Aplicação de Visual Question Answering em dados médicos - Projeto Ciência de Dados V

Carlos Dias Maia[1]

[1]Pontifícia Universidade Católica de Minas Gerais
Instituto de Ciências Exatas e Informática ICEI
Curso de Graduação em Ciência de Dados - Campus Praça da Liberdade
Belo Horizonte – MG – Brazil

Abstract. *This study explores the use of deep neural networks for the Visual Question Answering (VQA) task in the medical domain. We utilized a combination of ResNet-50 for visual feature extraction from medical images and BERT for interpreting natural language questions. The fusion of these features was performed to generate accurate answers. Data were sourced from the VQAMed2019 dataset, comprising thousands of images and question-answer pairs. The experiments evaluated the model's performance with and without fine-tuning, using the Adam optimizer and regularization techniques such as dropout. Results indicate that fine-tuning significantly enhances model accuracy, demonstrating the proposed approach's effectiveness for VQA in the medical context.*

Resumo. *O presente trabalho explora a utilização de redes neurais profundas para a tarefa de Visual Question Answering (VQA) no domínio médico. Utilizamos uma combinação de ResNet-50 para a extração de características visuais das imagens médicas e BERT para a interpretação de perguntas em linguagem natural. A fusão dessas características é realizada para produzir respostas precisas. Os dados foram extraídos do conjunto VQAMed2019, compreendendo milhares de imagens e pares de perguntas e respostas. Os experimentos conduzidos avaliaram o desempenho do modelo com e sem fine-tuning, utilizando o otimizador Adam e técnicas de regularização como dropout. Os resultados mostram que o fine-tuning melhora significativamente a precisão do modelo, demonstrando a eficácia da abordagem proposta para a tarefa de VQA no contexto médico.*

1. Introdução

Sistemas baseados em Visual Question Answering (VQA) formam um campo de pesquisa em aprendizado profundo que combina processamento de linguagem natural (NLP) e visão computacional para responder perguntas sobre imagens. O objetivo principal desse tipo de sistema é detectar as características presentes nas imagens e perguntas apresentadas ao modelo e produzir respostas de acordo com as características captadas.

No contexto médico, o Visual Question Answering Medical, ou VQA-Med, se destaca ao aplicar essas técnicas para interpretar imagens médicas e responder questões fundamentais para diagnósticos e tratamentos. A complexidade surge não apenas da diversidade das perguntas, que variam de binárias a contextuais, mas também da necessidade de integrar eficazmente informações visuais e textuais para uma resposta precisa. Esta integração é essencial para evitar simplificações excessivas de perguntas complexas ou complicações desnecessárias de questões simples, aspectos que frequentemente desafiam os sistemas atuais de VQA-Med [Bazi et al. 2023].

A relevância dessa tarefa não a torna menos complexa. A generalidade das perguntas que podem ser feitas pelo usuário final da aplicação é um exemplo dessa complexidade. As perguntas podem ter respostas de natureza binária (sim ou não) ou podem ser mais complexas, exigindo respostas detalhadas e dependentes de contexto. As técnicas de VQA-Med às vezes não conseguem diferenciar esses tipos de perguntas, resultando em um aumento desnecessário da complexidade de problemas simples ou em uma simplificação excessiva dos problemas mais complexos. Para lidar com essa variabilidade, é crucial implementar mecanismos de atenção que permitam compreender simultaneamente tanto o conteúdo visual das imagens quanto o conteúdo textual das perguntas. Esses mecanismos garantem uma contextualização adequada da pergunta em relação à imagem, aumentando a precisão e a relevância das respostas fornecidas [Yu et al. 2019, Gupta et al. 2021].

