VIII SIMPÓSIO DE INICIAÇÃO CIENTÍFICA

DO CONHECIMENTO ACADÊMICO À TRANSFORMAÇÃO SUSTENTÁVEL: INOVAÇÃO COM VALIDAÇÃO CIENTÍFICA

Campus Itabira *Campus* Itajubá

ESTRATÉGIAS DE GERENCIAMENTO DE KV CACHE: UM ESTUDO COMPARATIVO DE DESEMPENHO E MEMÓRIA

DOMICIANO, Mirela V., RIBEIRO, Ryan F. F., MULLER JUNIOR, Egon L.

UNIVERSIDADE FEDERAL DE ITAJUBÁ - CAMPUS ITAJUBÁ

INTRODUÇÃO

Modelos de linguagem de grande escala (LLMs), baseados na arquitetura Transformer, tornaram-se a base de diversas aplicações de IA. A geração de texto ocorre de forma autorregressiva. Para evitar a recomputação onerosa das representações de atenção, utiliza-se o KV Cache (POPE et al., 2023), uma técnica crucial que acelera a inferência.

Contudo, essa técnica demanda alta memória, especialmente em contextos longos, configurando um dos principais gargalos para a eficiência dos LLMs (LIU et al., 2024). Este trabalho realiza um estudo comparativo entre três estratégias de gerenciamento de KV Cache — ausência de cache, cache dinâmico e cache estático — avaliando seu impacto no desempenho e no consumo de memória. A pesquisa busca esclarecer os trade-offs entre eficiência e uso de recursos, contribuindo para o avanço de LLMs mais rápidos, escaláveis e acessíveis.

METODOLOGIA

O estudo foi conduzido no Google Colab, um ambiente de execução com GPU e implementado integralmente em Python. O modelo escolhido foi o de linguagem meta-llama/Llama-3.2-1B — um LLM de arquitetura Transformer representativa — carregado via o framework Hugging Face Transformers. Utilizou-se o PyTorch para operações de tensores e o psutil para monitoramento do uso de memória.

Para avaliação, desenvolveu-se um benchmark conversacional com múltiplos cenários do setor financeiro, simulando diálogos de vários turnos para observar o comportamento do KV Cache conforme o contexto aumenta. Foram comparadas três estratégias: sem cache, cache dinâmico e cache estático, analisando o impacto em tempo de geração, uso de memória e vazão (tokens/s). Todos os experimentos foram automatizados e seus resultados registrados em arquivos CSV para posterior análise.

RESULTADOS OBTIDOS E ANÁLISE

Os experimentos demonstraram diferenças significativas entre as estratégias de gerenciamento de KV Cache. Conforme resumido na Tabela 1, o uso de cache resultou em uma melhoria superior a dez vezes no tempo de geração e na vazão em relação à ausência de cache.

Estratégia	Tempo (s)	Memória (MB)	Tokens/s
None	54.275	0.0098	6.3
Dynamic	3.443	0.0098	34.1
Static	4.857	4.5137	34.4

Tabela 1 - Resumo dos Indicadores de Desempenho

A estratégia dinâmica apresentou o melhor equilíbrio entre desempenho e uso de memória, com tempo médio de 3,44s, vazão de 34 tokens/s e consumo mínimo de memória. A estratégia estática, embora tenha desempenho semelhante em velocidade, demandou maior uso de memória (4,51 MB). A configuração sem cachê, por sua vez, mostrou-se inviável para aplicações interativas devido ao tempo médio de geração elevado (54,27s) e baixa vazão (6,3 tokens/s).

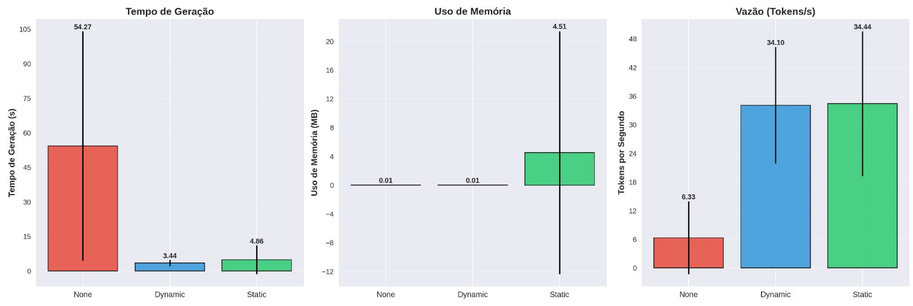

Figura 1 - Gráficos Comparativos das Estratégias

CONCLUSÕES

O estudo evidenciou a relevância do gerenciamento de KV Cache para o desempenho de LLMs, comparando as estratégias sem cache, dinâmico e estático. Os resultados indicaram que o uso de cache reduz significativamente o tempo de resposta, com desempenho até dez vezes superior em relação à ausência de cache. O cache dinâmico destacou-se por oferecer o melhor equilíbrio entre velocidade e consumo de memória, enquanto o cache estático, embora rápido, mostrou maior instabilidade e consumo de memória.

Conclui-se que o cache dinâmico é a abordagem mais eficiente e prática para LLMs, e recomenda-sse que estudos futuros a exploração de técnicas avançadas, como políticas de evicção e quantização, para otimizar ainda mais o uso de memória e ampliar a capacidade de contexto desses modelos.

AGRADECIMENTOS E FINANCIAMENTOS

Agradecemos aos colegas e professores do PET-TEC por promoverem um ambiente de inovação tecnológica que permitiu o desenvolvimento deste trabalho. Nossa gratidão também à UNIFEI e ao FNDE pelo suporte e incentivo, essenciais para a realização deste projeto.

REFERÊNCIAS

LIU, S., et al. KVQuant: Towards 10 Million Context Length LLM Inference with KV Cache Quantization. arXiv preprint arXiv:2405.09980, 2024. Disponível em: https://arxiv.org/abs/2405.09980. Acesso em: ago. 2025.

POPE, R., et al. Efficiently Scaling Transformer Inference. In: Proceedings of the 6th MLSys Conference, 2023. Disponível em: https://arxiv.org/abs/2211.05102. Acesso em: ago. 2025.

Diretoria de Pesquisa

CNPq

FAPEMIG