UNIVERSIDADE FEDERAL DO MARANHÃO

Curso de Ciência da Computação

Thalisson Jon Cutrim Silva

Detecção e diagnóstico automático de patologias na retina utilizando arquitetura baseada em Transformers

São Luís - MA

2024

Thalisson Jon Cutrim Silva

Detecção e diagnóstico automático de patologias na retina utilizando arquitetura baseada em Transformers

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. João Dallyson Sousa de Almeida

São Luís - MA

2024

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Silva, Thalisson Jon Cutrim.
Detecção e diagnóstico automático de patologias na retina utilizando arquitetura baseada em Transformers / Thalisson Jon Cutrim Silva. - 2024.
45 f.

Orientador(a): João Dallyson Sousa de Almeida.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, Auditório do NCA, 2024.

1. Classificação Multirrótulo. 2. Doenças da Retina. 3. Query2Label. 4. Transformers. I. Almeida, João Dallyson Sousa de. II. Título.

Trabalho aprovado em São Luís - MA, 04 de Dezembro de 2023:

Prof. Dr. João Dallyson Sousa de Almeida
Orientador
Universidade Federal do Maranhão

Prof. Dr. Tiago Bonini Borchatt
Universidade Federal do Maranhão

Prof. Dr. Darlan Bruno Pontes Quintanilha
Universidade Federal do Maranhão

Agradecimentos

Em primeiro lugar a Deus, e aos meus pais, por terem me oferecido todo o suporte até o momento, repassando suas experiências de vida e me criando com valores desde jovem.

À minha vó, que graças a ela pude ter uma educação de qualidade. Mesmo que ela não tenha oportunidade em vida atualmente, ela terá mais um neto formado.

Aos meus amigos que fiz durante a minha caminhada na UFMA, que guardarei para o resto da minha vida.

Ao professor João Dallyson, por ter a paciência de me orientar e a confiança de manter dois projetos comigo, além de me iniciar na pesquisa.

Aos demais professores pelos ensinamentos que me ajudaram a chegar no momento que estou.

"O objetivo fundamental dos sonhos não é o sucesso, mas nos livrar do fantasma do conformismo."

Augusto Cury

Resumo

Globalmente, pelo menos 2,2 bilhões de pessoas têm deficiência visual. Em pelo menos um bilhão, ou quase metade desses casos, a deficiência visual poderia ter sido evitada. Dessa forma, é de suma importância a realização de exames preventivos de fundo de olho com o intuito de diagnosticar previamente doenças, para evitar que tais patologias evoluam para estados irreversíveis, como cegueira. Assim, este estudo apresenta um novo método para detectar múltiplas patologias oculares em imagens de fundo de olho, com a utilização de uma arquitetura de rede neural baseada em *transformers*, denominada Query2Label. Inicialmente, a entrada da rede passa por métodos de tratamento de dados, com o intuito de filtrar ao máximo a informação desejada da imagem e resolver problemas de balanceamento de classes, assim podendo servir de entrada para a rede, que aplica técnicas como *Cross-Atention*, *Adaptive Feature Pooling*, *Feature Extracting*, entre outras demais citadas ao longo do trabalho. Os experimentos foram realizados em um *dataset* público denominado RFMiD, que possui exames com uma ou múltiplas patologias na retina. O método empregado apresentou resultados promissores, especialmente na categoria "D. Risk", alcançando uma precisão média de 99,8%. Em comparação com o estado da arte, o método demonstrou excelente desempenho na detecção da classe "ODP", previamente não detectada, e superou a precisão em categorias específicas, como "CSR","LS", entre outras. Esses resultados ressaltam a viabilidade e eficácia do estudo proposto para a classificação de patologias oftalmológicas específicas.

Palavras-chave: Classificação Multirrótulo; Doenças da Retina; Transformers; Query2Label.

Abstract

Globally, at least 2.2 billion people have visual impairments. The visual impairment could have been prevented in at least one billion, or nearly half of these cases. Therefore, it is of utmost importance to conduct preventive eye fundus exams to diagnose diseases in advance to prevent such pathologies from progressing to irreversible states, such as blindness. Thus, this research introduces a new method for detecting multiple ocular pathologies in fundus eye images using a neural network architecture not previously used in this context. The proposed method employs a new deep neural network architecture based on transformers. Initially, the network's input undergoes data preprocessing methods to filter the desired information from the image and address class imbalance issues. This preprocessed data is then used as input for the network, which applies techniques such as Cross-Attention, Adaptive Feature Pooling, and Feature Extraction, among others mentioned throughout the study. The experiments were conducted on a public dataset called RFMiD, containing examinations with one or multiple pathologies in the retina. The employed method yielded promising results, especially in the"D. Risk" category, achieving an average accuracy of 99.8%. Compared to the state-of-the-art approaches, the method showed outstanding performance in detecting the "ODP" class, previously undetected, and surpassed accuracy in specific categories such as "CSR", "LS", among others. These results underscore the proposed study's feasibility and effectiveness for classifying specific ophthalmic pathologies.

Keywords: Multi-Label classification; Retinal Diseases; Transformers; Query2Label.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CvT	Convolutional vision Transformer
FN	Falso Negativo
FP	Falso Positivo
FPR	False Positive Rate
LMT	Label Mask Training
NCA	Núcleo de Computação Aplicada
Q2L	Query2Labels
RFMiD	Retinal Fundus Multi-disease Image Dataset
ROC	Receiver Operating Characteristic Curve
TPR	True Positive Rate
UFMA	Universidade Federal do Maranhão
ViT	Vision Transformer

Sumário

1 Introdução

Globalmente, pelo menos 2,2 bilhões de pessoas têm deficiência visual e, no Maranhão, esse número é de quase 140 mil no estado (IBGE, 2010). Em pelo menos um bilhão, ou quase metade desses casos, a deficiência visual poderia ter sido evitada (WHO, 2021). Assim, é de extrema importância que aconteça avanços tecnológicos relacionados a diagnósticos preventivos e/ou automáticos, principalmente em doenças que atingem todos os anos uma larga escala de pessoas, como Degeneração Macular relacionada à idade (DMRI), Retinopatia Diabética (RD) e Glaucoma, que causam cegueira em mais de 10 milhões de pessoas em todo o mundo (MITTAL; RAJAM, 2020). O exame de fundo de olho é realizado com a visualização da região da retina, por meio de fotos coloridas do fundo de olho, oferecendo um exame não invasivo da microcirculação sistêmica da retina (PACHADE et al., 2021).

Já que múltiplas doenças podem acometer a retina de um único paciente, o trabalho em questão foi tratado como um caso de classificação Multirrótulo. No Aprendizado Multirrótulo temos um conjunto de treinamento composto por instâncias, nas quais uma única instância está associada a vários rótulos de diferentes classes simultaneamente (ZHANG et al., 2018). Tal método é de suma importância para o diagnóstico de doenças oculares, já que pacientes que, por exemplo, sofrem de RD podem também sofrer de outras doenças, como Glaucoma e DMRI. Portanto, a detecção de múltiplas doenças é essencial caso aja o risco de presença de mais de uma patologia em um paciente (PACHADE et al., 2021).

O Aprendizado Profundo utilizado para análise de imagens médicas surgiu no campo de Aprendizado de Máquina (LECUN; BENGIO; HINTON, 2015). Com isso em mente, foram aplicados diversos métodos de Aprendizado Profundo para detecção de patologias utilizando imagens do fundo do olho, realizando testes com diferentes modelos e parâmetros, os quais serão apresentados ao decorrer do trabalho.

1.1 Justificativa

Como bem citado, tais doenças podem levar à cegueira quando não são diagnosticadas previamente, logo é de extrema importância a detecção da anomalia em seus estágios primários. O exame de fundo de olho automatizado apresenta grande importância nesse cenário, já que um sistema automático consegue oferecer rastreamento em larga escala de forma padronizada e com baixo custo, pode ajudar a reduzir possíveis erros humanos e ainda fornecer atendimento médico em áreas remotas (MITTAL; RAJAM, 2020).

1.2 Objetivos

O objetivo deste estudo é propor um modelo de rede neural baseado em aprendizado profundo para classificar automaticamente tanto doenças oculares frequentes como patologias raras na retina.

1.2.1 Objetivos Específicos

No sentido de alcançar o objetivo geral pretendido, busca-se atingir os seguintes objetivos específicos:

- Adquirir bases públicas de imagens de retina com doenças frequentes e raras da retina;
- Avaliar a utilização de redes neurais profundas atuais mais utilizadas na tarefa de classificação de imagens;
- Propor novo modelo baseado em rede neural profunda para diferenciar imagens de retinas saudáveis e doentes;
- Propor um novo modelo baseado em rede neural profunda para classificar o tipo de patologia da retina.
- Avaliar o modelo proposto por meio de experimentos em base de imagens pública, utilizando métricas da literatura.

1.3 Contribuição

Destaca-se como principal contribuição deste trabalho a utilização de uma nova arquitetura de rede neural profunda, descrita na Seção 4.3, chamada de Query2Label, aplicada no problema de classificação de retinografias com múltiplos rótulos. Para tanto, utilizou-se do conjunto de imagens RFMiD juntamente com técnicas de tratamento de dados.

1.4 Organização do Trabalho

O restante do trabalho está estruturado da seguinte forma:

- O Capítulo 1 introduz o trabalho realizado, citando a justificativa para realização do mesmo e seus objetivos.
- O Capítulo 2 cita os principais trabalhos relacionados que contribuíram para a realização desta pesquisa.

- O Capítulo 3 trata da fundamentação teórica das técnicas utilizadas. É apresentado o conceito de Redes Neurais e *transformers* sendo abordados conceitos acerca do planejamento dos passos da arquitetura proposta.
- O Capítulo 4 apresenta as etapas que compõem a metodologia proposta para este trabalho. Descrevendo características do conjunto de dados, no pré-processamento e no método proposto.
- O Capítulo 5 trata dos resultados obtidos e discussões em relação aos experimentos realizados com a rede explicada.
- O Capítulo 6 apresenta as considerações finais sobre os resultados, contribuições e melhorias para trabalhos futuros.

