Aplicando técnicas de fine-tuning e compressão em Super Tiny Language Models na geração de pictograma para a comunicação alternativa

Lenon Anthony e Silas Augusto

Setembro, 2024

Abstract

Este artigo investiga a aplicação de técnicas de fine-tuning e compressão em Super Tiny Language Models (STLMs) para a geração de pictogramas, com o objetivo de aprimorar a comunicação alternativa. Focando no modelo TinyLlama com 1.1 bilhão de parâmetros, a pesquisa explora métodos que permitem otimizar modelos de linguagem de pequena escala para a criação eficiente de representações visuais, como pictogramas, sem comprometer a qualidade ou a performance. Os resultados demonstram como essas técnicas podem ser utilizadas para facilitar a comunicação visual em plataformas adaptativas.

1 Introdução

A comunicação alternativa [11] é essencial para indivíduos com dificuldades de fala ou limitações na comunicação verbal, oferecendo-lhes meios eficazes de interação. Com o avanço das tecnologias, aplicativos de comunicação alternativa, como o Livox, desempenham um papel crucial na promoção da inclusão social e na melhoria da qualidade de vida dessas pessoas, integrando inteligência artificial e outras inovações. A capacidade de personalização e adaptação desses aplicativos torna possível atender a uma diversidade de perfis de usuários, garantindo que cada indivíduo tenha uma ferramenta que realmente responda às suas particularidades.

O desenvolvimento de softwares e aplicativos tem desempenhado um papel crucial na promoção da inclusão social e na melhoria da qualidade de vida de pessoas com dificuldades de comunicação. Esses recursos tecnológicos oferecem soluções que vão além das abordagens tradicionais, permitindo a criação de interfaces intuitivas e adaptáveis que se moldam às capacidades e limitações dos usuários [3]. Aplicativos de comunicação alternativa, como o Livox, exemplificam esse avanço ao incorporar inteligência artificial e outros recursos inovadores, proporcionando a essas pessoas maior autonomia e eficácia na expressão de suas necessidades e desejos.

Apesar das inovações apresentadas, a otimização de aplicativos de comunicação alternativa para dispositivos com recursos limitados continua sendo um desafio significativo. Muitos usuários dependem de dispositivos móveis com capacidade de processamento e memória reduzidas, o que pode comprometer o desempenho do aplicativo. Nesse cenário, nossa pesquisa se concentra no desenvolvimento e fine-tuning de uma STLM (Super Tiny Language Model) [1], que será integrada ao Livox em uma fase futura. Essa integração tem como objetivo aprimorar ainda mais a funcionalidade e a eficiência da plataforma, garantindo que ela permaneça acessível e eficaz, independentemente das limitações de hardware dos dispositivos utilizados.

O objetivo deste trabalho é explorar e aplicar técnicas de fine-tuning e compressão em um modelo de linguagem de pequena escala, o TinyLlama, com o propósito de gerar atributos, que são *written_text*, *spoken_text* [2] e *image_id*, que representam o texto escrito, o texto falado e o identificador da imagem, respectivamente, em formato CSV. Através deste CSV gerado, futuramente pretendemos alimentar o JSON que representa um cartão dentro da aplicação Livox com estes atributos. Ao adotar essa estratégia, pretende-se melhorar a eficácia e a personalização da comunicação alternativa, oferecendo uma ferramenta que responda de maneira mais adequada e precisa às necessidades específicas de usuários com limitações na fala ou na comunicação verbal, contribuindo para a inclusão social e para a melhoria da qualidade de vida dessas pessoas.

2 Background

O Livox [8] representa um avanço significativo na comunicação aumentativa e alternativa (CAA) ao integrar inteligência artificial em uma plataforma altamente adaptável. Desenvolvido com o objetivo de oferecer uma interface profundamente personalizável, o Livox capacita indivíduos com dificuldades de comunicação a moldar o aplicativo conforme suas necessidades específicas, promovendo uma interação mais eficaz, independente e centrada no usuário.

Figure 1: Exemplo de pictograma criado no Livox

Uma das funcionalidades mais notáveis do Livox é a capacidade de criar cartões personalizados em formatos de pictogramas como mostrado na Figura 1. Esses pictogramas não apenas oferecem representações visuais simplificadas,

mas também refletem a realidade individual do usuário, facilitando uma comunicação mais intuitiva e alinhada com o contexto pessoal de cada indivíduo.

