Hossein Entezari Zarch

Thomas Lord Department of Computer Science ◇ Viterbi School of Engineering ◇ University of Southern California
Phone: +1 (213)709-9486 ◇ Email: entezari@usc.edu
Website hoenza Google Scholar hossein-entezari

Research Interests

- ◇ Large-Scale Machine Learning
- ◇ Large Language Models
- ◇ Retrieval-Augmented Generation
- ◇ Machine Learning Systems

Education

University of Southern California, Los Angeles, California — 2023 - Present
Ph.D. in Computer Science
Advisor: Prof. Murali Annavaram

University of Southern California, Los Angeles, California — 2023 - 2024
M.Sc. in Computer Science — GPA: 3.95/4.0

University of Tehran, Tehran, Iran — 2017 - 2022
B.Sc. in Computer Engineering[Software] — GPA: 18.59/20.0
Thesis: "**Incentive Mechanism for Reliable Coded Federated Learning**; Application in Distributed Edge Computation"
Advisors: Prof. Hamed Kebriaei & Prof. Pooya Shariatpanahi

Honors and Awards

- ◇ Among the top 10 percentile students in the UT Computer Engineering Students Class
- ◇ Ranked 258 (Top 0.2%) in Iranian University Entrance Exam among more than 137,000 participants
- ◇ Member of the **National Organization for Development of Exceptional Talents(NODET)**

Publications

◇ **Hossein Entezari Zarch**, Abdulla Alshabanah, Chaoyi Jiang, and Murali Annavaram."CADC: Encoding User-Item Interactions for Compressing Recommendation Model Training Data". *arXiv preprint arXiv:2407.08108.*

◇ Chaoyi Jiang*, Abdulla Alshabanah*, **Hossein Entezari Zarch**, Keshav Balasubramanian, and Murali Annavaram."HuffmanEmbed: Using Huffman Coding for Embedding Table Compression in Deep Learning Recommendation Models". *(Submitted to AAAI 2025) (*Equal Contribution)*

◇ Milad Soltany*, Hesam Mojtahedi*, **Hossein Entezari Zarch***, Amirhossein Kazerouni*, Alireza Morsali, Azra Abtahi, and Farokh Marvasti."Ensemble Neural Representation Networks". *arXiv preprint arXiv:2110.04124.* (*Equal Contribution)

◇ Seyed Masoud Rezaeijo, **Hossein Entezari Zarch**, Hesam Mojtahedi, Nahid Chegeni, and Amir Danyaei. "Feasibility study of synthetic DW-MR images at different b-values in patients with prostate cancer compared with real DW-MR images: qualitative and quantitative assessment of CycleGAN, Pix2PiX, and DC2Anet models". *Applied Magnetic Resonance*, 2022.

◇ Seyed Masoud Rezaeijo, Mohammadreza Ghorvei, Razzagh Abedi-Firouzjah, Hesam Mojtahedi, and **Hossein Entezari Zarch**. "Detecting COVID-19 in chest images based on deep transfer learning and machine learning algorithms". *Egyptian Journal of Radiology and Nuclear Medicine*, 2021.

Research Experience

Graduate Research Assistant, Super Computing In Pocket (SCIP), University of Southern California
Supervisor: Prof. Murali Annavaram — Jan. 2023 - Present

- Authored articles on optimizing DLRMs by reducing training data size and embedding table compression.
- Enhancing LLM inference efficiency through advanced caching and computation techniques.
- Researching RAG models to reduce inference latency for extended contexts.

Research Assistant, Smart Networks Lab (SNL), University of Tehran
Supervisor: Prof. Hamed Kebriaei & Prof. Pooya Shariatpanahi Sept. 2021 - July 2022

- Worked on utilizing incentive mechanisms in coded federated learning systems

Research Assistant, Multimedia & Signal Processing Lab (MSL), Sharif University of Technology
Supervisor: Prof. Farokh Marvasti Mar. 2021 - Oct. 2021
& Apr. 2019 - Mar. 2020

- Authored one article around INR networks based on the obtained results
- Investigated RNN models, like LSTM, for bilingual translation
- Explored attention-based models, like Transformer, GPT, and BERT

Research Assistant, University of Tehran
Supervisor: Dr. Seyed Masoud Rezaeijo Dec. 2020 - Sept. 2021

- Investigated and examined various GAN architectures to find the best that learns transferring MRI images
- Assisting in authoring two articles based on the obtained results

Research Intern, Computational Modeling & Machine Learning Lab, University of Tehran
Supervisor: Prof. Babak N. Araabi Jun. 2020 - Sept. 2020

- Revised a cognitive task implemented in Matlab
- Processed data derived from a cognitive task in R

Research Intern, Nojan Robotics and Artificial Intelligence, University of Tehran Science & Technology Park
Supervisor: Prof. Behnam Bahrak Jul. 2020 - Sept. 2020

- Explored object detection and classification models to be used in a real-time sorter robot
- Fine-tuned object detection models, like YoloV3, YoloV5, Fast-RCNN, MobileNet

Work Experience

Software Engineer Intern, Divar Company, Tehran, Iran
Sept. 2022 - Dec. 2022

- Maintaining and developing the search part of the Divar application's backend in the Search & Submit team
- Collaborating with a four-person team of junior and senior software developers

Relevant Courses

- ◇ **Natural Language Dialogue Systems**
- ◇ **Ethics in Natural Language Processing**
- ◇ **Convex Optimization**
- ◇ **Mathematics of High-Dimensional Data**
- ◇ **Frontiers of Machine Learning**
- ◇ **Machine Learning**

Teaching Assistantship (Graduate courses are indicated by †)

- ◇ **Fundamentals of Computation**
- ◇ **Database Systems**†
- ◇ **Explorations in Computing**

Skills

Programming Languages:

- Proficient in C/C++, Python, Java, Verilog
- Familiar with R, Matlab, LaTeX, HTML, CSS, JavaScript

Software & Frameworks:

- Proficient in PyTorch, NumPy, sci-kit learn, Pandas, CVX/CVXPy
- Familiar with TensorFlow, Keras, MySQL, MongoDB, Docker, Kubernetes