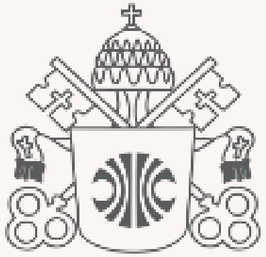


PUC Minas

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


Bruno Petrocchi de Sena Azevedo
Carlos Dias Maia
Nicolau Machado de Carvalho
Ranier Pereira Nunes de Melo
Samuel Fernandes Teixeira Lages

# Introdução

**Visual Question Answering (VQA):**

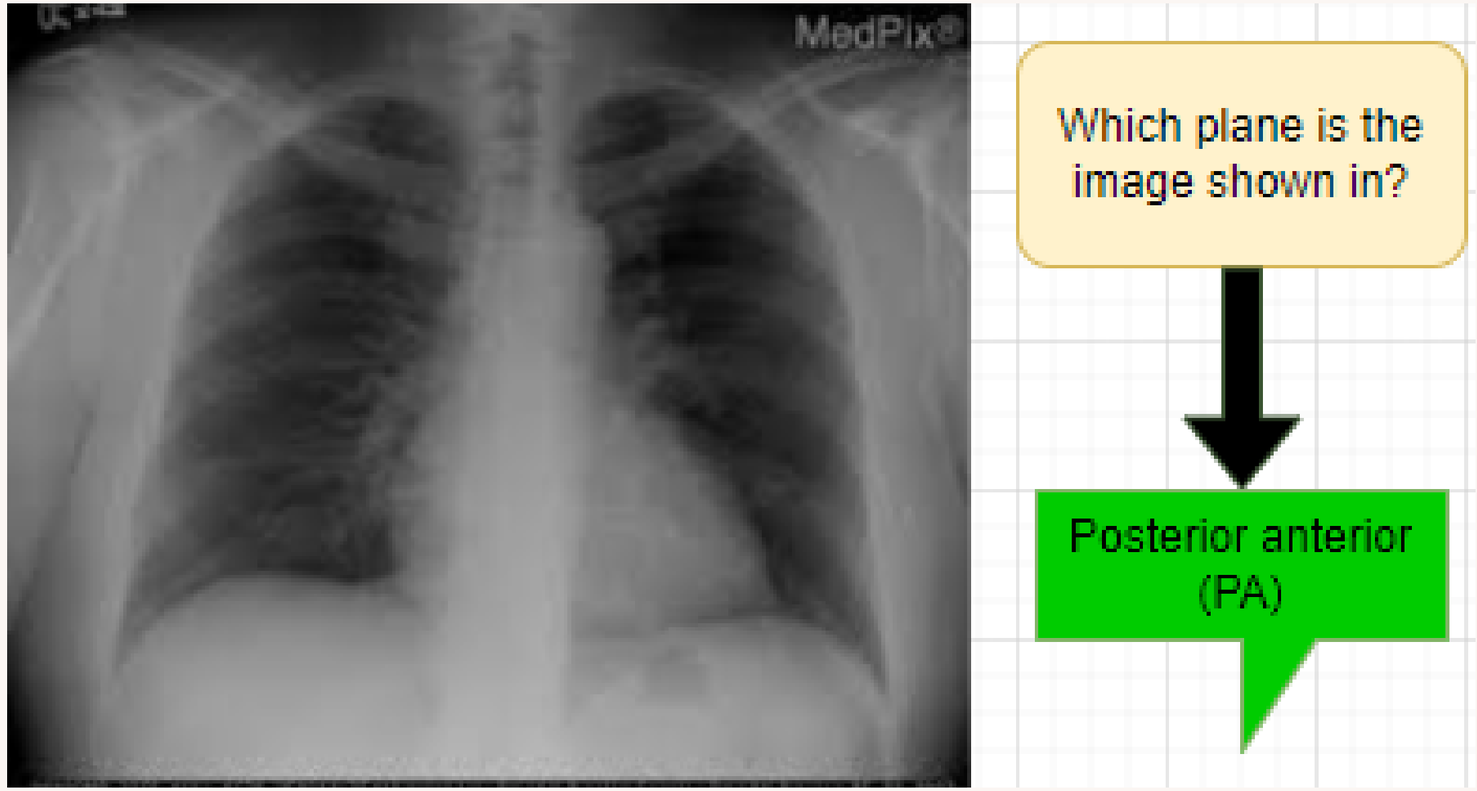
- **Combina processamento de linguagem natural (NLP) e visão computacional.**
- **Responde perguntas sobre imagens ao detectar características presentes nas imagens e perguntas.**