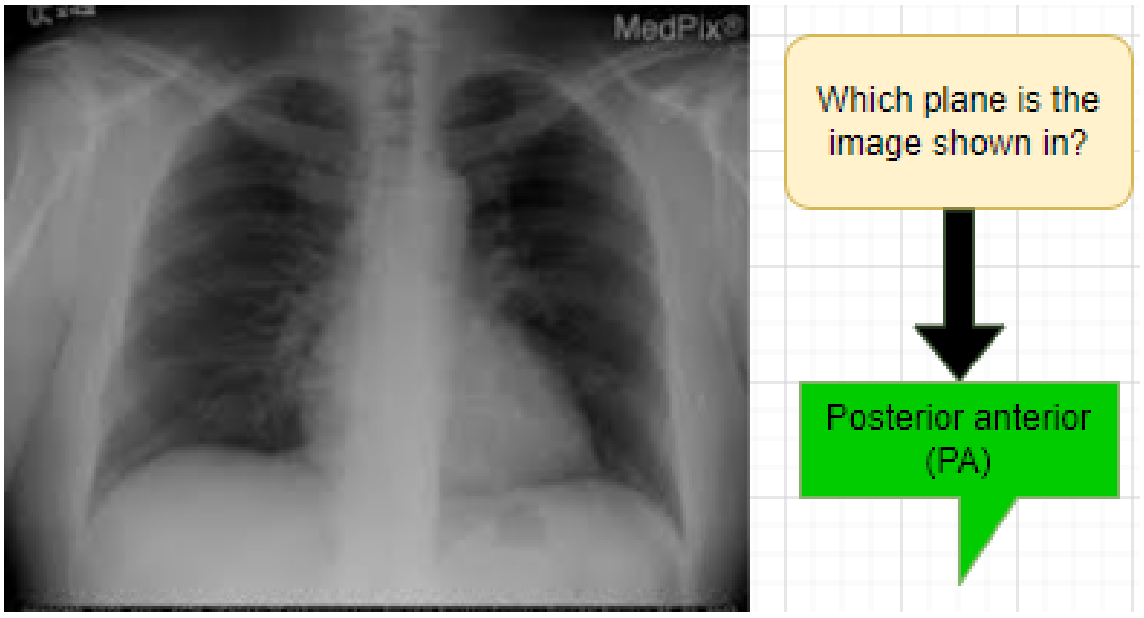

Figure 1. Exemplo de imagem do dataset com a pergunta e resposta correta

Neste trabalho, apresentamos uma aplicação de VQA-Med utilizando dados do ImageClef2019, com imagens exclusivamente do domínio médico, como demonstrado na Figura 1. Para a produção da aplicação, utilizamos o modelo pré-treinado de rede neural convolucional ResNet-50 para a tarefa de visão computacional e o modelo pré-treinado de linguagem natural BERT para a tarefa de perguntas e respostas. Utilizamos mecanismos de atenção em formato multi-head para recuperação de características ao longo da rede e uma fusão de atributos via concatenação ao final das duas redes. As features foram concatenadas e depois passaram por uma multi-head attention e, em seguida, por uma camada fully connected (FC). Além disso, realizamos uma avaliação do impacto de fazer um fine-tuning na rede de imagem.

O trabalho está estruturado nas seguintes seções: primeiramente, um referencial teórico sobre os modelos de VQA desenvolvidos para lidar especificamente com dados médicos e de forma geral é apresentado na Seção 2. Em seguida, na Seção 3, apresentamos o fluxograma da arquitetura do modelo, a origem dos dados utilizados, os modelos pré-treinados e seus ajustes (fine-tuning). Por fim, na Seção 4, apresentamos os resultados do modelo, incluindo a escolha de hiperparâmetros, métricas de perda e visualização de resultados.

2. Trabalhos Relacionados

Nesta seção, apresentamos um levantamento de modelos voltados para VQA-Med, organizando-os por temas para constituir um referencial teórico para o presente trabalho.

Trabalhos anteriores têm explorado diferentes abordagens para resolver desafios no VQA-Med, especialmente na fusão de atributos e no uso de mecanismos de atenção. Por exemplo, [Gupta et al. 2021] destacam a importância de hierarquizar perguntas através de SVM antes de integrá-las em redes neurais multimodais, enquanto [Vaswani et al. 2023] adaptam seus modelos com base na distribuição de dados específicos do domínio médico, utilizando fusão de atributos desde as primeiras camadas para melhorar a compreensão contextual das perguntas.

