XXXIII Congresso de Iniciação Científica da Unicamp

Framework para visualização e análise dos padrões de ativação das redes neurais

Palavras-Chave: redes neurais, modelos de linguagem, visualização

Autores:
Luiz Felipe Corradini Rego Costa, IC - Unicamp
Prof. Dr. André Santanchè (orientador), IC - Unicamp
Prof. Dr. Luiz Celso Gomes Jr. (coorientador), DAINF - UTFPR

INTRODUÇÃO:

O campo de Processamento de Linguagem Natural, área de pesquisa que objetiva possibilitar que máquinas processem textos em linguagem natural, tem sido revolucionado desde a introdução de LLMs (Modelos de Linguagem de Grande Escala). Impulsionados pelo treinamento em volumes substanciais de dados e avanços nas técnicas de aprendizado profundo, tais modelos têm apresentado capacidades comparáveis aos humanos em tarefas baseadas em linguagem natural, sendo utilizados em uma grande gama de aplicações, desde geração de texto até assistentes virtuais e ferramentas de criação de conteúdo (Zhao et al., 2024).

Contudo, apesar da ampla aplicação, a compreensão de como esses modelos operam e como funciona o processo de decisão ainda é um desafio importante. Isso ocorre porque, modelos baseados em arquiteturas de redes neurais são usualmente considerados “caixas-pretas”, devido a relação não-linear entre entradas e saídas, combinada com uma enorme dimensionalidade dos espaços de representação dos textos.

Diante da grande capacidade desses modelos em tarefas cotidianas, diversas aplicações críticas, tais como saúde, finanças e justiça, poderiam se beneficiar da utilização de tal tecnologia. Porém, a opacidade e complexidade estrutural dessas ferramentas prejudica a adoção segura e ética em contextos mais sensíveis, onde é necessário que os sistemas sejam auditáveis, confiáveis e eticamente explicáveis.

Pensando em buscar soluções para essa questão, o campo de pesquisa da explicabilidade tem se empenhado em elaborar metodologias e soluções que sejam capazes de elucidar de maneira mais clara como esses sistemas obtêm os resultados.

Uma das possíveis abordagens dentro desse contexto são os NRAGs (do inglês Neural Region Activation Graph), que consistem em uma metodologia voltada para a conversão das ativações internas de um LLM. Essas, por sua vez, são induzidas por um corpus, em grafos que representem as conexões entre diferentes regiões do espaço multidimensional das camadas da rede.

Os NRAGs proporcionam importantes aplicações, tais como: a comparação de subgrafos associados a categorias distintas, a análise das conexões entre regiões neurais ativadas em diferentes camadas do modelo e a investigação das propriedades dos grafos gerados por diferentes LLMs quando expostos ao mesmo corpus.

Com isso, os objetivos do projeto buscaram realizar um estudo sistemático acerca de métodos que auxiliam na interpretabilidade e explicabilidade das redes neurais, com foco principal no estudo dos NRAGs. Além disso, o projeto também teve como meta principal o desenvolvimento do *toolkit open-source LLM-MRI* (Costa et al., 2024), cuja proposta é simplificar o estudo de padrões de ativação nas redes neurais de LLMs, da mesma maneira que o MRI (magnetic ressonance image) faz para os cérebros humanos.

METODOLOGIA:

Em conformidade com os objetivos, o foco principal da pesquisa foi conceber e desenvolver o *framework* LLM-MRI, com a meta de simplificar o estudo dos padrões de ativações em LLMs, baseados na arquitetura *transformers*. Por isso, o primeiro passo no desenvolvimento da biblioteca foi a definição de sua estrutura e a pipeline a ser utilizada pelo usuário final, evidenciada na *Figura 1*.