**VQA-Med:**

- **Aplica técnicas de VQA para interpretar imagens médicas.**
- **Responde questões fundamentais para diagnósticos e tratamentos.**
- **Enfrenta complexidade devido à diversidade das perguntas e à integração de informações visuais e textuais.**

**Desafios do VQA-Med:**

- **Perguntas variam de binárias (sim/não) a contextuais e complexas.**
- **Necessidade de evitar simplificações excessivas de perguntas complexas e complicações desnecessárias de perguntas simples.**



# Trabalhos Relacionados

**Hierarquização e Fusões de Atributos:**

- **Gupta et al. 2021: Importância de hierarquizar perguntas com SVM antes de integrá-las em redes neurais multimodais.**
- **Vaswani et al. 2023: Adaptação de modelos com base na distribuição de dados médicos específicos e uso de fusão de atributos desde as primeiras camadas.**
- **Kornuta et al. 2019: Adaptação do modelo com base em análise estatística do conjunto de dados e uso de fusão de atributos na entrada dos modelos.**

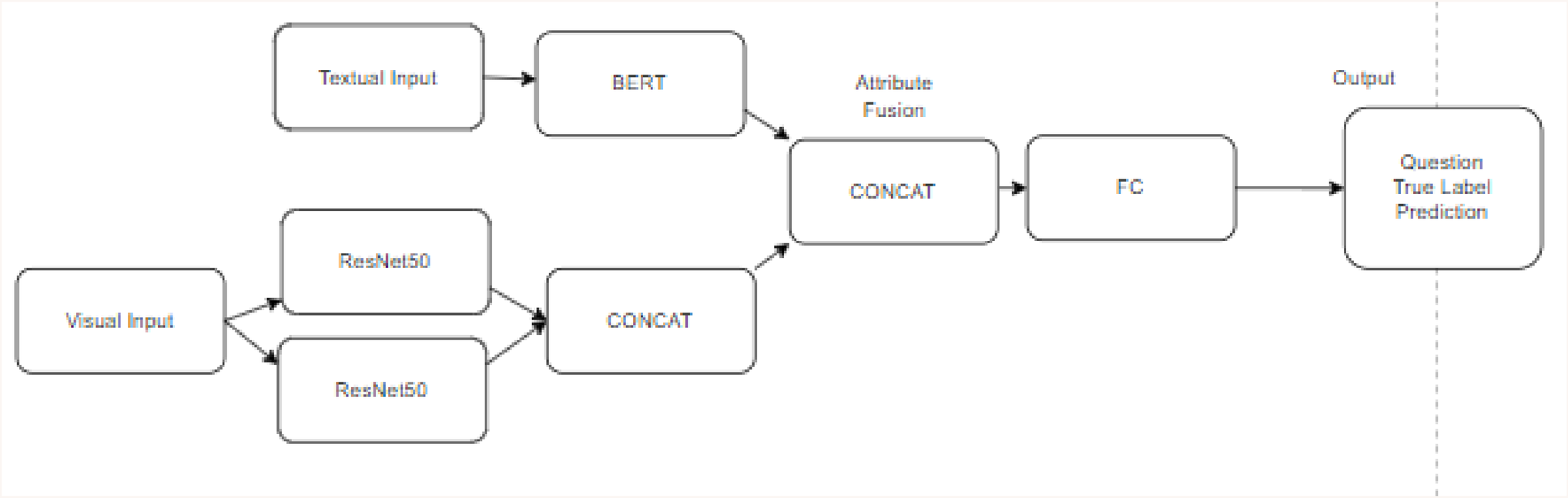
**Redes de Atenção:**

- **Yu et al. 2019: Abordagem de Rede Modular de Co-Atenção Profunda (MCAN) com camadas de Co-Atenção Modular (MCA) em cascata, modelando auto-atenção das perguntas e das imagens, além de atenção orientada das imagens.**