2 Trabalhos Relacionados

A literatura apresenta alguns trabalhos relacionados a esta pesquisa. Dentre os disponíveis, Müller e Kramer (2021), Liu et al. (2021), Rodriguez, AlMarzouqi e Liatsis (2022), propuseram métodos e técnicas computacionais para a detecção de patologias de fundo de olho que contribuíram para a realização deste trabalho.

No trabalho proposto por Müller e Kramer (2021), foi utilizado um método de Aprendizado em Conjunto após um *Up-Sampling* em todo o Dataset. O método consiste em utilizar dois modelos diferentes: um para a detecção da doença de fundo de olho e outro para a classificação em caso de patologia detectada. Para atingir um bom resultado, ambos os modelos foram pré-treinados no Dataset da ImageNet (DENG et al., 2009). Utilizando o *dataset* RFMiD no trabalho citado, foram utilizados quatro *backbones* diferentes juntamente do método proposto, sendo estes: DenseNet201, ResNet152, InceptionV3 e EfficientNetB4.

No trabalho realizado por Liu et al. (2021), foi desenvolvido um método de classificação utilizando *Transformers*. Motivados pelo sucesso do uso do *Transformers* em tarefas de visão computacional, apresentaram uma solução utilizando *Transformer Decoders* para questionar a existência de uma categoria. Utilizando um método de Atenção Cruzada para o *Adaptively Feature Pooling*, com o intuito de detectar diferentes partes importantes em uma imagem. Os *Transformer Decoders* são utilizados para extrair as características das imagens utilizando o método explicado previamente. Na arquitetura proposta utilizando o conjunto de dados RFMiD, antes de tudo a entrada passa pelo *backbone* que extrai a localidade das características encontradas, ou seja, as características espaciais. Uma grande vantagem do método proposto é que diferentes *backbones* podem ser utilizados.

No trabalho desenvolvido por Rodriguez, AlMarzouqi e Liatsis (2022), foi utilizado um método com base em *transformers*. Nesse estudo foi usada a arquitetura C-Tran proposta por Farnell et al. (2008) como modelo classificador. Esse modelo C-Tran consiste em um *transformer encoder* alimentado pelas características visuais extraídas de uma CNN e por um conjunto das máscaras das categorias. De uma imagem é extraída diversos atributos visuais com o uso de uma Rede Neural Convolucional *backbone*. De cada imagem, um conjunto de representações de rótulos (*label embeddings*) é gerada e, de cada uma, são geradas as predições utilizando redes independentes chamadas *redes Feed-forward*.

O trabalho apresentado por Müller e Kramer (2021), mostra um bom diferencial utilizando o método de *Ensemble Learning* para combinar as predições de vários modelos de redes convolucionais. Já os trabalho de Liu et al. (2021) e de Rodriguez, AlMarzouqi e Liatsis (2022) demonstram o diferencial do uso de *Transformers* para a tarefa de

classificação. O método de Rodriguez, AlMarzouqi e Liatsis (2022) em específico consegue obter um ótimo resultado mesmo sem conhecimento prévio, ou seja, pré-treinamento, utilizando o método LMT (label mask training), separando parte das categorias para aprender combinações entre as *labels*.

Porém, no trabalho aqui proposto foi baseado em sua maioria no método de Liu et al. (2021) (Query2Label), que apresenta uma grande vantagem: a grande flexibilidade no uso de diferentes *backbones* junto da arquitetura, utilizado para a extração de característica da imagem, podendo utilizar um modelo pré-treinado para atingir resultados ainda melhores. Dito isso, o motivo desse método ser utilizado como base neste método proposto apresentado é sua fácil adaptabilidade a diferentes *backbones* e o seu foco para classificação de múltiplos rótulos. Até o momento, o método de Liu et al. (2021) não foi previamente avaliado ou aplicado no contexto específico abordado por esta pesquisa, o que representa um fator relevante para a escolha desse trabalho como base.

3 Fundamentação Teórica

Na Seção 3.1 será explicado o conceito da classificação multirrótulo e a diferença dos diferentes tipos de classificação.

Na Seção 3.2 é fornecido uma breve visão sobre a evolução das redes neurais, destacando a influência no processamento de informações visuais. Além disso, também é abordado os principais conceitos sobre Redes Neurais Convolucionais.

A Seção 3.3 abordará a definição de *backbone* e a sua função na arquitetura utilizada neste trabalho, utilizando a ResNet para desempenhar tal cargo.

Na Seção 3.4 exploraremos os *Visions Transformers*, uma variação dos modelos baseados em *transformers* voltada para o processamento de imagens. Além disso, também será abordado etapas e conceitos importantes da arquitetura utilizada no trabalho.

Já na Seção 3.5 é mostrado as principais etapas e características do *Convolutional Vision Transformer*, destacando sua abordagem híbrida que integra elementos de Redes Neurais Convolucionais e modelos baseados em *transformers* para processamento eficiente de informações espaciais em imagens.

Finalizando o capítulo, na Seção 3.6 é abordada uma estratégia para otimização de hiperpârametros chamada de *sec:gridsearch*.

3.1 Classificação Multirrótulo

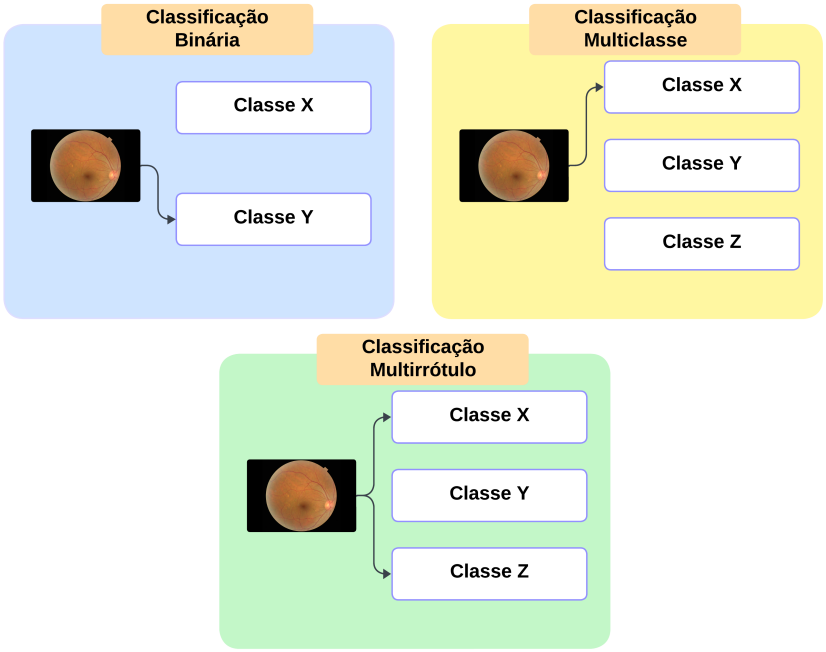
Um dos objetivos principais do trabalho apresentado é avaliar o desempenho do método apresentado na tarefa de classificação de imagens.

O objetivo da classificação é atribuir um rótulo a um conjunto de dados com base em suas características observadas. Esse processo visa identificar padrões e classificar posteriormente conjuntos de dados não rotulados.

Na classificação binária, há apenas duas classes possíveis para cada instância, e o objetivo é prever a qual dessas duas classes a instância pertence. Por outro lado, na classificação multiclasse, cada instância de dados é associada a exatamente uma categoria dentre várias classes presentes. No trabalho apresentado, o conjunto de dados utilizado é destinado à classificação multirrótulo, a qual é uma extensão da classificação padrão de um único rótulo, onde cada instância de dados pode estar associada simultaneamente a várias categorias (TAREKEGN; GIACOBINI; MICHALAK, 2021).

Na Figura 1 é mostrado a diferença entre os tipos de classificação previamente explicados:

Figura 1 – Tipos de problemas de classificação.



3.2 Redes Neurais

O trabalho em redes neurais tem sido motivado a muito tempo por conta do reconhecimento de que o cérebro humano processa informações de uma forma inteiramente diferente do computador digital convencional (HAYKIN, 2001).

Em sua forma mais geral, uma rede neural é um sistema projetado para modelar a maneira como o cérebro humano realiza uma tarefa particular, sendo normalmente implementada utilizando-se componentes eletrônicos ou é simulada por propagação em um computador digital (FLECK et al., 2016).

Entretanto, avanços significativos foram alcançados com o advento das Redes Neurais Convolucionais. As Redes Neurais Convolucionais revolucionaram a maneira como as informações visuais são interpretadas e processadas. Ao invés de depender exclusivamente de métodos tradicionais de processamento de imagem, as Redes Neurais Convolucionais utilizam uma abordagem de aprendizado de máquina para entender e extrair características complexas de imagens (VENKATESAN; LI, 2017).

3.2.1 Redes Neurais Convolucionais

Sendo uma das Redes Neurais Convolucionais (CNNs) pioneiras, LeCun et al. (1989) conseguiu atingir um grande sucesso com sua pesquisa e despertar interesse dos estudiosos da área de Redes Neurais. As Redes Neurais Convolucionais são uma classe especializada de redes neurais profundas que se destacaram significativamente no processamento de

dados visuais, como imagens e vídeos. Esses tipos de redes revolucionaram a capacidade de reconhecimento de padrões em imagens, tendo impacto em diversas áreas, desde visão computacional até processamento de linguagem natural.

Há também outros importantíssimos trabalhos que fizeram uso de Redes Neurais Convolucionais, como a arquitetura proposta por Krizhevsky, Sutskever e Hinton (2012). Esse trabalho introduziu a arquitetura conhecida como "AlexNet". Esta rede neural foi uma das primeiras a demonstrar o poder das Redes Neurais Convolucionais em tarefas de classificação de imagens em grandes conjuntos de dados.

Ela recebe esse nome de uma operação matemática linear entre matrizes chamada convolução. As Redes Neurais Convolucionais possuem várias camadas, incluindo camada convolucional, camada de não-linearidade, camada de *pooling* e camada totalmente conectada (ALBAWI; MOHAMMED; AL-ZAWI, 2017).

