Método automático para geração de laudos médicos em imagens de retinografia utilizando Transformer

**Eduardo F. P. Dutra, Victor H. B. de Lemos,
João D. S. Almeida, Anselmo C. de Paiva** [1]

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

{eduardo.felipe, victorhbl12, jdallyson, paiva}@nca.ufma.br

Abstract. *It is estimated that the number of people affected by retinal diseases will increase significantly in the coming decades. Traditional diagnosis of these pathologies involves visual analysis of retinal structures, is time-consuming and requires specialization. Therefore, it is applicable to use an automatic system to support specialists' diagnoses. This paper presents an automatic method for generating a medical report, using a convolutional neural network to extract features from the image and a Transformer network that suggests the initial medical report. The proposed method shows a 30% increase in BLEU compared to the best Image Captioning method on the DeepEyeNet database, which contains 265 different retinal diseases.*

Resumo. *Estima-se que o número de pessoas afetadas por doenças na retina aumentará significativamente nas próximas décadas. O diagnóstico tradicional dessas patologias envolve a análise visual das estruturas da retina, é um processo demorado e requer especialização. Assim, torna-se útil o uso de um sistema automático para suporte ao diagnóstico pelos especialistas. Neste trabalho é apresentado um método automático de geração de relatório médico, usando rede neural convolucional para extração de características da imagem, combinada a uma rede Transformer que sugere o relatório médico inicial. O método proposto apresenta aumento de 30% em BLEU comparado ao melhor método de Image Captioning na base DeepEyeNet, que tem 265 doenças de retina diferentes.*

1. Introdução

No mundo, pelo menos 2,2 bilhões de pessoas têm uma deficiência visual que dificulta a visão de perto ou de longe. Em pelo menos 1 bilhão desses casos, a deficiência visual poderia ter sido prevenida com o diagnóstico precoce [Organization et al. 2019]. Dentre esses casos, as principais condições que causam deficiência visual ou cegueira são catarata (94 milhões), erro refrativo (88,4 milhões), degeneração macular relacionada à idade (8 milhões), glaucoma (7,7 milhões), retinopatia diabética (3,9 milhões) [Steinmetz et al. 2021]. Cerca de 4,2 milhões de pessoas possuem Retinopatia diabética no mundo e estima-se que em 2050, mais de 16 milhões de pessoas acima de 40 anos terão diagnóstico confirmado apenas nas Américas [Hendrick AM 2015].

No diagnóstico tradicional de retina, a obtenção da imagem é frequentemente realizada por meio da oftalmoscopia indireta. O oftalmologista realiza o diagnóstico com

base na observação das estruturas da retina, como vasos sanguíneos, nervo óptico e a mácula. Com a crescente demanda por atendimento oftalmológico e a complexidade dos casos diagnosticados, é fundamental buscar métodos eficazes e rápidos de diagnóstico.

Desta maneira, a fotografia de fundo de olho (FFO) tem como objetivo obter imagens coloridas da retina. Pode ser realizado por um profissional médico com o uso do equipamento adequado, sendo um dos métodos com melhor relação custo-beneficio. Isso favorece a realização do diagnóstico em áreas com baixo número de oftalmologista por habitante. No Brasil, por exemplo, os oftalmologistas estão distribuídos em 29% dos municípios que possuem 79% da população do país, ficando 21% da população desassistida [Umbelino and Ávila 2023].

A principal abordagem para gerar relatório médico automático a partir de imagem, usando técnicas de *Deep Learning*, é o *Image Captioning*, que em sua abordagem mais simples, consiste em uma rede receber a imagem como entrada e produzir uma palavra por vez assim formando o texto do relatório. Esta técnica é amplamente utilizada no contexto de imagens de raio-x de tórax [Pavlopoulos et al. 2022]. As arquiteturas, em sua maioria, seguem o modelo *Encoder-Decoder*, usando *Backbones* e redes neurais recorrentes (RNN) [Shin et al. 2016] ,[Zhang et al. 2017]. Nesse sentido, o objetivo deste trabalho é propor um método para sugestão de laudos oftalmológicos em imagens de retina. Para tanto, utilizou-se da rede *EfficientNet* para extrair e mapear características e um módulo *Transformers* na etapa final de geração do laudo.

