Last updated in Feb 2025

Connor Gag

📍 San Diego, CA ✉ gagconnor@gmail.com ☎ (507) 766-6008 in connorgag connorgag

Summary

Data Scientist with a strong background in machine learning, AI, and scalable data pipelines, currently pursuing a Master's in Computer Science (AI specialization) at UC San Diego. Proven track record of deploying AI models, optimizing MLOps data workflows using GCP, BigQuery, and Python, and driving predictive analytics in the healthcare industry. Passionate about leveraging AI for real-world impact, with experience in NLP, geospatial data, and Retrieval-Augmented Generation (RAG) systems.

Skills

- Languages: Python (TensorFlow, PyTorch, Pandas, NumPy, Scikit-learn, Anaconda, Jupyter, Selenium, Matplotlib, PySpark) • SQL • Java • R • REST APIs • Bash • Linux command line
- Technologies: Git • Google Cloud Platform (GCP) • Azure • Snowflake • Docker • PowerBI • Unity • RAG • PostgreSQL • OracleDB • Jira

Education

UC San Diego *Sept 2024 – present*
MS in Computer Science, AI specialization

- GPA: 3.76/4.0
- Relevant coursework: Natural Language Processing (Transformers), Deep Learning (RNN, LSTM, CNN), Probabilistic Reasoning, Bayesian Networks, Algorithms, Search and Optimization Techniques

Gustavus Adolphus College *Sept 2018 – May 2022*
BA in Computer Science, Minor in Math

- GPA: 3.96/4.0
- Relevant Coursework: Algorithms, Theory of Computation, Database Systems, Compiler Design, Machine Learning, Linear Algebra, Multivariable Calculus, Statistics

Experience

Data Scientist *Remote*
UnitedHealth Group *May 2022 – June 2024*

- Developed and deployed an AI model to predict housing instability among Medicaid members. Research and analysis included prioritizing recall to identify at-risk individuals and prevent homelessness.
- Led data extraction and cleaning initiatives using complex SQL queries to process medical claims for 8M+ Medicaid members, collaborating with clinical and technical teams to ensure data accuracy.
- Designed and implemented a program to collect unstructured geospatial (GIS) SDOH (social determinants of health) data for individuals within a specified radius, supporting data-driven insights.
- Parallelized geospatial data collection, accelerating location data aggregation for integration into ML models and predictive analytics.
- Created and tested a machine learning model to identify Medicaid members at risk of transitioning to long-term care, enhancing resource allocation.
- Reduced security vulnerabilities by 40% through remediation of over 200 issues in the team's GitHub repositories.
- Migrated 20 GitHub repositories to GitHub Enterprise Cloud, increasing the visibility of our team's repositories within the company.
- Built and automated data pipelines using Google Cloud Run and BigQuery, enabling scalable and efficient data workflows.
- Spearheaded the team's transition to Agile workflows, achieving measurable gains in project delivery speed and team collaboration.

- Researched and designed a Retrieval-Augmented Generation (RAG) system enabling users to query company documents and data, providing quick access to insights via various LLMs (OpenAI and Google's APIs).

Data Science Intern *Remote*
UnitedHealth Group *June 2021 – Aug 2021*

- Achieved 86% test coverage for a large-scale ML data pipeline using Pytest, increasing the speed at which team members can make improvements.
- Refactored PySpark codebase to enhance modularity and readability, reducing unexpected bugs in the production environment.
- Organized and presented testing procedures to the team to improve team practices.

Computer Science Teaching Assistant *St. Peter, MN*
Gustavus Adolphus College *Feb 2021 – June 2021*

- Taught students introductory programming concepts through examples and projects in Python.
- Assisted professor in managing lab sessions and projects in Introduction to Programming II.
- Graded assignments and provided individual support to students to troubleshoot and understand concepts.

Projects

Analyzing Changing Trends of Podcasts Over Time Using BERTopic *GitHub Repo*

- Applied BERTopic to extract and analyze key topics from podcast transcripts over time, optimizing for both structured and conversational audio.
- Processed and segmented over 3,000 YouTube transcripts, handling missing timestamps and reducing conversational noise for improved topic modeling.
- Enhanced topic labels using GPT-4o and conducted time-series analysis to track evolving trends in major podcasts.
- Achieved a classification rate of 78.56% for news podcasts and 37.98% for conversational podcasts, demonstrating the model's effectiveness in identifying key topics from noisy data.

Semantic Segmentation of Images Using Deep Learning *GitHub Repo*

- Developed multiple CNN architectures for semantic segmentation on the PASCAL VOC-2012 dataset, including a baseline FCN, custom architecture, ResNet-34 based model, and U-Net implementation.
- Implemented various optimization techniques including cosine annealing learning rate scheduling, data augmentation (random cropping, flipping), and weighted loss functions to address class imbalance.
- Achieved significant performance improvements with the pretrained ResNet-based model, reaching 83.3% pixel accuracy and 0.188 mean IoU, compared to baseline accuracy of 73.77% and IoU of 0.0665.
- Experimented with different neural network architectures and compared their performance, including a custom network with strategic dropout layers and a U-Net implementation with batch normalization.
- Conducted comprehensive analysis of model performance using pixel accuracy and Intersection-over-Union (IoU) metrics, with detailed visualization of results and loss curves.

Image Classification with Low-Level Neural Network *GitHub Repo*

- Implemented a neural network from scratch using NumPy to classify pictures of clothing without relying on high-level machine learning libraries like TensorFlow or PyTorch.
- Experimented with different activation functions, regularization techniques, and hyperparameter tuning to improve model performance, achieving a test accuracy of 88%.
- Analyzed the effects of L1 vs. L2 regularization and momentum-based optimization on convergence and accuracy.
- Visualized and preprocessed dataset by normalizing pixel values and splitting data into training, validation, and test sets.

Transformer Encoder and Decoder Optimization *GitHub Repo*

- Implemented and analyzed attention mechanisms in a transformer, optimizing encoder and decoder attention patterns.
- Evaluated different positional embedding strategies (learned, sinusoidal, and ALiBi), reducing perplexity

from 178.20 to 112.86.
- Improved classification accuracy from 33.6% to 85.86% and conducted architecture exploration to enhance model efficiency.

Predictive Health Assessment Model

GitHub Repo ↗

- Used Azure Machine Learning Studio to train and hyperparameter-tune a model to predict individuals' general health given survey data.
- Deployed as a scalable API, enabling quick and secure health assessments via internal endpoints.

Expectation Maximization on Movie Reviews

GitHub Repo ↗

- Classified individuals into 4 types of movie watchers to predict each person's future reviews.
- Accomplished this through applying 256 iterations of Expectation Maximization on a dataset of movie reviews.

N-Gram Log Likelihood

GitHub Repo ↗

- Built a program that computes the unigram and bigram log likelihood of a sequence of characters, tokenizing by word.
- Combined the unigram and bigram models to compute the log likelihood of a sentence.