Universidade Estadual de Campinas
Instituto de Computação

Leodécio Braz da Silva Segundo

Classificação de Imagens Musculoesqueléticas Utilizando Aprendizado de Máquina Profundo

CAMPINAS
2021

Leodécio Braz da Silva Segundo

Classificação de Imagens Musculoesqueléticas Utilizando Aprendizado de Máquina Profundo

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Hélio Pedrini
Coorientador: Prof. Dr. Zanoni Dias

Este exemplar corresponde à versão final da Dissertação defendida por Leodécio Braz da Silva Segundo e orientada pelo Prof. Dr. Hélio Pedrini.

CAMPINAS
2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Se39c

Segundo, Leodécio Braz da Silva, 1997-
Classificação de imagens musculoesqueléticas utilizando aprendizado de máquina profundo / Leodécio Braz da Silva Segundo. – Campinas, SP : [s.n.], 2021.

Orientador: Hélio Pedrini.
Coorientador: Zanoni Dias.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Aprendizado profundo. 2. Classificação de imagem. 3. Redes neurais convolucionais. I. Pedrini, Hélio, 1963-. II. Dias, Zanoni, 1975-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Classification of musculoskeletal images using deep machine learning
Palavras-chave em inglês:
Deep learning
Image classification
Convolutional neural networks
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Hélio Pedrini [Orientador]
José Ramon Trindade Pires
Jacques Wainer
Data de defesa: 05-04-2021
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-9326-4122
- Currículo Lattes do autor: http://lattes.cnpq.br/0865120534739890

Universidade Estadual de Campinas
Instituto de Computação

Leodécio Braz da Silva Segundo

Classificação de Imagens Musculoesqueléticas Utilizando Aprendizado de Máquina Profundo

Banca Examinadora:

- Prof. Dr. Hélio Pedrini
  Instituto de Computação – Unicamp
- Dr. José Ramon Trindade Pires
  NeuralMind
- Prof. Dr. Jacques Wainer
  Instituto de Computação – Unicamp

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 05 de abril de 2021

Dedicatória

À minha mãe, por todo seu cuidado e dedicação durante toda a minha caminhada, meu pai por todos os seus esforços e por ser a figura que me motiva a ser, e minha irmã por ser minha fonte de inspiração em todos os momentos.

Nenhum esforço faz sentido, se você não acredita em si mesmo

(Maito Gai)

Agradecimentos

- Agradeço, primeiramente, a Deus e à minha família.
- Agradeço aos meus orientadores, professores Hélio Pedrini e Zanoni Dias por todo o aprendizado, tanto pessoal quanto acadêmico, que me proporcionaram nestes dois anos de mestrado.
- Agradeço a todos que fazem parte da Revolta da Cajuína, por toda a ajuda e o momentos compartilhados nestes dois últimos anos.
- Agradeço aos meus velhos amigos Alison, Erisson, Erly, Emy, Evandro, Jefferson, Matheus e Wendel, por todas as histórias vividas.
- Também agradeço à galera do FUT: Léo, o jogador, Yuri e Mateus, por todas as conversas descontraídas e as precisas análises sobre futebol. Vasco.
- Agradeço à Beatriz por todo o apoio que me dá, por todo o carinho e pela ajuda, por sua amizade que é muito especial para mim. Sem você, essa caminhada teria sido pior.
- Agradeço à Universidade Estadual de Campinas, ao Instituto de Computação e aos membros do Laboratório de Informática Visual (LIV), que me ajudaram em diversos momentos ao longo dessa caminhada, em especial, meus companheiros de orientação Gabriel, Vinícius, Tiago, Daiane e Marianna.
- Agradeço o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) - Código de Financiamento 001, pela bolsa de pesquisa concedida por meio do processo 88887.335985/2019-00, entre março de 2019 e julho de 2019. Agradeço também ao Centro de Estudos de Petróleo (CEPETRO) e à Fundação de Desenvolvimento da UNICAMP – FUNCAMP, pela bolsa de pesquisa concedida por meio do processo n° 2018/00607-6 entre outubro de 2019 e janeiro de 2021.
- A todos que fizeram parte da minha vida e que, de alguma maneira, contribuíram e me ajudaram a chegar aqui, o meu muito obrigado.

