Deep Speech: Scaling Up End-to-end Speech Recognition

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng

Baidu Research – Silicon Valley AI Lab
https://arxiv.org/abs/1412.5567

Presenter: Ching-Hsiang Chu
Outline

● Introduction
  ● The core of Deep Speech
  ● Optimizations on Multi-GPU
  ● Experiments
  ● Summary
  ● Live demo video
  ● Highlights of DeepSpeech 2
Introduction

- Speech Recognition Systems
  - Recognition and translation of spoken language into text by computers
  - Fundamental and essential component in our digital life
    - Interactive Voice Response, Healthcare, Driving, Smart Home...and more

Source: http://www.kathrynsadler.com/vui-design/
Source: https://www.techhive.com/article/3243581/speakers/amazon-echo-vs-google.html
What, When and Why we use voice search?

**WHEN do we use voice search?**

<table>
<thead>
<tr>
<th></th>
<th>TEENS</th>
<th></th>
<th>ADULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>57%</td>
<td>24%</td>
<td></td>
</tr>
<tr>
<td>with friends</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22% in the bathroom</td>
<td>15%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8% while cooking</td>
<td>23%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17% while exercising</td>
<td>15%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59% while watching tv</td>
<td>36%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**WHY do we use voice search?**

- It’s cool: 79% (Teens) / 63% (Adults)
- It’s the future: 89% (Teens) / 85% (Adults)
- It’s safer: 78% (Teens) / 76% (Adults)
- For multitasking: 78% (Teens) / 63% (Adults)
- Makes me more efficient: 73% (Teens) / 62% (Adults)

**What you WISH voice search could do**

- Send me pizza: 45% (Teens) / 36% (Adults)
- Find the remote: 34% (Teens) / 33% (Adults)
- Find my keys: 34% (Teens) / 44% (Adults)

Traditional Speech Recognition

- Multiple stages
  - Acoustic, Phoneme, and Language models
Why Deep Learning for Speech Recognition?

● Traditional approaches of Automatic Speech Recognition (ASR)
  ○ Specialized input features
  ○ Acoustic models
  ○ Hidden Markov Models (HMM)

● Issues
  ○ Low signal-noise ratio (SNR)
  ○ Speaker variability
    ■ E.g., Accent: https://www.youtube.com/watch?v=Sj3P8WYv61M
  ○ Natural/conversational speech

Source: http://www.kathrynsadler.com/vui-design/
Deep Learning in Speech Recognition

- Significant improvement in Word Error Rate (WER)
  - E.g., A WER of 5% roughly corresponds to 1 missed word for every 20
- Beating traditional approaches and even human?

Source:
https://www.slideshare.net/LuMa921/deep-learning-the-past-present-and-future-of-artificial-intelligence

Improvements in word error rate over time on the Switchboard conversational speech recognition benchmark. The test set was collected in 2000. It consists of 40 phone conversations between two random native English speakers.
Source: https://awni.github.io/speech-recognition/
Overview of Deep Speech

● Challenges
  ○ Overcomes noisy environments
  ○ Builds large & labeled training sets
  ○ Trains networks that are large enough to effectively utilize all of this data

● Contributions
  ○ Propose a RNN model to train speech data
  ○ Use Multi-GPU for training RNN
  ○ Synthesis noisy speech datasets
Outline

● Introduction

● The core of Deep Speech
  ○ Recurrent Neural Network

● Optimizations on Multi-GPU

● Experiments

● Summary

● Live demo video

● Highlights of DeepSpeech 2
Background - Recurrent Neural Network (RNN)

- Make use of sequential information
- Perform the same task for every element of a sequence, with the output being depended on the previous computations
  - A “memory” which captures information about what has been calculated so far

Deep Speech: RNN Training Setup

- Recurrent Neural Network (RNN)
  a. Ingest speech spectrograms
  b. Generate English text transcriptions
  c. Convert an input sequence $\mathbf{x}$ into a sequence of character probabilities for the transcription $\mathbf{y}$

