# Brian **Lester**

*Mountain View, CA.*

(734) 834-2546 | blester125@gmail.com | blester125.com | blester125 | blester125 | Brian Lester | Brian Lester

## Education

**University of Pittsburgh** — *Pittsburgh, PA*
Double major, 3.51. **Computer Science** with Honors, 3.86. **Neuroscience**, 3.19. — *2012–2016*

**Shanghai American School** — *Shanghai, China*
High School — *2008–2012*
Advanced Placement Curriculum. Team lead for FIRST Robotics Competition.

## Publications

**The Power of Scale for Parameter-Efficient Prompt Tuning**
**Brian Lester**, Rami Al-Rfou, Noah Constant
***EMNLP 2021***, Online. **94** citations

**Finetuned Language Models Are Zero-Shot Learners**
Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, **Brian Lester**, Nan Du, Andrew M Dai, Quoc V Le
***ICLR 2022***, Online. **34** citations

**An Effective Label Noise Model for DNN Text Classification**
Ishan Jindal, Daniel Pressel, **Brian Lester**, Matthew Nokleby
***NAACL 2019***, Minneapolis, Minnesota. **18** citations

**A Library for Rapid Modeling, Experimentation and Development of Deep Learning Algorithms targeting NLP**
Daniel Pressel, Sagnik Ray Choudhury, **Brian Lester**, Yanjie Zhao, Matt Barta
***ACL 2018***; *NLP-OSS Workshop*, Melbourne, Australia. **11** citations

**SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer**
Tu Vu, **Brian Lester**, Noah Constant, Rami Al-Rfou, Daniel Matthew Cer
***ACL 2022*** *submission; Preprint arXiv:2110.07904*. **5** citations

**Multiple Word Embeddings for Increased Diversity of Representation**
**Brian Lester**, Daniel Pressel, Amy Hemmeter, Sagnik Ray Choudhury, Srinivas Bangalore
*Preprint arXiv:2009.14394*, 2020. **5** citations

**Constrained Decoding for Computationally Efficient Named Entity Recognition Taggers**
**Brian Lester**, Daniel Pressel, Amy Hemmeter, Sagnik Ray Choudhury, Srinivas Bangalore
***EMNLP 2020***; *Findings*, Online. **3** citations

**iobes: Library for Span Level Processing**
**Brian Lester**
***ACL 2020***; *NLP-OSS Workshop*, Online. **1** citation

**Dynamically Adjusting a Voice Recognition System**
**Brian Lester**, Sorin M Panainte
*US Patent 9,984,688*, 2018. **1** citation

**Intent Features for Rich Natural Language Understanding**
**Brian Lester**, Sagnik Ray Choudhury, Rashmi Prasad, Srinivas Bangalore
***NAACL 2021***; *Industry Track*, Online.

**Baseline: Strong, Extensible, Reproducible, Deep Learning Baselines for NLP**
Daniel Pressel, **Brian Lester**, Sagnik Ray Choudhury, Matt Barta, Yanjie Zhao, Amy Hemmeter
***NuerIPS 2018***; *OSS Workshop*, Montreal Quebec.

**Leader: Prefixing a Length for Faster Word Vector Serialization**
**Brian Lester**
*Preprint arXiv:2009.13699*, 2020.

## Work Experience

**Google** — *Mountain View, California*
AI Resident — *2020–Present*
Deep Learning research with a focus on Natural Language Processing, large pre-trained models, and zero-shot transfer.

- Prompt Tuning: an efficient method of controlling large frozen pre-trained language models based on T5. Matches performance of full fine-tuning using only $0.003\%$ of the parameters.
- Flan: Multitask training for a 137 billion parameter transformer-based decoder-only language model to create a model that is more effective at zero-shot prompting and performs better using Prompt Tuning.
- SPoT: Using multitask prompts as strong initialization for Prompt Tuning resulting in increased performance. Also used prompt similarity to estimate task similarity and to predict transferability.
- Added partial network training, lazy loading, and pre-filling of the auto-regressive cache to t5x, the open-source reimplementation of T5 in Jax. This final change reduced inference latency from $30$ seconds to $2.4$.

