Achieving Nothing

Sam Davis

supervised by David Page, Myrtle
Carl Henrik Ek, University of Bristol

Overview

Increasing the efficiency of speech-to-text neural networks by sparsifying the large parameter matrices.

Motivation

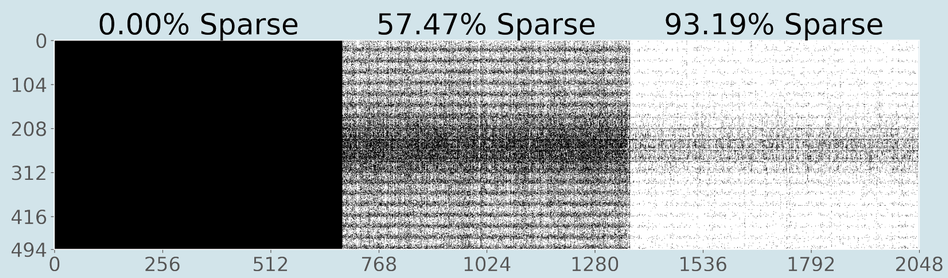
- Tens of millions of voice controlled devices have been sold. [1]
- Neural networks convert the speech data to text.
- Efficiency improvements will save companies millions of dollars.

Sparsity

- A neural network consists of the chained application of many nonlinear functions:

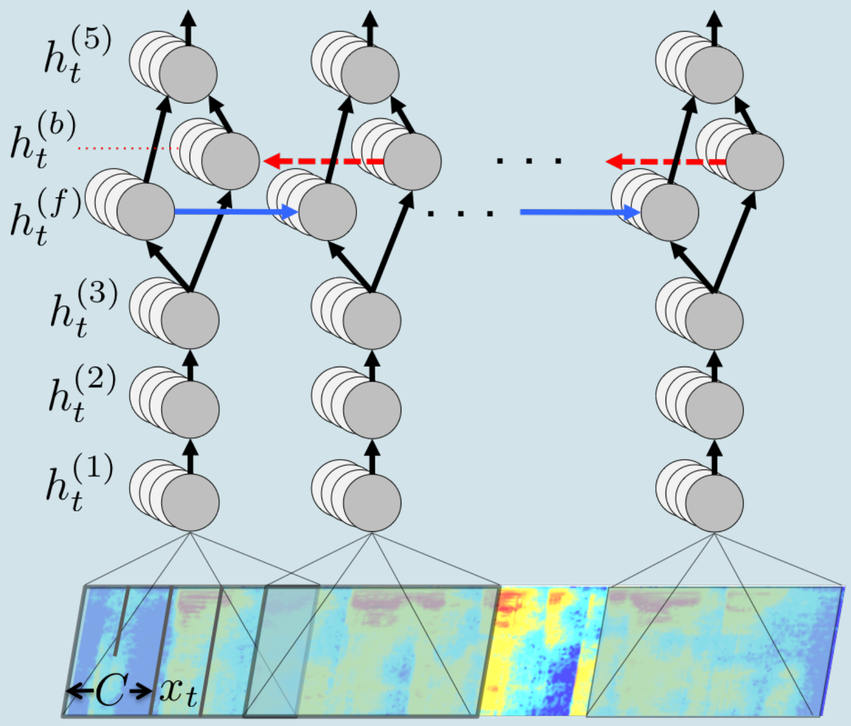
$$h^i = g(W^i h^{i-1} + b^i)$$

- The total FLOPs is dominated by large matrix-vector products.
- Inducing sparsity – fixing parameters to zero – in the matrices causes many operations to become NOPs.
- The FLOP requirement is reduced given hardware support.



