Ananya Harsh Jha

ananyahj@uw.edu https://ananyahjha93.github.io/

Education

Sept 2024 – Present **PhD Computer Science Engineering, University of Washington**

Sept 2018 – May 2020 **MS Computer Science, NYU Courant**

Aug 2012 – Dec 2016 **BTech Computer Science and Engineering, IIIT Delhi**

Employment History

Sept 2024 – Present **Graduate Research Assistant at UW**

- *Advisors*: **Luke Zettlemoyer** and **Hanna Hajishirzi**
- Collaborating with **Aaron Defazio** and **Fabian Schaipp** on using the last iterate bound on SGD to predict validation loss curves.
- Broadly interested in directional convergence, acceleration, generalization, and a better understanding of the river valley picture of optimization.

Jul 2022 – Aug 2024 **Predoctoral Researcher at Allen Institute for AI**

- *Advisors*: **Iz Beltagy** and **Emma Strubell**
- Worked on distillation and model compression.
- Ran experiments for **OLMo** modeling and the **dOLMa** dataset paper.
- Wrote the online zero-shot evaluation code for the OLMo project.

Jul 2020 – Mar 2022 **Research Engineer at PyTorch Lightning**

- Worked with **Kyunghyun Cho** to develop stochastic-autoencoders that utilize data augmentation for state-of-the-art self-supervised learning performance compared to VAEs.
- Staff contributor to Lightning ecosystem of repositories: **PyTorch Lightning**, **Lightning Flash**, **Torch Metrics** and **Bolts**.
- Responsible for setting up and maintaining a 48 GPU cluster; from hardware procurement and co-location negotiations, to setting up user accounts and installing libraries.
- Helped in creating data structure/algorithm and ML interview questions, conducted 20+ algorithm/data structure interview and 5+ ML interviews.

Jun 2019 – Jul 2019 **Research Intern at DeepMagic Inc.**

- *Advisor*: **Davi Geiger**
- Led the research on combining self-attention with convolutions in HRNets, for human pose-estimation, matching state-of-the-art results.

Jan 2017 – Jul 2018 **Research Assistant at IIIT-Delhi**

- *Advisor*: **Saket Anand**
- Incorporated cycle-consistency in VAEs to disentangle image representations better than the previously proposed adversarial methods. Work published at ECCV 2018.

Jan 2016 – Jun 2016 **Data Science Intern at Elucidata**

- Created 2 libraries in R (12k+ lines of code) for statistical analysis of metabolomic data and metabolite reaction network.

Research Publications

1. D. Groeneveld, I. Beltagy, P. Walsh, *et al.*, "Olmo: Accelerating the science of language models," *ArXiv*, vol. abs/2402.00838, 2024. URL: https://api.semanticscholar.org/CorpusID:267365485.
2. L. Soldaini, R. Kinney, A. Bhagia, *et al.*, "Dolma: An open corpus of three trillion tokens for language model pretraining research," *ArXiv*, vol. abs/2402.00159, 2024. URL: https://api.semanticscholar.org/CorpusID:267364861.
3. A. Jha, D. Groeneveld, E. Strubell, and I. Beltagy, "Just chop: Embarrassingly simple llm compression," 2023. URL: https://api.semanticscholar.org/CorpusID:258865382.
4. N. S. Detlefsen, J. Borovec, J. Schock, *et al.*, "Torchmetrics - measuring reproducibility in pytorch," *Journal of Open Source Software*, vol. 7, no. 70, p. 4101, 2022. DOI: 10.21105/joss.04101.
5. W. Falcon, A. H. Jha, T. Koker, and K. Cho, "Aasae: Augmentation-augmented stochastic autoencoders," *CoRR*, vol. abs/2107.12329, 2021. arXiv: 2107.12329.
6. A. H. Jha, S. Anand, M. Singh, and V. Veeravasarapu, "Disentangling factors of variation with cycle-consistent variational auto-encoders," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018.

Open-source Contributions

- **In-loop LLM evaluation suite**
  An evaluation suite for in-loop zero-shot evaluation of LLMs. End-tasks were selected from the **Eleuther-AI eval harness** in a way that they run under a minute (on 4 A100s), in-loop, during training and demonstrate non-random task performance for models as small as 300M parameters. The evaluation suite has been configured to match the results of the Eleuther-AI eval harness repository. This has also been adapted for **OLMo** to help with modeling and data ablation decisions.
- **LAMB AND LARS optimizers**
  Open-source implementations for **LAMB** and **LARS** optimizers that modify Adam and SGD, respectively, to adapt them to larger batch sizes by scaling optimizer updates by a ratio of weight norm to gradient norm for each layer in the network.
- **Gradient-sync for negative samples across GPUs**
  The first PyTorch open-source implementation of **SimCLR** to contain gradient-synchronization for negative samples across GPUs while applying the contrastive loss.
- **TorchMetrics**
  Co-wrote the current version of **TorchMetrics**, which allows users to write new metrics without exposing them to the complexity of distributed coding in PyTorch but maintains and synchronizes metric states across multiple GPUs in the background. Currently the repository has $100+$ implemented metrics and $1.7k+$ Github stars.
- **GLUE in HuggingFace Transformers**
  Added 5/9 GLUE tasks to complete the GLUE benchmark in **HuggingFace Transformers**, back when it was the BERT repository.