2.1. Fusão de Atributos

Em [Kornuta et al. 2019], os autores adaptaram o modelo de acordo com o conhecimento de domínio adquirido após analisar estatisticamente o conjunto de dados, moldando a arquitetura dos *encoders* para atender exclusivamente à distribuição de dados em conjunto com a tarefa apresentada. Eles delimitam os classificadores por categoria e apresentam uma fusão de atributos na entrada dos modelos de características visuais e textuais (early fusion). Os autores comentam que as respostas de algumas perguntas podem auxiliar no entendimento de outras perguntas mais complexas.

Situação semelhante é citada por [Gupta et al. 2021], que hierarquizam as perguntas/consultas dos usuários via SVM - Support Vector Machine antes de integrá-las ao modelo de rede neural profunda multimodal hierárquica, integrada por uma rede dupla de LSTMs - Long short-term memory. Contudo, diferentemente do primeiro trabalho, utiliza-se uma fusão tardia de atributos das imagens e das perguntas.

2.2. Estruturas Paralelas

Em [Liu et al. 2022], uma arquitetura semelhante é utilizada, mas com uma estrutura de ramificação em duas redes: a primeira para lidar com respostas candidatas de menor dificuldade de classificação (modelo de estrutura paralela) e, caso contrário, os dados são transferidos para o segundo ramo (modelo de recuperação de imagens).

2.3. Pré-processamento de Dados

O modelo de [Ren and Zhou 2020] utiliza pré-processamento de dados nas imagens e nos textos, além de extrair características de imagem com uma rede ResNet152 pré-treinada e utilizar várias *embeddings* para lidar com textos. Ele também reduz o número de parâmetros do transformador multi-head self-attention reduzindo o custo computacional.

2.4. Redes de Atenção

A abordagem de Rede Modular de Co-Atenção Profunda (MCAN) em [Yu et al. 2019] lida com o desafio utilizando camadas de Co-Atenção Modular (MCA) em cascata. Cada camada MCA modela a auto-atenção das perguntas e das imagens, além da atenção orientada das imagens, usando uma composição modular de duas unidades básicas de atenção.

No presente trabalho, optamos por abordar a tarefa proposta através de mecanismos de atenção, baseados no estudo seminal de Vaswani et al. [Vaswani et al. 2023], e aplicados especificamente em [Bazi et al. 2023]. Utilizamos uma arquitetura de codificador-decodificador, onde características da imagem são extraídas utilizando o modelo ViT (Vision Transformer), enquanto a pergunta é incorporada por meio de um codificador textual transformer. Posteriormente, as representações resultantes das modalidades visual e textual são concatenadas e alimentadas em um decodificador multimodal, permitindo a geração autoregressiva da resposta.

Os trabalhos relacionados constituem um referencial teórico crucial, fornecendo insights sobre como aplicar mecanismos de atenção eficazes em VQA. Eles destacam a importância do momento de fusão de atributos para integrar informações visuais e textuais de maneira coerente, além de oferecer estratégias para lidar com o contexto de informação em tarefas multimodais como esta. Esses estudos não apenas orientam a escolha e implementação dos componentes arquiteturais, mas também contribuem significativamente para a compreensão das melhores práticas em sistemas de Visual Question Answering.

3. Arquitetura da Rede Neural

Neste trabalho, é proposta uma arquitetura que utiliza redes neurais profundas para extrair as características visuais e textuais dos dados de entrada, que são então combinadas por meio de um módulo de fusão. A arquitetura combina uma ResNet-50 pré-treinada para extração de características visuais e um modelo BERT (Bidirectional Encoder Representations for Transformers) para processamento de linguagem natural. A ResNet-50 opera através de cinco convoluções principais, que produzem mapas de características em resoluções progressivamente menores, permitindo uma captura detalhada e hierárquica das características visuais nas imagens. Após o módulo de fusão, um classificador fully-connected gera a resposta final. Na Figura 2 é ilustrada a arquitetura da solução proposta.

