Integração do Contexto Clínico e Imagens de Raios-X do Tórax para Geração Automática de Laudos Radiológicos

Hériclys S. Borges[1], Pablo de A. Vieira[4], Flávio H. D. Araújo[1,2,3]
Antonio O. Carvalho Filho[1,2,3], Lilian R. G. Silva[5], Romuere R. V. e Silva[1,2,3]

[1]Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal do Piauí (UFPI) – Teresina – Brasil

[2]Programa de Pós-Graduação em Ciência da Computação
Universidade Federal do Piauí (UFPI) – Teresina – Brasil

[3]Sistemas de Informação - Campus Senador Helvídio Nunes de Barros
Universidade Federal do Piauí (UFPI) – Picos – Brasil

[4]Sistemas de Informação - Universidade Federal de Rondonópolis
Rondonópolis – Brasil

[5]Campus Valença - Instituto Federal do Piauí
Valença do Piauí – Brasil

hericlysdlarii@ufpi.edu.br

Abstract. *Automatic radiology report generation from chest X-rays has emerged as a promising approach to support clinicians and reduce the workload associated with medical image interpretation. This paper proposes a multimodal Transformer-based framework that integrates radiographic images (frontal and lateral views) with patient clinical history. Visual features are extracted using a ResNet-50 backbone with progressive fine-tuning, while clinical context is encoded using Bio_ClinicalBERT. These multimodal representations are fused within an encoder–decoder Transformer architecture to generate radiology reports autoregressively. Experiments conducted on the MIMIC-CXR dataset demonstrate that the proposed model is capable of producing structured and clinically coherent reports, achieving competitive performance in semantic similarity metrics.*

Resumo. *A geração automática de laudos radiológicos a partir de radiografias de tórax tem se destacado como uma estratégia para apoiar médicos e reduzir a carga de trabalho na interpretação de imagens. Este artigo propõe uma abordagem multimodal baseada em Transformers que integra imagens (frontal e lateral) e histórico clínico do paciente. As características visuais são extraídas com uma ResNet-50 com ajuste fino progressivo, enquanto o contexto clínico é codificado com Bio_ClinicalBERT. As representações são fundidas em uma arquitetura Transformer codificador-decodificador para geração autorregressiva dos laudos. Experimentos no MIMIC-CXR mostram que o modelo produz*

laudos estruturados e clinicamente coerentes, com desempenho competitivo em métricas de similaridade semântica.

1. Introdução

Entre todas as modalidades de imagem, o raio-X é a modalidade mais comum, rápida e barata usada para diagnosticar muitos distúrbios do corpo humano [Akhter et al. 2023]. A interpretação de imagens de raio-X requer experiência e conhecimento especializado, podendo estar sujeita a variações entre diferentes observadores. Além disso, a crescente demanda por exames de imagem gera desafios na análise e relato dos achados, impactando diretamente a eficiência dos serviços de saúde [Litjens et al. 2017, Bruno et al. 2015].

Muitos modelos utilizam apenas informações visuais das radiografias e empregam arquiteturas complexas, o que aumenta o custo computacional e dificulta a reprodutibilidade. Este trabalho apresenta um modelo para geração automática de laudos radiológicos baseado em *Transformers* [Vaswani et al. 2017], que integra características visuais de radiografias de tórax com informações de histórico clínico do paciente.

A abordagem utiliza uma representação híbrida global-local das imagens obtida pela rede neural convolucional *ResNet-50* [He et al. 2016] com *fine-tuning*, combinada à incorporação de dados clínicos processados pelo *Bio_ClinicalBERT* [Alsentzer et al. 2019]. O *decoder* emprega *embeddings BERT* [Devlin et al. 2019] pré-treinados para aprimorar a qualidade linguística dos laudos, resultando em uma arquitetura que combina eficientemente informações visuais e textuais para gerar descrições médicas precisas e contextualizadas.

Como contribuição, este trabalho integra imagens radiográficas e o histórico clínico do paciente para a geração automática de laudos. A abordagem proposta emprega uma estratégia híbrida de representação visual, combinando *tokens* visuais globais e locais extraídos de uma rede *ResNet-50* otimizada. Para capturar a semântica médica presente no histórico do paciente, o modelo incorpora o contexto clínico por meio do *Bio_ClinicalBERT*, alinhando essas informações com os achados visuais. A avaliação experimental realizada no conjunto de dados MIMIC-CXR demonstra a viabilidade da integração de informações multimodais para a geração de laudos radiológicos.