  - $\mathbf{x}^{(i)}$: a single utterance, a time-series of length $T^{(i)}$
    - Every time-slice is a vector of audio features
    
    $\mathbf{X} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots\}$

    $\hat{y}_t = \mathbb{P}(c_t | \mathbf{x})$, where $c_t \in \{a, b, c, \ldots, z, \text{space}, \text{apostrophe}, \text{blank}\}$

https://en.wikipedia.org/wiki/Spectrogram
RNN model - 5 layers of hidden units

\[ g(W^{(5)}_{h^{(4)}_t} + b^{(5)}) \]

where \( h^{(4)}_t = h^{(f)}_t + h^{(b)}_t \)

Bi-directional recurrent layer

\[ h^{(f)}_t = g(W^{(4)}h^{(3)}_t + W^{(f)}_r h^{(f)}_{t-1} + b^{(4)}) \]
\[ h^{(b)}_t = g(W^{(4)}h^{(3)}_t + W^{(b)}_r h^{(b)}_{t+1} + b^{(4)}) \]

Non-recurrent convolutional layers with ReLu activation function

\[ h^{(l)}_t = g(W^{(l)}h^{(l-1)}_t + b^{(l)}) \]

where \( g(z) = \min\{\max\{0, z\}, 20\} \)

Rectified-linear (ReLu) activation function
Another Look of the RNN model

- Output alphabet, space, and blank
Apply Connectionist Temporal Classification (CTC)*

- No Alignment needed
- Compute CTC loss to evaluate the gradient \( \nabla_y \mathcal{L}(\hat{y}, y) \)

Regularization - To Reduce Variance

- **Dropout** between 5-10%
  - In the feedforward layers, not the recurrent hidden activations

- Randomly jitter inputs
  - Translate the raw audio files by 5ms
    - To the left and right
  - Forward propagate the recomputed features
  - Average the output probabilities
  - At test time
    - Use an ensemble of several RNNs, averaging their outputs in the same way
**Language Model - Correct errors**

- Phonetically plausible renderings of English words

<table>
<thead>
<tr>
<th>RNN output</th>
<th>Decoded Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>what is the weather like in bostin right now</td>
<td>what is the weather like in boston right now</td>
</tr>
<tr>
<td>prime miniter nerenr modi</td>
<td>prime minister narendra modi</td>
</tr>
<tr>
<td>arther n tickets for the game</td>
<td>are there any tickets for the game</td>
</tr>
</tbody>
</table>

- Why we have these errors?
  - Words rarely or never appear in the training sets

- Solution: N-gram language model
  - Easily trained from huge unlabeled text corpora
  - 220 million phrases, supporting 495,000 words

\[
Q(c) = \log(P(c|x)) + \alpha \log(P_{lm}(c)) + \beta \text{word} \_ \text{count}(c)
\]
Outline

● Introduction
● The core of Deep Speech
● Optimizations on Multi-GPU
  ○ Hybrid Parallelism
● Experiments
● Summary
● Live demo video
● Highlights of DeepSpeech 2
Optimizations - Fast Training

- Homogeneous rectified linear networks
  - Simple to implement on a few highly-optimized BLAS calls
- Striding
  - Shorten the recurrent layers by taking “steps” of size 2 in the original input
  - Use of cuDNN to implement the first layer of convolution efficiently
- Efficient computation is critical
  - 5 billion connections for a typical utterance (Fully unrolled)
  - Solutions
    - Multi-GPU training: data and model parallelism
Hybrid Parallelism

● **Data parallelism**
  ○ Each GPU processes many examples in parallel
    ■ Concatenating many examples into a single matrix
  ○ Each GPU processing a separate minibatch of examples
    ■ Combining its computed gradient with its peers during each iteration
    ■ Up to the limit of GPU memory
  ○ Challenge: utterances have different lengths
    ■ Sorting training examples by the length
    ■ Combining only similarly-sized utterances into minibatches
    ■ Padding with silence when necessary