3 Trabalhos relacionados

A evolução contínua dos Modelos de Linguagem de Grande Escala (LLMs) tem catalisado avanços notáveis no campo do Processamento de Linguagem Natural (PLN). No entanto, apesar de sua eficácia, os LLMs enfrentam desafios significativos relacionados ao consumo elevado de recursos computacionais e energia. Em resposta a essas questões, têm emergido os Modelos de Linguagem Super Pequenos (STLMs) [1], projetados para oferecer desempenho comparável com uma fração dos parâmetros necessários pelos seus homólogos maiores.

O TinyLlama, discutido por Guertler, et al. [4] , é um exemplo de STLM que demonstra como modelos menores podem competir com grandes modelos de base, ao mesmo tempo em que reduzem a necessidade de recursos computacionais intensivos. O TinyLlama utiliza técnicas como tokenização em nível de byte com um mecanismo de pooling, weight tying e estratégias de treinamento eficientes para melhorar a eficiência de parâmetros e de amostras. Essas técnicas são de grande relevância para o nosso estudo, que busca otimizar a criação de pictogramas em plataformas de comunicação alternativa.

Além disso, a revisão abrangente apresentada em *"The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs"* [5], oferece uma visão detalhada sobre as práticas e tecnologias mais recentes em fine-tuning de modelos de linguagem. Esta revisão enfatiza a importância de técnicas como distilação de conhecimento e quantização, que são cruciais para adaptar modelos maiores para contextos mais restritos de recursos. Essas práticas são essenciais para garantir que o TinyLlama possa ser ajustado para tarefas específicas, como a geração de pictogramas, mantendo a performance em cenários de comunicação alternativa.

Edward J. Hu, et al. no artigo*"LoRA: Low-Rank Adaptation of Large Language Models"* [6] abordam o potencial das adaptações com o Low-Rank Adaptation (LoRA) para realizar ajustes finos eficientes em modelos de linguagem preexistentes, sem a necessidade de reajustar todos os parâmetros. Este método permite modificar apenas uma pequena porção dos parâmetros do modelo, reduzindo significativamente o custo computacional associado ao fine-tuning tradicional. Este trabalho é particularmente relevante para o nosso projeto, que visa integrar técnicas de compressão e fine-tuning em STLMs utilizados para a geração de pictogramas em sistemas de comunicação alternativa.

Apesar de não estarem no contexto de geração de texto para comunicação alternativa, esses trabalhos destacam a viabilidade de modelos menores e otimizados em aplicações que exigem eficiência tanto em termos de custo quanto de desempenho, alinhando-se com os objetivos deste estudo.

4 Metodologia

Nossa abordagem utiliza técnicas avançadas de geração de texto para potencializar a eficácia dos sistemas de comunicação aumentativa e alternativa através de um Modelo de Linguagem Super Pequeno (STLM) para geração de pictogramas. Assim, passamos por etapas de desenvolvimento, treinamento e implementação.

4.1 Processo de Criação de Dataset

Utilizamos uma abordagem onde o ChatGPT-3.5 foi configurado para assumir o papel de um assistente de pessoas com dificuldades de comunicação. Solicitamos que o modelo gerasse uma lista de cartões que representassem tarefas do dia, onde o output, em formato CSV, irá conter atributos essenciais como *written_text, spoken_text e image_id*. Esses cartões são projetados para serem utilizados como base de dados para o treinamento do nosso STLMs, assegurando que o conteúdo seja diretamente aplicável e relevante para o contexto da comunicação alternativa. O dataset gerado foi submetido a uma revisão humana para garantir a qualidade e a pertinência dos dados.

Em um futuro próximo, estamos planejando alimentar nosso dataset com dados reais de usuários, permitindo que nosso modelo de linguagem se torne mais preciso e eficaz. Isso permitirá que o modelo aprenda com exemplos concretos, se adapte às nuances da linguagem natural e forneça respostas mais personalizadas e satisfatórias. Com essa abordagem, pretendemos criar um sistema robusto e capaz de lidar com situações imprevisíveis, tornando-se um dos mais avançados e eficazes do mercado.