2. Trabalhos Relacionados

A geração automática de laudos radiológicos tem recebido crescente atenção da comunidade científica, com diversos trabalhos propondo abordagens inovadoras para este desafio. [Chen et al. 2020] propuseram o *Memory-driven Transformer* (R2Gen), uma arquitetura que incorpora memória relacional ao decodificador Transformer por meio de uma normalização condicional de camada. Esta abordagem permitiu capturar padrões recorrentes em laudos radiológicos, resultando em melhorias significativas nas métricas de avaliação nos *datasets* IU X-Ray e MIMIC-CXR.

Posteriormente, [Chen et al. 2021] estenderam este trabalho propondo as *Cross-modal Memory Networks* (CMN), que introduzem uma matriz de memória compartilhada para alinhar informações visuais e textuais. Por meio de mecanismos de consulta e resposta à memória, o modelo estabelece correspondências explícitas entre regiões das imagens e termos dos laudos, melhorando a acurácia clínica das descrições geradas.

Mais recentemente, [Nicolson et al. 2022] investigaram o impacto da inicialização com pesos pré-treinados para geração de laudos de raios-X de tórax. Seu estudo comparou diferentes *checkpoints* de visão computacional e processamento de linguagem natural, concluindo que a combinação CvT-21 com DistilGPT2 alcança os melhores resultados, superando abordagens anteriores em métricas de geração de linguagem e eficácia clínica.

Paralelamente, [Yang et al. 2022] propuseram uma abordagem baseada em conhecimento para geração de laudos radiológicos, distinguindo entre conhecimento geral e conhecimento específico. O conhecimento geral, extraído do RadGraph, fornece informações médicas amplas e independentes da imagem de entrada. O conhecimento específico, por sua vez, é obtido por meio da recuperação de relatórios similares com base na distribuição de rótulos da imagem atual, fornecendo conhecimento contextualizado e personalizado. Para integrar estas informações, os autores desenvolveram um mecanismo de atenção *multi-head* melhorado por conhecimento, que combina as características visuais com as informações estruturais do grafo de conhecimento. Os resultados experimentais demonstraram que ambos os tipos de conhecimento contribuem significativamente para a qualidade dos laudos gerados.

Em contraste, a abordagem proposta neste trabalho integra características visuais de imagens de raios-X de tórax com informações contextuais derivadas do histórico clínico do paciente, utilizando representações *Bio_ClinicalBERT*. Essa integração multimodal permite que o modelo capture o contexto clinicamente relevante, mantendo uma arquitetura baseada em transformadores relativamente simples e reprodutível.

3. Metodologia

A metodologia proposta para o desenvolvimento do sistema de geração automática de laudos radiológicos é estruturada em seis etapas principais: aquisição de dados; pré-processamento; extrator de características visuais; codificação do histórico clínico; arquitetura *Transformer Encoder-Decoder*; e treinamento e testes. A Figura 1 ilustra a arquitetura do modelo proposto.

3.1. Aquisição dos Dados

A etapa inicial do projeto consistiu na aquisição do conjunto de dados *MIMIC-CXR-JPG v2.0.0* [Johnson et al. 2019], um dos maiores bancos públicos de radiografias de tórax disponíveis para pesquisa em inteligência artificial aplicada à saúde. O *dataset* contém 377.110 imagens de raio-X associadas a 227.827 laudos radiológicos, derivados de exames realizados no *Beth Israel Deaconess Medical Center*. As imagens estão disponíveis em formato JPG, derivadas dos exames originais em *DICOM*, e são acompanhadas por relatórios textuais que descrevem os achados radiológicos. Além disso, o *dataset* inclui diferentes vistas radiográficas, principalmente projeções frontais (PA/AP) e laterais, que foram utilizadas neste trabalho para enriquecer a representação visual dos exames.

O conjunto de dados consiste em imagens de raios-X de tórax e seus respectivos laudos radiológicos extraídos de registros eletrônicos de saúde. As imagens incluem vistas frontais (póstero-anterior ou ântero-posterior) e laterais. A Figura 2 mostra um exemplo de imagens de raios-X de tórax associadas a um laudo radiológico. As imagens ilustram exames radiográficos típicos disponíveis no conjunto de dados. O laudo radiológico correspondente às imagens apresentadas na Figura 2 é mostrado na Tabela 1.