O'Shea e Nash (2015) também afirmam que as funcionalidades básicas de uma Rede Neural Convolucional podem ser descritas em 4 etapas:

- Camada de Entrada: Como em outras Redes Neurais Artificiais, a camada de entrada em uma *CNN* é responsável por armazenar os valores dos píxeis da imagem. Cada neurônio nesta camada representa um píxel da imagem de entrada.
- Camada Convolucional e Função de Ativação ReLu: A camada convolucional é responsável por realizar operações de convolução em toda a rede neural convolucional. Ela aplica um conjunto de filtros aos dados de entrada para extrair características, multiplicando os pesos desses filtros pelas regiões da imagem. A ReLu é uma função de ativação frequentemente usada para introduzir não linearidade, transformando os valores negativos em zero e mantendo os valores positivos.
- Camada de *Pooling*: Esta camada executa a operação de *pooling*, realizando uma redução da dimensionalidade espacial da entrada. Isso ajuda a diminuir a complexidade computacional e a reduzir o número de parâmetros, tornando o processo de aprendizado mais eficiente.
- Camada totalmente conectada *(fully-connected)*: Na arquitetura de uma rede neural, as camadas totalmente conectadas recebem a saída da camada anterior como entrada de dados. Essas camadas têm como objetivo final produzir saídas correspondentes às classes do modelo para a tarefa de classificação.

Resumidamente, as Redes Neurais Convolucionais são estruturadas de maneira a atender às demandas específicas de problemas relacionados ao processamento de imagem. Elas se destacam por serem altamente adaptáveis às diferenças das imagens, conseguindo aprender automaticamente padrões complexos e realizar tarefas como reconhecimento de

objetos, segmentação e classificação de imagens. Neste trabalho, a ênfase está na utilização da *CNN* para a classificação de imagens.

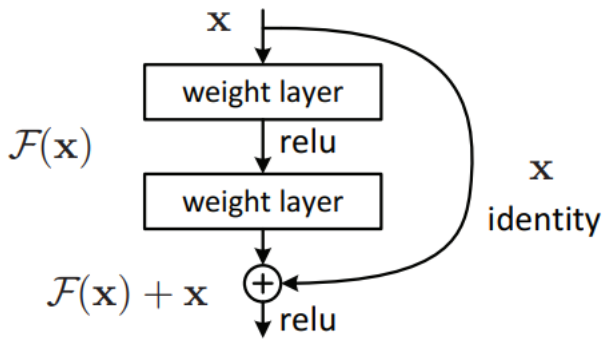
3.3 Backbone

O termo *backbone* refere-se à arquitetura de rede neural convolucional (CNN) ou estrutura semelhante que serve como a base principal para a extração de características dos dados de entrada. O *backbone* da rede desempenha um papel vital na extração e aprendizado das características importantes do conjunto de dados, sendo comumente utilizado como componente inicial em modelos que utilizam *transformers* para tarefas de classificação.

Neste trabalho, utilizou-se a rede ResNet como backbone. Esta rede foi inicialmente proposta por He et al. (2016), seguindo uma abordagem em adicionar conexões de atalho a cada duas camadas para uma rede em estilo VGG, estilo este apresentado inicialmente por Simonyan e Zisserman (2014).

O atalho feito pela ResNet conecta as ativações de uma camada a outras camadas, o que acaba pulando algumas camadas entre elas, resultando na formação de um bloco residual. A rede ResNet é construída através do empilhamento desses blocos em conjunto. Na figura 2 é demonstrado a conexão de atalho presente nesta rede.

Figura 2 – Demonstração da conexão de atalho presente na rede ResNet.



Um dos problemas mais famosos resolvidos pela ResNet é o *vanishing gradient*, que ocorre quando a rede é muito profunda, na qual os gradientes de onde a função de perda é calculada encolhem para zero após várias execuções da regra da cadeia, resultando na não atualização do valor dos pesos e, portanto, nenhum aprendizado está sendo feito.

A versão da rede utilizada em um dos testes que serão demonstrados é a ResNet101 (HE et al., 2016), apresentando 101 camadas de profundidade.

3.4 Vision Transformers

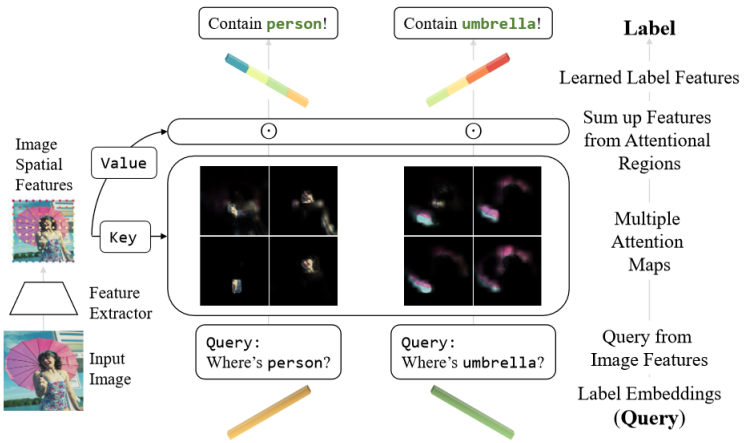
Descritos pela primeira vez por Vaswani et al. (2017), o uso de modelos baseado em *transformers* se mostram extremamente eficientes quando utilizado para reconhecimento de imagens (DOSOVITSKIY et al., 2020) e já demonstraram um grande potencial nessa área (YUAN et al., 2021). A utilização de *transformers* tem suas raízes na necessidade de extrair características discriminativas locais adaptativamente para diferentes rótulos, o que é uma propriedade fortemente desejada devido à existência de múltiplas características em uma única imagem (LIU et al., 2021).

Os *Visions Transformers* são uma adaptação dos modelos baseados *transformers* para a tarefa de processamento de imagens. O uso de *Vision Transformers* (ViT) no método proposto acontece da seguinte forma: a imagem é dividida em múltiplos *patches* (uma subseção da imagem) sendo alimentada em uma arquitetura de múltiplos blocos de *transformers* para classificação (DOSOVITSKIY et al., 2020).

Na arquitetura Query2Label apresentada por Liu et al. (2021) é feito o uso de *transformers*, com o uso juntamente de um método de *Cross Attention*, uma abordagem que permite ao modelo focar em partes específicas da entrada durante o processo de aprendizado, junto dos *transformers decoders*, explicado na subseção 3.4.1, com o intuito de extrair características de cada categoria separadamente.

A Figura 3 ilustra a arquitetura Query2Label juntamente do método de *Cross-Attention*, utilizado para realizar o *Adaptive Feature Pooling*, focando em diferentes partes da imagem.

Figura 3 – Ilustração da Query2Label (LIU et al., 2021).



Nas subseções adiante serão explicadas etapas e conceitos importantes da arquitetura Q2L.

3.4.1 Transformers encoders e transformers decoders

Para melhor entendimento de como funciona a arquitetura utilizada no trabalho, é preciso apresentar os *transformers encoders* e *transformers decoders*:

- *Transformers encoders* (codificador): É uma parte de um modelo *transformer* que processa e codifica a entrada, sendo a imagem de fundo de olho no trabalho em questão. Ele aprende a representar e entender os padrões nos dados de entrada.
- *Transformers decoders* (decodificador): É outra parte do modelo de Transformer que usa a representação aprendida pelo codificador para gerar uma saída, sendo as predições para cada categoria de patologia ocular associada.

3.4.2 Feature Extracting

Dada uma imagem como entrada, são extraídas as *spatial features* (características localizadas em espaços da imagem) por meio do *backbone*. Após isso, é adicionada uma camada linear de projeção para projetar as *features* e realizar a comparação com a dimensão da característica procurada.

3.4.3 Query Updating

Após obter as características espaciais das imagens utilizadas como entrada, é utilizado *label embeddings* (representação da classe como vetor numérico) como *queries* (aquilo que quer ser buscado), sendo executado o método de *cross-atention* para buscar características relacionadas das categorias das características espaciais utilizando Transformers decoders de múltiplas camadas. É utilizado uma arquitetura de *Transformers* padrão, que apresenta um módulo de *self-attention*, de *cross-atention* e a FFN (*position-wise feed-forward network*).

Ambos os módulos de *self-attention* e *cross-atention* são implementados utilizando a mesma função *MultiHead*. Cada *label embeddings* é tratado como um parâmetro aprendível, dessa forma as *embeddings* podem ser aprendidas de ponta a ponta a partir dos dados e das categorias relacionadas no modelo de forma implícita.

3.4.4 Feature Projection

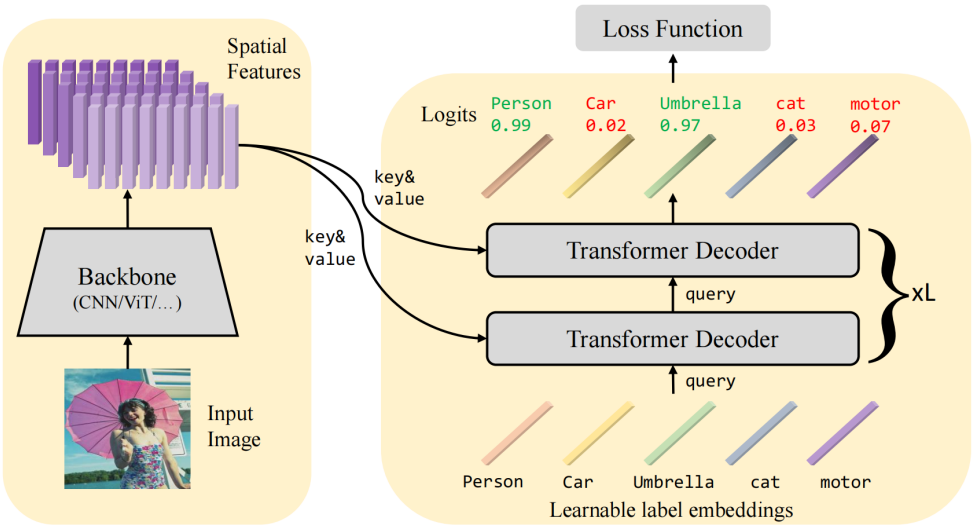
Para realizar uma classificação multirrótulo cada predição das categorias é tratada como uma classificação binária, sendo projetada a característica de cada classe para um *logit value* (representa valores de probabilidade de 0 a 1) utilizando uma camada de projeção linear seguida da seguinte função sigmóide:

$$pk = Sigmoid(W_k^T Q_{L,k} + b_k) \tag{3.1}$$

onde $W_k \in \Re, W = [W_1, ..., W_K]^T \in \Re^{Kxd}$ e $b_k \in \Re, b = [b_1, ..., b_k]^T \in \Re^K$ são parâmetros na camada linear, e $p = [p_1, ..., p_k]^T \in \Re^K$ são as probabilidades previstas de cada classe.