Resumo

Distúrbios musculoesqueléticos, muitas vezes caracterizados por dores agudas ou crônicas, limitam a capacidade de um indivíduo em realizar atividades de rotina. Os distúrbios ocorrem com frequência e determinar se um estudo radiográfico apresenta a estrutura óssea "normal" ou "anormal" é uma tarefa crítica na radiologia.

Sistemas de auxílio ao diagnóstico possuem o potencial de aliviar a sobrecarga de médicos e diminuir a ocorrência de diagnósticos incorretos, entretanto, formular métodos de sucesso para a classificação automática apresenta vários desafios. Técnicas de aprendizado de máquina profundo têm sido cada vez mais aplicadas nesse tipo de problemas, com o objetivo de fornecer resultados mais precisos nas tarefas de classificação, tanto em imagens quanto em textos médicos, assim como na geração automática de relatórios médicos.

Neste estudo, avaliamos o uso de diferentes redes neurais convolucionais para detectar anormalidades musculoesqueléticas por meio de exames de raios-X e averiguar o impacto de diversas técnicas aplicadas de forma incremental ao modelo. Baseado nos experimentos realizados, os melhores resultados foram obtidos por meio de uma estratégia de *ensemble*, em que empregamos uma máquina de vetores de suporte (SVM) para combinar as saídas dos diferentes modelos utilizados.

Nesta pesquisa, investigamos a influência da fusão de características de diferentes modalidades visando melhorar a detecção de anormalidades musculoesqueléticas. Os resultados obtidos indicam que utilizar uma abordagem multimodal apresenta uma melhora nos resultados quando comparado com os resultados de classificação utilizando imagens e textos individualmente.

Duas bases de dados médicos disponíveis publicamente foram utilizadas nos nossos experimentos para validar as arquiteturas propostas. Diversas métricas quantitativas foram calculadas para demonstrar a eficácia da classificação de imagens musculoesqueléticas.

Abstract

Musculoskeletal disorders, often characterized by acute or chronic pain, limit an individual's ability to perform routine activities. Musculoskeletal disorders occur frequently and determining whether a radiographic study presents the "normal" or "abnormal" bone structure is a critical task in radiology.

Diagnostic aid systems have the potential to relieve physicians' overload and reduce the occurrence of incorrect diagnoses, however, formulating successful methods for automatic classification presents several challenges. Deep learning techniques have been increasingly applied to these types of problems, with the purpose of providing more accurate results in the classification tasks, both in images and medical texts, as well as in the automatic generation of medical reports.

In this study, we evaluate the use of different convolutional neural networks to detect musculoskeletal abnormalities by means of X-ray examinations and investigate the impact of various techniques applied incrementally to the model. Based on the experiments carried out, the best results were obtained through an ensemble strategy, in which we employ a support vector machine (SVM) to combine the outputs of the different models used.

In this research, we investigated the influence of the fusion of characteristics of different modalities in order to improve the detection of musculoskeletal abnormalities. The results obtained indicate that using a multimodal approach presents an improvement in the results when compared to the classification results using images and texts individually.

Two publicly available medical databases were used in our experiments to validate the proposed architectures. Several quantitative metrics were calculated to demonstrate the effectiveness of the musculoskeletal image classification.

