Investigação de redes convolucionais para classificação de imagens de radiografia do tórax

Danilo Henrique da Silva Santana

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2024

Danilo Henrique da Silva Santana

Investigação de redes convolucionais para classificação de imagens de radiografia do tórax

Monografia apresentada ao curso Ciência de dados e inteligência artificial
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito para a obtenção do grau de Bacharel em titulo

Orientador: Leonardo Batista Vidal

Maio de 2024

Catalogação na publicação
Seção de Catalogação e Classificação

S232i Santana, Danilo Henrique da Silva.
Investigação de redes convolucionais para classificação de imagens de radiografia do tórax / Danilo Henrique da Silva Santana. - João Pessoa, 2024.
43 f. : il.

Orientação: Leonardo Batista.
TCC (Graduação) - UFPB/CI.

1. CNNs. 2. ViT. 3. Swin transformer. 4. Classificação de imagem. I. Batista, Leonardo. II. Título.

UFPB/CI CDU 004.932.2

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Ciência de dados e inteligência artificial intitulado ***Investigação de redes convolucionais para classificação de imagens de radiografia do tórax*** de autoria de Danilo Henrique da Silva Santana, aprovada pela banca examinadora constituída pelos seguintes professores:

Prof. Dr. Nome do Professor A
Instituicao do Professor A

Prof. Dr. Nome do Professor B
Instituicao do Professor B

Prof. Dr. Nome do Professor C
Instituicao do Professor C

Coordenador(a) do Departamento Nome do Departamento
Nome do Coordenador
CI/UFPB

João Pessoa, 16 de maio de 2024

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

Se as pessoas não acreditam que a matemática é simples, é apenas porque não percebem o quão complicada é a vida.

AGRADECIMENTOS

Primeiramente agradeço a Deus por ter me dado paciência nessa jornada tão difícil, agradeço a minha família pelo amor, apoio e ensinamentos. Agradeço a meus amigos que participaram dessa etapa comigo, em momentos de alegria, tristeza e sucesso, sem eles não seria possível enfrentar tantos obstáculos, todas as horas gastas em tutoria, discord e league of legends, com certeza nos aproximaram bastante e nos ajudou a criar um elo bem forte. Agradeço a minha esposa por estar comigo nessa etapa final me apoiando e ajudando nos meus momentos mais desesperados mas também por ter me dado tantos momentos felizes.

RESUMO

A identificação de doenças no corpo humano muitas vezes é desafiadora, uma vez que muitas delas permanecem invisíveis aos olhos. Em tais situações, a doença pode permanecer latente no organismo, dificultando a identificação do problema. Nesse contexto, os exames de raio-X desempenham um papel fundamental na identificação e diagnóstico desses problemas, permitindo a visualização interna do organismo e a avaliação da integridade dos órgãos. A radiografia do tórax, em particular, é um dos exames mais comuns e eficazes nesse sentido. Por meio desse exame, é possível identificar anomalias no tamanho ou na forma do coração, distúrbios nas veias e artérias, além de problemas nos pulmões, como pneumonia e outras doenças respiratórias. O trabalho em questão investiga a relevância da classificação de imagens de radiografias torácicas, em meio ao crescente uso de redes neurais convolucionais (CNNs), além de explicar as mudanças que foram feitas durante os anos que permitiram as CNNs ficarem tão famosas não apenas para a classificação de imagens, mas também para a detecção de objetos e segmentação. Nesse contexto, as arquiteturas de *transformers*, particularmente o *vision transformer* (ViT), emergem como modelo dominante. O objetivo principal é compreender e avaliar as inovações no desempenho de um *visual transformer* hierarquico chamado *swin transformer*. Para isso, introduzimos algumas dessas modificações em CNNs clássicas, como VGG16 e ResNet-50, e treinamos esses modelos utilizando a base de dados CheXpert, avaliando sua acurácia. A partir de uma série de ajustes sequenciais, mantendo apenas as modificações que demonstraram impacto positivo na acurácia, alcançamos uma precisão de 86,15% para VGG16 e 80,05% para ResNet-50, enquanto o *swin transformer* obteve 78,75%. Esses resultados sugerem que, apesar da robustez do *swin transformer*, certas adaptações nas CNNs podem resultar em desempenho superior, evidenciando a importância de explorar diferentes abordagens arquiteturais no contexto da classificação de imagens médicas.

Palavras-chave: CNNs, ViT, swin transformers, Classificação de imagem, CheXpert

ABSTRACT

The identification of diseases in the human body is often challenging, as many of them remain invisible to the naked eye. In such situations, the disease may remain latent in the organism, making it difficult to identify the problem. In this context, X-ray exams play a fundamental role in identifying and diagnosing these issues, allowing for the internal visualization of the body and the assessment of organ integrity. Chest X-rays, in particular, are one of the most common and effective exams in this regard. Through this exam, it is possible to identify anomalies in the size or shape of the heart, disorders in the veins and arteries, as well as problems in the lungs, such as pneumonia and other respiratory diseases. The work in question investigates the relevance of classifying images of chest X-rays amidst the growing use of convolutional neural networks (CNNs), while also explaining the changes that have been made over the years that have allowed CNNs to become famous not only for image classification but also for object detection and segmentation. In this context, transformer architectures, particularly vision transformers (ViTs), emerge as dominant models. The main goal is to understand and evaluate the performance innovations of a hierarchical visual transformer called swin transformer. To do this, we introduce some of these modifications to classic CNNs, such as VGG16 and ResNet-50, and train these models using the CheXpert database, evaluating their accuracy. Through a series of sequential adjustments, keeping only the modifications that demonstrated a positive impact on accuracy, we achieved a precision of 86.15% for VGG16 and 80.05% for ResNet-50, while the swin transformer obtained 78.75%. These results suggest that, despite the robustness of the swin transformer, certain adaptations in CNNs can result in superior performance, highlighting the importance of exploring different architectural approaches in the context of medical image classification.

Key-words: CNNs, ViT, swin transformers, Image classification, CheXpert

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

VIT – Visual transformers

VGG16 – Visual Geometry Group

CNN – redes neurais convolucionais

Sumário

1 INTRODUÇÃO

A identificação de doenças no corpo humano muitas vezes é desafiadora, uma vez que muitas delas permanecem invisíveis aos olhos. Em tais situações, a doença pode permanecer latente no organismo, dificultando a identificação do problema. O corpo humano reage de diversas formas diante de condições adversas, manifestando-se através de sintomas como dor, sensação de frio ou calor (RAJADANURAKS et al., 2021). Nesse contexto, os exames de raio-X desempenham um papel fundamental na identificação e diagnóstico desses problemas, permitindo a visualização interna do organismo e a avaliação da integridade dos órgãos.

A radiografia do tórax, em particular, é um dos exames mais comuns e eficazes nesse sentido. Por meio desse exame, é possível identificar anomalias no tamanho ou na forma do coração, distúrbios nas veias e artérias, além de problemas nos pulmões, como pneumonia e outras doenças respiratórias (ANIS et al., 2020). Uma das principais vantagens da radiografia torácica é sua acessibilidade, baixo custo e procedimento relativamente simples. Além disso, a quantidade de radiação envolvida é considerada segura, tornando-o uma opção segura para pacientes de todas as idades (RAJADANURAKS et al., 2021).

Nos últimos anos, os avanços tecnológicos no campo da inteligência artificial, especialmente no aprendizado profundo, têm impulsionado significativamente a pesquisa em visão computacional. Nesse contexto, a análise de radiografias torácicas emergiu como um tema de grande relevância. Identificar patologias e doenças a partir dessas imagens representa um desafio significativo. Para abordar essa questão, muitos estudos têm recorrido ao uso de Redes Neurais Convolucionais (CNNs) para a classificação e detecção dessas condições médicas. Automatizar o processo de diagnóstico traz consigo uma série de vantagens, especialmente em ambientes onde radiologistas podem não estar prontamente disponíveis ou em regiões com recursos limitados. Além disso, a automação pode contribuir para uma melhor priorização do fluxo de trabalho médico, reduzindo o tempo gasto em tarefas tediosas e permitindo que os profissionais de saúde se concentrem em aspectos mais críticos do atendimento ao paciente (IRVIN et al., 2019).

A evolução das arquiteturas de redes neurais convolucionais (CNNs) foi impulsionada por competições como a ILSVRC (ImageNet Large Scale Visual Recognition Challenge), onde modelos são avaliados utilizando o extenso conjunto de dados da ImageNet, composto por mais de 14 milhões de imagens e mais de 20 mil categorias. A vitória inicial da AlexNet na ILSVRC-2012 marcou um ponto de virada, incentivando pesquisadores a explorarem e aprimorarem ainda mais as arquiteturas de CNNs (KRIZHEVSKY; SUTSKEVER; HINTON, 2017). Subsequentemente, modelos como VGG (SIMONYAN; ZISSERMAN, 2015) e GoogLeNet (SZEGEDY et al., 2014) apresentaram desempenhos

notáveis ao aumentar tanto a profundidade quanto o tamanho das redes.

A introdução das ResNets foi um marco significativo, ao propor blocos residuais que permitiram a construção de redes ainda mais profundas, evitando o problema de desvanecimento do gradiente (WANG et al., 2019). Essa abordagem inovadora possibilitou a criação de modelos mais complexos e capazes de capturar representações mais abstratas dos dados.

A DenseNet trouxe uma abordagem única ao redefinir a conexão entre as camadas convolucionais. Em vez de simplesmente agregar informações de camadas anteriores, cada camada em uma DenseNet recebe como entrada os mapas de características de todas as camadas anteriores, resultando em uma rede densamente conectada que promove a reutilização eficiente de informações em todas as camadas (HUANG et al., 2018).

