Date: 05/12/2023

Graph Mining Presentation

Hate Speech Detection using LLMs

BITS Pilani

Pilani Campus

Ipshita Mahapatra - 2022H1030128P

Supervisor : Dr. Vinti Agarwal

Problem Statement

The objective of this task is to create an automatic classification system that predicts whether a given text contains caste/Immigration hate speech or not on Social media.

The dataset provided for this task contains texts(social media comments) in Tamil language.

Proposed Solution to Solve this problem-
Hate Speech Identification using LLMs

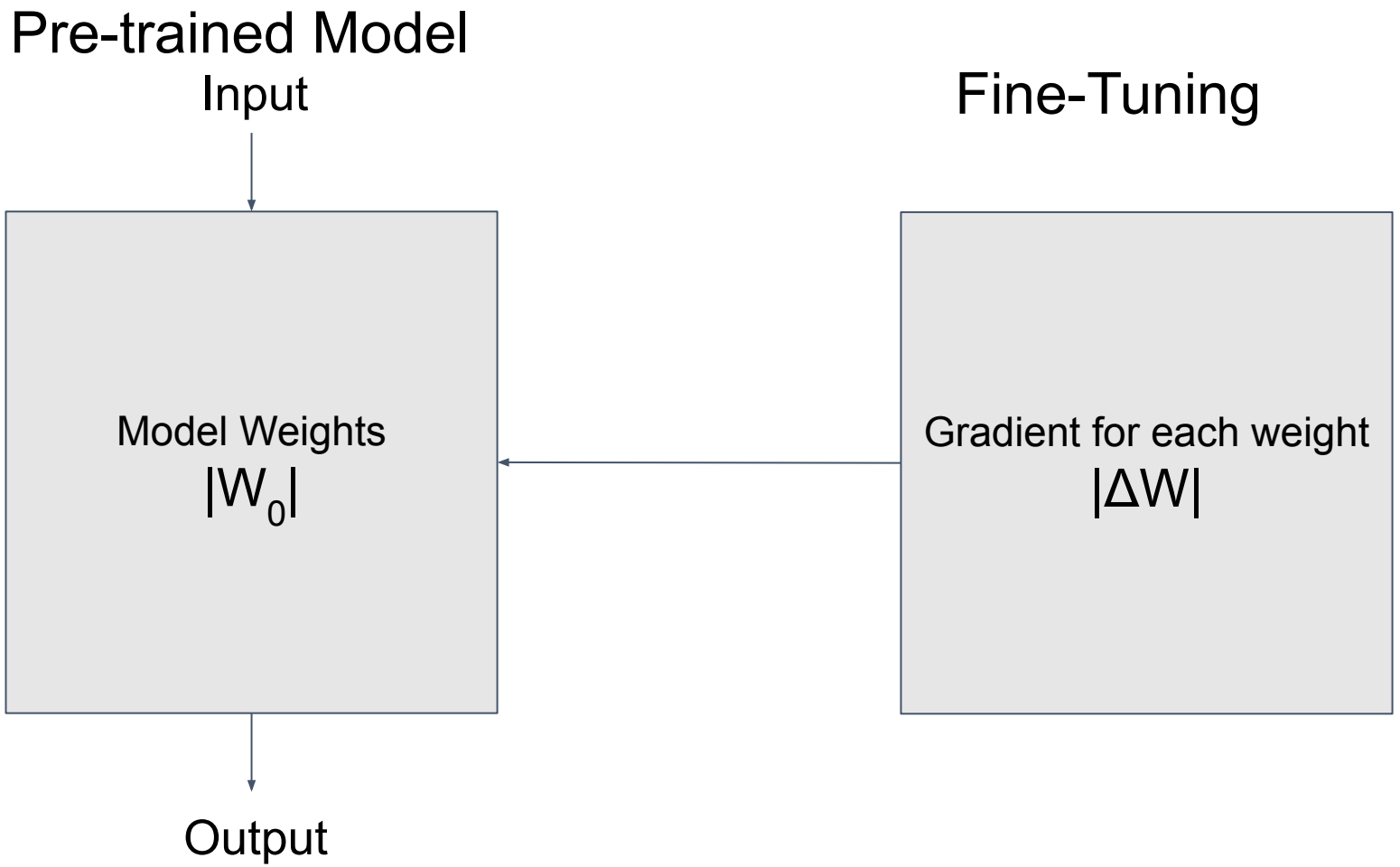
Fine Tuning LLMs using LoRA

LLMs - Large Language Models

- Advanced NLP Models
- Trained on Massive amounts of diverse text data
- Key Features -
  - Scale for Parameters
  - Pre-Training on Broad Data
  - Transfer Learning
