Efficient Compression of Clinical Language Models via Distillation and Structured Pruning

Abstract

We propose an efficient model compression framework for clinical language models targeting multi-label diagnosis classification tasks. Using BioLinkBERT fine-tuned on MIMIC-III discharge summaries, our approach combines knowledge distillation with a novel structured pruning strategy, Smart GRIFFIN. Unlike conventional pruning techniques, Smart GRIFFIN selects important feed-forward network neurons based on L2 norm statistics and incorporates a stability-aware heuristic to skip pruning in structurally stable layers. We systematically evaluate three experimental pipelines that differ in the timing and application of pruning relative to distillation. Results show that jointly compressing both the teacher and student models yields the highest performance, with Smart GRIFFIN achieving superior micro and macro F1 scores, a 3.11× compression ratio, and up to 1.79× real-time inference speedup—highlighting its effectiveness for deployment in latency- and resource-constrained clinical environments.

1. Introduction

Transformer-based language models have significantly advanced the field of clinical natural language processing, enabling high-performance tasks such as diagnosis classification, entity recognition, and symptom extraction. However, their substantial computational demands—both in memory and latency—pose serious challenges for real-time deployment in resource-constrained healthcare settings. These constraints are particularly critical when deploying models at the point of care, where interpretability and efficiency are as important as predictive accuracy. To address these challenges, model compression techniques such as knowledge distillation and structured pruning have emerged as effective strategies. Distillation transfers knowledge from a large, overparameterized teacher model into a compact student, often yielding models with similar predictive power but reduced computational footprint. Pruning, on the other hand, seeks to remove redundant or low-importance components of the model architecture, further reducing inference cost.

In this work, we investigate the combined effect of these two complementary techniques within the clinical NLP domain. Building on recent advances in structured pruning, we propose Smart GRIFFIN, an extension of the GRIFFIN method designed for encoder-based models and multi-label classification. Smart GRIFFIN introduces a neuron importance ranking mechanism based on L2 norm statistics, a heuristic layer-skipping criterion to preserve stable transformer blocks, and post-pruning recovery to ensure robustness. Through extensive experimentation on the MIMIC-III discharge summaries dataset using BioLinkBERT [1], we demonstrate that our approach yields highly efficient models with minimal degradation—or in some cases, improvement—in generalization performance.

2. Methods

2.1 Dataset

We utilize 10,000 discharge summaries from the MIMIC-III dataset [2]. To ensure balanced training, samples lacking any of the top 20 most frequent labels were excluded, and a greedy label-aware subsampling algorithm was employed to enforce uniform representation across classes.

2.2 Teacher and Student Models

For this project, we adopt BioLinkBERT-large as the teacher model, which is fine-tuned using binary cross-entropy loss. The student model is BioLinkBERT-base, trained via knowledge distillation to replicate the teacher's predictions while being more compact and efficient. The distillation objective:

$$\mathcal{L}_{total} = \alpha \cdot \mathrm{KL}\big(\sigma(z_s/T), \sigma(z_t/T)\big) + (1-\alpha) \cdot \mathrm{BCE}(z_s, y)$$

where α balances the contribution of the distillation loss and the supervised loss, KL is the Kullback–Leibler divergence between the softened output distributions of the student ($\mathrm{z_s}$) and teacher (z_t) logits, $\sigma(\cdot)$ denotes the softmax function with temperature T, and $BCE(\cdot,\cdot)$ represents the binary cross-entropy loss between the student logits and the ground-truth labels y. In this work, we set $\alpha = 0.3$, temperature $T = 1.0$.

2.3 Smart GRIFFIN Structured Pruning

In this project, experiments use **Smart GRIFFIN**, our structured pruning extension of the original GRIFFIN [3], adapted to encoder-only architectures and multi-label classification. Key improvements introduced in Smart GRIFFIN include:

- **Structured slicing over masking**: Instead of zeroing out weights, we perform **actual pruning of weight tensors**, reducing parameter count and runtime memory use.
- **Neuron importance ranking via L2 norm**: FFN hidden units are ranked using the L2 norm of their input weights, preserving the most impactful activations.
- **Layer-skipping heuristic**: We skip pruning in layers with low standard deviation in importance scores (< 1e-3), which protects against pruning stable or already-optimized layers.
- **Post-pruning recovery**: We introduce a lightweight one-epoch fine-tuning phase to recover from potential performance drops, minimizing retraining cost.

These enhancements make Smart GRIFFIN robust, deployment-friendly, and effective for structured pruning of classification heads in encoder models like BioLinkBERT.

Table I. Details of the Conducted Experiments

Experiment ID	Description
A	Distillation Only: Student is trained from full teacher without pruning.
B	Distill → Prune Student: Student is trained from teacher, then pruned via Smart GRIFFIN.
C	Prune Teacher → Distill → Prune Student: Teacher is pruned first; student is distilled and then pruned.

3. Experimental Design

Three experimental pipelines (Table I) were designed to systematically investigate the impact of structured pruning when applied at different stages of the distillation and training process. All models were trained using a batch size of 1 (to accommodate long clinical notes) and a learning rate of 1e-5, with early stopping based on validation performance. To comprehensively assess each approach, we measured accuracy, micro and macro F1 scores, parameter count, model compression ratio, and inference latency (in milliseconds per sample). All experiments were conducted on a Windows workstation equipped with an NVIDIA RTX 4070 GPU (12GB VRAM), using PyTorch library and HuggingFace Transformers to ensure reproducibility and compatibility with real-world deployment pipelines.