● **Model parallelism**
  ○ Challenge: the sequential nature of the recurrent layers
    ■ Divide the model in half along the time dimension
Hybrid Parallelism (Cont’)

Exchange intermediate activations and swap roles

Source: https://www.slideshare.net/baiduusa/deep-speech-recent-progress-on-mandarin-speech-recognition
Outline

- Introduction
- The core of Deep Speech
- Optimizations on Multi-GPU
- **Experiments**
  - Constructing Training Data
  - Evaluation
- **Summary**
- Live demo video
- Highlights of DeepSpeech 2
Training Data

- Few public datasets of sufficient scale

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Hours</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ</td>
<td>read</td>
<td>80</td>
<td>280</td>
</tr>
<tr>
<td>Switchboard</td>
<td>conversational</td>
<td>300</td>
<td>4000</td>
</tr>
<tr>
<td>Fisher</td>
<td>conversational</td>
<td>2000</td>
<td>23000</td>
</tr>
<tr>
<td>Baidu</td>
<td>read</td>
<td>5000</td>
<td>9600</td>
</tr>
</tbody>
</table>

- Synthesis by superposition
  - Creating noisy environments
  - To avoid using a single noise source

- Lombard Effect
  - Speakers actively change the pitch or inflections of their voice to overcome noise around them

\[
\hat{x}^{(i)} = x^{(i)} + \xi^{(i)}
\]

\[
\hat{x}^{(i)} = x^{(i)} + \xi_1^{(i)} + \xi_2^{(i)} + \ldots
\]
Training Data (Cont’)

- Demonstration of Synthesized noisy data

https://youtu.be/P9GLDezYVX4?t=3444 (57m24s)
Experiments

● Training & Inferencing
  ○ Use the proposed RNN model trained from a selection of the datasets
  ○ Predict **character-level** transcriptions
  ○ Predicted probability vectors and language model are then fed into decoder to yield a **word-level** transcription

● Word Error Rate (WER)
  ○ Compared with the ground truth transcription

● Testing sets
  ○ Conversational speech: Switchboard Hub5’00
  ○ Synthesized noisy speech
Experiments - Conversational speech

● Training
  ○ 300-hour Switchboard, and 2000-hour Fisher conversational telephone speech dataset
  ○ A full pass over 2300 hours of data in a few hours

● Decoding: 4-gram language model with 30,000 word vocabulary (trained)

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

5 hidden layers each with 2048 neurons on only 300 hour SWB data
4 RNNs each with 5 hidden layers of 2304 neurons on only 2300 hour combined corpus
Experiments - Noisy Speech

● Constructed evaluation set
  ○ 100 noisy and 100 noisy-free utterances from 10 speakers
  ○ Utterance text
    ■ Web search queries and text messages, news clippings, phone conversations...
  ○ Noise environments
    ■ background radio or TV; washing dishes; a crowded cafeteria;...
  ○ Low signal-to-noise ratio (SNR)

● Decoding: 5-gram language model
  ○ Trained on 220 million phrases of Common Crawl, Kept 495,000 words

<table>
<thead>
<tr>
<th>System</th>
<th>Clean (94)</th>
<th>Noisy (82)</th>
<th>Combined (176)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple Dictation</td>
<td>14.24</td>
<td>43.76</td>
<td>26.73</td>
</tr>
<tr>
<td>Bing Speech</td>
<td>11.73</td>
<td>36.12</td>
<td>22.05</td>
</tr>
<tr>
<td>Google API</td>
<td>6.64</td>
<td>30.47</td>
<td>16.72</td>
</tr>
<tr>
<td>wit.ai</td>
<td>7.94</td>
<td>35.06</td>
<td>19.41</td>
</tr>
<tr>
<td><strong>Deep Speech</strong></td>
<td><strong>6.56</strong></td>
<td><strong>19.06</strong></td>
<td><strong>11.85</strong></td>
</tr>
</tbody>
</table>
Summary

