

Work Experience

Amazon Web Services (AWS)

Mar 2024 – Present

Applied Scientist II

San Diego, CA

- Leading Research & Development initiatives and implementing Model fine-tuning, parameter-efficient fine-tuning, continued pre-training, Reinforcement Learning (e.g. PPO, GRPO) solutions across various model families (e.g. Llama, Mistral, Qwen, Amazon Nova, Anthropic Claude, Titan Text Premier or custom etc.) for enterprise clients (e.g. in LLM Information Retrieval, Low Resource Language LLMs and LLM Code Assistants, outperforming other industry leaders).
- Architected and optimized distributed training pipelines scaling over 128 GPUs, over 128,000 context lengths (e.g. Sequence Parallelism, Tensor Parallelism, Pipeline Parallelism, ZeRO stage Optimizations etc.) and over 1 Trillion tokens streaming datasets.
- Optimized multi-threaded dataset preprocessing pipelines scaling to over 1 Trillion token datasets, including optimizing for full utilization of context length (e.g. packing)
- Lead contributor to Research Initiatives and publications.
- Collaborating with Anthropic for improving Claude for offerings to enterprise customers.
- Key contributor in improving the performance of internal AWS models hosted on AWS Bedrock.

Center for Voice Intelligence and Security

Feb 2023 – Feb 2024

Research Associate

Pittsburgh, PA

- Research and development activities related to Representation Learning and Explainable/Interpretable AI.
- Proposed and implemented the Multi-Head Gaussian Adaptive Attention mechanism, improving performance across multiple modalities, fine-tuning Llama 2 for text classification, WavLM for speech emotion recognition and BEiT for image classification [[Paper](#)] [[Code](#)]

NASA Ames Research Center

Aug 2022 – Dec 2022

Research Associate – Simulation & Analysis of Airport Surface Operations

Mountain View, CA

- Led the research in a team of 7, and applied Optimization and Reinforcement Learning to aircraft traffic control
- Reduced aircraft taxiway waiting time and increased throughput by 2x (worst-case) to 40x (best-case)

Amazon – Alexa Perceptual Technologies

May 2022 – Aug 2022

Applied Scientist II Intern

Boston, MA

- Proposed and developed a novel method using Multi-Head Self-Attention for efficient layer utilization of large self-supervised models for speech encoding in the downstream task of Speech Emotion Recognition (SER)
- Proposed, developed & evaluated novel methods for disentanglement of linguistic and paralinguistic speech representations and for the task of SER
- Outperformed previous approaches in the literature by +4% in accuracy [[Paper](#)]

Imperial College London

Jul 2020 – Jul 2021

Research Associate

London, UK

- Proposed a new learning algorithm for biologically-plausible Artificial Neural Networks treated as Complex Systems composed of Small-World Modular Networks.
- Performed computational experiments and simulations for Computational Neuroscience. [[Paper](#)] [[Code](#)]

ImpacTech – Big Data & AI startup

Apr 2019 – Feb 2020

Machine Learning Engineer

London, UK

- Proposed, developed and evaluated a novel Automatic Speech Recognition (ASR) system architecture working with data on the scale of Petabytes.
- Contributed to open-source OpenSeq2Seq project [[link](#)]

- Outperformed Word Error Rate of state-of-the-art ASR by a margin of 20% for AWS and by a margin of 2% for Microsoft Azure.

National Guard of Cyprus – Infantry

Jul 2014 – Jul 2016

Mortars Specialist Nicosia, Cyprus

- Collaborated and reported to the United Nations and worked in the Computing sector at the General Headquarters.
- Improved my leadership skills and self-discipline

Higher Education

Carnegie Mellon University, United States

Aug 2021 – Dec 2022

Master of Science in Computer Science and Engineering (MSIT-IS)

GPA: 3.6 / 4.0

Focus areas: Machine Learning, Natural Language Processing, Optimization, Privacy in Machine Learning

Imperial College London, United Kingdom

Oct 2016 – Jun 2020

Master of Engineering (incorporating the Bachelor) in Electronic and Information Engineering (EE-CS)

GPA: First Class Honours (equivalent to 4.0 / 4.0)

Focus areas: Machine Learning, Optimization, Computer Vision, Speech Processing

Selected Publications

[1] JEPA as a Neural Tokenizer: Learning Robust Speech Representations with Density Adaptive Attention

Georgios Ioannides, Christos Constantinou, Aman Chadha, Aaron Elkins, Linsey Pang, Ravid Shwartz-Ziv, Yann LeCun - NeurIPS 2025 (UniReps workshop) – arXiv:2512.07168v1, 2025

[2] Gaussian Adaptive Attention is All You Need: Robust Contextual Representations Across Multiple Modalities

Georgios Ioannides, Aman Chadha, Aaron Elkins - arXiv preprint arXiv:2401.11143, 2024

[3] *Towards End-To-End Disentanglement of Linguistic and Paralinguistic Contributions to Speech Emotion Recognition*

Georgios Ioannides, Michael Owen, Andrew Fletcher, Viktor Rozgic, & Chao Wang, Interspeech 2023.

[4] Compressed Models for Co-reference Resolution: Enhancing Efficiency with Debaised Word Embeddings, **Georgios**

Ioannides, Aishwarya Jadhav, Aditi Sharma, Alan W. Black, 03 July 2023, Scientific Reports, Springer Nature, 2023

Awards

Carnegie Mellon University scholarship - Awarded upon admission to the MS program for academic excellence **2021**

Governor's MEng Prize - Outstanding student in final year of the MEng program **2020**

Dean's List - Top 10% of the graduating class **2020**

Skills

Coding: Python, C++, C, Bash, MATLAB, Unit and Regression Testing.

Frameworks & Tools: PyTorch, TensorFlow, MXNet, scikit-learn, networkx, Statsmodels, NumPy, AWS, EC2, S3, Git, Hugging-Face, GPU/CUDA, DeepSpeed, Neuron, SageMaker HyperPod, AWS Bedrock, Slurm, FSDP

Computational Methods: Supervised Learning, Unsupervised Learning, Reinforcement Learning, Deep Learning, Time-Series Forecasting, Natural Language Processing (NLP), Computer Vision, Signal Processing, Speech Processing, Multimodal Machine Learning, Statistics, Large Language Models (LLM).

Languages: English (native level), Greek (native level)