Como contribuição destaca-se a avaliação de *Backbones* aplicada no *dataset DeepEyeNet* e proposição da combinação da *EfficientNet* [Tan and Le 2019] como *Backbone* para a rede *Transformer* [Vaswani et al. 2017]. O método proposto superou o estado da arte na métrica BLEU [Papineni et al. 2002a] para a tarefa de *Image Captioning* no *dataset DeepEyeNet*

Este artigo está organizado da seguinte forma: na Seção 2, são apresentados os trabalhos relacionados a *Image Captioning*. Na Seção 3, são descritos os materiais e o método. Na Seção 4, são apresentados os experimentos e resultados para validar o método proposto, além de um estudo comparativo com trabalhos relacionados. Finalmente, as conclusões e trabalhos futuros são apresentados na Seção 5.

2. Trabalhos Relacionados

O problema de *Image Captioning* através de *Deep Learning* é geralmente abordado por métodos que seguem a arquitetura *Encoder-Decoder* [Vinyals et al. 2015]. Onde o *Encoder* extrai características visuais da imagem, enquanto o *Decoder* gera a legenda com base nessas características. A seguir, são apresentados trabalhos de *Image Captioning* desenvolvidos para a base *DeepEyeNet* e, em seguida, trabalhos aplicados a bases de imagens em geral.

Em [Huang et al. 2022], a inclusão de palavras-chave relacionadas ao tema (*retinopathy*, *miopic*, etc.) na entrada do modelo demonstrou melhora na assertividade. Para isso, propuseram um módulo multimodal capaz de combinar características de imagem e texto. As características da imagem e do texto pré-processado são utilizadas como entrada para uma camada de *Self-Attention*, que permite capturar a relação entre o problema e os elementos da imagem. Em seguida, as características da imagem são reforçadas usando

conexão residual e utilizadas como entrada para uma rede *LSTM-Bidirecional*. A cada iteração, além da saída do módulo multimodal, as características da imagem são fornecidas como entrada para a *LSTM*. A avaliação do método foi conduzida na base proposta por [Huang et al. 2021b], na qual foram obtidos resultados promissores até o momento da publicação de seu estudo.

No método proposto por [Liu et al. 2021], a imagem é dividida em pequenos blocos de tamanhos iguais, chamados de *patches*. Cada pedaço é reduzido para dimensionalidade única. Em seguida, é aplicado o *Positional Encoding*, no qual cada elemento da imagem é mapeado para seu correspondente em texto. Isso serve como entrada para uma rede *Transformer* [Vaswani et al. 2017]. Essa combinação de arquiteturas permite ao modelo capturar informações de contexto mais ricas e gerar textos mais significativos. Os resultados do método superaram outras arquiteturas no teste de avaliação do *dataset* MSCOCO [Lin et al. 2014]. Uma dessas abordagens se baseava em usar duas redes diferentes, em que a primeira encontra partes importantes da imagem enquanto a segunda identifica os elementos presentes, e então ambos servem de entrada para outra rede que faz a correlação entre as saídas e gera o texto [Li et al. 2019]. Em outra abordagem, primeiro é realizado a detecção dos elementos da imagem com a característica de fornecer elementos geométricos e de aparência, para que o *Encoder* proposto correlacione seus elementos e o *Decoder* produza o texto [Herdade et al. 2020].

Os trabalhos apresentados anteriormente, indicam avanços promissores na tarefa de *Image Captioning*, mecanismos de atenção e a combinação de diferentes tipos de rede tem sido cada vez mais comum. Neste sentido, o objetivo deste trabalho é propor um método de *Image Captioning* capaz de produzir relatório médico relevante, usando dos principais *Backbones* da atualidade para extrair características da imagem em combinação com uma rede *Transformer*.

3. Materiais e Método

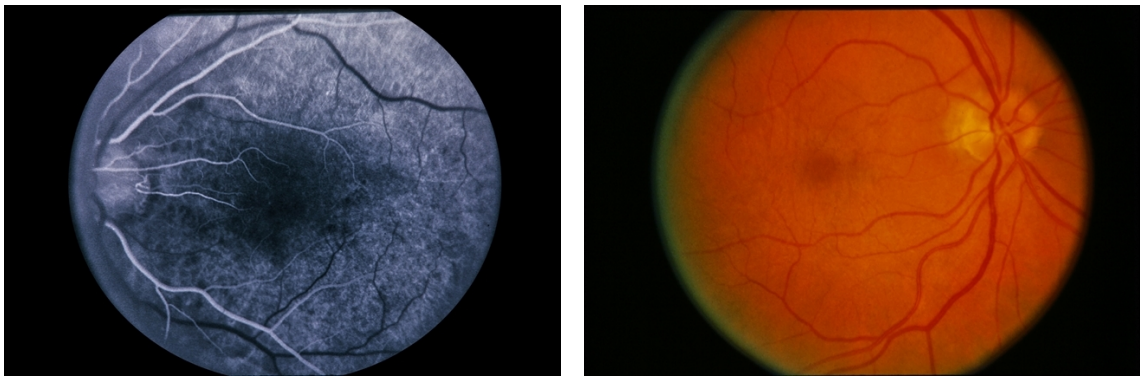
Nesta Seção, são descritos os procedimentos adotados para o desenvolvimento do estudo comparativo proposto para aprimorar geração de relatório médico, em imagens de retina. Na primeira etapa, é feita a aquisição de imagens da base *DeepEyeNet* [Huang et al. 2021b]. Em seguida, são definidas as métricas comumente utilizadas para avaliação dos resultados. Então é apresentado o método de *Image Captioning* proposto.