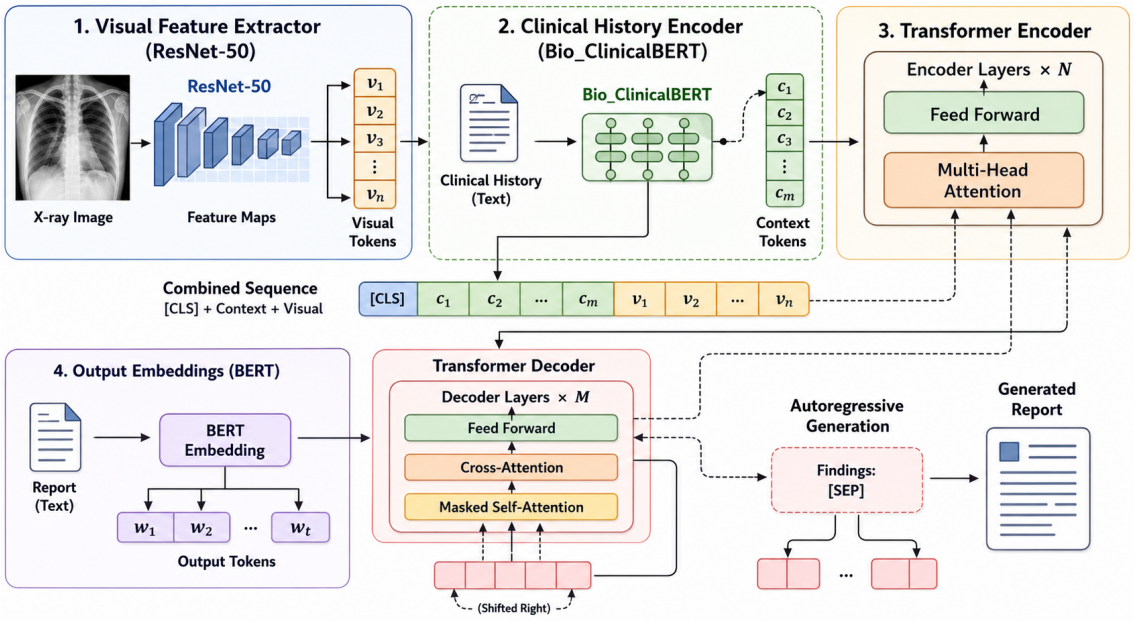

Figura 1. Arquitetura proposta para geração automática de laudos radiológicos.

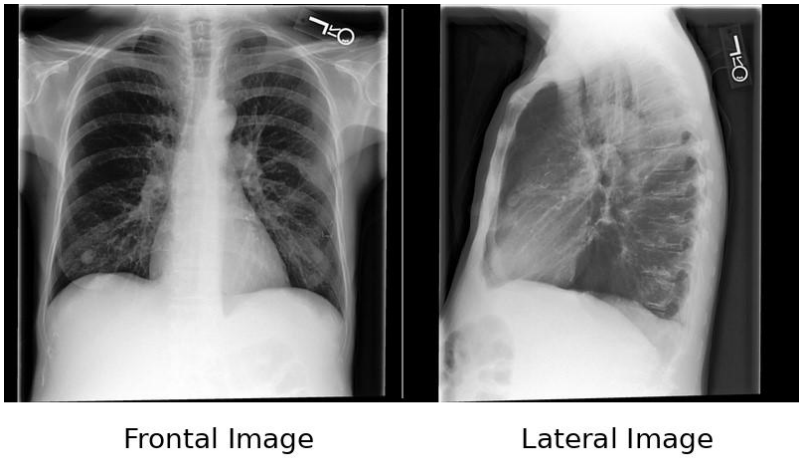

Figura 2. Exemplo de imagens radiográficas utilizadas no *dataset*, contendo as projeções frontal e lateral do tórax.

Tabela 1. O laudo radiológico correspondente às imagens apresentadas na Figura 2.

INDICATION:	New onset ascites; eval for infection.
FINDINGS:	There is no focal consolidation, pleural effusion or pneumothorax. Bilateral nodular opacities that most likely represent nipple shadows. The cardiomediastinal silhouette is normal. Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs are noted.
IMPRESSION:	No acute cardiopulmonary process.

