# Geração de laudos de retinografia utilizando Contrastive Captioner

**Patrik O. Pimentel**[1], **Mauricio M. Almeida**[1], **João D. S. Almeida**[1], **Victor H. B. de Lemos**[1], **Luis Eduardo S. C. Martins**[1]

[1]Núcleo de Computação Aplicada – Universidade Federal do Maranhão (UFMA)
CEP 65085-580 – São Luís – MA – Brasil

{patrikop, mauricio.almeida,

luisescm, victorhbl12, jdallyson}@nca.ufma.br

***Abstract.*** *Automatic generation of fundus reports acts as medical support, allowing the diagnosis of eye diseases more quickly when compared to traditional methods, reducing the waiting time of patients with eye diseases and contributing to the reduction of cases of visual impairment. Recent models of report generation propose new methods for integrating visual and textual information, presenting dependence on keywords for the generation of clinical descriptions. In this work, we explore the pre-trained Contrastive Captioner (CoCa), aiming to correlate image and text by combining the two loss functions present in the model, with the objective of generating fundus reports without depending on keywords. In the experiments performed on the DeepEyeNet dataset, the method achieved a BLEU-4 of 0.230, CIDEr of 0.517, and METEOR of 0.423.*

***Resumo.*** *A geração automática de laudos de retinografia atua como suporte médico, permitindo o diagnóstico de doenças oculares com maior agilidade se comparado a métodos tradicionais, reduzindo o tempo de espera dos pacientes com doenças oculares e contribuindo para a diminuição de casos de deficiência visual. Modelos recentes de geração de laudos propõem novos métodos para integração de informações visuais e textuais, apresentando dependência de palavras-chave para a geração das descrições clínicas. Neste trabalho, exploramos o Contrastive Captioner (CoCa) pré-treinado, visando correlacionar imagem e texto por meio da combinação das duas funções de perda presentes no modelo, visando gerar laudos de retinografias sem depender de palavras-chave. Nos experimentos realizados no dataset DeepEyeNet o método alcançou um BLEU-4 de 0,230, CIDEr de 0,517, e METEOR de 0,423.*

## 1. Introdução

As doenças oculares ocorrem por diversos fatores, incluindo envelhecimento, causas genéticas, hábitos e estilo de vida. Algumas dessas doenças podem causar deficiência visual, como glaucoma e retinopatia diabética. Além disso, estima-se que pelo menos 2,2 bilhões de pessoas possuam deficiência visual e que, em pelo menos metade desses casos, a deficiência visual poderia ter sido evitada ou ainda não foi tratada [Organization et al. 2019]. Espera-se que, até 2045, o número de indivíduos adultos com retinopatia diabética aumente para 160,5 milhões [Teo et al. 2021]. É necessário que doenças como essas sejam diagnosticadas o mais breve possível, pois levam a um mau

funcionamento da visão e representam a principal causa de cegueira na população em idade ativa [Zheng et al. 2012].

Nos métodos tradicionais, a detecção de doenças oculares é manual, o que a torna cara, demorada e difícil de adotar em larga escala [Iqbal et al. 2022]. Além disso, em países em desenvolvimento, oftalmologistas se concentram em áreas mais favorecidas, resultando em acesso desigual aos serviços [Hong et al. 2016].

Com os avanços do aprendizado profundo, novas abordagens de triagem oftalmológica utilizam inteligência artificial para aprimorar a detecção precoce e o tratamento de doenças. A integração entre Visão Computacional e Processamento de Linguagem Natural (PLN) tem se destacado nos modelos Vision-Language (VLMs), como o LLaVA-Med [Li et al. 2023], e em modelos semelhantes ao M3 Transformer [Shaik et al. 2024]. Esses modelos possibilitam interpretações automatizadas de exames e aumentam a eficiência nos diagnósticos.

Modelos mais recentes para geração de descrições de retinografias enfrentam dificuldades na correspondência entre pares de dados. Para atenuar isso, o modelo Contrastive Captioner (CoCa) [Yu et al. 2022] utiliza uma abordagem híbrida de aprendizado contrastivo e generativo, melhorando a associação entre imagens e descrições. O CoCa gera legendas mais ricas e supera modelos como FLAVA [Singh et al. 2022] e CLIP [Radford et al. 2021] em benchmarks de geração de legendas para imagens, como o Flickr30K [Plummer et al. 2015], alcançando $92,5\%$ de $R@1$ (image-to-text) em avaliação zero-shot.