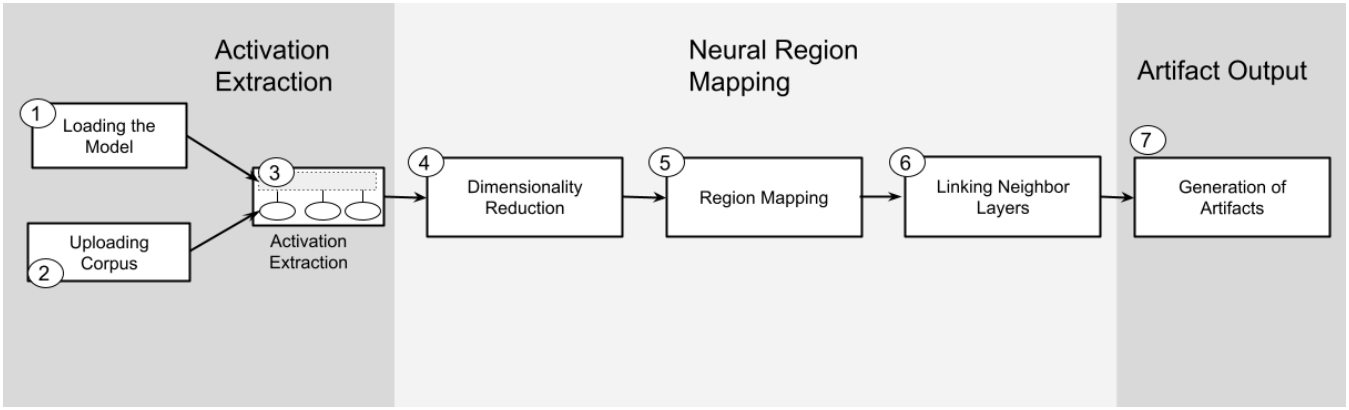


Figura 1: Diagrama representando o diagrama de funcionalidades da biblioteca

A ideia central consiste em dividir o funcionamento da biblioteca em 3 etapas: Extração das Ativações, Mapeamento das Regiões Neurais e Geração de Artefatos. Durante a primeira etapa, o modelo e o *corpus* são carregados e as ativações referentes ao estímulo do modelo pelos documentos. Nas etapas seguintes, a dimensionalidade das ativações obtidas é reduzida, de modo a consolidar as novas regiões neurais em um espaço reduzido. No próximo passo, é realizado a conexão entre as diferentes regiões, a fim de agregar os diversos mapas em um único grafo. Por fim, na sétima etapa as renderizações para os NRAGs são obtidas, de modo a permitir que o usuário tenha acesso tanto ao grafo em si, quanto à imagem que estrutura as camadas e diferentes categorias presentes no *corpus.*

Definida a pipeline e estrutura, os passos seguintes consistiram em estudar maneiras de obter as ativações e as estruturas mencionadas. Como primeira etapa, buscou-se compreender e replicar a implementação descrita no segundo capítulo do livro *Natural Language Processing with Transformers* (Turnstall et al., 2022), utilizando a base de dados *fake.BR Corpus* (Silva et al., 2020) e o modelo *distilbert-base-uncased* (Sanh et al., 2019), com um formato diferente do proposto, com quadrados ao invés de hexágonos. Essa aplicação permitiu a extração de *grids* correspondentes às ativações de cada camada induzidas pelo *corpus* passado para o modelo.

Como próximo passo, buscou-se desenvolver o mapeamento, agora da rede completa, no formato de grafo. A ideia consiste em representar cada região ativada, em cada camada, por um documento específico, em um nó distinto no grafo, e conectar nós que foram ativados pelos mesmos documentos em camadas adjacentes. O

resultado da imagem referente á criação do grafo é evidenciada na *Figura 2,* representando as ativações para base de dados contendo notícias verdadeiras e falsas. No grafo em questão, nós (representando regiões neurais) foram coloridos de acordo com a categoria que mais ativou a região específica.

