Adaptive Activation Functions

Does the choice of the activation function matter in small LMs?

Filip Ignijic
f.ignijic@student.tudelft.nl

Aral de Moor
Supervisor

Maliheh Izadi, Arie van Deursen
Profesors

TU Delft

1 Introduction

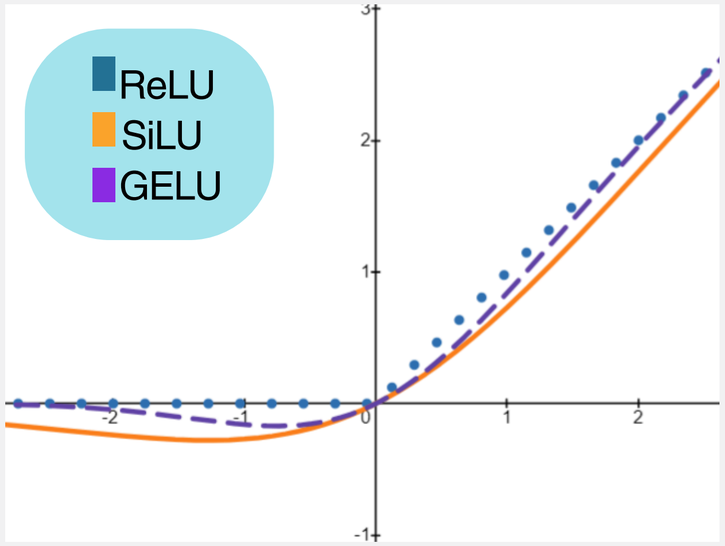
This study examines the role of activation functions in smaller-scale language models(10M).

Literature suggests that the impact of activation functions diminishes as the model size increases. We hypothesize that reducing the model size to 10 million parameters will enhance the influence of activation functions.

There is a noticeable gap in the literature regarding the use of adaptive activation functions in language models, presenting a novel research direction.

2 Activations

This research evaluates the performance of various activation functions. We begin by comparing baseline **GELU**, the default in many modern models, with its predecessor **ReLU** and **Adaptive GELU**, a novel parameterized version of GELU. Additionally, we compare ReLU with **PReLU**, which introduces parameterization to ReLU. Next, we examine the differences between **SiLU** and its parameterized counterpart, **Swish**. Finally, we compare all these activation functions against a model using the **KAN Network**.



3 The Experimental setup

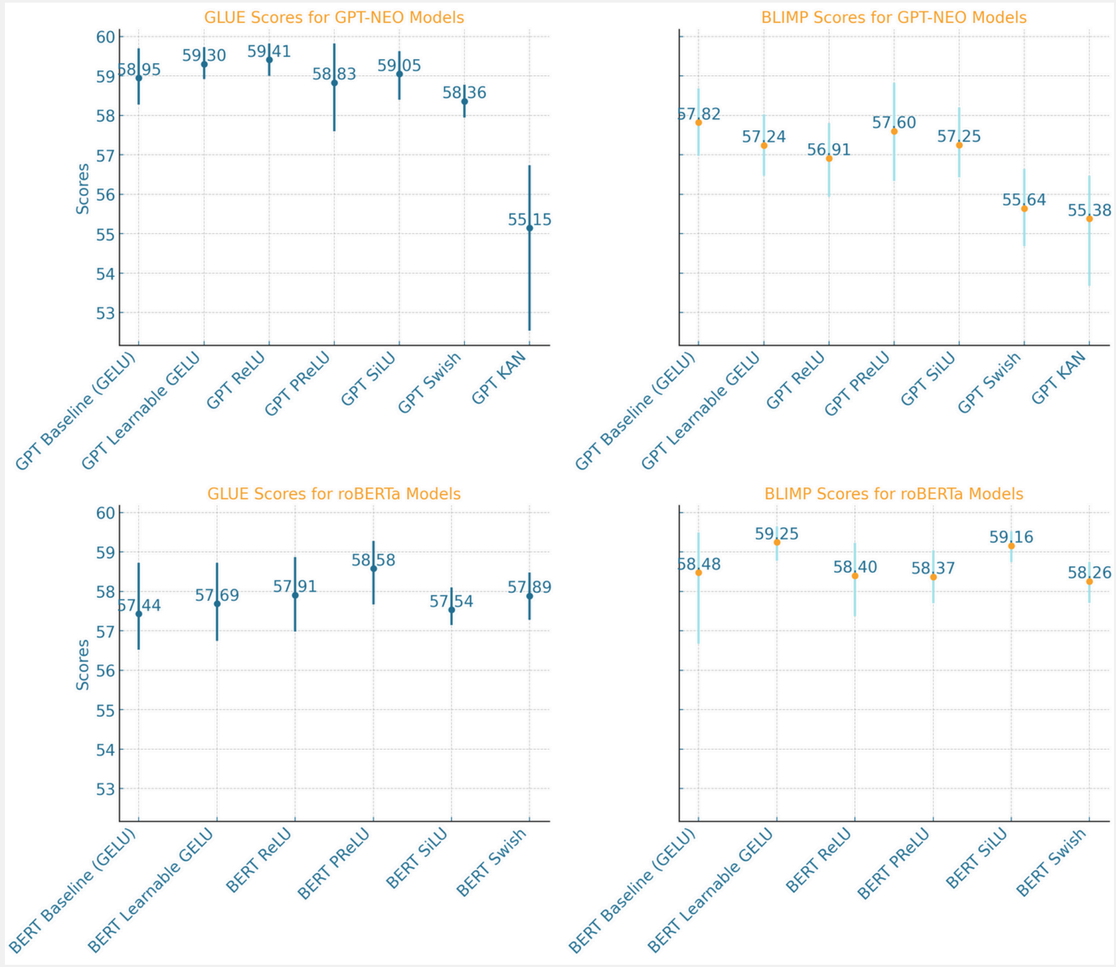
Pre-Train
On 6 different seeds

Evaluate
On (SUPER)GLUE and BLIMP

Statiscial significance testing
Bootstrap for 10k samples

4 Results

- Models using the KAN network are significantly worse than the rest.
- There is no significant difference between static functions and their adaptive counterpart
- No significant difference between the default GELU and its predecessor ReLU



5 Conclusions

This study highlights some potential flaws in previous works. Our findings suggest that the benefits of activation functions are statistically insignificant.
The reduced intermediate size and limited computation resources constrained the KAN model's performance, therefore it's premature to dismiss it. We encourage future research in that direction.