4.2 Escolha do Modelo

A escolha do TinyLlama [4] foi fundamentada em seu desempenho destacado nas métricas avaliadas, demonstrando que, apesar de possuir um número de parâmetros significativamente menor (1B) em comparação com modelos maiores, como o Phi-3 (3B), ele alcança resultados superiores em benchmarks de conjuntos de dados de perguntas e respostas como BLiMP e *ARC_easy*. Esses resultados indicam que, mesmo sendo mais compacto, o TinyLlama é capaz de compreender e aplicar conhecimentos complexos de forma eficiente. Isso o torna particularmente adequado para aplicações em dispositivos com recursos limitados, onde a eficiência computacional deve ser equilibrada com a precisão e a velocidade de processamento

4.3 Treinamento e Otimização

Para treinar o modelo de maneira eficiente e otimizar seu desempenho, adotamos uma série de estratégias específicas. Estas incluíram técnicas de redução de complexidade, ajuste cuidadoso de hiperparâmetros e testes contínuos para maximizar a eficácia e a velocidade do modelo.

4.3.1 Quantização

Para otimizar ainda mais o uso de memória, implementamos técnicas de quantização [11] de 4 bits usando a configuração BitsAndBytes. Isso permitiu reduzir significativamente o footprint de memória do modelo, mantendo um bom equilíbrio entre desempenho e eficiência.

4.3.2 Técnica Low-Rank Adaptation

Empregamos a técnica LoRA [7] para fine-tuning 2, que nos permitiu treinar apenas um pequeno conjunto de parâmetros adicionais, mantendo a maior parte do modelo base congelada. A configuração LoRA ajustada para otimizar o processo de aprendizagem como mostrado na tabela 1.

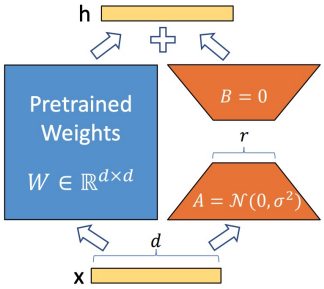

Figure 2: Arquitetura LORA

Table 1: Tabela de parâmetros do LoraConfig

parâmetro	valor
r	8
lora_alpha	16
lora_dropout	0.07

4.3.3 Otimização do Treinamento

Várias estratégias foram implementadas para otimizar o processo de treinamento:

Utilizamos *gradient_accumulation_steps=16* para simular batches maiores, permitindo treinar efetivamente com datasets maiores mesmo com recursos de hardware limitados. Optamos pelo otimizador AdamW com precisão de 32 bits *paged_adamw_32bit*, que oferece um bom equilíbrio entre precisão e eficiência de memória. Definimos uma taxa de aprendizado inicial de 1e-3 com um scheduler cosseno, o que permite um ajuste fino do learning rate ao longo do treinamento.

Além disso, o treinamento foi realizado em FP16 (fp16=True) para aumentar a eficiência computacional sem comprometer significativamente a precisão.

4.3.4 Preparação e Formatação dos Dados

Os dados de treinamento foram cuidadosamente preparados e formatados. Utilizamos um formato específico para as entradas e saídas:

```
(<|user|> Eu estou escrevendo isso.)
(<|assistant|> E eu estou respondendo aqui.)
```

Dessa forma, o modelo irá compreender melhor a estrutura das conversas e a gerar respostas mais apropriadas.

4.3.5 Configuração do Treinamento

O treinamento foi configurado para 3 épocas, com um máximo de 1200 steps. Definimos um tamanho máximo de sequência de 2048 tokens, adequado para o modelo TinyLlama e crucial para controlar o uso de memória. O tamanho do batch por dispositivo foi ajustado para 8, com passos de acumulação de gradiente definidos em 16, efetivamente simulando um batch size maior. Estas estratégias combinadas nos permitiram treinar um modelo eficiente e otimizado, mantendo um bom equilíbrio entre desempenho, tamanho do modelo e velocidade de inferência. Para contexto, a máquina que utilizamos para treinar possui 6GB de VRAM.

4.4 Geração de atributos em CSV

Após o treinamento do modelo, salvamos os parâmetros ajustados em um arquivo. Posteriormente, esses parâmetros são integrados ao modelo base por meio de um processo de merge. Uma vez que o modelo base é combinado com os parâmetros treinados, ele passa a gerar os valores em formato CSV, como *written_text, spoken_text, e image_id*. Esses dados serão eventualmente convertidos para um formato JSON, que é utilizado para criar cartões no Livox, como ilustrado na Figura 3. Este formato JSON padronizado permite que o aplicativo se adapte dinamicamente às necessidades do usuário, melhorando a interatividade e a acessibilidade do sistema. Além disso, a estrutura do JSON facilita atualizações e manutenções futuras, garantindo que o sistema permaneça eficiente e relevante para os usuários.

Figure 3: Formato JSON

4.5 Métricas

No contexto do desenvolvimento de STMLs para a geração de pictogramas em comunicação alternativa, as métricas BLEU e TER emergem como ferramentas essenciais para a avaliação quantitativa das saídas geradas.