A Figura 4 mostra o esquema do modelo Query2Label, utilizando todos os conceitos previamente explicados nas subseções 3.4.2, 3.4.3 e 3.4.4.

Figura 4 – Esquema do modelo Query2Label (LIU et al., 2021).



Após extrair as características espaciais de uma imagem, cada *label embedding* (representação da classe como vetor numérico) é encaminhada para os decodificadores do *Transformer*. Isso é feito para realizar uma busca, comparando a *label embedding* com as características em cada localização espacial para gerar mapas de atenção, e reunir de maneira adaptativa a característica desejada, combinando linearmente as características espaciais com base nos mapas de atenção. A característica reunida é então usada para prever a existência da categoria "questionada"(*queried label*) (LIU et al., 2021).

3.5 Modelo CvT

A proposta do Convolutional Vision Transformer (CvT) é a integração estratégica de vantagens de redes convolucionais e *transformers*, mantendo as características positivas de ambos os paradigmas (AMJOUD; AMROUCH, 2020).

Os resultados obtidos com o CvT demonstram um ótimo desempenho de quando pré-treinado com o conjunto de dados ImageNet. Além disso, o modelo se destaca por ser leve e eficiente, superando o desempenho de modelos baseados em Redes Neurais Convolucionais.

O CvT divide os Transformers em estágios, formando uma estrutura hierárquica. No início de cada estágio, há um passo chamado *convolutional token embedding*. Essa etapa executa uma operação de convolução sobre um mapa de tokens 2D, a qual é basicamente uma representação espacial da informação da imagem, e em seguida aplica uma normalização. Essa abordagem resulta em um *downsampling* espacial, um processo semelhante ao realizado em Redes Neurais Convolucionais, que permite que o modelo capture informações locais da imagem.

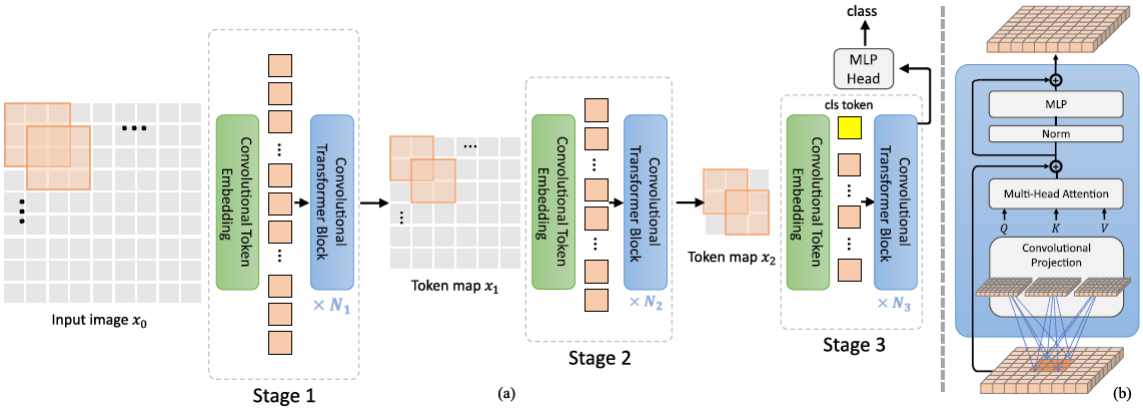
O CvT utiliza duas operações baseadas em convolução na arquitetura do *Vision Transformer*. Há três estágios do CvT:

- 1 - *Convolutional Token Embedding*: No início de cada estágio, a imagem de entrada (ou mapas de tokens 2D) passa por uma camada chamada "Convolutional Token Embedding". Aqui, é feita uma operação de convolução com sobreposição de partes da imagem, e essas partes são transformadas em tokens organizados em uma grade espacial 2D. Isso ajuda a reduzir o número de tokens (ou resolução das características) enquanto aumenta a largura dos tokens (ou dimensão das características), proporcionando um downsampling espacial, semelhante ao processo que ocorre em Redes Neurais Convolucionais.

- 2 - *Convolutional Transformer Blocks*: Depois do processo inicial, cada estágio é composto por vários *Convolutional Transformer Blocks*. Esse bloco usa uma técnica chamada *Convolutional Projection* para as etapas de consulta, chave e valor. Essa técnica permite capturar melhor informações locais, auxiliando o modelo a entender detalhes mais específicos e reduzir ambiguidades na atenção (dando importância corretamente a determinadas partes dos dados de entrada), tornando o processamento mais eficiente para certos tipos de informações espaciais em imagens, que necessitam a atenção detalhes muito específicos, como estruturas finas, texturas ou pequenas características localizadas em regiões específicas da imagem.

- 3 - Token de Classificação e *MLP Head*: O token de classificação, que contém informações resumidas e contextualizadas sobre a imagem processada ao longo das camadas anteriores, é adicionado apenas no último estágio. Por fim, um MLP (*Multi-Layer Perceptron*) é usado para prever a classe com base no token de classificação produzido na saída do último estágio.

Figura 5 – Pipeline da arquitetura CvT proposta por Amjoud e Amrouch (2020).

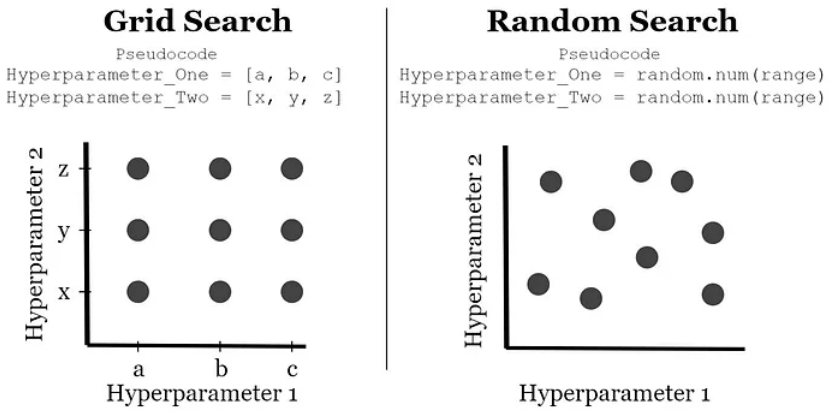


3.6 Grid Search

O *Grid Search*, juntamente da busca manual, é a estratégia mais amplamente utilizada para otimização de hiperparâmetros (BERGSTRA; BENGIO, 2012). A mudança de hiperparâmetros com o *Grid Search* refere-se a uma técnica de ajuste de hiperparâmetros em que é definido uma grade de valores para os hiperparâmetros no modelo e é testado cada combinação possível desses valores. Também existem outros mecanismos de buscas de hiperparâmetros, como o *Random Search*, que explora aleatoriamente valores de hiperparâmetros, enquanto o *Grid Search* explora exaustivamente todas as combinações em uma grade predefinida. Neste trabalho foi utilizado o *Grid Search* devido a já existir recomendações predefinidas de hiperparâmetros na arquitetura e modelos utilizados, sendo possível filtrar mais facilmente os valores em uma grade para realizar a busca.

A figura 6 demonstra o comparativo entre o método de *Grid Search* e o método de *Random Search*:

Figura 6 – Comparativo entre o *Grid Search* e o *Random Search*. Fonte: (STALFORT, 2019).



Os parâmetros inseridos nessa grade foram o *learning rate* (taxa de aprendizado) e o *weight decay* (decaimento de peso).

4 Materiais e Método

Na Seção 4.1 o conjunto de dados utilizado será descrito, apresentando estrutura desse conjunto de imagens relacionadas a diversas patologias oftalmológicas.

Na Seção 4.2, descrevem-se todas as etapas executadas durante o processamento das imagens e a implementação da rede neural. Serão detalhados desde o corte das imagens e equilíbrio das classes através do Up-Sampling até a aplicação de técnicas de aumento de dados durante o treinamento.

A Seção 4.3 apresenta detalhes sobre a Query2Label, uma arquitetura de rede neural adotada para o processamento de imagens multirrótulo com o uso de transformers.

Por fim, a Seção 4.4 descreve as métricas utilizadas para avaliar o desempenho da rede neural, com foco na métrica de precisão e AUC-ROC Curve, destacando suas definições e relevância na análise dos resultados obtidos durante os experimentos.

4.1 Dataset RFMiD

O conjunto de dados RFMiD é uma base de imagens de retina de acesso público, publicada no início de 2021. A versão do Dataset utilizada neste trabalho apresenta 1920 imagens no subconjunto de treinamento e 640 imagens no subconjunto de validação, com 29 patologias presentes.

A Tabela 1 mostra a distribuição de amostras para cada classe presente no subconjunto de treinamento.

Tabela 1 – Frequência das patologias presentes no subconjunto de dados de treinamento do Dataset RFMiD.

