Eric Wallace

E-mail: ericwallace@berkeley.edu
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 NLP, RISELab, BAIR) — Berkeley, California
Research Assistant — Aug 2019 - Present
Advisors: Dan Klein, Dawn Song

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

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

Industry 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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Interpreting Neural Networks With Nearest Neighbors
Eric Wallace*, Shi Feng*, and Jordan Boyd-Graber.
EMNLP Workshop on Analyzing and Interpreting Neural Networks (BlackboxNLP), 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 (LIME, input reduction, Anchors), and adversarial attacks (SEARs, universal adversarial triggers).

Mentoring

Tony Zhao (2020-Present), UC Berkeley Undergraduate.
Nikhil Kandpal (2019), Independent Researcher. Published [7]. Now PhD Student at UNC.
Jens Tuyls (2019-2020), UC Irvine Undergraduate. Published [4,8]. Now PhD Student at Princeton.
Junlin Wang (2019-2020), UC Irvine Undergraduate. Published [4,8]. Now Masters Student at UC Irvine.

Talks

November 2020. *Imitation Attacks and Defenses for Black-box Machine Translation Systems.* Empirical Methods in Natural Language Processing (EMNLP) in Virtual.

July 2020. *Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers.* International Conference in Machine Learning (ICML) in Virtual.

July 2020. *Pretrained Transformers Improve Out-of-Distribution Robustness.* Association for Computational Linguistics (ACL) in Virtual.

November 2019. *Universal Adversarial Triggers for Attacking and Analyzing NLP.* Empirical Methods in Natural Language Processing (EMNLP) in Hong Kong.

November 2018. *Pathologies of Neural Models Make Interpretation Difficult.* Empirical Methods in Natural Language Processing (EMNLP) in Brussels, Belgium.

Academic Service

Program Committee Member

- North American Chapter of the Association for Computational Linguistics (NAACL): 2021
- Association for Computational Linguistics (ACL): 2020
- Neural Information Processing Systems (NeurIPS): 2020
- Empirical Methods in Natural Language Processing (EMNLP): 2020, 2019, 2018 (*Top Reviewer*)

Open Source Software

AllenNLP (Contributor)
A software library with abstractions for NLP research, written on top of PyTorch. Developer of the AllenNLP Interpretation Toolkit [8].

Press & Media

Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers [5], Twitter, TWiML Talk Podcast, Towards Data Science, Henry AI Labs Video, Synced, BAIR Blog, NLP Newsletter

Universal Adversarial Triggers for Attacking and Analyzing NLP [7], Twitter, Wired, qbitai, Synced, NLP Newsletter, Freethink.

AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models [8], Twitter, InfoQ, UC Irvine, NLP Newsletter, Freethink

Trick Me If You Can: Human-in-the-loop Generation of Adversarial Examples for Question Answering [13], Front page of Reddit, Dukakis Shaping Futures, UMD Press Release, UMD Podcast, AI2 NLP Highlights Podcast.

Pathologies of Neural Models Make Interpretations Difficult [14]. AI2 NLP Highlights Podcast, TWiML Talk Podcast, UCI NLP, UMD.