● RNN model for speech recognition
  ○ Conversational speech
  ○ Speech in noisy environments

● Multi-GPU training
  ○ Data and model parallelism

● Build large training sets
  ○ Synthesized noisy speech data
Outline

- Introduction
- The core of Deep Speech
- Optimizations on Multi-GPU
- Experiments
- Summary
- Live demo video
- Highlights of DeepSpeech 2
Demo of Deep Speech

Live Demo @ GTC 2015: https://youtu.be/qP9TOX8T-kl?t=2520 (42m05s)
Deep Speech 2

- Extended model to recognize English and Mandarin
  - Required essentially no expert knowledge
- Enhanced training performance
  - Faster communication
    - Customized Baidu All-Reduce
  - Fast implementation of CTC for GPUs
  - Custom memory allocators
    Weeks => days

Deep Speech 2 - Accents and Noise

Comparison of human transcribers to Baidu’s Deep Speech 2 model on various types of speech. Notice the humans are worse at transcribing the non-American accents. This is probably due to an American bias in the transcriber pool. I would expect transcribers native to a given region to have much lower error rates for that region’s accents.

Source: https://awni.github.io/speech-recognition/
# Deep Speech 2 - Languages

- **English**
  - Up to 68 million parameters

<table>
<thead>
<tr>
<th>Test set</th>
<th>Ours</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ eval’92</td>
<td>3.10</td>
<td>5.03</td>
</tr>
<tr>
<td>WSJ eval’93</td>
<td>4.42</td>
<td>8.08</td>
</tr>
<tr>
<td>LibriSpeech test-clean</td>
<td>5.15</td>
<td>5.83</td>
</tr>
<tr>
<td>LibriSpeech test-other</td>
<td>12.73</td>
<td>12.69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test set</th>
<th>Ours</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>VoxForge American-Canadian</td>
<td>7.94</td>
<td>4.85</td>
</tr>
<tr>
<td>VoxForge Commonwealth</td>
<td>14.85</td>
<td>8.15</td>
</tr>
<tr>
<td>VoxForge European</td>
<td>18.44</td>
<td>12.76</td>
</tr>
<tr>
<td>VoxForge Indian</td>
<td>22.89</td>
<td>22.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test set</th>
<th>Ours</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHiME eval real</td>
<td>21.59</td>
<td>11.84</td>
</tr>
<tr>
<td>CHiME eval sim</td>
<td>42.55</td>
<td>31.33</td>
</tr>
</tbody>
</table>

- **Mandarin**
  - ~80 million parameters

<table>
<thead>
<tr>
<th>Test set</th>
<th>Human</th>
<th>RNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 utterances / committee</td>
<td>4.0</td>
<td>3.7</td>
</tr>
<tr>
<td>250 utterances / individual</td>
<td>9.7</td>
<td>5.7</td>
</tr>
</tbody>
</table>

Table 6: We benchmark the best Mandarin system against humans on two randomly selected test sets. The first set has 100 examples and is labelled by a committee of 5 Chinese speakers. The second has 250 examples and is labelled by a single human transcriber.
• Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng, “Deep Speech: Scaling up end-to-end speech recognition”
  ○ https://arxiv.org/abs/1412.5567
• Stanford Seminar - Deep Speech: Scaling up end-to-end speech recognition
  ○ https://www.youtube.com/watch?v=P9GLDezYVX4
• https://web.stanford.edu/class/cs224s/lectures/224s.17.lec8.pdf
  ○ Slides #32-45
• Live demo#GTC2015
  ○ https://youtu.be/qP9TOX8T-kI?t=2561
• A TensorFlow implementation of Baidu's DeepSpeech architecture
  ○ https://github.com/mozilla/DeepSpeech
• Deep Speech 3: Even more end-to-end speech recognition
  ○ http://research.baidu.com/Blog/index-view?id=90
Thank you

Questions?