Patologia	**Amostra**	**Patologia**	**Amostra**	**Patologia**	**Amostra**
D.Risk	1519	MS	15	PT	11
DR	376	CSR	37	RT	14
ARMD	100	ODC	282	RS	43
MH	317	CRVO	28	CRS	32
DN	138	TV	6	EDN	15
MYA	101	AH	16	RPEC	22
BRVO	73	ODP	65	MHL	11
TSLN	186	ODE	58	RP	6
ERM	14	ST	5	OTHER	34
LS	47	AION	17		

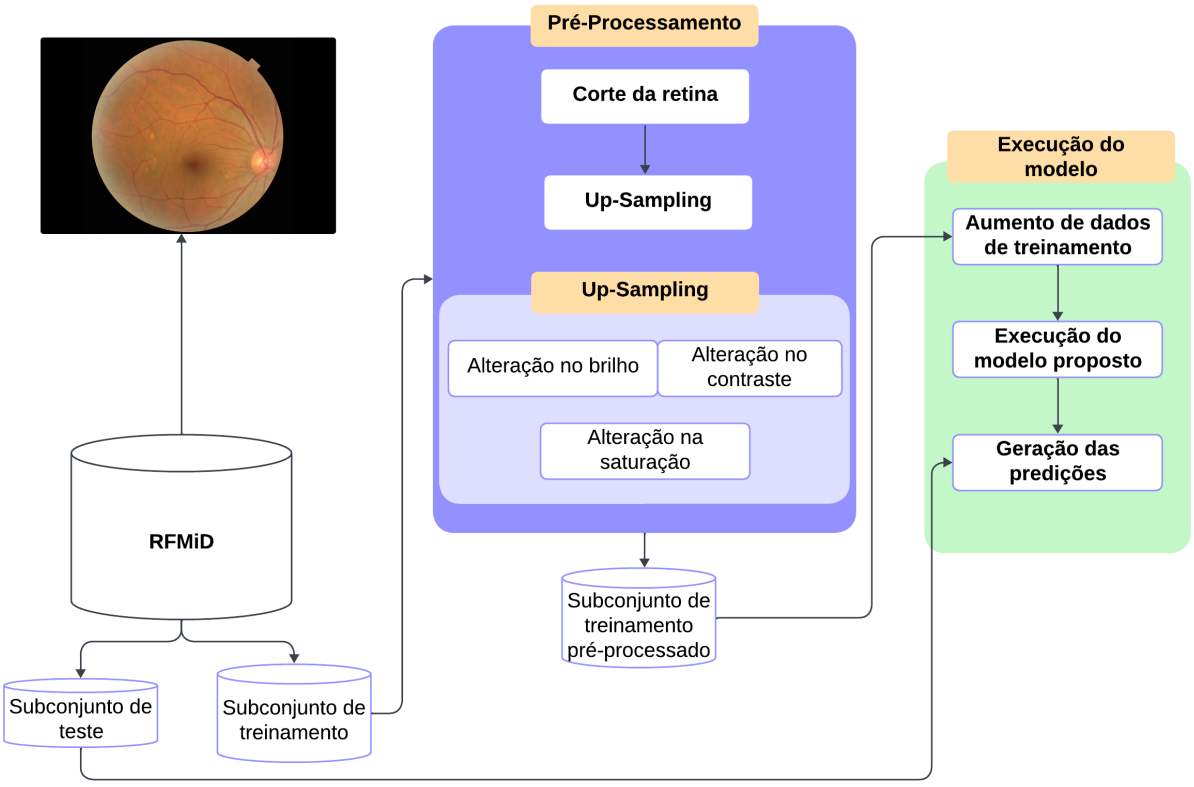
Para melhor entendimento das siglas correspondentes às patologias oculares presentes no conjunto de dados, as mesmas serão listadas a seguir:

1. D.Risk (Disease Risk): Presença de doença/anormalidade
2. DR: Retinopatia Diabética
3. ARMD: Degeneração Macular Relacionada à Idade
4. MH: Neblina na mídia
5. DN: Drusas
6. MYA: Miopia
7. BRVO: Oclusão de Veia Retiniana de Ramo
8. TSLN: Tesselação
9. ERM: Membrana Epirretiniana
10. LS: Cicatrizes de Laser
11. MS: Cicatriz Macular
12. CSR: Retinopatia Serosa Central
13. ODC: Escavação do Disco Óptico
14. CRVO: Oclusão de Veia Retiniana Central
15. TV: Vasos Tortuosos
16. AH: Hialose Asteróide
17. ODP: Palidez do Disco Óptico
18. ODE: Edema do Disco Óptico
19. ST: Derivação Optociliar
20. AION: Neuropatia Óptica Isquêmica Anterior
21. PT: Telangiectasia Parafoveal
22. RT: Tração Retiniana
23. RS: Retinite
24. CRS: Coriorretinite
25. EDN: Exsudação
26. RPEC: Alterações no Epitélio Pigmentado da Retina
27. MHL: Buraco Macular
28. RP: Retinite Pigmentosa
29. Other: Outras patologias oculares

Este conjunto de dados possibilitará o desenvolvimento de técnicas e redes para detecção e classificação de doenças na presença de um número reduzido de casos de amostra (PACHADE et al., 2021).

4.2 Método proposto

Figura 7 – Etapas do método proposto.



Após a coleta de dados, ocorre a etapa de pré-processamento do modelo. O subconjunto passa por uma divisão holdout de 85:15, com 85% das imagens para treinamento e 15% para validação. Primeiramente, apenas as imagens do conjunto de dados de treinamento passam pelo processo pela função de corte da retina e de *Up-Sampling*, sendo criado outro conjunto de dados de treinamento melhorado com a utilização dos métodos propostos. Dessa forma, o subconjunto de treinamento contém 3.354 imagens que passaram pelas etapas de pré-processamento, e o subconjunto de validação apresenta 640 imagens. Após essas etapas, acontece processo de execução do modelo, passando por uma etapa de aplicação de augmentations que não foram aplicadas durante o processo de *Up-Sampling*, podendo ser vistas na subseção 4.2.3. O pré-processamento de dados é muitas vezes de grande importância para se obter resultados razoáveis caso se esteja preocupado em construir um modelo de previsão bom e robusto (RINNAN et al., 2009).

A seguir serão mostrados os passos realizados na fase de pré-processamento do conjunto de dados, sendo realizada uma técnica de corte da imagem, *up-sampling* em todo o *dataset* e aplicações de *augmentation* em todo esses dados alterados durante o treinamento.

4.2.1 Função de corte da retina

A ideia de aplicação dessa função no conjunto de dados RFMiD foi primeiramente citada por Müller e Kramer (2021). As imagens foram cortadas com o intuito de manter a imagem do fundo do olho centralizada na imagem e preservar a proporção após o redimensionamento das imagens.

O corte foi feito individualmente nas imagens de acordo com cada resolução do equipamento utilizado na aquisição da imagem, que correspondem as seguintes dimensões: 1424x1424, 1536x1536 e 3464x3464 pixels. As imagens posteriormente foram redimensionadas para treinamento da rede, passando para 384x384 pixels, tamanho esse esperado pelo modelo para as imagens. Na figura 8 é possível ver a mudança da imagem após o Up-Sampling e a função de corte.

Figura 8 – Imagem após passar pelo Up-Sampling e função de corte.

4.2.2 Up-Sampling

O redimensionamento ou *resampling* de imagens é uma das operações mais essenciais, suportada por praticamente todos os softwares de edição de imagem, sendo usado para muitos fins (FATTAL, 2007). No trabalho em questão, foi utilizado um método de *Upsampling* para diminuir o extremo desbalanceamento presente no *dataset*.

Treinar um bom classificador multirrótulo é uma tarefa complexa, no entanto, o grande desequilíbrio entre classes se torna um grande desafio para a construção de um bom modelo (KAUR; GOSAIN, 2018). Com o intuito de resolver tal desbalanceamento, foi realizado um método de *upsampling*.

O método utilizado e a técnica de balanceamento de pesos das classes ajudou a resolver o desbalanceamento do conjunto de dados, aumentando o número de amostras em patologias mais raras, apesar de algumas classes em específico continuarem a apresentar fortes dificuldades por conta do número muito pequeno de amostras.

Por meio do *up-sampling*, foi garantido que cada categoria apresentasse pelo menos 100 amostras no conjunto de dados. A técnica foi realizada diretamente no *dataset* RFMiD antes do treinamento para mitigar parcialmente o problema de desbalanceamento de classes.

Vale ressaltar que essa técnica foi aplicada apenas ao subconjunto de treinamento do conjunto de dados RFMiD. O método de aumento de dados utilizado no *up-sampling* foram: alteração no brilho, tonalidade, contraste e saturação. Essas transformações foram aplicadas, pois as imagens podem apresentar diferenças naturais em suas propriedades visuais devido às condições de captura ou às características individuais das amostras, como luminosidade e outros aspectos visuais comuns em conjuntos de dados reais. A Tabela 2 mostra a frequência das categorias após o *Up-Sampling*:

Tabela 2 – Frequência das patologias presentes no subconjunto de dados de treinamento do Dataset RFMiD após o up-sampling.

Patologia	**Amostra**	**Patologia**	**Amostra**	**Patologia**	**Amostra**
D.Risk	2953	MS	105	PT	104
DR	566	CSR	110	RT	103
ARMD	168	ODC	571	RS	103
MH	455	CRVO	100	CRS	104
DN	300	TV	102	EDN	105
MYA	124	AH	101	RPEC	102
BRVO	124	ODP	272	MHL	101
TSLN	345	ODE	109	RP	100
ERM	108	ST	101	OTHER	188
LS	139	AION	101		

4.2.3 Augmentations usadas durante o treinamento.

Durante o treinamento também foram aplicadas algumas técnicas de aumento de dados, sendo essas citadas a seguir na tabela 3:

Tabela 3 – Lista de operações de aumento de dados aplicadas.