Este trabalho propõe um modelo de geração de laudos de retinografia utilizando o CoCa. Avaliamos seu desempenho na tarefa de geração de legendas na base DeepEyeNet [Huang et al. 2021], considerando tanto a versão pré-treinada quanto a ajustada via fine-tuning e seu desempenho comparado a outros modelos. O modelo pré-treinado reduz o custo computacional associado ao treinamento de VLMs [Bordes et al. 2024]. Os resultados mostram que o CoCa pré-treinado e ajustado supera outros modelos em métricas como BLEU-4 e METEOR, demonstrando a eficácia do aprendizado transferido e a redução do custo computacional, além de mitigar a limitação de bases de dados de retinografias.

Este artigo descreve trabalhos relacionados com DeepEyeNet (Seção 2), metodologia (modelo, base de dados - Seção 3), resultados e comparação (Seção 4), e conclusões/trabalhos futuros com CoCa para descrições clínicas (Seção 5).

## 2. Trabalhos Relacionados

Os avanços recentes em modelos de visão e linguagem impulsionam a geração automática de descrições para imagens médicas, permitindo uma melhor interpretação e análise clínica. Diversos modelos são propostos para a geração de legendas em retinografias. Alguns adotam arquiteturas mais simples em comparação com modelos baseados em *Transformers* [Vaswani et al. 2023], como a combinação de *Convolutional Neural Networks* (CNNs) com redes *Long Short Term Memory* (LSTM) [Schmidhuber et al. 1997]. Em Huang et al. (2022) , uma rede CNN é utilizada como *backbone* para a extração de características e uma rede LSTM bidirecional como *decoder* sequencial. O diferencial desse modelo é o mecanismo baseado em *non-local attention* [Wang et al. 2018], denominado

*TransFuser*, que realiza a fusão multimodal ao combinar imagens e palavras-chave definidas por especialistas, permitindo ao modelo gerar relatórios mais precisos e contextualizados.

A rede LSTM enfrenta o problema do *vanishing gradient* [Pascanu et al. 2013], limitando sua eficácia em contextos que demandam maior capacidade de memória e processamento contínuo. Atualmente, os modelos baseados em *Transformer* representam o estado da arte em diversas tarefas de geração de texto [Topal et al. 2021], utilizando mecanismos de atenção que permitem o processamento eficiente de sequências longas e complexas. O método proposto por Dutra et al. (2024) demonstra a capacidade desses modelos na geração automática de descrições clínicas, combinando a *EfficientNet* como *backbone* e uma camada *Transformer* para processar o texto e realizar a integração multimodal de maneira confiável.

Nesse mesmo domínio, Krishna Cherukuri et al. (2024) propõem o modelo GCS-M3VLT, que utiliza um mecanismo de autoatenção guiada por contexto e integra informações visuais e textuais de forma contextualizada. Ao final, o *Language Generation Decoder*, baseado na arquitetura do GPT-2 [Radford et al. 2019], processa os dados multimodais para gerar descrições médicas coerentes e detalhadas. Os resultados demonstram que o GCS-M3VLT supera modelos anteriores na tarefa de *Image Captioning* para imagens de retina, alcançando um BLEU-4 de 0,231, indicando uma melhora na geração de relatórios clínicos.

Os trabalhos apresentados demonstram que modelos baseados em *Transformer*, com mecanismos de atenção especializados, melhoram o alinhamento entre imagens e legendas, avançando na geração de legendas. Este trabalho explora o uso do *Vision-Language Model* (VLM) CoCa para gerar legendas em retinografias, propondo um modelo que combina aprendizado contrastivo e um mecanismo generativo para produzir descrições mais precisas.

## 3. Materiais e Método

Esta seção apresenta o método proposto para a geração de descrições clínicas de retinografia, incluindo o conjunto de dados, a arquitetura utilizada, o pré-processamento aplicado para a padronização dos dados, e, por fim, a técnica empregada para reduzir a necessidade de mais épocas de treinamento, por meio do treino parcial do modelo.

### 3.1. Base de Retinografias

Para a avaliação da abordagem proposta, foi escolhido o conjunto de dados *DeepEyeNet* [Huang et al. 2021], frequentemente utilizado para validação de modelos de geração de descrições de retinografia, que contém 15.709 pares de imagens e legendas descritivas, sendo estas definidas por especialistas em retina ou oftalmologistas, além desses pares de dados, contém palavras-chave que podem indicar a doença, o tipo de exame realizado ou outras informações adicionais. Além disso, cada imagem possui palavras-chave associadas que descrevem informações relevantes para o diagnóstico.