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviações

AUC	Area Under the Curve
AVG-MAX	Média-Máxima
BERT	Bidirectional Encoder Representations from Transformers
CAD	Computer-Aided Diagnosis
CAM	Class Activation Mapping
CNN	Convolutional Neural Network
CT	Computed Tomography
CUI	Concept Unique Identifiers
DCNN	Deep Convolutional Neural Network
DenseNet	Dense Convolutional Network
FN	Falsos Negativos
FP	Falsos Positivos
GMU	Gated Multimodal Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
ISIC	International Skin Imaging Collaboration
KNN	K-Nearest Neighbors
LIV	Laboratório de Informática Visual
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
MURA	Musculoskeletal Radiographs
NLP	Natural Language Processing
NN	Neural Network
PET	Positron Emission Tomography
PNG	Portable Network Graphics
RF	Random Forests
RGB	Red-Green-Blue
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
ROCO	Radiology Objects in COntext
SVHN	Street View House Numbers
SVM	Support Vector Machine
TFP	Taxa de Falsos Positivos
TVP	Taxa de Verdadeiros Positivos
VGG	Visual Geometry Group
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos

Sumário

Capítulo 1

Introdução

Neste capítulo, descrevemos o problema a ser investigado nesta dissertação, os objetivos, as contribuições, as questões de pesquisa que serviram para guiar este trabalho, bem como a organização do restante do texto.

1.1 Descrição do Problema

Distúrbios musculoesqueléticos, muitas vezes caracterizados por dores agudas ou crônicas, limitam significativamente a mobilidade, a aptidão e a capacidade funcional de um indivíduo, reduzindo, por exemplo, sua capacidade de desenvolver atividades de rotina, trabalho ou lazer. Lesões musculoesqueléticas ocorrem com frequência e, por isso, representam um problema relevante, pois determinar se um estudo radiográfico apresenta a estrutura óssea "normal" ou "anormal" é uma tarefa crítica na área de radiologia, dado o fato de que, por exemplo, ao interpretar um estudo como "normal", elimina-se a necessidade de os pacientes se submeterem a procedimentos ou diagnósticos adicionais, o que pode representar um grave risco em caso de erro [75].

Imagens de raios-X são um dos exames radiológicos mais acessíveis e comumente utilizados para detectar e localizar anormalidades em estudos radiográficos. O processo de interpretar exames em imagens é uma tarefa normalmente complexa. Os profissionais médicos são responsáveis por examinar e interpretar uma grande quantidade de imagens diariamente para realizar diagnósticos [51]. Métodos automáticos podem reduzir erros médicos e beneficiar departamentos médicos, reduzindo o custo por exame [51, 56].

Para facilitar o processo de diagnóstico e interpretação automática de imagens biomédicas, muitos sistemas de auxílio ao diagnóstico (*Computer Aided Diagnosis* - CAD), que fazem uso de técnicas de processamento de imagens e visão computacional combinadas com os recentes avanços de aprendizado de máquina profundo, têm sido desenvolvidos para esses procedimentos médicos [50, 67].

Há várias tarefas aplicadas a imagens médicas que podem auxiliar médicos e especialistas da área de saúde durante os exames, tais como detecção e segmentação de lesões [61]. Para a tarefa de classificação, diversas metodologias têm sido exploradas e alguns dos trabalhos de maior sucesso incluem classificadores como K-Vizinho mais Próximo (*K-Nearest Neighbors* - KNN) [57], Máquinas de Vetores de Suporte (*Support Vector Machines* -

SVM) [22] e Florestas Aleatórias (*Random Forests* - RF) [62]. Muitos estudos também utilizam Redes Neurais (*Neural Networks* - NN) [64,94], combinando uma ampla gama de técnicas, como aprendizado profundo, processamento de imagens e visão computacional, para interpretar automaticamente imagens radiográficas [25, 50].

A classificação de raios-X encontra aplicação em uma variedade de distúrbios médicos. Ela é normalmente utilizada em estudos que envolvem a região do tórax [7, 101] para classificação de câncer de mama [27, 100] ou câncer de pulmão [70], além de estudos em diferentes regiões do corpo, como detecção de osteoartrite de joelho [49] ou anormalidades musculoesqueléticas em geral [71, 75].

Métodos de aprendizado profundo são o estado da arte em tarefas de classificação na área médica. Redes Neurais Convolucionais Profundas (*Deep Convolutional Neural Networks* - DCNN) são amplamente utilizadas principalmente em domínios de imagem [16, 75] e, recentemente, muitos métodos de Processamento de Linguagem Natural (*Natural Language Processing* - NLP) foram desenvolvidos para extrair rótulos a partir de textos de relatórios médicos [3, 73].

