

Brian Lester

Mountain View, CA.

☎ (734) 834-2546 | ✉ blester125@gmail.com | 🏠 blester125.com | 📧 blester125 | 🌐 blester125 | 📺 Brian Lester | 📺 Brian Lester

Education

University of North Carolina

PH.D. STUDENT WORKING WITH COLIN RAFFEL

University of Pittsburgh

DOUBLE MAJOR, 3.51. **COMPUTER SCIENCE** WITH HONORS, 3.86. **NEUROSCIENCE**, 3.19.

Shanghai American School

HIGH SCHOOL

Advanced Placement Curriculum. Team lead for FIRST Robotics Competition.

Chapel Hill, NC

2022–Present

Pittsburgh, PA

2012–2016

Shanghai, China

2008–2012

Selected Publications

h-index: 6

The Power of Scale for Parameter-Efficient Prompt Tuning

BRIAN LESTER, RAMI AL-RFOU, NOAH CONSTANT

EMNLP 2021, Online. ffd9d7f0b03b85747b001b4734d5ee31b5229aa4 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. 66c641587e6efa2d50a0e3b5a2fdf8ec2f529982 citations

SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer

TU VU, BRIAN LESTER, NOAH CONSTANT, RAMI AL-RFOU, DANIEL MATTHEW CER

ACL 2022, Dublin, Ireland and Online. c28b7dfe341f1e13a5a98efbce7946ef795cf9b8 citations

An Effective Label Noise Model for DNN Text Classification

ISHAN JINDAL, DANIEL PRESSEL, BRIAN LESTER, MATTHEW NOKLEBY

NAACL 2019, Minneapolis, Minnesota. 8b222b3915ce5115bcae189fccd6526125f079c2 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. 1db083076dff68bf43307b72720e8d06e9e73993 citations

Scaling Up Models and Data with t5x and seqio

ADAM ROBERTS, HYUNG WON CHUNG, ANSELM LEVSKAYA, GAURAV MISHRA, JAMES BRADBURY, DANIEL ANDOR, SHARAN NARANG, BRIAN LESTER, ET AL. (8/42)

Preprint arXiv:2203.17189, 2022. 1ed66e048bb025e75aa5ea660545285212e5341f 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. 988542a34d923ae96957bc692c67f120706ea37e citations

iobes: Library for Span Level Processing

BRIAN LESTER

ACL 2020; NLP-OSS Workshop, Online. d701ef7637f499b8c3145cac07deec3f94195ec4 citations

Dynamically Adjusting a Voice Recognition System

BRIAN LESTER, SORIN M PANAINTE

US Patent 9,984,688, 2018. 1 citation

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. 5be88bd5ab23d1ce85419d9aceec4588278e1ae1 citations

Reducing Retraining by Recycling Parameter-Efficient Prompts

BRIAN LESTER*, JOSHUA YURTSEVER*, SIAMAK SHAKERI, NOAH CONSTANT

Preprint arXiv:2208.05577, 2022. 2d3b6058370b804009fd7866098f4fca2d1894ca citations

Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation

TU VU, ADITYA BARUA, BRIAN LESTER, DANIEL MATTHEW CER, MOHIT IYER, NOAH CONSTANT

EMNLP 2022 submission; Preprint arXiv:2205.12647, 2022. 7fdec63d153c6a9f86ac51f72b98aa1a24c9b2af citations

Intent Features for Rich Natural Language Understanding

BRIAN LESTER, SAGNIK RAY CHOUDHURY, RASHMI PRASAD, SRINIVAS BANGALORE

NAACL 2021; Industry Track, Online. 7f6f7e56da14711b2645007492aece0af38ddd06 citations

Leader: Prefixing a Length for Faster Word Vector Serialization

BRIAN LESTER

Preprint arXiv:2009.13699, 2020. 25375a645d4633dd2c062e019cdc15823833f29d citations

Work Experience

Google Brain

SENIOR RESEARCH ENGINEER

Mountain View, California

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. Open-sourced our codebase and it has enabled 3 published papers, 1 product launch, and at least 5 in-flight papers.
- 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

MACHINE LEARNING ENGINEER

Ann Arbor, Michigan

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

LEAD MACHINE LEARNING RESEARCH ENGINEER

Ann Arbor, Michigan

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

SOFTWARE ENGINEERING INTERN

Van Buren, Michigan

2015

Designed an adaptive system to minimize voice recognition errors based on ASR confidence scores. We patented this system and it is used in Mazda Cars.

Presentations

Prompt Tuning

DEEP LEARNING

University of Michigan
University of North Carolina

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

DEEP LEARNING

NeurIPS OSS Workshop

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

DEEP LEARNING

EMU ML Conference

March 2020¹

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

NATURAL LANGUAGE PROCESSING

A2D-NLP

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

NUMERICAL COMPUTATION

Michigan Python Meetup

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

DEEP LEARNING

PyData Ann Arbor

October 2017

Using learned character-compositional input representations to create Deep Neural Networks with an open vocabulary.

Skills

Deep Learning

Extensive experience building novel Neural Network architectures, generally for NLP. High-performance training with Data and Model Parallelism, including multihost distributed training on TPU.

Infrastructure

Build and deploy with Kubernetes, Docker, Flux, MongoDB, Apache Nifi, Github Actions, and GitLab CI/CD. Experience training large neural networks on Google Cloud (GCP)

Toolkits

Jax, Flax, PyTorch, Tensorflow, NumPy, Pandas, Faiss, SpaCy, NLTK, Tensorflow-Datasets, Seaborn, and Matplotlib.

Languages

Python, Cython, Java, C, Javascript, C++, Lisp, and ~~TeX~~.

¹Canceled due to COVID-19

Service & Public Scholarship

REVIEWING

- 2022 IEEE Transactions on Affective Computing
- 2022 ARR: ACL Rolling Review
- 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