

Avaliação de Modelos Transformers para Recuperação Multimodal de Informações

Leonardo O. Campos¹, Mirela T. Cazzolato¹

¹Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos, SP – Brasil

leonardocampos@usp.br, mirela@icmc.usp.br

Abstract. *In multimodal information retrieval, extracting discriminative features is essential for search systems. This research analyzes latent representations from Transformer-based architectures applied to multimodal medical information retrieval. The main idea is to explore explainable metrics and approaches for understanding content-based retrieval results, grounded in generated representations, variations in multimodal data characteristics, and similarity search operators. The current state of the study focuses on instantiating the evaluation pipeline, including curating datasets, extractors, and similarity operators. Preliminary results show consistent patterns when comparing Von Neumann Entropy values across the embeddings and the retrieval results. Currently, feature spaces generated from different datasets are being evaluated, while approaches and metrics are being investigated to help understand retrieval quality based on the characteristics represented in the feature space.*

Resumo. *Na recuperação de informação multimodal, a extração de características discriminativas é essencial para os sistemas de busca por similaridade. Esta pesquisa analisa representações latentes obtidas por arquiteturas baseadas em Transformers aplicadas à recuperação de informação médica multimodal. A principal ideia é explorar métricas e abordagens explicáveis para compreender os resultados da recuperação baseada em conteúdo, considerando as representações geradas, as variações nas características dos dados multimodais e os operadores de busca por similaridade. O estágio atual do estudo concentra-se na implementação do pipeline de avaliação, incluindo a curadoria de conjuntos de dados, extratores de características e operadores de similaridade. Resultados preliminares mostram padrões consistentes na comparação dos valores de Entropia de von Neumann entre os embeddings e os resultados de recuperação. Na continuação do estudo, estão sendo avaliados espaços gerados por diferentes conjuntos e buscadas abordagens e métricas que auxiliem no entendimento da qualidade do resultado de recuperação com base nas características representadas no espaço de características.*

Informações Gerais. **Nível:** Mestrado. **Aluno:** Leonardo O. Campos. **Orientador:** Mirela T. Cazzolato. **Programa:** Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional da Universidade de São Paulo. **Admissão:** Março 2025. **Exame de qualificação:** Abril 2026. **Defesa prevista:** Abril 2027. **Etapas concluídas:** Créditos obrigatórios em disciplinas, revisão bibliográfica, definição do problema, exame de qualificação, exame de proficiência em língua estrangeira. **Publicações relacionadas:** [Campos et al. 2026].

1. Introdução

A Recuperação de Informação é um campo de pesquisa da Computação que tem forte ligação com métodos de extração de características, o que impacta diretamente os resultados de recuperação [Zhang et al. 2024]. Historicamente, os primeiros métodos eram conhecidos como “*handcrafted*”, ou métodos de primeira ordem. Com o tempo, houve a adaptação desses métodos para que fossem capazes de trabalhar com estatística e reconhecimento de contexto, chegando até os métodos atuais, que fazem extrações por meio de modelos de Inteligência Artificial [Attallah 2025].

Dentre os métodos utilizados para extração de características mais conhecidos estão as Redes Neurais Recorrentes e suas variações, como as Long-Short-Term Memory; e as baseadas em convoluções para imagens, como as Redes Convolucionais. No entanto, com o surgimento recente da arquitetura *Transformer*, a matriz de atenção foi introduzida como nova forma de correlacionar todas as informações de entrada do modelo, melhorando a capacidade de capturar características semânticas e informações distantes na mesma representação [Vaswani et al. 2017, Xu et al. 2023, Liu and Deng 2015].

Neste contexto, este trabalho de mestrado propõe uma arquitetura de avaliação, utilizando métodos baseados na arquitetura *Transformer* como extratores de características, combinados com operadores de busca aproximada e métodos de explicabilidade. A ideia é estudar o espaço vetorial de representação, buscando entender o comportamento desse espaço com diferentes representações multimodais e suas influências nos resultados de recuperação.

Os resultados desta pesquisa visam oferecer uma análise do comportamento das arquiteturas *Transformers* no contexto da recuperação de informação multimodal. É possível utilizar as interpretações dos resultados desta pesquisa em diferentes domínios. Assim, espera-se contribuir para o avanço da literatura sobre recuperação de informação explicável em domínios multimodais.

