Eric Wallace

E-mail: ericwallace@berkeley.edu
Scholar: scholar.google.com/ericwallace
Twitter: twitter.com/Eric_Wallace_
Website: ericswallace.com

Education

UC Berkeley 2019 - Present
Ph.D. in Computer Science
GPA: 4.0/4.0

University of Maryland 2014 - 2018
B.S. in Computer Engineering
GPA: 3.9/4.0, GRE: 170/170Q, 168/170V, 6/6W

Research Experience

UC Berkeley Berkeley, California
Research Assistant Aug 2019 - Present
Advisors: Dan Klein, Dawn Song

Facebook AI Research (FAIR) Menlo Park, California
Research Intern June 2021 - Sept 2021
Advisors: Robin Jia, Douwe Kiela

Allen Institute for Artificial Intelligence (AI2) Irvine, California
Research Intern Jan 2019 - Aug 2019
Advisors: Matt Gardner, Sameer Singh

University of Maryland College Park, MD
Undergraduate Research Assistant Jan 2018 - Dec 2018
Advisor: Jordan Boyd-Graber

SWE Experience

Lyft, Self Driving Team Palo Alto, California
Software Engineering Intern June - Aug 2018

Intel Folsom, California
Software Engineering Intern Aug - Dec 2017

Fellowships, Awards & Honors

AI2 Intern of the Year, 2019
EMNLP Best Demo Award, 2019
EMNLP Travel Award 2018
EMNLP Best Reviewer Award, 2018
AIAA Student Conference Best Paper, 2017
Eagle Scout, 2012

Publications [Scholar]

[1] Calibrate Before Use: Improving Few-shot Performance of Language Models
Tony Z. Zhao*, **Eric Wallace***, Shi Feng, Dan Klein, and Sameer Singh
International Conference in Machine Learning (ICML), 2021.

[2] Extracting Training Data from Large Language Models
Nicholas Carlini, Florian Tramèr, **Eric Wallace**, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel
USENIX Security Symposium, 2021.

[3] Concealed Data Poisoning Attacks on NLP Models
Eric Wallace*, Tony Z. Zhao*, Shi Feng, and Sameer Singh
North American Chapter of the Association for Computational Linguistics (NAACL), 2021.

[4] Detoxifying Language Models Risks Marginalizing Minority Voices
Albert Xu, Eshaan Pathak, **Eric Wallace**, Maarten Sap, Suchin Gururangan, and Dan Klein
North American Chapter of the Association for Computational Linguistics (NAACL), 2021.

[5] Imitation Attacks and Defenses for Black-box Machine Translation Systems
Eric Wallace, Mitchell Stern, and Dawn Song.
Empirical Methods in Natural Language Processing (EMNLP), 2020.

[6] Evaluating Models' Local Decision Boundaries via Contrast Sets
Matt Gardner, Yoav Artzi, . . . , **Eric Wallace**, Ally Zhang, and Ben Zhou.
Findings of Empirical Methods in Natural Language Processing (EMNLP Findings), 2020.

[7] AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts
Taylor Shin*, Yasaman Razeghi*, Robert L Logan IV*, **Eric Wallace**, and Sameer Singh.
Empirical Methods in Natural Language Processing (EMNLP), 2020.

[8] Gradient-based Analysis for NLP Models is Manipulatable
Junlin Wang*, Jens Tuyls*, **Eric Wallace**, and Sameer Singh
Findings of Empirical Methods in Natural Language Processing (EMNLP Findings), 2020.

[9] Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers
Zhuohan Li*, **Eric Wallace***, Sheng Shen*, Kevin Lin*, Kurt Keutzer, Dan Klein, and Joseph E. Gonzalez
International Conference in Machine Learning (ICML), 2020.

[10] Pretrained Transformers Improve Out-of-Distribution Robustness
Dan Hendrycks*, Xiaoyuan Liu*, **Eric Wallace**, Adam Dziedzic, Rishabh Krishnan,
and Dawn Song.
Association for Computational Linguistics (ACL), 2020.

[11] Universal Adversarial Triggers for Attacking and Analyzing NLP
Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh.
Empirical Methods in Natural Language Processing (EMNLP), 2019.

[12] AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models
Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh.
Demo at Empirical Methods in Natural Language Processing (EMNLP), 2019.
Best Demo Award

[13] Do NLP Models Know Numbers? Probing Numeracy in Embeddings
Eric Wallace*, Yizhong Wang*, Sujian Li, Sameer Singh, and Matt Gardner.
Empirical Methods in Natural Language Processing (EMNLP), 2019.

[14] Misleading Failures of Partial-input Baselines
Shi Feng, **Eric Wallace**, and Jordan Boyd-Graber.
Association for Computational Linguistics (ACL), 2019.