Essas diferentes arquiteturas representam avanços significativos na área de visão computacional, cada uma contribuindo com ideias inovadoras para aprimorar o desempenho e a eficiência das CNNs. Essa contínua evolução demonstra o comprometimento dos pesquisadores em desenvolver modelos cada vez mais sofisticados para lidar com desafios complexos de reconhecimento de padrões em imagens.

A introdução dos *transformers* no domínio do processamento de linguagem natural marcou uma transição significativa, substituindo as redes neurais recorrentes como a estrutura fundamental (LIU et al., 2022). Essa mudança proporcionou escalabilidade e eficiência computacional, permitindo o treinamento de modelos com bilhões de parâmetros sem comprometer o desempenho (MAURÍCIO; DOMINGUES; BERNARDINO, 2023). No entanto, seu impacto no campo da visão computacional foi inicialmente limitado. Contudo, com a introdução da camada de *Patchify*, que divide a imagem em uma sequência de blocos, e a incorporação do mecanismo de atenção, os *vision transformers* (ViT) ganharam relevância nesse domínio. Com apenas algumas modificações, essa abordagem tornou-se o uso da arquitetura não apenas para problemas de classificação de imagens, mas também para segmentação e detecção de objetos (DOSOVITSKIY et al., 2020).

A escolha da arquitetura adequada para resolver o problema de classificação de imagens é uma etapa crucial que demanda uma avaliação meticulosa de diversas variáveis. Aspectos como limitações técnicas, métricas de desempenho, consumo de memória e eficácia da rede precisam ser cuidadosamente considerados para determinar a estrutura neural mais adequada à problemática em questão. Nesse contexto, o presente trabalho se propõe a realizar um estudo comparativo das arquiteturas de redes neurais desenvolvidas ao longo dos anos, incluindo a VGG16, ResNet-50 e um ViT hierárquico chamado *swin transformer*, com o intuito de investigar como as inovações trazidas pelo *swin transformer* impactam o desempenho das redes convolucionais. Essa análise aprofundada permitirá compreender melhor as vantagens e desafios de cada abordagem, possibilitando uma escolha mais

informada e eficaz da arquitetura a ser empregada em tarefas de classificação de imagens.

1.1 Definição do Problema

O problema reside em compreender a importância da seleção da arquitetura na determinação do desempenho do modelo em tarefas específicas de aprendizado de máquina como a classificação de imagens. O estudo em questão propõe uma análise entre três arquiteturas amplamente reconhecidas: VGG16, ResNet-50 e o *Swin Transformer*. A pesquisa busca aprimorar as arquiteturas convolucionais em relação ao *swin transformer* por meio da exploração de ajustes nos hiperparâmetros e da implementação de estratégias de treinamento otimizadas. Essa abordagem abrangente visa oferecer conhecimentos valiosos para a seleção e otimização eficaz de arquiteturas de rede neural em diferentes contextos de aplicação.

1.1.1 Objetivo geral

Busca-se identificar as inovações no *swin transformer* que influenciaram positivamente as CNNs no problema de classificação de imagens de radiografia do tórax, evidenciando que, apesar da crescente popularidade das novas arquiteturas baseadas em *transformers*, as CNNs continuam desempenhando um papel significativo devido à sua simplicidade e capacidade de generalização elevada.

1.1.2 Objetivos específicos

A seguir, são apresentados os objetivos específicos.

- Treinar as seguintes redes convolucionais: VGG16 e Resnet-50; e o ViT hierárquico, *swin transformer*.
- Treinar as redes convolucionais com as mudanças.
- Comparar e avaliar o impacto das mudanças nos experimentos.

1.2 Estrutura da monografia

A estrutura deste trabalho foi organizada em cinco capítulos, cada um desempenhando um papel específico na apresentação e desenvolvimento da pesquisa. Iniciando com a Introdução, onde uma breve contextualização sobre o problema de classificação

de radiografias do tórax na área de aprendizado profundo é apresentada, seguida pela exposição dos objetivos gerais e específicos do estudo.

No capítulo subsequente, intitulado "Conceitos Gerais e Revisão de Literatura", é fornecida a fundamentação teórica do trabalho, destacando-se a revisão da literatura existente sobre o tema. Neste contexto, vários artigos relevantes que abordaram o mesmo problema são discutidos, proporcionando uma base sólida para a escolha das técnicas utilizadas ao longo da pesquisa.

A metodologia, abordada no terceiro capítulo, detalha o processo passo a passo adotado para resolver o problema proposto. Aqui, são explicados em detalhes o pré-processamento dos dados, as decisões tomadas em relação às arquiteturas e hiperparâmetros dos modelos, bem como os resultados alcançados.

Posteriormente, no capítulo de "Apresentação e Análise dos Resultados", é realizada uma análise minuciosa dos resultados obtidos a cada modificação na arquitetura, levando em consideração aspectos como acurácia, performance e eficiência dos modelos desenvolvidos.

Por fim, as "Conclusões e Trabalhos Futuros"encerram o trabalho, onde os resultados são sintetizados e as conclusões são tiradas com base nos achados da pesquisa. Além disso, são apontadas direções para futuros estudos e possíveis extensões do trabalho desenvolvido.

2 CONCEITOS GERAIS E REVISÃO DA LITERATURA

2.1 Conceitos

2.1.1 Camadas convolucionais

As camadas convolucionais são compostas por diversos mapas de características, obtidos após aplicar os filtros convolucionais nas imagens como visto na figura 1. Esses filtros, representados por matrizes de pesos, podem ter tamanhos variados, como 5x5 ou 7x7, para imagens 2-D de canal único. A convolução propõem uma forma de extrair as características utilizando os filtros convolucionais resultando na extração de características importantes das imagens (WANG et al., 2019).

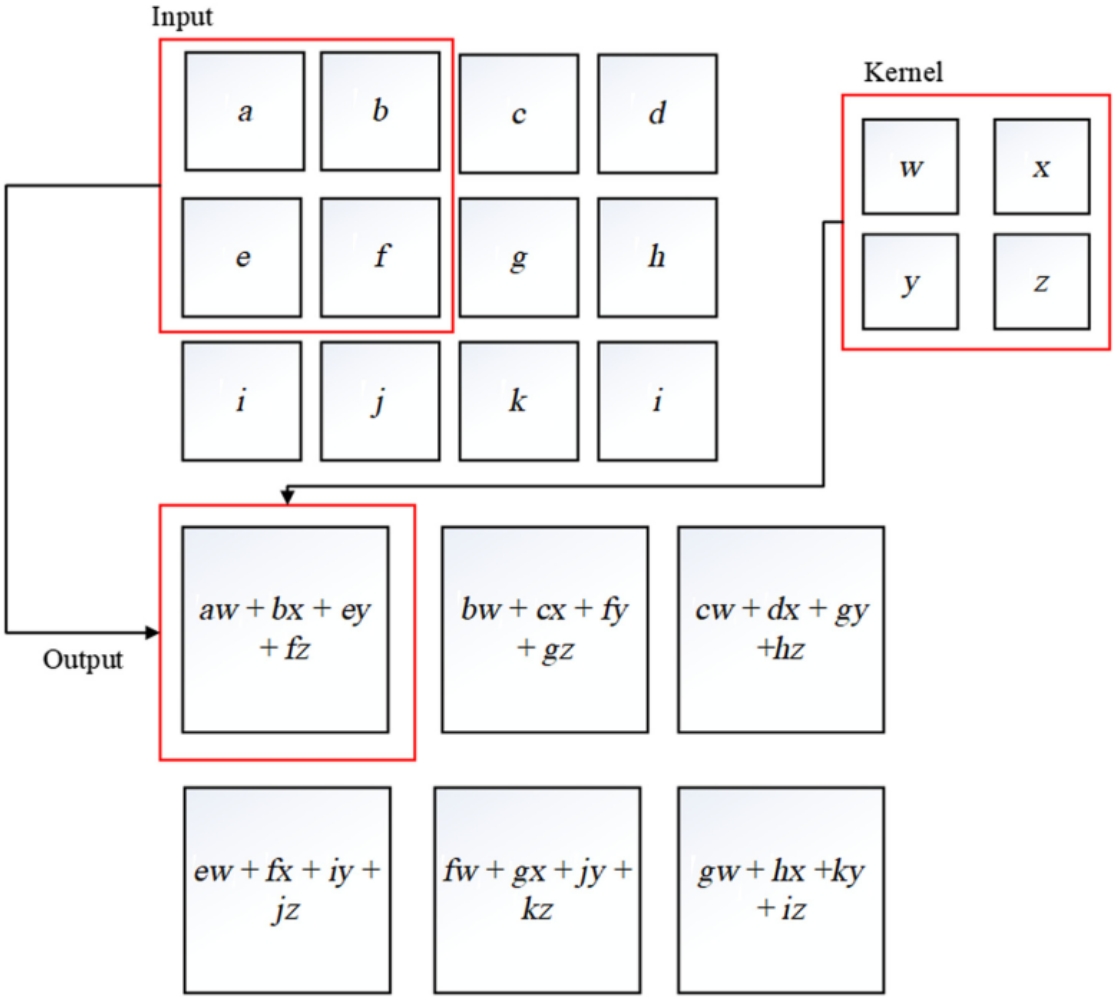

Figura 1: Convolução 2-D.
Adaptado de (WANG et al., 2019)