A métrica BLEU [9] por sua natureza, é amplamente reconhecida por sua capacidade de avaliar a qualidade das traduções produzidas por sistemas de tradução automática, comparando a saída gerada pela máquina com uma ou mais traduções de referência humanas. Esta métrica se baseia na contagem de n-gramas coincidentes e incorpora uma penalidade de brevidade para evitar pontuações elevadas em textos curtos que, apesar da correspondência nos n-gramas, não capturam adequadamente o sentido integral do texto original.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

onde:

- BP é a Penalidade de Brevidade (Brevity Penalty),
- w_n são os pesos para os diferentes n-gramas,
- p_n é a precisão dos n-gramas correspondentes entre a tradução e a referência.

A métrica TER [10] quantifica o número de edições necessárias — inserções, exclusões, substituições e trocas de palavras adjacentes — para modificar a tradução gerada de modo que esta corresponda exatamente a uma tradução de referência. Valores inferiores de TER indicam uma maior proximidade ao texto de referência, refletindo traduções de maior qualidade.

$$\text{TER} = \frac{\text{Número de edições}}{\text{Número de palavras na tradução de referência}}$$

onde:

- Número de edições inclui inserções, exclusões e substituições de palavras necessárias para alinhar a tradução gerada com a tradução de referência.

5 Resultados

A aplicação da técnica LoRA ao nosso modelo de linguagem demonstrou melhorias significativas em termos quantitativos e qualitativos. As métricas utilizadas para avaliar o desempenho do modelo ressaltam a eficácia do LoRA em aprimorar a qualidade das traduções e gerações de texto.

As métricas de BLEU indicam um aumento da pontuação média de 10% para 62%, como mostrado no gráfico 5 no cenário em que o modelo é mergeado aos parâmetros treinados do LoRA que foram salvos, destacando uma melhoria substancial na precisão. Esses valores quantitativos são complementados por uma análise das métricas de TER, onde observamos uma redução significativa de 142.6931 para 54.9780. Isso sugere que as traduções com LoRA estão muito mais próximas do texto de referência, visto que quanto menor o valor nas métricas do TER, significa que melhor é a precisão das mensagens, diminuindo os erros de tradução em quase 61.52%, como mostrado no gráfico 6.

No entanto, é crucial não se limitar apenas a uma avaliação quantitativa. Qualitativamente, os outputs gerados sem o uso do LoRA frequentemente resultam em textos que estão fora de contexto ou que não mantêm a essência do conteúdo original, como mostrado na figura 4. Isso reforça a importância de considerar a qualidade contextual e a coesão do texto, aspectos que são significativamente melhorados com a integração do LoRA.

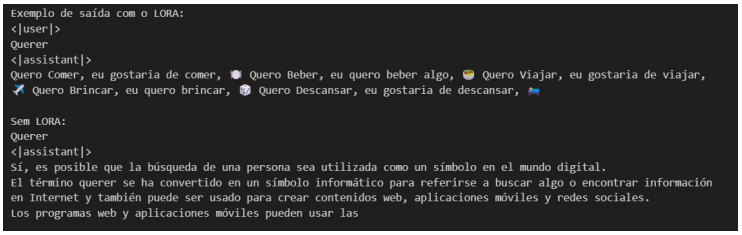

Figure 4: Comparação antes e depois dos parâmetros treinados serem mergeados

Portanto, a incorporação do LoRA não apenas melhora o desempenho do modelo conforme medido por métricas padrões, mas também garante que o conteúdo gerado seja contextualmente relevante e mais útil para aplicações práticas. Assim, o LoRA se prova uma ferramenta valiosa para o desenvolvimento de modelos de linguagem mais eficientes e eficazes, especialmente em aplicações que requerem alta fidelidade e contextualização do texto gerado.

