ALPEREN GORMEZ

alperengormez@gmail.com | alperengormez.github.io | linkedin.com/in/alperengormez | github.com/alperengormez | Google Scholar

CURRENT POSITION

•Meta **Menlo Park, CA, USA**
Research Scientist Dec 2024 - Present

- ◊ Drove efficiency and model-hardware co-design efforts for Meta's proprietary MTIA chip, successfully migrating production-scale recommendation models from NVIDIA GPUs, achieving a 2.5x performance increase and validating accuracy, resulting in millions of dollars in savings.
- ◊ Pioneered an LLM-driven agentic workflow for model development and debugging, substantially improving developer experience by automating complex, iterative root-cause analysis.

EDUCATION

University of Illinois Chicago **Chicago, IL, USA**
Doctor of Philosophy in Electrical and Computer Engineering; Cumulative GPA: 4.0/4.0 Aug 2019 - Oct 2024
Advisor: Assoc. Prof. Erdem Koyuncu
Ph.D. Dissertation: Efficient Neural Network Inference and Training Using Early Exit Strategies

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

Nagoya University **Nagoya, JAPAN**
School of Informatics Apr 2018 - Jul 2018

WORK EXPERIENCE

•Google **Mountain View, CA, USA**
Research Intern May 2024 - Aug 2024

- ◊ Designed real-time streaming sound separation models for Project Starline.
- ◊ Worked on audio-visual modeling, used Gemini for sound classification.
- ◊ Created a new dataset using Gemini API with the end goal of fine-tuning a pre-trained audio-visual model.

•Apple **Seattle, WA, USA**
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, USA**
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.

•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.

PUBLICATIONS

5. **A. Görmez** and E. Koyuncu, “Class-aware Initialization of Early Exits for Pre-training Large Language Models,” in *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024.
4. **A. Görmez** and E. Koyuncu, “Class Based Thresholding in Early Exit Semantic Segmentation Networks,” in *IEEE Signal Processing Letters*, vol. 31, pp. 1184-1188, 2024. Also in IEEE MLSP 2024.
3. **A. Görmez** and E. Koyuncu, “Dataset Pruning Using Early Exit Networks,” *ICML Workshop on Localized Learning (LLW)*, 2023. Also in M2L and Cohere for AI - ML Efficiency Group. Accepted for publication in ACM Transactions on Intelligent Systems and Technology.
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, “E^{2}CM: Early Exit via Class Means for Efficient Supervised and Unsupervised Learning,” *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1-8. Top-voted poster award in EEML.

RESEARCH EXPERIENCE

•University of Illinois Chicago **Chicago, IL**
Research Assistant Aug 2019 - Oct 2024

- Developed a novel weight initialization technique for early exit large language models (LLMs) to accelerate pre-training.
- Designed experiments to reduce the memory footprint of mixture of experts (MoE) based models.
- 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.
- Supervised a MSc student’s thesis on early exit networks for deep reinforcement learning. Held weekly meetings, suggested research directions and experiments.
- 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.
- Helped students in ECE 317 - Digital Signal Processing I, ECE 311 - Communication Engineering, ECE/CS 559 - Neural Networks, ECE 407 - Pattern Recognition courses.

•Nagoya University **Nagoya, 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^{2}CM.
•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

•University of Illinois Chicago **Chicago, IL, USA**
Supervisor May 2022 - Oct 2024

- Advised a MSc student on their thesis, which investigated early exit networks in deep reinforcement learning. Through weekly meetings, I helped shape their research direction and proposed specific experiment ideas.
- Supervised an undergraduate student’s research, focusing on neural networks, knowledge distillation, conditional computation and early exit networks.
- Mentored an undergraduate student in building an object detection system, starting from the conceptualization phase to the final implementation.

•Deep Learning Indaba
Mentor Jan 2021 - Jan 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.