1.1. Descrição do Problema

A literatura sobre Recuperação de Informação utilizando *Transformers* como extratores vem se desdobrando em novos trabalhos, demonstrando bons resultados em domínios unimodais e multimodais, especialmente em trabalhos que visam melhorar modelos que trabalham com alinhamento semântico como pré-treinamento. No entanto, estes estudos focam em adaptações desses modelos para fins de recuperação, e seus resultados são métricas de acurácia desses modelos junto a buscadores. Há uma lacuna de análise do espaço vetorial para entender as implicações dos diferentes tipos de arquiteturas baseadas em *Transformers* no espaço de representações criado e o que essas características implicam no resultado da busca por similaridade [Zhang et al. 2024, Xu et al. 2023, Amudhavalli 2025].

Tendo isso em vista, o levantamento da Tabela 1 categoriza os principais trabalhos do estado da arte da literatura.

É possível notar que, embora a multimodalidade seja uma característica recorrente, a maioria dos trabalhos negligencia métodos de explicabilidade ou métricas de Teoria da Informação. A ausência de análises baseadas em entropia ou de visualização do espaço impossibilita a compreensão de fenômenos como o colapso de dimensionalidade

Tabela 1. Taxonomia dos trabalhos relacionados, comparando a proposta com relação à categoria de estudo, ao domínio multimodal, à análise do espaço vetorial e o emprego da entropia na análise.

Estudo	Estudo de Avaliação	Domínio Multimodal	Análise do Espaço Vetorial	Entropia na Análise
[Kim et al. 2021]	×	✓	✓	✓
[Dubey et al. 2022]	×	×	✓	✓
[Song et al. 2023]	×	×	×	×
[Zheng et al. 2024]	×	✓	×	×
[Fiel and Sablatnig 2024]	×	×	×	×
[Jiang et al. 2025]	×	×	×	×
[Peng 2025]	×	✓	✓	✓
[Amudhavalli 2025]	✓	✓	×	×
Proposta	✓	✓	✓	✓

ou o viés de agrupamento, elementos que fundamentam a originalidade da proposta deste trabalho ao integrar a análise do espaço vetorial como métrica central de avaliação.

1.2. Solução Proposta

O objetivo deste trabalho de mestrado é analisar diferentes modelos baseados em *Transformer*, atuando como extratores, em conjunto com diferentes arquiteturas de busca por similaridade aproximada, incorporando explicabilidade no espaço vetorial para compreender os resultados da busca no domínio médico. Na explicabilidade, serão utilizadas diferentes abordagens das entropias de Shannon e de Von Neumann. Os objetivos são: 1) Analisar o comportamento de diferentes modelos baseados em *Transformers* no contexto de Recuperação de Informação. 2) Utilizar Entropia, métricas e abordagens de explicabilidade e técnicas de redução de dimensionalidade para avaliações do espaço vetorial gerado por estes modelos. 3) Avaliar diferentes abordagens de busca aproximada, para recuperação de conteúdo. 4) Comparar os resultados obtidos pelos algoritmos baseados em Transformers com os de algoritmos que não utilizam mecanismos de atenção.

2. Metodologia

O desenvolvimento desta proposta de pesquisa envolve os seguintes passos: 1) Curadoria de bases de dados (principalmente médicas) e análise exploratória. 2) Escolha e implementação de arquiteturas baseadas em *Transformer*. 3) Análise do espaço vetorial utilizando Entropia, métricas e abordagens de explicabilidade e redução de dimensionalidade para auxiliar na visualização do espaço de características. 4) Escolha e implementação de arquiteturas de busca por similaridade aproximada. 5) Implementação de métricas de avaliação. 6) Avaliação dos resultados e discussões.

