

StoryWeaverGPT

Model Update, Code Explanations and Training

Group1

December 12, 2024



1 Model Update

2 Code Explanations



Model Update

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



GeLU Activation

- GeLU is given as

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

- 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

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

1

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



- With approximation implemented, the equation becomes

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

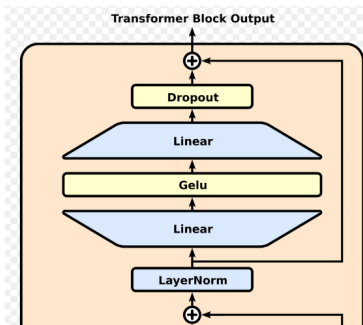
- Replacing $\sqrt{\frac{2}{\pi}}$ with 0.7978845608, we get

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



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 gradients and avoid vanishing gradients.



Shakespeare Dataset

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



Code Explanations



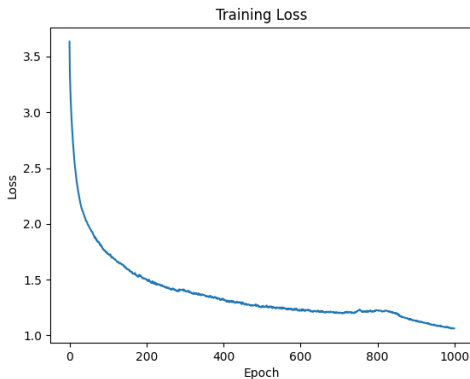
Model Recreation with torch.nn.Module

- Recreated the model with torch.nn.Module under assumption that our code lacks optimization.
- With same configuration, original model took 7 hours for 300 epoch, whereas torch.nn.Module took 50 minutes for 500 epoch.
- Before testing our actual model, we will test the torch.nn.Module model.



Model Training

- Trained on 1000 epoch, with learning rate 0.0001.
- takes approximately 2 hours to train on rtx 3080.



Remaining Tasks

- 1. Finish training on the original model.
- 2. Implement a probabilistic sampling method with temperature.
- 3. (If possible) Implement a repetition penalty.