3.1. Base de Imagens

Para validar o método, foi utilizada a base de imagens *DeepEyeNet* [Huang et al. 2021b], em sua composição cada imagem possui palavras-chave que descrevem a doença além de uma descrição clinica do paciente (Figura 1). Ambos os campos referentes a cada imagem foram manualmente preenchidos por especialistas em retina ou oftalmologistas.

A base contém 15.709 imagens, sendo que 1.811 são referentes ao método Retinografia Fluorescente (FA) e 13.898 obtidas pelo método Fotografia de Fundo de Olho (FFO). Devido à natureza da FA, suas imagens estão em escala de cinza, enquanto as imagens obtidas por meio da FFO estão coloridas (Figura 1). A base possui 265 doenças de retina diferentes, com imagens de diferentes resoluções e descrições com tamanhos variados. A ampla variedade de imagens é importante para tornar a rede generalista, capaz de gerar texto para outras imagens de FFO e FA de outros *dataset*, no entanto, a base

possui imagens que combinam mais de um exame diferente além de apresentar tipos de imagem que se repetem poucas vezes o que torna difícil a geração de relatórios relevantes. A Figura 3.1 apresenta exemplos desse tipo de ruído. A base está dividida em três partes da seguinte maneira, 60% separado para Treinamento, 20% para Validação e 20% para Teste.



Keywords: acute macular neuroretinopathy.
Clinical Description: 26-year-old female, amn/macular neuroretinopathy.

Keywords: branch retinal vein occlusion (brvo), macular exudates.
Clinical Description: 81-year-old male with brvo with macular exudate.

Figura 1. Exemplos de elementos do DeepEyeNet, cada imagem possui dois campos que a descrevem de maneiras diferentes [Huang et al. 2021b].

3.2. Método proposto

A arquitetura proposta se divide em duas partes principais (Figura 3), a *EfficientNet*, para extrair e mapear características de imagens para o espaço latente definido, e um módulo *Transformers*, para interpretar e correlacionar cada elemento da imagem.

Para extrair características visuais de alto nível das imagens, utiliza-se a rede *EfficientNet*. As camadas convolucionais iniciais mantiveram-se congeladas, aproveitando sua capacidade de identificar padrões visuais complexos em imagens coloridas, uma vez que as imagens no conjunto de dados *DeepEyeNet*, consistem predominantemente em imagens coloridas.

3.2.1. Encoder

O *Transformer* é dividido em duas partes distintas. A primeira parte, o *Encoder*, é responsável por interpretar a imagem. Composto por uma camada *Self-Attention* com múltiplas cabeças (MultiHead), que permite ao modelo focar em diferentes partes da imagem simultaneamente, seguida por uma camada de *FeedForward*. Após cada camada, existe uma conexão residual, permitindo que o modelo aprenda com base na representação original da imagem, e em seguida, uma camada de normalização que garante a estabilidade do treinamento. Este processo pode ser definido da seguinte maneira:

$$MultiHead(Q, K, V) = Concat(h_1, ..., h_n)W^O \tag{1}$$

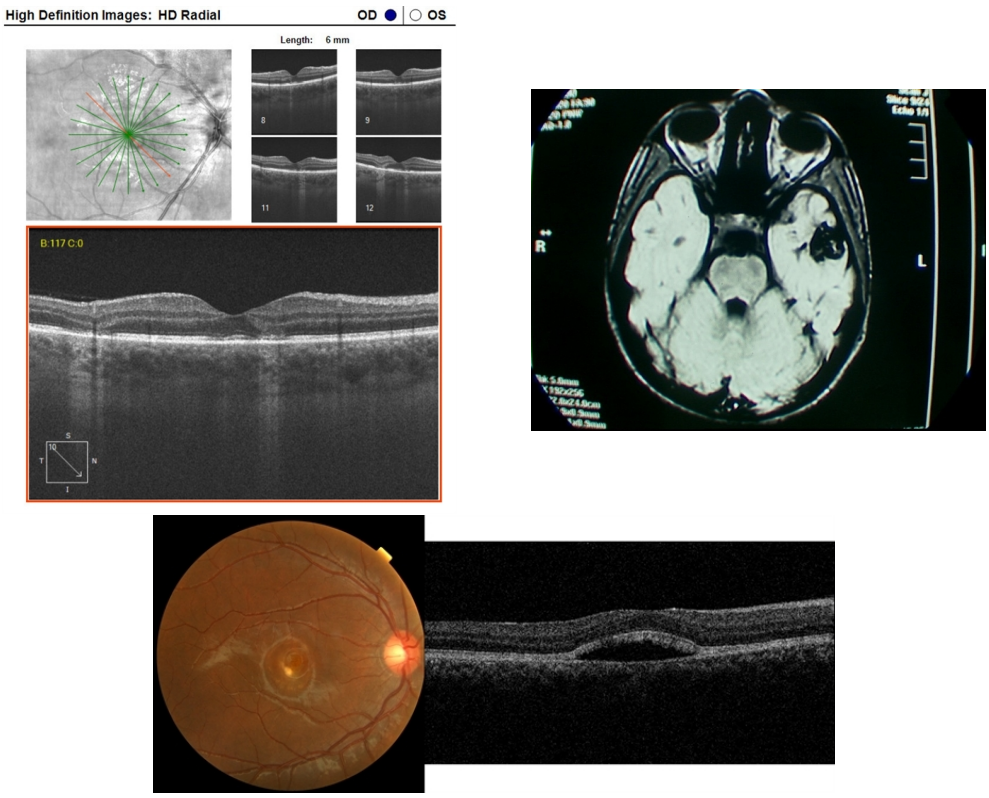


Figura 2. Exemplos de 3 imagens com ruído. [Huang et al. 2021b]

Onde h representa diferentes partes da imagem e W^O uma camada totalmente conectada. O cálculo do produto escalar na operação de atenção segue o padrão estabelecido por [Vaswani et al. 2017].

$$FeedForward(x) = Linear(GELU(Linear(x))) \tag{2}$$

Ao término das operações (1) e (2), aplica-se adição a uma conexão residual que segue uma camada de normalização. Onde x^{in} e x^{out} representam a entrada e saída da camada, respectivamente.

$$x^{\text{out}} = \text{Norm}(x^{\text{in}} + \text{camada}(x^{\text{in}})) \tag{3}$$

3.2.2. Decoder

O *Decoder* pode ser divido em três partes. Na primeira, uma camada de *MultiHead* com máscara. Na segunda, relaciona-se a saída do *Encoder* com a saída da camada anterior em outra *MultiHead*, aproveitando a integração de informações multimodais de maneira eficaz. O processo pode ser descrito da seguinte maneira, um texto Q tem sua similaridade calculada com a saída do *Encoder* K que pondera cada parte da saída do *Encoder* V. Com isso, é garantido a correlação entre elementos da imagem com o texto produzido.

Na terceira, uma camada de *Feed Forward* descrita como na equação 4:

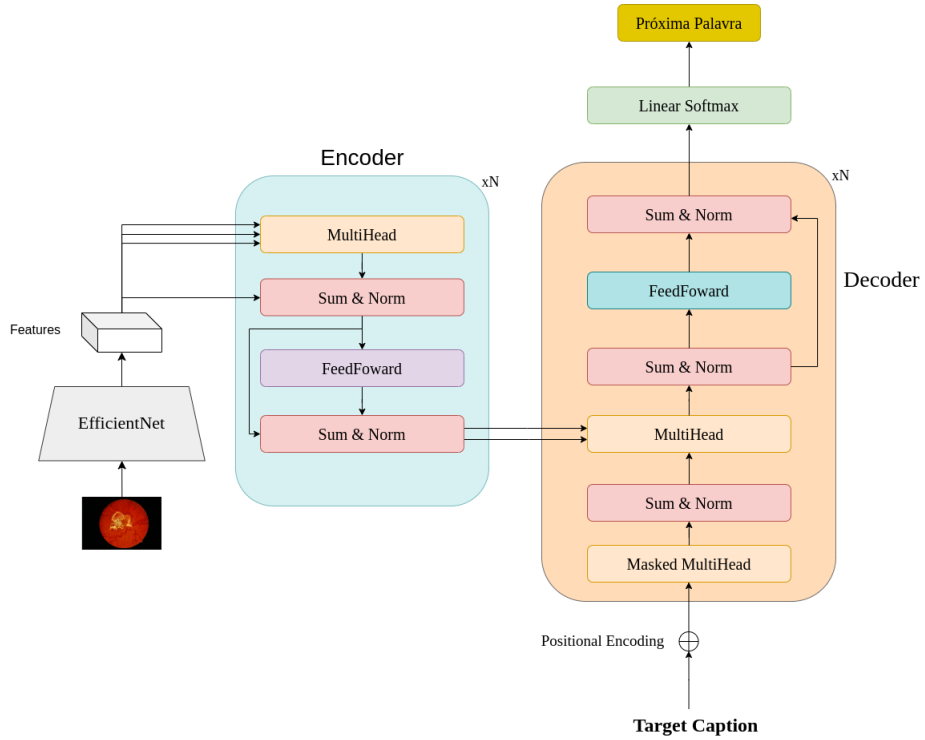


Figura 3. Etapas da arquitetura utilizada.

$$FeedForward = Linear(ReLU(Linear(x))) \tag{4}$$

E finalmente, a palavra seguinte é gerada por meio de uma camada totalmente conectada seguido da função de ativação *Softmax*.

A função de *Loss* do método pode ser definida pela entropia cruzada entre a palavra gerada pelo modelo e a referência. Além disso, utiliza-se a técnica de *DoublyStochastic Attention*, proposto em [Xu et al. 2016].

Para mapear o texto de entrada para o espaço latente do *Decoder*, foi utilizado o método proposto por [Vaswani et al. 2017] de *Positional Encoding*. No qual o modelo aprende a posição relativa dos elementos na sequência, sem depender de *tokens* especiais.

3.3. Métricas de avaliação

Para validar os resultados, é necessário comparar o texto gerado e o relatório escrito por especialistas. No entanto, esta tarefa demanda tempo, visto que o conjunto de teste (Seção 3.1) possui mais de 3.000 elementos. Para avaliar a qualidade do texto produzido pelo método, utilizam-se das seguintes métricas de comparação textual: BLEU 1 a 4 [Papineni et al. 2002b], METEOR [Lavie and Denkowski 2009] e RougeL [Lin 2004]. Cada métrica possui nuances e focos diferentes, todas se baseiam no conceito fundamental de n-grama (Figura 4).

O *BLEU* (Bilingual Evaluation Understudy) [Papineni et al. 2002b] é amplamente utilizada para avaliar a qualidade de traduções de maneira automática. A métrica compara o texto gerado automaticamente com o texto feito por humanos. Sua pontuação é calculada com base na sobreposição de n-gramas entre a saída do modelo e as referências,

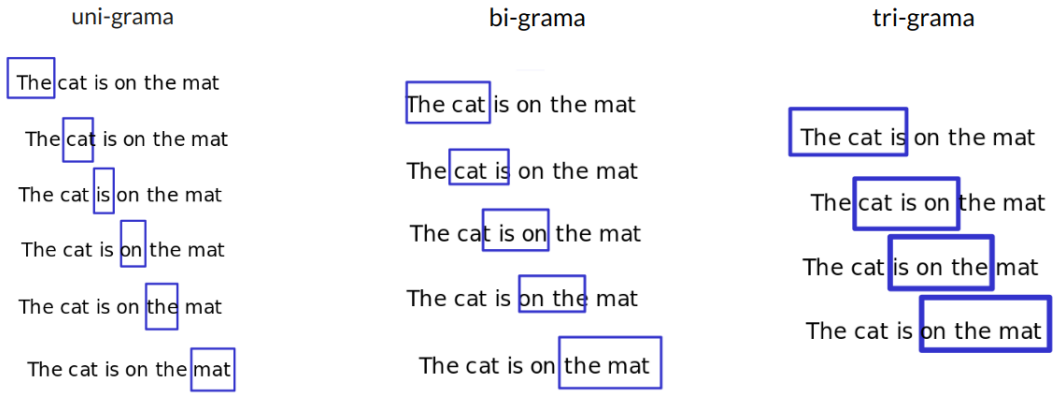


Figura 4. Visualização de N-gramas.

diminuindo a pontuação em caso de diferença entre as palavras presentes em cada uma das frases, também considerando sua ordem de disposição. O *METEOR* se destaca por sua abordagem única, mapeando pares de palavras iguais de maneira que se obtenha o menor número de colisões (Figura 5). O *METEOR* calcula sua pontuação considerando a precisão, além de priorizar resultados de melhor *recall*. O *ROUGE-L* gera o resultado através da maior subsequência comum, a precisão e *recall* são obtidas pelo tamanho da subsequencia dividido pelo tamanho da frase candidata no caso da precisão, e pelo tamanho da frase referencia no caso do *recall*, então o resultado será a média harmônica desses dois valores. Todas as métricas apresentadas pontuam os textos em um intervalo de 0 a 1, sendo que o valor 0 indica que os textos são completamente distintos, enquanto o valor 1 indica a similaridade entre os textos.

