Cornell University
Cornell Tech,
New York, NY

arush@cornell.edu
http://rush-nlp.com
@srush_nlp

# Alexander M. Rush

## Appointment

*Cornell University Computer Science Department*
2019-. Associate Professor of Computer Science

*Hugging Face*
2019-. NLP Open-Source Startup

*Harvard University School of Engineering and Applied Sciences*
2015-2019. Assistant Professor of Computer Science

*Facebook Artificial Intelligence Research Lab*
2015. Post-Doctoral Fellowship Advisor: Yann LeCunn

## Education

*Massachusetts Institute of Technology*
2009-2014. Ph.D, Computer Science. Advisor: Michael Collins
Dissertation: *Relaxation Methods for Natural Language Decoding*.

*Harvard University*
2007. B.A., Computer Science.

## Awards

2023 Best Paper Runner-up - NeurIPS
Outstanding Paper - EMNLP
2021 Best Demo Paper - EMNLP
Outstanding Short Paper - NAACL
Sloan Fellowship
2020 Best Demo Paper (Runner-Up), ACL
Best Paper - DAC (Hardware)
Best Demo Paper - EMNLP
2019 NSF Career Award
Best Demo Paper - Nominee, ACL

## Grants

2018 Senior Program Chair, ICLR
Best Paper - Runner-Up, VAST (Visualization)
2017 Best Demo - Runner-Up, ACL
Invitation IJCAI Early Research Spotlight
Best Paper - Runner-Up, EMNLP
2015 NIPS Deep Learning Symposium (Invited Paper)
2012 Best Paper Award, NAACL
2010 Best Paper Award, EMNLP

2019 NSF Career Award
Sony Faculty Awards
2018 Google, Facebook, and Amazon AWS Faculty Awards
2017 Bloomberg and Intel AI Collaboration Faculty Awards
2016 Microsoft Azure and Samsung AI Award
2015 Google Faculty Award

Publications

(Full list: http://bit.do/alexander-rush)

**Highly Cited Publications (Google Scholar Metrics)**

Alexander M. Rush, Sumit Chopra, and Jason Weston. *A Neural Attention Model for Abstractive Sentence Summarization.* *EMNLP 2015. (1600 citations, 3rd Most Cited AAAI Paper 2015-2020)*

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. *Character-Aware Neural Language Models.* *AAAI 2016, (1250 citations, 3rd Most Cited AAAI Paper 2015-2020)*

Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush. *LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks.* *InfoVis 2017, (139 citations, 12th Most Cited IEEE Transactions on Visualization and Computer Graphics Paper 2015-2020)*

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush. *OpenNMT: Open-Source Toolkit for Neural Machine Translation.* *ACL Demo 2017 (900 citations, Best Demo Runner-up, 5th Most Cited ACL Paper 2015-2020)*

**All Conference Papers**

[1] J. O. Yin, Alexander Rush. *Compute-Constrained Data Selection.* *Preprint 2024*

[2] J. X. Morris, Alexander Rush. *Contextual Document Embeddings.* *Preprint 2024*

[3] Y. Lu, J. N. Yan, S. Yang, J. T. Chiu, S. Ren, F. Yuan, W. Zhao, Z. Wu, Alexander Rush. *A controlled study on long context extension and generalization in llms.* *Preprint 2024*

[4] J. Wang, D. Paliotta, A. May, Alexander Rush, T. Dao. *The mamba in the llama: Distilling and accelerating hybrid models.* *NeurIPS 2024*

[5] S. Geng, W. Zhao, Alexander Rush. *Great Memory, Shallow Reasoning: Limits of NN-LMs.* *Preprint 2024*

[6] J. N. Yan, T. Liu, J. Chiu, J. Shen, Z. Qin, Y. Yu, C. Lakshmanan, Y. Kurzion, Alexander Rush. *Predicting text preference via structured comparative reasoning.* *ACL 2024*

[7] W. Zhao, G. Gao, C. Cardie, Alexander Rush. *I Could've Asked That: Reformulating Unanswerable Questions.* *EMNLP 2024*

[8] Yash Akhauri, Ahmed F AbouElhamayed, Jordan Dotzel, Zhiru Zhang, Alexander M Rush, Safeen Huda, Mohamed S Abdelfattah. *ShadowLLM: Predictor-based Contextual Sparsity for Large Language Models.* *EMNLP 2024*

