Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2024, 11(10):35-38**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Comparing Efficiency of Large and Small Language Models for Spam Text Detection

Rahul Kavi, Jeevan Anne

Independent Researcher, USA

Abstract Are Language Models efficient spam detectors? This study aims to test language models such as LLAMA2 and Phi2 as spam classification models. Language Models have shown incredible capability to understand human language, context and meaning. These have shown to carry coherent conversations with users across several domains. This study aims at checking whether SLMs (like LLMs) have capability that transfers to using them as spam detectors as language models already understand human language and context. LLAMA2 [1] is a popular Large Language Model (LLM) and Phi2 [2] is a popular Small Language Model (SLM). This study shows that both SLMs and LLMs are quite capable to be adapted to spam detection domain even though they weren't trained originally for this domain. This study show how efficient SLMs are compared to LLMs on common language classification tasks such as Spam Classification. We show that even with fewer parameters SLMs perform very well (with less than 1% difference in accuracy and precision).

Keywords Spam Detection, Small Language Models, Large Language Models, Machine Learning, Natural Language Processing.

1. Introduction

Large Language Models have taken over the world by storm. They have been used in text generation, text classification, sentiment analysis, named entity recognition, etc. They have become a buzz word, and they changed the way we interact with text. These models however have become very large and needs large GPU compute to train them. Running them on CPU takes much longer time than GPU (due to the nature of the operations underlying inside its architecture). Small Language Models have become a credible alternative that run with relatively fewer resources. However, training Small Language Models is a challenging task as it needs much cleaner datasets. SLMs are trained on specific domains and on smaller text corpus. However, they are designed to be efficient and compatible with resource constrained devices such as embedded systems, phones, etc. LLMs have been used to tackle spam detection problem. Researchers have focused on spam detection for several decades. More recently, SLMs have emerged as a credible and resourceful approach to tackle spam detection. SLMs are known to be able to handle and carry out coherent language conversations (such as Phi2, Phi3).

LLMs such as LLAMA2, Mistral [3] and LLAMA3 [4] are popular choices for several NLP tasks such as text classification, text generation, NER, etc. SLMs such as Phi2 and Phi3 are also popular for text classification, text generation and question answering. In this paper, we explore how LLMs and SLMs compare in terms of classification accuracy in spam detection problems. Specifically, we pick and popular spam dataset freely available on UCI repository [5].

Both LLMs and SLMs have vast knowledge of language understanding and generation. This can be used to identify commonly used keywords in spam text, context, etc. This knowledge is immensely helpful in identifying spam text from on-spam text. Language Models are inherently good at text classification. These models can be quantized or distilled to be run on low resource devices such as smartphones, etc. Quantization as

an approach to make Language Models run on low resource devices is a recent trend. This trend has picked up significantly once it was realized that language models can carry out conversations with humans or bots in a coherent manner.

This study was performed using 2 popular language models such as LLAMA2 and Phi2. These are well established models and quite popular in the research community to solve a variety of tasks in NLP. For this study, we take one LLM (LLAMA2) and one SLM (Phi2) and fine tune it on spam detection dataset. The finetuning was performed using PEFT(LORA) approaches. LORA is parameter efficient fine-tuning approach (PEFT) introduced by Microsoft. Since LLMs (and SLMs) have billions of trainable parameters, all of them cannot possibly be updated while transfer learning. Instead, new set of trainable parameters (low-rank matrices) are introduced inside transformer layers. These parameters are trained to adapt the model to newer tasks (instead of solving the original task the Language Model was trained on). LoRA is shown to improve model performance for other downstream tasks. This is a popular approach used to adapt the Language Model to new domains.

2. Related Work

LLMs such as LLAMA, LLAMA2 and LLAMA3 have shown excellent performance in language generation, language understanding, question answering, etc. These were trained on large text corpus using several hundreds of GPU hours on high end machines. There is immense potential in these models as they are designed to be able to carry out natural text generation on wide variety of tasks. Transformer architecture has been the one single most influential piece of work in Machine Learning. Application of this architecture has spanned Natural Language, Vision, etc.