2.1.2 Subamostragem

Geralmente, uma camada de subamostragem, como apresentado na figura 2 é intercalada periodicamente entre as camadas de convolução. Essa camada tem como função reduzir gradualmente o tamanho espacial dos dados, o que contribui para a diminuição do número de parâmetros na rede e reduz o consumo de recursos computacionais. Além disso, as camadas de subamostragem têm a capacidade de aprender algumas características das entradas. Os métodos mais comuns utilizados para isso são o subamostragem máxima (*max pooling*) e o subamostragem média global (*global average pooling*). Essas técnicas desempenham um papel fundamental na redução da dimensionalidade dos dados ao longo da rede, permitindo uma representação mais compacta e eficiente das características relevantes.(WANG et al., 2019)

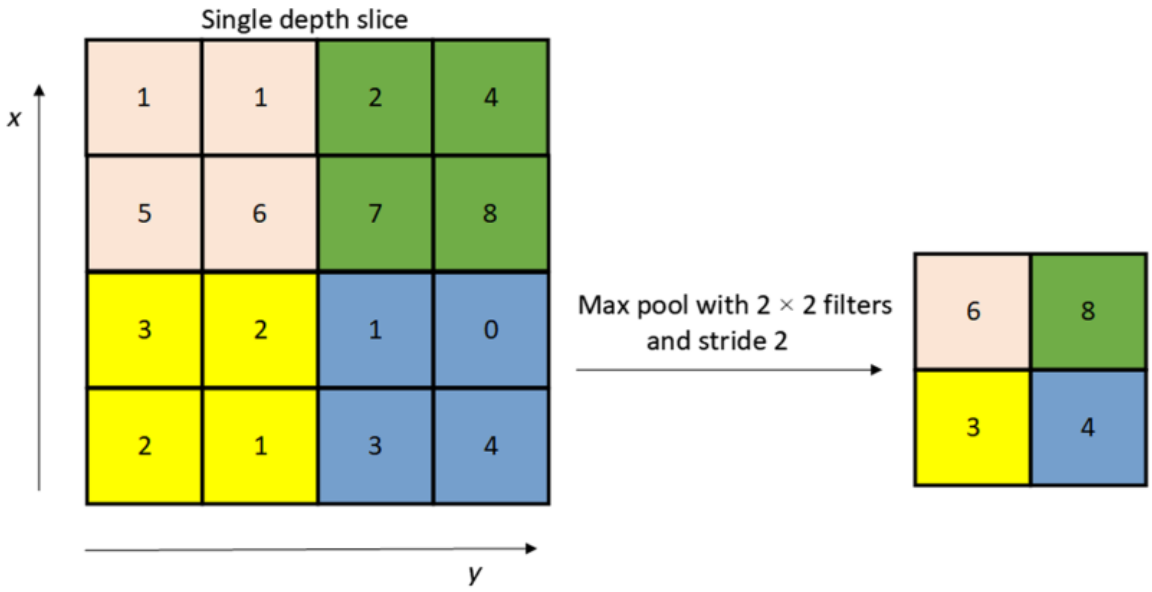

Figura 2: Max pooling.
Adaptado de (WANG et al., 2019)

2.1.3 Funções de ativação

A função de ativação desempenha um papel crucial na introdução de não linearidades na rede neural, permitindo capturar padrões e relacionamentos complexos nos dados de entrada. Antes da popularização da função de ativação ReLU, as redes neurais tradicionais predominantemente empregavam a função sigmoide. Em geral, as funções sigmoides podem ser categorizadas em sigmoides logísticas e sigmoides tangente hiperbólica. No entanto, devido aos problemas apresentados pela função sigmoide, como saturação e gradiente instável, a ReLU emergiu como uma alternativa amplamente adotada, trazendo consigo diversas vantagens, tais como inibição unilateral, ampla faixa de excitação,

ativação esparsa e mitigação do problema do gradiente desvanecente (WANG et al., 2019).

Contudo, a função ReLU também apresenta desafios, nos quais uma grande fração dos neurônios pode se tornar inativa ou permanecer não responsiva, comprometendo o processo de aprendizado. Recentemente, a função de ativação Gaussian Error Linear Unit (GELU) emergiu como uma alternativa à ReLU, oferecendo gradientes mais suaves, diferenciabilidade e a capacidade de aproximar amplamente a função ReLU amplamente utilizada (LEE et al., 2023). Essa nova função de ativação mostra-se promissora para lidar com os desafios encontrados pela ReLU, contribuindo para uma melhor eficiência e desempenho das redes neurais. As comparações podem ser vistas na figura 3.

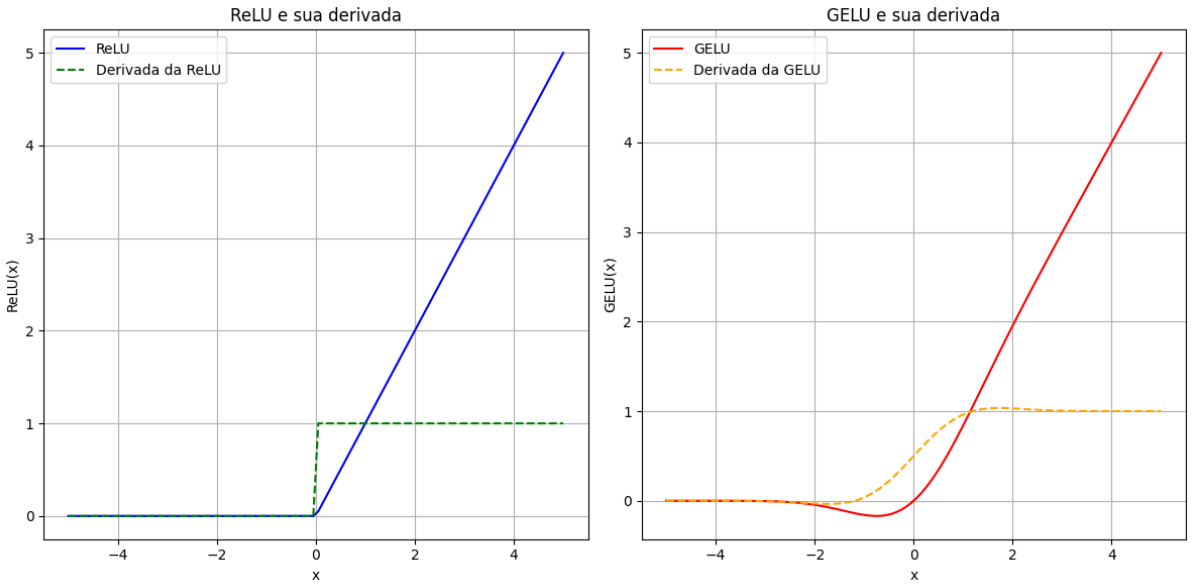

Figura 3: Funções de ativações.

2.1.4 *Batch Normalization* e *Layer Normalization*

No processo de treinamento de redes neurais profundas, a distribuição de entrada na camada oculta desempenha um papel crucial. Quando essa distribuição não é equilibrada, gradualmente a distribuição global se aproxima dos limites superior e inferior da faixa de valores da função não linear, resultando em uma convergência de treinamento lenta. Para contornar esse problema, é necessário normalizar os dados (WANG et al., 2019).

Os métodos de normalização têm como objetivo mitigar a variação interna das variáveis em redes neurais profundas, realizando a normalização das entradas em cada camada. Essas técnicas promovem dinâmicas de treinamento mais estáveis, possibilitando uma convergência mais rápida e reduzindo a dependência dos gradientes na distribuição de entrada. Com isso, os métodos de normalização se tornaram componentes essenciais das

arquiteturas modernas de aprendizagem profunda, viabilizando o treinamento de redes mais profundas com maiores taxas de aprendizagem (LEE et al., 2023).

Ao empregar a técnica de Batch Normalization (BN), em qualquer camada oculta h, as entradas são submetidas a uma ativação não linear para obter a saída. Para cada neurônio (ativação) em uma dada camada, as pré-ativações são normalizadas para terem média zero e desvio padrão unitário. O BN é uma técnica amplamente utilizada que reduz a variação interna das variáveis, normalizando ativações em um lote durante o treinamento (LEE et al., 2023).

Por outro lado, a Layer Normalization (LN) é uma outra técnica de normalização que aborda algumas limitações do BN, como a dependência do tamanho do minibatch e o desempenho reduzido em redes recorrentes. Diferentemente do BN, que normaliza as ativações em lotes durante o treinamento, o LN normaliza as ativações na dimensão do recurso em cada camada (LEE et al., 2023). Os exemplos podem ser visto na figura 4

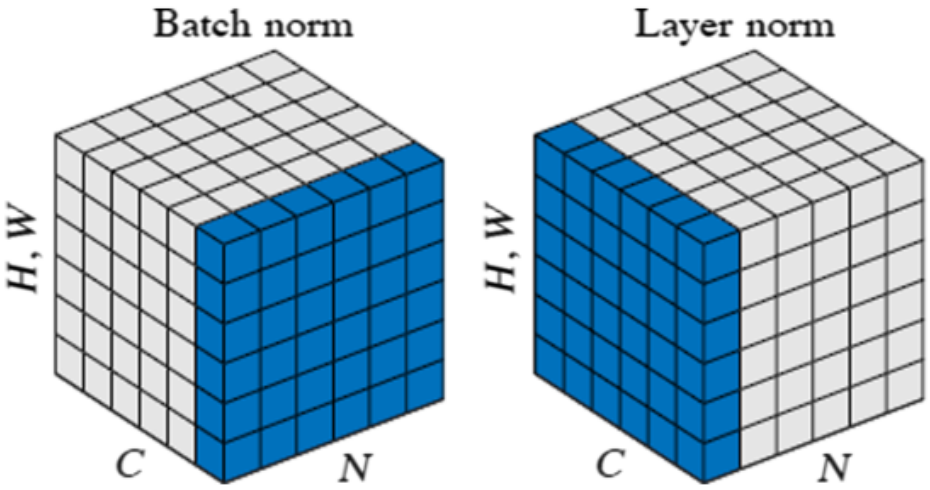

Figura 4: Normalizações.
Fonte: Adaptado de (WANG et al., 2019)

2.2 Arquiteturas

A seguir, serão explicadas todas as arquiteturas que foram utilizadas no trabalho, incluindo suas contribuições para área de aprendizado profundo e classificação de imagens.

