ALPEREN GORMEZ

agorme2@uic.edu | alperengormez.github.io | linkedin.com/in/alperengormez | github.com/alperengormez | Google Scholar

EDUCATION

University of Illinois Chicago — **Chicago, IL**
Doctor of Philosophy in Electrical and Computer Engineering; Cumulative GPA: 4.0/4.0 — Aug 2019 - Present (2024)
Advisor: Asst. Prof. Erdem Koyuncu

Bilkent University — **Ankara, TURKEY**
Bachelor of Science in Electrical and Electronics Engineering — Aug 2015 - Jun 2019

PUBLICATIONS

4. **A. Görmez** and E. Koyuncu, "Dataset Pruning Using Early Exit Networks," *ICML Workshop on Localized Learning (LLW)*, 2023.
3. **A. Görmez** and E. Koyuncu, "Class Based Thresholding in Early Exit Semantic Segmentation Networks," arXiv:2210.15621 [cs.CV], 2022.
2. **A. Görmez** and E. Koyuncu, "Pruning Early Exit Networks," *2022 Sparsity in Neural Networks*, 2022.
1. **A. Görmez**, V. R. Dasari and E. Koyuncu, "E2CM: Early Exit via Class Means for Efficient Supervised and Unsupervised Learning," *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1-8.

WORK EXPERIENCE

•Apple — **Seattle, WA**
AIML Intern — May 2023 - Aug 2023
◊ Implemented 2 post training quantization and pruning algorithms in PyTorch in a production-ready and modular way for the on-device team to compress large language models. My branch got merged.
◊ Enhanced the model compression algorithms by implementing 3 new features resulting in a notable 4% further memory reduction improvement.
◊ Conducted extensive analysis by testing 366 different compression configurations across 11 open source and internal models on 13 datasets, evaluating 12 compression parameters.
◊ Fostered collaboration with research and hardware teams, exploring quantization, weight clustering and adapter approaches for further optimization.
◊ Identified and presented the optimal compression configuration, achieving 71% model size reduction without compromising performance. Delivered findings to the director for review.

•Roku — **San Jose, CA**
Machine Learning Intern — May 2021 - Aug 2021
◊ Led efforts to reduce the inference time of a CTR prediction model within the Advertising Engineering team.
◊ Leveraged mlpy for cross-feature generation and feature transformation, Apache Spark for large-scale data processing, and TFX for streamlining data pipelines.
◊ Attained a notable 0.03 improvement in AUC while adhering to stringent inference time requirements.
◊ Conducted in-depth experimentation with TensorFlow, exploring early exit networks and applying knowledge distillation techniques.

•University of Illinois Chicago — **Chicago, IL**
Teaching Assistant — Aug 2019 - Present
◊ Taught the ECE/CS 559 - Neural Networks course using PyTorch (2021 Fall, 2022 Fall).
◊ Instructed students MATLAB for the ECE 311 - Communication Engineering course (2020 Fall, 2021 Spring, 2022 Spring).
◊ Helped students in the ECE 317 - Digital Signal Processing I course (2019 Fall, 2020 Spring, 2023 Spring).

•ASELSAN — **Ankara, TURKEY**
Candidate Engineer — Feb 2019 - Jun 2019
◊ Designed neural networks in TensorFlow to achieve precise sound classification for passive sonar applications.
◊ Employed Python and Julia to visualize data acquired from ultrasonic sensors. Successfully identified a faulty sensor through insightful data analysis.
◊ Implemented sonar signal processing algorithms in MATLAB for the Acoustics Signal Processing Department.

RESEARCH EXPERIENCE

•University of Illinois Chicago   **Chicago, IL**
Research Assistant   Aug 2019 - Present (2024)

◇ For the first time in the literature, applied early exit networks to the task of dataset pruning and achieved a 60% reduction in deep learning model training costs.

◇ Leveraged the neural collapse phenomenon in early exit semantic segmentation models, resulting in a 23% reduction in computational costs while maintaining accuracy for edge devices.

◇ Investigated the combined impact of early exiting, pruning, and sparsity through PyTorch experimentation.

◇ Worked on early exit neural networks, adaptive inference, and model compression, which led to a 50% reduction in computational costs while preserving the performance.

◇ Conducted experiments on efficient distributed neural network training techniques.

◇ Provided mentorship and supervision to undergraduate students in early exit, knowledge distillation, conditional computation and object detection research projects.

◇ Participated in the following communities: EEML, tinyML, SNN, M2L.

•Nagoya University   **Aichi, JAPAN**
Research Student   Apr 2018 - Jul 2018

◇ Engaged in advanced research on pattern recognition and anomaly detection with guidance from Prof. Kenji Mase.

HONORS AND AWARDS

•Mediterranean Machine Learning Summer School 2023: Selected to attend the M2L.

•IEEE Computational Intelligence Society Travel Grant: Received a travel grant to attend IEEE WCCI 2022.

•Eastern European Machine Learning Summer School 2022: Received the top-voted poster award for E^2CM.

•Bilkent University Honor Student: High academic standing, 2015 - 2019.

•Bilkent University Comprehensive Scholarship: Full tuition waiver and stipend during the B.S. program, 2015 - 2019.

•LYS Degree: Ranked 341st in Turkey's National University Entrance Exam among over 2 million students, 2015.

OUTREACH AND MENTORING

•Lacmus Foundation
Contributor   Jul 2022 - Jan 2023

◇ Developed deep learning solutions for the open source Lacmus project to support search-and-rescue operations aimed at missing people.

•University of Illinois Chicago   **Chicago, IL**
Supervisor   May 2022 - May 2023

◇ Supervised an undergraduate student in early exit networks research.

◇ Managed an undergraduate student's research in the GPIP program, focusing on neural networks, knowledge distillation, and conditional computation.

◇ Mentored an undergraduate student in the creation of an object detection system, starting from the conceptualization phase to the final implementation.

•Deep Learning Indaba
Mentor   Jan 2021 - May 2023

◇ Volunteered as a mentor, providing guidance to students on research projects, industry applications, and graduate school pursuits to foster the growth of machine learning and artificial intelligence in Africa.

PROFESSIONAL ACTIVITIES

•Reviewer

◇ IEEE Global Communications Conference (GLOBECOM), 2023.

◇ IEEE Transactions on Computational Imaging, 2022.

•Member of the Organizing Team

◇ IEEE International Conference on Network Protocols (ICNP), 2019.

•Attendee

◇ Learning on Graphs 2022.

◇ PyTorch Conference 2022.

◇ NeurIPS 2022.

◇ AWS Summit Chicago 2022.

◇ Google PhD Summit Chicago 2020.