Listen, Attend and Spell

A Neural Network for Large Vocabulary Conversational Speech Recognition

William Chan, Navdeep Jaitly, Quoc Le, Oriol Vinyals

williamchan@cmu.edu

{ndjaitly,qvl,vinyals}@google.com

Carnegie Mellon University

Google

*work done at Google Brain.

September 13, 2016

Outline

1. Introduction and Motivation
2. Model: Listen, Attend and Spell
3. Experiments and Results
4. Conclusion

Introduction and Motivation

Automatic Speech Recognition

Input

- Acoustic signal

Output

- Word transcription

State-of-the-Art ASR is Complicated

- Signal Processing
- Pronunciation Dictionary
- GMM-HMM
- Context-Dependent Phonemes
- DNN Acoustic Model
- Sequence Training
- Language Model

- Many proxy problems, (mostly) independently optimized
  - Disconnect between proxy problems (i.e., frame accuracy) and ASR performance
  - Sequence Training solves some of the problems

HMM Assumptions

- Conditional independence between frames/symbols
- Markovian
- Phonemes

- We make untrue assumptions to simply our problem
  - Almost everything fallback to the HMM (and phonemes)

Goal: Model Characters directly from Acoustics

Input

- Acoustic signal (e.g., filterbank spectra)

Output

- English characters

- Don't make assumptions about the our distributions

End-to-End Model

- Signal Processing
- Listen, Attend and Spell (LAS)
- Language Model?

- One model optimized end-to-end
- learn pronunciation, acoustic, dicationary all in one end-to-end model

Model: Listen, Attend and Spell

Sequence-to-Sequence (and Attention)

Machine Translation:

- Sutskever et al., "Sequence to Sequence Learning with Neural Networks," in NIPS 2014.
- Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in EMNLP 2014.
- Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate," in ICLR 2015.

TIMIT:

- Chorowski et al., "Attention-Based Models for Speech Recognition," in NIPS 2015.

Listen, Attend and Spell

Let $\mathbf{x}$ be our acoustic features, and let $\mathbf{y}$ be the sequence we are trying to model (i.e., character sequence):

$$\mathbf{h} = \text{Listen}(\mathbf{x}) \tag{1}$$

$$P(y_i|\mathbf{x}, y_{<i}) = \text{AttendAndSpell}(y_{<i}, \mathbf{h}) \tag{2}$$

Implicit Language Model

- HMM/CTC have conditional independence assumption
- seq2seq models have a conditional dependence on the previously emitted symbols:

$$P(y_i|\mathbf{x}, y_{<i}) \tag{3}$$

Listen, Attend and Spell

- $\text{Listen}(\mathbf{x})$ can be a RNN (i.e., LSTM).
  - Transform our input features $\mathbf{x}$ into some higher level feature $\mathbf{h}$
- AttendAndSpell is an attention-based RNN decoder.

Listen, Attend and Spell

AttendAndSpell is an attention-based RNN decoder:

$$c_i = \text{AttentionContext}(s_i, \mathbf{h}) \tag{4}$$

$$s_i = \text{RNN}(s_{i-1}, y_{i-1}, c_{i-1}) \tag{5}$$

$$P(y_i|\mathbf{x}, y_{<i}) = \text{CharacterDistribution}(s_i, c_i) \tag{6}$$

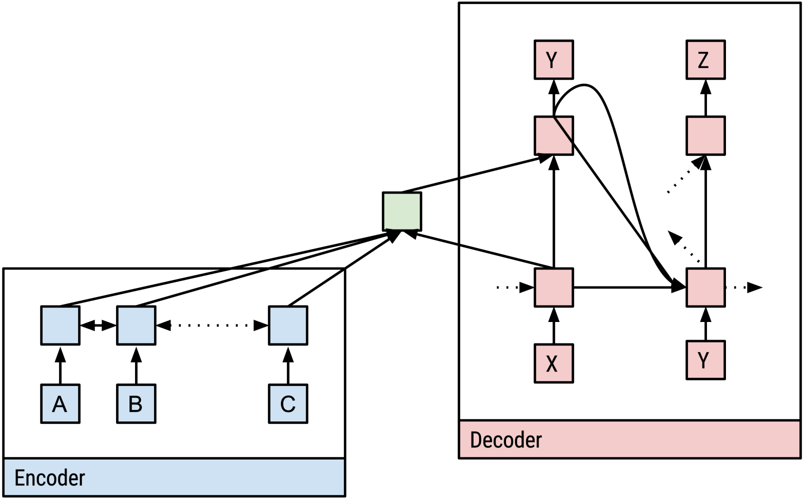
The AttentionContext creates an alignment and context for each timestep:

$$e_{i,u} = \langle \phi(s_i), \psi(h_u) \rangle \tag{7}$$

$$\alpha_{i,u} = \frac{\exp(e_{i,u})}{\sum_{u'} \exp(e_{i,u'})} \tag{8}$$

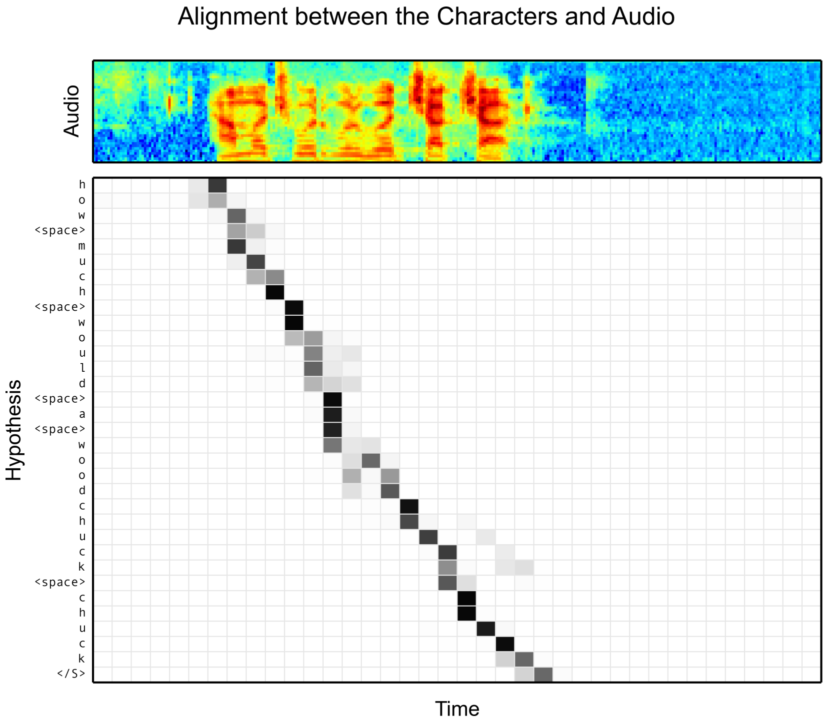
$$c_i = \sum_u \alpha_{i,u} h_u \tag{9}$$

Listen, Attend and Spell



Listen, Attend and Spell

- Attention mechanism creates a short circuit between each decoder's output and the acoustic
- More efficient information/gradient flow!
- Creates an explicit alignment between each character and the acoustic features
  - CTC's alignment is latent

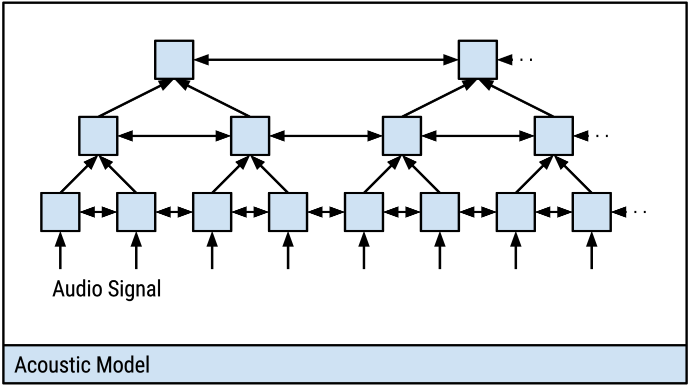


Alignment between the Characters and Audio
Audio
Hypothesis
h
o
w
<space>
m
u
c
h
<space>
w
o
u
l
d
<space>
a
<space>
w
o
o
d
c
h
u
c
k
<space>
c
h
u
c
k
</S>
Time