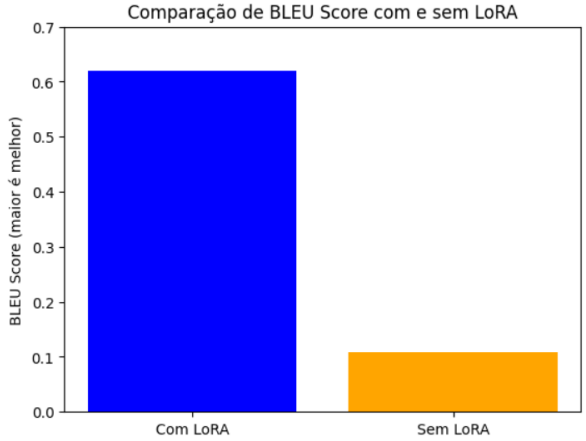

Figure 5: Métricas BLEU

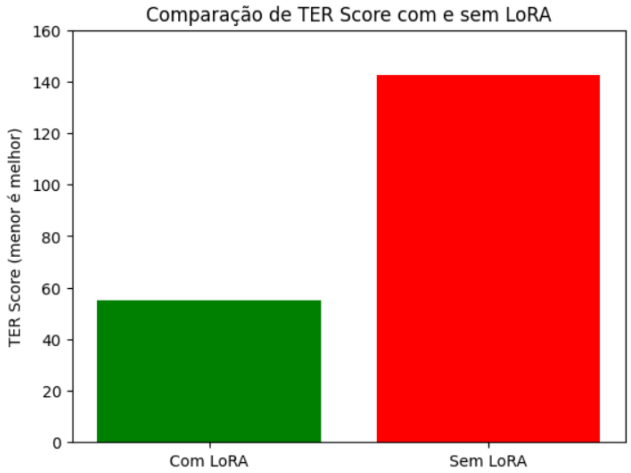

Figure 6: Métricas TER

6 Conclusão

Este trabalho validou a eficácia da técnica LoRA no contexto de um Super Tiny Language Model (STLM) para comunicação aumentativa e alternativa, destacando notáveis avanços em termos de métricas tanto quantitativas quanto qualitativas. Os resultados obtidos demonstram que a implementação de LoRA não apenas aprimora a precisão das previsões e a relevância contextual das saídas, mas também reforça a utilidade prática do modelo em aplicações reais,

tornando-o um instrumento valioso para melhorar a comunicação para usuários com necessidades especiais.

Para trabalhos futuros, uma direção promissora será aplicar o modelo de STLMs desenvolvido, que já utiliza técnicas de fine-tuning e compressão, para operar eficazmente em dispositivos Android, no aplicativo do Livox. Esta aplicação visa melhorar a geração de pictogramas para comunicação alternativa, tornando a tecnologia não apenas mais acessível, mas também mais integrada com dispositivos móveis de uso diário. Ao adaptar o modelo para funcionar em plataformas Android, poderemos explorar como as limitações de hardware influenciam a eficiência do modelo e como as otimizações podem compensar essas restrições.

Neste contexto, incluirá ajustes no modelo para garantir que ele mantenha uma performance de alta qualidade, mesmo em dispositivos com menos capacidade de processamento e armazenamento. A implementação bem-sucedida desse modelo em dispositivos Android pode significativamente expandir seu alcance e utilidade, fornecendo suporte de comunicação essencial para usuários que dependem de métodos alternativos de comunicação no seu dia a dia.

References

[1] Toloka AI. Balancing power and efficiency: The rise of small language models, 2024.

[2] Android Developers. Texttospeech. `https://developer.android.com/reference/android/speech/tts/TextToSpeech`, 2024.

[3] Jennifer B. Ganz, Margaret B. Boles, Florence D. Goodwyn, and Margaret M. Flores. Efficacy of handheld electronic visual supports to enhance vocabulary in children with asd. *Focus on Autism and Other Developmental Disabilities*, 29(1):3–12, 2014.

[4] Leon Guertler, Dylan Hillier, Palaash Agrawal, Chen Ruirui, Bobby Cheng, and Cheston Tan. Super tiny language models. *arXiv preprint arXiv:2405.14159*, 2024.

[5] Dylan Hillier, Leon Guertler, Palaash Agrawal, Chen Ruirui, Bobby Cheng, and Cheston Tan. Enhancing parameter efficiency in small language models. *arXiv preprint arXiv:2408.13296*, 2024.

[6] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zhiyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Zhiqing Sun, Lu Wei, Vivek Natarajan, and Ramesh Raskar. Lora: Low-rank adaptation of large language models, 2021. GitHub repository.

[8] Livox. Livox - tecnologia assistiva para comunicação alternativa. https://livox.com.br/br/, 2024.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[10] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas.

[11] Oregon Health & Science University. Challenges and opportunities in augmentative and alternative communication: Research and technology development to enhance communication. https://www.ohsu.edu/sites/default/files/2019-05/Challenges%20and%20opportunities%20in%20augmentative%20and%20alternative%20communication%20Research%20and%20technology%20development%20to%20enhance%20communication%20and.pdf, 2019. Accessed: date-of-access.