2.2.1 VGG16

Redes neurais mais rasas enfrentam certas limitações em tarefas de reconhecimento de imagens em larga escala. Com o intuito de explorar o potencial das redes neurais profundas, Simonyan e Zisserman (2015) propuseram a Visual Geometry Group, conhecida

como VGG como apresentado na figura 5. A principal contribuição da VGG é uma implementação de uma rede neural profunda utilizando filtros 3x3 produzindo uma melhoria em relação as arquiteturas anteriores.

Uma das inovações da VGG reside principalmente no uso de campos receptivos bem menores em toda a rede em comparação com campos receptivos maiores com diferentes strides. Ao utilizar duas convoluções 3x3, obtém-se um efeito semelhante a uma convolução 5x5; e ao empregar três convoluções 3x3, tem-se um efeito similar de uma convolução 7x7. Essa estratégia é adotada pela rede por duas razões principais: em primeiro lugar, ela contém três camadas utilizando a função ReLU em vez de apenas uma, tornando a função de decisão mais discriminatória; em segundo lugar, isso possibilita a redução do número de parâmetros. Na competição de classificação de imagens ILSVRC-2014, a VGG alcançou o segundo lugar, com uma taxa de erro de top-5 de 7.3%, demonstrando que redes neurais mais profundas podem obter melhores desempenhos (WANG et al., 2019).

Figura 5: VGG16 arquitetura.

2.2.2 ResNet-50

Apesar da constatação de que aumentar a profundidade e tamanho das redes neurais pode melhorar sua performance, descobriu-se que isso resulta na ocorrência de um problema conhecido como o desaparecimento do gradiente durante o *backpropagation*. Essa questão foi atribuída à inicialização normalizada e às camadas intermediárias de normalização. Mesmo com a aplicação de métodos para mitigar esse problema, o aumento do número de camadas na rede levou à saturação ou mesmo à redução da acurácia do treinamento, não devido ao *overfitting*, mas sim devido à dificuldade de otimização de redes neurais profundas. Nesse contexto, a ResNet-50 foi desenvolvida para abordar essas questões, evitando o desaparecimento do gradiente ao aumentar o tamanho e a profundidade das redes, resultando em uma melhoria de desempenho significativa. A ResNet, construída com blocos residuais, é capaz de superar a limitação de 100 camadas e até mesmo alcançar 1000 camadas.

A arquitetura da ResNet-50 é predominantemente composta por blocos residuais. No bloco de aprendizagem residual, assume-se que a função original a ser aprendida é

H(x), e o bloco de aprendizagem residual é então definido como F(x) = (x) - x como apresentado na figura 6. O problema da degradação sugere que os solucionadores podem ter dificuldades em aproximar mapeamentos de identidade por múltiplas camadas não lineares. Com a reformulação da aprendizagem residual, se os mapeamentos de identidade forem ótimos, os solucionadores podem simplesmente direcionar os pesos das múltiplas camadas não lineares para zero para aproximar os mapeamentos de identidade (HE et al., 2016).

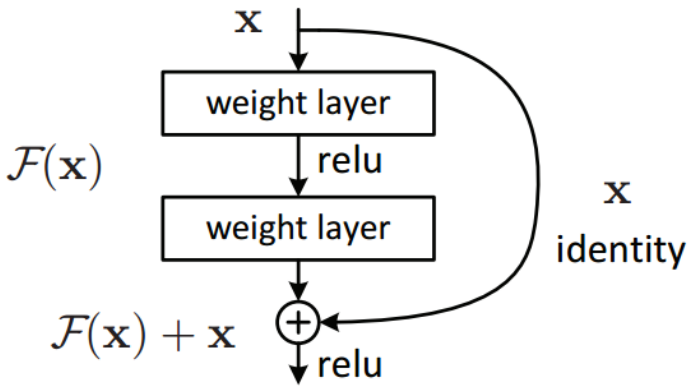

Figura 6: Bloco residual da ResNet-50.
Fonte: (HE et al., 2016)

A ResNet-50 utiliza dois tipos de blocos residuais para construir sua arquitetura. O primeiro, denominado bloco de *bottleneck*, consiste em uma sequência de três camadas convolucionais: 1x1, 3x3 e 1x1. Essa estrutura é aplicada recorrentemente várias vezes dentro de cada bloco, facilitando a extração e transformação de características. O segundo componente é responsável por conectar esses blocos convolucionais. Semelhante ao bloco de *bottleneck*, ele consiste em três camadas convolucionais: 1x1, 3x3 e 1x1. No entanto, uma camada adicional é incluída para preservar a entrada original. Essa conexão permite a combinação de informações derivadas dos cálculos residuais com a entrada original, facilitando o fluxo de gradientes e aumentando a capacidade do modelo de aprender representações complexas de forma eficaz.

A ResNet, como apresentada na figura 7, construída com blocos de aprendizagem residual, alcançou resultados impressionantes, conquistando o primeiro lugar na competição de classificação de imagens ILSVRC-2015, com *top-5 error* de apenas 3,57% (HE et al., 2016).

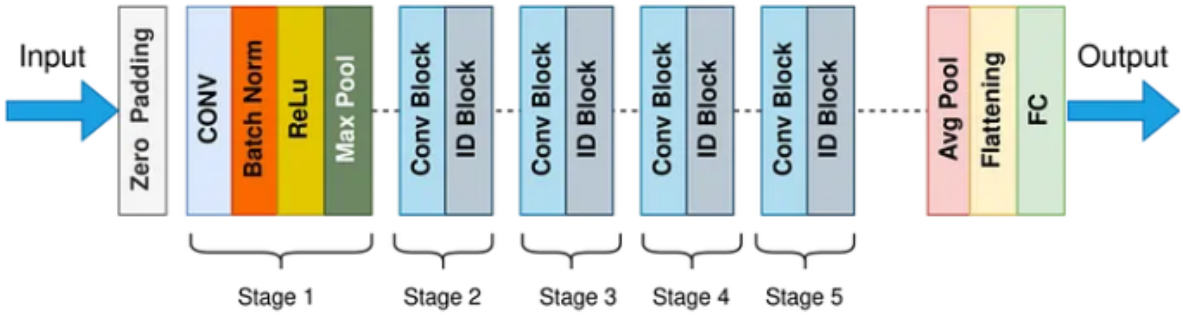

Figura 7: ResNet-50 arquitetura.
Fonte: https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at-the-model-architecture-and-code-implementation-d8d8fa67e46f

2.2.3 Vision Transformer

A abordagem do Vision Transformer (ViT) representa uma inovação significativa em relação às redes neurais convolucionais tradicionais. Enquanto as CNNs convencionais processam imagens em sua totalidade, o ViT adota uma estratégia diferente, tratando as imagens como uma sequência de blocos de imagens 2D, cada um com dimensões de 16x16 pixels. Esses blocos são então convertidos em vetores numéricos, conhecidos como embeddings, que servem como entrada para o modelo.

O processo de codificação do ViT é composto por uma série de N blocos, alternando camadas de auto-atenção, o que permite a troca de informações entre diferentes partes da imagem. Após a aplicação da auto-atenção, são utilizadas camadas de normalização e uma camada linear (MLP), que opera sobre os vetores que contêm os pesos calculados durante o processo de atenção. Essa operação resulta na saída ou serve como entrada para as próximas camadas do modelo. Os detalhes podem ser visualizados na figura 8

O sucesso do ViT é notável quando pré-treinado em grandes conjuntos de dados, como o ImageNet-21k ou JFT-300M, e transferido para tarefas com menor volume de dados. Os resultados obtidos pelo melhor modelo ViT são impressionantes: uma precisão de 88,55% no ImageNet, 90,72% no ImageNet-ReaL, 94,55% no CIFAR-100 e 77,63% no conjunto VTAB, que consiste em 19 tarefas diferentes. Além disso, o ViT demanda consideravelmente menos recursos computacionais em comparação com as CNNs tradicionais, mantendo um desempenho competitivo em tarefas de reconhecimento de imagem (DOSOVITSKIY et al., 2020).