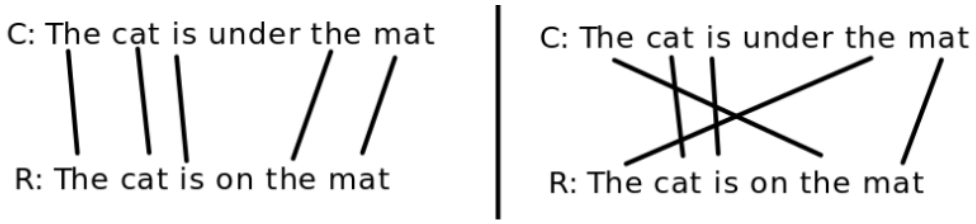


Figura 5. Mapeamento de uni-gramas.

4. Experimentos e Resultados

Nesta Seção, são apresentados os resultados obtidos pelo método proposto e verifica-se sua efetividade em relação a outros métodos já apresentados na literatura. Destaca-se que o propósito dos experimentos é aprimorar a qualidade dos relatórios médicos de retina gerados automaticamente.

Foi utilizada a base de dados *DeepEyeNet* (Seção 3.1) em sua divisão original para realizar uma comparação fiel com os demais métodos. A entrada da rede é dividida em pares de imagem e texto correspondente (Figura 1). As imagens foram redimensionadas utilizando a técnica de interpolação bilinear [Monasse 2019] para um tamanho padrão 256x256 com 3 canais de cores, o que garante a compatibilidade do modelo com todas as imagens da base utilizada.

Para realizar a escolha do *Backbone EfficienteNet*, foi realizado um estudo comparativo entre modelos pré-treinados de CNNs que demonstraram desempenho superior na tarefa de classificação de imagens na base de dados *ImageNet* [Deng et al. 2009].

Inicialmente, todas as redes foram treinadas para a tarefa de classificação na base de imagens *ImageNet*. As camadas finais de classificação foram removidas e foi realizado *fine-tuning* nas últimas camadas. O módulo *Transformer* é composto por 2 camadas de *Encoder* e 4 camadas de *Decoder*. A dimensão das características é de 512 e o número de cabeças de atenção é 8 tanto para ambos. O modelo foi treinado por 40 épocas, configurado com *EarlyStopping* em 10 épocas consecutivas sem melhoria na métrica *Bleu* 4, a redução de 25% da taxa de aprendizado ocorre a cada 5 épocas consecutivas onde o *Bleu* 4 não melhora. Ambas são feitas apenas no conjunto de validação. Foi utilizado o otimizador Adam [Kingma and Ba 2014] com *batch size* de 32.

As métricas apresentadas anteriormente (Seção 3.3) foram utilizadas para avaliar a efetividade do método proposto.

A Tabela 1, apresenta os resultados de cada *Backbone* em combinação com a rede *Transformer* apresentada. O método proposto apresenta a *EfficientNet* como Backbone, pois esta obteve os melhores resultados dentre os testes realizados. Além disso, os resultados mostram a estabilidade da rede *Transformer*, dado que cada *Backbone* possui arquitetura e desempenho específico.

Tabela 1. Diferentes redes convolucionais e desempenho como extrator de características.

Módulo extrator de características	Bleu 1	Bleu 2	Bleu 3	Bleu 4	Bleu Avg	ROUGE-L	METEOR
ResNext[Xie et al. 2017]	0,326	0,269	0,231	0,198	0,256	0,301	0,341
VGG16[Simonyan and Zisserman 2015]	0,333	0,274	0,235	0,201	0,261	0,307	0,353
DenseNet[Huang et al. 2018]	0,342	0,282	0,243	0,208	0,269	0,311	0,355
MobileNet[Howard et al. 2017]	0,347	0,285	0,245	0,210	0,272	0,313	0,361
ResNet101[He et al. 2015]	0,343	0,282	0,242	0,207	0,269	0,302	0,358
EfficientNet-v2[Tan and Le 2021]	0,355	0,292	0,250	0,212	0,277	0,323	0,366
Método proposto	**0,365**	**0,300**	**0,257**	**0,220**	**0,286**	**0,329**	**0,378**