[9] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, Volodymyr Kuleshov. *Simple and Effective Masked Diffusion Language Models.* *NeurIPS 2024*

[10] Junxiong Wang, Ali Mousavi, Omar Attia, Ronak Pradeep, Saloni Potdar, Alexander M Rush, Umar Farooq Minhas, Yunyao Li. *Entity disambiguation via fusion entity decoding.* *NAACL 2024*

[11] Junxiong Wang, Tushaar Gangavarapu, Jing Nathan Yan, Alexander M. Rush. *MambaByte: Token-free Selective State Space Model.* *COLM 2024*

[12] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, Thomas Wolf. *Zephyr: Direct Distillation of LM Alignment.* *COLM 2024*

[13] Celine Lee, Abdulrahman Mahmoud, Michal Kurek, Simone Campanoni, David Brooks, Stephen Chong, Gu-Yeon Wei, Alexander M. Rush. *Guess and Sketch: Language Model Guided Transpilation.* *ICLR 2024*

[14] Justin T. Chiu, Wenting Zhao, Derek Chen, Saujas Vaduguru, Alexander M. Rush, Daniel Fried. *Symbolic Planning and Code Generation for Grounded Dialogue.* *EMNLP 2023*

[15] Zhiying Xu, Francis Y. Yan, Rachee Singh, Justin T. Chiu, Alexander M. Rush, Minlan Yu. *Teal: Learning-Accelerated Optimization of WAN Traffic Engineering.* *SIGCOMM 2023*

[16] John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, Alexander M. Rush. *Text Embeddings Reveal (Almost) As Much As Text.* *EMNLP 2023*

[17] John X. Morris, Chandan Singh, Alexander M. Rush, Jianfeng Gao, Yuntian Deng. *Tree Prompting: Efficient Task Adaptation without Fine-Tuning.* *EMNLP 2023*

[18] Wenting Zhao, Justin T. Chiu, Claire Cardie, Alexander M. Rush. *HOP, UNION, GENERATE: Explainable Multi-hop Reasoning without Rationale Supervision.* *EMNLP 2023*

[19] Junxiong Wang, Jing Nathan Yan, Albert Gu, Alexander M. Rush. *Pretraining Without Attention.* *EMNLP 2023 Findings*

[20] Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, Colin Raffel. *Scaling Data-Constrained Language Models.* *NeurIPS 2023 (Oral)*

[21] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, Victor Sanh. *OBELISC: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents.* *NeurIPS 2023 Dataset*

[22] Wenting Zhao, Justin T. Chiu, Claire Cardie, Alexander M. Rush. *Abductive Commonsense Reasoning Exploiting Mutually Exclusive Explanations.* *ACL 2023*

[23] Yuntian Deng, Noriyuki Kojima, Alexander M. Rush. *Markup-to-Image Diffusion Models with Scheduled Sampling.* *ICLR 2023*

[24] Thierry Tambe, Jeff Zhang, Coleman Hooper, Tianyu Jia, Paul N. Whatmough, Joseph Zuckerman, Maico Cassel dos Santos, Erik Jens Loscalzo, Davide Giri, Kenneth L. Shepard, Luca P. Carloni, Alexander M. Rush, David Brooks, Gu-Yeon Wei. *A 12nm 18.1TFLOPs/W Sparse Transformer Processor with Entropy-Based Early Exit, Mixed-Precision Predication and Fine-Grained Power Management.* *ISSCC 2023*

[25] BigScience Workshop. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.* *Arxiv Preprint*

[26] David Chiang, Alexander M. Rush, Boaz Barak. *Named Tensor Notation.* *TMLR 2022*

[27] Zhiying Xu, Sivaramakrishnan Ramanathan, Alexander Rush, Jelena Mirkovic, Minlan Yu. *Xatu: boosting existing DDoS detection systems using auxiliary signals.* *CoNEXT 2022*

[28] John X Morris, Justin T Chiu, Ramin Zabih, Alexander M Rush. *Unsupervised Text Deidentification.* *EMNLP Findings 2022*

