activation for a label $\boldsymbol{l}$ at time step $\boldsymbol{t}$
  - $\boldsymbol{u}$: lattice states aligned with label $\boldsymbol{l}$ at time $\boldsymbol{t}$
  - $\boldsymbol{\alpha_{x,z^l}(t,u)}$: forward variable, the summed probability of all paths in the lattice z^l starting in the initial state at time 0 and ending in state $\boldsymbol{u}$ at time $\boldsymbol{t}$
  - $\boldsymbol{\beta_{x,z^l}(t,u)}$: backward variable

Basic Model

- Bidirectional LSTM RNN
  - 5x600
  - 7x1000
- Layer Connections in Bidirectional LSTM



- Input: mel-spaced log filterbank features
- Output: word posterior probabilities
- Distributed Training
  - Asynchronous SGD
  - Optimized Native TensorFlow CPU kernel

Data

- Youtube
  - Test Set
    - Videos from Google Preferred channels
    - 296 videos from 13 categories (avg. 5 min)
    - ~25 hours, 250k words
  - Training Set (semi-supervised)
    - Leverage user-uploaded captions for labels
      - select only audio segments in a video where the user uploaded caption matches the transcript produced by an ASR system
    - ~125k hours, 1.2B words, vocabulary of 1.7M
    - Spoken Vocabulary
      - >100 times, 82k words, OOV 0.63%
    - Written Vocabulary
      - > 80 times, 98k words, OOV 0.7%

State vs. Phone

- Conventional State/Phone based Models
  - CD triphone states
  - CD single-state phone units

Table 1: Bidirectional-LSTM acoustic models trained on data sets of varying sizes.

Model	Training Criterion	Size	Data (hrs)	WER(%)
CD states	CE	5x600	650	29.0
	CE	5x600	5000	21.2
CD phones	CE	5x600	5000	20.3
	CE	5x600	50000	17.7
	CE	5x600	125000	16.7
	CTC	5x600	125000	16.5
	CTC, multi_lstm_op[1]	5x600	125000	15.5
	CTC, multi_lstm_op[1]	7x1000	125000	14.2

- There is **little difference** between CE and CTC training criteria.
  - Asynchronous SGD gives better results with faster parameter updates

State vs. Phone

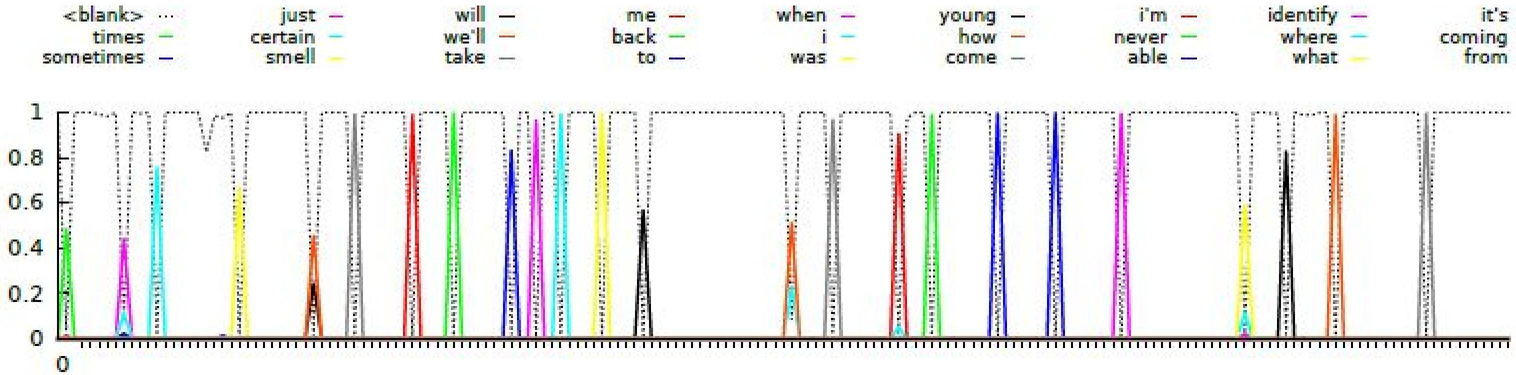


Figure 1: The word posterior probabilities as predicted by the NSR model at each time-frame (30 msec) for a segment of music video 'Stressed Out' by Twenty One Pilots. We only plot the word with highest posterior and the missing words from the correct transcription: '*Sometimes a certain smell will take me back to when I was young, how come I'm never able to identify where it's coming from*'.

Phone vs. Word

- Word Models Compared with Phone Models
  - Word model can be used **without decoding or language model** → end-to-end recognizer

Table 2: CTC CD phone models compared with CTC word models.

Model	Layers	Outputs	Params	Vocab	OOV(%)	WER(%) w/ LM	WER(%) w/o LM
CTC CD phone	5x600	6400	14m	500000	0.24	15.5	—
	7x1000	6400	43m	500000	0.24	14.2	—
	7x1000	35326	75m	500000	0.24	14.5	—
	7x1000	6400	43m	82473	0.63	14.7	—
CTC spoken words	5x600	82473	57m	82473	0.63	14.5	15.8
	7x1000	82473	116m	82473	0.63	13.5	14.8
CTC written words	7x1000	97827	137m	97827	0.70	13.4	13.9

- Capable of accurate speech recognition with no LM or decoding involved

Phone vs. Word

- Error Rate Correction for Spoken Word Model
  - References are in written domain while model output is in spoken domain
    - errors like “three” vs. “3”
  - Force align the utterances with a graph
    - C * L * project (V * T)
      - C: context transducer
      - L: lexicon transducer
      - V: spoken-to-written transducer
      - project: map the input symbols to the output symbols
    - project (V * G)
      - convert written language model G to a spoken form
      - use the spoken LM to build the decoding graph

Phone vs. Word

Table 3: Comparison of CD phone with spoken word models in spoken domain.

Model	Layers	Outputs	Params	Vocab	OOV(%)	Spoken WER(%) w/ LM	Spoken WER(%) w/o LM
CTC CD phone	7x1000	6400	43m	500000	0.24	12.3	—
CTC spoken words	7x1000	82473	116m	82473	0.63	11.6	12.0

- Word models without use of any language model or decoding performs at 12.0% WER, slightly better than the CD phone model that uses an LVCSR decoder and incorporates a 30m 5-gram language model.
- Adding LM for the CTC spoken word model improves the error rate from 12.0% to 11.6%, not too much.

Summary

- The final system performs better than a well-trained, conventional CD phone-based system on a difficult YouTube video transcription task
  - word model of bidirectional LSTM plus CTC loss having 7x1000 layers with 116 parameters and 82k vocabulary size
  - 13.4% WER for written domain with LM
  - 11.6% spoken WER for spoken domain with LM