Apesar da importância dessas tarefas, muitas vezes elas são consideradas complexas, pois se faz necessária uma grande quantidade de dados e de recursos que nem sempre estão disponíveis ou facilmente acessíveis. Além desse fato, o treinamento de modelos de aprendizado profundo tipicamente requer alto poder computacional, que muitas vezes pode ser limitado.

1.2 Objetivos

Os principais objetivos deste trabalho incluem a avaliação de técnicas e processos que melhorem a detecção de anormalidades em imagens musculosqueléticas, além da proposição e implementação de um método que combina informações multimodais de imagens e textos para realizar uma classificação.

Para alcançar esses objetivos gerais, os seguintes tópicos foram definidos:

- Avaliação de desempenho da classificação utilizando modelos clássicos já existentes na literatura.
- Definição de técnicas e dos processos a serem aplicados nas imagens.
- Avaliação de diferentes abordagens de *ensemble* dos modelos.
- Geração de textos artificiais a partir das imagens.
- Avaliação da fusão entre métodos multimodais.

1.3 Contribuições

As principais contribuições que foram derivadas a partir do desenvolvimento deste trabalho de pesquisa incluem:

- Desenvolvimento e avaliação de diferentes métodos de *ensemble*.

- Avaliação de um método multimodal no problema de classificação de imagens musculoesqueléticas.

1.4 Questões de Pesquisa

Para guiar o desenvolvimento deste trabalho de pesquisa, além de auxiliar o refinamento dos resultados para os problemas investigados, algumas questões de pesquisa (QP) foram formuladas:

- **QP1:** Qual é o impacto de diferentes formas de *ensemble* no processo de classificação?
- **QP2:** A utilização de uma abordagem multimodal, no contexto de dados médicos, pode produzir resultados melhores que os resultados utilizando apenas imagens e textos?
- **QP3:** É possível gerar, com boa qualidade, dados textuais de legendas artificiais para imagens de radiografias musculoesqueléticas?

1.5 Publicações

Os seguintes artigos foram publicados a partir do desenvolvimento deste trabalhos de pesquisa:

- L. Braz, V. Teixeira, H. Pedrini, Z. Dias. Image-Text Integration Using a Multimodal Fusion Network Module for Movie Genre Classification. 11th International Conference on Pattern Recognition Systems (ICPRS). Curicó, Chile, March 17-19, 2021.
- L. Braz, V. Teixeira, H. Pedrini, Z. Dias. ImTeNet: Image-Text Classification Network for Abnormality Detection and Automatic Reporting on Musculoskeletal Radiographs. Brazilian Symposium on Bioinformatics (BSB). Online Meeting, pp. 150-161, November 23-27, 2020.
- G. Sato, L. Braz, Z. Dias. Classification of Musculoskeletal Abnormalities with Convolutional Neural Networks. Brazilian Symposium on Bioinformatics (BSB). Online Meeting, pp. 69-80, November 23-27, 2020.
- V. Teixeira, L. Braz, H. Pedrini, Z. Dias. DuaLAnet: Dual Lesion Attention Network for Thoracic Disease Classification in Chest X-Rays. 27th International Conference on Systems, Signals and Image Processing (IWSSIP). Rio de Janeiro-RJ, Brazil, pp. 69-74, June 3-5, 2020.

1.6 Organização do Texto

O restante do texto está organizado da seguinte forma. No Capítulo 2, revisamos alguns conceitos relevantes relacionados aos problemas investigados. No Capítulo 3, apresentamos alguns trabalhos relacionados à classificação de imagens e textos médicos e às abordagens multimodais. No Capítulo 4, descrevemos as bases de dados e as métricas de avaliação utilizadas nos experimentos. No Capítulo 5, descrevemos o método proposto em que tratamos o problema de classificação de anormalidades musculoesqueléticas utilizando uma abordagem multimodal com imagens e textos médicos. No Capítulo 6, apresentamos e discutimos os resultados obtidos após uma série de experimentações com redes neurais convolucionais combinadas com diversas técnicas para detectar anormalidades em imagens de raio-X. Os resultados obtidos a partir da aplicação do método proposto são também reportados e analisados. No Capítulo 7, apresentamos as conclusões e possíveis tópicos e abordagens para trabalhos futuros.