SLMs have been a more recent phenomenon. BERT has laid the foundational work in this area. GPT like models have shown that large models have nearly human like reasoning, text generation capabilities [7]. Architectures like ELMO [8] and BERT [9] have paved way to decoder only models such as GPT, LLAMA, CLAUDE, Mistral. The popular models have distilled and quantized versions of the original full capability models (30 billion, 80 billion parameters). Microsoft's groundbreaking work on Phi2, Phi3[10] have shown that with quality data (textbook data), small models have the capability to comprehend human language, text understanding, classification, etc. This work has shown that larger number of parameters don't always mean higher efficiency on all text related tasks (such as generation, classification, summarization, etc.).

SLMs have shown remarkable capability compared to LLMs in specific tasks such as MMLU [10]. This means that SLMs can be leveraged to be run on smaller compute devices, low memory constraint environments. Quantizing these models has led to even more push towards running these incredible innovations on devices with smaller memory and compute capabilities.

LLMs have been used to solve several text classification tasks such as spam detection, rating measurements, toxic content detection. Language models once trained on large text data can learn to distinguish between the text that is relevant to the context and spam. Traditionally spam detection was done with the use of TF-IDF (features), then these features were passed onto classification models such as Naïve Bayes, SVM, Neural Networks (including BERT like architectures). Language Models use encoding approaches using approaches like sentencepiece[5], etc. Embeddings from these SLMs can be extracted to be passed to simple neural network like architectures. However, LLMs and SLMs are trained on large text corpus that is not specific to text classification. These models need to be adapted to text classification using approaches such as LoRA (PEFT) [11].

In our approach, we use LLAMA2 and Phi2 along with LoRA (PEFT) [11] to classify spam text from regular text. This study is primarily aimed towards non-prompt-based approaches to classifying text. We show that with this work SLMs are powerful alternative to LLMs when it comes to classifying spam text from regular text.

In this section, we describe our current approach to solving the spam text classification. Firstly, we take one of the most popular LLMs such as LLAMA2. The LLAMA2 tokenizer is loaded and encoded input of the training dataset is obtained. The encoded text is passed into the model and embedded is obtained. A simple classification model is trained on top of this, and logits are obtained. PEFT approach such as LoRA[11] is adopted to add a few trainable parameters across all layers of the LLAMA2 model. This helps with efficient fine tuning on the dataset. The approach is described in the following diagram.

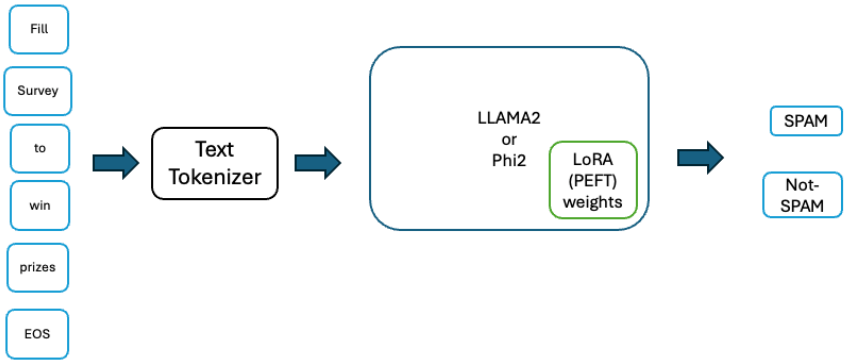

Figure 1: LLM and SLM Model Training/Inference with LoRA

Next, we take the popular SLM such as Phi2. The Phi2 tokenizer is used to extract tokenized text. The encoded text is passed into the model and embedding is obtained. Like previously described approach, PEFT is used to train a binary classification model to classify text from spam text. Due to the size of the dataset, model, we have fine-tuned the model for 3 epochs (as the loss curve flattened on the validation dataset). The dataset consists of 5572 examples. The dataset was split into 60% train, 20% test, 20% validation datasets.