**Metodologia do Presente Trabalho:**

- **Utilização de mecanismos de atenção baseados em Vaswani et al. 2023, aplicados especificamente em Bazi et al. 2023.**
- **Arquitetura de codificador-decodificador com características da imagem extraídas pelo modelo ViT (Vision Transformer) e perguntas incorporadas por meio de um codificador textual transformer.**
- **Representações visuais e textuais concatenadas e alimentadas em um decodificador multimodal para geração autoregressiva da resposta.**

# Arquitetura da Rede Neural



# Arquitetura da Rede Neural

- **A arquitetura utiliza redes neurais profundas para extrair características visuais e textuais dos dados de entrada, que são  combinadas por meio de um modulo de fusão;**

- **Utiliza-se 2 redes ResNet50 pré-treinadas e soma os gradientes, concatenando a saída de ambas e utilizando como entrada para o módulo de fusão;**

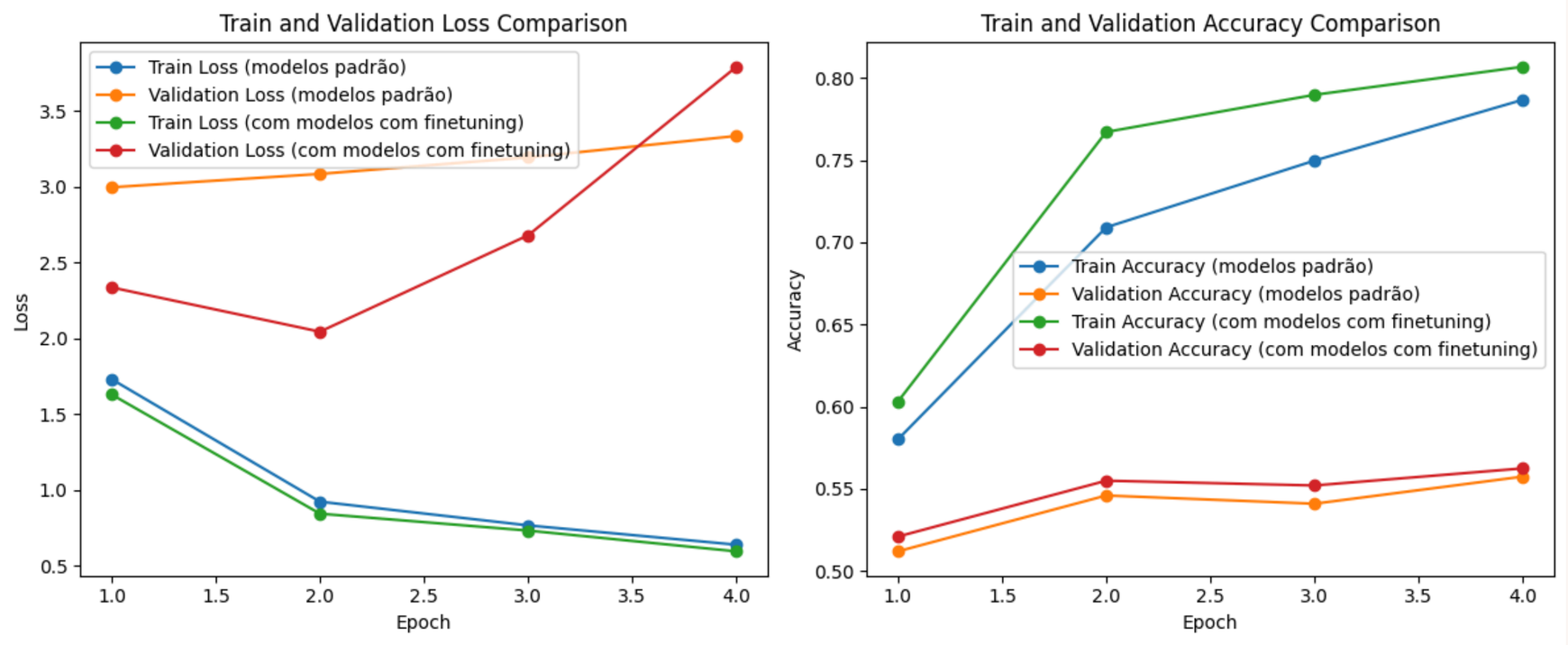
- **Para processamento da linguagem natural, utilizamos um modelo BERT (Bidirectional Encoder Representations for Transformers)**

- **Processo de Extração e Fusão de Características: Após concatenar as características, aplica-se Multi-Head Attention para capturar relações complexas e finaliza com uma camada linear para uma representação combinada.**