Capítulo 2

Conceitos Relacionados

Neste capítulo, apresentamos conceitos importantes que serão utilizados ao longo do trabalho, de modo a facilitar a compreensão do método proposto. Estes conceitos estão associados desde a temas mais gerais, tais como Radiologia, Visão Computacional e Processamento de Linguagem Natural, até temas mais específicos, tais como Sistemas de Auxílio ao Diagnóstico, Redes Neurais, *Transformers*, Aumentação de Dados e Transferência de Aprendizado.

2.1 Radiologia

A radiologia é um ramo da medicina que utiliza imagens dos órgãos do corpo para detecção, interpretação e diagnóstico de doenças [66]. Estas imagens são geradas por meio do uso de radiações, permitindo a visualização de ossos, órgãos ou estruturas do corpo [65].

A radiologia é vital para o diagnóstico de uma variedade de doenças, como câncer e pneumonia, sendo utilizada em diversos procedimentos médicos, incluindo cirurgia, pediatria, obstetrícia, análise de traumas, medicina emergencial, entre outros. As imagens radiológicas incluem a radiografia, a tomografia computadorizada, a ressonância magnética, a ultrassonografia e a mamografia [17].

Uma das principais vantagens das radiografias é o seu baixo custo de produção [35], sendo facilmente acessíveis para exames. Imagens radiográficas são produzidas por meio do uso de radiação ionizante [17]. Várias bases de dados disponíveis publicamente, tais como MURA [75] e CheXpert [39], são formadas por radiografias.

2.2 Sistemas de Auxílio ao Diagnóstico

Sistemas de Auxílio ao Diagnóstico (*Computer-Aided Diagnosis* - CAD) são mecanismos que dão assistência a médicos na interpretação de dados médicos. Sistemas CAD combinam técnicas de processamento de imagens, processamento de linguagem natural, aprendizado de máquina e reconhecimento de padrões, para auxiliar especialistas médicos no processo de diagnóstico.

Sistemas CAD são principalmente utilizados no contexto de imagens médicas, abrangendo uma variedade de modalidades de imagens incluindo radiografia, tomografia com-

putadorizada, ressonância magnética e ultrassonografia, em tarefas como detecção, classificação ou segmentação de lesões, tumores e fraturas [67].

Diversos fatores podem afetar e comprometer os laudos e diagnósticos realizados por um ser humano, como a presença de ruído nas imagens dos exames, a similaridade entre observações normais e anormais (por exemplo, tumores benignos e malignos) e o estado físico e mental dos radiologistas [77].

Devido ao grande volume de dados médicos gerados de exames em pacientes, a tarefa de diagnóstico pelos especialistas é complexa e suscetível a erros. Sistemas CAD possuem, como finalidade, melhorar a precisão dos diagnósticos, buscando detectar e classificar anormalidades ou doenças presentes nos exames, podendo ser utilizados como uma "segunda opinião", os quais têm demonstrado ser úteis ao reduzir erros [77].

2.3 Visão Computacional

Visão computacional é um ramo da Ciência da Computação que visa produzir descrições precisas de objetos físicos a partir de imagens [8]. A visão computacional busca descrever o mundo real, reconstruindo suas propriedades, tais como forma, iluminação e cores, de modo a fornecer às máquinas a habilidade de reconhecer e interpretar imagens de maneira tão eficiente quanto os seres humanos [88, 92].

A área de visão computacional possui uma ampla variedade de aplicações, por exemplo, navegação de robôs, inspeção industrial, recuperação de imagens, análise de imagens médicas, reconhecimento e detecção de faces, segurança e vigilância, entre outras.

