1 Introduction

Over the last three decades, researchers have proposed approximately 400 different activation functions [1], suggesting a vast landscape of possibilities for neural network optimization. Historically, models such as those based on the transformer architecture, introduced in initial transformer paper [8], predominantly utilized Rectified Linear Unit (ReLU). However, the landscape began shifting when other activation functions started being considered.

A pivotal moment in the evolution of activation functions in language models was marked by the introducction of the Gaussian Error Linear Unit (GELU)[2]. GELU has become the popular choice for language models and it's also the default activation function RoBERTa and GPT-Neo, implemented by Hugging Face which are the ones I will be using as my baseline. This function's popularity underscores its perceived utility over traditional functions like ReLU in specific contexts, particularly in models with parameters on the scale of hundreds of millions.

Despite these advancements, continuous innovation leads to alternatives like GeGLU, noted for its effectiveness [7], also used in last year's winner of the BabyLM challenge [6]. Yet, a significant research gap persists—a comprehensive comparison of multiple activation functions under consistent conditions is notably absent. This gap could be explained by findings from another paper, which suggests that the impact of activation functions diminishes as the model size increases, evident in models with over a billion parameters [4]. This also explains the initial move away from ReLU, since all the research on alternatives was done on models with the size of approximately 100 million parameters.

Given these insights, this research will explore the impact of various activation functions on smaller-scale language models with around 10 million parameters. The hypothesis posits that at smaller scales, the choice of activation function is crucial, potentially leading to significant performance variations.

Further, this research will delve into an area of adaptive activation functions. It has been shown that adaptive activation functions outperform static ones in text-to-text machine translation [5], but there seems to be a lack of further research into adaptive function in language models, likely due to an expected tradeoff between additional trainable parameters and impact on performance. Additionally, recent developments in KAN: Kolmogorov-Arnold Networks [3] suggest a shift towards using activation functions on edges instead of nodes, but due to its recency, it has yet to be tested on a language model. This research will also experiment with this concept and apply it to language modeling to assess its efficacy at smaller scales.

This paper will structure its discussion starting with a review of historical and current activation functions, followed by methodology, experimental setup, results, and conclusions. By addressing these facets, the study aims to illuminate how different activation functions can enhance or compromise the performance of scaled-down language models, ultimately contributing to the optimization of neural network design.

References

[1] V. Kunc and J. Kléma. Three decades of activations: A comprehensive survey of 400 activation functions for neural networks. *arXiv*, arXiv:2402.09092, 2024.

[2] M. Lee. Gelu activation function in deep learning: A comprehensive mathematical analysis and performance. *arXiv*, arXiv:2305.12073, Aug 2023.

[3] Z. Liu and Others. Kan: Kolmogorov-arnold networks. *arXiv*, arXiv:2404.19756, May 2024.

[4] I. Mirzadeh and Others. Relu strikes back: Exploiting activation sparsity in large language models. *arXiv*, arXiv:2310.04564, Oct 2023.

[5] A. Rajanand and P. Singh. Erfrelu: Adaptive activation function for deep neural network.

[6] D. Samuel, A. Kutuzov, L. Øvrelid, and E. Velldal. Trained on 100 million words and still in shape: Bert meets british national corpus. *arXiv*, arXiv:2303.09859, May 2023.

[7] N. Shazeer. Glu variants improve transformer. *arXiv*, arXiv:2002.05202, Feb 2020.

[8] A. Vaswani and Others. Attention is all you need. *arXiv*, arXiv:1706.03762v5, Dec 2017.