**Interactions** *Ann Arbor, Michigan*
Machine Learning Engineer *2018–2020*
Built production grade deep learning solutions and lead research efforts to push the boundaries of performance.
- Designed novel neural network architectures for calibrated intent detection, slot filling, and named entity linking using ConvNets, bLSTM-CRFs, ranking models, and transformer-based seq2seq models.
- Designed label space, annotation guidelines, and data collection method for NLU component of dialogue systems.
- Created a cloud-native model training platform based on declarative pipelines and kubernetes. Built a deployment platform that powers NLU for multiple production dialogue systems.
- Built efficient, batched implementations of complex neural network architectures such as Beam Search. My CRF implementation reduced training time by a factor of 10.

**Trove** *Ann Arbor, Michigan*
Lead Machine Learning Research Engineer *2017–2018*
Created a model training and serving platform that processed 200 million emails per day. Provided technical leadership to the ML team.
- Designed ConvNets for text classification to find sentences that contain questions. This powered a user-facing feature and was used to featurize the social graph created from email.
- Created neural ranking model was used to find coreferent mentions in the text and provide context to users.
- Used lexical features, as well as connectivity information in the email social graph, to identified bot accounts.

**Visteon Corporation** *Van Buren, Michigan*
Software Engineering Intern *2015*
Designed an adaptive system to minimize voice recognition errors. We patented this system and it is used in Mazda Cars.

# Presentations

**Prompt Tuning** *University of Michigan*
*University of North Carolina*
Deep Learning *2021–2022*
An overview of my work on Prompt Tuning, as well as our work—Flan and SPoT—directly built on Prompt Tuning. The talk includes a collection of insights about the behavior of soft prompts aggregated from others' followup work.

**NeurIPS Spotlight Talk on Mead-Baseline** *NeurIPS OSS Workshop*
Deep Learning *December 2018*
A spotlight talk, at the Open Source Software workshop at NeurIPS 2018, about our open-source toolkit, Mead-Baseline.

**Confidence and Calibration of Neural Network Models** *EMU ML Conference*
Deep Learning *March 2020* [1]
An overview of techniques used to adjust model calibration, evaluation of models that have the ability to "reject" decision with low confidence, and their uses in the NLU unit of a production dialogue system.

**Padding in Neural Networks for Natural Language Processing** *A2D-NLP*
Natural Language Processing *February 2020*
A survey of NLP building blocks with a focus on correctness and the need for padding in complex situations as well as places it is unexpected, like max-pooling following a 1D convolution.

**Optimization via NumPy and Cython** *Michigan Python Meetup*
Numerical Computation *January 2020*
I use a series of optimizations for computing pairwise Manhattan distance to introduce core NumPy concepts and Cython to reduce the runtime from multiple hours to just seconds.

**Input Representations of Deep Neural Networks** *PyData Ann Arbor*
Deep Learning *October 2017*
Using learned character-compositional input representations to create Deep Neural Networks with an open vocbaulary.

# Skills

| | |
|---|---|
| **Deep Learning** | Extensive experience building novel Neural Network architectures with Jax, Flax, Pytorch, and Tensorflow, generally for NLP. High-performance training with Data and Model based Parallelism, including multi-host distributed training on TPUs. |
| **Infrastructure** | Build and deploy with Kubernetes, Docker, Flux, MongoDB, Apache Nifi, Github Actions, and GitLab CI/CD. |
| **Toolkits** | NumPy, Pandas, Faiss, Gensim, SpaCy, NLTK, SciPy, Tensorflow-Datasets, Seaborn, and Matplotlib. |
| **Languages** | Python, Cython, Java, C, Javascript, C++, Elisp, and LaTeX. |

# Service & Public Scholarship

## Reviewing

| | |
|---|---|
| 2022 | NAACL |
| 2020–2021 | Computer Speech and Language |
| 2019 | CoNLL |

## Public Scholarship

2020 [1] **a²-dlearn:** Helped organize logistics, recruit speakers, and acquire funding through sponsorships

[1] Canceled due to COVID-19