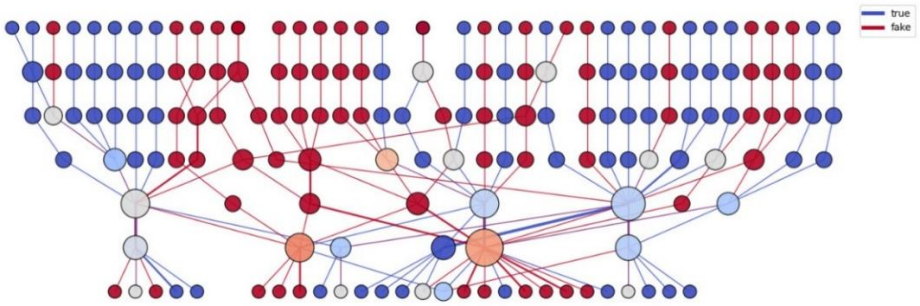


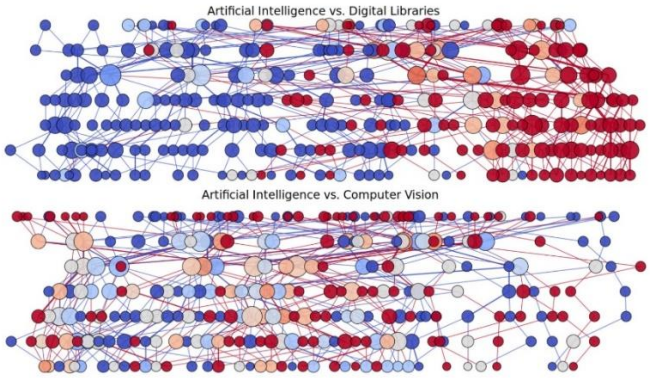
Figura 2: Visualização como imagem do grafo obtido a partir da base de dados Fake.Br Corpus e o modelo distilbert-base-uncased

Definidas as principais funções que compõem a pipeline evidenciada na *Figura 1,* foi necessário estruturar as funções definidas de modo a disponibilizá-las como *toolkit open-source*, conforme especificado nos objetivos do projeto. Por fim, com a biblioteca publicada e o código e a documentação disponibilizados no *Github*[1], o trabalho voltou-se à exploração, validação e otimização dos artefatos obtidos.

RESULTADOS E DISCUSSÃO:

Após a estruturação e publicação da versão inicial da biblioteca, foram realizados testes mais extensos, a fim de compreender com mais profundidade o seu funcionamento. A primeira abordagem utilizada consistiu em obter os grafos a partir de diferentes bases de dados e modelos, de modo a observar o seu comportamento e realizar uma comparação mais extensa das visualizações.

Para isso, foram realizados testes utilizando uma base de dados com mais de 500 mil artigos (Clement et al., 2019). No presente estudo, foram selecionados materiais publicados entre 2023 e 2024. Essa escolha se deu pelo fato de que buscou-se avaliar menos artigos e comparar os grafos gerados para três categorias, são elas: Inteligência Artificial, Visão Computacional e Bibliotecas Digitais. Isso foi feito para checar se é possível inferir correlação entre a proximidade de duas áreas de pesquisa e as regiões ativadas em comum, conforme evidenciado na *Figura 3*.



[1] https://github.com/explic-ai/LLM-MRI

Figura 3: Comparação das visualizações obtidas com artigos de Inteligência Artificial e Bibliotecas Digitais e Inteligência Artificial com Visão Computacional

Com o resultado, é possível observar no grafo de artigos relacionados à Inteligência Artificial com Visão Computacional um número considerável de interseções das regiões neurais (nós com cor mais clara), indicando que documentos das duas categorias ativaram as mesmas regiões. Porém, ao comparar artigos de Inteligência Artificial com artigos da área de Bibliotecas Digitais, por exemplo, é notável que as áreas de interseção são consideravelmente menores. Tal resultado condiz com o comportamento esperado pela biblioteca, tendo em vista que Inteligência Artificial é fortemente relacionada com a área de Visão Computacional do que com Bibliotecas Digitais, e portanto, tem uma interseção maior nas regiões ativadas dentro do modelo.

Além disso, foram realizados testes mais extensos para compreender as complexidades de espaço e de tempo de execução dos artefatos para diferentes modelos. Para a realização dos testes, a biblioteca foi testada para 3 modelos distintos, com diferentes números de parâmetros, além de utilizar diferentes números de instâncias dos datasets como *corpus* junto a cada modelo.

