Fatih İlhan

fatih-ilhan.github.io

Research Interests

Full-Stack Efficiency Optimizations for Inference/Finetuning of Large Generative AI Models, Distributed Systems, Heterogeneous Computing, Large Language Models, Multimodal Learning, Multi-agent AI Systems

Work Experience

Apple **Seattle, WA**

Machine Learning Researcher, AI/ML Dec 2025 – present

- On-device machine learning efficiency optimizations.
- Mobile-cloud hybrid computing for scalable and efficient AI applications.

IBM Research **Yorktown Heights, NY**

Research Scientist Intern May-Aug 2022/23/24/25

- Implemented codegen module for PyTorch eager mode integration through C++ front-end API extension for IBM's AI accelerator, Spyre.
- 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, led to a publication at CVPR24.
- Worked on computation-efficient federated learning under heterogeneous settings with on-premise deployments, filed a patent and 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 system, within a team of three engineers. Analyzed retail data from customer businesses to provide procurement and logistics insights, reduced consumption forecast errors by 40%.

Academic Experience

Georgia Institute of Technology **Atlanta, GA, USA**

Ph.D. in Computer Science, CGPA: 3.84/4.00, Advisor: Prof. Ling Liu Aug 2021 – Dec 2025

- Thesis Topic: Resource-Adaptive Efficiency Optimizations for Large Vision and Language Models
- Published 20 papers (7 as first author) in top venues such as CVPR, NeurIPS, ICLR, EMNLP.
- Georgia Tech Head TA of the Year (2025).

Bilkent University **Ankara, Türkiye**

M.Sc. in EEE, CGPA: 3.58/4.00, Advisor: 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 PAMI, TNNLS and 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
- Exchange student at Nagoya University, Japan (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

Conference Papers

[C24] **F. Ilhan**, G. Liu, R, Kompella, S. F. Tekin, T. Huang, Z. Yahn, Y. Xu, and L. Liu, "Attention-aware Inference Optimizations for Large Vision-Language Models with Memory-efficient Decoding", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[C23] Y. Xu, G. Liu, R. Kompella, T. Huang, S. Hu, **F. Ilhan**, S. F. Tekin, Z. Yahn and L. Liu, "A Multi-Agent Perception-Action Alliance for Efficient Long Video Reasoning", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[C22] S. F. Tekin, **F. Ilhan**, S. Hu, T. Huang, Y. Xu, Z. Yahn, and L. Liu, "H^3 Fusion : Helpful, Harmless, Honest Fusion of Pretrained-LLMs", *European Chapter of the Association for Computational Linguistics (EACL)*, 2026.

[C21] **F. Ilhan**, S. F. Tekin, T. Huang, G. Liu, R. Kompella, G. Eisenhauer, Y. C. Lin, C. Pu, and L. Liu, "FedHFT: Efficient Federated Finetuning with Heterogeneous Edge Clients", *IEEE International Conference on Cognitive Machine Intelligence (IEEE CogMI)*, 2025.

[C20] Z. Yahn, S. F. Tekin, **F. Ilhan**, S. Hu, T. Huang, Y. Xu, M. Loper, and L. Liu, " Adversarial Attention Perturbations for Large Object Detection Transformers", *International Conference on Computer Vision (ICCV)*, 2025.

[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", *International Conference on Learning Representations (ICLR)*, 2025. *(oral)*

[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", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[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", *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[C16] T. Huang, S. Hu, **F. Ilhan**, S. F. Tekin and L. Liu, "Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning", *Thirty-eighth Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[C15] S. F. Tekin, **F. Ilhan**, T. Huang, S. Hu and L. Liu, "LLM-TOPLA: Efficient LLM Ensemble by Maximising Diversity", *ACL Conference on Empirical Methods in Natural Language Processing (EMNLP Findings)*, 2024.

[C14] KH. Chow, Sihao Hu, Tiansheng Huang, **Fatih Ilhan**, Wenqi Wei, and Ling Liu, "Diversity-driven Privacy Protection Masks Against Unauthorized Face Recognition", *Privacy Enhancing Technologies Symposium (PETS)*, 2024

[C13] **F. Ilhan**, G. Su, Q. Wang and L. Liu, "Scalable Federated Learning with System Heterogeneity", *IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2023. *(demo)*

[C12] **F. Ilhan**, G. Su and L. Liu, "ScaleFL: Resource-Adaptive Federated Learning with Heterogeneous Clients", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[C11] **F. Ilhan**, S. F. Tekin, S. Hu, T. Huang, KH Chow, L. Liu, "Hierarchical Deep Neural Network Inference for Device-Edge-Cloud Systems", *ACM International World Wide Web Conference (WWW)*, 2023. *(poster)*

[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", *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[C9] KH. Chow, L. Liu, W. Wei, **F. Ilhan** and Y. Wu, "STDLens: Securing Federated Learning Against Model Hijacking Attacks", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[C8] W. Wei, L. Liu, KH. Chow, **F. Ilhan** and Y. Wu, "Model Cloaking against Gradient Leakage", *IEEE International Conference on Data Mining (ICDM)*, 2023.

