Assessing Compression Strategies in Large Language Models for Symptom Extraction from Clinical Notes

Francis Boabang

1 Introduction

Large language models (LLMs) such as BERT [2], BioBERT [8], and Bio_ClinicalBERT [1] have demonstrated strong performance across a wide range of clinical NLP tasks, including named entity recognition (NER), relation extraction, and clinical document classification. Bio_ClinicalBERT, trained specifically on clinical notes from the MIMIC-III dataset, has been shown to outperform general-domain models on healthcare-specific benchmarks.

Symptom extraction, a subtask of NER, focuses on identifying and labeling mentions of patient-reported symptoms in unstructured clinical narratives. Earlier methods relied heavily on rule-based systems and clinical lexicons, such as cTAKES and MetaMap, which had limited generalizability. More recently, transformer-based models like BioBERT and ClinicalBERT have significantly improved symptom tagging performance by learning contextual representations of medical terminology [5].

To address the computational challenges of deploying large models, several model compression techniques have been proposed. Knowledge distillation [4], pruning [3], quantization [6], and low-rank matrix factorization [9] are widely adopted to reduce the size, memory footprint, and inference latency of neural networks without a significant drop in accuracy. For instance, DistilBERT [10] uses knowledge distillation to create a smaller, faster version of BERT that retains most of its predictive performance.

Although model compression is well studied in general NLP, relatively few works have explored its impact in clinical settings. Si et al. [11] conducted a benchmarking study on clinical NLP tasks and found that distilled and pruned models can achieve performance close to their uncompressed counterparts on tasks like concept extraction and sentence classification. However, there remains a gap in the literature regarding compression strategies specifically for symptom extraction using real-world clinical datasets such as MIMIC-III.

This study addresses this gap by systematically comparing multiple compression techniques applied to Bio_ClinicalBERT for the task of symptom extraction. To our knowledge, this is one of the first comprehensive evaluations of the trade-offs between model efficiency and clinical symptoms extraction task.

This report focuses on developing efficient transformer-based models for automated symptom extraction from clinical text using the MIMIC-III dataset [1]. We fine-tune Bio-ClinicalBERT on structured clinical notes to identify and classify patient-reported symptoms, aiming to enhance clinical decision support. To improve computational efficiency, we apply model compression techniques including pruning, quantization, and knowledge distillation. We benchmarked five different modeling strategies:Baseline (full fine-tuning of Bio-ClinicalBERT), Knowledge Distillation, Low-Rank Matrix Factorization, Weight Pruning and Quantization. We evaluate the models using F1 score, precision, recall, and AUROC across training epochs. The results demonstrate that optimized models can reliably extract symptoms from clinical narratives, making them suitable for deployment in resource-constrained healthcare environments. https://github.com/boabangf/Compressed-LLM-Bio_ClinicalBERT-/tree/main

2 Experimental Setup

Each model was trained for five epochs, and the results presented here focus on performance at the **5th epoch**, where all models had stabilized.

[1] https://www.kaggle.com/datasets/bilal1907/mimic-iii-10k

2.1 Metric Definitions

To evaluate model performance, we used the following metrics:

- **Accuracy**: Proportion of correct predictions
- **Precision**: Proportion of true positives among predicted positives
- **Recall**: Proportion of true positives among actual positives
- **F1 Score**: Harmonic mean of Precision and Recall
- **AUROC**: Area Under the Receiver Operating Characteristic Curve

Method	Accuracy	Precision	Recall	F1 Score	AUROC
Baseline	**0.9356**	**0.9365**	**0.9380**	**0.9372**	0.9878
Distillation	0.9333	0.9333	0.9357	0.9342	**0.9896**
Low-Rank	0.9218	0.9262	0.9230	0.9240	0.9871
Pruning	0.9287	0.9297	0.9310	0.9302	0.9883
Quantization	0.9126	0.9189	0.9135	0.9152	0.9872

Table 1: Performance comparison of different compression techniques at epoch 5.

