Joseph E. Gonzalez
Co-director of the RISE Lab
jegonzal@cs.berkeley.edu

Natural Language Processing (NLP) Systems

Applications?

- Speech recognition
  - Predict text from sound
- Machine translation
  - Predict text from text
- Text to speech
  - Predict sound from text
- Sentiment analysis
  - Predict sentiment from text
- Automated question answering
  - Predict text from (text, images, datasets)
- Information retrieval
  - Ranking/clustering (text, images, and data) according a query

What makes language challenging?

- **Representation** (inputs)
  - Images are easy to represent as tensors
  - Word sequences are more complex
- Often **structured prediction** (outputs)
  - Example: predicting complete sentences
- Requires **semantic knowledge** and **logical reasoning**
  - "One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know."
- Substantial **variation across languages**
  - Each language historically required different methods

Two NLP Systems Papers

Speech Recognition

Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

Baidu Research – Silicon Valley AI Lab*
Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu

Machine Translation

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com

Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

- Describe **production solutions**
- Present a **full systems approach** spanning training to inference
- Present the **interaction** between **systems** and **model design**

Context

2015

Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

Baidu Research – Silicon Valley AI Lab*
Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu

2014

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

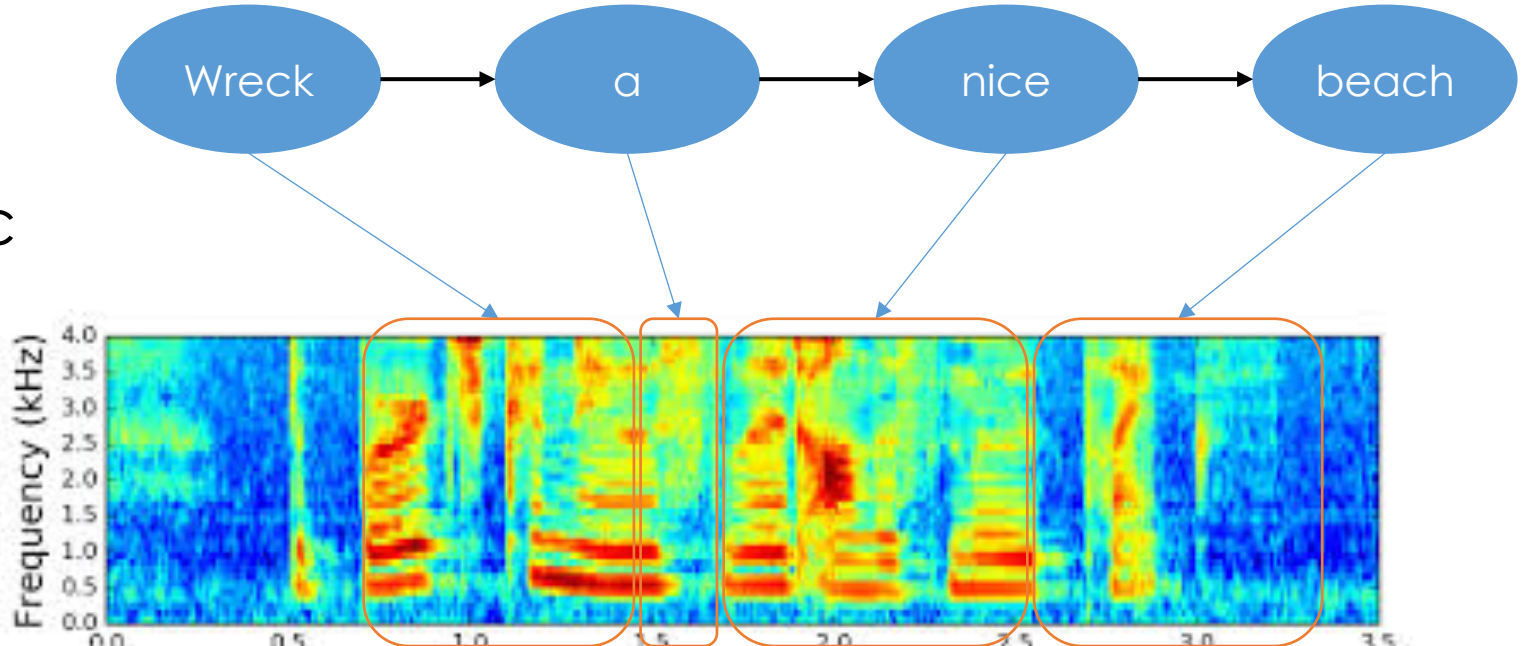
2014

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

- Demonstrated the viability of an end-to-end deep learning approach to speech recognition
- Previously: HMMs with (deep) acoustic models.



Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

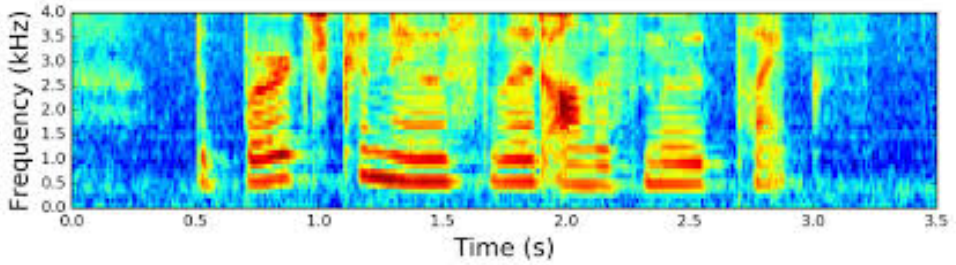
Baidu Research – Silicon Valley AI Lab*
Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu

- DeepSpeech2:
  - Improves the model architecture (8x Larger)
  - Elaborates on the training and deployment process
  - Provides a more complete description and analysis of the system

Big Ideas

- Demonstrates **super human performance** from an **end-to-end deep learning** approach
  - Input spectrograms, output letters (graphemes, bi-graphemes)
- Presents a **single model** and system for both **English** and **Mandarin** Speech Recognition
  - Previous work required separate approaches
- Well presented/evaluated ideas:
  - Data collection and preparation
  - System training and inference performance
  - Ablation studies for many of the design decisions

Model Innovations



- Predict letters (pairs of letters) directly from spectrograms
  - Emerging trend at this point
- Deeper Models enabled by:
  - Heavy use **of batch normalization**
- SortaGrad curriculum learning
  - Train on shorter (easier) inputs first → improves/stabilizes training
- 2D convolution on spectrogram inputs
  - Striding to reduce runtime

- Simple 5-Gram language model

RNN output	Decoded Transcription
what is the weather like in bostin right now	what is the weather like in boston right now
prime miniter nerenr modi	prime minister narendra modi
arther n tickets for the game	are there any tickets for the game

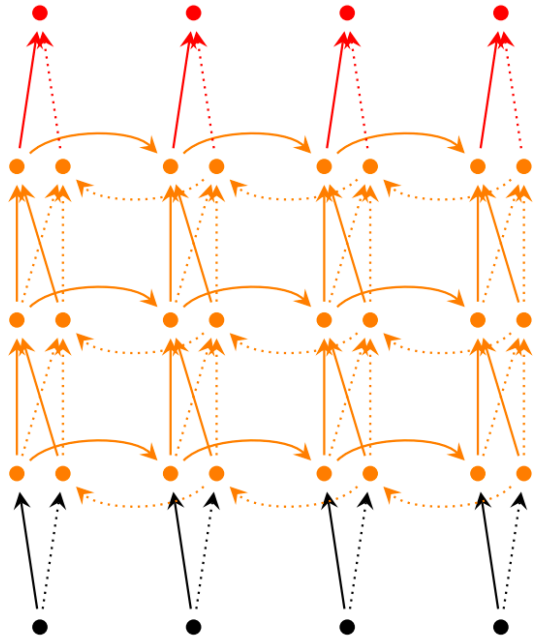
- Prediction requires beam search to maximize

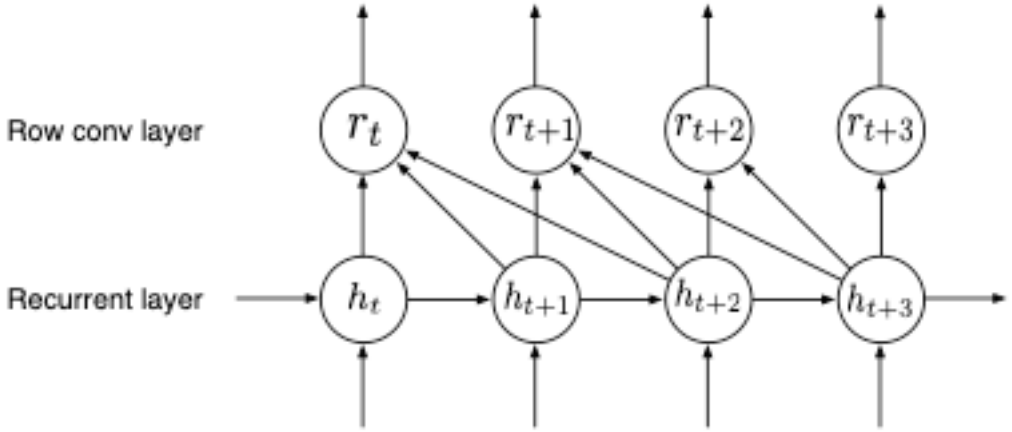
$$Q(y) = \log(p_{\text{ctc}}(y|x)) + \alpha \log(p_{\text{lm}}(y)) + \beta \, \text{word\_count}(y)$$

Model and System Co-Design

- Minimize bi-directional layers to enable faster inference
  - Introduce row convolutions