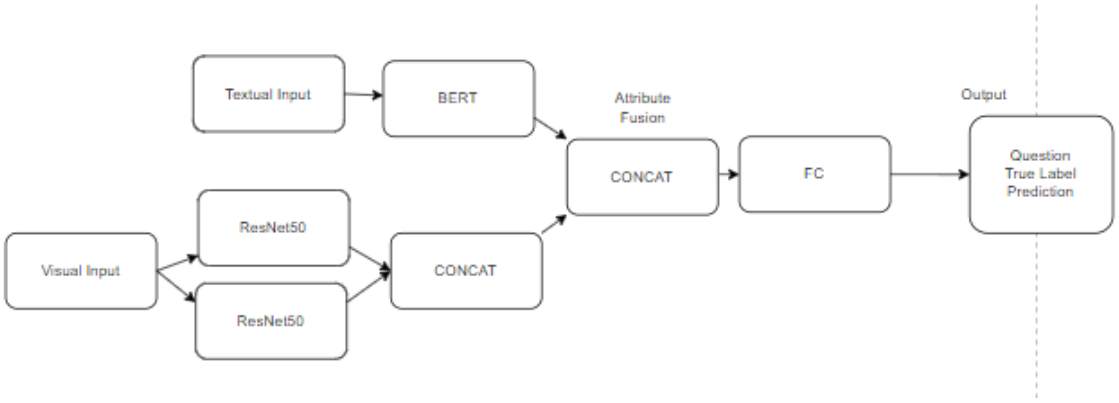

Figure 2. Arquitetura Básica da Rede

Após a extração de características visuais e textuais, essas são concatenadas, permitindo uma fusão de atributos. Essa concatenação é alimentada em um decodificador multimodal, que realiza a geração autoregressiva de respostas, garantindo que a saída seja coerente e contextualizada tanto em termos visuais quanto textuais. A eficácia dessa fusão de atributos é crucial para garantir uma interpretação precisa das imagens médicas no contexto de perguntas e respostas, conforme destacado por [Yu et al. 2019].

As abordagens empregadas foram projetadas com o objetivo de maximizar a precisão e a generalização do modelo de Visual Question Answering no domínio médico. Isso é alcançado através da utilização de técnicas robustas e adaptáveis, que respondem aos desafios específicos apresentados pelos conjuntos de dados médicos e pela complexidade das questões visuais médicas. Ao combinar essas duas técnicas de extração de características, a arquitetura proposta é capaz de fornecer respostas precisas e contextualmente relevantes, potencializando o uso de VQA em aplicações médicas e facilitando a tomada de decisões clínicas.

3.1. Conjunto de Dados

O conjunto de dados utilizado, proveniente do *Visual Question Answering in the Medical Domain - VQAMed2019*, consiste em três conjuntos principais: treinamento, validação e teste. O conjunto de treinamento contém 3.200 imagens médicas, cada uma acompanhada por múltiplos pares de perguntas e respostas (QA), totalizando 12.792 pares. Para validação, utilizamos 500 imagens com 2.000 pares QA, enquanto o conjunto de teste inclui 500 imagens com 500 perguntas. Para garantir uma avaliação significativa, as perguntas foram categorizadas em quatro tipos principais: Modalidade, Plano, Sistema de Órgãos e Anormalidade.

3.2. Extração de Características Visuais

Para a tarefa de visão computacional, empregamos a ResNet-50 pré-treinada. Inicialmente, as imagens são alimentadas na ResNet-50, que extrai características visuais. Uma camada totalmente conectada converte as características de entrada em um vetor de 256 dimensões, seguida por uma função de ativação ReLU para introduzir não-linearidade. Para mitigar o sobreajuste, aplicamos uma camada de dropout com uma probabilidade de 0.2 durante o treinamento. Após outra camada totalmente conectada para refinamento,

uma função LogSoftmax é aplicada para produzir saídas probabilísticas, ideal para tarefas de classificação.

3.3. Extração de Características Textuais

Utilizamos o BERT para a tarefa de perguntas e respostas. Esta arquitetura compreende duas partes principais:

O **Tokenizer** é responsável por dividir o texto em tokens e convertê-los em índices compreendidos pelo modelo. Isso é crucial para processar textos em linguagem natural de maneira eficiente.

O **BERT** é um modelo de linguagem baseado em transformadores, pré-treinado em grandes volumes de texto para entender o contexto das palavras em uma frase considerando sua sequência em ambas as direções. Essa capacidade de interpretação contextual permite ao BERT processar o texto de entrada de maneira altamente precisa e informada. Utilizamos o modelo 'bert-base-uncased' da Hugging Face, treinado em texto sem diferenciação de maiúsculas e minúsculas (uncased).

