Sashank Gondala

☎ +1(703)712-2706 • ✉ sgondala@gatech.edu • 🌐 www.sgondala.com/

Education

Georgia Institute of Technology **Atlanta, Georgia**
Master of Science, Computer Science with specialization in Machine Learning; **GPA 3.8/4** May 2021

- MS Project Advisor: *Prof. Dhruv Batra*
- TA for CS 7643 - Deep Learning, CS 7641 - Machine Learning, CS8803 - Systems for ML Research
- Selected Coursework: Deep Learning, Deep Learning for Text, Reinforcement Learning, Machine Learning, Machine Learning for Trading, Graduate Algorithms

Indian Institute of Technology Bombay **Mumbai, India**
Bachelor of Technology, Computer Science and Engineering; **GPA 8.24/10** May 2016

- IIT-JEE 2012 **All India Rank 14** out of 500k candidates
- **All India Rank 8** in Nationwide Education and Scholarship Test 2014. Received scholarship for the same
- Selected Coursework: Computer Graphics, Operating Systems, Compilers, Digital Geometry Processing

Conference Submissions

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

Abstract: Language models (LMs) for virtual assistants (VAs) are typically trained on large amounts of data, resulting in prohibitively large models which require excessive memory and/or cannot be used to serve user requests in real-time. Entropy pruning results in smaller models but with significant degradation of effectiveness in the tail of the user request distribution. We customize entropy pruning by allowing for a keep list of infrequent n-grams that require a more relaxed pruning threshold, and propose three methods to construct the keep list. Each method has its own advantages and disadvantages with respect to LM size, ASR accuracy and cost of constructing the keep list. Our best LM gives 8% average Word Error Rate (WER) reduction on a targeted test set, but is 3 times larger than the baseline. To reduce the size of the keep list and the resulting LM, we propose two discriminative methods to select a subset of n-grams, one based on recognition errors of synthesized audio and one based on a text-only model that approximates the recognition errors. We show that the approximate version, while being much cheaper than the 'real' versiona, can retain the majority of the observed WER reductions.

Work Experience

Apple (Cupertino, CA) **Jun 2020 - Aug 2020**
Machine Learning Research Internship (AI/ML)

- Worked as a research intern in the Language Modeling team
- Explored ways to maximize speech recognition accuracy while keeping the size of the language model small
- Obtained 10% reduction in LM size with negligible increase in WER (3x better than random)
- Currently under review at ICASSP 2021 (Shared first authorship with one other)
- Got a full-time offer for the role of 'Language Modeling Scientist'

Oracle HQ (Redwood City, CA) **Jul 2016 - Aug 2019**
Senior Member of Technical Staff

- **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.

Amazon **May 2015 – Jul 2015**
Software Developer Internship

- Worked with Amazon Custom, the team that deals with customized products
- Built an API test suite for services of Amazon Custom using TestNG in Java.

Housing.com **May 2014 – Jul 2014**
Software Developer Internship

- Deployed a linear time conversion funnel application to analyze user drop-off points on the website.

Projects

Training without reference captions — Jan 2020 - May 2020

- Worked on the idea of *training without reference captions* to improve performance of image captioning models on test-time out-of-distribution data as measured by CIDEr metric.
- Modeled a CIDEr predictor using a pretrained multi-modal transformer to predict the CIDEr score without needing access to reference captions
- Incorporated VIFIDEL, SLOR and predicted CIDER values as rewards and trained an image captioning model to optimize for these using policy-gradient methods

Automatic Augmentations for NLP — Aug 2020 - Current

- Using Neural Architecture Search to select automated augmentation strategies for scarce-data text classification.

Question Answering using Deep Learning — Oct 2019 - Dec 2019

- Worked on various Question Answering tasks - Google's Natural Question Answering and Stanford's SQuAD 2.0
- Implemented approaches such as LSTM based co-attention models, augmented BERT models, Ensembles, etc.

Neural Machine Translation — Mar 2019 - Apr 2019

- Implemented a sequence-to-sequence (Seq2Seq) network in PyTorch to translate Spanish text to English.
- Used a Bidirectional LSTM with multiplicative attention as Encoder and a Unidirectional LSTM as Decoder.

ML trading bot — Mar 2019 - Apr 2019

- Created a random forests-based trading algorithm which takes in the stock price and market indicators to predict the movement of a stock.

OpenAI Agents — Jan 2019 - Feb 2019

- Successfully trained agents to solve several of the OpenAI challenges, using a mix of Reinforcement Learning (RL) techniques such as Q-Learning, DQN, DDQN, and Policy Iteration

Companies Co-founded

Decentralized CDN — Feb 2018 - Dec 2018

- **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.
- The project was eventually discontinued due to a lack of product-market fit.

Technical Skills

- **Languages:** C++ (Expert) | Python (Expert) | Java (Intermediate) | Bash (Intermediate)
- **Others:** Scikit-learn (Expert) | PyTorch (Expert) | TensorFlow (Intermediate) | SQL (Intermediate)

Scholastic Achievements

- Secured **All India Rank 14** in IIT-JEE out of 500k test-takers **2012**
- Secured **All India Rank 59** in EAMCET out of 300k test-takers **2012**
- Obtained **7th position** in State Mathematics Olympiad (APAMT) **2009**
- Was placed **National Top 1%** in several Astronomy, Physics, and Junior Science Olympiads **2010-12**
- Attended (**30 students across the country**) Indian National Astronomy Olympiad (INAO), Indian National Junior Science Olympiad (INJSO) Olympiad Camps held by HBCSE **2010**

Extra-Curriculars

- Worked as the Class Representative for a batch of 97 people
- Worked as a member of Insight (IIT-B newsletter) and contributed to an article on *Academic Ethics*