[29] Yuntian Deng, Volodymyr Kuleshov, Alexander M Rush. *Model Criticism for Long-Form Text Generation.* *EMNLP 2022*

[30] Leandro von Werra et al.. *Evaluate and Evaluation on the Hub: Better Best Practices for Data and Model Measurement.* *EMNLP Demos 2022 (Best Demo)*

[31] Hendik Strobelt et al.. *Interactive and Visual Prompt Engineering for Ad-hoc Task Adaptation with Large Language Models.* *IEEE Trans on Visualization 2022*

[32] Thierry Tambe et al.. *A 16-nm SoC for Noise-Robust Speech and NLP Edge AI Inference With Bayesian Sound Source Separation and Attention-Based DNNs.* *IEEE Solid-State Circuits 2022*

[33] Stephen Bach et al.. *Promptsource: An integrated development environment and repository for natural language prompts.* *ACL Demo 2022*

[34] Samantha Petti, et al.. *End-to-end learning of multiple sequence alignments with differentiable Smith-Waterman.* *Bioinformatics*

[35] Victor Sanh, et al.. *Multitask prompted training enables zero-shot task generalization.* *ICLR 2022*

[36] Stanislav Lukyanenko et al.. *Developmental Stage Classification of Embryos Using Two-Stream Neural Network with Linear-Chain Conditional Random Field.* *MICCAI 2021*

[37] Keyon Vafa, Yuntian Deng, David Blei, Alexander Rush. *Rationales for sequential predictions.* *EMNLP 2021*

[38] Justin Chiu, Yuntian Deng, and Alexander M. Rush. *Low-Rank Constraints for Fast Inference in Structured Models.* *NeurIPS 2021*

[39] Matthe Skiles et al.. *Conference demographics and footprint changed by virtual platforms.* *Nature Sustainability*

[40] Yuntian Deng and Alexander M. Rush. *Sequence-to-Lattice Models for Fast Translation.* *EMNLP Findings Short 2021*

[41] Quentin Lhoest et al. *Datasets: A Community Library for Natural Language Processing.* *EMNLP Demos 2021 (Best Demo)*

[42] Thierry Tambe and Others. *EdgeBERT: Sentence-Level Energy Optimizations for Latency-Aware Multi-Task NLP Inference.* *IEEE MICRO 2021*

[43] Hendrik Strobelt, Jambay Kinley, Robert Krueger, Johanna Beyer, Alexander M. Rush, Hanspeter Pfister. *GenNI: Human-AI Collaboration for Data-Backed Text Generation.* *IEEE VIS 2021*

[44] Demi Guo, Alexander M. Rush, Yoon Kim. *Parameter-efficient transfer learning with diff pruning.* *ACL 2021*

[45] Teven Le Scao, Alexander M. Rush. *How many data points is a prompt worth?.* *NAACL Short 2021 (Best Paper - Runner-Up)*

[46] François Lagunas, Ella Charlaix, Victor Sanh, Alexander M Rush. *Block pruning for faster transformers.* *ACL 2021*

[47] Steven Cao, Victor Sanh, Alexander M. Rush. *Low-Complexity Probing via Finding Subnetworks.* *NAACL Short 2021*

[48] Xinya Du, Alexander M. Rush, Claire Cardie. *Template Filling with Generative Transformers.* *NAACL Short 2021*

[49] Thierry Tambe, En-Yu Yang, Glenn G Ko, Yuji Chai, Coleman Hooper, Marco Donato, Paul N Whatmough, Alexander M Rush, David Brooks, Gu-Yeon Wei. *9.8 A 25mm2 SoC for IoT Devices with 18ms Noise-Robust Speech-to-Text Latency via Bayesian Speech Denoising and Attention-Based Sequence-to-Sequence DNN Speech Recognition in 16nm FinFET.* *IEEE International Solid-State Circuits Conference 2021*

[50] Yuntian Deng, Alexander M. Rush. *Cascaded Text Generation with Markov Transformers.* *NeurIPS 2020*

[51] Yao Fu, Chuanqi Tan, Bin Bi, Mosha Chen, Yansong Feng, Alexander Rush. *Latent Template Induction with Gumbel-CRFs.* *NeurIPS 2020*

[52] Victor Sanh, Thomas Wolf, Alexander M. Rush. *Movement Pruning: Adaptive Sparsity by Fine-Tuning.* *NeurIPS 2020*