O conjunto de dados contém imagens obtidas por Fotografia de Fundo de Olho, totalizando 13.898 imagens coloridas, e por Retinografia Fluorescente, com 1.811 imagens em escala de cinza. Para os experimentos, foi utilizada a divisão da base original, dividida em três partes, sendo 60% para treinamento, 20% para validação e 20% para

teste, resultando em 9.425 imagens para treinamento, 3.142 para validação e 3.142 para teste. A Figura 1 mostra exemplos de imagens de retinografia e angiografia presentes no dataset.

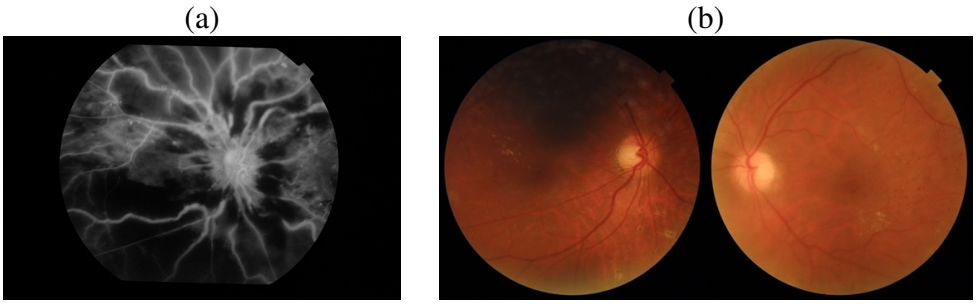



**Figura 1. Exemplo de imagem obtida por meio de Retinografia Fluorescente (a) e Fotografia de Fundo de Olho (b), presentes no *dataset* [Huang et al. 2021].**

### 3.2. Método proposto

O método proposto é ilustrado na Figura 2. Após a aquisição do conjunto de dados, é realizado o pré-processamento dos dados. O treinamento foi realizado em duas etapas: na primeira, apenas uma parte do modelo é ajustada para adaptá-lo ao domínio das descrições clínicas, e na segunda, é realizado um fine-tuning completo do modelo, sobre a base de dados, em ambas as etapas é utilizada apenas os dados de treino da divisão original do *dataset*. Em uma etapa posterior, é realizada a geração das descrições clínicas utilizando a divisão de teste da base e em seguida, as métricas de similaridade entre a predição e a referência são calculadas com base nas mesmas adotadas por outros trabalhos da literatura.

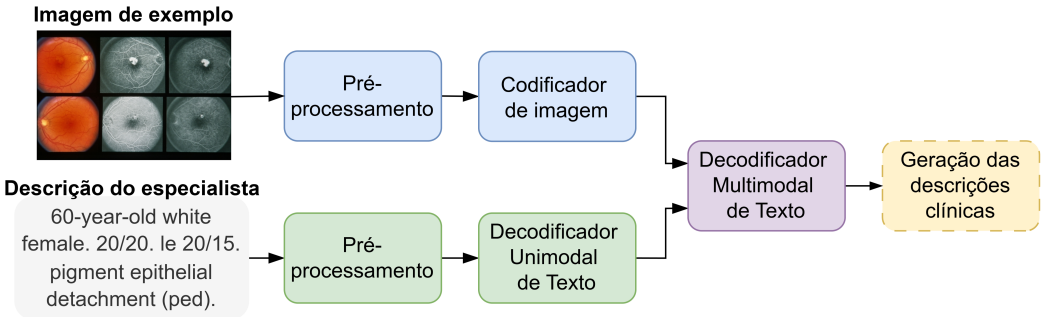



**Figura 2. Etapas do método.**

### 3.3. Pré-processamento das imagens e texto

Para o fine-tuning do modelo, foram utilizadas as imagens e as legendas do *dataset* (Subseção 3.1). A seguir, é descrito o processamento aplicado a cada um desses dados.

As imagens foram redimensionadas para a dimensão de 224x224 utilizando a técnica de interpolação bilinear [Monasse 2019], possuindo 3 canais de cores. Ao final, é realizada uma normalização de canais.

Para o texto, foi realizada a remoção de caracteres não alfanuméricos, como adotado por [Huang et al. 2022] e os números foram substituídos pelos três primeiros caracteres da palavra *number* para marcar sua posição. Em seguida, é utilizada a técnica de

tokenização *Byte Pair Encoding* (BPE) [Sennrich et al. 2016], que permite criar um vocabulário aberto ao dividir as palavras em sub palavras frequentes, agrupando de maneira iterativa as sub palavras mais comuns ou pares de caracteres, formando novas palavras. A técnica permite tratar palavras raras ou novas, permitindo que modelos pré-treinados se adaptem melhor ao novo vocabulário.