O MIMIC-CXR-JPG apresenta desafios substanciais para modelagem computacional devido ao desbalanceamento entre classes, à ausência de imagens laterais para 69,8% dos pacientes, à variabilidade estrutural e linguística dos laudos radiológicos, bem como à presença de ruído nos rótulos derivados automaticamente. Adicionalmente, a heterogeneidade nos protocolos de aquisição e a presença de dispositivos médicos introduzem vieses potenciais e correlações espúrias. Tais fatores aumentam a complexidade do treinamento, dificultam a generalização e exigem estratégias arquiteturais e metodológicas específicas para mitigação desses problemas.

3.2. Pré-processamento

Para compatibilidade com a arquitetura *ResNet-50* pré-treinada no *ImageNet* [Deng et al. 2009], as imagens de raio-X passaram por um pré-processamento específico que inclui redimensionamento e normalização. Inicialmente, todas as imagens foram redimensionadas para 224×224 *pixels* utilizando interpolação bilinear, garantindo a dimensão esperada pela rede convolucional.

Em seguida, aplicou-se a normalização com os valores de média e desvio padrão originalmente utilizados no treinamento do *ImageNet*, respectivamente *média* = [0.485, 0.456, 0.406] e *desvio padrão* = [0.229, 0.224, 0.225], procedimento essencial para manter a compatibilidade com os pesos pré-treinados e assegurar que os gradientes durante o *fine-tuning* operem em uma escala adequada, facilitando a adaptação ao domínio radiológico sem comprometer as características visuais previamente aprendidas.

Para enriquecer a diversidade dos dados de treinamento e mitigar o risco de *overfitting*, foram aplicadas técnicas de aumento de dados durante o processo de treinamento do modelo. Utilizou-se *flip* horizontal; rotação da imagem em até ±10 graus; zoom entre 91% e 109%; translação horizontal de até ±8% de largura e vertical até ±10 de altura; ajuste de contraste; ajuste de matiz; e adição de ruído gaussiano entre 0,0 e 0,1. Todas essas transformações foram usadas com probabilidade de 50%.

O processamento textual dos laudos e históricos clínicos foi realizado utilizando o tokenizador do modelo *Bio_ClinicalBERT*, assegurando consistência com o vocabulário médico especializado composto por 28.996 *tokens*, incluindo *tokens* especiais. Antes da *tokenização*, os textos passaram por etapas de limpeza e normalização que incluíram conversão para letras minúsculas, remoção de caracteres especiais não alfanuméricos com preservação de pontuação básica, substituição de múltiplos espaços por um único espaço e exclusão de laudos vazios ou com menos de 5 *tokens*.

Para garantir dimensionalidade uniforme no processamento, adotou-se comprimento máximo de sequência de 768 *tokens* (contemplando os *tokens* [CLS] e [SEP]), com truncamento para laudos mais longos e preenchimento com o *token* [PAD] para os mais curtos, viabilizando o treinamento em *batches* do modelo *transformer*. O histórico clínico do paciente é extraído das seções "indicação", "comparação"e "histórico"do laudo original, quando disponíveis, e passa por um pré-processamento específico para integração ao modelo. Utiliza-se o mesmo tokenizador do *Bio_ClinicalBERT*, com comprimento máximo limitado a 128 *tokens* para padronização das entradas.

3.3. Extrator de Características Visuais

O módulo extrator de características visuais é baseado na arquitetura ResNet-50, pré-treinada no *ImageNet* e submetida a uma estratégia de *fine-tuning* controlado para adaptação ao domínio radiológico. Este componente é responsável por processar as imagens de raio-X e convertê-las em uma representação híbrida global-local, que combina um *token* global obtido pelo *flatten* completo do mapa de características com *tokens* espaciais representando regiões específicas da imagem. Essa abordagem possibilita que o modelo capture tanto informações contextuais amplas quanto detalhes localizados relevantes para a identificação de achados clínicos, preservando o conhecimento prévio da rede enquanto a ajusta para as particularidades das imagens radiológicas.

3.4. Codificação do Histórico Clínico

Além das informações visuais, o modelo também incorpora informações textuais provenientes do histórico clínico do paciente. Esses textos foram processados utilizando o modelo *Bio_ClinicalBERT*, um modelo de linguagem pré-treinado em textos biomédicos. O codificador transforma o histórico clínico em um vetor de contexto que representa informações relevantes sobre sintomas, condições prévias e indicações clínicas do exame. Esse vetor é posteriormente projetado no mesmo espaço latente das representações visuais, viabilizando a integração multimodal das informações.

