AML Proposal

mm20079

March 2024

1 Overview

We aim to investigate how replacing the standard dot product mutli-head attention in transformer architecture

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

[1]

We will use a baseline gaussian attention mecahnism

$$GCT(x) = \exp(\frac{-(x-\mu)^2}{2\sigma^2})$$

as described in [2]. And compare its effectiveness with a laplacian distribution of the form $\frac{1}{2b}\exp -(\frac{|x-u|}{b})$ where our learned parameters will be a diversity parameter b, and ϕ which is a learnable offset of our sample mean $\hat{\mu} = \frac{1}{n}\sum i = 1^N x_i$ giving the model the ability to approximate the population mean from the sample mean, $\mu_{pop} = \hat{\mu} + \phi$

We will use a pretained model (Llama 2) and train an attention module within the decoder, we aim to compare both the GAAM (Gaussain Adaptive Attention Mechanism) and the Laplacian Attention mechanism on a News classification Dataset.

1.1 Formally

Let X be the set of input news articles and Y be the set of corresponding categories. We aim to learn a model $M : X \to Y$ that maps each news article $x \in X$ to its correct category $y \in Y$ using a Laplacian Attention Mechanism. Our goal is to minimize the classification error defined as:

$$\min_f \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(M(x_i) \neq y_i)$$

where N is the total number of news articles in the dataset, x_i is the i-th news article, y_i is its correct category.

2 Theoretical Intuition

Recent research by Ioannides et al. (2024) has shown that the usage probabilistic attention functions can enhance the capabilities of transformers in a variety of domains. Specifically, Ioannides et al. (2024) have proposed a Gaussian adapative attention mechanism [GAAM] that has reached state of the art performance on emotion recognition in speech, text classification, and image classification. Thus, the usage of probabilitstic attention functions within transformers seems very promising, which is why we want to investigate it in more detail.

With regards to the paper by Ioannides et al., it caught our attention that the authors give no explicit reasons with regards to why a Gaussian distribution was chosen for the attention mechanism (instead of e.g. a Laplacian or a t-distribution). Instead, they claim that when GAAM is applied independently to multiple attention heads, then they can jointly learn any possible probability distribution. The goal of our project is to investigate whether this claim holds and if adapting the attention function to another type of distribution potentially leads to an additional improvements in the results. Concretely, we intend to mimic the overall transformer architecture proposed by Ioannides et al., but adapt the attention function to different distributions in order to learn how the type of distribution chosen for the attention function affects the accuracy of the model. Alternative distributions we currently consider exploring are a Laplacian distribution, a chi-squared distribution, and a t-distribution.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[2] G. Ioannides, A. Chadha, and A. Elkins, "Gaussian adaptive attention is all you need: Robust contextual representations across multiple modalities," 2024.