W^1 Non-Zero Parameter Heatmap

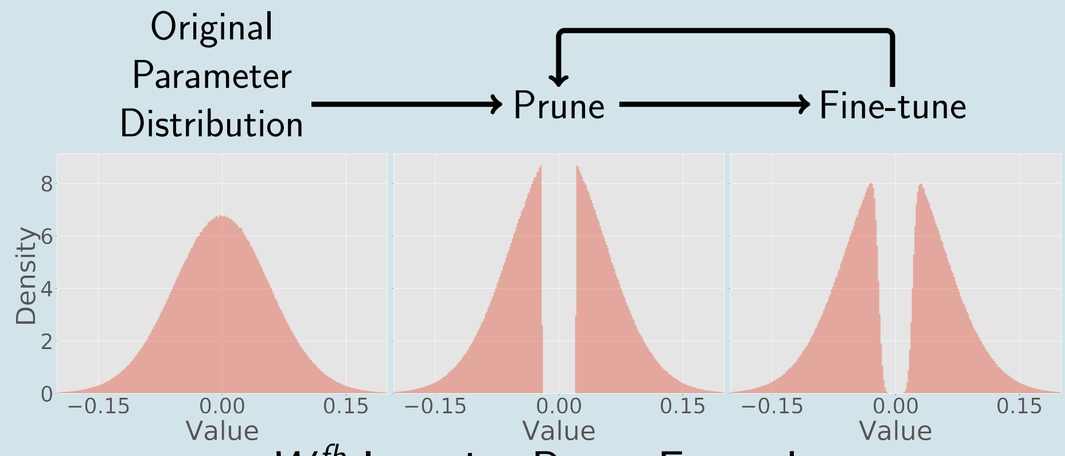
Model



Deep Speech Architecture [2]

Task	Speech-to-text
Network Input	Mel Frequency Cepstral Coefficients (MFCC)
Network Output	Character Sequence
Network Loss	Connectionist Temporal Classification (CTC)
Accuracy Metric	Word Error Rate (WER)
BLSTM Parameters	100 M
Total Parameters	122 M
Sequence Length	500 steps/10 s audio
Training Set	960 hours
Validation Set	5.4 hours
Test Set	5.4 hours
Training Setup	4 * NVIDIA Tesla V100
Hours/Training Epoch	2

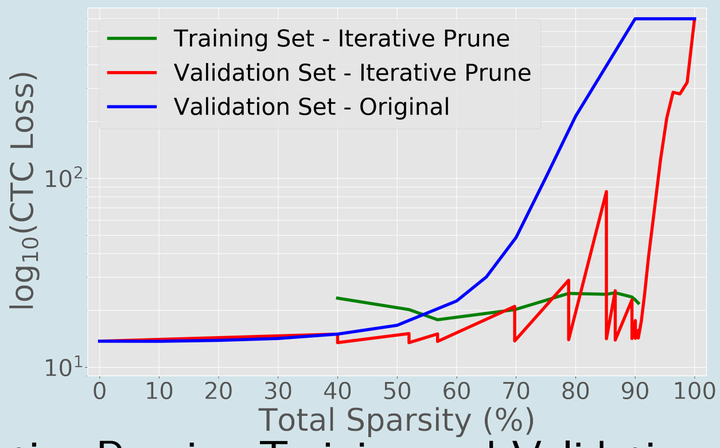
Baseline Approach



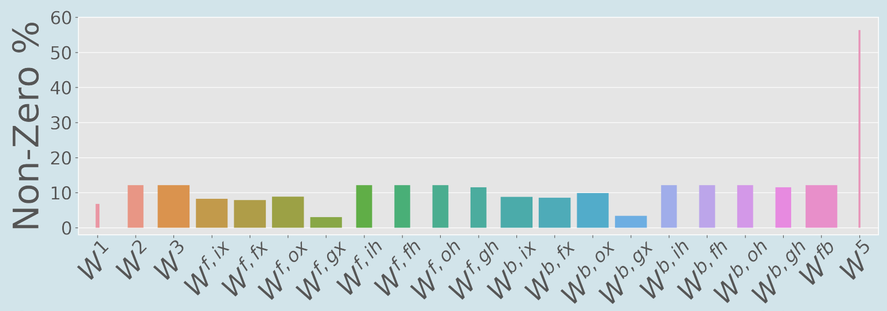
W^{fb} Iterative Prune Example

Baseline Results

Model	Total Sparsity (%)	Word Error Rate (WER)
Human [3]	-	5.83
Deep Speech	0.00	5.12
Deep Speech	90.54	5.39



Iterative Pruning Training and Validation Loss



90.54% Total Sparsity, Non-Zero Parameter % Per-Layer

Future Work

- Reach sparsity limit with baseline method.
- Investigate sparsity patterns within the network.
- Experiment with alternative methods including those that induce sparsity during the initial training process.

Acknowledgements

Myrtle
www.myrtle.ai
@myrtleai

With thanks to Myrtle for supporting the project by providing expert supervision, an infrastructure for experiments, and an FPGA platform for deployment.

Citations

[1] "Amazon.com announces fourth quarter sales up 38% to $60.5 billion," *Amazon Press Room - Press Release*, Feb 2018.

[2] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, pp. 173–182, 2016.