Gabriel Mongaras

gmongaras@smu.edu • https://www.linkedin.com/in/gmongaras
gabriel@mongaras.com • https://gabrielm.cc/
512 – 659 – 5405 • https://github.com/gmongaras
https://www.youtube.com/@gabrielmongaras

OBJECTIVE: Enthusiastic artificial intelligence engineering student seeking to do research and development in the machine learning field and become a leading contributor.

EDUCATION:

Southern Methodist University – Lyle School of — Dallas, TX
Masters/Bachelors of Science in Computer Science — Expected Grad Date: May 2025
Bachelors of Science in Statistical Science — GPA: 3.86
Bachelor of Science in Data Science
Bachelor of Arts in Mathematics

Austin Community College — Austin, TX
Associates of Science in Computer Programming — Grad Date: May 2021
Occupational Skills Award – Computer Programming — GPA: 3.9

RELEVANT COURSES: Graduate Artificial Intelligence, Graduate Machine Learning 2, Graduate Algorithm Engineering, Assembly Programming, Algorithms, Calculus I, II & III, Graduate OS and System Software, Digital Logic Design, Linear Algebra, Digital History, Discrete Computational Structures, Applied Statistics, Engineering Design, Math Modeling, Math of ML, Applied Machine Learning, Data Structures

SKILLS:

Coding: Python, C++, CUDA, C, JavaScript, SQL, PL/SQL, AWS, Linux, Arduino, ARM, Android SDK, Java, Django, Flask, HTML, CSS

AI: Neural Networks, Generative models, PyTorch, scikit-learn, Reinforcement Learning, NumPy, CNNs, Transformers, GANs, NEAT, Diffusion Models, Object Detection, Audio Processing, Huggingface, TensorFlow

Blockchain: Smart Contracts, Solidity, Remix IDE

EXPERIENCE:

Google, *Student Researcher,* Dallas, TX — October 2024-December 2024
- Diffusion models are slow during inference. I researched methods to improve diffusion model inference speed performance. Some tests can be found here: https://github.com/gmongaras/Token_Merging_Tests

Google, *Software Engineering Intern,* Seattle, WA — May 2024-August 2024
- On the Google labs team, I researched video editing using inversion techniques.
- Performed a literature review search on current SOTA video editing techniques.
- Implemented these techniques in JAX for future researches at Google to use.

Hotshot, *AI Engineer,* Virtual — April 2024-May 2024
- Helped make changes to the new model which improves upon the Act 1 model.

Amazon, *Applied science Intern,* Sunnyvale, CA — May 2023-August 2023
- On the Amazon Alexa ESP (Echo Spatial Perception) team, worked to improve the algorithm that detects which Alexa is closest to a user after saying the wake word based on audio signals coming from all devices in a household using deep learning techniques.
- Researched different methods to keep the model smaller, faster, and more accurate at the same time.
- Looked into different types of data that can be fed into the model to improve model accuracy.

Meta, *Intern,* Menlo Park, CA — May 2022-August 2022
- Created a working mobile app using the Android SDK for a project assigned by Meta University.
- Researched and created an AI model to generate random sentences from Gaussian noise for the app.
- Worked with team members to implement rules and strategies to deal with security on data and database applications.

Southern Methodist University, *Undergraduate Research Assistant,* Dallas, TX — Fall 2021-May 2024
- Molecules have various stable equilibrium positions. When changing from one of these states to another, the molecule goes through a transition state which is observed using MP3 (Moller-Plesset Perturbation Theory) which is an accurate computation, but also very expensive. However it can be approximated by THC (Tensor Hypercontraction) which reduces the computation time, but also reduces the accuracy. We correct the error with an MLP which is faster than doing the MP3 computation, but more accurate than the THC measures. Our results have shown to achieve 2 orders of magnitude better than THC in terms of the MP3 value.