### 3.4. Arquitetura CoCa

Nesta seção, são apresentadas a arquitetura do modelo e a função de *loss*, com um detalhamento sobre o Codificador de Imagem, o Decodificador de Texto e a função de *loss*.

O *Contrastive Captioner* (CoCa) utiliza a arquitetura *Endoder-Decoder*. O diferencial do modelo é que o decodificador é dividido em 2 partes, permitindo o uso do método contrastivo, o módulo unimodal que realiza o processamento apenas dos textos, sendo retirado o mecanismo de atenção cruzada dessa parte, e o módulo multimodal que possui na saída o mecanismo de atenção cruzada. A arquitetura é dividida em três partes: Codificador de Imagem, Decodificador Unimodal de Texto e Decodificador Multimodal de Texto, responsável por integrar as informações obtidas através do Codificador de Imagem e Decodificador Unimodal de Texto. A Figura 3 ilustra a arquitetura Coca utilizada, sendo que a camada Multi-Head, presente no bloco *Transfomer* pode utilizar mascaramento dependendo do módulo em que é utilizado.

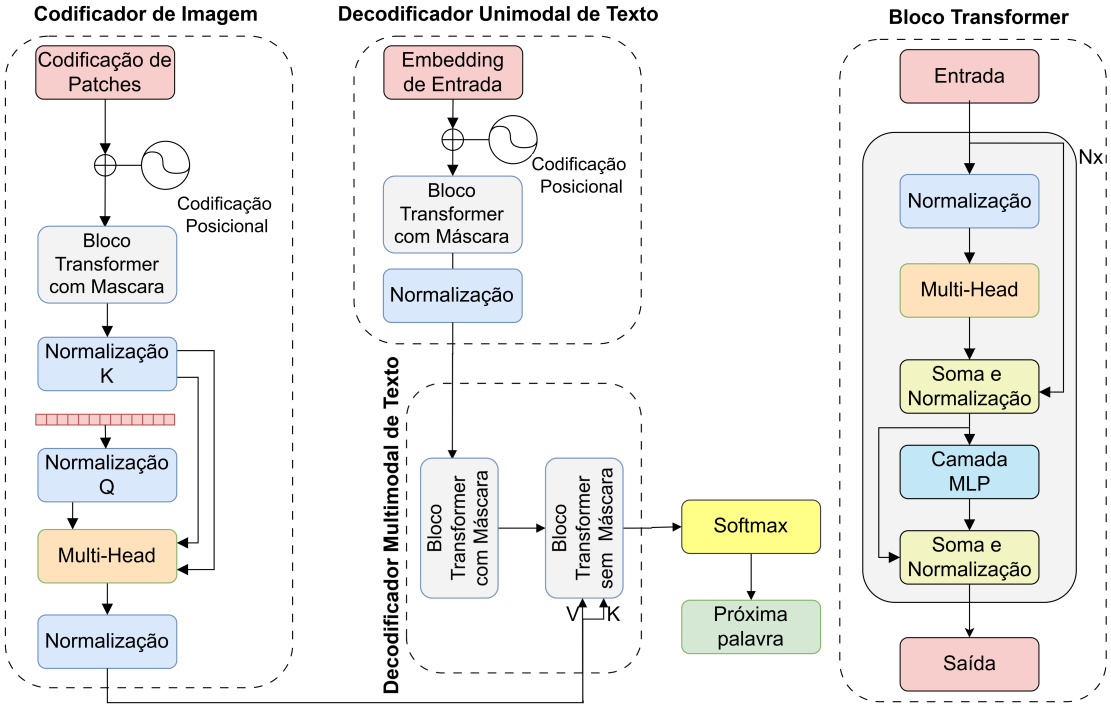



**Figura 3. Arquitetura do modelo CoCa.**

#### 3.4.1. Codificador de Imagem

O Codificador de Imagem recebe uma imagem como entrada e segue um fluxo de processamento semelhante ao do modelo ViT [Dosovitskiy et al. 2020]. Inicialmente, a imagem

é dividida em *patches*, que são achatados e transformados em *embeddings*. A esses vetores, adicionam-se *positional embeddings* aprendíveis, permitindo que o *Image Encoder* preserve a estrutura espacial da imagem.