[53] Justin T. Chiu, Alexander M. Rush. *Scaling Hidden Markov Language Models.* *EMNLP 2020*

[54] Congzheng Song, Alexander M. Rush, Vitaly Shmatikov. *Adversarial Semantic Collisions.* *EMNLP 2020*

[55] Demi Guo, Yoon Kim, Alexander M. Rush. *Sequence-Level Mixed Sample Data Augmentation.* *EMNLP 2020*

[56] Thierry Tambe, En-Yu Yang, Zishen Wan, Yuntian Deng, Vijay Janapa Reddi, Alexander Rush, David Brooks, Gu-Yeon Wei. *AdaptivFloat: A Floating-point based Data Type for Resilient Deep Learning Inference.* *DAC 2020 (Best Paper)*

[57] Thomas Wolf et al. *Transformers: State-of-the-art Natural Language Processing.* *EMNLP Demos 2020 (Best Demo)*

[58] Alexander Rush. *Torch-Struct: Deep Structured Prediction Library.* *ACL Demos 2020 (Best Demo Honorable Mention)*

[59] Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M. Rush, Yoav Artzi. *What is Learned in Visually Grounded Neural Syntax Acquisition.* *ACL 2020 (Short)*

[60] Xiang Lisa Li, Alexander M. Rush. *Posterior Control of Blackbox Generation.* *ACL 2020*

[61] Jiawei Zhou, Zhiying Xu, Alexander M. Rush, Minlan Yu. *Automating Botnet Detection with Graph Neural Networks.* *AutoML for Networking and Systems Workshop*

[62] Georgios A. Tritsaris, Yiqi Xie, Alexander M. Rush, Stephen Carr, Marios Mattheakis, Efthimios Kaxiras. *LAN – A materials notation for 2D layered assemblies.* *None*

[63] Udit Gupta, Brandon Reagen, Lillian Pentecost, Marco Donato, Thierry Tambe, Alexander M. Rush, Gu-Yeon Wei, David Brooks. *MASR: A Modular Accelerator for Sparse RNNs.* *PACT 2019*

[64] Joe Davison, Joshua Feldman and Alexander Rush. *Commonsense Knowledge Mining from Pretrained Models.* *EMNLP 2019*

[65] Zachary Ziegler, Yuntian Deng and Alexander Rush. *Neural Linguistic Steganography.* *EMNLP 2019*

[66] Yoon Kim, Chris Dyer, Alexander M. Rush. *Compound Probabilistic Context-Free Grammars for Grammar Induction.* *ACL 2019*

[67] Gehrmann S, Strobelt H, Krueger R, Pfister H, and Alexander M. Rush. *Visual Interaction with Deep Learning Models through Collaborative Semantic Inference.* *InfoVis 2019*

[68] Jiawei Zhou, Alexander M. Rush. *Simple Unsupervised Summarization by Contextual Matching.* *ACL 2019*

[69] Sebastian Gehrmann, Hendrik Strobelt, Alexander M Rush. *GLTR: Statistical Detection and Visualization of Generated Text.* *ACL Demo 2019 (Best Demo Honorable Mention)*

[70] Yoon Kim, Alexander M. Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, Gabor Melis. *Unsupervised Recurrent Neural Network Grammars.* *NAACL 2019*

[71] Adji B. Dieng, Yoon Kim, Alexander M. Rush, David M. Blei. *Avoiding Latent Variable Collapse With Generative Skip Models.* *AISTATS 2019*

[72] Fritz Obermeyer, Eli Bingham, Martin Jankowiak, Justin Chiu, Neeraj Pradhan, Alexander Rush, Noah Goodman. *Tensor Variable Elimination for Plated Factor Graphs.* *ICML 2019*

[73] Zachary M. Ziegler, Alexander M. Rush. *Latent Normalizing Flows for Discrete Sequences.* *ICML 2019*

[74] Yoon Kim, Sam Wiseman, Alexander M. Rush. *Deep Latent-Variable Models for Natural Language.* *EMNLP 2018 (Tutorial)*

[75] Sebastian Gehrmann, Falcon Z. Dai, Henry Elder, Alexander M. Rush. *End-to-End Content and Plan Selection for Data-to-Text Generation.* *INLG 2018*