Listen, Attend and Spell

- Model works but...
  - Takes "forever" to train, after $>$ 1 month model still not converged : (
  - WERs in the $>$20s (CLDNN-HMM is 8ish)
  - Attention mechanism must focus on a long range of frames

Pyramid



Pyramid

- Build higher level features with each layer
- Reduce number of timesteps for attention to attend to
- Computational efficiency
- 8 filterbank frames $\rightarrow$ 1 pyramid frame feature

$$h_i^j = \mathrm{pBLSTM}(h_{i-1}^j, \left[h_{2i}^{j-1}, h_{2i+1}^{j-1}\right]) \tag{10}$$

Pyramid

1. Sequence-to-Sequence
2. Attention
3. Pyramid

- 16 to 20-ish WERs (w/o LM)
- Takes around 2-3 wks to train, overfitting a HUGE problem
- Mismatch between train and inference conditions

Sampling Trick

Machine Translation, Image Captioning and TIMIT:

- Bengio et al., "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks," in NIPS 2015.

Sampling Trick

- Training is conditioned on ground truth
- We don't have access to ground truth during inference!

$$\max_{\theta} \sum_{i} \log P(y_i|\mathbf{x}, y^*_{>i}; \theta) \quad (11)$$

Sampling Trick

- Sample from our model
- Condition on sample for next step prediction

$$\tilde{y}_i \sim \mathrm{CharacterDistribution}(s_i, c_i) \tag{12}$$

$$\max_{\theta} \sum_i \log P(y_i | \mathbf{x}, \tilde{y}_{>i}; \theta) \tag{13}$$

Listen, Attend and Spell

$$P(\mathbf{y}|\mathbf{x}) = \prod_i P(y_i|\mathbf{x}, y_{<i}) \tag{14}$$

$$\mathbf{h} = \text{Listen}(\mathbf{x}) \tag{15}$$

$$P(\mathbf{y}|\mathbf{x}) = \text{AttendAndSpell}(\mathbf{h}, \mathbf{y}) \tag{16}$$

Language Model Rescoring

- Leverage on vast quantities of text!
- Normalize our LAS model by number of characters in utterance – LAS has bias for short utterances.

$$s(\mathbf{y}, \mathbf{x}) = \frac{\log P(\mathbf{y}|\mathbf{x})}{|\mathbf{y}|_c} + \lambda \log P_{\mathrm{LM}}(\mathbf{y}) \tag{17}$$

Experiments and Results

Dataset

- Google voice search
- 2000 hrs, 3M training utterances
- 16 hrs, 22K test utterances
- Mixed Room Simulator, artificially increase acoustic data by x20 (i.e., YouTube and environmental noise)
- Clean and noisy test set

Training

- Stochastic Gradient Descent
- DistBelief 32 replicas, minibatch size of 32
- 2-3 weeks of training time

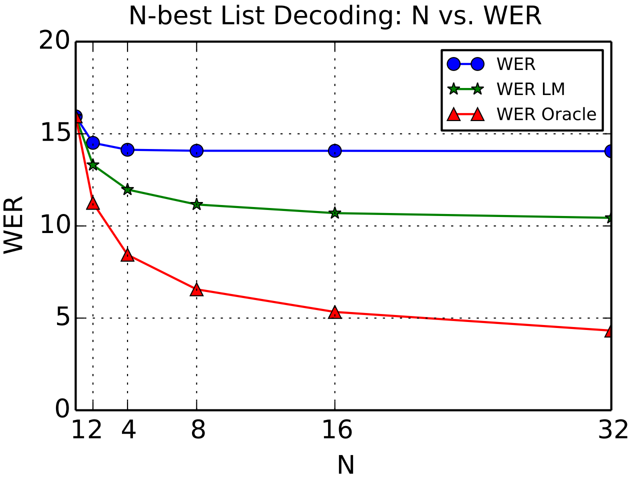
Results

Model	**Clean WER**	**Noisy WER**
CLDNN-HMM (Tara et al., 2015)	8.0	8.9
LAS	16.2	19.0
LAS + LM	12.6	14.7
LAS + Sampling	14.1	16.5
LAS + Sampling + LM	10.3	12.0