Os vetores resultantes são então processados por múltiplas camadas empilhadas do bloco *Transformer* [Vaswani et al. 2023]. A saída final passa por uma camada de *Attentional Pooler*, que utiliza uma camada de consulta aprendível [Arar et al. 2022] para gerar consultas com menor custo computacional.

### 3.4.2. Decodificador de Texto: Unimodal e Multimodal

O Decodificador de Texto é composto por dois módulos que possuem tarefas distintas. O Decodificador Unimodal é treinado exclusivamente com texto, codificando-o em um vetor latente de informações. Assim como o Decodificador de Imagem, o módulo unimodal possui um token [CLS] aprendível na saída final.

O Decodificador Multimodal de Texto é responsável por combinar as saídas do Codificador de Imagem e do Decodificador Unimodal. Na primeira etapa, a saída do Decodificador Unimodal passa por um mecanismo de autoatenção mascarada. Em seguida, é utilizado o mecanismo de atenção cruzada na saída do decodificador juntamente com a saída da primeira parte do módulo, permitindo a integração das informações textuais e visuais.

### 3.4.3. Calculo da Loss

O modelo CoCa utiliza a combinação de duas funções de perda para treinar pares de imagem e texto. A primeira é a perda de entropia cruzada [Mao et al. 2023], que calcula a diferença entre a referência e os tokens gerados, permitindo que o modelo gere legendas coerentes e detalhadas com base nas informações visuais. A segunda é a perda contrastiva [Yang et al. 2022], cujo objetivo é alinhar embeddings de texto e imagem no mesmo espaço latente utilizando os tokens [CLS] das saídas dos módulos. Essa abordagem permite que o modelo atue tanto em tarefas discriminativas quanto generativas.

O cálculo da perda total do modelo é dado na Equação 1, sendo $w_{cap}$ e $w_{con}$ os hiperparâmetros de ponderação das funções de perda.

$$L_{CoCa} = w_{cap} \cdot L_{cap} + w_{con} \cdot L_{con} \tag{1}$$

## 3.5. Adaptação prévia do Decodificador Unimodal

Antes do ajuste do modelo completo, o Decodificador Unimodal é treinado com as legendas do *dataset* Subseção 3.1 utilizando entropia cruzada como função de perda. Essa etapa acelera a convergência do modelo completo, reduz a necessidade de mais épocas de fine-tuning e contribui para alcançar melhores métricas nos experimentos.

## 3.6. Métricas de Avaliação

Para realizar a avaliação das descrições geradas, são escolhidas as mesmas métricas utilizadas em trabalhos anteriores, que são: BLEU [Post 2018], METEOR [Banerjee and Lavie 2005], ROUGE-L [Lin 2004] e CIDEr [Vedantam et al. 2015].

A métrica BLEU, assim como as outras métricas utilizadas, é calculada com base nas predições do modelo e o texto de referência. Ela calcula a sobreposição de n-gramas, que são sequências contínuas de tokens, caracteres ou palavras.A pontuação pode diminuir quando há mudanças na ordem das palavras, quando curtas ou não correspondem com a referência, uma limitação que apresenta é que não consegue captar nuances semânticas corretamente.O METEOR supera o BLEU ao permitir sinônimos e maior flexibilidade na ordem. Sua pontuação é calculada pelo F-Mean, que combina a precisão e o *recall* por meio de uma média harmônica ponderada, priorizando o *recall* e aplicando penalização para sequências desalinhadas.

Métricas como a ROUGE-L, não exigem correspondência exata na ordem. A pontuação do ROUGE-L se dá pela avaliação das Subsequências Comuns mais Longas (LCS), ponderando precisão e recall. A métrica CIDEr foi proposta para avaliar a qualidade das descrições de imagens geradas, permitindo analisar a semântica, atribuindo pesos a termos informativos e aplicando diferentes tamanhos de n-gramas para produzir a pontuação final. As métricas BLEU, METEOR e ROUGE-L possuem intervalos entre 0 e 1 e a CIDEr entre 0 e 10. Quanto maior o valor, melhor a correspondência.

## 4. Experimentos e Resultados

Nesta seção, são apresentados hiperparâmetros e resultados, incluindo uma comparação entre resultados do modelo CoCa e a literatura. Destaca-se também a comparação entre o uso de legendas e palavras-chave, analisando diferentes abordagens na utilização dos dados na etapa de treinamento adotadas em trabalhos anteriores.