3.5. Transformer Encoder-Decoder

O *Transformer Encoder* atua como o componente central de processamento e fusão de informações, recebendo a sequência combinada de *tokens* que inclui tanto as representações visuais híbridas quanto o vetor de contexto clínico projetado. Por meio de múltiplas camadas de atenção *multi-head* e redes *feed-forward*, este módulo estabelece dependências de longo alcance entre os diferentes elementos da sequência, possibilitando que o modelo relacione achados visuais específicos com informações do histórico do paciente. A arquitetura do *encoder* é fundamental para criar uma representação integrada e contextualizada que servirá como base para a geração do laudo textual.

O *Transformer Decoder* é responsável pela geração autoregressiva do laudo radiológico, palavra por palavra, utilizando *embeddings BERT* pré-treinados que fornecem representações linguísticas ricas para o vocabulário médico especializado. Este módulo incorpora mecanismos de atenção cruzada que possibilitam consultar dinamicamente as representações produzidas pelo *encoder* durante o processo de decodificação, garantindo que cada *token* gerado seja semanticamente alinhado com as características visuais e clínicas da entrada. O uso de *embeddings* pré-treinados acelera a convergência do modelo e melhora a qualidade linguística dos laudos produzidos, resultando em textos coerentes, fluentes e terminologicamente precisos.

3.6. Treinamento

O modelo foi treinado utilizando o otimizador *AdamW* [Loshchilov and Hutter 2019], com regularização por *label smoothing* e um *scheduler* de taxa de aprendizado baseado em *ReduceLROnPlateau*. Durante o treinamento, o objetivo foi minimizar a função de perda baseada em entropia cruzada entre os *tokens* previstos e os *tokens* de referência dos laudos.

Tabela 2. Hiperparâmetros do modelo.

Parâmetro	Valor
Dimensão dos embeddings	768
Número de cabeças de atenção	8
Taxa de dropout	0.3
Comprimento máximo da sequência	768
Tamanho do vocabulário	28996
Número de épocas	1000
Critério de parada	*early stopping*
Otimizador	AdamW
Função de perda	Cross-entropy
Taxa de aprendizado inicial	0,00001
Decaimento de pesos	0.01
Suavização de rótulos	0.1

Durante a etapa de inferência, a geração do texto é realizada utilizando a estratégia de *Beam Search* [Sutskever et al. 2014], que mantém múltiplas hipóteses de sequências candidatas ao longo do processo de decodificação. Diferentemente do método *greedy*, que seleciona apenas o token de maior probabilidade em cada passo, o *Beam Search* preserva as k sequências mais prováveis com base na probabilidade acumulada. A cada nova etapa de geração, essas hipóteses são expandidas e reavaliadas, possibilitando que o modelo explore diferentes combinações de *tokens* antes de selecionar a sequência final com maior probabilidade global. Essa estratégia contribui para a produção de laudos mais coerentes e semanticamente consistentes.

4. Experimentos

Nossos experimentos foram conduzidos utilizando o conjunto de dados MIMIC-CXR, com 227.835 imagens pareadas a 65.379 laudos textuais. Devido ao desbalanceamento do conjunto de dados, os experimentos foram conduzidos com uma partição de 7.020 amostras para treinamento, 2.006 para validação e 1.003 para teste.

As configurações adotadas abrangem desde a otimização e regularização até estratégias específicas para *fine-tuning* da *ResNet-50* e do *Bio_ClinicalBERT*, buscando equilibrar a capacidade de aprendizado com a prevenção de *overfitting*, considerando a natureza multimodal da tarefa e o tamanho do conjunto de dados disponível. A Tabela 2 sumariza os principais hiperparâmetros utilizados no treinamento do modelo proposto.

A configuração de hiperparâmetros utilizada neste trabalho segue práticas comumente adotadas em modelos de geração de texto baseados em *Transformers* [Vaswani et al. 2017, Devlin et al. 2019]. A dimensão de incorporação foi definida como 768 para corresponder ao espaço de representação do Bio_ClinicalBERT [Alsentzer et al. 2019]. O modelo emprega 8 cabeças de atenção e regularização *dropout* com uma taxa de 0,3 para mitigar o sobreajuste, consistente com configurações comumente usadas em arquiteturas *Transformers*.

