Finetuning Transformer Encoders for Classifying Movie Reviews

Moses Addai, Jakob Simmons, Emma Vejcik

Dataset: Kaggle IMDB Movie Reviews 2021

Features:

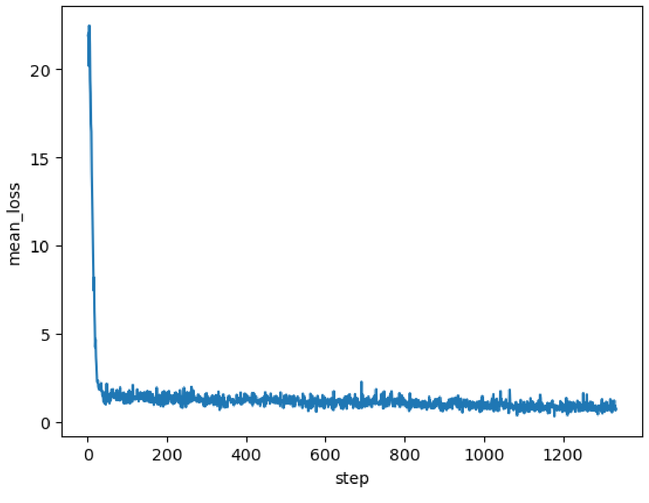
- **id**: <*discrete, int*> a key for identifying the review, discrete, unique
- **review**: <*str*> the full text of the review
- **rating**: <*discrete, int*> a rating between 1 and 10 for the movie
- **author**: <*str*> the author of the review
- **title**: <*str*> the title of the review

Target Feature: **rating**

Training on: review

A transformer model is used to create the encodings of the review texts. These encodings are then used as the feature to predict on.

Preliminary Results



The above loss graph shows the Gemini model(5T) scores a 35% accuracy when tested with 20 samples. After finetuning, this score increases to 87% for 100 samples.

*Multi-Layer Perceptron (MLP)

Methodology: Finetuning LLMs with MLP* Multiclass Classifiers

We start with a pre-trained language model, Google's 5T Gemini, that defines a nonlinear mapping $\boldsymbol{\phi(x;\theta)}$ where x are the input tokens (our data) and θ are the pre-trained parameters (that we do not have access to, found by Google). These pre-learned parameters θ encapsulate the language understanding from the training data. The functions ϕ can be interpreted as a set of basis functions learned from the large amounts of data that Google used to train the model from that will now be applied to our data inputs.

On top of these basis functions, we used a custom classification head, $h(\cdot;\varphi)$, implemented as an MLP with configurable hidden layers and hidden size).

The final model is:

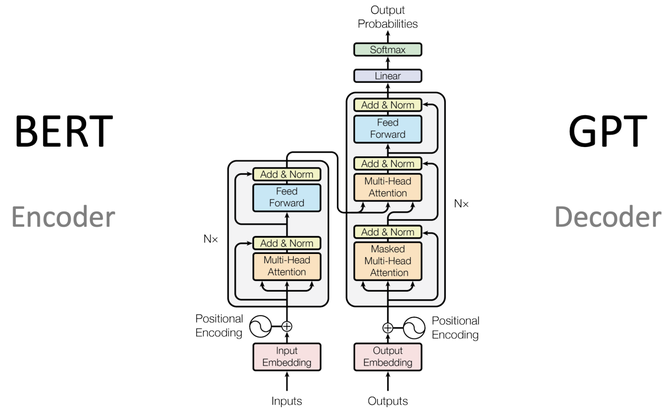
$$\boldsymbol{F(x;\theta,\varphi)=h(\phi(x;\theta);\varphi)},$$

which outputs K logits for K classes, in our case, $K = 10$.
We minimize the negative log-likelihood (aka the cross-entropy).

Our approach involves freezing a portion of the pre-trained encoder parameters while leaving others trainable. Mathematically, we impose the constraint

$$\boldsymbol{\theta_i \text{ is fixed for } i \in F \text{ and trainable for } i \in T},$$

where T and F are the sets of indices corresponding to trainable and frozen parameters. The fraction p_{tune} is determines the percentage of parameters frozen in T.



Hyperparameters		
Gemini	Tuned Hyperparameters	Fixed Hyperparameters
Model Size		770 M
% Finetuned	0, 0.1, 0.5, 1	
Custom MLP Head # of Layers	256, 512, 1024	
Custom MLP Head Hidden Layer Size	1,2,3	
Learning Rate		2e-5
# Training Epochs	5	
Classification Head Activation Function		ReLU

Planned Work

The remainder of the work that we will complete consists of determining the best hyperparameters for our models.

References

Gemini Team et. al. Gemini: A Family of HighlyCapable Multimodal Models. arXiv e-prints, pagearXiv:2312.11805, December 2023.Marge Simpson (2010). “Blue hair looks nice.”. In: Nature communications 1, p. 622.

Deshpande, D. (2021). IMDB Movie Reviews 2021, Version 1. Retrieved February 28th, 2025 from https://www.kaggle.com/datasets/darshan1504/imdb-movie-reviews-2021/data

Niklas Heidloff. Foundation models, transformers, bert and gpt, 02 2023.