Para reduzir a necessidade de grandes quantidades de dados, utilizou-se o modelo pré-treinado no *dataset* LAION-2B [Schuhmann et al. 2022], uma base de dados de grande escala que contém mais de 2 bilhões de pares imagem-texto coletados da web. Foram realizados diversos experimentos de fine-tuning, variando o congelamento do Codificador de Texto, do Decodificador Unimodal de Texto e o treinamento completo com todos os módulos ajustáveis.

Os experimentos foram realizados com uma taxa de aprendizado de $1 \cdot 10^{-5}$, batch 16 e dropout de 0,2. Aplicou-se um warmup inicial de 8000 steps, seguido por um agendador de taxa de aprendizado (learning rate scheduler) baseado em decaimento cosseno e tendo como otimizador o AdamW [Loshchilov and Hutter 2017]. Seguindo Yu et al. (2022), utilizamos $w_{cap} = 2$ e $w_{con} = 1$.

O fine-tuning do modelo foi realizado com a seguintes configurações de hiperparâmetros. A entrada do Codificador de Imagem é de 32×32, gerando embeddings de 768 dimensões para cada patch. O Decodificador Unimodal utiliza embeddings de 512 dimensões. As três partes do modelo (Seção 3.4) possuem 12 camadas cada e utilizam 8 cabeças de atenção. Ao final de cada parte, é aplicada uma projeção para 512 dimensões de saída, garantindo compatibilidade entre os módulos. O modelo utiliza um vocabulário de 49.408 tokens e um comprimento máximo de contexto de 76 tokens. Durante o treinamento, foram utilizadas 150 épocas com early stopping e paciência de 12.

Os experimentos utilizaram 2 GPUs NVIDIA Tesla T4 (2x16 GB RAM). Com nossa abordagem, o tempo médio de treinamento foi de 4 horas, em contraste com a abordagem sem adaptação prévia do Decodificador Unimodal, que necessita em média de 11 horas.

A Tabela 1 apresenta os resultados para as variações no congelamento dos módulos do modelo para o fine-tuning. Nela, o módulo Decodificador de Texto Unimodal é representado como Decodificador de Texto. O modelo *zero-shot* obteve resultados superiores em relação aos modelos com congelamentos parciais, sendo que o congelamento do Decodificador de Imagem apresenta o resultado inferior ao do Decodificador de Texto, indicando a necessidade de adaptação as novas imagens, especialmente em um conjunto variado de imagens, como o *DeepEyeNet*. Além disso, o melhor desempenho é alcançado quando todos os módulos permanecem descongelados, reforçando a importância do fine-tuning completo para a geração de laudos mais precisos.

**Tabela 1. Resultados das métricas de avaliação para diferentes configurações de congelamento dos módulos do modelo.**

| Congelado | BLEU 1 | BLEU 2 | BLEU 3 | BLEU 4 | CIDEr | ROUGE-L | METEOR |
|---|---|---|---|---|---|---|---|
| Codificador de Imagem + Decodificador de Texto | 0,046 | 0,036 | 0,027 | 0,021 | 0,044 | 0,078 | 0,061 |
| Codificador de Imagem | 0.055 | 0.039 | 0,030 | 0,020 | 0,089 | 0,074 | 0,059 |
| Decodificador de Texto | 0.126 | 0.093 | 0,071 | 0,067 | 0,243 | 0,157 | 0,128 |
| Todos (*zero-shot*) | 0,18 | 0,13 | 0,10 | 0,08 | 0,20 | 0,21 | 0,17 |
| Metade do Decodificador de Texto | 0,19 | 0,15 | 0,12 | 0,09 | 0,26 | 0,24 | 0,23 |
| Sem congelamento (todos treináveis) | **0,379** | **0,329** | **0,274** | **0,229** | **0,517** | **0,428** | **0,423** |

Na Tabela 2, é apresentada a comparação dos dados necessários para o treinamento de cada modelo. Isso se reflete na necessidade de alguns modelos utilizarem imagens e palavras-chave para gerar melhores descrições. Métodos como o proposto por [Dutra et al. 2024] requerem apenas a imagem para a geração da legenda, demonstrando um resultado promissor na geração de legendas para imagens retinais e podendo ser integrados na triagem de pacientes, assim como o modelo CoCa, utilizado neste trabalho.