O treinamento foi realizado utilizando o otimizador *AdamW* com uma taxa de aprendizado inicial de 1×10^{-5}, uma configuração frequentemente adotada em cenários de

ajuste fino envolvendo modelos de linguagem pré-treinados [Loshchilov and Hutter 2019, Devlin et al. 2019]. Para melhorar a generalização do modelo, foi aplicada uma suavização de rótulos com um fator de 0,1, seguindo a estratégia introduzida na arquitetura original do *Transformer*. Durante a inferência, os laudos radiológicos foram gerados utilizando busca em feixe com tamanho igual a 5, uma estratégia de decodificação amplamente utilizada em tarefas de legendagem de imagens e geração de texto neural [Vinyals et al. 2015, Chen et al. 2020].

4.1. Métricas de Avaliação

A métrica BLEU (*Bilingual Evaluation Understudy*) mede a precisão de *n-gramas* (por exemplo, unigramas, bigramas e trigramas) nas legendas geradas em comparação com as legendas de referência. Ela avalia o quão bem o texto gerado pelo modelo se alinha com as legendas anotadas por humanos, calculando a proporção de *n-gramas* correspondentes entre o texto gerado e as referências, enquanto penaliza saídas excessivamente curtas. As pontuações BLEU variam de 0 a 1, sendo que valores mais altos indicam melhor alinhamento entre a legenda gerada e as legendas de referência. Formalmente, a métrica BLEU pode ser definida de acordo com a Equação1.

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

onde, p_n representa a precisão modificada de *n-gramas*, w_n é o peso atribuído a cada ordem de *n-grama* (geralmente uniforme, de modo que $\sum w_n = 1$), e BP é o *brevity penalty*, responsável por penalizar sentenças geradas que sejam muito curtas em relação às referências. O *brevity penalty* é definido pela Equação 2.

$$BP = \begin{cases} 1, & \text{se } c > r \\ e^{(1-r/c)}, & \text{se } c \leq r \end{cases} \tag{2}$$

onde, c e r são, respectivamente, o comprimento da sentença gerada e de referência.

A métrica ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) foca na revocação (*recall*), avaliando a sobreposição de n-gramas entre as legendas geradas pelo modelo e as legendas de referência. Essa métrica enfatiza a capacidade do modelo em capturar palavras e frases relevantes presentes nas anotações de verdade fundamental.

Entre suas variações mais utilizadas estão o ROUGE-N, que mede a sobreposição de n-gramas, e o ROUGE-L, que considera a maior subsequência comum (*Longest Common Subsequence* – LCS) entre os textos comparados, possibilitando avaliar similaridades estruturais entre as sentenças. A formulação geral do ROUGE-N é dada pela Equação 3.

$$ROUGE\text{-}N = \frac{\sum_{S \in \{\text{Referências}\}} \sum_{\text{gram}_n \in S} \text{Count}_{match}(\text{gram}_n)}{\sum_{S \in \{\text{Referências}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}, \tag{3}$$

onde, $\text{Count}_{match}(\text{gram}_n)$ representa o número de n-gramas que aparecem tanto na legenda gerada quanto na legenda de referência, e $\text{Count}(\text{gram}_n)$ representa o número total de n-gramas presentes nas legendas de referência.

Valores mais altos de ROUGE indicam maior similaridade entre os textos gerados e os textos de referência, sugerindo que as legendas produzidas pelo modelo capturam de forma mais abrangente as informações relevantes presentes nas anotações originais.

A métrica METEOR (*Metric for Evaluation of Translation with Explicit ORdering*) foi originalmente desenvolvida para avaliação de tradução automática, mas tem se mostrado eficaz na avaliação de laudos radiológicos por considerar aspectos linguísticos mais sofisticados que o BLEU. Diferentemente de métricas baseadas puramente em precisão de *n-grams*, o METEOR estabelece correspondência entre *unigrams* considerando não apenas *matching* exato, mas também sinônimos, *stemming* e paráfrases, o que resulta em melhor correlação com o julgamento humano. Assim, a pontuação do METEOR tende a refletir de forma mais fiel a similaridade semântica entre o laudo gerado e o laudo de referência, sendo que valores mais altos indicam maior similaridade entre os textos comparados. Formalmente, a métrica pode ser definida pela Equação 4.

$$METEOR = F_{mean} \times (1 - Penalty) \tag{4}$$

onde, F_{mean} é a média harmônica ponderada entre precisão e revocação (Equação 5).

$$F_{mean} = \frac{10 \cdot P \cdot R}{R + 9P}, \tag{5}$$

onde, P representa a precisão e R representa a revocação entre os *unigrams* do texto gerado e do texto de referência. O termo de penalização é definido pela Equação 6.

$$Penalty = 0.5 \left(\frac{chunks}{matches} \right)^3, \tag{6}$$

onde $matches$ é o número de *unigrams* alinhados entre os textos, e $chunks$ é o número de sequências contíguas de correspondências. Esse fator penaliza correspondências fragmentadas, favorecendo traduções (ou descrições) com melhor ordenação estrutural.

