Sashank Gondala

+1-703-712-2706
sgondala2@gmail.com

ABOUT ME

I'm a Research Engineer at Apple, where I build Language Models for Speech Recognition. I spend most of my time on various aspects of language modeling such as data selection, modeling architecture, scaling the training, and evaluating approaches to improve Word Error Rate on rare words. Before this, I was a Master's student at Georgia Tech at Prof. Devi Parikh's lab where I worked on problems in Language Grounding such as Image Captioning, Vision and Language Navigation. I also love systems engineering. Before my Masters, I was a senior C++ engineer working on areas such as profiling, caching, and optimizing low-level mechanisms such as pointer retrieval to improve wall time.

ACADEMIC BACKGROUND

Georgia Tech, Atlanta, USA — Aug 2019 - May 2021
Master of Science, Computer Science

- Worked with Prof. Devi Parikh on Language Grounding tasks
- Teaching Assistant for Deep Learning (2x) and Machine Learning
- Co-organized the VQA workshop at CVPR 2021

IIT Bombay, Mumbai, India — Aug 2012 - May 2016
Bachelor of Technology, Computer Science and Engineering

- IIT-JEE 2012 **All India Rank 14** out of 500k candidates
- All India Rank 8 in Nationwide Education and Scholarship Test 2014.

PUBLICATIONS

Error-driven Pruning of Language Models for Virtual Assistants — **ICASSP 2021**
Sashank Gondala*, Lyan Verwimp*, Ernest Pusateri, Manos Tsagkias, C Van Gysel

Explored ways to maximize accuracy of an English speech recognition model while keeping the size of it's language model small. Obtained a 10% reduction in language model size with negligible increase in Word Error Rate (3x better than random)

EMPLOYMENT HISTORY

Language Modeling Scientist — May 2021 - Present
Apple, Cupertino, CA

- Working as a Language Modeling Scientist in Siri Speech Recognition team
- Replaced existing word level RNN with subword based transformer architecture which resulted in a ∼7% improvement in WER for the same model size on disk
- Experimented with various subword tokenization and data selection techniques and obtained a ∼1.5% improvement in WER on the rare word recognition with no increase in model's size or latency
- Obtained a 10% reduction in NGram language model size with negligible increase in Word Error Rate by implementing a custom NGram pruning algorithm. This was published at ICASSP 2021.
- Regularly prototype and ship ideas from the latest literature to both improve the accuracy and reduce the memory footprint
- Have experience scaling training to ∼hundred GPUs and handling hundreds of millions of rows of data
- Have expertise working on scale across classical models (NGram), autoregressive models (LSTMs), and attention-based models (Transformers)

Cofounder Feb 2018 - Dec 2018
Decentralized CDN Startup

- **Co-founded a startup** to provide decentralized CDN services by sharing the spare bandwidth and hard drive space of Internet users over blockchain.
- IP developed include a prototype to support HLS video streams and a whitepaper that describes the challenges and solutions based on the SOTA techniques including Service Certificates, Probabilistic Micropayments, etc.

Senior Member Technical Staff June 2016 - June 2019
Oracle, Redwood City, CA

- Worked on performance optimizations for Oracle BI, a data analytics product
- **Improved sorting time** of a C++ in-memory query engine **by 15%** by identifying bottlenecks and enhancing the code to use compile time code generation techniques (C++11 Variadic templates).
- **Improved query run time** of benchmark set **by 20%** by enhancing caching algorithm logic modifying cache seed logic to cache the data post relevant processing rather than raw data.
- Introduced a new query syntax to enable auto discovery of backend tables bypassing the current requirement of manual import. **Reduced each ongoing release time by a few weeks**. Used YACC, LEX, and C++.
- Improved cache hit rate by changing the internal load balancer logic to create a deterministic server-user mapping instead of a session based allocation.

PROJECTS

Vision and Language Navigation Sep 2020 - Mar 2021
Research exploration with Prof. Devi Parikh

- Built an agent for the task of following English language navigation instructions in a simulated house environment (VLN-CE)
- Built a hierarchical planner + controller architecture - planner predicts an intermediate waypoint and controller navigates to the waypoint. This improved sample complexity and overcame problems with long-range planning
- Incorporated semantic maps of the environment for better grounding and trained via both Imitation Learning and Reinforcement Learning
- Obtained a Success Rate of 24.6%, compared to the baseline of 23%

Test-time training for novel-object image captioning Jan 2020 - May 2020
Research exploration with Prof. Devi Parikh

- Built a model to generate captions for images containing novel objects (objects not mentioned in train data) using non-paired data
- Trained a multi-modal transformer model (ViLBERT) to predict the CIDEr score (measures image-text match) without needing access to reference captions
- Using VIFIDEL, SLOR, and predicted CIDER values as rewards, trained a captioning model to optimize for these metrics using policy-gradient methods

SELECTED COURSEWORK

- Deep Learning • Deep Learning for Text • Reinforcement Learning
- Machine Learning • Machine Learning for Trading
- Computation and Brain • Graduate Algorithms