Augmentation	**Descrição**	**Parâmetro**
RandomRotation	Rotaciona aleatoriamente a imagem de acordo com o ângulo selecionado dentro do limite máximo	degrees = 90
RandomHorizontalFlip	Vira a imagem horizontalmente	-
RandomVerticalFlip	Vira a imagem verticalmente	-
ColorJitter	Altera randomicamente a saturação, brilho, contraste e a tonalidade da imagem de acordo com o limite máximo	brightness = 0.2, constrast = 0.2, saturation = 0.2, hue = 0,2

4.3 Arquitetura utilizada

A arquitetura utilizada foi a Query2Label (LIU et al., 2021), uma arquitetura designada para classificação multirrótulo com o uso de *transformers*. Na etapa de treinamento são executados os passos também já descritos anteriormente na seção 3.4: na primeira etapa o *backbone* extrai as características das imagens com o método de *Multi-Head Attention*, focando em diferentes partes das imagens. Na segunda etapa terá um bloco de múltiplas camadas de *Transformers Decoder* para realizar o *Query Updating* (3.4.3) e *Adaptive*

Feature Pooling, onde são comparadas as *label embeddings* com as *features* coletadas em cada localização para gerar mapas de atenção (3.4.2), além de uma camada linear de projeção para computar as predições dadas.

O *dataset* padrão realizado na arquitetura "Q2L" foi o "MS-COCO". Para carregar o conjunto de dados RFMiD, criou-se uma classe customizada para carregar o *dataset* desejado, utilizando "pytorch", um *framework* baseado na biblioteca "Torch". Após isso, foram feitas customizações diretamente no código disponibilizado pela arquitetura para adaptação da rede com o intuito de receber o *dataset* RFMiD, como a mudança no carregamento da quantidade de classes.

4.4 Métricas de Avaliação

Para avaliação dos resultados, foram utilizadas as métricas de *Precision* e a *AUC-ROC Curve*. A métrica de precisão mede a porcentagem de exemplos classificados corretamente em uma categoria específica dentro do conjunto de dados. A métrica de precisão é descrita pela Equação 4.1.

$$Precisão = \frac{TP}{TP + FP} \tag{4.1}$$

onde TP é a quantidade de verdadeiros positivos e FP é a quantidade de falsos positivos. O verdadeiro positivo seria quando a rede prevê corretamente a presença de uma patologia ocular na imagem. Já o falso positivo acontece quando o modelo prevê a presença de uma patologia em uma imagem, mas essa categoria não está realmente presente nessa imagem.

A *AUC-ROC Curve* é importante para entender como o modelo está performando principalmente na questão de distinção de diferentes classes. No gráfico em questão, é mostrado o desempenho do modelo de classificação nos diferentes limiares.

A curva demonstra dois parâmetros, o *True Positive Rate (TPR)* e o *False Positive Rate (FPR)*, descritos nas Equações 4.2 e 4.3.

$$TPR = \frac{TP}{TP + FN} \tag{4.2}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.3}$$

onde TP é a quantidade de verdadeiros positivos, FN é a quantidade de falsos negativos e TN é a quantidade de verdadeiros negativos.

5 Resultados e Discussão

Esta seção apresenta e discute os resultados que foram alcançados realizando os experimentos no *dataset* RFMiD 4.1.

A versão utilizada da ResNet como backbone, sendo estas: ResNet18, ResNet34, ResNet50 e ResNet101. As técnicas utilizadas em cada backbone da ResNet foram as mesmas utilizadas na versão da ResNet101, assim como os mesmos parâmetros já apresentados. Entretanto, a ResNet101 foi o Backbone que apresentou melhores resultados comparado as outras arquiteturas.

Os dois primeiros experimentos foram realizados com um treinamento de 250 épocas, apresentando 64 de *batch size*, o *learning rate* inicial de $1x10^{-}4$ e utilizando o "AdamW" como otimizador. No último teste apresentado, é aplicado uma taxa de *weight decay* e inserido a técnica de *Grid Search* para otimização de hiperparâmetros, onde a taxa de aprendizado e o *weight decay* são alterados até encontrar os melhores valores para cada parâmetro. Foi seguido uma estratégia de divisão holdout na proporção de 85:15, onde 85% dos dados foram designados para treinamento e os restantes 15% para validação.

Os resultados mostrados na Tabela 4 foram alcançados utilizando a arquitetura base apresentada juntamente com o backbone da ResNet101.

Tabela 4 – Precisão de cada categoria (Q2L-ResNet101).

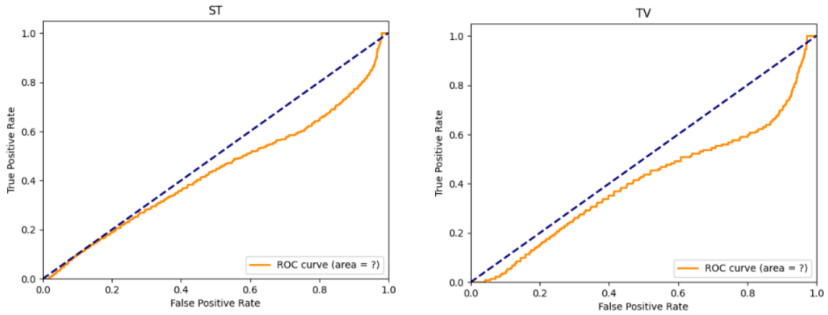
Patologia	**Precisão**	**Patologia**	**Precisão**	**Patologia**	**Precisão**
D.Risk	98,7	MS	8,5	PT	2,6
DR	88,7	CSR	28,1	RT	73,8
ARMD	60,8	ODC	55,2	RS	82,5
MH	86	CRVO	73,3	CRS	27,1
DN	46	TV	0,6	EDN	16,6
MYA	87,3	AH	47,5	RPEC	7,1
BRVO	57,7	ODP	37	MHL	43,1
TSLN	64,9	ODE	72,9	RP	8
ERM	6	ST	1	OTHER	17,5
LS	53,9	AION	20,6		

De início, a classe de D.Risk já obteve uma taxa de precisão muito alta, atingindo 98,7%. Apesar de algumas classes apresentarem valores satisfatórios, outras como "ST", "RP", "ERM", "TV"e "MS"apresentaram valores de precisão muito baixos. O motivo principal do baixo desempenho dessas categorias é por conta do pequeno número de amostras presentes no conjunto de treinamento dessas classes, somando com a dificuldade de detectar as difíceis características presentes nessas patologias.

Apesar do *Up-Sampling* feito para evitar a baixa precisão em categorias com pouca quantidade de amostra, continuou se apresentando dificuldades para aumentar a precisão

das mesmas.

Figura 9 – Curva ROC das patologias ST e TV.



Para dados mais específicos, a classe "TV" apresenta apenas 6 amostras iniciais e a classe "ST" menos ainda, havendo apenas 5 amostras. Tais classes com números extremamente pequenos de imagens se mostraram um grande desafio para se conseguir uma boa precisão.

Como visto na Figura 9, a *ROC Curve* se manteve abaixo e próximo de 0.5 de AUC, apresentando grande dificuldade no modelo de distinguir se a imagem é uma classe negativa ou positiva se tratando da categoria em questão, classificando erroneamente grande parte dos testes.

Além de testes realizados com a arquitetura base, foi utilizado um modelo da mesma arquitetura pré-treinada utilizando *Vision Transformers*, os quais são implementados justamente para tarefas de processamento de visão (como reconhecimento de imagens), sendo feito um redimensionamento para 384x384 após toda a etapa de pré-processamento para ser utilizado como entrada da arquitetura modificada.

Com a arquitetura modificada, foram feitas diversas alterações na fase de pré-processamento para obter um melhor uso da memória disponível, abrindo possibilidade para aumento do tamanho do *batch* na execução da rede nos testes posteriores. Juntamente com a modificação da arquitetura, foi utilizado um modelo disponível chamado "CVT_W24"(WU et al., 2021) como *feature extractor*, sendo pré-treinado na Imagenet. Por esse motivo, a arquitetura nesse teste será chamada de "Q2L-CVT_W24".

A Tabela 5 que será apresentada irá mostrar a precisão de cada categoria do conjunto de dados RFMiD após o treinamento do método em questão.

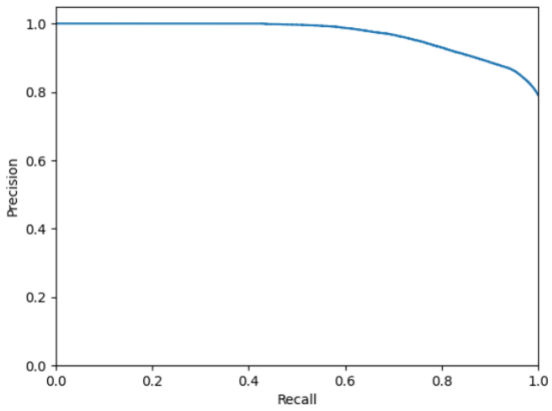
Tabela 5 – Precisão de cada categoria (Q2L-CVT_W24).

Patologia	Precisão	Patologia	Precisão	Patologia	Precisão
D.Risk	99,8	MS	14,4	PT	2,1
DR	87,5	CSR	69,1	RT	77,3
ARMD	55,3	ODC	41,8	RS	73
MH	86,3	CRVO	80,9	CRS	22,6
DN	29,4	TV	0,3	EDN	21,3
MYA	83,2	AH	24,1	RPEC	2,3
BRVO	57,7	ODP	22,1	MHL	7,4
TSLN	65	ODE	72,5	RP	45,3
ERM	11,5	ST	0,8	OTHER	19,2
LS	30	AION	32,6		

Apesar das mudanças realizadas, não foram obtidos grandes avanços se comparado aos testes realizados com o método anterior, se mostrando superior apenas em casos específicos, como na patologia "RP", onde obteve melhor salto de precisão. Visto isso, uma possível alternativa seria utilizar os dois métodos em paralelos com separação dos melhores resultados de cada um.

Como estado da arte, foi utilizado o trabalho de Rodriguez, AlMarzouqi e Liatsis (2022), que também utilizou *transformers* no *dataset* RFMiD. Apesar de usar o mesmo conjunto de dados, o trabalho em questão não utilizou todas as classes do *dataset*, já que ele usou apenas classes que também apresentavam em outro conjunto de dados estudado no trabalho em questão. Comparando apenas a detecção da patologia, foi obtido um resultado superior: no trabalho em questão, foi utilizado a categoria dos normais, que basicamente seria a mesma categoria do "D.Risk", porém havendo falsos positivos ao invés de verdadeiros positivos, ou seja: a mesma categoria apenas utilizando os casos normais como métrica ao invés de casos com doença. Na classe dos normais foi obtido uma precisão de 85,9% no trabalho comparado, enquanto no método proposto se obteve 99,8%. A Figura 10 mostra o gráfico de relação entre a precisão e o *recall* da categoria de "D.Risk".

