

Pruning Dinâmico Baseado em Ativação para Especialização por Domínio em Modelos de Linguagem

Claudio Honório da Silva Junior¹, Gabriel Aragão Sanches²

¹Faculdade Donaduzzi - Toledo, PR - Brazil

claudiohonorio.ac@gmail.com, gabriel-aragao@outlook.com.br

Abstract. *This paper investigates activation-based dynamic pruning as a strategy for specializing large language models in specific domains. The method estimates component relevance from activation magnitude and recurrence, enabling the adaptive removal of less useful units during inference. Partial results with the Llama-2-7B model show pruning rates of 30% for Law, 25% for Medicine, and 20% for Literature, motivating a broader evaluation of response quality, latency, and computational efficiency.*

Resumo. *Este artigo investiga o pruning dinâmico baseado em ativação como estratégia para especializar modelos de linguagem de larga escala em domínios específicos. O método estima a relevância dos componentes a partir da magnitude e da recorrência das ativações, permitindo a remoção adaptativa de unidades menos úteis durante a inferência. Resultados parciais com o modelo Llama-2-7B mostram taxas de poda de 30% para Direito, 25% para Medicina e 20% para Literatura, motivando uma avaliação mais ampla da qualidade das respostas, latência e eficiência computacional.*

1. Introdução

Os avanços recentes em modelos de linguagem de larga escala (LLMs) têm demonstrado desempenho expressivo em diversas tarefas de processamento de linguagem natural. Esses modelos são capazes de executar atividades complexas a partir de grandes volumes de dados e de um número elevado de parâmetros, alcançando resultados relevantes em múltiplos domínios [Brown et al. 2020].

Entretanto, esse desempenho está diretamente associado ao aumento da escala dos modelos, o que implica custos computacionais elevados e desafios práticos de implantação. Conforme discutido por [Zhu et al. 2024], as capacidades desses modelos vêm acompanhadas de limitações relacionadas ao seu tamanho e às suas exigências computacionais. De forma complementar, [Strubell et al. 2019] destacam que os ganhos de acurácia em modelos profundos de PLN dependem, em muitos casos, da disponibilidade de recursos computacionais substanciais. Esse cenário evidencia uma tensão entre desempenho e viabilidade operacional, especialmente em ambientes com restrições de hardware e aplicações especializadas por domínio.

Nesse contexto, estratégias de compressão e especialização tornam-se relevantes para aproximar a capacidade dos LLMs das exigências de uso em cenários com recursos limitados. Entre essas estratégias, o *pruning* desponta como alternativa promissora para reduzir a complexidade do modelo sem descaracterizar completamente seu comportamento funcional.

2. Objetivo

O objetivo deste trabalho é investigar se o *pruning* dinâmico guiado por ativações internas pode especializar um modelo de linguagem para domínios específicos sem exigir novo treinamento completo. Especificamente, busca-se identificar componentes pouco relevantes a partir da magnitude e da recorrência de ativação, aplicar a poda de forma adaptativa nos domínios de Literatura, Medicina e Direito e analisar, ainda que de maneira preliminar, os efeitos dessa redução estrutural sobre as respostas geradas.

3. Referencial Teórico

O avanço dos modelos de linguagem de larga escala está ligado à evolução das arquiteturas neurais e ao aumento da escala de treinamento. Nesse contexto, o Transformer tornou-se um marco ao substituir estruturas recorrentes por mecanismos de atenção, permitindo maior paralelização e eficiência no treinamento [Vaswani et al. 2017].

A literatura também demonstra que o desempenho desses modelos está fortemente correlacionado com sua escala. Estudos sobre leis de escala mostram que a perda em modelos de linguagem segue relações do tipo *power-law* em função do número de parâmetros, da quantidade de dados e do orçamento computacional [Kaplan et al. 2020]. Esse comportamento levou ao desenvolvimento de modelos cada vez maiores, sob a premissa de que o aumento da escala resulta em ganhos sistemáticos de desempenho. No entanto, trabalhos mais recentes indicam que o desempenho ótimo depende de um equilíbrio entre tamanho do modelo e volume de dados [Hoffmann et al. 2022].

Modelos com bilhões de parâmetros demandam grandes quantidades de memória e recursos de processamento, tornando sua utilização restrita a ambientes com infraestrutura avançada [Zhu et al. 2024]. Esse cenário motiva a investigação de técnicas de compressão que permitam reduzir o custo de inferência sem comprometer substancialmente o desempenho.

Em modelos de linguagem, métodos recentes de *pruning* exploram redundâncias estruturais para remover componentes pouco relevantes. O SparseGPT mostra que é possível alcançar alta esparsidade com impacto controlado na qualidade [Frantar and Alistarh 2023]. Já métodos como Wanda combinam informações de pesos e ativações para estimar a importância dos componentes da rede [Sun et al. 2023].

Embora esses avanços representem um progresso significativo, a maioria das abordagens ainda depende de métricas estáticas ou de informações limitadas do modelo. Nesse sentido, permanece em aberto a investigação de métodos que considerem de forma mais abrangente o comportamento do modelo durante a inferência, especialmente em cenários específicos de domínio. A análise de ativações internas surge, portanto, como uma alternativa promissora para capturar padrões de uso do modelo, permitindo identificar componentes relevantes e possibilitando estratégias de *pruning* mais adaptativas.

4. Metodologia

Foi adotado um delineamento experimental exploratório com o modelo Llama-2-7B para investigar a eficácia de estratégias de *pruning* dinâmico, com foco na especialização por domínio. Foram considerados três domínios específicos — Literatura, Medicina e Direito — escolhidos por apresentarem vocabulário técnico e estruturas conceituais suficientemente distintas. Para cada domínio, foi construído um conjunto com aproximadamente

100 questões abertas, distribuídas em diferentes níveis de complexidade, com o objetivo de avaliar a capacidade de especialização do modelo em cenário *zero-shot*, sem ajuste adicional [Bommasani et al. 2021, Brown et al. 2020].