5. Resultados e Discussão

Apesar das métricas automáticas moderadas, a análise qualitativa mostra que o modelo captura padrões radiológicos e a estrutura clínica dos laudos. Porém, os laudos gerados apresentam erros como alucinações factuais, omissão de achados, inconsistências semânticas, desalinhamento imagem-texto, padrões genéricos, erros na intensidade dos achados e confusão entre conceitos clínicos, revelando limitações no alinhamento multimodal e na consistência factual. Trabalhos futuros propõem aprendizado multitarefa com classificação de patologia, decodificação guiada por fatos e geração hierárquica para reduzir alucinações e melhorar a consistência clínica.

Observa-se que os modelos baseados em memória (R2Gen e CMN) estabelecem um *baseline* sólido para a tarefa, enquanto a estratégia de *warm starting* proposta por [Nicolson et al. 2022] eleva substancialmente o desempenho em todas as métricas. O modelo proposto, embora apresente métricas BLEU inferiores, demonstra resultados competitivos em ROUGE-L e METEOR, sugerindo que diferentes *trade-offs* podem ser explorados no projeto de arquiteturas para geração de laudos radiológicos.

A Tabela 3 apresenta uma comparação quantitativa dos resultados reportados nestes trabalhos com a abordagem proposta neste artigo.

Tabela 3. Comparação de métricas entre diferentes abordagens para geração de laudos radiológicos.

Modelo	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
R2Gen [Chen et al. 2020]	0,353	0,218	0,145	0,103	0,277	0,142
CMN [Chen et al. 2021]	0,353	0,218	0,148	0,106	0,278	0,142
CvT2DistilGPT2 [Nicolson et al. 2022]	**0,392**	**0,248**	**0,171**	**0,127**	**0,286**	0,155
[Yang et al. 2022] - MIMIC-CXR	0,363	0,228	0,156	0,115	0,284	-
Modelo Proposto	0,310	0,185	0,127	0,089	0,263	**0,282**

A Tabela 4 apresenta exemplos comparativos entre laudos gerados e referências.

Tabela 4. Exemplos comparativos entre laudos de referência e laudos gerados pelo modelo proposto

Laudo de Referência	Laudo Gerado
"Os pulmões estão claros, sem consolidação focal. Não há derrame pleural nem pneumotórax. A silhueta cardíaca está levemente aumentada. Os contornos do mediastino são normais. Há persistente leve proeminência da vasculatura pulmonar central, sugerindo congestão vascular pulmonar central, sem edema pulmonar evidente. Sem consolidação focal que sugira pneumonia."	"A silhueta cardíaca permanece aumentada. Os contornos mediastinais e hilares são normais. A vasculatura pulmonar não está ingurgitada. Não há grande derrame pleural nem pneumotórax. Há leve congestão vascular pulmonar, sem edema pulmonar evidente. Sem consolidação focal que sugira pneumonia."
"Os pulmões estão hiperinsuflados, com achatamento do diafragma, sugerindo doença pulmonar obstrutiva crônica (DPOC). Há mínima cicatriz no pulmão direito, na região média. Não há consolidação focal, derrame pleural ou evidência de pneumotórax. A silhueta cardíaca está levemente aumentada. A aorta encontra-se calcificada e tortuosa. Leve aumento da silhueta cardíaca, sem edema pulmonar evidente."	"Aumento moderado da silhueta cardíaca, inalterado. Os contornos mediastinais e hilares estão inalterados. A vasculatura pulmonar não está ingurgitada. Os pulmões permanecem hiperinsuflados, com achatamento das cúpulas diafragmáticas, sugestivo de DPOC. Não há consolidação focal, derrame pleural ou pneumotórax. Há leves alterações degenerativas na coluna torácica. Cardiomegalia moderada, sem edema pulmonar evidente."

