Bom dia/tarde a todos. Meu nome é **Felipe Estrada Nunes da Silva**, e hoje apresentarei meu Trabalho de Conclusão de Curso, cujo tema é **"Implementação e Avaliação de Redes Neurais Compactas para Detecção de Catarata com Dados Limitados"**.

A **catarata** é uma doença caracterizada pela **opacificação do cristalino**, reduzindo a passagem de luz até a retina e comprometendo a visão. É a **principal causa de cegueira reversível no mundo** e está associada a fatores como:

* **Envelhecimento**
* **Diabetes**
* **Predisposição genética**
* **Exposição prolongada à radiação ultravioleta**

A doença tem um impacto severo na qualidade de vida, dificultando tarefas simples como leitura e direção. Embora **a cirurgia seja a solução definitiva**, um **diagnóstico precoce** pode permitir medidas preventivas, retardando sua progressão.

Os métodos convencionais de diagnóstico de catarata dependem de exames clínicos presenciais realizados por **oftalmologistas especializados**. No entanto:

* Há **déficit de especialistas** em diversas regiões, principalmente no interior e em países em desenvolvimento.
* A necessidade de exames **presenciais limita o diagnóstico precoce**.
* Soluções baseadas em **Inteligência Artificial (IA)** podem oferecer um diagnóstico rápido, barato e acessível.

Os modelos de aprendizado profundo têm grande potencial na **detecção automatizada de doenças oculares**, porém, **muitos exigem grandes volumes de dados e poder computacional elevado**. Meu trabalho busca avaliar o desempenho de **redes neurais compactas**, que operam com **baixo custo computacional** em aplicações médicas em cenários de **dados limitados**.

Este trabalho tem como objetivo principal **avaliar a viabilidade de redes neurais compactas para a detecção de catarata em um cenário clínico realista, com disponibilidade restrita de dados**.

Os objetivos específicos incluem:

* Implementar e testar **diferentes arquiteturas de redes neurais**, incluindo:
  + **Redes Convolucionais (CNNs)**
  + **Vision Transformers (ViTs)**
  + **Modelos híbridos (CNN + Transformer)**
* Comparar o desempenho dos modelos considerando métricas como **acurácia, sensibilidade e eficiência computacional**.
* Validar a robustez dos modelos em um contexto de **dados limitados**, avaliando **técnicas de generalização**.

A **Inteligência Artificial (IA)** é um campo da computação que busca desenvolver sistemas capazes de **realizar tarefas que tradicionalmente exigem inteligência humana**.

Dentro da IA, temos o **Aprendizado de Máquina (Machine Learning, ML)**, que se baseia na ideia de que algoritmos podem aprender **padrões a partir de dados**, sem a necessidade de programação explícita.

O **Aprendizado Profundo (Deep Learning, DL)** é um subconjunto do ML que utiliza **redes neurais profundas** para processar informações de maneira hierárquica.

As **Redes Neurais Convolucionais (CNNs - Convolutional Neural Networks)** são arquiteturas projetadas para processar **dados bidimensionais**, preservando suas relações espaciais e sendo amplamente utilizadas em **análise de imagens médicas**.

O funcionamento de uma CNN baseia-se em três tipos principais de camadas: **convolução, subamostragem (pooling) e camadas totalmente conectadas**. A **camada de convolução** é responsável pela extração de características das imagens, utilizando filtros conhecidos como **kernels**. Cada kernel percorre a imagem aplicando uma **operação de convolução**, onde os valores do filtro são multiplicados pelos pixels correspondentes e somados, resultando em um mapa de características. Diferentes filtros podem aprender a detectar bordas, texturas e padrões específicos dentro da imagem.

Após a convolução, os mapas de características passam por uma **camada de ativação**, sendo a mais comum a **ReLU (Rectified Linear Unit)**, que introduce **não-linearidade** no modelo ao substituir valores negativos por zero. Isso melhora a capacidade da rede de aprender padrões complexos.

A próxima etapa é a **subamostragem (pooling)**, que reduz a dimensão dos mapas de características e, consequentemente, o número de parâmetros da rede. O método mais utilizado é o **Max Pooling**, que seleciona o valor máximo dentro de uma determinada região da imagem. Essa técnica torna o modelo mais eficiente e ajuda a reduzir **overfitting**, melhorando a generalização para novos dados.

