

Fatih İlhan

Resume

School of Computer Science, College of Computing
Georgia Institute of Technology, Atlanta, GA, USA

e-mail: filhan@gatech.edu
web: fatih-ilhan.github.io
github: github.com/fatih-ilhan
google scholar profile: [DHB3X18AAAAJ](https://scholar.google.com/citations?user=DHB3X18AAAAJ)
orcid id: [0000-0002-0173-7544](https://orcid.org/0000-0002-0173-7544)

RESEARCH INTERESTS Efficient Inference/Fine-tuning for Large Language/Multi-modal Models, Systems for AI, Computer Vision, Distributed/Federated Learning, Ensemble Learning, Reinforcement Learning, AI Safety and Alignment

ACADEMIC EXPERIENCE **Georgia Institute of Technology** **Atlanta, GA, USA**
Ph.D. in Computer Science, CGPA: 3.84/4.00, Supervisor: [Prof. Ling Liu](#) Aug 2021 – Present

- Thesis Topic: Resource-adaptive Efficiency Optimizations for Large Vision-Language Models
- Coursework Focus: Machine Learning, Database Systems, Algorithms
- Published 13 papers (5 as first author) in top venues such as CVPR, NeurIPS, ICLR, EMNLP.
- Served as reviewer for CVPR, ICCV, AAAI, ICML, ICDCS, IEEE PAMI and IEEE TOIT.
- Head TA for the Advanced Internet Systems course with 5 TAs and 100-150 students, selected as the outstanding Head TA for OMSCS program (2024) and Georgia Tech Head TA of the Year (2025).

Bilkent University **Ankara, Türkiye**
M.Sc. in EEE, CGPA: 3.58/4.00, Supervisor: [Prof. Serdar Kozat](#) Sep 2019 – Aug 2021

- Thesis: Nonstationary Time Series Prediction with Markovian Switching RNNs
- Published 3 papers in top IEEE journals, served as reviewer for IEEE TNNLS and IEEE TSP.
- TA for the courses: Statistical Learning and Data Analytics, Neural Networks.

B.Sc. in Electrical and Electronics Engineering, CGPA: 3.81/4.00 Aug 2014 – Jun 2019

- Senior Project: GPS-independent outdoor localization system
- Specialization in signal processing, machine learning, communications
- Attended exchange program at Nagoya University, Japan (Spring 2018) and studied intelligent automobile systems.

Ankara Science High School **Ankara, Türkiye**
High School Degree, Science Track, CGPA: 95.26/100 Sep 2010 – Jun 2014

WORK EXPERIENCE **IBM Thomas J. Watson Research Center** **Yorktown Heights, NY**
Researcher Intern, Mentors: Dr. Gong Su, Dr. Donna Dillenger May-Aug 2022/23/24

- Worked on memory-efficient decoding with KV cache compression for long-context inference with LLMs, filed a patent application.
- Researched efficient pruning for LLM fine-tuning through CPU/GPU workload balancing. Our work led to a publication at CVPR24.
- Worked on computation-efficient federated learning under heterogeneous settings with on-premise deployments, filed a patent. Our work led to two publications at CVPR23 and ICDCS23.

DataBoss Analytics **Ankara, Türkiye**
Machine Learning Engineer Aug 2018 – Jul 2021

- Built [Predy.AI](#), an end-to-end pipeline for real-time spatio-temporal prediction, anomaly detection and recommendation systems, within a team of three engineers. Analyzed retail data from customer businesses to provide procurement and logistics insights, reduced consumption forecast errors by up to 40%.

Roketsan **Ankara, Türkiye**
Electronics Engineer Intern Jun 2017 – Jul 2017

- Integrated GPS and INS data using Extended Kalman Filter for navigation systems and enhanced localization precision by 85%. Built a Labview application for fast and simultaneous communication with eight GPS receivers.