[76] Yuntian Deng*, Yoon Kim*, Justin Chiu, Demi Guo, Alexander M. Rush. *Latent Alignment and Variational Attention.* *NIPS 2018*

[77] Sam Wiseman, Stuart M. Shieber, Alexander Rush. *Learning Neural Templates for Text Generation.* *EMNLP 2018*

[78] Sebastian Gehrmann, Yuntian Deng, Alexander Rush. *Bottom-Up Abstractive Summarization.* *EMNLP 2018*

[79] Luke Melas-Kyriazi, George Han, Alexander Rush. *Training for Diversity in Image Paragraph Captioning.* *EMNLP 2018 (Short)*

[80] Luong Hoang, Sam Wiseman, Alexander Rush. *Entity Tracking Improves Cloze-style Reading Comprehension.* *EMNLP 2018 (Short)*

[81] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush. *Seq2Seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models .* *VAST 2018, EMNLP-BlackBox 2018 (Best Paper - Honorable Mention)*

[82] Alexander M. Rush. *The Annotated Transformer.* *ACL NLP-OSS 2018*

[83] Jean Senellart, Dakun Zhang, Bo Wang, Guillaume Klein, J.P. Ramatchandirin, Josep Crego, Alexander M. Rush. *OpenNMT System Description for WNMT 2018: 800 words/sec on a single-core CPU.* *WNMT 2018 (First-Place CPU Speed/Memory)*

[84] Yoon Kim, Sam Wiseman, Andrew C. Miller, David Sontag, Alexander M. Rush. *Semi-Amortized Variational Autoencoders.* *ICML 2018*

[85] Brandon Reagen, Udit Gupta, Robert Adolf, Michael M. Mitzenmacher, Alexander M. Rush, Gu-Yeon Wei, David Brooks. *Compressing Deep Neural Networks with Probabilistic Data Structures.* *ICML 2018, SysML 2018*

[86] Allen Schmaltz, Yoon Kim, Alexander M. Rush, Stuart M. Shieber. *Adapting Sequence Models for Sentence Correction.* *EMNLP 2017*

[87] Sam Wiseman, Stuart M Shieber Alexander M. Rush. *Challenges in Data-to-Document Generation.* *EMNLP 2017*

[88] Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, Yann LeCun. *Adversarially Regularized Autoencoders.* *ICML 2018, NIPS 2017 Workshop*

[89] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush. *OpenNMT: Open-Source Toolkit for Neural Machine Translation.* *ACL Demo 2017 (Best Demo Runner-up)*

[90] Ankit Gupta, Alexander M. Rush. *Dilated Convolutions for Modeling Long-Distance Genomic Dependencies.* *ICML CompBio 2017 (Best Poster)*

[91] Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M. Rush. *Image-to-Markup Generation with Coarse-to-Fine Attention.* *ICML 2017*

[92] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush. *LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks.* *InfoVis 2017*

[93] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. *Structured Attention Networks.* *ICLR 2017*

[94] Greg Yang and Alexander M. Rush. *Lie-Access Neural Turing Machines.* *ICLR 2017*

[95] Yoon Kim and Alexander M. Rush. *Sequence-Level Knowledge Distillation.* *EMNLP 2016*

[96] Sam Wiseman and Alexander M. Rush. *Sequence-to-Sequence Learning as Beam-Search Optimization.* *EMNLP 2016 (Best Paper Runner-Up)*

[97] Peter Kraft, Hirsh Jain, and Alexander M. Rush. *An Embedding Model for Predicting Roll-Call Votes.* *Proceedings of EMNLP 2016*

[98] Allen Schmaltz, Alexander M. Rush, and Stuart M. Shieber. *Word Ordering Without Syntax.* *EMNLP 2016*

[99] Allen Schmaltz, Yoon Kim, Alexander M. Rush, and Stuart M. Shieber. *Sentence-Level Grammatical Error Identification as Sequence-to-Sequence Correction.* *Workshop Submission for AESW 2016 (Top Performing System)*

[100] Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. *Learning Global Features for Coreference Resolution.* *NAACL 2016*

[101] Sumit Chopra, Michael Auli, and Alexander M. Rush. *Abstractive Sentence Summarization with Attentive Recurrent Neural Networks.* *NAACL 2016*