Por fim, os mapas de características são achatados e passados por uma ou mais **camadas totalmente conectadas (Fully Connected Layers - FC)**, onde cada neurônio recebe informações de todos os neurônios da camada anterior. A última camada possui um número de neurônios igual ao número de classes do problema, sendo responsável pela **classificação final**.

No contexto deste trabalho, foram utilizadas algumas arquiteturas de CNNs já bem estabelecidas na literatura. A **EfficientNet-B0**, por exemplo, é um modelo que otimiza a relação entre profundidade (número de camadas), largura (número de canais) e resolução da imagem, garantindo um **desempenho eficiente com menor custo computacional**.

Outra arquitetura avaliada foi a **ResNet (Residual Network)**, que introduz **conexões residuais** para evitar o problema do **desaparecimento do gradiente**. Isso permite a construção de redes mais profundas sem prejudicar o aprendizado. Modelos como **ResNet6, ResNet8, ResNet10 e ResNet18** foram testados neste estudo, variando o número de camadas convolucionais para analisar o impacto da profundidade no desempenho da rede.

Os **Transformers** surgiram como uma inovação no campo do **Processamento de Linguagem Natural (NLP)**, revolucionando a forma como modelos de aprendizado profundo processam informações sequenciais. Tradicionalmente, redes neurais recorrentes (RNNs) eram utilizadas para lidar com sequências, mas apresentavam limitações, como dificuldades na modelagem de dependências de longo prazo e alto custo computacional. Os Transformers resolveram essas limitações ao introduzir o **mecanismo de autoatenção (Self-Attention)**, permitindo que diferentes partes da entrada fossem processadas simultaneamente.

Esse mecanismo de autoatenção funciona atribuindo **pesos dinâmicos** a diferentes partes da entrada, permitindo que o modelo determine quais informações são mais relevantes em um determinado contexto. Para calcular essa relevância, o Transformer utiliza três componentes principais:

* **Query (Q)**: Representação da palavra (ou elemento) que precisa encontrar relações dentro da sequência.
* **Key (K)**: Representação das demais palavras (ou elementos) que podem ter relevância para a Query.
* **Value (V)**: Informação associada a cada palavra (ou elemento), que será ponderada com base na similaridade entre Query e Key.

A similaridade entre Query e Key é obtida por meio de um **produto escalar**, gerando uma matriz de pesos que define a importância relativa de cada elemento no contexto da entrada. Esse processo permite que o Transformer capture relações complexas entre os dados, sem a necessidade de processamento sequencial, tornando o modelo altamente eficiente para grandes volumes de dados.

Os transformers são estruturados com atenção multicabeça, que nada mais é o mecanismo de atenção em múltiplas cabeças em paralelo, da camada feedfoward de múltiplos perceptons, esta sendo a camada aonde é armazenado o conhecimento, uma camada de normalização e a camada de saída que porcessa os token para a tarefa final.

Apesar do grande sucesso dos Transformers no NLP, sua aplicação em **Visão Computacional** só ocorreu recentemente, com a introdução do **Vision Transformer (ViT)**. Diferente das redes convolucionais, que utilizam filtros para extrair características locais da imagem, o ViT trata imagens de maneira semelhante ao processamento de texto, dividindo-as em **pequenos patches** e processando cada um como um **token** dentro do modelo de atenção.

O primeiro passo do ViT é a **divisão da imagem em patches de tamanho fixo**. Cada patch é transformado em um vetor por meio de uma camada chamada **Patch Embedding**, que projeta a informação espacial para um espaço vetorial adequado para o Transformer. Para garantir que a ordem dos patches seja preservada, adiciona-se uma informação adicional chamada **Positional Encoding**, que fornece um indicativo da posição original de cada patch na imagem.

Depois dessa etapa, os vetores são passados para várias camadas do Transformer, onde ocorre o processamento de autoatenção. A diferença fundamental entre o ViT e uma CNN é que, enquanto as redes convolucionais extraem características locais da imagem por meio de filtros especializados, o ViT processa a imagem de forma **global**, permitindo que cada parte da imagem interaja diretamente com todas as outras partes. No final do modelo, um **Token de Classe** é adicionado e utilizado para representar a saída final da rede, equivalente à predição do modelo.