  - Versatility

Examples - GPT, BERT

Fine Tuning LLMs



The issue with this model is that ΔW is as huge as W_0 which make it computationally expensive and heavy in terms of memory as well.

Fine Tuning using LoRA

Pre - trained Model = $P_{\Phi}(y|x)$

Fine Tuning using LoRA:

Initial weights of the pretrained Model = W_0
Updated weights -

$$W_0 + \Delta W = W_0 + BA$$

Where $W_0 \in R^{d \times k}$, $B \in R^{d \times r}$, $A \in R^{r \times k}$ and rank r = min(d, k)

For the Output $h = W_0 x$, our modified forward pass yields:

$h = W_0 x + \Delta W x = W_0 x + BAx$

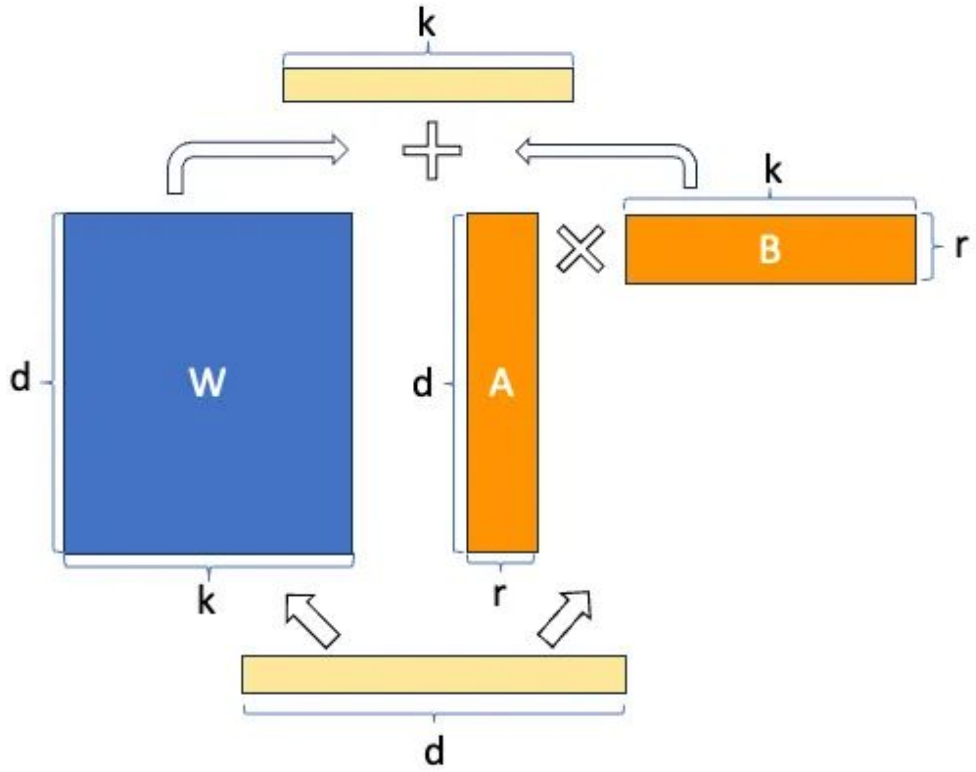


Image Reference - https://dataman-ai.medium.com/fine-tune-a-gpt-lora-e9b72ad4ad3

Tamil-LLaMa

Pre-training and Fine tuning LLaMa-2 over Tamil language corpus.