A visão computacional é um dos campos mais impactados com o advento das técnicas de aprendizado profundo [89]. Com a necessidade de classificar e interpretar imagens automaticamente, grandes volumes de dados têm sido criados e disponibilizados. O surgimento de estruturas, como camadas convolucionais em redes neurais, também tem contribuído para o avanço da área.

Muitos modelos de aprendizado de máquina profundo para visão computacional têm sido utilizados para o reconhecimento ou a detecção de objetos e formas, transcrição de símbolos presentes em uma imagem, segmentação e classificação de objetos [29].

2.4 Processamento de Linguagem Natural

Processamento de Linguagem Natural é um campo da Inteligência Artificial que consiste na manipulação de elementos da linguagem natural, como fala e texto. A partir de meios para lidar com esse tipo de informação, os computadores podem extrair informações relevantes de textos como, por exemplo, na tarefa de sumarização, em que, ao receber um texto completo como entrada, o modelo pode construir um resumo a partir do contexto passado como entrada [45].

Para Ruder et al. [79], a área de processamento de linguagem natural tem como objetivo ensinar os computadores a ler, compreender e inferir significado das línguas naturais. Muito mais do que uma simples ferramenta de manipulação simbólica, um modelo de

inteligência artificial pode permitir uma compreensão do contexto a partir de um modelo de linguagem.

O processamento de linguagem natural utiliza abordagens que se baseiam em um conjunto de teorias e tecnologias, aplicando abordagens computacionais para analisar e representar elementos linguísticos, com o propósito de alcançar uma capacidade de processamento de linguagem semelhante ao ser humano para desempenhar inúmeras tarefas ou aplicações [58].

Os fundamentos de processamento de linguagem natural estão ligados a várias disciplinas da Ciência da Computação, Linguística, Matemática, Psicologia, entre outras áreas [45]. Suas aplicações englobam uma série de domínios de estudo, tais como sumarização de texto, recuperação de informações, reconhecimento de voz e tradução automática. Esta última consiste no modelo receber uma frase (sequência de palavras) em uma determinada língua natural e emitir como resposta uma frase equivalente em uma outra língua humana [29, 45].

Os métodos de processamento de linguagem natural são baseados em modelos de linguagem, responsáveis por definir uma distribuição de probabilidade sobre as sequências de palavras ou caracteres em dados de linguagem natural [29].

2.5 Classificação

Modelos baseados em redes neurais, em especial, redes neurais profundas, têm apresentado ótimos resultados em tarefas de classificação em geral e representam atualmente o estado da arte nesse tipo de atividade [34, 91], sendo utilizados em inúmeros problemas envolvendo não apenas imagens, tais como classificação [53], recuperação [83], reconhecimento e detecção de faces e objetos [11, 86], mas também textos para classificação automática [37, 105], categorização [43, 44], tradução [96], entre outras aplicações [59, 84].

A tarefa de classificação é utilizada para predizer a qual classe (ou categoria) uma instância de dados pertence. Algoritmos de aprendizado de máquina para classificação são utilizados em problemas como filtragem de *spam* em e-mails, categorização de documentos, reconhecimento de imagens e reconhecimento de manuscritos. Um exemplo de problema de classificação comumente encontrado na literatura é o reconhecimento de objetos, em que a entrada é uma imagem e a saída geralmente é um valor numérico que identifica cada objeto presente na imagem [29].

No paradigma de aprendizado supervisionado, um modelo de classificação recebe como entrada um conjunto de dados rotulados, em que cada rótulo representa a classe à qual o dado pertence. O aprendizado supervisionado consiste em, a partir de dados de treinamento, associar uma informação de "resposta" como um rótulo ou classe a um conjunto de amostras. Assim, o termo supervisionado origina-se da ideia de que um rótulo y é fornecido por um instrutor ou professor que ensina o modelo de aprendizado de máquina o que deve ser feito [29]. Se esses rótulos forem textuais, podem ser atribuídas transformações que os convertem para vetores numéricos. Em uma classificação binária, por exemplo, esses rótulos geralmente são representados pelos valores inteiros 0 e 1.