SKILLS	<p>Programming: Python, C++, CUDA, Triton, SQL, R, Java, MATLAB, Assembly (8051), VHDL</p> <p>Tools: Deep Learning Frameworks (PyTorch, Keras, Tensorflow, vLLM), MLOps Tools (Kubernetes, Polyaxon, MLFlow), Other Tools (Docker, Flask, Django, Kafka, Spark), Agile (Gitlab, Atlassian Tools)</p> <p>Test Scores: TOEFL iBT: 108, GRE: 149/170/3.5</p>
PROJECTS	<p>Source codes with more details are available on: github.com/git-disl and github.com/fatih-ilhan</p> <p>Efficient Inference/Fine-tuning:</p> <ul style="list-style-type: none"> - Memory-efficient federated learning/fine-tuning: ScaleFL [C12, C13], RECAP [C18], Fed4LM [P1] - Adaptive inference: HiDEC [C11], EENet [C17] - KV cache compression for long-context inference [P2, P3] - Efficient ensemble learning: LLM-TOPLA [C15] <p>Robust Deep Learning Systems:</p> <ul style="list-style-type: none"> - Defense algorithms against adversarial attacks. [C8, C9, C10, C14] - LLM safety and alignment [C16, C19] <p>Time Series Prediction and Anomaly Detection:</p> <ul style="list-style-type: none"> - Online time series analysis [J3, C1, C2, C5] - Spatio-temporal event prediction [J1, C6] - Anomaly detection [J2, C3]
CONFERENCE PAPERS	<p>[C19] T. Huang, S. Hu, F. Ilhan, S. F. Tekin, and L. Liu, “Booster: Tackling Harmful Fine-tuning for Large Language Models via Attenuating Harmful Perturbation”, <i>International Conference on Learning Representations (ICLR)</i>, 2025. (oral)</p> <p>[C18] F. Ilhan, G. Su, S. F. Tekin, T. Huang, S. Hu, and L. Liu, “Resource-Efficient Transformer Pruning for Fine-tuning of Large Models”, <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i>, 2024.</p> <p>[C17] F. Ilhan, KH. Chow, S. Hu, T. Huang, S. F. Tekin, W. Wei, Y. Wu, M. Lee, R. Kompella, H. Latapie, G. Liu, L. Liu, “Adaptive Deep Neural Network Inference Optimization with EENet”, <i>IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)</i>, 2024.</p> <p>[C16] T. Huang, S. Hu, F. Ilhan, S. F. Tekin and L. Liu, “Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning”, <i>Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)</i>, 2024.</p> <p>[C15] S. F. Tekin, F. Ilhan, T. Huang, S. Hu and L. Liu, “LLM-TOPLA: Efficient LLM Ensemble by Maximising Diversity”, <i>ACL Conference on Empirical Methods in Natural Language Processing (EMNLP Findings)</i>, 2024.</p> <p>[C14] KH. Chow, Sihao Hu, Tiansheng Huang, Fatih Ilhan, Wenqi Wei, and Ling Liu, “Diversity-driven Privacy Protection Masks Against Unauthorized Face Recognition”, <i>Privacy Enhancing Technologies Symposium (PETS)</i>, 2024</p> <p>[C13] F. Ilhan, G. Su, Q. Wang and L. Liu, “Scalable Federated Learning with System Heterogeneity”, <i>IEEE International Conference on Distributed Computing Systems (ICDCS)</i>, 2023. (demo)</p> <p>[C12] F. Ilhan, G. Su and L. Liu, “ScaleFL: Resource-Adaptive Federated Learning with Heterogeneous Clients”, <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i>, 2023.</p> <p>[C11] F. Ilhan, S. F. Tekin, S. Hu, T. Huang, KH Chow, L. Liu, “Hierarchical Deep Neural Network Inference for Device-Edge-Cloud Systems”, <i>ACM International World Wide Web Conference (WWW)</i>, 2023. (poster)</p> <p>[C10] T. Huang, S. Hu, KH. Chow, F. Ilhan, S. F. Tekin and L. Liu, “Lockdown: Backdoor Defense for Federated Learning with Isolated Subspace Training”, <i>Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)</i>, 2023.</p> <p>[C9] KH. Chow, L. Liu, W. Wei, F. Ilhan and Y. Wu, “STDLens: Securing Federated Learning Against Model Hijacking Attacks”, <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i>, 2023.</p> <p>[C8] W. Wei, L. Liu, KH. Chow, F. Ilhan and Y. Wu, “Model Cloaking against Gradient Leakage”, <i>IEEE International Conference on Data Mining (ICDM)</i>, 2023.</p>

