

Vaishnavi Shrivastava

| | | |
|------------------------------|--|--|
| BASIC INFORMATION | Master's in Computer Science Applicant Homepage: https://vshrivas.github.io/ | Email: vashri@microsoft.com Phone Number: (+1) 408-477-5322 |
| EDUCATION | California Institute of Technology (Caltech) Bachelor of Science, Computer Science | Sep'15 – Jun'19 3.9/4.0 |
| RESEARCH INTERESTS | Natural Language Processing: Transfer Learning & Language Models, Question Answering, Abstractive Summarization, Dialog Systems, Multi-lingual Models Machine Learning: Few-shot Learning, Federated Learning, Deep Reinforcement Learning, Model Interpretability, Multi-modal Learning | |
| TECHNICAL SKILLS | Languages: <i>Proficient:</i> Python, Java, C, C++ <i>Basic:</i> C#, SQL Toolkits: PyTorch, Keras, Tensorflow | |
| PUBLICATIONS | [1] (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 [2] (Under review) F. Miresghallah, V. Shrivastava , M. Shokouhi, T. Berg-Kirkpatrick, R. Sim, D. Dimitriadis. 2021. UserIdentifier: Implicit User Representations for Simple and Effective Personalized Sentiment Analysis. arXiv: 2110.00135 | |
| WORK AND RESEARCH EXPERIENCE | Applied Scientist: <ul style="list-style-type: none">● Microsoft AI: Suggested Replies & Summarization (Sep'19 - Present) <i>Themes: Dialog Systems, Model Compression, Personalization, Summarization</i> Software Engineering Intern: <ul style="list-style-type: none">● Microsoft AI: Knowledge Mining and Graphs Group (Jul'18 - Sep'18) <i>Themes: Key-Phrase Extraction, Part-of-Speech Tagging, Email Search</i>● Microsoft: Substrate Data Store Group (Jun'17 - Sep'17) <i>Themes: Multi-threading, Backend, Thread-Safe Caching</i>● Dell-EMC: (Jun'16 - Sep'16) <i>Themes: Distributed Computing Algorithms, Concurrent Services</i> | |
| TEACHING EXPERIENCE | Teaching Assistant: <ul style="list-style-type: none">● Caltech: Machine Learning & Data Mining, CS 155 (Jan'19 - Mar'19)● Caltech: Database System Implementation, CS 122 (Jan'18 - Mar'18) | |
| RECENT PROJECTS | Personalized Language Models (Jul'21 - Present) <ul style="list-style-type: none">– Aim is building user-level personalized generative reply suggestion dialog systems with GPT-2.– Developed a modified <i>Prefix-Tuning</i> based approach to learn user-embeddings to condition GPT-2 model for personalization, improving validation perplexity by 9% over vanilla prefix-tuning.– Jointly trained a network with GPT-2 to generate embeddings as a function of user n-gram language signals, to solve the cold-start problem of personalizing responses for unseen users.– Using <i>LoRA: Low-Rank Adaptation of Large Language Models</i> 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 <ul style="list-style-type: none">– 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.

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.

Dialog System Triggering

(Feb - Jun'20)

- Trained a light-weight biLSTM classifier to decide which messages to trigger core reply suggestion system on, to prevent suggesting irrelevant responses to open-ended questions.
- Shipping the model led to reductions in latency and improved user engagement metrics, since fewer suggestions were shown to users for open-ended questions that the model struggled with.

SELECTED
PREVIOUS
PROJECTS

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

Milad Shokouhi, *Partner Applied Scientist, Microsoft*
Dan Schwartz, *Principal Applied Scientist, Microsoft*
Abhishek Jha, *ML Engineering Manager, Stripe*
Donnie Pinkston, *Lecturer, Caltech*