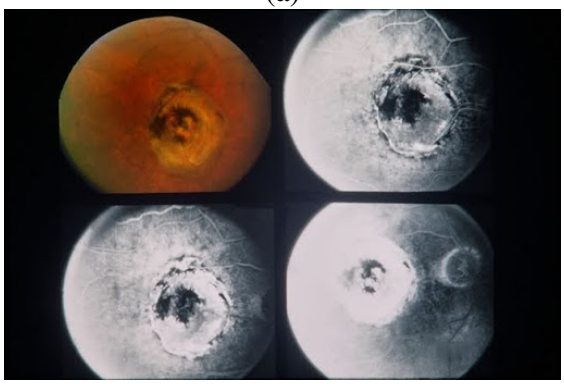
Na Tabela 2 compara-se o método com melhor resultado apresentado neste artigo com a proposta de outros autores, ambas avaliadas no mesmo conjunto de teste. É evidente o aprimoramento das métricas BLEU 1-4 em relação aos outros métodos. Isso mostra que a arquitetura proposta obteve progresso significativo em sugerir relatórios similares a um especialista.

Tabela 2. Comparação dos resultados.

Modelo	Bleu 1	Bleu 2	Bleu 3	Bleu 4	Bleu Avg	ROUGE-L	METEOR
[Huang et al. 2021b]	0.184	0.114	0.068	0.032	0.100	0.232	-
[Huang et al. 2021a]	0.219	0.134	0.074	0.035	0.116	0,252	-
[Huang et al. 2022]	0.230	0.150	0.094	0.053	0.132	0,291	-
[Shaik and Cherukuri 2024]	0,297	0,230	0,214	0,142	0.221	**0,391**	-
Método proposto	**0,365**	**0,300**	**0,257**	**0,220**	**0,286**	**0,329**	**0,378**

Na Figura 6 (a) e (b) são apresentados casos de sucesso onde o método apresenta a doença correta e produz texto com alta semelhança, em relação a um especialista, para

(a)

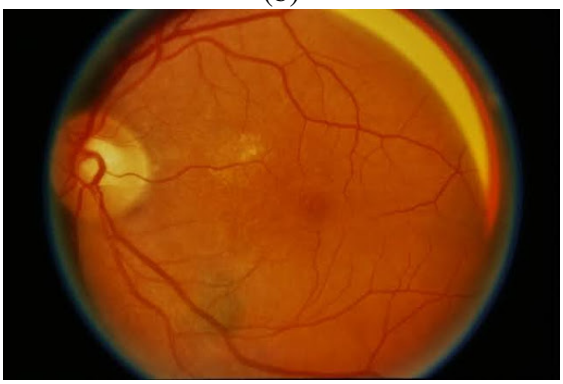


Predição: 74-year-old white female. rpe tear and srnv-md. re 20/100 le 10/200.

Referência: 73-year-old white female. srnv-md. re 10/200 le 10/200.

Bleu 4: 0,543

(b)

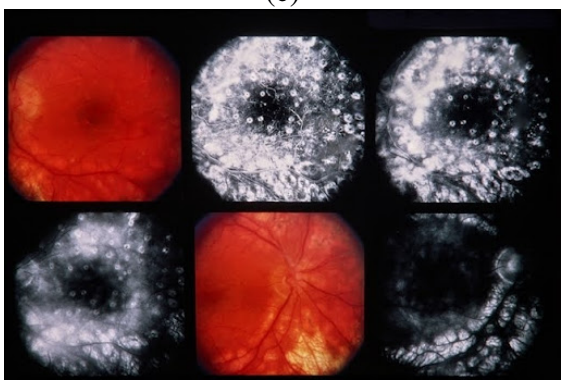


Predição: 66-year-old white male, icsc/severe.

Referência: 90-year-old white male, icsc/severe.

Bleu 4: 0,903

(c)



Predição: 22-year-old white male. coloboma. 20/13 and 20/25.

Referência: 53-year-old white female. 20/25 ou. proliferative diabetic retinopathy.

Bleu 4: 0,319

(d)

Predição: central serous retinopathy.

Referência: 38-year-old white male. re 20/70 le 20/20. central serous retinopathy.

Bleu 4: 0,014

Figura 6. Exemplos de predições geradas a partir do método proposto. Imagens e Referências por [Huang et al. 2021b]

um dos casos, errando apenas na idade do paciente. Na Figura 1 (c), é apresentado um exemplo em que o modelo falha quase que por completo, pontuando nas métricas apenas para palavras genéricas relacionadas à idade e sexo. A Figura 1 (d) apresenta um caso específico. A predição da doença foi realizada de forma adequada, no entanto, obteve uma métrica desfavorável por não seguir o mesmo padrão de escrita da referência. O método apresentado se limita em gerar texto conforme a base de treinamento, sem fazer a distinção entre palavras mais importantes que outras, como é o caso do nome das doenças, isso implica que haverá casos nos quais apesar de uma boa pontuação das métricas o método erra a doença. Por outro lado, os exemplos apresentados na Figura 6 ajudam a justificar a dificuldade de realizar *Image Captioning* no *DeepEyeNet*. A base não apresenta um padrão

definido para descrição feita por especialista, que atrapalha o treinamento do modelo e prejudica a obtenção de bons resultados no conjunto de testes.