**Tabela 2. Comparação dos dados utilizados usado para treinamento**

| Model | Imagem | Legenda | Palavras chaves |
|---|---|---|---|
| [Huang et al. 2022] | ✓ | ✓ | ✓ |
| [Dutra et al. 2024] | ✓ | ✓ | ✗ |
| [Yu et al. 2022] | ✓ | ✓ | ✗ |
| [Shaik et al. 2024] | ✓ | ✓ | ✓ |
| [Krishna Cherukuri et al. 2024] | ✓ | ✓ | ✓ |

A Tabela 3 apresenta o método com melhor desempenho em comparação com outros trabalhos que utilizam a *DeepEyeNet* como base de avaliação. Os resultados destacam um desempenho notável, especialmente nas métricas BLEU-4, CIDEr e METEOR, alcançando 0,230, 0.517 e 0,423 respectivamente, indicando que o modelo gera legendas próximas às referências de especialistas, enquanto apresenta uma boa correspondência semântica, dependendo apenas das imagens para gerar uma nova descrição clínica.

Em relação aos modelos recentes da literatura, o modelo superou algumas abordagens e obteve o segundo melhor desempenho nas métricas BLEU-2 e BLEU-4, sendo superado apenas pelo estado da arte [Krishna Cherukuri et al. 2024], que utiliza palavras-chave para o treinamento, ficando somente 0,001 BLEU-4 abaixo. Esses resultados demonstram a eficiência da abordagem utilizada e sugerem que ajustes adicionais como a

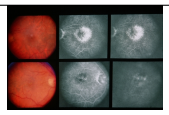
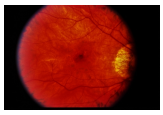
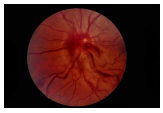
utilização de técnicas de *Data Augmentation*, melhorias na arquitetura do modelo e uma otimização de hiperparâmetros, podem auxiliar o modelo a alcançar o estado da arte na geração de laudos de retinografia.

**Tabela 3. Comparação entre o resultado obtido e os modelos propostos na literatura**

| Model | BLEU 1 | BLEU 2 | BLEU 3 | BLEU 4 | CIDEr | ROUGE-L | METEOR |
|---|---|---|---|---|---|---|---|
| [Huang et al. 2022] | 0,230 | 0,150 | 0,094 | 0,053 | 0,370 | 0,291 | - |
| [Dutra et al. 2024] | 0,365 | 0,300 | 0.257 | 0.220 | - | 0,329 | 0,378 |
| **CoCa *fine-tuning DeepEyeNet*** | **0,379** | **0,329** | **0,274** | **0,230** | **0,517** | **0,428** | **0,423** |
| [Shaik et al. 2024] | 0,394 | 0,312 | 0,291 | 0,208 | 0,537 | 0,493 | - |
| **[Krishna Cherukuri et al. 2024]** | **0,430** | **0,345** | **0,319** | **0,231** | **0,559** | **0,497** | - |

A Tabela 4 apresenta descrições geradas pelo modelo CoCa, comparando-as com as legendas de referência padronizadas no pré-processamento descrito na Subseção 3.3 e avaliando a correspondência entre elas. Algumas predições são próximas as referências como nos exemplos (a) e (b), enquanto o exemplo (c) pontua apenas por acertar palavras genéricas do conjunto de dados e o exemplo (d) falha completamente, gerando uma descrição incorreta para a imagem.

**Tabela 4. Exemplos das descrições geradas pelo modelo.**

| ID | Imagem | Referência | Predição | BLEU-4 |
|---|---|---|---|---|
| (a) |  | [num] year old white female pe folds re [num] [num] le [num] [num] | num year old white female re num num le num num pe folds | 0.717 |
| (b) |  | [num] year old white female myopic degeneration | num year old white female retinitis cmv | 0.614 |
| (c) |  | [num] year old with pdr with nvd | num year old white male with macular scar cnvm armd | 0.167 |
| (d) |  | acute appearance with optic nerve hemorrhage and congestion | macular hole | 0.0 |

## 5. Conclusão

Esta pesquisa demonstra a eficácia do modelo Contrastive Captioner (CoCa) na geração de legendas relevantes para imagens de retinografia. A combinação de seus aspectos contrastivo e generativo permite a produção de descrições que auxiliam na análise automatizada desses exames oftalmológicos. A adaptação do CoCa, através de um método de treinamento de baixo custo, representa uma contribuição significativa ao fornecer um suporte valioso para o trabalho do oftalmologista, que pode avaliar e complementar as legendas geradas conforme necessário.