Figura 10 – Gráfico Precision/Recall da categoria "D.risk".



Ainda comparando com o método proposto por Rodriguez, AlMarzouqi e Liatsis

(2022), houve algumas categorias que o estado da arte não adquiriu nenhum acerto, obtendo 0% de precisão, como o caso da classe "ODP". Com o método apresentado neste trabalho, tal categoria obteve 37% de precisão no primeiro teste apresentado e 22,1% no segundo teste. Outras patologias também apresentaram resultados semelhantes e outros com pequenas melhoras, como "MYA", com uma pequena superioridade de 6% utilizando o primeiro teste e com 2,2% no segundo teste e "CRVO", com uma melhora de 20,9% utilizando o segundo teste.

Apesar disso, houve também classes com resultados inferiores em comparação com a arquitetura comparada, como a classe "DN"com uma piora de 27% se comparada com o primeiro método; "RS", o qual foi obtido 100% e "BRVO"com uma piora de 35,2% comparado com quaisquer dos métodos. Na Tabela 6 será mostrado o último teste apresentado comparado com o método proposto:

Tabela 6 – Precisão de cada categoria (Q2L-CVT_W24) comparado com o estado da arte (método de Rodriguez, AlMarzouqi e Liatsis (2022)).

Método proposto		**Estado da arte**	
Patologia	**Precisão**	**Patologia**	**Precisão**
D.Risk	99,8	D.Risk	85,9
ARMD	55,3	ARMD	80,0
MH	86,3	MH	87,5
DN	29,4	DN	70,8
MYA	83,2	MYA	81,0
BRVO	57,7	BRVO	92,9
TSLN	65,0	TSLN	80,0
LS	30,0	LS	50,0
CSR	69,1	CSR	44,4
ODC	41,8	ODC	66,1
CRVO	80,9	CRVO	60,0
ODP	22,1	ODP	0,00
ODE	72,5	ODE	83,3
RS	73,0	RS	100,0
CRS	22,6	CRS	40,0
OTHER	19,2	OTHER	58,7

Em geral, houve classes que apresentaram resultados superiores no método apresentado, enquanto outras apresentaram resultados inferiores em comparação com o estado da arte no teste em questão. O método proposto superou o estado da arte no "D.Risk", "MYA", "CSR", "CRVO"e "ODP";

No próximo teste, foi utilizado a mesma arquitetura e o mesmo *backbone*. Porém, agora foi adicionado uma função de *weight decay* para as camadas do modelo, que é uma técnica de regularização que afeta os parâmetros do modelo durante o treinamento, penalizando os pesos da rede durante o processo de otimização. Além disso, foi aplicado a otimização de hiperparâmetros com *Grid Search*, explicado na seção 3.6.

A tabela que será apresentada irá mostrar a precisão de cada categoria do conjunto de dados RFMiD após as mudanças apresentadas:

Tabela 7 – Precisão de cada categoria (Q2L-CVT_W24) com *weight decay* e *Grid Search*.

Patologia	Precisão	Patologia	Precisão	Patologia	Precisão
D.Risk	99,8	MS	17	PT	2,4
DR	89,5	CSR	65,4	RT	78,7
ARMD	67,1	ODC	35,8	RS	85,9
MH	86,2	CRVO	82,4	CRS	29,7
DN	29,2	TV	0,6	EDN	21,9
MYA	83,2	AH	35,6	RPEC	4,5
BRVO	65,3	ODP	20,7	MHL	7,4
TSLN	75,2	ODE	72,8	RP	51,5
ERM	10,3	ST	0,6	OTHER	20,7
LS	65,9	AION	35,6		

Comparando com o segundo teste, algumas classes permaneceram com a precisão muito semelhante, onde a maioria apresentou pequenos aumentos. No entanto, classes como "AH", "TSLN"e "LS"apresentaram um salto maior. Na Tabela 8 será comparado o último teste apresentado com o estado da arte.

Tabela 8 – Precisão de cada categoria (Q2L-CVT_W24) com *weight decay* e *Grid Search* comparado com o estado da arte (método de Rodriguez, AlMarzouqi e Liatsis (2022)).

Método proposto		Estado da arte	
Patologia	**Precisão**	**Patologia**	**Precisão**
D.Risk	99,8	D.Risk	85,9
ARMD	67,1	ARMD	80,0
MH	86,2	MH	87,5
DN	29,2	DN	70,8
MYA	83,2	MYA	81,0
BRVO	65,3	BRVO	92,9
TSLN	75,2	TSLN	80,0
LS	65,9	LS	50,0
CSR	65,4	CSR	44,4
ODC	35,8	ODC	66,1
CRVO	82,4	CRVO	60,0
ODP	20,7	ODP	0,00
ODE	72,8	ODE	83,3
RS	85,9	RS	100,0
CRS	29,7	CRS	40,0
OTHER	20,7	OTHER	58,7

Apesar do pequeno aumento de precisão na maioria das classes, o último teste proposto pelo método apresentado conseguiu superar apenas mais uma categoria, sendo essa a "LS", além dos outros rótulos já superados nos testes anteriores.

Na tabela 9 é mostrado o comparativo das precisões das categorias dos experimentos realizados. Nessa tabela é possível observar que houveram algumas patologias que apresentaram saltos de precisão expressivos, como: "AION", "LS"e "RP". Apesar disso, a maioria das doenças oculares mantiveram ou obtiveram pequenos saltos de precisão.

Tabela 9 – Comparativo dos resultados dos três experimentos apresentados.

Experimento 1		**Experimento 2**		**Experimento 3**	
Patologia	**Precisão**	**Patologia**	**Precisão**	**Patologia**	**Precisão**
D.Risk	98,7	D.Risk	99,8	D.Risk	99,8
DR	88,7	DR	87,5	DR	89,5
ARMD	60,8	ARMD	55,3	ARMD	67,1
MH	86	MH	86,3	MH	86,2
DN	46	DN	29,4	DN	29,2
MYA	87,3	MYA	83,2	MYA	83,2
BRVO	57,7	BRVO	57,7	BRVO	65,3
TSLN	64,9	TSLN	65	TSLN	75,2
ERM	6	ERM	11,5	ERM	10,3
LS	53,9	LS	30	LS	65,9
MS	8,5	MS	14,4	MS	17
CSR	28,1	CSR	69,1	CSR	65,4
ODC	55,2	ODC	41,8	ODC	35,8
CRVO	73,3	CRVO	80,9	CRVO	82,4
TV	0,6	TV	0,3	TV	0,6
AH	47,5	AH	24,1	AH	35,6
ODP	37	ODP	22,1	ODP	20,7
ODE	72,9	ODE	72,5	ODE	72,8
ST	1	ST	0,8	ST	0,6
AION	20,6	AION	32,6	AION	35,6
PT	2,6	PT	2,1	PT	2,4
RT	73,8	RT	77,3	RT	78,7
RS	82,5	RS	73	RS	85,9
CRS	27,1	CRS	22,6	CRS	29,7
EDN	16,6	EDN	21,3	EDN	21,9
RPEC	7,1	RPEC	2,3	RPEC	4,5
MHL	43,1	MHL	7,4	MHL	7,5
RP	8	RP	45,3	RP	51,5
OTHER	17,5	OTHER	19,2	OTHER	20,7

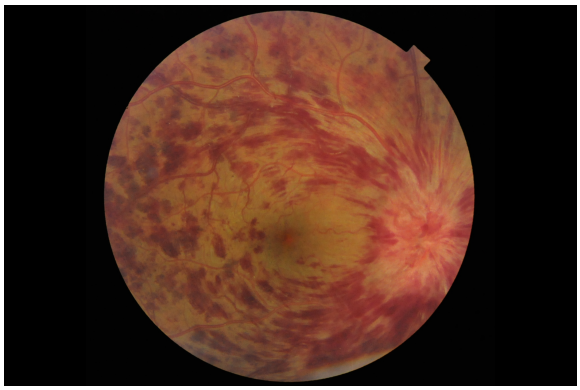
5.0.1 Estudos de Caso

Houve categorias que atingiram valores de precisão muito abaixo do esperado, como o caso da categoria "TV". Essa patologia é caracterizada por um padrão de grande tortuosidade nos vasos sanguíneos, que parece dilatado e segue um caminho sinuoso, como visto na Figura 11. Além do padrão ser de difícil detecção, o conjunto de dados apresentou apenas seis amostras dessa doença ocular, sendo os principais motivos para a baixa precisão nesta patologia.

Figura 11 – Vasos tortuosos presentes na categoria "TV".

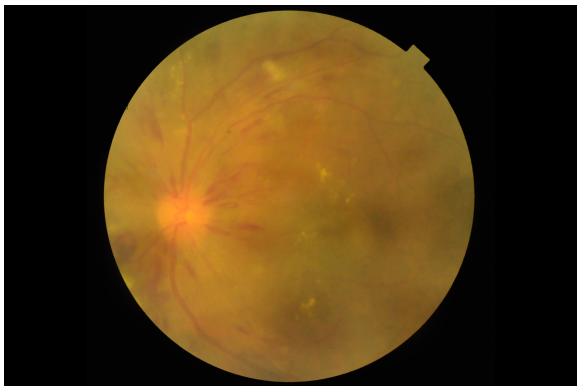
Apresentando bons resultados, a categoria que mais apresentou distância de valor de precisão de forma positiva do método de Rodriguez, AlMarzouqi e Liatsis (2022) foi a categoria "CRVO". Essa patologia se apresenta quando a veia principal da retina fica bloqueada perto do começo ou no nervo óptico. Os sinais clínicos incluem hemorragias em formato de chama, observado na Figura 12. Embora tenha sido treinado com um conjunto pequeno de amostras, apenas 28 no total, o método demonstrou habilidade para aprender os padrões das figuras. Isso se deve à sua capacidade de identificar facilmente a característica principal, que se destaca consideravelmente em relação às demais amostras.

