Andrew Drozdov

mrdrozdov.github.io | andrew.drozdov@databricks.com | mrdrozdov | RsU18o4AAAAJ

Summary

My research interests are in large language models and information retrieval. In particular, I develop algorithms for effective training and inference of deep learning models. Applications of my work include passage ranking, text generation, semantic parsing, among others.

Education

Ph.D. in Computer Science, *University of Massachusetts Amherst* Sep 2018 - Feb 2024
Focus: Machine Learning; Natural Language Processing (NLP); Information Retrieval
Thesis title: Unlocking Natural Language Generalization with Adaptive Retrieval-based Methods
Committee: Mohit Iyyer (Co-advisor), Andrew McCallum (Co-advisor), Hamed Zamani, Kyunghyun Cho, Jonathan Berant

M.S. in Computer Science, *New York University* Sep 2015 - Dec 2016
Mentors: Sam Bowman, Kyunghyun Cho

M.Eng. in Computer Science, *Cornell University* Sep 2013 - Dec 2013
Member of the inaugural cohort at Cornell Tech in NYC. Transitioned to full-time at Okta before completing the program.

B.S.E. in Computer Science, *University of Michigan* Sep 2009 - May 2013

Work Experience

Research Scientist, *Databricks - Mosaic Research (RAG Team), w/ Michael Carbin* Feb 2024 - Present

Student Researcher, *Google - Google Research, w/ Kai Hui & Don Metzler* Apr 2023 - Sep 2023

Research Intern & Student Researcher, *Google - Google Brain, w/ Xinying Song & Denny Zhou* Summer 2022

Research Intern, *IBM - IBM Research, w/ Ramon Astudillo, Tahira Naseem & Yoon Kim* Summer 2021

Research Intern & Student Researcher, *Google - Google AI Language* Summer 2019

Research Engineer, *eBay - Deep Learning Recommendation Systems* Aug 2017 - Aug 2018

Data Engineer, *Datadog* Summer 2015

Software Engineer, *Okta* Jun 2013 - Feb 2015

Selected Publications

Fused Text Generation for Retrieval-augmented Long-form Question Answering (Working Title)
First-author paper from Ph.D. work. Not yet published.

PaRaDe: Passage Ranking using Demonstrations with Large Language Models
A. Drozdov, H. Zhuang, Z. Dai, Z. Qin, R. Rahimi, X. Wang, D. Alon, M. Iyyer, A. McCallum, D. Metzler, K. Hui
EMNLP 2023 (Findings).

You can't pick your neighbors, or can you? When and how to rely on retrieval in the kNN-LM
A. Drozdov, S. Wang, N. Rahimi, A. McCallum, H. Zamani, M. Iyyer
EMNLP 2022 (Findings).

Compositional Semantic Parsing with Large Language Models
A. Drozdov, N. Schärli, E. Akyürek, N. Scales, X. Song, X. Chen, O. Bousquet, D. Zhou
ICLR 2022.

Additional Research

Multistage Collaborative Knowledge Distillation from Large Language Models
J. Zhao, W. Zhao, **A. Drozdov**, B. Rozonoyer, M. Sultan, J. Lee, M. Iyyer, A. McCallum
ACL 2024.

kNN-LM Does Not Improve Open-ended Text Generation
S. Wang, Y. Song, **A. Drozdov**, A. Garimella, V. Manjunatha, M. Iyyer
EMNLP 2023.

Inducing and Using Alignments for Transition-based AMR Parsing
A. Drozdov, J. Zhou, R. Florian, A. McCallum, T. Naseem, Y. Kim, R. Astudillo
NAACL 2022.

Improved Latent Tree Induction with Distant Supervision
A. Drozdov, Z. Xu, J. Lee, T. O'Gorman, S. Rongali, M. Iyyer, A. McCallum
EMNLP 2021.

Unsupervised Parsing with S-DIORA: Single Tree Encoding for DIORA
A. Drozdov, S. Rongali, Y. Chen, T. O'Gorman, M. Iyyer, A. McCallum
EMNLP 2020.