[C7] S. Hu, T. Huang, **F. Ilhan**, S. F. Tekin, L. Liu, "Large Language Model-Powered Smart Contract Vulnerability Detection: New Perspectives", *IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (IEEE TPS-ISA)*, 2023.

[C6] **F. Ilhan**, S. F. Tekin and B. Aksoy, "Spatio-Temporal Crime Prediction via Temporally Hierarchical Convolutional Neural Networks", *28th IEEE Signal Processing and Communications Applications Conference*, 2020.

[C5] **F. Ilhan**, N. M. Vural and S. S. Kozat, "LSTM-Based Online Learning with Extended Kalman Filter Based Training Algorithm", *28th IEEE Signal Processing and Communications Applications Conference*, 2020.

[C4] **F. Ilhan** and E. Mumcuoglu, "Performance Analysis of Semi-Supervised Learning Methods under Different Missing Label Patterns", *28th IEEE Signal Processing and Communications Applications Conference*, 2020.

[C3] **F. Ilhan**, S. F. Yilmaz and S. S. Kozat, "A Two-Stage Multi-Class Classification Approach Based on Anomaly Detection", *28th IEEE Signal Processing and Communications Applications Conference*, 2020. *(poster)*

[C2] N. M. Vural, B. Altas, **F. Ilhan** and S. S. Kozat, "Shortest Path Learning in Non-Stationary Environments via Online Convex Optimization", *28th IEEE Signal Processing and Communications Applications Conference*, 2020.

[C1] N. M. Vural, B. Altas, **F. Ilhan** and S. S. Kozat, "Online Shortest Path Learning via Convex Optimization", *28th IEEE Signal Processing and Communications Applications Conference*, 2020.

Journal Papers

[J5] S. Hu, T. Huang, **F. Ilhan**, S. F. Tekin, G. Eisenhauer, M. Looper, and L. Liu, "Matching Accounts on Blockchain via Pseudo Fine-tuning of Language Models", *ACM Transactions on Intelligent Systems and Technology (ACM TIST)*, 2026.

[J4] S. F. Tekin, **F. Ilhan**, T. Huang, S. Hu, M. Loper, and L. Liu, " Robust Few-Shot Ensemble Learning with Focal Diversity-Based Pruning", *ACM Transactions on Intelligent Systems and Technology (ACM TIST)*, 2025.

[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", *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[J2] N. M. Vural, **F. Ilhan**, S. F. Yilmaz, S. Ergüt and S. S. Kozat, "Achieving Online Regression Performance of LSTMs with Simple RNNs", *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[J1] **F. Ilhan** and S. S. Kozat, "Modeling of Spatio-Temporal Hawkes Processes with Randomized Kernels", *IEEE Transactions on Signal Processing*, 2020.

Preprints

[P6] **F. Ilhan**, G. Su, S. F. Tekin, T. Huang, S. Hu, Z. Yahn, Y. Xu and L. Liu, "AttentionZip: Memory-Efficient Decoding for Fast LLM Inference in Long-Context Tasks", *under review*, 2026.

[P5] S. Hu, T. Huang, **F. Ilhan**, S. F. Tekin and L. Liu, "A Survey on Large Language Model-Based Game Agents", *under review*, 2025.

[P4] S. F. Tekin, **F. Ilhan**, T. Huang, S. Hu, Z. Yahn and L. Liu, " Vision Verification Enhanced Fusion of VLMs for Efficient Visual Reasoning", *in progress*, 2026.

[P3] S. F. Tekin, **F. Ilhan**, G. Liu, R. Kompella, and L. Liu, "Dynamic Optimizations of LLM Ensembles with Two-Stage Reinforcement Learning Agents", *in progress*, 2026.

[P2] Z. Yahn, **F. Ilhan**, T. Huang, S. F. Tekin, S. Hu, Y. Xu, M. Loper, and L. Liu, " Personalized Face Privacy Protection from a Single Image", *under review*, 2026.

[P1] Y. Xu, G. Liu, R. Kompella, S. Hu, **F. Ilhan**, S. F. Tekin, Z. Yahn and L. Liu, "A Neurosymbolic Agent System for Compositional Visual Reasoning", *in progress*, 2026.

Patents

[T3] **F. Ilhan**, G. Su, “Memory-Efficient Decoding with KV Cache Compression for Large Language Models”, *(filed)*, 2025.

[T2] **F. Ilhan**, G. Su, “Resource-Efficient and CPU-Assisted Pruning for Large Language Model Finetuning”, *(filed)*, 2024.

[T1] **F. Ilhan**, G. Su, “Computation-Efficient Federated Learning System for Resource Heterogeneity”, P20240403701, 2023.

Skills

Programming: Python, C++, CUDA, Triton, SQL, R, Java, Matlab, Assembly (8051), VHDL
Tools: Deep Learning Frameworks (PyTorch, Keras, Tensorflow, vLLM), MLOps Tools (Kubernetes, Polyaxon, MLFlow), Other Tools (Docker, Flask, Django, Kafka, Spark), Agile (Gitlab, Atlassian Tools)
Test Scores: TOEFL iBT: 108, GRE: 149/170/3.5

Awards and Honors

- Online Head TA of the Year Award by Georgia Tech (2025).
- Outstanding Head TA Award in OMSCS 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.