2.6 Redes Neurais

As Redes Neurais Artificiais, comumente denominadas Redes Neurais, tiveram a sua motivação na neurociência, onde surgiram inspiradas nos neurônios do cérebro humano, que são interconectados e transferem informações entre si por meio de impulsos elétricos. Dessa forma, uma rede neural é composta por unidades de processamento chamadas de "neurônios artificiais" ou perceptrons [78], que estão conectadas por meio de pesos sinápticos. No processo de aprendizado da rede, esses pesos são ajustados a fim de alcançar o objetivo desejado. Perceptron multicamadas (do inglês *Multilayer Perceptron* - MLP) é a denominação dada às redes neurais compostas por várias camadas de neurônios.

As camadas de uma MLP são denominadas (i) camada de entrada, em que são recebidos os valores dos dados de entrada e repassados para a próxima camada da rede; (ii) camadas intermediárias, também chamadas de camadas ocultas, que realizam o processamento dos valores de entrada; (iii) camada de saída, que é a última camada da rede, em que o resultado proveniente das camadas anteriores é apresentado. Na Figura 2.1, ilustramos uma rede neural contendo 4 neurônios na camada de entrada, 8 neurônios em cada uma das duas camadas intermediárias e 2 neurônios na camada de saída.

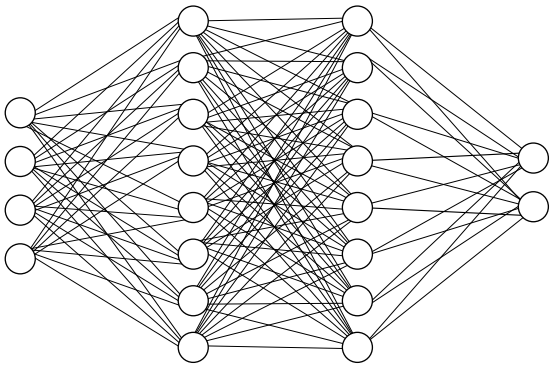


Figura 2.1: Ilustração de uma rede neural artificial contendo quatro camadas.

No processo de treinamento de uma rede neural, os pesos sinápticos são atualizados a fim de reduzir o erro na saída da rede baseado em uma função de perda. No algoritmo de retropropagação (*backpropagation*), o erro é calculado de acordo com os desvios entre o que é esperado como saída e o que é obtido pelos neurônios da camada de saída da rede. O valor de erro é propagado de volta à rede, atualizando os pesos das conexões entre cada neurônio, de modo a minimizar o erro resultante ao final [32].

2.6.1 Redes Neurais Convolucionais

Redes Neurais Convolucionais (do inglês *Convolutional Neural Networks* - CNN) [55] são um tipo de redes neurais profundas que empregam filtros convolucionais em suas camadas, que permitem extrair características de partes específicas de uma imagem, possibilitando também uma representação mais eficaz dos dados [29]. As CNNs representam atualmente o estado da arte em inúmeros problemas de Visão Computacional como classificação de imagens [53], reconhecimento e detecção de objetos [28, 76], segmentação de imagens [33], entre outros.

As camadas convolucionais consistem em diversos neurônios responsáveis por aplicar filtros em partes específicas da imagem de entrada. Cada neurônio está conectado a um conjunto de neurônios da camada anterior e, para cada conexão, é atribuído um peso, chamado de peso sináptico. Os pesos da entrada de cada neurônio são combinados entre si e passados para a camada seguinte [95]. Estes pesos atribuídos às conexões entre os neurônios desempenham o papel de um filtro convolucional aplicado no domínio espacial. Dessa forma, na etapa de treinamento da rede, esses filtros são ajustados para que sejam ativados na presença de características relevantes identificadas na entrada.