Pre-Training - Casual Language Modeling Method

CLM method suggests that the Model is trained to predict the next word provided it is given a sequence of previous words.

$$P(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} P(x_t | x_1, x_2, \ldots, x_{t-1})$$

Training infrastructure-

- Nvidia A100 GPU with 80GB of VRAM.
- The models were trained for 1 epoch on the entire dataset
- Microsoft Azure's Standard NC24adsA100v4 instance was used.
- Training time
  - 7B model - 48 hours
  - 13B model - 60 hours
- The hyper parameters for pre-training as mentioned in Table 1

Table 1: Pre-Training Hyperparameters

Configurations	7B	13B
Training Data	12GB	4GB
Epochs	1	1
Batch Size	64	64
Initial Learning Rate	2e-4	2e-4
Max Sequence Length	512	512
LoRA Rank	64	64
LoRA Alpha	128	128
LoRA Target Modules	QKVO, MLP	QKVO, MLP
Training Precision	FP16	FP16

Tamil-LLaMa (Contd…)

Fine tuning - Using LoRA

The Hyperparameters used to finetune LLaMa 2 with Tamil Language corpus are mentioned in Table 2

Hardware Requirements-
The same A100 GPU with 80GB of VRAM was utilized.

Table 2: Fine-tuning Hyperparameters

Configurations	7B	13B
Training Data	145k	145k
Epochs	2	1
Batch Size	64	64
Dropout Rate	0.1	0.1
Initial Learning Rate	2e-4	2e-4
Max Sequence Length	512	512
LoRA Rank	64	64
LoRA Alpha	128	128
LoRA Target Modules	QKVO, MLP	QKVO, MLP
Training Precision	FP16	FP16

Tamil-LLaMa Comparative Results

- In manual examinations, the Tamil LLaMA models outperformed gpt-3.5-turbo and achieved outstanding ratings in GPT-4 evaluations, indicating superior performance.
- It should be highlighted, however, that GPT-4 may automatically favour replies from its own model lineages, and certain areas, such as ethics, exhibit limits due to the lack of alignment efforts.
- Despite issues in literature and entertainment due to data restrictions, Tamil-LLaMA models provide a solid platform for future improvements and advancements in big language models for Tamil.

Table 3: GPT-4 rated performance scores for different models on Tamil instructions