**Desafios do Vision Transformer**

Embora o ViT tenha demonstrado resultados impressionantes em grandes bases de dados, como **ImageNet-21k**, ele apresenta alguns desafios significativos. O principal deles é a necessidade de **grandes quantidades de dados para generalizar bem**, pois, ao contrário das CNNs, o ViT não possui filtros convolucionais para aprender representações visuais eficientes em pequenos conjuntos de dados. Isso faz com que o ViT tenha **desempenho inferior em bases pequenas**, tornando-se uma escolha menos viável para aplicações médicas onde a disponibilidade de imagens rotuladas pode ser limitada.

Para contornar essa limitação, foram desenvolvidas abordagens híbridas que combinam **CNNs e Transformers**. Um exemplo dessa estratégia é o **ResNet-ViT**, onde uma ResNet é utilizada para extrair características da imagem antes de alimentar o Transformer, substituindo o tradicional Patch Embedding. Outra alternativa é o **Compact Convolutional Transformer (CCT)**, que incorpora uma camada convolucional na entrada do ViT, garantindo que a rede aprenda características locais antes de aplicar a autoatenção global.

Para avaliar a viabilidade de redes neurais compactas na detecção de catarata, utilizamos um conjunto de dados especializado contendo imagens oftalmológicas classificadas em quatro categorias: **normal, catarata, glaucoma e doenças de retina**. No entanto, como o foco do estudo é a detecção de catarata, os experimentos foram conduzidos em duas abordagens distintas: uma de **classificação binária**, diferenciando apenas entre olhos normais e olhos com catarata, e outra de **classificação multiclasse**, incluindo todas as categorias presentes no dataset.

Antes do treinamento dos modelos, foi realizada uma **etapa de pré-processamento das imagens** para garantir a consistência dos dados. Todas as imagens foram **redimensionadas** para um tamanho padrão compatível com os modelos utilizados, normalizadas para valores entre 0 e 1 e submetidas a técnicas de **aumento de dados (Data Augmentation)**, como rotações aleatórias, mudanças de brilho e espelhamento horizontal. Esse procedimento visa aumentar a variabilidade do conjunto de treinamento e reduzir o risco de **overfitting**.

A base de dados foi então dividida em três partes:

* **80% para treinamento**: conjunto de imagens usado para ajuste dos pesos dos modelos.
* **12% para validação**: utilizado para monitoramento do desempenho durante o treinamento e ajuste de hiperparâmetros.
* **8% para teste**: conjunto reservado para a avaliação final dos modelos.

Dado que a disponibilidade de imagens médicas rotuladas é frequentemente limitada, este trabalho busca avaliar como redes neurais compactas se comportam nesse cenário, verificando se sua eficiência pode ser mantida mesmo com um volume reduzido de dados.

Foram implementadas e comparadas diversas arquiteturas de redes neurais, divididas em três grupos principais: **Redes Neurais Convolucionais (CNNs), Vision Transformers (ViTs) e modelos híbridos** que combinam as duas abordagens.

No grupo das CNNs, foi testado o modelo EfficientNet-B0, uma versão compacta da família EfficientNet e também modelos da **família ResNet**, como **ResNet6, ResNet8, ResNet10 e ResNet18** foram escolhidas para avaliar como diferentes profundidades influenciam o desempenho da detecção de catarata.

Na abordagem baseada em **Transformers**, utilizamos o **Vision Transformer Lite (ViT-Lite)**, uma versão compacta do ViT original, adaptada para cenários com menor disponibilidade de processamento. Ele reduz o número de **camadas no encoder (de 12 para 4)** e o número de **cabeças de atenção (de 12 para 2)**, buscando minimizar a complexidade computacional sem comprometer significativamente o desempenho.

Por fim, os modelos híbridos testados incluem o **Compact Convolutional Transformer (CCT)** e o **ResNet-ViT**, que combinam a eficiência das convoluções na extração de características locais com o poder dos Transformers na modelagem de relações globais na imagem. Enquanto o **CCT** incorpora convoluções na fase inicial do ViT para otimizar a extração de características, o **ResNet-ViT** substitui o tradicional Patch Embedding do ViT por uma **ResNet**, permitindo que a extração de características seja feita por uma CNN antes da aplicação do mecanismo de autoatenção.