	<p>[C7] S. Hu, T. Huang, F. Ilhan, S. F. Tekin, L. Liu, “Large Language Model-Powered Smart Contract Vulnerability Detection: New Perspectives”, <i>IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (IEEE TPS-ISA)</i>, 2023.</p> <p>[C6] F. Ilhan, S. F. Tekin and B. Aksoy, “Spatio-Temporal Crime Prediction via Temporally Hierarchical Convolutional Neural Networks”, <i>28th IEEE Signal Processing and Communications Applications Conference</i>, 2020.</p> <p>[C5] F. Ilhan, N. M. Vural and S. S. Kozat, “LSTM-Based Online Learning with Extended Kalman Filter Based Training Algorithm”, <i>28th IEEE Signal Processing and Communications Applications Conference</i>, 2020.</p> <p>[C4] F. Ilhan and E. Mumcuoglu, “Performance Analysis of Semi-Supervised Learning Methods under Different Missing Label Patterns”, <i>28th IEEE Signal Processing and Communications Applications Conference</i>, 2020.</p> <p>[C3] F. Ilhan, S. F. Yilmaz and S. S. Kozat, “A Two-Stage Multi-Class Classification Approach Based on Anomaly Detection”, <i>28th IEEE Signal Processing and Communications Applications Conference</i>, 2020. (poster)</p> <p>[C2] N. M. Vural, B. Altas, F. Ilhan and S. S. Kozat, “Shortest Path Learning in Non-Stationary Environments via Online Convex Optimization”, <i>28th IEEE Signal Processing and Communications Applications Conference</i>, 2020.</p> <p>[C1] N. M. Vural, B. Altas, F. Ilhan and S. S. Kozat, “Online Shortest Path Learning via Convex Optimization”, <i>28th IEEE Signal Processing and Communications Applications Conference</i>, 2020.</p>
JOURNAL PAPERS	<p>[J3] F. Ilhan, O. Karaahmetoglu, I. Balaban and S. S. Kozat, “Markovian RNN: An Adaptive Time Series Prediction Network with HMM-based Switching for Nonstationary Environments”, <i>IEEE Transactions on Neural Networks and Learning Systems</i>, 2021.</p> <p>[J2] N. M. Vural, F. Ilhan, S. F. Yilmaz, S. Ergüt and S. S. Kozat, “Achieving Online Regression Performance of LSTMs with Simple RNNs”, <i>IEEE Transactions on Neural Networks and Learning Systems</i>, 2021.</p> <p>[J1] F. Ilhan and S. S. Kozat, “Modeling of Spatio-Temporal Hawkes Processes with Randomized Kernels”, <i>IEEE Transactions on Signal Processing</i>, 2020.</p>
PREPRINTS	<p>[P3] F. Ilhan, S. F. Tekin, S. Hu, T. Huang and L. Liu, “Neural Cache Compression for Memory-Efficient Inference with Large Vision-Language Models”, <i>in progress</i>, 2025.</p> <p>[P2] F. Ilhan, G. Su and L. Liu, “Memory-Efficient Decoding with KV Cache Compression for Long-Context LLMs”, <i>in progress</i>, 2025.</p> <p>[P1] F. Ilhan, S. F. Tekin, S. Hu, T. Huang and L. Liu, “Fed4LM: Efficient Federated Fine-tuning under Data and Resource Heterogeneity with a Mixture of Masked Adapters”, <i>in progress</i>, 2025.</p>
PATENTS	<p>[T2] F. Ilhan, G. Su, “Memory-Efficient Decoding with KV Cache Compression for Large Language Models”, (filed), 2025.</p> <p>[T1] F. Ilhan, G. Su, “Computation-Efficient Federated Learning System for Resource Heterogeneity”, P20240403701, 2023.</p>
AWARDS AND HONORS	<ul style="list-style-type: none"> - Online Head TA of the Year Award by Georgia Tech (2025). - Outstanding Head TA Award in OMS CS program by Georgia Tech (2024). - Full Scholarship from the Scientific and Technological Research Council of Türkiye for M.Sc. studies. - Full Scholarship from Bilkent University during B.Sc. and M.Sc. Studies. - 80th among 0.2M university graduates in ALES (Turkish National GRE). - JASSO Scholarship for Exchange Program at Nagoya University. - Bilkent University High Honor Student during B.Sc. Studies. - 191st among 2M high school graduates in University Entrance Examination.