2.2 Datasets

The data used in this study were obtained from the MIMIC-III database [7]. We used 10% of the **NOTEEVENTS** dataset from the **MIMIC-III** clinical database [2]. The NOTEEVENTS dataset comprises unstructured clinical notes such as discharge summaries, nursing notes, and physician reports. These notes provide rich textual information regarding patients' symptoms making them valuable for clinical natural language processing (NLP) tasks. After extracting 10% (i.e., this value is user-adjustable.) of the full dataset, we performed standard preprocessing and randomly split the data into training and testing sets for supervised classification. Our classification task aimed to extract symptoms from the free-text notes using **Bio-ClinicalBERT-based models**.

2.3 Analysis

At the beginning of the experiments, we fine-tuned the model by selecting optimal hyperparameters such as the learning rate, regularization strength, matrix rank, and the percentage of weights to prune. The algorithm dynamically identifies the five most frequent symptoms (this value is user-adjustable) such as *edema*, *pain*, *cough*, *fever*, and *bleeding* for training and classification using the `NOTEEVENTS` file from the MIMIC-III dataset. It should be noted that this list can vary depending on the portion of the dataset selected for training.

The baseline model achieved the best overall performance in terms of accuracy (93.56%), precision, recall, and F1 score, although it yielded a slightly lower AUROC compared to the distilled model as shown in Table 1. The distilled model achieved a higher AUROC than the baseline because it introduces regularization that reduces overfitting.

The pruned model outperformed both the low-rank matrix factorization and quantization methods. This is attributed to its implicit regularization, which helps reduce overfitting and improves generalization. The low-rank matrix factorization model, even after selecting an appropriate rank of 32, lagged slightly behind the baseline, distilled, and pruned models. This underperformance is likely due to the model requiring more training epochs beyond the five used in this study to reach optimal performance.

[2] https://www.kaggle.com/datasets/bilal1907/mimic-iii-10k

The quantized model achieved the lowest performance across accuracy, precision, recall, and F1 score. Quantization often results in reduced model precision because it lowers the bit-width of weights and activations, which can impair the model's ability to capture subtle clinical signals. This limitation is especially impactful for tasks involving symptom extraction. However, it is worth noting that the quantized model slightly outperformed the low-rank model in terms of AUROC as shown in Table 1.

3 Conclusion

This study evaluated several model compression techniques such as knowledge distillation, low-rank matrix factorization, pruning, and quantization against a baseline Bio-ClinicalBERT model for clinical text classification on the MIMIC-III NOTEEVENTS dataset.

Among all models, the baseline consistently delivered the highest performance across most metrics, including accuracy, precision, recall and F1 score, validating the effectiveness of full fine-tuning. However, the distilled model demonstrated nearly equivalent performance while achieving a higher AUROC, indicating improved generalization and robustness to overfitting. Pruning also emerged as a strong contender, balancing compression and predictive power effectively. Low-rank matrix factorization showed moderate performance and would likely benefit from additional training epochs to fully realize its potential. Quantization, while offering the most significant efficiency gains, suffered the largest drop in classification performance, particularly for precision-dependent clinical tasks.

Overall, this study highlights that while compression methods can significantly reduce model complexity, careful tuning and selection are required to maintain clinical utility. For real-world deployments where resources are limited, knowledge distillation and pruning provide practical alternatives to baseline models without severe trade-offs in accuracy.

References

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[5] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[6] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *arXiv preprint arXiv:1712.05877*, 2018.

[7] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[9] Xuezhe Ma, Jing Liu, and Eduard Hovy. Tensorized lstms for sequence learning. *arXiv preprint arXiv:1901.10717*, 2019.

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[11] Yiqing Si, Jin Wang, Hua Xu, and Kirk Roberts. Benchmarking transformer-based models on clinical nlp tasks. *Journal of Biomedical Informatics*, 117:103773, 2021.