O treinamento dos modelos foi realizado utilizando a **biblioteca PyTorch**, com uma abordagem otimizada para cenários de dados limitados. Para garantir um aprendizado eficiente, utilizamos o **otimizador AdamW**, que combina o método Adam com regularização baseada em decaimento de peso (Weight Decay), reduzindo o impacto de parâmetros menos relevantes. A taxa de aprendizado inicial foi definida como **4 × 10⁻⁶** e ajustada dinamicamente durante o treinamento por meio da técnica **ReduceLROnPlateau**, que reduz a taxa de aprendizado quando o modelo para de melhorar na validação.

Além disso, aplicamos **Early Stopping**, interrompendo o treinamento caso o desempenho na validação não melhorasse após um número determinado de épocas, evitando o sobreajuste.

Para avaliação do desempenho dos modelos, utilizamos as seguintes métricas:

* **Acurácia**: percentual de predições corretas.
* **Perda CrossEntropyLoss**: métrica de erro utilizada em classificações multiclasse.
* **Sensibilidade (Recall)**: métrica fundamental para garantir que o modelo seja eficaz na identificação de casos positivos de catarata.

Os resultados dos experimentos demonstraram que as **Redes Convolucionais (CNNs) foram mais eficazes do que os Vision Transformers na detecção de catarata em um cenário de dados limitados**.

Na **classificação binária (normal vs. catarata)**, os melhores resultados foram obtidos com **ResNet e ResNet-ViT**, que atingiram **100% de precisão no conjunto de testes** e mais de **95% na validação**. O **Compact Convolutional Transformer (CCT)** também teve um bom desempenho, alcançando **98,44% de precisão**. Já o **EfficientNet-B0**, apesar de obter uma boa acurácia de **95,31%**, exigiu um tempo de treinamento significativamente maior (100 épocas).

Na **classificação multiclasse**, o modelo **ResNet-ViT** novamente obteve o melhor resultado, com **95,31% de precisão no teste**, seguido pela **ResNet (89,84%)**. No entanto, os modelos puramente baseados em Transformers, como **ViT-Lite**, apresentaram dificuldades para generalizar, com um desempenho abaixo de **70% na validação**. Isso confirma a **sensibilidade dos Transformers ao tamanho da base de dados**, reforçando a importância de utilizar modelos híbridos ou técnicas adicionais para melhorar a generalização

Outro ponto relevante é o impacto da profundidade da rede no desempenho. **Redes muito profundas, como ResNet18, não apresentaram ganhos significativos em comparação com versões mais leves, como ResNet6 e ResNet8**, o que sugere que modelos mais compactos podem ser mais adequados para este tipo de aplicação.

Com base nesses resultados, podemos afirmar que **modelos baseados em CNNs são a melhor escolha para a detecção de catarata em um contexto de dados limitados, enquanto Transformers precisam de grandes volumes de dados para alcançar resultados competitivos**.

Este estudo demonstrou que **redes neurais compactas são viáveis e eficazes para a detecção de catarata**, sendo que as **CNNs superam os Transformers puros quando o volume de dados disponíveis é restrito**. Modelos híbridos como **ResNet-ViT** apresentaram desempenho competitivo, combinando a robustez das convoluções com a capacidade de autoatenção dos Transformers.

Os principais achados deste trabalho foram:

1. **Redes convolucionais compactas, como ResNet6 e ResNet8, são suficientes para detectar catarata com alta precisão e baixo custo computacional**.
2. **Modelos puramente baseados em Transformers não são ideais para pequenas bases de dados**, devido à sua dependência de grandes volumes de dados para aprendizado adequado.
3. **Modelos híbridos (ResNet-ViT e CCT) podem ser uma alternativa promissora**, mas sua adoção deve ser analisada caso a caso.

Como trabalho futuro, sugerimos **testar técnicas de aprendizado por transferência (Transfer Learning)** e **expandir o dataset com novas imagens oftalmológicas**, permitindo uma melhor avaliação do impacto dos Transformers no diagnóstico de catarata.