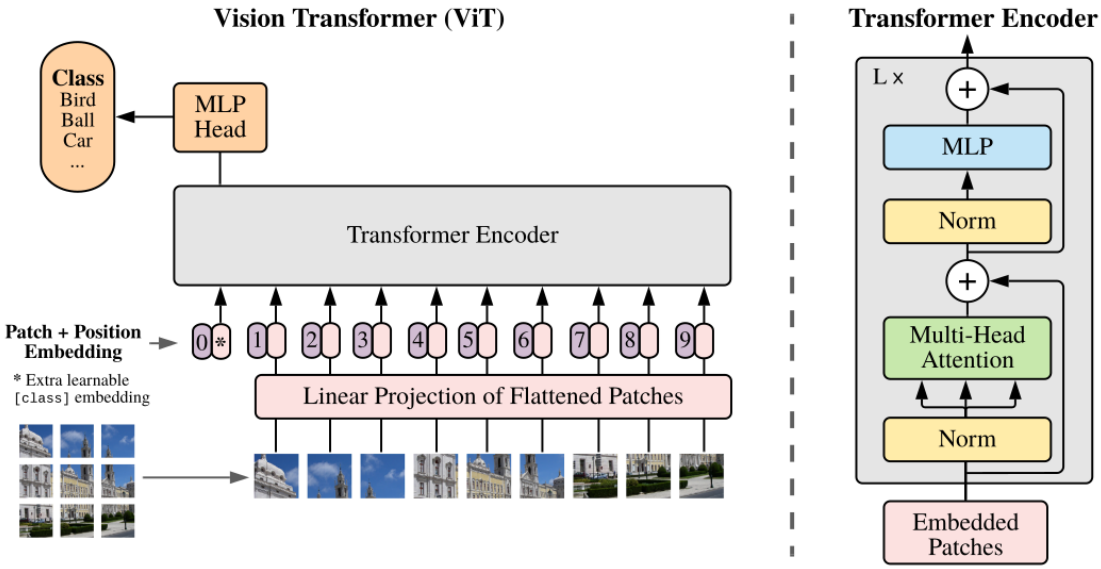

Figura 8: Transformers arquitetura.
Fonte: Adaptado de (DOSOVITSKIY et al., 2020)

2.2.4 Swin Transformers

Um dos desafios enfrentados ao utilizar o *vision transformer* é o sistema de atenção global, o qual possui uma complexidade quadrática conforme o tamanho da entrada. Essa abordagem pode ser adequada para a classificação de imagens do conjunto de dados ImageNet, porém apresenta dificuldades quando aplicada a imagens de resolução mais elevada. Em resposta a esse desafio, foi proposto o *swin transformer*, como apresentado na figura 9, uma arquitetura que constrói mapas de características hierárquicos e possui uma complexidade computacional linear em relação ao tamanho da imagem.

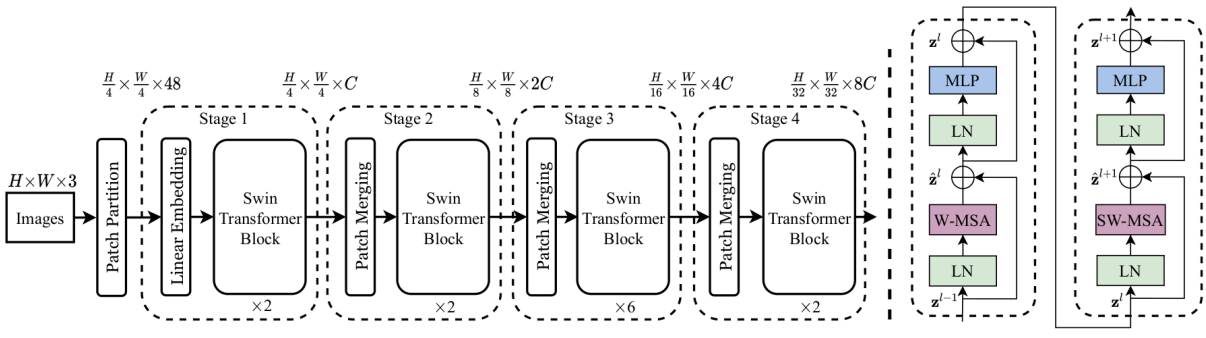

Figura 9: Arquitetura do *swin transformer*
Fonte: Adaptado de (LIU et al., 2021)

A complexidade computacional linear é alcançada ao calcular a autoatenção localmente em janelas não sobrepostas que particionam uma imagem, conforme ilustrado na

figura 10. O número de *patches* em cada janela é fixo, o que resulta em uma complexidade linear em relação ao tamanho da imagem. Além disso, uma das principais inovações dessa arquitetura é o deslocamento da partição da janela entre as camadas de *self-attention* consecutivas. Esse método permite uma melhor conexão entre as janelas da camada anterior, melhorando assim a capacidade de modelagem Liu et al. (2021).

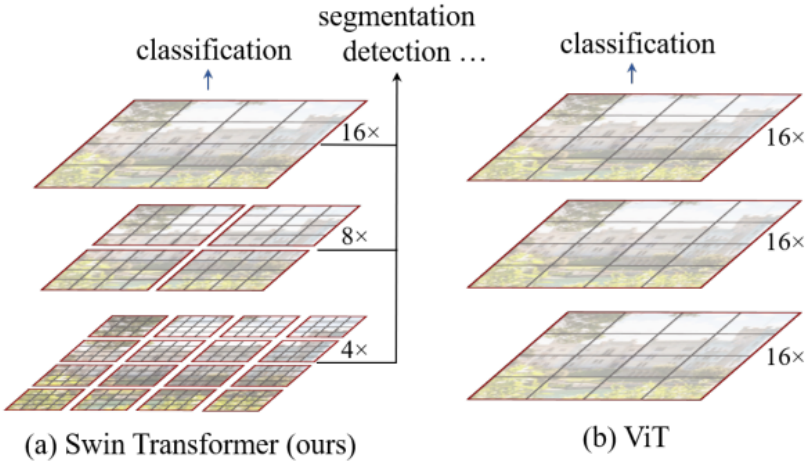

Figura 10: Exemplo de cálculo das imagens no swin transformers.
Fonte: Adaptado de (LIU et al., 2021)

2.3 CheXpert

O conjunto de dados do CheXpert é uma vasta coleção de imagens de radiografias de tórax, essencial para a interpretação médica. Este conjunto abrange um total de 224.316 radiografias, provenientes de 65.240 pacientes, todas elas rotuladas conforme a presença ou ausência de 14 observações radiográficas comuns, a figura 11 apresenta essas observações. Com o intuito de automatizar a detecção dessas observações nos relatórios radiológicos, foi desenvolvido um rotulador especializado. Tal ferramenta não apenas agrega eficiência ao processo de interpretação, mas também aborda as inerentes incertezas presentes na interpretação radiográfica. (IRVIN et al., 2019).

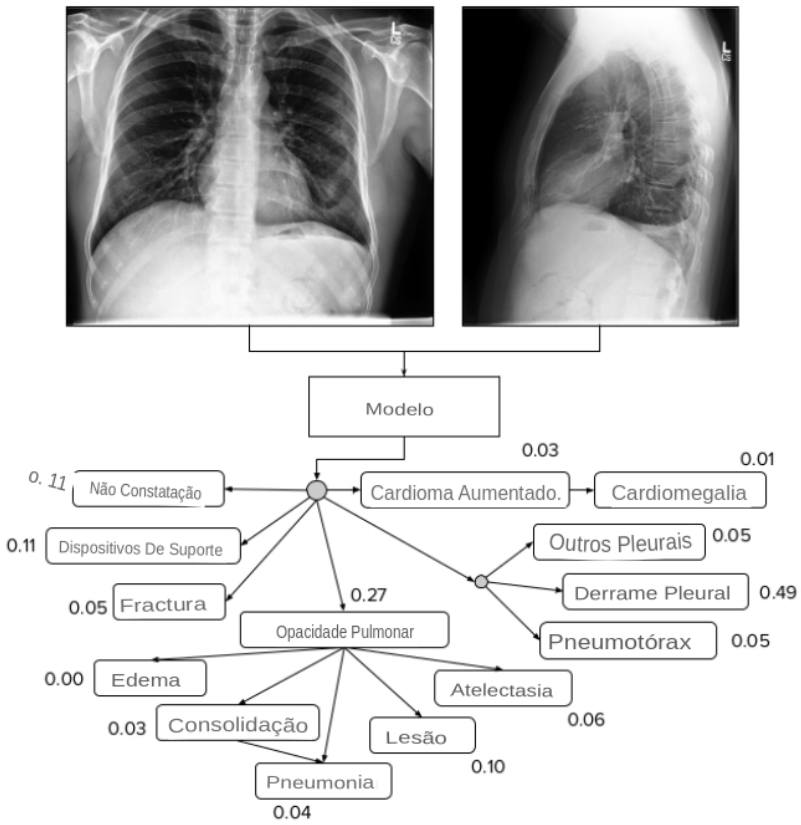

Figura 11: Conjunto de dados *chexpert* e as probabilidades de cada classe para a imagem avaliada.

Fonte: Adaptador de (IRVIN et al., 2019)

2.4 Acurácia

Para o cálculo da acurácia, foi utilizado a biblioteca scikit-learn (sklearn). Esta biblioteca oferece um método chamado *accuracy_score*, o qual mensura as instâncias classificadas corretamente em relação ao total de instâncias. Matematicamente, a acurácia é calculada pela seguinte fórmula:

$$\text{Acurácia} = \frac{\text{Número de predições corretas}}{\text{Número total de predições}} \times 100\% \tag{1}$$

2.5 Trabalhos relacionados

Bressem et al. (2020) avaliaram dezesseis arquiteturas, incluindo a VGG16 e a ResNet-50, em dois conjuntos de dados distintos: CheXpert e COVID-19. A pesquisa destaca que a escolha da arquitetura exerce um impacto significativo na performance da classificação. Observou-se que redes neurais profundas, como a ResNet-101 e a ResNet-50, tendem a alcançar valores elevados na métrica *areas under the receiver operating characteristics curves* (AUROC) no conjunto de dados do CheXpert. No entanto, o estudo

também revela que redes neurais mais rasas, exemplificadas pela AlexNet e VGG16, com menos camadas convolucionais, conseguem atingir um bom resultado na análise de radiografias do tórax, avaliando na métrica *area under the precision-recall curve* (AUPRC).