# Resultados

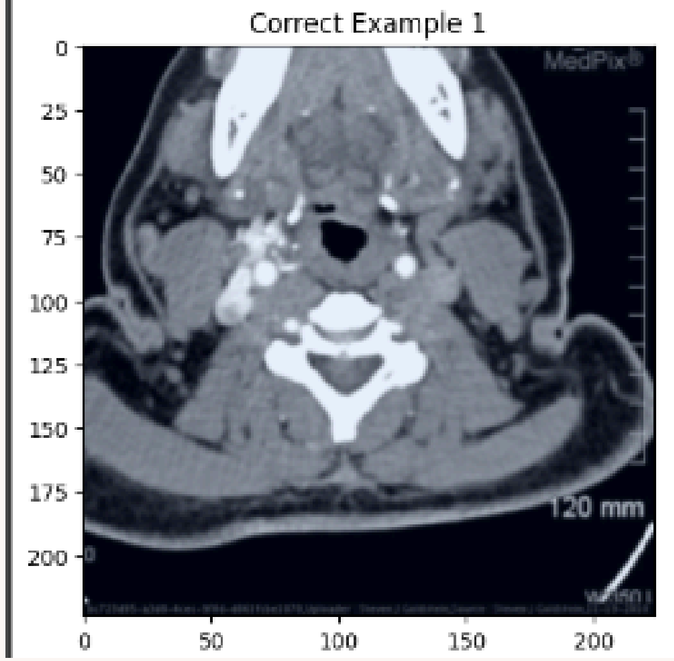
| Modelo - Sem Fine-Tuning | | | | |
|---|---|---|---|---|
| Epochs | Train Loss | Train Accuracy | Validation Loss | Validation Accuracy |
| 1 | 1.7290 | 0.5801 | **2.9945** | 0.5120 |
| 2 | 0.9214 | 0.7091 | 3.0827 | 0.5460 |
| 3 | 0.7648 | 0.7497 | 3.1916 | 0.5410 |
| 4 | **0.6373** | 0.7867 | 3.3331 | 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 |

