NNMax: Nearest Neighbor Max for Attention Mechanisms in Transformer Architecture

Project Focus and Background

Building on our research from the 2024 Bowdoin Summer Research Fellowship, Professor Farias and I are continuing our work on *ConvNN: Convolutional Nearest Neighbor*, a new approach to deep learning that improves how computers recognize and classify images. As we finalize our findings and prepare to submit a paper to a deep learning conference, we have discovered a deeper connection between **Convolutional Neural Networks (CNNs)** - which are commonly used in image recognition and **Transformers**, the architecture behind artificial intelligence systems like ChatGPT.

This discovery has led us to an exciting research question: **How can we incorporate the nearest neighbor concept into the way Transformers process information?**

Our goal for Summer 2025 is to develop **NNMax**, a new method for improving how Transformers pay attention to information. Transformers rely on a process called **softmax**, which helps decide which words or images are most important in a given context. However, we believe we can improve this process by incorporating ideas from **nearest neighbor algorithms**—a simple but powerful method used in machine learning to find the most similar pieces of information. By combining these concepts, we hope to enhance AI models such as ChatGPT and other large-scale AI systems.

This research sits at the intersection of two major areas of artificial intelligence: **CNNs and Transformers**. CNNs are great at identifying patterns in images, while Transformers have revolutionized language understanding by determining which words or pieces of data should be prioritized. However, few studies have attempted to unify these two powerful methods. Our approach builds on **K-Nearest Neighbors (KNN)**—one of the most fundamental techniques in machine learning—to develop a new way for AI to process information.

At its core, our method replaces the traditional softmax function with **NNMax**, which selects the most relevant pieces of information based on their similarity rather than simply assigning probability scores. This raises key research questions:

- **How does NNMax compare to softmax in accuracy and efficiency?**
- **What advantages and challenges come with using NNMax in AI models?**
- **How can NNMax be applied to real-world AI applications, such as large language models (LLMs) and ChatGPT?**

To explore these questions, we will build on existing research, including the influential paper *Attention Is All You Need*, which introduced Transformers. Our study aims to offer fresh insights into how AI models can be improved.

Methodology and Timeline

Over 10 weeks, we will carefully develop and test NNMax following a structured plan:

- Weeks 1-3: Design and build the NNMax function. We will study previous research and implement our new approach.
- Weeks 4-5: Test NNMax to see how it performs compared to the standard softmax function. We will analyze its strengths and weaknesses.
- Weeks 6-8: Integrate NNMax into a custom-built Transformer model and evaluate its effectiveness.
- Weeks 9-10: Write a research paper summarizing our findings, with the goal of submitting it to a machine learning conference.

Preparation and Background

I have been preparing for this research since my sophomore year by studying deep learning techniques and applying them in real-world projects. Last summer, I helped develop *ConvNN*, a new deep learning method that showed promising results. This spring, I am refining *ConvNN* for submission to a research conference.

Through this experience, I have built a strong foundation in **computer science, mathematics, and artificial intelligence**, equipping me with the necessary skills to take on this research project.

Impact on My Academic and Career Goals

This research aligns with my long-term goal of **pursuing a Ph.D. in Computer Science**, specializing in artificial intelligence and machine learning. Conducting advanced research this summer will give me valuable experience in AI model development, testing, and academic writing. Additionally, presenting our findings at a conference will help me build connections in the research community and strengthen my graduate school applications.

By dedicating this summer to computer science research, I will be better prepared for **Ph.D. applications in Fall 2025**.