Liu et al. (2022), destacam-se a importância das ConvNets no contexto da visão computacional e introduzem-se os *swin transformers*, uma arquitetura hierárquica baseada nos *vision transformers*. Esses *transformers* têm sido amplamente adotados como espinha dorsal para resolver problemas de visão computacional. O estudo aborda os componentes específicos dos *swin transformers* e aplica suas decisões de design em uma rede convolucional conhecida como ResNet-50. Uma comparação é feita com base em métricas de acurácia, desempenho e tamanho do modelo. Os resultados indicam que a nova rede convolucional, denominada ConNeXt, pode alcançar uma acurácia top-1 satisfatória no conjunto de dados ImageNet e também demonstra um desempenho superior em outras tarefas, como detecção de objetos e segmentação, quando comparada a um modelo ViT.

Wang et al. (2019), traçam a evolução das CNNs, destacando as decisões tomadas ao longo do tempo que possibilitaram o desenvolvimento de novas arquiteturas capazes de lidar eficientemente com uma variedade de tarefas em visão computacional, tais como detecção de objetos, segmentação e classificação de imagens. O estudo realiza uma análise comparativa de diversas arquiteturas para o problema específico de classificação de imagens, ressaltando a importância das redes neurais profundas, dos blocos residuais para mitigar o problema do desvanecimento do gradiente e das redes densas. Além disso, o estudo destaca a relevância contínua das CNNs no futuro, com ênfase em técnicas como transferência de aprendizado e mecanismos de atenção visual, bem como a necessidade de infraestrutura computacional robusta para treinar esses modelos com eficácia.

Maurício, Domingues e Bernardino (2023) apresentam uma análise abrangente de diversos estudos que comparam arquiteturas convolucionais com o uso do (ViT) em problemas de classificação de imagens com o objetivo de responder questões como: A arquitetura do ViT pode ter uma performance melhor do que a CNN independente das características do dataset e o que influencia as CNNs para não ter uma performance tão boa quanto um ViT? A conclusão do estudo aponta que o ViT pode ser mais eficaz em problemas de classificação de imagens em conjuntos de dados pequenos, devido à sua camada de *self-attention*. No entanto, quando treinado com conjuntos de dados limitados, o ViT pode apresentar menor capacidade de generalização e desempenho inferior em comparação com as CNNs. Essa análise ressalta a importância de considerar o contexto específico do conjunto de dados e as características da tarefa ao escolher entre diferentes arquiteturas de rede neural para problemas de classificação de imagens.

Darapaneni, Krishnamurthy e Paduri (2020) abordam mais de 18 arquiteturas amplamente utilizadas, todas avaliadas no conjunto de dados CIFAR-10 com relação a desempenho, acurácia e consumo de memória. O artigo conclui que arquiteturas como

VGG16 e AlexNet apresentaram bons resultados quando consideramos apenas o consumo de memória. Por outro lado, ResNet-50 e MobileNetV2 são arquiteturas recomendadas, levando em conta não apenas a acurácia, mas também o consumo de memória e o desempenho geral. Essas descobertas evidenciam a importância de considerar múltiplos aspectos ao escolher uma arquitetura de rede neural para um determinado problema de visão computacional.

Tabela 1: Trabalhos Relacionados

Título do Trabalho	**Base Utilizada**	**Arquiteturas**	**Métricas Utilizadas**
(BRESSEM et al., 2020)	Covid-19 e CheXpert	AlexNet, ZFNet, VGG, GoogLeNet, Inception-v2, SqueezeNet, Densenet e ResNet	ROC e AUROC
(LIU et al., 2022)	ImageNet-1k e ImageNet-22K	ResNet-50 e *swin transformers*	acurácia top-1 IN-1K
(WANG et al., 2019)	ImageNet e CIFAR-10	AlexNet, ZFNet, VGG, GoogLeNet, Inception-v2, (BN-inception), PReLU, Inception-v3 e ResNet	Top-5 (test) error
(MAURÍCIO; DOMINGUES; BERNARDINO, 2023)	Coleção de artigos que comparam ViT com CNNs	CNNs no geral e ViT	-
(DARAPANENI; KRISHNAMURTHY; PADURI, 2020)	CIFAR-10	LeNet-50, AlexNet, Zfnet, VGG16, Inception V1, ResNet-50, SqueezeNet entre outras.	top-5 error

3 METODOLOGIA

Nesta seção, é descrito a abordagem adotada para aplicar as modificações inspiradas no *swin transformer* nas CNNs como a VGG16 e ResNet-50. O treinamento foi conduzido de maneira consistente, utilizando o mesmo conjunto de dados e hiper-parâmetros semelhantes, o que garantiu uma comparação equitativa entre os modelos modificados e os originais. Adicionalmente, foram inseridas camadas de Batch Normalization (BN) entre as camadas da VGG16 por razões metodológicas, uma vez que uma das modificações propostas envolve a redução do uso de BN. Esta escolha metodológica foi feita para garantir a integridade do experimento e possibilitar uma análise comparativa precisa dos resultados.

3.1 Pré-processamento

Como mencionado anteriormente, o conjunto de dados do CheXpert consiste em 14 observações, cada uma representando a presença ou ausência de 12 patologias, além de dispositivos de suporte e a ausência de doenças. Cada paciente possui imagens frontais e laterais, nas quais as patologias são classificadas em quatro valores distintos: 1 quando a patologia está presente, 0 quando está ausente, -1 quando há incerteza quanto à sua presença e um espaço em branco quando não há observações sobre a condição naquela imagem.

A partir disso, foi realizado um pré-processamento para determinar as classes com maior representatividade no conjunto de dados. Após análise, as classes de interesse escolhidas foram *Pleural Effusion* (PE) e *No Finding* (NF). Para balancear os dados, foi utilizada a função do Pandas para selecionar aleatoriamente 5000 amostras de cada classe, totalizando 10 mil amostras de treinamento. Quanto às incertezas e valores em branco, adotou-se uma abordagem semelhante ao trabalho de (IRVIN et al., 2019), substituindo esses valores por 0, indicando a ausência da patologia. Além disso, as imagens frontais e laterais foram mantidas, com 8000 imagens frontais e 2000 imagens laterais. Do total, 6000 imagens correspondem a pacientes do sexo masculino e 4000 a pacientes do sexo feminino. As imagens foram normalizadas mantendo o tom RGB. O conjunto de teste consiste em 2000 imagens no total, com 1000 imagens para cada classe.

3.2 Treinamento

Feito o treinamento inicial das redes convolucionais sem nenhuma modificação e também do *swin transformer*, pois dessa forma podemos ter a acurácia base para comparar com as arquiteturas convolucionais modificadas, com isso foi obtido os seguintes resultados de acurácia no conjunto de teste para os 3 modelos além de seus respectivos

tamanhos. Para a VGG16 o resultado inicial foi de 83,30% com o tamanho de 68.41MB, a Resnet-50 obteve 77,30% com 89,79MB e por fim o *swin transformer* conseguiu uma acurácia de 78,25% com 3MB. Com base nesses resultados, foi possível comparar as CNNs, modificadas com as inovações do *swin transformer* e avaliar o impacto da performance e custo das redes para o problema de classificação de imagens.

Posteriormente, uma série de decisões foi tomada, abrangendo: 1) Aumento da profundidade, 2) Modificação do bloco inicial, 3) Variação no tamanho dos filtros, 4) Mudança de função de ativação, 5) Redução do número de ativações, 6) Minimização do uso de normalizações, e 7) Modificação do método de normalização.

O treinamento foi realizado utilizando as configurações apresentadas na tabela 2. O *weight decay* foi escolhido baseado no treinamento feito pelo trabalho de (LIU et al., 2022), otimizador utilizado foi Adam devido sua popularidade e uso em problemas de visão computacional, a taxa de aprendizado foi escolhido um valor menor, pois quando esse valor era aumentado resultava em uma acurácia reduzida, mas no caso do *swin transformers* aconteceu o contrário. O tamanho do *batch* e a quantidade de épocas foi escolhida devido às limitações da máquina que foi o Google Colab utilizando a GPU L4 de 22GB de memória da gpu.

Configuração de (pre-)treinamento	VGG16	ResNet-50	Swin
weight decay	-	0.05	0.05
Otimizador	Adam	Adam	Adam
Taxa de aprendizado	1,00E-04	1,00E-04	2,00E-04
Tamanho do *batch*	16	16	16
Épocas de treinamento	15	15	15

Tabela 2: Configurações utilizadas no treinamento das redes escolhidas

3.3 Aumento da profundidade

Considerando a importância de aumentar a profundidade da rede e a observação de que o Swin Transformer segue uma proporção diferente, conforme apresentado na figura 9, onde possui blocos no tamanho de (2, 2, 6, 2), optou-se por aumentar a profundidade da VGG16, que consiste em 5 blocos, além das últimas camadas densas. Cada bloco possui a seguinte quantidade de camadas convolucionais: de (2, 2, 3, 3, 3) para (3, 3, 9, 3, 3) na VGG16 e de (3, 4, 6, 3) para (3, 3, 9, 3) na ResNet-50.

Vale ressaltar que o aumento foi realizado com base no seguinte princípio: ao dividir o tamanho dos blocos do Swin Transformer por 2, obtemos (1, 1, 3, 1). Aplicando a mesma lógica à ResNet-50, mas dividindo por 3, chegamos a (1, 1, 3, 1). Como a VGG16 possui um bloco adicional, resultou em um tamanho semelhante de (1, 1, 3, 1, 1). No entanto, ao comparar a acurácia do modelo original, constatou-se que a VGG16 não

conseguiu generalizar e, portanto, não aprendeu. Quanto à ResNet-50, houve uma redução na acurácia, levando à decisão de descartar essa modificação. Além disso, observou-se um leve aumento no tamanho da rede tanto para a VGG16 quanto para a ResNet-50.