[102] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. *Character-Aware Neural Language Models.* *AAAI 2016*

[103] Alexander M. Rush, Sumit Chopra, and Jason Weston. *A Neural Attention Model for Abstractive Sentence Summarization.* *EMNLP 2015.*

[104] Jason Weston, Antoine Bordes, Sumit Chopra, Tomas Mikolov, and Alexander M. Rush. *Towards AI-Complete Question Answering A Set of Prerequisite Toy Tasks.* *ArXiv Preprint*

[105] Sam Wiseman, Alexander M. Rush, Jason Weston, and Stuart M. Shieber. *Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution.* *ACL 2015.*

[106] Yacine Jernite, Alexander M. Rush, and David Sontag. *A Fast Variational Approach for Learning Markov Random Field Language Models.* *ICML 2015.*

[107] Lingpeng Kong, Alexander M. Rush, and Noah A. Smith. *Transforming Dependencies into Phrase Structures.* *NAACL 2015.*

*Journal Papers*

Alexander M. Rush and Michael Collins. *A Tutorial on Dual Decomposition and Lagrangian Relaxation for Inference in Natural Language Processing*. Journal of Artificial Intelligence Research 2012.

Professional Service

President / Founder : COLM 2024-
Secretary: ICLR Board 2021-2024
General Chair: ICLR (Deep Learning) 2020
Senior Program Chair: ICLR (Deep Learning) 2019
Senior Area Chair: TACL 2020; NeurIPS 2019; ACL (Generation) 2019; ACL (Generation) 2020
Area Chair: NAACL (Machine Learning) 2020; NIPS 2018; NAACL (Machine Learning) 2016; ICLR (NLP) 2017, 2018; EMNLP (Generation/Summarization) 2017; ACL (Parsing and Tagging) 2017
Tutorial Chair: NAACL 2016

Graduate Field Activity

Field Member - Computer Science
PhD Students
Graduated Students

Justin Chiu (2024); Cohere AI

Jiawei Zhou (2023); Assistant Professor, Stonybrook

Yuntian Deng (2023); Assistant Professor, Waterloo

Sam Wiseman (2019); Assistant Professor, Duke University

Yoon Kim (2020); Assistant Professor, MIT

Sebastian Gehrmann (2020); Google NLP

Current Cornell Students

Nathan Yan

Junxiong Wang

Jack Morris

Woojeong Kim

Celine Lee

Wenting Zhao

University Service

Program Director 2020-; Computer Science Program, the largest Masters program at Cornell Tech.

Admissions Committee 2020-2024

Teaching

| | |
|---|---|
| 2020-2024 | Instructor, Machine Learning Engineering (150 students) , Cornell Tech, Spring. |
| 2021 | Instructor, Topics in Machine Learning and NLP (10 students) , Cornell Tech, Spring. |
| 2020 | Instructor, Topics in Machine Learning and NLP (100 students), Cornell Tech, Spring. |
| | Instructor, Machine Learning Engineering (100 students) (Rated 4.75/5), Cornell Tech, Fall. |
| 2019 | Instructor, Machine Learning for NLP (50 students) , Harvard University, Spring. |
| 2018 | Instructor, Machine Learning for NLP (50 students) (Rated 4.8/5) , Harvard University, Spring. |
| | Instructor, Advanced Machine Learning (100 students), Harvard University, Fall. |
| 2017 | Instructor, Machine Learning (250 students), Harvard University, Spring. |
| | Instructor, Advanced Machine Learning (100 students), Harvard University, Fall. |
| 2016 | Instructor, Machine Learning for NLP (50 students) (Rated 4.9/5), Harvard University, Spring. |
| 2015 | Instructor, Artificial Intelligence (100 students), Harvard University, Fall. |
| 2013 | Instructor (with Michael Collins), Natural Language Processing, Columbia University, Fall. |
| | Head Teaching Assistant, Natural Language Processing , Michael Collins, Columbia University, Spring (taught on Coursera, 30,000+ registered students). |
| 2012 | Head Teaching Assistant, Natural Language Processing, Michael Collins, Columbia University, Fall. |

Patents

A neural attention model for abstactive summarization (Facebook). Alexander M. Rush, Sumit Chopra, Jason Weston 2017.