Table II. Final Comparison Across Pipelines

Pipeline	Accuracy (%)	Micro F1 (%)	Macro F1 (%)	Latency (ms)	Parameters (million)	Compression	Speedup
Exp. A	64.18	40.88	41.86	12.59	108	3.05×	1.0×
Exp. B	**89.74**	53.20	40.47	**5.88**	105	3.09×	1.76×
Exp. C	89.57	**57.16**	**45.42**	6.10	**104**	**3.11×**	**1.79×**

Results in bold and red indicate the best performance.

4. Results

This section analyzes the empirical results of the experimental pipelines summarized in Table I, which examine the effects of distillation and pruning applied at different stages of model compression for multi-label diagnosis classification. Each pipeline is assessed across predictive performance (accuracy, micro/macro F1), model size (parameter count and compression ratio), and computational efficiency (latency and speedup). The obtained results are tabulated in Table II.

4.1 Accuracy and F1 Analysis

Experiment A (Distillation Only) served as the baseline. The student model, distilled directly from the uncompressed teacher, achieved 64.18% accuracy, 40.88% micro-F1, and 41.86% macro-F1. While this pipeline demonstrates strong compression (3.08×) and forms a lightweight model (108M parameters), its predictive performance is substantially lower than the other configurations. This underscores a limitation of distillation alone: although the student benefits from softened label distributions, it does not receive regularization from structural constraints and may underfit or learn suboptimal representations.

Experiment B (Distill → Prune Student) improves upon this baseline significantly. Post-distillation pruning using our Smart GRIFFIN method increases the student model's accuracy to 89.74% and micro-F1 to 53.20%, a large performance leap over

Experiment A. While macro-F1 declines slightly to 40.47%, this suggests a mild degradation in performance on rare classes—a trade-off common when aggressively pruning without modifying the teacher. Notably, this pipeline also reduces inference latency to 5.88 ms, yielding a 1.76× speedup, despite only a modest compression gain (1.03×). This demonstrates that even when total parameter count is nearly unchanged, smart structured pruning of FFN layers can yield real hardware-level benefits.

Experiment C (Prune Teacher → Distill → Prune Student) offers the most balanced performance overall. Although its accuracy (89.57%) is slightly lower than Experiment B by a marginal 0.17%, it achieves the highest micro-F1 (57.16%) and macro-F1 (45.42%). These improvements demonstrate that pruning the teacher model before distillation results in a compressed yet informative supervision signal, guiding the student to learn more generalizable representations. Moreover, this configuration achieves the highest compression ratio (3.11×) and fastest inference latency (6.10 ms, 1.79× speedup) while using fewer parameters (104M) than both prior experiments. These results confirm that compressing both teacher and student models provides compounded benefits in performance and efficiency.

4.2 Latency and Compression

From an efficiency standpoint, Experiments B and C offer the strongest gains in both compression and inference speed, while Experiment A, despite achieving a solid 3.05× compression, offers no speedup and remains the slowest in latency. Experiment C stands out as the most efficient configuration, achieving the highest compression ratio (3.11×) and the fastest inference time (6.10 ms), alongside robust predictive performance across all metrics. These results demonstrate that Smart GRIFFIN pruning, when applied consistently to both the teacher and student models, not only preserves informative neurons but also eliminates architectural redundancy—yielding structural simplification that accelerates downstream inference. Interestingly, although Experiments B and C have nearly identical parameter counts (105M vs. 104M), Experiment C achieves a higher speedup (1.79× vs. 1.76×). This discrepancy is likely since pruning the teacher model propagates architectural sparsity into the distilled student, thereby reducing cumulative transformer computation during inference. In contrast, Experiment A, which skips pruning entirely, incurs the longest latency (12.59 ms), confirming that compression alone does not equate to hardware-level efficiency.

4.3 Impact of Smart GRIFFIN Pruning

Across both Experiments B and C, Smart GRIFFIN consistently yielded gains without degrading performance. The use of L2-norm neuron ranking, combined with a statistical layer-skipping heuristic, prevented over-pruning of stable layers — a common risk in naïve structured pruning. Our method also avoids the pitfalls of unstructured sparsity (e.g., masking-based methods), delivering true speedups and smaller models in practice. Moreover, the slight performance edge of Experiment C in both micro and macro F1 suggests that a pruned teacher retains essential generalization capabilities and can be used effectively to supervise a compact student. This observation has significant implications for low-resource deployments, where training large teachers may not be feasible.

5. Conclusion

This work presents an effective compression strategy for clinical natural language processing that integrates multi-label knowledge distillation with a novel variant of structured pruning, referred to as Smart GRIFFIN. Unlike traditional unstructured approaches, our method performs explicit slicing of FFN layers based on L2-norm activation statistics and incorporates a heuristic for layer-wise pruning stability by skipping layers with low score variance. This results in models that are not only more compact, but also more computationally efficient and robust in generalization. Among the evaluated pipelines, Experiment C—which applies Smart GRIFFIN pruning to both the teacher and the student—achieves the best overall performance, balancing accuracy, micro/macro F1, model size, and latency. These findings underscore the importance of compressing the supervision source (teacher) alongside the student to maximize both effectiveness and efficiency. All experiments were conducted in a reproducible and modular framework, with complete code, metrics, and trained checkpoints made available to support future extensions and deployment in real-world clinical systems.

References

[1] Y. Yoon, M. T. Bahadori, C. Zhang, Y. Luo, and J. Sun, "BioLinkBERT: Pretrained Biomedical Language Representation Model for Biomedical Text Mining," *Bioinformatics Advances*, vol.

[2] A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016. [Online]. Available: https://doi.org/10.1038/sdata.2016.35

[3] Y. Wang et al., "GRIFFIN: Efficient FFN Compression for Language Models via Structured Pruning," *arXiv preprint arXiv:2402.19427*, 2024.