3.4 Modificação do bloco inicial

O bloco inicial de uma arquitetura é crucial para determinar como as imagens serão processadas no início da rede. Tanto as CNNs quanto o *swin transformer* seguem uma abordagem semelhante, reduzindo progressivamente a imagem para um mapa de características correspondente. Na arquitetura padrão da ResNet-50, o bloco inicial é composto por uma camada de convolução com filtros 7x7 com passo 2, seguida por um *pooling máximo*, resultando em uma redução da resolução das imagens de entrada em 4x. Por sua vez, o bloco inicial da VGG16 consiste em duas camadas convolucionais 3x3 consecutivas com passo 1 e uma camada de *max pooling*, reduzindo pela metade as dimensões espaciais das imagens de entrada.

No caso do *swin transformer*, uma estratégia de "*patchify*"mais agressiva é adotada no bloco inicial. Nesse caso, os blocos iniciais foram atualizados para uma camada convolucional com filtro 4x4 e passo 4 seguido de uma camada de BN e uma ativação ReLu. Embora essas atualizações tenham sido implementadas, a VGG16 não obteve uma melhoria mas sim uma leve redução na acurácia e a Resnet-50 obteve seu melhor resultado. No entanto, observou-se uma leve redução no tamanho dos modelos. Essas alterações destacam a importância do bloco inicial na arquitetura da rede e como diferentes abordagens podem afetar os resultados finais e o tamanho do modelo.

3.5 Tamanho do filtro

Considerando a abordagem do *swin transformer*, que se destaca por seu sistema de atenção, permitindo que cada camada tenha um campo receptivo global, decidimos explorar o uso de filtros de tamanho variado em modelos convolucionais como a VGG16 e a ResNet-50. Ao realizar uma série de experimentos com filtros de tamanhos 5x5, 7x7 e 9x9, observamos resultados distintos. Embora a aplicação desses filtros à ResNet-50 não tenha gerado melhorias em relação ao modelo anterior modificado, notamos que a VGG16 obteve um desempenho superior com o filtro 7x7. Além disso, constatamos que o aumento do tamanho do filtro resulta em um aumento considerável no tamanho da rede, sendo o maior tamanho observado na VGG16 e na ResNet-50 com o filtro 9x9. Essas descobertas ressaltam a importância de considerar o impacto do tamanho do filtro na arquitetura e no desempenho dos modelos convolucionais.

3.6 Relu para Gelu

Embora a função de ativação ReLU seja amplamente utilizada devido à sua simplicidade e eficiência, a função Gelu é adotada em diversos modelos de transformers, incluindo o Bert (DEVLIN et al., 2019) e o GPT-2 (RADFORD et al., 2019). Diante disso, as ativações ReLU empregadas na ResNet-50 e na VGG16 foram substituídas. Entretanto, não houve melhoria na ResNet-50 e a VGG16 obteve o melhor resultado, e essa alteração não teve um impacto significativo na quantidade de parâmetros.

3.7 Poucas ativações

Além do ViT, outras arquiteturas também usam poucas ativações como a EfficientNet Tan e Le (2020) que utiliza um coeficiente composto que permite aumentar a profundidade, a largura e a resolução da rede sem a necessidade de usar muitas camadas de normalização comparado a outras redes tradicionais. MobileNet Howard et al. (2017) foi projetado especificamente para aplicações de visão computacional e incorporadas, onde os recursos computacionais são limitados. Essa arquitetura utiliza convoluções separáveis em profundidade para reduzir o número de parâmetros e ativações, mantendo o desempenho. Levando em consideração que os *transformers* tem apenas um bloco de ativação na camada de multi *perceptron* e da implementação dessa mudança nas redes anteriores, foi decidido aplicar essa mesma condição as arquiteturas escolhidas.

3.8 Poucas normalizações

Enquanto as arquiteturas baseadas em *transformers* geralmente incluem um número limitado de normalizações, há casos em que redes mais compactas, como a SqueezeNet Iandola et al. (2016), dispensam o uso dessas técnicas para manter a eficiência, reduzindo assim o número de camadas de *batch normalization*. No entanto, ao remover essas camadas, não houve uma melhora significativa nas redes mencionadas.

3.9 Mudança de normalização

A introdução da normalização de batch (BN) como um componente essencial nas redes neurais convolucionais (CNN) foi um avanço significativo, uma vez que contribui para melhorar a convergência durante o treinamento e prevenir o sobreajuste. Apesar de existirem outras técnicas de normalização disponíveis, como a normalização de camada (LN), a BN continua sendo a escolha predominante. No entanto, ao lidar com o *swin transformer*, observou-se que a LN tem sido preferencialmente adotada como a técnica principal de normalização. Substituir as camadas de BN por LN nas arquiteturas de

CNNs não produziu os resultados esperados em termos de acurácia. Em ambos os casos, os modelos não foram capazes de aprender de forma eficaz, indicando assim a importância de escolher a técnica de normalização mais adequada para o tipo específico de arquitetura de rede neural.

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

A seguir, é feito uma anaĺise de resultados dos experimentos nas tabelas 3 e 4 Para a VGG16, é observado que ao aumentar a profundidade da rede resultou no não aprendizado. Isso pode ser atribuído ao fato de que, com o aumento da profundidade, existe a possibilidade de acontecer o desaparecimento do gradiente. No entanto, apesar da robustez da ResNet-50 é possível que tenha reduzido tanto a informação das imagens que tornou o aprendizado das características mais difícil ao longo de todas as etapas da rede.

Ao modificar o bloco inicial das arquiteturas, observamos uma redução significativa na dimensão da imagem. Isso ocorre porque o passo é definido como 4 e o filtro da convolução é de 4x4, resultando em um único mapa de características de saída com dimensões espaciais aproximadas de 56x56. O passo de 4 move a janela do filtro 4 *pixels* de cada vez, diminuindo efetivamente as dimensões espaciais do mapa de características de saída em um fator de 4 em comparação com a imagem de entrada. A VGG16 obteve uma redução na acurácia para 82,25% enquanto a ResNet-50 obteve seu melhor resultado com 80,05%.

Aplicar filtros maiores, como mencionado anteriormente, permite que cada neurônio capture informações de uma área maior da imagem de entrada, o que pode facilitar o reconhecimento de padrões maiores e mais complexos. Além disso, com filtros maiores, as CNNs podem aprender características mais complexas e abstratas a partir dos dados de entrada, o que pode resultar em uma representação e discriminação de características aprimoradas. Ao aplicar essa mudança na VGG16, observamos uma melhora com filtros maiores, sendo o 7x7 o mais eficaz, e o uso do kernel 9x9 resultou em uma acurácia superior em comparação com o modelo original. No entanto, ao aplicar filtros maiores na ResNet-50, observamos que o melhor resultado foi alcançado com um kernel 9x9, mas, de maneira geral, a acurácia foi inferior à do modelo original.

Como mencionado anteriormente, as vantagens de utilizar a função de ativação ReLU são amplamente reconhecidas devido à sua simplicidade e eficiência. No entanto, a ReLU não aborda diretamente a regularização de *Dropout*, que consiste em aleatoriamente anular algumas funções de ativação em determinados nós das camadas. Inspirada nessa ideia, foi desenvolvida a função de ativação Gelu, que busca combinar essas vantagens. Dado que a Gelu é amplamente utilizada no *swin transformer*, foi decidido aplicá-la nas arquiteturas estudadas. No entanto, foi observado que a VGG16 obteve seu melhor resultado com 86,15% porém com o aumento do tamanho da rede para 317,85MB e a ResNet-50 para 73,50% com o tamanho da rede de 89,77MB.

Levando em consideração que o *swin transformer* emprega poucas ativações, a simplificação dos modelos é buscada, o que implica em camadas mais lineares. Essa abor-

dagem não apenas reduz o custo computacional, mas também diminui a probabilidade de *overfitting*. Ao aplicar essa modificação, observou-se que a VGG16 teve uma melhora em relação ao modelo original, embora tenha perdido um pouco de acurácia em comparação ao modelo atual, que utiliza uma filtros 7x7 e função de ativação Gelu, resultando em uma acurácia de 83,55% e um tamanho da rede se mantave. Já para a ResNet-50, essa mudança resultou em uma acurácia de 72,55% e o tamanho da rede se manteve igual ao modelo que teve a melhor acurácia.

Conforme mencionado na metodologia, o *swin transformer* emprega poucas normalizações, o que também reduz o custo computacional ao evitar cálculos adicionais e proporciona uma maior flexibilidade ao modelo para generalizar para dados não observados. Com essa modificação, verificou-se que a VGG16 não apresentou melhora com a acurácia de 78,94% e a ResNet-50 com 67,4%. Isso se deve ao fato de que os modelos originalmente incluem funções de normalização, que facilitam a convergência rápida e reduzem o *overfitting*. Ao remover essas camadas, pode ocorrer uma baixa capacidade de aprendizado.

Considerando que o *swin transformer* adotam a LN em vez da BN, essa alteração foi aplicada nas redes selecionadas. Enquanto a BN normaliza as ativações dentro dos *batches*, calculando média e variância, a LN não depende do *batch* escolhido e aplica a média e variância em todas as camadas da rede. Com base nessa distinção, as redes foram atualizadas, substituindo o uso de BN por LN. Entretanto, observou-se que ambas as redes não foram capazes de aprender, o que sugere que, ao aplicar a normalização nas camadas em vez dos *batches*, a rede pode não ter capturado adequadamente os padrões necessários para cada *batch* selecionado, resultando na ausência de aprendizado.