No primeiro passo, foi explorada a base de dados MIMIC Chest X-ray (MIMIC-CXR), disponibilizada publicamente pelo *PhysioNet*. Trata-se de uma base

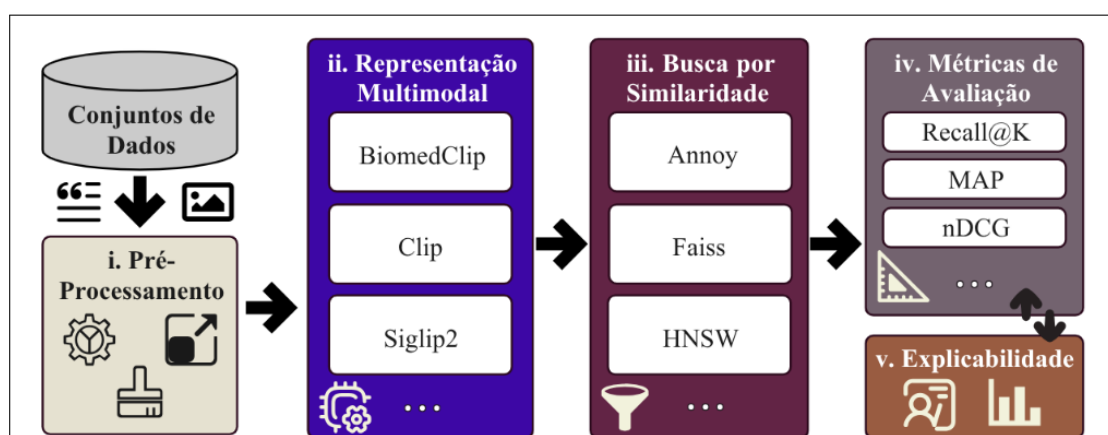


Figura 1. Arquitetura proposta para avaliar modelos baseados na arquitetura Transformer no contexto da recuperação de informação multimodal. Serão realizados os seguintes passos: (i) pré-processamento; (ii) representação dos dados utilizando modelos baseados em Transformers; (iii) consultas por similaridade; validação experimental por meio de (iv) métricas de avaliação e (v) abordagens de explicabilidade.

de dados médica multimodal de larga escala, utilizada em pesquisas em IA médica [Johnson et al. 2024]. Em sua versão 2.1.0, a base de dados é composta por 377.110 imagens de radiografias de tórax associadas a 227.835 exames radiográficos realizados no *Beth Israel Deaconess Medical Center*, em Boston, no período de 2011 a 2016. Os exames são fornecidos no formato DICOM e vêm acompanhados de seus respectivos laudos radiológicos em texto livre, de forma totalmente anonimizada, permitindo o desenvolvimento de arquiteturas de RI rotulada, o que é o objetivo deste trabalho. Novas bases de dados serão incorporadas ao projeto ao longo do desenvolvimento.

Para o passo 2, foram inicialmente escolhidos os seguintes modelos: ALBEF, BiomedCLIP, BLIP, CLIP, OmniVec2, PubMedCLIP e SigLIP 2. Já no passo 3, será utilizada, inicialmente, a Entropia de Von Neumann, para encontrar o nível de dispersão de informação no espaço vetorial, junto à redução de dimensionalidade, para facilitar a visualização destes resultados.

No passo 4, serão utilizados os algoritmos Annoy, Faiss e HNSW, dadas as suas diferenças nos trabalhos sobre busca aproximada. Também será considerada a aplicação do operador k NN, utilizando assim a busca na avaliação dos espaços gerados. No passo 5, serão utilizadas as métricas de Revocação, NDCG, MRR e Map para capturar todas as perspectivas sobre o ranqueamento gerado pela recuperação, observando qual a taxa de acertos por ponderação ou por média.

Por fim, serão realizadas análises e discussões dos resultados, atribuindo a cada um uma explicação do comportamento dos resultados de recuperação em relação aos resultados das análises do espaço vetorial e às bases multimodais utilizadas.

3. Resultados Preliminares

Em experimentos preliminares, utilizou-se a base de dados ChestX-ray8 [Wang et al. 2017], um repositório público de escala hospitalar, amplamente utilizado na literatura para tarefas de visão computacional em saúde. O conjunto de dados

é composto por 112.120 imagens de radiografia de tórax em visão frontal, provenientes de 30.805 pacientes distintos. Além das imagens, a base de dados fornece anotações textuais e dados tabulares que categorizam a presença de oito patologias torácicas comuns: Atelectasia, Cardiomegalia, Efusão, Infiltração, Massa, Nódulo, Pneumonia e Pneumotórax. Esses rótulos foram extraídos automaticamente dos relatórios radiológicos originais por meio de técnicas de mineração de texto e de processamento de linguagem natural (PLN).

