

Toronto, CA

□ (734) 834-2546 | Selester125@gmail.com | Ablester125.com | Oblester125 | Iblester125 | Iblester12

# **Education**

**University of Toronto** 

Ph.D. Student working with Colin Raffel. One year was at the University of North Carolina

**University of Pittsburgh** 

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

**Shanghai American School** 

HIGH SCHOOL, TEAM LEAD IN FIRST ROBOTICS COMPETITION.

Toronto, Canada 2022-Present Pittsburgh, PA 2012-2016 Shanghai, China 2008-2012

h-index: 10

# Selected Publications

# The Power of Scale for Parameter-Efficient Prompt Tuning

BRIAN LESTER, RAMI AL-RFOU, NOAH CONSTANT

EMNLP 2021, Online. 3873 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. 3596 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. 278 citations

#### Scaling Up Models and Data with t5x and segio

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, **Brian** Lester, Et Al. (8/42)

JMLR 2023, 2023. 195 citations

#### Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation

Tu Vu, Aditya Barua, Brian Lester, Daniel Matthew Cer, Mohit Iyyer, Noah Constant

EMNLP 2022, Abu Dhabi, UAE and Online. 64 citations

#### An Effective Label Noise Model for DNN Text Classification

ISHAN JINDAL, DANIEL PRESSEL, BRIAN LESTER, MATTHEW NOKLEBY

NAACL 2019, Minneapolis, Minnesota. 48 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. 16 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. **11** citations

#### **Reducing Retraining by Recycling Parameter-Efficient Prompts**

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

*Preprint arXiv:2208.05577*, 2022. **10** citations

# Git-Theta: A Git Extension for Collaborative Development of Machine Learning Models

Nikhil Kandpal\*, **Brian Lester\***, Mohammed Muqeeth, Anisha Mascarenhas, Monty Evans, Vishal Baskaran, Tenghao Huang, Haokun Liu. Colin Raffel

ICML 2023, Honolulu Hawaii. 10 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. 8 citations

## iobes: Library for Span Level Processing

**BRIAN LESTER** 

ACL 2020; NLP-OSS Workshop, Online. 7 citations

#### **Training LLMs over Neurally Compressed Text**

**BRIAN LESTER**, JAEHOON LEE, ALEX ALEMI, JEFFREY PENNINGTON, ADAM ROBERTS, JASCHA SOHL-DICKSTEIN, NOAH CONSTANT *TMLR* 2024</del>, 2024. 6 citations

#### 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. 2 citations

## **Intent Features for Rich Natural Language Understanding**

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

NAACL 2021; Industry Track, Online.

# **Work Experience**

**Google DeepMind** Mountain View, California SENIOR RESEARCH ENGINEER 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 of our codebase and it has enabled 5 published papers, 1 product launch, and at least 3 more in-flight papers.
  Flan: Multitask training for a 137 billion parameter transformer-based decoder-only language model to create a model that is more effec-
- tive 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

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

2017-2018

LEAD MACHINE LEARNING RESEARCH ENGINEER

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

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**

University of Michigan

University of North Carolina

DEEP LEARNING

**Prompt Tuning** 

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

A spotlight talk, at the Open Source Software workshop at NeurIPS 2018, about our open-source toolkit, Mead-Baseline.

December 2018

#### **Confidence and Calibration of Neural Network Models**

EMU ML Conference

DEEP LEARNING

March 2020 <sup>1</sup>

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-NIP

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

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**

PvData 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, 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++, Elisp, and MFX.

MAY 23, 2025 BRIAN LESTER · CURRICULUM VITAE

<sup>&</sup>lt;sup>1</sup>Canceled due to COVID-19

# Service & Public Scholarship.

# REVIEWING

2023 NeurIPS

2022 IEEE Transactions on Affective Computing

2022 ARR: ACL Rolling Review

2022-2024 NAACL

2020–2021 Computer Speech and Language

2019 CoNLL

# PUBLIC SCHOLARSHIP

2020 <sup>1</sup> **a<sup>2</sup>-dlearn:** Helped organize logistics, recruit speakers, and acquire funding through sponsorships