Figura 12 – Hemorragias em formato de chama da categoria "CRVO".



Essa amostra apresentada na Figura 13 contém a patologia "CRVO", já apresentada anteriormente, junto de outra, a "MH". A opacidade visual presente na imagem com "MH"pode ser um indicativo da presença de catarata, opacidades no vítreo, edema de córnea ou pupilas pequenas. A patologia "MH"também atingiu uma boa precisão pela rede, atingindo 86,2% de precisão, chegando muito próximo do método proposto por Rodriguez, AlMarzouqi e Liatsis (2022).

Figura 13 – Amostra apresentando a patologia "MH"e "CRVO".



6 Conclusão

Este estudo apresentou um modelo alternativo para classificação multirrótulo de patologias na retina, onde tais técnicas podem ser utilizadas em trabalhos futuros e a arquitetura apresentada sendo também útil em casos específicos. O conjunto de dados utilizado no trabalho foi o RFMiD, onde se destacou a detecção de patologias (Disease Risk) e na classificação de alguns dos rótulos do *dataset*, como "ODP","CRVO"e outras categorias presentes nesse conjunto de dados.

Nos testes realizados, o método proposto conseguiu atingir 99,8% na categoria de D.Risk, ou seja, uma detecção quase perfeita da patologia ocular. Outras categorias também merecem destaque, como a "CRVO", atingindo 82,4% de precisão com apenas 28 imagens no subconjunto de treinamento; "LS", alcançando 65,9% de precisão com apenas 47 imagens no subconjunto de treinamento e "CSR", atingindo 65,4% de precisão com 37 imagens no mesmo subconjunto. Nos comparativos realizados, o método apresentado conseguiu superar algumas classes do método de Rodriguez, AlMarzouqi e Liatsis (2022), utilizado como estado da arte, e se equiparar em diversas outras. Nas categorias em que não superou, nenhum salto muito expressivo foi apresentado, como na categoria "ODP", no qual nenhum dado conseguiu ser classificado pelo estado da arte, diferentemente do método proposto.

As contribuições deste trabalho se apresentam no teste de uma nova arquitetura de rede baseada em *Transformers*, aplicada para classificar doenças em retinografias da base RFMiD e nas técnicas de pré-processamento utilizadas. Acrescenta-se que o trabalho realizado conseguiu atingir um salto expressivo em se tratando da detecção de patologia no paciente, mostrando uma precisão quase que perfeita, e em outras categorias.

Por fim, propõe-se para trabalhos futuros a aplicação de outras arquiteturas como *backbone*, como a DenseNet (HUANG et al., 2017) ou a TResNet (RIDNIK et al., 2021). Combinando o uso de outras redes (caso seja necessário) com o método apresentado em (OH; PARK, 2022), pode-se utilizar um extrator de características com dois classificadores diferentes, dependendo do desempenho de cada um em classes específicas, separando um subconjunto do conjunto de dados inteiro para cada classificador, assim como foi utilizado no trabalho referenciado. Por exemplo: utilizar o método proposto para a detecção de patologia (D.Risk) e "CRVO", que supera a precisão do estado da arte, e utilizar o método do estado da arte para classes como "BRVO"e "RS", que apresenta melhores resultados se comparado com o método proposto, cada arquitetura sendo utilizada em um classificador, sendo treinados separadamente. Dessa forma, utilizando o melhor de dois classificadores diferentes, pode-se alcançar um melhor resultado. Espera-se que a junção do método

proposto e o trabalho citado possa atingir um nível de desempenho ligeiramente superior no conjunto de dados utilizado para o estudo.

Referências

ALBAWI, S.; MOHAMMED, T. A.; AL-ZAWI, S. Understanding of a convolutional neural network. In: IEEE. *2017 international conference on engineering and technology (ICET)*. [S.l.], 2017. p. 1–6. Citado na página 20.

AMJOUD, A. B.; AMROUCH, M. Convolutional neural networks backbones for object detection. In: SPRINGER. *Image and Signal Processing: 9th International Conference, ICISP 2020, Marrakesh, Morocco, June 4–6, 2020, Proceedings 9*. [S.l.], 2020. p. 282–289. Citado 3 vezes nas páginas 8, 24 e 26.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *Journal of machine learning research*, v. 13, n. 2, 2012. Citado na página 26.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In: IEEE. *2009 IEEE conference on computer vision and pattern recognition*. [S.l.], 2009. p. 248–255. Citado na página 16.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. Citado na página 22.

FARNELL, D. J.; HATFIELD, F. N.; KNOX, P.; REAKES, M.; SPENCER, S.; PARRY, D.; HARDING, S. P. Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. *Journal of the Franklin institute*, Elsevier, v. 345, n. 7, p. 748–765, 2008. Citado na página 16.

FATTAL, R. Image upsampling via imposed edge statistics. In: *ACM SIGGRAPH 2007 papers*. [S.l.: s.n.], 2007. p. 95–es. Citado na página 31.

FLECK, L.; TAVARES, M. H. F.; EYNG, E.; HELMANN, A. C.; ANDRADE, M. A. d. M. Redes neurais artificiais: Princípios básicos. *Revista Eletrônica Científica Inovação e Tecnologia*, v. 1, n. 13, p. 47–57, 2016. Citado na página 19.

HAYKIN, S. *Redes neurais: princípios e prática*. [S.l.]: Bookman Editora, 2001. Citado na página 19.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778. Citado na página 21.

HUANG, G.; LIU, Z.; MAATEN, L. V. D.; WEINBERGER, K. Q. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 4700–4708. Citado na página 42.

IBGE. *Censo Demográfico*. 2010. Disponível em: <https://cidades.ibge.gov.br/brasil/ma/pesquisa/11/0>. Citado na página 13.

KAUR, P.; GOSAIN, A. Issues and challenges of class imbalance problem in classification. *International Journal of Information Technology*, Springer, p. 1–7, 2018. Citado na página 31.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, v. 25, 2012. Citado na página 20.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015. Citado na página 13.

LECUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W.; JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation*, MIT Press, v. 1, n. 4, p. 541–551, 1989. Citado na página 19.

LIU, S.; ZHANG, L.; YANG, X.; SU, H.; ZHU, J. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021. Citado 6 vezes nas páginas 8, 16, 17, 22, 24 e 32.

MITTAL, K.; RAJAM, V. Computerized retinal image analysis-a survey. *Multimedia Tools and Applications*, Springer, v. 79, n. 31, p. 22389–22421, 2020. Citado na página 13.

MüLLER, I. S.-R. D.; KRAMER, F. Multi-disease detection in retinal imaging based on ensembling heterogeneous deep learning models. 2021. Citado 2 vezes nas páginas 16 e 31.

OH, Y.-t.; PARK, H. End-to-end two-branch classifier for retinal imaging analysis. In: IEEE. *2022 International Conference on Electronics, Information, and Communication (ICEIC)*. [S.l.], 2022. p. 1–3. Citado na página 42.

O'SHEA, K.; NASH, R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015. Citado na página 20.

PACHADE, S.; PORWAL, P.; THULKAR, D.; KOKARE, M.; DESHMUKH, G.; SAHASRABUDDHE, V.; GIANCARDO, L.; QUELLEC, G.; MÉRIAUDEAU, F. Retinal fundus multi-disease image dataset (rfmid): a dataset for multi-disease detection research. *Data*, MDPI, v. 6, n. 2, p. 14, 2021. Citado 2 vezes nas páginas 13 e 29.

RIDNIK, T.; LAWEN, H.; NOY, A.; BARUCH, E. B.; SHARIR, G.; FRIEDMAN, I. Tresnet: High performance gpu-dedicated architecture. In: *proceedings of the IEEE/CVF winter conference on applications of computer vision*. [S.l.: s.n.], 2021. p. 1400–1409. Citado na página 42.

RINNAN, Å.; NØRGAARD, L.; BERG, F. van den; THYGESEN, J.; BRO, R.; ENGELSEN, S. B. Data pre-processing. *Infrared spectroscopy for food quality analysis and control*, Academic Press San Diego, California, p. 29–50, 2009. Citado na página 30.

RODRIGUEZ, M.; ALMARZOUQI, H.; LIATSIS, P. Multi-label retinal disease classification using transformers. *arXiv preprint arXiv:2207.02335*, 2022. Citado 8 vezes nas páginas 9, 16, 17, 36, 37, 38, 40 e 42.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. Citado na página 21.

STALFORT, J. *Hyperparameter tuning using grid search and random search*. 2019. Disponível em: <https://medium.com/@jackstalfort/hyperparameter-tuning-using-grid-search-and-random-search-f8750a464b35>. Citado 2 vezes nas páginas 8 e 26.

TAREKEGN, A. N.; GIACOBINI, M.; MICHALAK, K. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, Elsevier, v. 118, p. 107965, 2021. Citado na página 18.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 22.

VENKATESAN, R.; LI, B. *Convolutional neural networks in visual computing: a concise guide*. [S.l.]: CRC Press, 2017. Citado na página 19.

WHO. *Blindness and vision impairment*. 2021. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>. Citado na página 13.

WU, H.; XIAO, B.; CODELLA, N.; LIU, M.; DAI, X.; YUAN, L.; ZHANG, L. Cvt: Introducing convolutions to vision transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2021. p. 22–31. Citado na página 35.

YUAN, L.; CHEN, Y.; WANG, T.; YU, W.; SHI, Y.; JIANG, Z.-H.; TAY, F. E.; FENG, J.; YAN, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. [S.l.: s.n.], 2021. p. 558–567. Citado na página 22.

ZHANG, M.-L.; LI, Y.-K.; LIU, X.-Y.; GENG, X. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, Springer, v. 12, n. 2, p. 191–202, 2018. Citado na página 13.