The Phi2 LoRA model consists of 5 million trainable parameters (around 20% of 2.7 billion parameters of Phi2 model). The LLAMA2 LoRA model consists of 8 million trainable parameters (around 12% of the 7 billion parameter LLAMA model). The final prediction is obtained by computing an argmax of the logits from the final layer. The below table describes the data distribution employed for training, eval and testing datasets.

Table 1: Spam Text Detection Dataset Distribution

Data Split	Number of examples	Percentage of the dataset
Train	3343	60%
Eval	1115	20%
Test	1114	20%

3. Results & Discussion

In this section, we look at results from the training of Phi2, LLAMA2 model. Table 2 describes the results from LLAMA2-LoRA model results and Table 3 describes results from the Phi2-LoRA model.

Table 2: LLAMA2-LoRA model performance

Split	Accuracy	Precision	Recall	F1 Score
Train	0.9940	0.9939	0.9940	0.9940
Eval	0.9847	0.9846	0.9847	0.9846
Test	0.9910	0.9909	0.9910	0.9910

Table 3: Phi2-LoRA model performance

Split	Accuracy	Precision	Recall	F1 Score
Train	0.9886	0.9885	0.9886	0.9885
Eval	0.9802	0.9980	0.9802	0.9799
Test	0.9829	0.9827	0.9829	0.9827

We can see that the results from LLAMA2 and Phi2 models are nearly identical. However, they differ in one key fact. The Phi2 model requires significantly fewer models for training and the memory footprint is relatively lower. This highlights an important fact that the Phi2-LoRA model performs almost identically compared to LLAMA- LoRA model. When it comes to situations with lower memory resources and available compute, it is an easy choice to go with Phi2 like SLMs instead of full-fledged LLMs (like LLAMA2).

4. Conclusion

We have demonstrated the SLMs perform nearly as good as LLMs for simple language classification tasks such as spam text classification. SLMs require relatively few resources to train and are more efficient in terms of compute and memory requirements. They continue to show remarkable capabilities for several NLP tasks like text classification. SLMs like Phi2, and now Phi3 (multilingual-MoE) are showing remarkable capabilities for human language understanding and continue to be very popular among the research community for wide variety of tasks.

References

[1]. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample (2024). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971. 2023 Feb 27.

[2]. Javaheripi M, Bubeck S, Abdin M, Aneja J, Bubeck S, Mendes CC, Chen W, Del Giorno A, Eldan R, Gopi S. Phi-2: The surprising power of small language models. Microsoft Research Blog. 2023.

[3]. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas DD, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR. Mistral 7B. arXiv preprint arXiv:2310.06825. 2023 Oct 10.

[4]. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, Goyal A. The llama 3 herd of models. arXiv preprint arXiv:2407.21783. 2024 Jul 31.

[5]. Almeida T, Hidalgo J. SMS Spam Collection [dataset]. 2011. UCI Machine Learning Repository. Available from: https://doi.org/10.24432/C5CC84.

[6]. Elkins K, Chun J. Can GPT-3 pass a writer's Turing test?. Journal of Cultural Analytics. 2020 Sep 14;5(2).

[7]. Choo S, Kim W. A study on the evaluation of tokenizer performance in natural language processing. Applied Artificial Intelligence. 2023 Dec 31;37(1):2175112.

[8]. Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer. Deep Contextualized Word Representations. ArXiv abs/1802.05365 (2018).

[9]. Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.

[10]. Abdin M, Jacobs SA, Awan AA, Aneja J, Awadallah A, Awadalla H, Bach N, Bahree A, Bakhtiari A, Behl H, Benhaim A. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219. 2024 Apr 22.

[11]. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685. 2021 Jun 17.