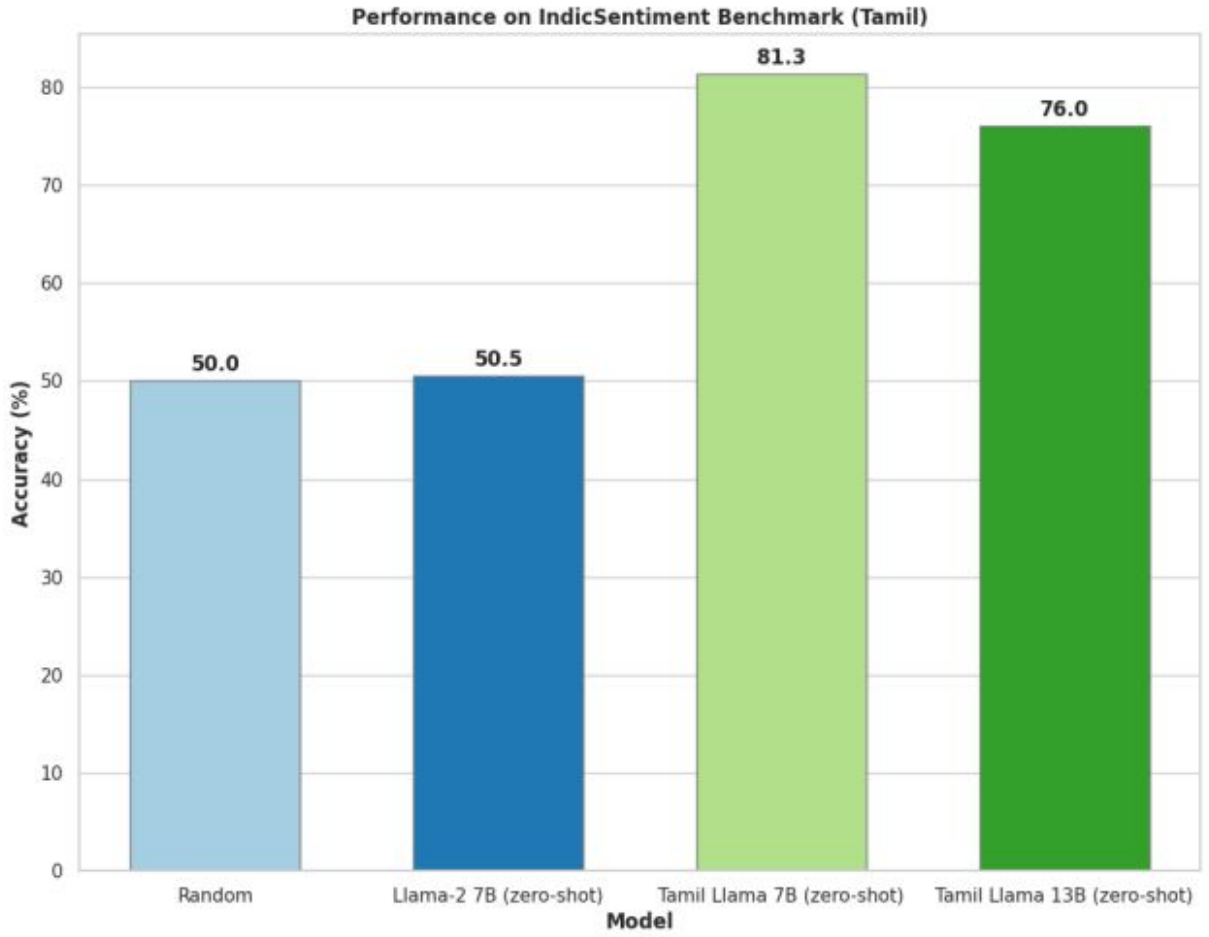
Task Type	**Tamil-LLaMA-7B**	**Tamil-LLaMA-13B**	***gpt-3.5-turbo***
Question Answering	**77.00**	75.33	54.33
Open-ended QA	84.47	**85.26**	58.68
Reasoning	47.50	**64.25**	63.50
Literature	45.50	40.00	**71.00**
Entertainment	43.33	50.00	**60.00**
Creative Writing	92.50	**95.62**	59.69
Translation	60.56	66.67	**92.78**
Coding	63.57	**76.07**	57.14
Ethics	23.75	**57.50**	40.00
Overall	63.83	**71.17**	61.33

Tamil-LLaMA results

Comparative analysis of the following models

- Random
- LLaMA 2(7b parameters)
- Tamil-LLaMA 7B
- Tamil-LLaMA 13B

Over the IndicSentiment Benchmark(Tamil) shows that Tamil-LLaMA 7B outperforms all the other models, followed very closely by Tamil-LLaMA 13B



Performance comparison on the IndicSentiment-7B dataset

References

[1] Hu, E. J. (2021, June 17). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv.org. https://arxiv.org/abs/2106.09685

[2] Balachandran, A. (2023, November 10). *Tamil-Llama: A New Tamil Language Model Based on Llama 2*. arXiv.org. https://arxiv.org/abs/2311.05845

Thank You