Unsupervised Labeled Parsing with DIORA
A. Drozdov, P. Verga, Y. Chen, M. Iyyer, A. McCallum
EMNLP 2019 (Short Paper).

Unsupervised Latent Tree Induction with Deep Inside-Outside Recursive Auto-Encoders (DIORA)
A. Drozdov, P. Verga, M. Yadav, M. Iyyer, A. McCallum
NAACL 2019 (Oral).

Emergent Communication in a Multi-Modal, Multi-Step Referential Game
K. Evtimova, **A. Drozdov**, D. Kiela, K. Cho
ICLR 2018.

Do latent tree learning models identify meaningful structure in sentences?
A. Williams, **A. Drozdov**, S. Bowman
TACL 2018.

Professional Service

Conference Leadership
ARR - Apr '24 - Senior Area Chair
ARR - Feb '24 - Area Chair

Invite-only Workshops
Information Retrieval Research in the Age of Generative AI CCC Workshop (co-located with SIGIR) - Jul '24
CIIR Brainstorming Session - Dec '23

Reviewing
I've reviewed over 100 research papers across a variety of AI and NLP conferences:
AAAI '19, '23–24
ACL '21
CoNLL '20–23
COLM '24
EMNLP '22–23
ICLR '22–24
ICML '20–23
Neurips '19–23
SIGIR '22–23
WSDM '24

Teaching & Mentorship

Hosted Interns:
Mathew Jacob @ Databricks w/ Michael Carbin and Matei Zaharia — Summer '24

UMass Amherst, *Teaching Assistant*
Industry Mentorship Course (CS-696DS) with Andrew McCallum. — Spring '22, Spring '23
Advanced Natural Language Processing (CS-685) with Mohit Iyyer. — Spring '22

Cornell University, *Teaching Assistant*
Data Science in the Wild (CS-5304) with Giri Iyengar at Cornell Tech. — Spring '18

Academic Mentorship: I have mentored 18 MS students and 1 BS student on research projects at UMass, primarily through independent studies with IESL and the industry mentorship course. Among others, topics have included knowledge distillation, cross-lingual training, and data mining. On these projects I've partnered with Amazon (Saleh Sulton), Bloomberg (Amanda Stent), and Chan Zuckerberg Initiative (Boris Veytsman).

Invited Talks

MosaicX Spotlight NYC (Lightning Talk). Evaluating RAG Systems. May 22, '24
NYU, Tal Linzen's lab. Unsupervised parsing, success and failures. Spring '22
UMass Amherst, Neural Networks (CS-682) taught by Erik Learned-Miller. Using transformers for NLP. Fall '21
MIT, NLP lab meeting invited by Yoon Kim. Neural alignments for AMR. Fall '21
CMU, Algorithms for NLP (CS-11711) taught by Emma Strubell. Unsupervised parsing with S-DIORA. Fall '20
IBM, NLP reading group, organized by Ramon Astudillo. Unsupervised parsing with DIORA. Spring '20

Awards

Top Reviewer, Neurips '22

Expert Reviewer, ICML '21

Top-33% Reviewer, ICML '20

Best Deep Learning Project (Jointly with K. Evtimova) Fall '16
NYU's Center of Data Science Award Ceremony. Award selected by Yann Lecun.
Project Title: Understanding Mutual Information and its Use in InfoGAN

Activities

RAG Reading Group @ Databricks, *Organizer* Spring '24 - Present

Data Science Tea, *Co-Organizer* Fall '18, Fall '19
Weekly speaker series covering a range of domains (attendance between 30-100 people depending on the topic).

Recurse Center, *Attendee* Spring '15
Participated in a three-month long *writer's retreat for programmers*, organizing seminars on artificial intelligence and distributed systems. Noteable community members include Michael Nielsen (Neural Networks and Deep Learning), Greg Brockman (OpenAI), among others.

Athletics
I was captain of my track & field team in high school, setting team mid-distance records and racing in the pentathlon. After recovering from an unrelated leg injury, I've since restored my passion in running.
High School PRs: 400m (0:51), 800m (2:10)
Recent PRs: 400m (1:03, May 2024), 800m (2:26, May 2024)