As perguntas foram submetidas diretamente ao modelo base e às versões podadas, e as respostas geradas antes e depois da poda foram comparadas qualitativamente. O objetivo foi observar se a redução estrutural comprometeria a capacidade de resposta ou, alternativamente, favoreceria maior especialização no domínio avaliado. A proposta parte da hipótese de que componentes com baixa ativação recorrente durante a inferência tendem a apresentar menor contribuição funcional para tarefas especializadas.

A importância de cada componente foi estimada a partir de duas métricas complementares: magnitude média de ativação e recorrência de ativação. A magnitude média quantifica a intensidade com que uma unidade é ativada ao longo de N amostras de entrada, já a recorrência quantifica a frequência com que essa ativação ultrapassa o limiar $\tau = 0,01$:

$$R_l = \frac{1}{N} \sum_{i=1}^N I(|a_l^{(i)}| > \tau)$$

em que $a_l^{(i)}$ representa a ativação da unidade na camada l para a amostra i e que I é a função indicadora. A combinação dessas métricas foi utilizada para definir a importância final do componente:

$$I_l = M_l \cdot (1 + R_l)$$

Essa formulação prioriza unidades cuja contribuição é, ao mesmo tempo, intensa e consistente, penalizando ativações esporádicas e favorecendo componentes funcionalmente mais estáveis durante a inferência.

5. Resultados

Nesta seção, apresentam-se resultados parciais obtidos com o modelo de 1.1B parâmetros, considerando a taxa de poda alcançada em cada domínio específico.

Tabela 1. Taxa de poda por domínio para o modelo de 1.1B parâmetros

Domínio	Taxa de poda	Percentual
Direito	0.30	30%
Medicina	0.25	25%
Literatura	0.20	20%

Os resultados indicam que o domínio de Direito apresentou a maior taxa de poda, com 0.30, seguido por Medicina, com 0.25, e Literatura, com 0.20. Esses valores correspondem, respectivamente, à remoção de 30%, 25% e 20% dos componentes considerados menos relevantes em cada domínio. De forma preliminar, esse comportamento sugere que o conjunto de ativações observado nas questões jurídicas concentrou uma quantidade maior de componentes elegíveis à remoção segundo a métrica proposta.

Como métrica complementar de desempenho, avaliou-se a perplexidade no conjunto WikiText-2. O modelo base apresentou perplexidade de 6.52, enquanto a versão podada apresentou valores entre 7.878 e 7.980. Esse aumento, aproximadamente entre 20% e 25%, indica degradação moderada da capacidade preditiva após a poda. Assim, embora a técnica reduza a quantidade de componentes ativos, há um custo mensurável em desempenho, reforçando a necessidade de equilibrar taxa de poda e qualidade das respostas.

Tabela 2. Exemplos qualitativos de perguntas e respostas por domínio

Domínio	Pergunta representativa	Síntese da resposta gerada
Literatura	Qual é a diferença entre narrador-personagem e narrador onisciente?	A resposta distinguiu os dois tipos de narrador ao contrastar perspectiva limitada e perspectiva onisciente, destacando diferenças de ponto de vista, acesso às informações e grau de subjetividade na narrativa.
Medicina	Explique o processo de fotossíntese em células vegetais e sua relevância para a medicina.	A resposta apresentou as etapas centrais da fotossíntese e relacionou o tema a aplicações médicas, como obtenção de compostos bioativos, desenvolvimento de fármacos e usos terapêuticos de derivados vegetais.
Direito	Em quais situações um contrato pode ser considerado nulo, anulável ou inexigível?	A resposta diferenciou os três conceitos jurídicos e associou cada caso a situações típicas, como ilicitude, incapacidade, coação, fraude, ausência de requisitos legais e limitações de exigibilidade em juízo.

Os exemplos qualitativos apresentados na Tabela 2 sugerem que o modelo preserva capacidade de resposta em diferentes domínios mesmo após a aplicação da poda. Embora os resultados ainda sejam parciais, a diferença entre as taxas de poda indica que a especialização por domínio pode responder de forma distinta à estratégia baseada em ativação.

6. Conclusão

Os resultados apresentados indicam que o *pruning* dinâmico via ativação é uma abordagem promissora para a especialização de modelos de linguagem em domínios específicos. Mesmo em caráter parcial, a análise mostrou que a taxa de poda varia conforme o domínio avaliado, com destaque para Direito, sem descaracterizar a proposta de comparação qualitativa das respostas geradas. Ainda assim, a análise qualitativa baseada em aproximadamente 100 perguntas por domínio é inconclusiva para garantir segurança e confiabilidade em aplicações críticas, como a supracitada, em que respostas incorretas podem produzir interpretações sensíveis.

Como continuidade, pretende-se ampliar a avaliação com métricas quantitativas de qualidade das respostas, tempo de inferência e consumo de memória, além de comparar a abordagem com outras estratégias de compressão. Esses desdobramentos podem contribuir para consolidar a viabilidade do método como alternativa prática para adaptação eficiente de LLMs.

Referências

- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Frantar, E. and Alistarh, D. (2023). Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pages 10323–10337. PMLR.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *Conference on Neural Information Processing System*.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3645–3650.
- Sun, M., Liu, Z., Bair, A., and Kolter, J. Z. (2023). A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Zhu, X., Li, J., Liu, Y., Ma, C., and Wang, W. (2024). A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.