3.4. Fusão de Atributos

As características visuais e textuais são extraídas das imagens utilizando a ResNet-50 e o BERT, respectivamente. Após a extração, as características resultantes são achatadas para facilitar a manipulação e, em seguida, concatenadas em um vetor unificado. Este vetor unificado é então processado por um mecanismo de Multi-Head Attention, que permite a combinação de informações de diferentes partes da entrada, aumentando a capacidade do modelo de capturar relações complexas entre as características visuais e textuais. Finalmente, o vetor passa pela camada linear de fusão, que produz uma representação final combinada das imagens, incorporando tanto os aspectos visuais quanto linguísticos.

4. Resultados

4.1. Setup de Experimentos

Nesta subseção, descrevemos o protocolo de treinamento da rede, os otimizadores utilizados, a escolha e valores dos hiperparâmetros, entre outros detalhes importantes.

Para treinar o modelo de Visual Question Answering (VQA) no domínio médico, adotamos um protocolo estruturado para maximizar o desempenho e a generalização do modelo. Inicialmente, a rede neural foi inicializada com pesos pré-treinados da ResNet-50 para a extração de características visuais e do BERT ('bert-base-uncased') para o processamento de linguagem natural.

Durante o treinamento, aplicamos o fine-tuning nos pesos da ResNet-50 e do BERT para adaptá-los às características específicas do conjunto de dados VQAMed2019. Isso permitiu que os modelos aprendessem representações mais eficazes para as imagens médicas e as perguntas associadas.

O otimizador escolhido para ajustar os pesos da rede foi o AdamW, com uma taxa de aprendizado inicial de 0.001. O AdamW foi selecionado por sua eficácia comprovada em otimização de redes neurais profundas, permitindo uma convergência mais rápida durante o treinamento.

A função de perda utilizada foi a entropia cruzada, adequada para problemas de classificação como o VQA. Também incorporamos uma camada de dropout com probabilidade de 0.2 para regularizar a rede e mitigar o overfitting, garantindo que o modelo não se ajustasse excessivamente aos dados de treinamento.

O modelo foi treinado ao longo de 10 épocas. Após 4 épocas, o incremento de ganho tornou-se insignificante, e para economizar recursos computacionais e evitar overfitting, utilizamos uma função de parada antecipada (early stopping). Essa técnica interrompe o treinamento quando o modelo atinge um desempenho satisfatório, evitando o desperdício de tempo e recursos. Esses aspectos foram cuidadosamente ajustados para garantir resultados robustos e generalizáveis, essenciais para a tarefa desafiadora de responder questões visuais em um contexto médico.

Este protocolo experimental foi fundamental para estabelecer uma base sólida de desempenho do modelo, preparando-o para as análises e resultados detalhados apresentados nas seções seguintes.

4.2. Resultados de Treinamento e Validação

Modelo - Sem Fine-Tuning				
Epochs	Train Loss	Train Accuracy	Validation Loss	Validation Accuracy
1	1.7290	0.5801	**2.9945**	0.5120
2	0.9214	0.7091	3.0827	0.5460
3	0.7648	0.7497	3.1916	0.5410
4	**0.6373**	0.7867	3.3331	0.5575
Modelo - Com Fine-Tuning				
Epochs	Train Loss	Train Accuracy	Validation Loss	Validation Accuracy
1	1.6288	0.6031	2.3341	0.5210
2	0.8421	0.7671	**2.0421**	0.5550
3	0.7306	0.7897	2.6785	0.5521
4	**0.5934**	0.8068	3.7896	0.5625