No contexto deste trabalho, esses dados foram preparados para o ajuste dos modelos baseados em *Transformers* BiomedCLIP, CLIP, PubMedCLIP e SigLIP2. Após as etapas de treinamento e alinhamento semântico, os pesos dos modelos foram atualizados para a extração de *embeddings* a partir do conjunto de avaliação (dados não vistos durante o treinamento).

A partir dos vetores obtidos, aplicou-se a Entropia de Von Neumann normalizada com o objetivo de quantificar o nível de utilização e a dispersão do espaço vetorial, permitindo correlacionar essa métrica com o desempenho dos operadores de busca. No contexto de RI, é melhor que as representações tenham características discriminativas que as diferenciem no momento em que o operador aplica o cálculo da distância. Ou seja, valores de entropia mais próximos de um indicam que os vetores apontam para posições diferentes no espaço geométrico, enquanto valores próximos de zero indicam que os documentos compartilham muitas características, tornando difícil a distinção entre os objetos.

Os valores de entropia revelam-se baixos para todas as arquiteturas, situando-se abaixo de 0,22. Nota-se que, mesmo os modelos pré-treinados em dados médicos, como o BiomedCLIP (0,212) e o PubMedCLIP (0,077), mantiveram um comportamento de baixa entropia, similar àqueles expostos estritamente a dados de domínio aberto, como o SigLIP2 (0,149) e o CLIP (0,056). Esse cenário sugere uma subutilização das dimensões do espaço latente (também conhecido como colapso dimensional) na base de dados ChestX-ray8.

Após a estruturação do espaço vetorial, os indexadores Annoy e HNSW foram empregados para conduzir o processo de recuperação de informação multimodal. Os resultados obtidos por esses operadores e suas respectivas métricas de avaliação estão consolidados na Tabela 2. As métricas utilizadas foram: NDCG, que mede a qualidade do *ranking* ponderando a relevância graduada dos itens em função de sua posição na lista; e MRR, que avalia o sistema com base no inverso da posição do primeiro item relevante recuperado. A escolha dessas métricas justifica-se pela necessidade de avaliar não apenas a capacidade dos indexadores de encontrar os documentos corretos, mas também a precisão na ordenação dos resultados mais prioritários no topo da lista recuperada.

A análise das métricas indica que o nível de assertividade na recuperação ainda é incipiente para a base de dados ChestX-ray8. Esse desempenho modesto decorre do fato de a base de dados não disponibilizar os relatórios clínicos textuais na íntegra, limitando-se a rótulos tabulares estruturados. Essa característica reduz a eficácia do alinhamento semântico latente. Modelos do tipo *Dual-Encoder*, como os avaliados nesta pesquisa, demandam descrições textuais ricas em contexto para estabelecer correlações robustas com os atributos visuais das imagens. A hipótese é que a conversão das variáveis tabulares em estruturas sintéticas frasais não forneceu o contexto semântico necessário para um

Tabela 2. Médias de NDCG e MRR para consultas de imagem com $k = 5, 10$ e 15 .

Alg.	Modelo	NDCG@5	NDCG@10	NDCG@15	MRR@5	MRR@10	MRR@15
Annoy	biomedclip	0,119	0,170	0,204	0,291	0,292	0,293
Annoy	clip	0,234	0,231	0,230	0,317	0,328	0,331
Annoy	pubmedclip	0,160	0,190	0,219	0,236	0,236	0,237
Annoy	siglip2	0,126	0,177	0,209	0,306	0,309	0,309
HNSW	biomedclip	0,018	0,093	0,124	0,032	0,085	0,089
HNSW	clip	0,216	0,221	0,221	0,297	0,312	0,315
HNSW	pubmedclip	0,058	0,170	0,167	0,101	0,101	0,104
HNSW	siglip2	0,371	0,321	0,302	0,371	0,377	0,379

alinhamento otimizado. Diante dessa limitação, estendeu-se a investigação ao repositório MIMIC Chest X-Ray (MIMIC-CXR), que inclui relatórios clínicos textuais completos e semiestruturados. Também estão sendo avaliadas bases de dados multimodais de outros domínios, complementando assim o estudo experimental proposto.