Embora o modelo proposto apresente pontuações BLEU inferiores em comparação com arquiteturas baseadas em memória, como R2Gen e CMN, ele demonstra desempenho competitivo nas métricas ROUGE-L e METEOR. Por exemplo, enquanto o modelo proposto alcança um BLEU-4 de 0,089 — valor inferior aos reportados por trabalhos como R2Gen (0,103) e CMN (0,106) — ele obtém valores relativamente próximos em ROUGE-L (0,263) e apresenta um valor elevado em METEOR (0,282), indicando maior similaridade semântica entre os laudos gerados e os laudos de referência.

Dessa forma, mesmo com menor sobreposição exata de n-gramas, os resultados sugerem que o modelo consegue preservar a coerência clínica e a estrutura típica dos laudos radiológicos. Em outras palavras, o modelo frequentemente descreve os mesmos

achados clínicos presentes nos laudos de referência, ainda que utilize variações linguísticas ou estruturas sintáticas diferentes.

Um aspecto importante da abordagem proposta é a sua simplicidade arquitetural. Enquanto diversos modelos de última geração dependem de módulos de memória complexos, grafos de conhecimento ou mecanismos baseados em recuperação, a arquitetura proposta mantém um *design* de transformador multimodal simples. Isso facilita o treinamento e a reprodução do modelo, ao mesmo tempo que permite a captura de informações clinicamente relevantes.

Além disso, a integração do histórico clínico do paciente fornece pistas contextuais adicionais que auxiliam na geração do laudo. Os exemplos qualitativos apresentados na Tabela 4 indicam que o modelo é capaz de capturar achados radiológicos importantes, como cardiomegalia, congestão pulmonar e ausência de pneumotórax ou derrame pleural. Apesar disso, algumas limitações persistem. O modelo ocasionalmente gera inconsistências factuais ou achados alucinatórios, um problema comum em sistemas de geração de laudos radiológicos. Essas questões ressaltam a necessidade de mecanismos adicionais para aprimorar a consistência factual e o alinhamento visual-textual.

6. Conclusão

O modelo proposto combina um codificador visual *ResNet-50* otimizado com representações *Bio_ClinicalBERT* e uma estrutura de codificador-decodificador *Transformer* para gerar laudos radiológicos de forma autorregressiva. Os resultados experimentais no conjunto de dados MIMIC-CXR demonstram que a abordagem proposta é capaz de gerar descrições clinicamente coerentes, mantendo uma arquitetura simples e reproduzível.

Embora as pontuações de avaliação automática obtidas permaneçam abaixo das de alguns métodos de última geração recentes, a análise qualitativa indica que o modelo captura com sucesso diversas descobertas clinicamente relevantes e preserva a estrutura geral dos laudos radiológicos.

Trabalhos futuros se concentrarão em aprimorar a consistência factual e a precisão clínica por meio da incorporação de tarefas auxiliares de classificação de patologias, estratégias de decodificação guiadas por conhecimento e mecanismos hierárquicos de geração de laudos. Além disso, pesquisas futuras investigarão o uso de estruturas visuais mais avançadas, como os transformadores visuais, para aprimorar ainda mais o aprendizado da representação visual.

Referências

Akhter, Y., Singh, R., and Vatsa, M. (2023). Ai-based radiodiagnosis using chest x-rays: A review. *Frontiers in big data*, 6:1120989.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics.

Bruno, M. A., Walker, E. A., and Abujudeh, H. H. (2015). Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *RadioGraphics*, 35(6):1668–1676.

Chen, Z., Shen, Y., Song, Y., and Wan, X. (2021). Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914.

Chen, Z., Song, Y., Chang, T.-H., and Wan, X. (2020). Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449. arXiv:2010.16056.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. (2019). Mimic-cxr-jpg dataset v2.0.0. Available at: https://physionet.org/content/mimic-cxr-jpg/2.0.0/.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.

Nicolson, A., Dowling, J., and Koopman, B. (2022). Improving chest x-ray report generation by leveraging warm starting. *arXiv preprint*, arXiv:2201.09405. Submitted to Elsevier.

Sutskever, I., Vinyals, O., and Le, Q. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE CVPR*, pages 3156–3164.

Yang, S., Wu, X., Ge, S., Zhou, S. K., and Xiao, L. (2022). Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, xx:xxx–xxx. arXiv:2112.15009.