Modelo	Modificações	Acurácia	Tamanho(MB)
VGG16	Original	83,30	68,41
	Aumento da profundidade	Não aprendeu	142,26
	Mudança do bloco inicial	82,25	58,28
	Filtro 5x5	83,90	168.20
	Filtro 7x7	85,45	317.85
	Filtro 9x9	85,40	517.37
	Filtro 7x7 + Gelu	86,15	317.85
	Filtro 7x7 + Gelu + poucas ativações	83,55	317.85
	Filtro 7x7 + Gelu + poucas normalizações	78,94	317.83
	Filtro 7x7 + Gelu + BN ->LN	Não aprendeu	317.85
Swin transformer	Original	78,75%	3,57

Tabela 3: Resultados dos experiementos com a VGG16 no conjunto de teste.

Modelo	Modificações	Acurácia	Tamanho(MB)
Resnet-50	Original	77,30	89.79
	Aumento da profundidade	76,15	101.52
	Mudança do bloco inicial	80,05	89.77
	Mudança do bloco inicial + Filtro 5x5	74,45	166.52
	Mudança do bloco inicial + Filtro 7x7	75,90	281.64
	Mudança do bloco inicial + Filtro 9x9	75,30	435.14
	Mudança do bloco inicial + Relu ->Gelu	73,50	89.77
	Mudança do bloco inicial + poucas ativações	72,55	89.77
	Mudança do bloco inicial + poucas normalizações	67,40	89.62
	Mudança do bloco inicial + BN ->LN	Não aprendeu	89.77
Swin transformer	Original	78,75%	3,57

Tabela 4: Resultados dos experimentos com a ResNet-50 no conjunto de teste.

5 CONCLUSÕES E TRABALHOS FUTUROS

Com o propósito de investigar as redes neurais utilizadas na classificação de imagens e analisar o impacto do *swin transformer* em comparação com as CNNs, este trabalho abordou conceitos importantes para o avanço dos estudos em redes neurais. Vale salientar que buscou-se identificar as inovações no *swin transformer* que poderiam influenciar positivamente as CNNs, evidenciando que, apesar da crescente popularidade das novas arquiteturas baseadas em *transformers*, as CNNs continuam desempenhando um papel significativo devido à sua simplicidade e capacidade de generalização elevada.

Com base nos experimentos realizados, podemos inferir que, embora o *swin transformer* seja uma arquitetura robusta e eficiente, as CNNs ainda desempenham um papel importante no campo de visão computacional. Isso se deve à facilidade de implementação, simplicidade de algumas redes e capacidade de generalização elevada das CNNs para o problema de classificação de imagem. Ao analisar as modificações aplicadas, observamos um aumento significativo na quantidade de parâmetros para filtros de grande porte, enquanto para as demais alterações, a estabilidade foi mantida. Destaca-se que a VGG16 alcançou uma acurácia de 86,16% e a ResNet-50, 80,05%, enquanto o *swin transformer* obteve 78,75%. Esses resultados reforçam a importância contínua das CNNs e sua capacidade de competir efetivamente com arquiteturas mais recentes, mesmo em um cenário em rápida evolução de técnicas de aprendizado profundo.

De acordo com (LIU et al., 2022), nos anos 2020, os *vision transformers*, especialmente os hierárquicos como os *swin transformers*, emergiram como alternativas cada vez mais populares em problemas de visão computacional, superando gradualmente as CNNs como a escolha preferida. Embora os *vision transformers* sejam reconhecidos por sua eficiência e escalabilidade, as CNNs ainda mantêm sua relevância e competitividade. Este fenômeno é evidenciado pelo fato de que as CNNs modificadas, como demonstrado no experimento realizado, podem competir de maneira eficaz com esses novos modelos do estado da arte, alcançando resultados superiores aos *vision transformers* em determinadas circunstâncias. Esses achados reforçam a importância de considerar não apenas a novidade das arquiteturas de redes neurais, mas também sua adaptabilidade e capacidade de se ajustar às demandas específicas de cada problema.

É importante destacar que além das inovações utilizadas no *swin transformer*, existem outras técnicas que podem ser empregadas, como o uso da convolução *depthwise*, convolução fatorada e inversão de *bottleneck*, entre outras. Adicionalmente, devido às limitações de espaço e memória da máquina, é possível treinar esses modelos por mais épocas utilizando um maior tamanho de *batch*. Ao aplicar essas novas estratégias e prolongar o tempo de treinamento, é possível alcançar resultados mais promissores e refinados.

Em síntese, este estudo contribui para o entendimento do cenário atual das arqui-

teturas de redes neurais aplicadas à classificação de imagens. Ao comparar as tradicionais redes convolucionais com as emergentes *vision transformers*, evidenciamos a importância de considerar não apenas a eficácia de uma arquitetura específica, mas também seu contexto de aplicação e as possíveis modificações que podem ser feitas para melhorar seu desempenho.

Embora os *vision transformers* tenham ganhado destaque recentemente, este trabalho mostra que as CNNs ainda são uma opção viável e competitiva, especialmente quando adaptadas e aprimoradas para atender às necessidades específicas do problema em questão. O avanço contínuo na pesquisa de arquiteturas de redes neurais promete trazer ainda mais inovações e aprimoramentos para o campo da visão computacional, permitindo o desenvolvimento de sistemas cada vez mais eficientes e precisos para uma ampla gama de aplicações práticas.

REFERÊNCIAS

ANIS, S.; LAI, K. W.; CHUAH, J. H.; ALI, M.; MOHAFEZ, H.; HADIZADEH, M.; DING, Y.; CHAO, O. An overview of deep learning approaches in chest radiograph. **IEEE Access**, v. 8, n. 1, p. 182347 – 182354, 10 2020.

BRESSEM, K. K.; ADAMS, L. C.; ERXLEBEN, C.; HAMM, B.; NIEHUES, S. M.; VAHLDIEK, J. L. Comparing different deep learning architectures for classification of chest radiographs. **Scientific reports**, Nature Publishing Group UK London, v. 10, n. 1, p. 13590, 2020.

DARAPANENI, N.; KRISHNAMURTHY, B.; PADURI, A. R. Convolution neural networks: A comparative study for image classification. In: IEEE. **2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)**. [S.l.], 2020. p. 327–332.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019.

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.

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.

HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDREETTO, M.; ADAM, H. **MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications**. 2017.

HUANG, G.; LIU, Z.; MAATEN, L. van der; WEINBERGER, K. Q. **Densely Connected Convolutional Networks**. 2018.

IANDOLA, F. N.; HAN, S.; MOSKEWICZ, M. W.; ASHRAF, K.; DALLY, W. J.; KEUTZER, K. **SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and ¡0.5MB model size**. 2016.

IRVIN, J.; RAJPURKAR, P.; KO, M.; YU, Y.; CIUREA-ILCUS, S.; CHUTE, C.; MARKLUND, H.; HAGHGOO, B.; BALL, R.; SHPANSKAYA, K.; SEEKINS, J.; MONG, D. A.; HALABI, S. S.; SANDBERG, J. K.; JONES, R.; LARSON, D. B.; LANGLOTZ, C. P.; PATEL, B. N.; LUNGREN, M. P.; NG, A. Y. **CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison**. 2019.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Communications of the ACM**, AcM New York, NY, USA, v. 60, n. 6, p. 84–90, 2017.

LEE, M. et al. Mathematical analysis and performance evaluation of the gelu activation function in deep learning. **Journal of Mathematics**, Hindawi, v. 2023, 2023.

LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S.; GUO, B. **Swin Transformer: Hierarchical Vision Transformer using Shifted Windows**. 2021.

LIU, Z.; MAO, H.; WU, C.-Y.; FEICHTENHOFER, C.; DARRELL, T.; XIE, S. A convnet for the 2020s. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2022. p. 11976–11986.

MAURÍCIO, J.; DOMINGUES, I.; BERNARDINO, J. Comparing vision transformers and convolutional neural networks for image classification: A literature review. **Applied Sciences**, MDPI, v. 13, n. 9, p. 5521, 2023.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. et al. Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, 2019.

RAJADANURAKS, P.; SURANUNTCHAI, S.; PECHPRASARN, S.; TREEBUPACHATSAKUL, T. Performance comparison for different neural network architectures for chest x-ray image classification. In: **2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST)**. [S.l.: s.n.], 2021. p. 49–53.

SIMONYAN, K.; ZISSERMAN, A. **Very Deep Convolutional Networks for Large-Scale Image Recognition**. 2015.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. **Going Deeper with Convolutions**. 2014.

TAN, M.; LE, Q. V. **EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks**. 2020.

WANG, W.; YANG, Y.; WANG, X.; WANG, W.; LI, J. Development of convolutional neural network and its application in image classification: a survey. **Optical Engineering**, Society of Photo-Optical Instrumentation Engineers, v. 58, n. 4, p. 040901–040901, 2019.

ANEXO A – ANEXOS E APÊNDICES 1

Anexos e apêndices são materiais adicionais, utilizados para complementar o texto, acrescentados ao final do trabalho, com a finalidade de esclarecimento ou de comprovação.

Apêndices são elaborados pelo autor e visam complementar uma argumentação. Os Anexos não são elaborados diretamente pelo autor e servem de fundamentação teórica, comprovação e ilustração (ex. mapas, leis, estatutos entre outros). Os apêndices devem aparecer antes dos anexos.