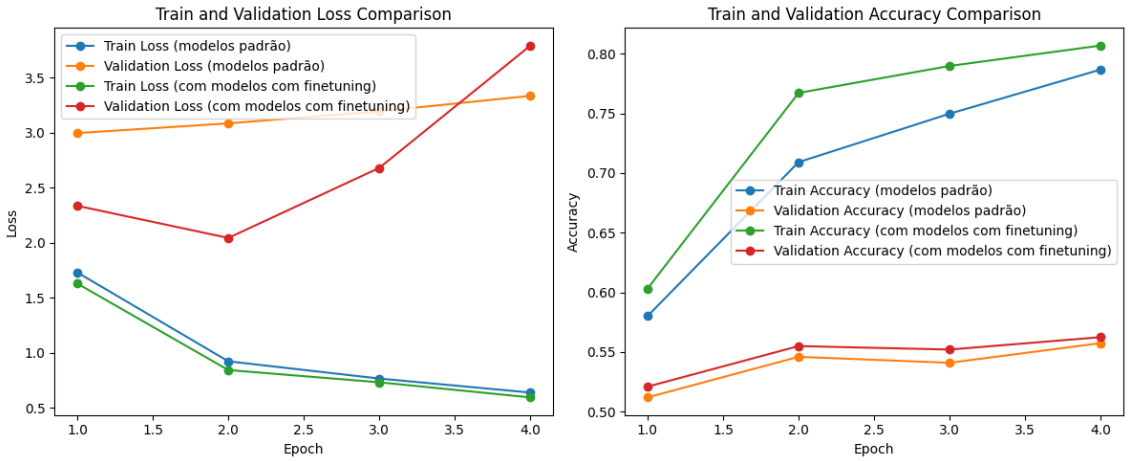

Figure 3. Apresentação gráfica dos dados tabulares

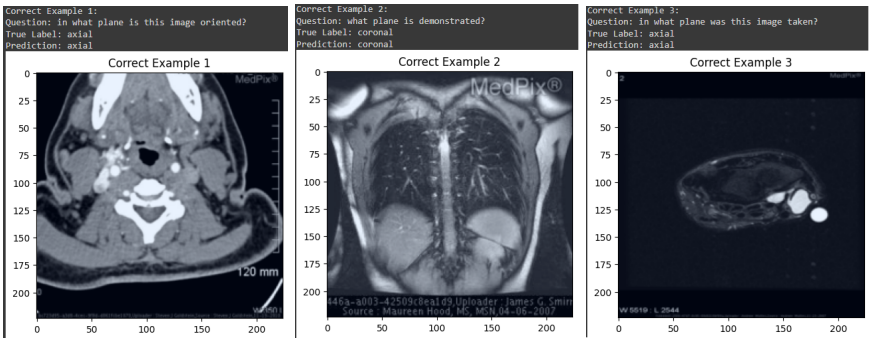

Figure 4. Resultados de acerto do Modelo

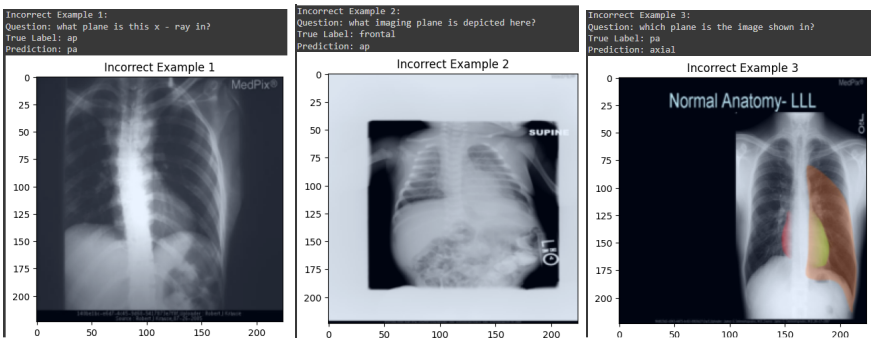

Figure 5. Resultados de erro do Modelo

4.3. Discussão dos Resultados

Os experimentos revelaram que o fine-tuning das redes ResNet-50 e BERT resultou em melhorias significativas na acurácia de validação ao longo das épocas de treinamento. A escolha do otimizador AdamW, com uma taxa de aprendizado adaptável, e a incorporação de dropout para regularização foram fundamentais para mitigar o overfitting e garantir a generalização do modelo.

Observa-se que o modelo com fine-tuning apresenta uma melhoria consistente na acurácia de validação em comparação ao modelo sem fine-tuning. As Figuras 3, 4, e 5 mostram as curvas de perda, exemplos de acertos e erros do modelo, respectivamente, proporcionando uma visão abrangente sobre o comportamento e a eficácia do modelo treinado. A figuração gráfica dos resultados detalha a evolução das métricas de treinamento e validação, fornecendo insights valiosos sobre o desempenho do modelo em diferentes fases do treinamento.