Por fim, pontua-se que o estudo [Campos et al. 2026], publicado até o momento, abordou a representação de imagens utilizando Transformers e o impacto de dados faltantes na recuperação por conteúdo. O trabalho foi desenvolvido com colegas do grupo de pesquisa, e relaciona-se marginalmente ao mestrado do aluno, uma vez que envolveu a representação e o estudo do espaço de representações dentro de diferentes condições.

Agradecimentos

O trabalho possui o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código de Financiamento 001, da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP – proc. #2025/20457-2, #2024/13328-9, #2024/15430-5, #2026/11130-2), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq–proc. #133021/2025-4), e da Universidade de São Paulo (PRPI 1032-proc. #207).

Declaração de uso de IA generativa

A ferramenta de IA generativa Gemini foi utilizada para revisão linguística e para auxiliar na estruturação de texto.

Referências

- Amudhavalli, L. (2025). Exploring transformer-based architectures for large-scale multi-modal information retrieval systems. *International Journal of Computer Science and Engineering Innovations*, 1(1):18–26. DOI: 10.64137/31079458/IJCSEI-V1I1P103.
- Attallah, O. (2025). Multi-domain feature incorporation of lightweight convolutional neural networks and handcrafted features for lung and colon cancer diagnosis. *Technologies*, 13(5). DOI: 10.3390/technologies13050173.
- Campos, L. O., Favoreto, G. S., Traina-Jr., C., Traina, A. J. M., and Cazzolato, M. T. (2026). Incomplete information retrieval without imputation: Exploiting the correlation among deep features. In *Proceedings of the 39th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*. To appear.

- Dubey, S. R., Singh, S. K., and Chu, W.-T. (2022). Vision Transformer Hashing for Image Retrieval. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. DOI: 10.1109/ICME52920.2022.9859900.
- Fiel, S. and Sablatnig, R. (2024). Self-supervised vision transformers for writer retrieval. In *Document Analysis and Recognition - ICDAR 2024*, pages 380–396. Springer. DOI: 10.1007/978-3-031-70536-6_23.
- Jiang, X., Tang, H., Pan, Y., and Li, Z. (2025). Rethinking vision transformer for large-scale fine-grained image retrieval. *arXiv preprint arXiv:2504.16691*, 18(9). DOI: 10.48550/arXiv.2504.16691.
- Johnson, A., Pollard, T., Mark, R., Berkowitz, S., and Horng, S. (2024). MIMIC-CXR Database. *PhysioNet*. Version 2.1.0.
- Kim, W., Son, B., and Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. In *Proc. of the International Conference on Machine Learning*, volume 139 of *PMLR*. PMLR. DOI: <https://doi.org/10.48550/arXiv.2102.03334>.
- Liu, S. and Deng, W. (2015). Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734. IEEE. DOI: 10.1109/ACPR.2015.7486599.
- Peng, Y. (2025). A clip-based cross-modal matching model for image-text retrieval. *Information Technology and Control*, 54(3):1030–1048.
- Song, C. H., Yoon, J., Choi, S., and Avrithis, Y. (2023). Boosting vision transformers for image retrieval. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 107–117. DOI: 10.48550/arXiv.2210.11909.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Curran Associates, Inc. DOI: 10.48550/arXiv.1706.03762.
- Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106. DOI: 10.1109/CVPR.2017.369.
- Xu, P., Zhu, X., and Clifton, D. A. (2023). Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11993–12013. DOI: 10.1109/TPAMI.2023.3274228.
- Zhang, Z., Li, J., Zhu, L., Wang, T., and Shen, H. T. (2024). Cross-modal retrieval: A review of methodologies, datasets, and future perspectives. *IEEE Transactions on Multimedia*. DOI: 10.1109/ACCESS.2024.3444817.
- Zheng, J., Liang, M., Yu, Y., Li, Y., and Xue, Z. (2024). Knowledge graph enhanced multimodal transformer for image-text retrieval. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 70–82. IEEE. DOI: 10.1109/ICDE60146.2024.00013.