5. Conclusão

Image Captioning de retina é uma tarefa desafiadora devido à escassez de bases públicas. Nesse contexto, métodos eficazes são necessários para produzir resultados precisos.

Este artigo propõe um método que aprimora a qualidade dos relatórios de imagem de retina gerados automaticamente para imagens do *DeepEyeNet*. Apresenta-se o *Backbone* da *EfficientNet* e a rede *Transformer* descrita na Seção 3.2.1.

Os resultados obtidos demonstram um desempenho promissor ao utilizar técnicas que melhoram a captura de informações obtidas por meio de uma CNN, combinando mecanismos de *Self-Attention* com convoluções. Isso permite que o *Decoder* interprete e produza textos relevantes para os especialistas. O método proposto pode ser integrado a um sistema que auxilie o oftalmologista, de forma que, durante a análise da retina, o método produza uma descrição inicial, permitindo que o especialista avalie e complemente o relatório médico.

Como trabalhos futuros, sugere-se explorar outras maneiras de representar a imagem, tratando *Image Captioning* de maneira semelhante a *Sequence-to-Sequence* [Liu et al. 2021]. Nesta abordagem não há necessidade de realizar a etapa de *Encoding*, isso retira do modelo uma série de convoluções reduzindo a complexidade e tempo de treinamento. Além disso, sugere-se utilizar das *Keywords* fornecidas pela base para garantir que o método aprenda também a acertar qual a doença presente na retina.

Agradecimentos

Ocultado para a revisão Os autores agradecem Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES),Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), e a Fundação de Amparo à Pesquisa Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) (Termo: 000527/2024), Empresa Brasileira de Serviços Hospitalares (Ebserh) Brasil (Grant number 409593/2021-4) pelo financiamento.

Referências

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

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Hendrick AM, Gibson MV, K. A. (2015). Diabetic retinopathy. *Prim Care*.

Herdade, S., Kappeler, A., Boakye, K., and Soares, J. (2020). Image captioning: Transforming objects into words.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2018). Densely connected convolutional networks.

Huang, J.-H., Wu, T.-W., Yang, C.-H. H., Shi, Z., Lin, I.-H., Tegner, J., and Worring, M. (2022). Non-local attention improves description generation for retinal images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1606–1615.

Huang, J.-H., Wu, T.-W., Yang, C.-H. H., and Worring, M. (2021a). Deep context-encoding network for retinal image captioning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3762–3766.

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. (2021b). 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.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lavie, A. and Denkowski, M. J. (2009). The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2–3):105–115.

Li, G., Zhu, L., Liu, P., and Yang, Y. (2019). Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8928–8937.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Liu, W., Chen, S., Guo, L., Zhu, X., and Liu, J. (2021). Cptr: Full transformer network for image captioning.

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.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002a). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002b). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pavlopoulos, J., Kougia, V., Androutsopoulos, I., and Papamichail, D. (2022). Diagnostic captioning: a survey. *Knowledge and Information Systems*, 64(7):1691–1722.

Shaik, N. S. and Cherukuri, T. K. (2024). Gated contextual transformer network for multi-modal retinal image clinical description generation. *Image and Vision Computing*, page 104946.

Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., and Summers, R. M. (2016). Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Steinmetz, J. D., Bourne, R. R., Briant, P. S., Flaxman, S. R., Taylor, H. R., Jonas, J. B., Abdoli, A. A., Abrha, W. A., Abualhasan, A., Abu-Gharbieh, E. G., et al. (2021). Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to vision 2020: the right to sight: an analysis for the global burden of disease study. *The Lancet Global Health*, 9(2):e144–e160.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Tan, M. and Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training.

Umbelino, C. C. and Ávila, M. P. (2023). As condições de saúde ocular no brasil. *São Paulo: Conselho Brasileiro de Oftalmologia*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

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

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2016). Show, attend and tell: Neural image caption generation with visual attention.

Zhang, Z., Xie, Y., Xing, F., McGough, M., and Yang, L. (2017). Mdnet: A semantically and visually interpretable medical image diagnosis network.