ACTIVITIES: Artificial Intelligence Club, President
Cybersecurity Club, Member
Computer Science Club, Member
Commons Council, Member

AWARDS: Hunt Scholars
Rotunda Scholars
University Honor Role
Hyer Society Member
Hilltop Scholar
Discovery Scholar
Accelerated Pathways Masters Degree Program

ENGINEERING PROJECTS:

Senior Thesis Fall 2023/Spring 2024

- Worked on a method called Cottention for making transformers linear in time and memory. Linear transformers have the goal of making the memory usage from quadratic to linear, thus saving resources.
- Paper found here: https://arxiv.org/abs/2409.18747

Diffusion Models From Scratch Fall 2022/Spring 2023

- Coded a Diffusion Model from pure PyTorch that learns how to produce images given random noise from a Gaussian distribution.
- On top of the basic DDPM model, I improved the speed of image generation by converting the model to a DDIMs, which removes the Markov chain restriction of the basic DDPM model.
- Added Classifier-Free guidance to improve model FID score.
- Saved several pre-trained models that generate images with a minimum FID score of around 30
- https://github.com/gmongaras/Diffusion_models_from_scratch

MetaU Capstone Summer 2022

- Created an app that gave daily fortunes to users which can be shared with friends found on the app.
- Built a model using a Transformer WGAN to generate random fortunes from Gaussian noise.
- https://github.com/gmongaras/MetaU_Capstone

YOLOX From Scratch Spring 2022/Summer 2022

- Coded an AI from scratch that learns how to detect objects given an image by putting bounding boxes around objects in the image.
- To detect objects, the algorithm predicts three attributes: The location of a bounding box to put around an object, how confident the model is that there's an object in that bounding box, and what object is in that bounding box.
- The project can be found here: https://github.com/gmongaras/YOLOX_From_Scratch
- Additionally, I wrote an article series explaining all the parts to this algorithm: https://gmongaras.medium.com/list/yolox-explantation-1bff11aa9911

Visualizing Gradient Descent Summer/Fall 2021

- Using only NumPy in Python, a neural network with forward and backward methods classifies a given point (x1, x2) to a color of red or blue based on the training data
- The network is trained using gradient descent which I coded from scratch with basic NumPy operations
- The model created represents how other models in the real world learn as they use the same algorithm
- The project can be found here: https://github.com/gmongaras/Visualizing_Gradient_Descent_For_BCE_Loss

PAPERS/ARTICLES:

Diffusion Models — DDPMs, DDIMs, and Classifier Free Guidance

- Developed a method called "Cottention", a linear complexity attention algorithm that has similar accuracy to classic softmax attention while being faster and more memory efficient.
- Code found here: https://github.com/gmongaras/Cottention_Transformer
- Paper found here: https://arxiv.org/abs/2409.18747

Diffusion Models — DDPMs, DDIMs, and Classifier Free Guidance

- Wrote about the evolution of base Diffusion Models and how they work.
- This article has been published by Better Programming
- https://betterprogramming.pub/diffusion-models-ddpms-ddims-and-classifier-free-guidance-e07b297b2869

Coding An AI Girlfriend

- Explains how I coded a virtual AI girlfriend using an assortment of AI technologies
- https://medium.com/mlearning-ai/coding-a-virtual-ai-girlfriend-f951e648aa46

How Do Self-Attention Masks Work?

- How do masks in the self-attention function work? This article attempts to explain how they work.
- This article has been published by MLearning.ai
- https://medium.com/mlearning-ai/how-do-self-attention-masks-work-72ed9382510f

YOLOX Explanation Series:

- Explains how the YOLOX object detection algorithm works through 4 different articles
- These articles have been published by MLearning.ai
- https://gmongaras.medium.com/list/yolox-explantation-1bff11aa9911

Community Detection with Neural Networks:

- Explains how neural networks can be used to detect communities in a graph and how this algorithm performs against the Girvan Newman algorithm.
- https://medium.com/smucs/community-detection-with-neural-networks-2e6c79a28d0c