Techniques for discriminative dependency parsing (Google). Slav Petrov, Alexander M. Rush, 2015.

Efficient parsing with structured prediction cascades (Google). Slav Petrov, Alexander M. Rush, 2013

Determining user affinity towards applications on a social networking website (Facebook., Thomas S. Whitnah, Alexander M. Rush, Ding Zhou, Ruchi Sangvhi, 2010.

Personal Libraries

Llama2 Rust.

llama2 in rust.

LLM Training Puzzles.

puzzles for learning about distributed training.

Thinking Like Transformers.

learn to think like a transformers.

GPU-Puzzles.

A series of puzzles for learning about the core aspects of modern deep learning coding. Includes puzzles for tensors, gpu's, and auto-differentiation..

Annotated S4.

Annotated S4 is a pedagogical implementation of the S4 model for very long range sequnece modeling utilizing JAX as a method for explaining mathematically complex code..

PromptSource.

PromptSource is an IDE for producing natural language prompts on real datasets. It was the basis of the T0 model for large-scale multitask training..

Break Through AI.

Break Through AI is a free summer program for supporting female undergraduates to learn AI and ML skills in an applied environment. I teach an 8 week summer program on the core elements on ML in a coding first environment..

MiniConf.

MiniConf is a project developed for ICLR as an easy-to-use tool for hosting fully remote asynchronous virtual conferences. It was heavily used in 2020 to host ACL, ICML, AKBC, AIStats, EMNLP, NeurIPS, and many other virtual conferences..

MiniTorch.

MiniTorch is a DIY teaching library to walkthrough the process of building a tensor, autodifferentiation library from scratch. It is used to teach machine learning engineering at Cornell Tech..

Streambook.

Streambook is a literate programming environment designed to make it easy to write publishable Jupyter notebooks without ever having to open a browser or break your github flow..

### Named Tensor Notation.

Named Tensor Notation was a follow-up to the named tensor proposal to develop a mathematical notation for more explicit multi-dimensional dot products when describing neural network interactions..

### NLP Browser.

NLP Browser is a web app that lets any easily browse through more than 150 datasets used in NLP and hosted by Hugging Face. The app is a pretty addictive way to casually learn about new datasets and challenges..

### NamedTensor (Tensor Considered Harmful).

Named Tensor is a proposal for adding a new datastructure to mathematical libraries to tread tensors more like dicts and less like tuples. This blog post had the impact of getting PyTorch to add a NamedTensor annotation in v1.3 of the libary..

### Torch Struct.

Torch-Struct is a passion project of mine to test out whether deep learning libraries can be used to implement classical structured prediction. It includes heavily-tested reference reimplementations of many core NLP algorithms..

### OpenNMT.

A full service open-source neural machine translation system. Originally developed in Lua with Systran, since ported to PyTorch and TensorFlow and maintained externally..

### The Annotated Transformer.

The annotated transformer was an experiment in blogging based on literate papers. The idea was to teach researchers how an important model in NLP works by aligning the paper line-by-line with an implementation. The blog post was widely distributed, and there have been many follow-ups for new model..

Academic Internships

Research Intern, *Google Research*, 2011 – 2013 , New York, NY. Advisor: Slav Petrov.

Research Intern, *USC/ISI* , Summer 2010, Marina Del Rey, CA. Advisor: Liang Huang.

Industry

Lead Engineer (Platform Team), *Facebook*, 2007 – 2009, Palo Alto, CA.

Developed compiler for Facebook Markup Language (FBML) to sanitize user content.

Developed system for crowd-sourced translation of Facebook user text.