Bidirectional RNN





Placed at top of network

Variations in approach for Mandarin

Key Observation: *Minimal changes required*

- Did not need to explicitly model language specific pronunciation
- Output requires extension to 6000 characters + Roman alphabet
- Shorter beam searches

Data Preparation

- Used models to align transcription with audio
- Used models to identify bad transcriptions
- Data augmentation
  - Adding synthetic noise
- Constructed human baseline

Systems Innovations in Training

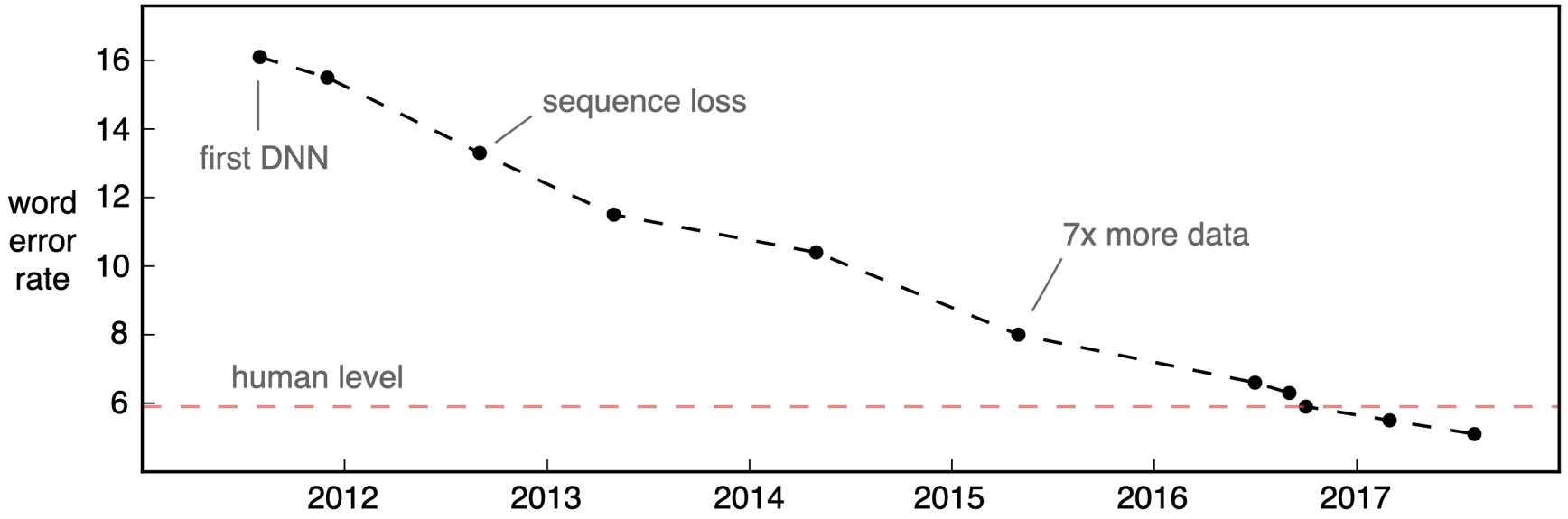
- Synchronous Distributed SGD for better reproducibility
- Custom GPU implementation of CTC Loss
  - Combinatorial evaluation
- Optimized all-reduce with GPUDirect communication
  - Substantial performance gains
- Custom memory allocators and fallback for long utterances
- 45% of peak theoretical performance per node.

Innovations during Inference

- Batch Dispatch:
  - Combines queries when possible to improve throughput
- Low bit precision (16bit) floating point arithmetic
- Optimized Small Batch GPU GEMM Kernels
- Reduced beam search to most likely characters
  - Speeds up Mandarin by 150x
- P98 latency of 67 milliseconds with 10 streams per server
  - Is this good?

Metrics of Success

- Word error rate
  - Similar to edit distance measures frequency of substitutions, deletions, and insertions



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.

https://awni.github.io/speech-recognition/

Big Results

- Outperforms humans on several benchmarks

Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69

Accented Speech			
Test set	DS1	DS2	Human
VoxForge American-Canadian	15.01	7.55	4.85
VoxForge Commonwealth	28.46	13.56	8.15
VoxForge European	31.20	17.55	12.76
VoxForge Indian	45.35	22.44	22.15

Noisy Speech			
Test set	DS1	DS2	Human
CHiME eval clean	6.30	3.34	3.46
CHiME eval real	67.94	21.79	11.84
CHiME eval sim	80.27	45.05	31.33

Limitations and Future Impact

Future Impact

- **Provided substantial evidence of end-to-end approach**
- May be in use at Baidu?
- Batching techniques appear in other inference systems

Limitations

- Not clear how to personalize?
- Substantial computational costs at prediction time
- Difficult to reproduce results (no code or data released)