[15] Compositional Questions Do Not Necessitate Multi-hop Reasoning
Sewon Min*, **Eric Wallace***, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi,
and Luke Zettlemoyer.
Association for Computational Linguistics (ACL), 2019.

[16] Understanding Impacts of High-Order Loss Approximations and Features in
Deep Learning Interpretation
Sahil Singla, **Eric Wallace**, Shi Feng, and Soheil Feizi.
International Conference in Machine Learning (ICML), 2019.

[17] Trick Me If You Can: Human-in-the-loop Generation of Adversarial Examples
for Question Answering
Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber.
Transactions of the Association for Computational Linguistics (TACL), 2019.

[18] Pathologies of Neural Models Make Interpretations Difficult
Shi Feng, **Eric Wallace**, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez,
and Jordan Boyd-Graber.
Empirical Methods in Natural Language Processing (EMNLP), 2018.

Teaching Experience

EMNLP 2020 Tutorial—*Interpreting Predictions of NLP Models* November 2020
Eric Wallace, Sameer Singh, Matt Gardner
A tutorial on interpretability methods for NLP, e.g., saliency maps, input perturbations, influence functions, and adversarial attacks.

Mentoring

Tony Z. Zhao (2020-2021), UC Berkeley Undergrad. Published [1, 3]. Now PhD student at Stanford.
Albert Xu (2020-2021), UC Berkeley Undergrad. Published [4]. Now PhD student at USC.
Eshaan Pathak (2020-2021), UC Berkeley Undergrad. Published [4].
Jens Tuyls (2019-2020), UC Irvine Undergrad. Published [8,12]. Now PhD student at Princeton.
Junlin Wang (2019-2020), UC Irvine Undergrad. Published [8,12]. Now Masters student at UCI.
Nikhil Kandpal (2019), UMD Undergrad. Published [11]. Now PhD at UNC.

Talks

March 2021. *What Can We Learn from Vulnerabilities of NLP Models?* Stanford NLP Seminar, Cornell NLP Seminar, DeepMind.

Conference Oral Presentations: ICML 2021 Virtual [1], NAACL 2021 Virtual [3], EMNLP 2020 Virtual [5], ICML 2020 Virtual [9], ACL 2020 Virtual [10], EMNLP 2019 Hong Kong [11], EMNLP 2018 Brussels [18].

Academic Service

Program Committee Member

- North American Chapter of the Association for Computational Linguistics (NAACL): 2021
- Association for Computational Linguistics (ACL): 2020, 2021
- International Conference on Machine Learning (ICML): 2021
- Neural Information Processing Systems (NeurIPS): 2020, 2021
- Empirical Methods in Natural Language Processing (EMNLP): 2021, 2020, 2019, 2018 (Top Reviewer)
- Workshops: RobustML Workshop (ICLR 2021), MRQA (EMNLP 2021), NLP for Positive Impact (ACL 2021), SRW (NAACL 2021)

Press & Media

Calibrate Before Use: Improving Few-shot Performance of Language Models [1], Sebastian Ruder Review, Synced, Venture Beat, Gwern Newsletter

Extracting Training Data from Large Language Models [2], Twitter #1, Twitter #2, Twitter #3, Google Blog, BAIR Blog, Nature News, Henry AI Labs, Yannic Kilcher, Top of Hacker News, Top of ML Reddit, Venture Beat, Full Scan (Spanish), WebBigData (Japanese), AI Times (Korean), Analytics India Mag, Sebastian Ruder Highlights.

Concealed Data Poisoning Attacks on NLP Models [3], Import AI, Twitter Discussion 1, Twitter Discussion 2, Sebastian Ruder Highlights

Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers [9], Twitter, TWiML Talk Podcast, Sebastian Ruder Highlights, Towards Data Science, Henry AI Labs Video, Synced, BAIR Blog, Sebastian Ruder Newsletter

Universal Adversarial Triggers for Attacking and Analyzing NLP [11], Twitter, Wired, qbitai, Synced, Sebastian Ruder Newsletter, Freethink, Lilian Weng Blog

AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models [12], Twitter, InfoQ, UC Irvine, Sebastian Ruder Newsletter, Freethink

Trick Me If You Can: Human-in-the-loop Generation of Adversarial Examples for Question Answering [17], Front page of Reddit, Dukakis Shaping Futures, UMD Press Release, UMD Podcast, AI2 NLP Highlights Podcast.

Pathologies of Neural Models Make Interpretations Difficult [18]. AI2 NLP Highlights Podcast, TWiML Talk Podcast, UCI NLP, UMD.