Decoding

- We didn't decode with a dictionary!
  - LAS implicitly learnt the dictionary during training
  - Rare spelling mistakes!
- We didn't decode with a LM! (only rescored)
  - n-best list decoding where $n = 32$
- CLDNN-HMM is convolutional and unidirectional, LAS is not convolutional and bidirectional

Decoding

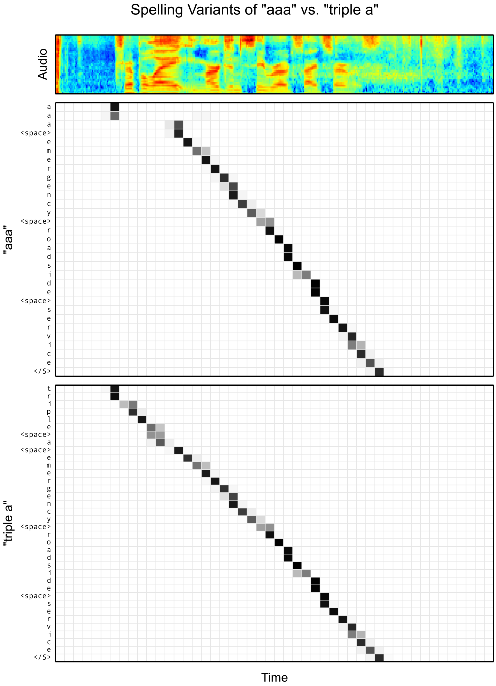

N-best List Decoding: N vs. WER
WER
WER LM
WER Oracle
20
15
10
5
0
1
2
4
8
16
32
N
WER

Decoding

- 16% WER without any searching (and LM) - just take the greedy path!
- LAS does "reasonably" well even with $n = 4$ in n-best list decoding
- Not much to gain after $n > 16$
- LM rescoring recovers less than $1/2$ of the oracle – need to improve LM?

Results: Triple A

N	**Text**	$\log P$	**WER**
Truth	call aaa roadside assistance	-	-
1	call aaa roadside assistance	-0.57	0.0
2	call triple a roadside assistance	-1.54	50.0
3	call trip way roadside assistance	-3.50	50.0
4	call xxx roadside assistance	-4.44	25.0


Spelling Variants of "aaa" vs. "triple a"
Audio
"aaa"
a a a <space> e m e r g e n c y <space> r o a d s i d e <space> s e r v i c e </S>
"triple a"
t r i p l e <space> a <space> e m e r g e n c y <space> r o a d s i d e <space> s e r v i c e </S>
Time

Conclusion

Conclusion

- Listen, Attend and Spell (LAS)
  - End-to-end speech recognition model
  - No conditional independence, Markovian assumptions, or proxy problems
  - Sequence-to-Sequence + Attention + Pyramid
  - One model: integrate all traditional components of an ASR system into one model (acoustic, pronunciation, language, etc...)
- Competitive to state-of-the-art CLDNN-HMM system
  - 10.3 vs. 8.0 WER
- Time to throw away HMM and phonemes!
- Independently proposed by Bahandau et al., 2016 on WSJ (next next talk, checkout their paper too!)

Acknowledgements

Nothing is possible without help from many friends...

- Google Speech team: Tara Sainath and Babak Damavandi
- Google Brain team: Andrew Dai, Ashish Agarwal, Samy Bengio, Eugene Brevdo, Greg Corrado, Andrew Dai, Jeff Dean, Rajat Monga, Christopher Olah, Mike Schuster, Noam Shazeer, Ilya Sutskever, Vincent Vanhoucke and more!