No entanto, devido à presença de amostras na validação que não estão presentes no treinamento, a perda de validação aumenta com as épocas. Isso ocorre porque o modelo começa a ajustar-se cada vez mais às características específicas das amostras de treinamento, resultando em um fenômeno conhecido como overfitting. Como consequência, o modelo se torna menos capaz de generalizar para novas amostras não vistas durante o treinamento. Esse aumento na perda de validação indica que, à medida que o treinamento avança, o modelo se adapta demais aos dados de treinamento, perdendo sua capacidade de responder adequadamente a dados novos e não familiares.

5. Conclusão

Este estudo apresentou uma abordagem para a tarefa de Visual Question Answering (VQA) no contexto médico, combinando a potência da ResNet-50 para extração de características visuais com as capacidades interpretativas do BERT para linguagem natural. Através de uma análise rigorosa dos resultados, verificamos que o fine-tuning significativamente aprimorou o desempenho do modelo, resultando em maior precisão nas respostas a perguntas baseadas em imagens médicas.

Os experimentos destacaram que a integração eficaz de características visuais e textuais, seguida de uma fusão bem-sucedida, é uma estratégia promissora para resolver desafios complexos de VQA no domínio médico. Além disso, a aplicação de técnicas de regularização, como o dropout, e a seleção criteriosa de hiperparâmetros, desempenharam um papel crucial na robustez do modelo, mitigando o risco de overfitting.

Os resultados obtidos estabelecem uma base sólida para futuras investigações nesta área, sugerindo que avanços adicionais podem ser alcançados explorando arquiteturas mais avançadas e ampliando os conjuntos de dados disponíveis. Esta pesquisa contribui não apenas para o campo específico de VQA médico, mas também para o desenvolvimento de aplicações de inteligência artificial que integram dados multimodais de forma eficiente e precisa.

Para futuras melhorias, propomos a implementação das seguintes estratégias:

1. **Pré-processamento de Dados**: Implementar técnicas avançadas de pré-processamento de dados para melhorar a qualidade das entradas do modelo. Isso pode incluir normalização, augmentação de dados e remoção de ruídos.
2. **Batch Normalization**: Substituir o *dropout* por *batch normalization* e adicionar mais camadas de *batch normalization* ao longo da rede. Esta abordagem pode ajudar a estabilizar e acelerar o treinamento, além de melhorar a regularização do modelo.
3. **Ensemble de Modelos**: Realizar um *ensemble* de modelos utilizando diferentes modelos pré-treinados. Isso inclui a realização de testes abrangentes com outros modelos pré-treinados, com foco especial no *Inception*, que tem demonstrado excelente desempenho em várias tarefas de visão computacional.
4. **Alteração no Tipo de Saída**: Alterar o tipo de saída do modelo de um rótulo único para uma representação embutida (*embedded*), permitindo uma resposta mais rica e contextual às perguntas visuais.
5. **Mitigar o Problema da Perda de Validação**: A implementação de técnicas como validação cruzada e ajuste de hiperparâmetros pode ajudar com o problema onde a perda de validação aumenta com as épocas, promovendo uma melhor generalização e um desempenho mais robusto em dados de validação.

References

Bazi, Y., Rahhal, M. M. A., Bashmal, L., and Zuair, M. (2023). Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10(3).

Gupta, D., Suman, S., and Ekbal, A. (2021). Hierarchical deep multi-modal network for medical visual question answering. *Expert Systems with Applications*, 164:113993.

Kornuta, T., Rajan, D., Shivade, C., Asseman, A., and Ozcan, A. S. (2019). Leveraging medical visual question answering with supporting facts.

Liu, S., Zhang, X., Zhou, X., et al. (2022). Bpi-mvqa: a bi-branch model for medical visual question answering. *BMC Med Imaging*, 22(1):79.

Ren, F. and Zhou, Y. (2020). Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access*, 8:50626–50636.

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

Yu, Z., Yu, J., Cui, Y., Tao, D., and Tian, Q. (2019). Deep modular co-attention networks for visual question answering.