StoryWeaverGPT

Model Update, Code Explainations and Training

Group1

December 11, 2024

1. Model Update

2. Code Explainations

Model Update

- Changed Activation in FeedForward block from ReLU to GeLU.
- Changed Dataset from WritingPrompts to Shakesphere.

GeLU Activation

- GeLU is given as

$$GeLU(x) = \frac{1}{2}(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right))$$

- Where the erf is the error function, given as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

- And approximated as

$$\text{erf}\left(\frac{x}{\sqrt{2}}\right) \approx \tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right)$$

1

[1] Hendrycks, D., & Gimpel, K. (2016). Gaussian Error Linear Units (GELUs).

GEeU Activation

- With approximation implemented, the equation becomes

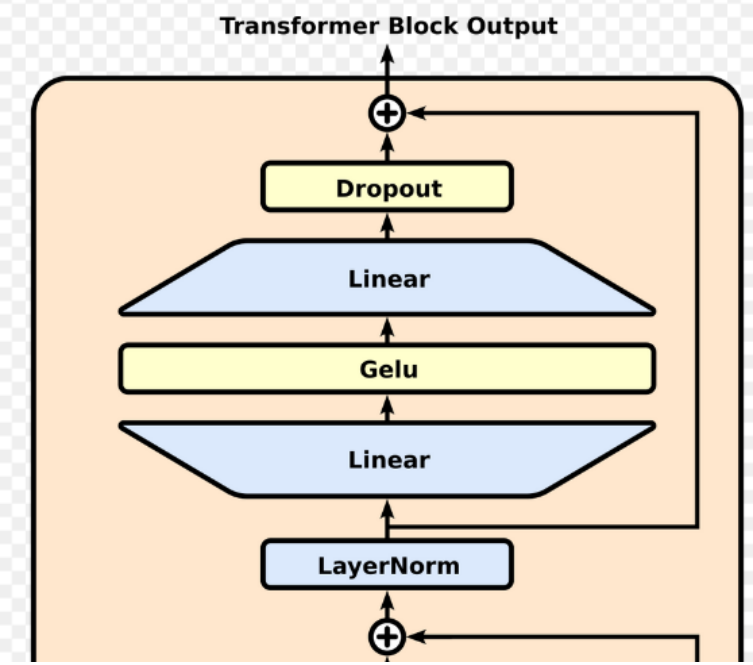
$$GeLU(x) = \frac{1}{2}(1 + tanh(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)))$$

- Replacing $\sqrt{\frac{2}{\pi}}$ with 0.7978845608, we get

$$GeLU(x) = \frac{1}{2}(1 + tanh(0.7978845608(x + 0.044715x^3)))$$

GeLU Activation: Intuition

- GeLU is much smoother, where ReLU has abrupt changes at 0.
- Inputs around zero are partially activated, where ReLU would be off.
- It is analytically differentiable, which may yield smoother gradiants and avoid vanishing gradiants.



Shakesphere Dataset

- The Shakesphere dataset is a collection of Shakesphere plays.
- While it is still large, it is much smaller than the WritingPrompts dataset.

Code Explainations