Vaishnavi Shrivastava

Basic Information

Email: vaish.shrivastava@stanford.edu — Pronouns: *she/her/hers*
Homepage: https://vshrivas.github.io/ — Phone Number: (+1) 408-477-5322

Education

Stanford University — **Sep'22 – Jun'24 (projected)**
Master of Science, Computer Science
Advisor: Prof. Percy Liang

California Institute of Technology (Caltech) — **Sep'15 – Jun'19**
Bachelor of Science, Computer Science — **3.9/4.0**

Research Interests

Natural Language Processing: Question Answering, Commonsense Reasoning, Retrieval Augmentation, Prompting, Question Decomposition, Grounded Language Learning
Machine Learning: Few-shot Learning, Federated Learning, Deep Reinforcement Learning, Model Interpretability, Multi-modal Learning

Technical Skills

Languages: *Proficient*: Python, Java, C, C++ | *Basic*: C#, SQL
Toolkits: PyTorch, Keras, Tensorflow

Publications

[1] (NAACL 2022) F. Mireshghallah, **V. Shrivastava**, M. Shokouhi, T. Berg-Kirkpatrick, R. Sim, D. Dimitriadis. 2021. UserIdentifier: Implicit User Representations for Simple and Effective Personalized Sentiment Analysis. *https://aclanthology.org/2022.naacl-main.252/*

[2] (Preprint) **V. Shrivastava***, R. Gaonkar*, S. Gupta*, A. Jha. 2021. Exploring Low-Cost Transformer Model Compression for Large-Scale Commercial Reply Suggestions. *arXiv: 2111.13999*

Work and Research Experience

Research Assistant:

- **Stanford University**: Advised by Prof. Percy Liang — *(Sep'22 - Current)*
  Themes: Large language models, Retrieval Augmentation, Reasoning, Question Decomposition

Applied Scientist:

- **Microsoft AI**: Suggested Replies & Summarization — *(Sep'19 - Aug'22)*
  Themes: Dialog Systems, Model Compression, Personalization, Summarization

Software Engineering Intern:

- **Microsoft AI**: Knowledge Mining and Graphs Group — *(Jul'18 - Sep'18)*
  Themes: Key-Phrase Extraction, Part-of-Speech Tagging, Email Search
- **Microsoft**: Substrate Data Store Group — *(Jun'17 - Sep'17)*
  Themes: Multi-threading, Backend, Thread-Safe Caching
- **Dell-EMC**: — *(Jun'16 - Sep'16)*
  Themes: Distributed Computing Algorithms, Concurrent Services

Teaching Experience

Teaching Assistant:

- **Caltech**: Machine Learning & Data Mining, CS 155 — *(Jan'19 - Mar'19)*
- **Caltech**: Database System Implementation, CS 122 — *(Jan'18 - Mar'18)*

Recent Projects

Prompt-based Reasoning — *(Jul'21 - Present)*
Advisor: Prof. Percy Liang, Stanford University

- Developing novel techniques to integrate chain-of-thought prompting, question decomposition, and retrieval for more robust and reliable reasoning for question answering.

Personalized Language Models *(Jul'21 - Present)*

- Aim is building user-level personalized generative reply suggestion dialog systems with GPT-2.
- Developed a modified *Prefix-Tuning* based approach to learn user-embeddings to condition GPT-2 model for personalization, improving validation perplexity by 9% over vanilla prefix-tuning.
- Using *LoRA: Low-Rank Adaptation of Large Language Models* technique for more fine-grained personalization by directly personalizing weight updates to GPT-2's attention matrices.

Implicit Personalized User Representations *(Jul - Sep'21)*
Paper

- Investigated using uniformly distributed, non-trainable, user-specific prompts for user-personalization, instead of trainable embeddings, to circumvent periodically training embeddings per user.
- Demonstrated that we can outperform SOTA prefix-tuning based results on a suite of sentiment analysis by up to 13%, resulting in a paper.

Federating Adapters *(Jul - Aug'21)*

- To reduce communication overhead for large language models (LMs) during federated learning, proposed inserting bottleneck adapter layers and sharing client-server updates only on those layers.
- Improved communication costs by 121x on sentiment analysis, without significant accuracy drops.
- Proposed a user clustering mechanism to leverage *AdapterFusion* and further improve accuracy.

Factual Consistency for Abstractive Summarization *(Mar - Jun'21)*

- Developed an automated metric for evaluating factual consistency of summaries by few-shot tuning GPT-3 for question generation (QG) and question answering (QA).
- Generated questions on the summary using QG model, and answers to those questions first based on the source and then based on the summary using the QA model.
- Evaluated answer similarity between source and summary using an F1 score.

Selected Previous Projects

Multi-turn Conversation Modeling *(Nov'20 - Feb'21)*

- Modeled multi-turn conversations for contextualized response suggestions in dialog systems.
- Implemented shared-weight Hierarchical Transformers to encode prior utterances separately and aggregate them using a self-attention layer to form contextualized input representations.
- Saw substantial gains in offline metrics compared to previous single-turn model and baseline concatenating previous utterances as new input.

Low-Cost Transformer Model Compression *(Jul - Nov'20)*
Paper

- Experimented with low-cost methods to compress Transformer bi-encoder based reply suggestion system, reducing training and inference times by 42% and 35% respectively.
- Investigated how dataset size, pre-trained model use, and domain adaptation of the pre-trained model affected the performance of compression techniques.
- Discovered that large-data settings allow low-cost techniques to be very effective in compressing pre-trained model based architectures. Insights led to a paper and a talk.

Talks

"Supercharging Reply Suggestions: Model Compression Solutions and Insights from a Real-World Setting". Microsoft Machine Learning, AI and Data Science Conference (MLADS) 2021

Selected Leadership Positions

- Corporate Vice President, *Caltech IEEE*
- Treasurer, *Caltech Society of Women Engineers*
- Secretary, *Caltech Robogals*

References

Percy Liang, *Associate Professor, Stanford University*
Milad Shokouhi, *Partner Applied Scientist, Microsoft*
Dan Schwartz, *Principal Applied Scientist, Microsoft*
Donnie Pinkston, *Lecturer, Caltech*