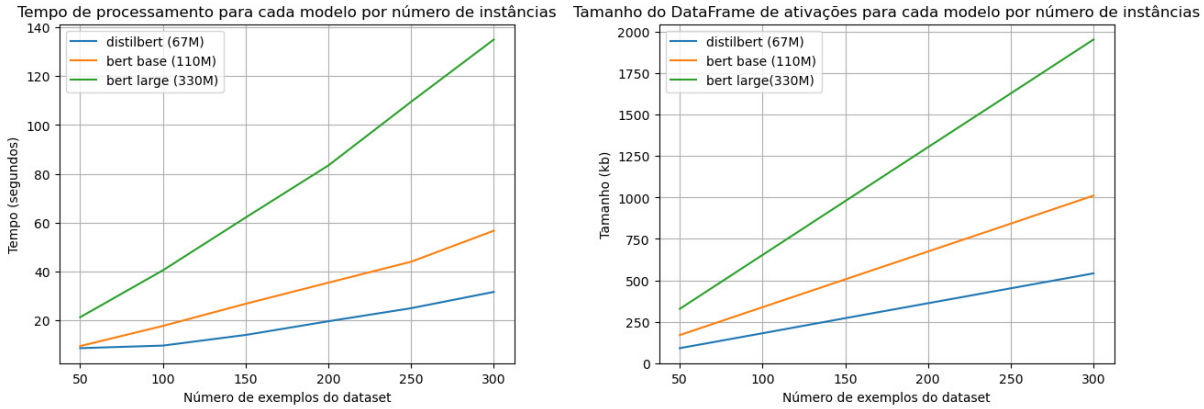


Figura 4: Gráficos correspondentes ao tempo e memória necessários para o processamento de documentos de diferentes tamanhos com 3 modelos distintos pela biblioteca LLM-MRI

A partir dos resultados, é possível notar que tanto o tempo necessário para processar todos os documentos é similar quando são utilizados poucos documentos, mas torna-se mais discrepante conforme mais instâncias são passadas para a biblioteca. Um comportamento similar é possível de ser avaliado quando analisada a quantidade de memória utilizada por cada modelo, o que é relevante de ser considerado, visto que todas as ativações são armazenadas na memória do computador. Esses comportamentos reforçam a necessidade de buscar melhor aproveitamento de GPUs nos processamentos da biblioteca.

CONCLUSÕES:

Como conclusão do projeto, espera-se que a abordagem de NRAGs seja um mecanismo relevante para explicabilidade de LLMs, se aproveitando das vastas pesquisas na área de análise de redes complexas. A biblioteca LLM-MRI busca habilitar os usuários a atingir uma compreensão mais intuitiva no tocante de como informações e padrões refletem nas redes neurais presentes dentro dos LLMs. A biblioteca ainda está em constante desenvolvimento, e tem como próximos passos permitir ao usuário obter métricas de redes complexas dentro da

biblioteca, agregadas às atuais possibilidades de visualização, além de gerar visualizações mais interativas, identificando palavras e conceitos referentes a cada região neural.

BIBLIOGRAFIA

CLEMENT, Colin B.; BIERBAUM, Matthew; O'KEEFFE, Kevin P.; ALEMI, Alexander A. **On the Use of ArXiv as a Dataset. 2019**. Disponível em: https://arxiv.org/abs/1905.00075. Acesso em: 30 Jul, 2025.

Costa, L., Figênio, M., Santanchè, A., and Gomes-Jr, L. (2024). Llm-mri python module: a brain scanner for llms. In Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados, pages 125–130, Porto Alegre, RS, Brasil. SBC.

SANH, Victor; DEBUT, Lysandre; CHAUMOND, Julien; WOLF, Thomas. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, [S.l.], 2019. Disponível em: https://arxiv.org/abs/1910.01108. Acesso em: 30 Jul, 2025.

SILVA, Renato M.; SANTOS, Roney L. S.; ALMEIDA, Tiago A.; PARDO, Thiago A. S. Towards automatically filtering fake news in Portuguese. Expert Systems with Applications, [S.l.], v. 146, p. 113199, 2020. ISSN 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.113199.

TUNSTALL, Lewis; VON WERRA, Leandro; WOLF, Thomas; et al. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, 2022. Disponível em: https://www.oreilly.com/library/view/natural-language-processing/9781098103231/. Acesso em: 30 Jul, 2025.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2024b). A survey of large language models.