# Validação e Treino - Métricas

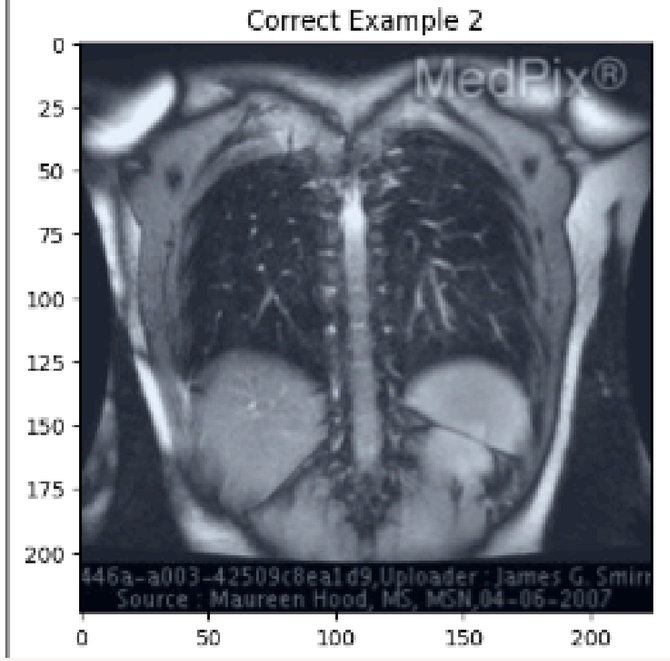


# Acertos do modelo

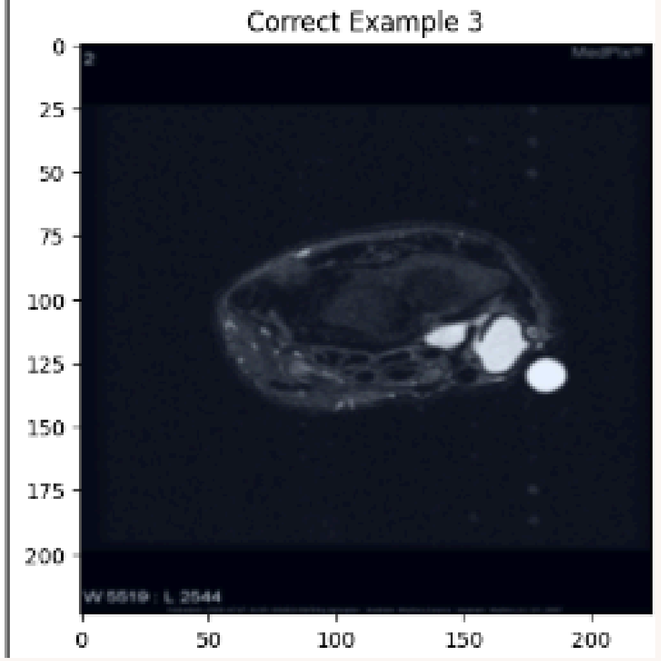
```
Correct Example 1:
Question: in what plane is this image oriented?
True Label: axial
Prediction: axial
```



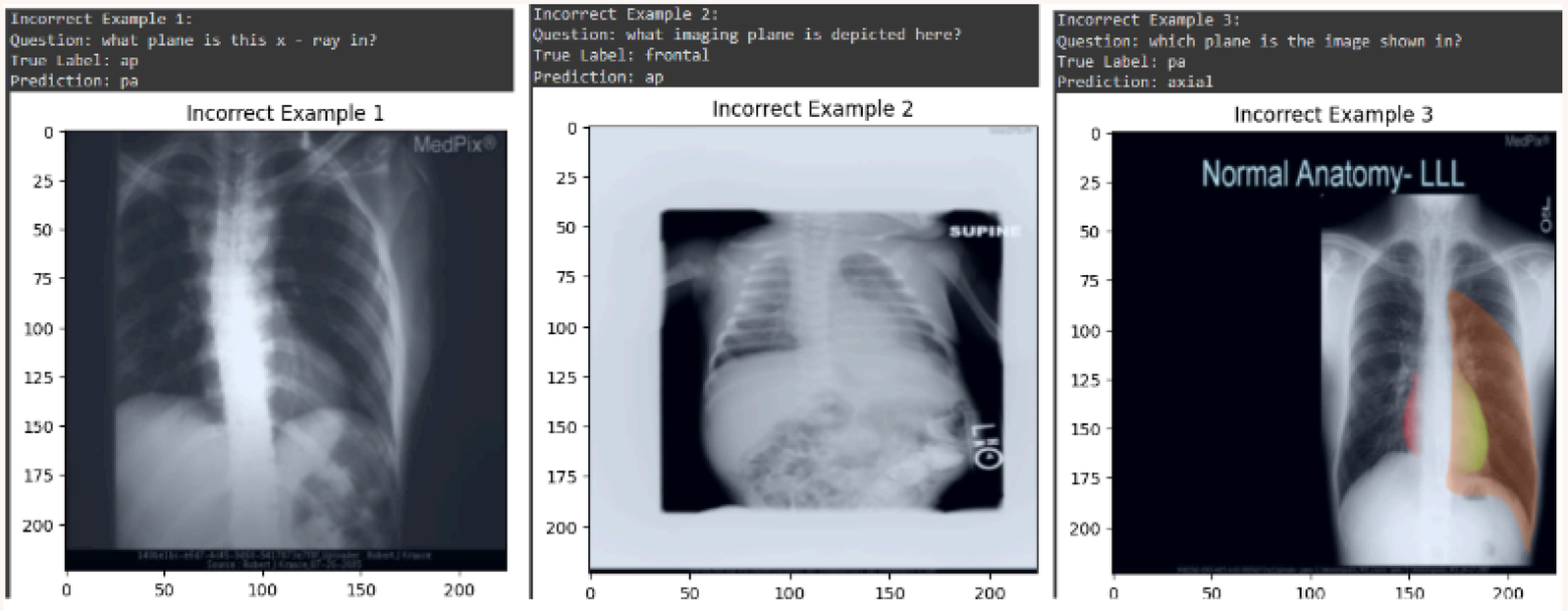

```
Correct Example 2:
Question: what plane is demonstrated?
True Label: coronal
Prediction: coronal
```




```
Correct Example 3:
Question: in what plane was this image taken?
True Label: axial
Prediction: axial
```



# Erros do modelo



# Possíveis Melhorias

- Pré-processamento de dados
- Batch Normalization
- Ensemble de modelos
- Alteração no tipo de saída
- Mitigar o problema da perda de validação