Invited Talks

2024 Panel, EMNLP Keynote.
Richard Karp Lecture, Simons Institute.
Keynote, IEEE Big Data.
2023 Panel, NeurIPS Keynote.
Invited Talk, JHU.
Invited Talk, NYU.
Keynote Talk, MLSys 2023.
Keynote Talk, Simon Workshop on LLMs .
Invited Talk, Dagstuhl Seminar.
Invited Talk, UCSD AI Seminar.
Invited Talk, Penn NLP Seminar.
Invited Talk, Stanford NLP Seminar.
2022 Invited Talk, SoCal NLP.
Invited Talk, LXMLS 2022.
Invited Talk, MASC 2022.
Invited Talk, Rutgers Efficient ML.
Invited Talk, Georgia Tech NLP Seminar.
Invited Talk, Pacific Research Lab NLP.
2021 Invited Talk, NeurIPS Crowd Source ML.
Invited Talk, NeurIPS AIPLANs Workshop.
Invited Talk, Microsoft Efficient ML.
Invited Talk, Stanford SysML.
Invited Talk, Lisbon Machine Learning.
Invited Talk, University of Cambridge.
Invited Talk, Oracle.
Invited Talk, UCSB.
2020 Invited Talk, ByteDance.
Invited Talk, Baidu.
Colloquium, UMass Lowell.
Invited Talk, London Machine Learning.
Invited Talk, UCSB.
Invited Talk, Baidu.
Invited Talk, Google AI.
Invited Talk, Oracle AI.
Invited Talk, EMNLP Structured Prediction WS.
Invited Talk, EMNLP SustaiNLP WS.

2019 Colloquium, University of Edinburgh, Spring.
Colloquium, Tel Aviv University, Spring.
Invited Talk, Conversational Intelligence Summer.
Invited Talk, GANocracy, MIT, Summer.
Invited Talk, OpenAI, MIT, Summer.
Invited Talk, NeuralGen Workshop NAACL Summer.
Invited Talk, Berkeley NLP, Fall.
Keynote, PyTorch Developers Conference, Fall.

2018 Invited Talk, University of Washington, Spring.
Invited Talk, Allen Institute for AI, Spring.
Invited Talk, MSR, Spring.
Keynote, American Machine Translation Association, Spring.
Invited Talk, University of Texas, Spring.
Invited Talk, University of Maryland, Spring.
Invited Talk, Georgetown, Spring.
Invited Talk, Lisbon ML Summer School, Summer.
Invited Talk, Columbia University, Fall.
Invited Talk, New York University - Text as Data, Fall.
Tutorial, EMNLP, Fall.

2017 Invited Talk, Google Faculty Day, Spring.
Invited Talk, New England Machine Learning Day, Spring.
Invited Talk, Google, Spring.
Invited Talk, Berkeley CS, Spring.
Invited Talk, Notre Dame, Spring.
Colloquium, TTI-Chicago, Spring.
Invited Talk, Apple, Siri Team, Spring.
Colloquium, Samsung Global AI Forum, Fall.
Invited Talk, AMD, Fall.

2016 Invited Talk, NYU, Fall.
Invited Talk, BBN Research, Fall.
Invited Talk, Bloomberg, Fall.
Invited Talk (Speech Group), MIT, Fall.
Invited Talk, IBM Research, Fall.
Colloquium, CMU, Fall.
Invited Talk, Stanford NLP, Summer.
Invited Talk, Oracle Labs, Summer.
Invited Talk, Twitter, Summer.
Colloquium, John Hopkins University, Spring.
Colloquium, Rakuten, Spring.

2014 Colloquium, University of Washington, Spring.
Colloquium, NYU, Spring.
Colloquium, CMU, Spring.
Colloquium, MIT, Spring.
Colloquium, Harvard, Spring.
Colloquium, TTIC, Spring.
Colloquium, University of Maryland, Spring.

2013 Invited Tutorial, UMBC, October.
Invited Talk, CS and Social Science Seminar, UMass Amherst, October.
Talk, NLP Seminar, Columbia University, October.
Invited Talk, ML Seminar, UMass Amherst, October.
Invited Talk, Johnson Research Labs, NY, August.
Invited Talk, Society for Historians of American Foreign Relations, Arlington, June.
Invited Talk, Columbia University, Spring.
Invited Talk, NLP Seminar, City University of New York, Spring.

2012 Invited Tutorial, Neural Information Processing Systems (NIPS), December.

2011 Invited Tutorial, Google Research, Mountain View, August.
Tutorial. Association of Computational Linguistic (ACL), June.
Invited Talk, ML Seminar, University of Massachusetts, Amherst, Spring.
ML Tea, MIT, January.

2010 NLP Seminar, USC/ISI, Summer.

2006 Invited Talk, Computational Linguistics Seminar, University of Pennsylvania, November.