Trabalhos futuros podem explorar o uso de CNNs pré-treinadas como *Image Encoder* para redução de custo, mantendo o mesmo desempenho obtido com a ViT. Outra

possibilidade é utilizar modelos ViT treinados em conjuntos de dados de classificação de retinografias com alto nível de ruído como codificadores de imagem, com o objetivo de melhorar a eficácia do modelo em distinguir entre imagens, em cenários com dados anotados escassos ou inconsistentes. Alguns grupos de doenças, são mais prevalentes em faixas etárias, como por exemplo, catarata em idosos. Com isso, sugere-se incluir essas condições como palavras-chave, juntamente com as já existentes no *dataset*. Somado a isso, pode-se treinar um módulo em pares de imagem e palavras-chave da base que seja capaz de gerar essas palavras com base na imagem, utilizado-o tanto na fase de treinamento quanto na geração dos laudos, a fim de fornecer maior contexto ao modelo gerador.

Por fim, este trabalho apresenta uma limitação em relação ao dataset *DeepEyeNet*, pois é estrangeiro e pode não refletir totalmente as especificidades do contexto clínico brasileiro, como diferenças na incidência de doenças e na qualidade das imagens. Além disso, em situações clínicas específicas , o modelo pode apresentar baixa assertividade ao lidar com retinografias agrupadas em apenas uma imagem, comprometendo a precisão das descrições geradas.

## Agradecimentos

Os autores agradecem o apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), e Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA).

## Referências

Arar, M., Shamir, A., and Bermano, A. H. (2022). Learned queries for efficient local attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10841–10852.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., et al. (2024). An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dutra, E. F., de Lemos, V. H., Almeida, J. D., and de Paiva, A. C. (2024). Método automático para geração de laudos médicos em imagens de retinografia utilizando transformer. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 507–518. SBC.

Hong, H., Mújica, O. J., Anaya, J., Lansingh, V. C., López, E., and Silva, J. C. (2016). The challenge of universal eye health in latin america: distributive inequality of ophthalmologists in 14 countries. *BMJ open*, 6(11):e012819.

Huang, J.-H., Wu, T.-W., Yang, C.-H. H., Shi, Z., Lin, I., Tegner, J., Worring, M., et al. (2022). Non-local attention improves description generation for retinal images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1606–1615.

Huang, J.-H., Yang, C.-H. H., Liu, F., Tian, M., Liu, Y.-C., Wu, T.-W., Lin, I., Wang, K., Morikawa, H., Chang, H., et al. (2021). Deepopht: medical report generation for retinal images via deep models and visual explanation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2442–2452.

Iqbal, S., Khan, T. M., Naveed, K., Naqvi, S. S., and Nawaz, S. J. (2022). Recent trends and advances in fundus image analysis: A review. *Computers in Biology and Medicine*, 151:106277.

Krishna Cherukuri, T., Shareef Shaik, N., Devi Bodapati, J., and Hye Ye, D. (2024). Gcs-m3vlt: Guided context self-attention based multi-modal medical vision language transformer for retinal image captioning. *arXiv e-prints*, pages arXiv–2412.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. (2023). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mao, A., Mohri, M., and Zhong, Y. (2023). Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR.

Monasse, P. (2019). Extraction of the Level Lines of a Bilinear Image. *Image Processing On Line*, 9:205–219. `https://doi.org/10.5201/ipol.2019.269`.

Organization, W. H. et al. (2019). World report on vision.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Post, M. (2018). A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Schmidhuber, J., Hochreiter, S., et al. (1997). Long short-term memory. *Neural Comput*, 9(8):1735–1780.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shaik, N. S., Cherukuri, T. K., and Ye, D. H. (2024). M3t: Multi-modal medical transformer to bridge clinical context with visual insights for retinal image medical description generation. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3037–3043. IEEE.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. (2022). Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650.

Teo, Z. L., Tham, Y.-C., Yu, M., Chee, M. L., Rim, T. H., Cheung, N., Bikbov, M. M., Wang, Y. X., Tang, Y., Lu, Y., Wong, I. Y., Ting, D. S. W., Tan, G. S. W., Jonas, J. B., Sabanayagam, C., Wong, T. Y., and Cheng, C.-Y. (2021). Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. *Ophthalmology*, 128(11):1580–1591.

Topal, M. O., Bas, A., and van Heerden, I. (2021). Exploring transformers in natural language generation: Gpt, bert, and xlnet. *arXiv preprint arXiv:2102.08036*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803.

Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., and Gao, J. (2022). Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19163–19173.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models.

Zheng, Y., He, M., and Congdon, N. (2012). The worldwide epidemic of diabetic retinopathy. *Indian journal of ophthalmology*, 60(5):428–431.