Enquanto em redes neurais tradicionais cada neurônio de uma camada está conectado a todas as unidades de neurônios da camada seguinte, chamadas de camadas totalmente conectadas (*Fully Connected Layer*), os neurônios das camadas convolucionais utilizam uma conectividade local, como ilustrado na Figura 2.2, em que cada neurônio na camada N está conectado a apenas alguns neurônios da camada $N+1$, ao invés de se conectarem a todos os neurônios da camada. Neurônios de uma mesma camada são agrupados e suas saídas formam mapas de características, como representado pela região em azul na Figura 2.2. Um mapa de características é produzido a partir da aplicação de operações de convolução da imagem de entrada com os filtros das camadas convolucionais.

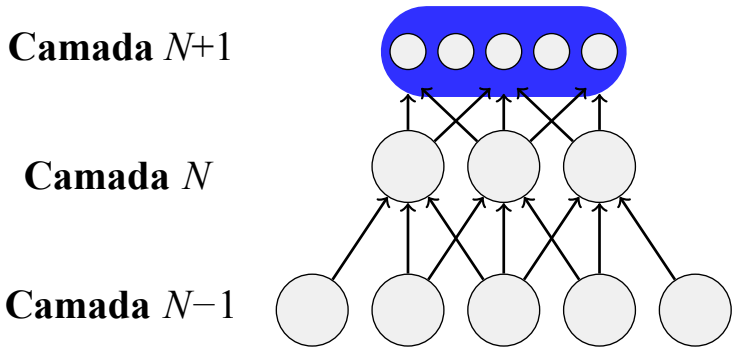


Figura 2.2: Exemplo da conectividade local de uma rede neural convolucional. A região em azul na camada $N+1$, destacada na imagem, representa um mapa de características, definido como o agrupamento de neurônios de uma mesma camada.

Além das camadas convolucionais, há dois outros tipos importantes de camadas que compõem uma CNN: camadas de *pooling* e camadas totalmente conectadas. A camada de *pooling*, visa reduzir o tamanho da entrada. A operação de *pooling* reduz a dimensionalidade da informação de entrada, mantendo informações relevantes. Há diferentes tipos de *pooling*, tais como a soma (*sum pooling*), a média (*average pooling*) e o valor máximo (*max pooling*). Este último consiste em substituir os valores de uma região pelo valor máximo da mesma. Na Figura 2.3, ilustramos o resultado da aplicação da operação de *max pooling*, em que é possível observar a redução no tamanho da entrada, causando uma redução na quantidade de processamento necessário nas próximas camadas da rede.

Algumas das redes (ou família de redes) neurais convolucionais bastante comuns na literatura que apresentam bons resultados em competições como *ImageNet* [81] são: VGG-Net [85], ResNet [34], DenseNet [38], EfficientNet [93] e Inception [14, 90].

A arquitetura de rede DenseNet (do inglês *Dense Convolutional Network*) conecta cada camada a todas as outras camadas subsequentes. Enquanto muitas redes convoluci-

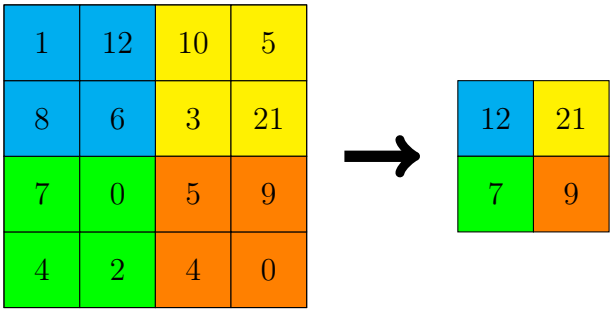


Figura 2.3: Aplicação da operação de *max pooling* em uma imagem 4×4 utilizando uma máscara 2×2.

onais tradicionais contendo N camadas possuem N conexões, as redes do tipo DenseNet possuem $\frac{N(N+1)}{2}$ conexões e, para cada camada, os mapas de características das camadas anteriores são usados como entrada e o mapa de característica gerado é utilizado como entrada para as camadas seguintes.

As arquiteturas de redes do tipo DenseNet apresentaram melhorias significativas comparado ao estado da arte em CIFAR-10 e CIFAR-100 [52], SVHN [69] e ImageNet [20], sendo necessário menos memória e poder computacional para atingir alto desempenho [38].

2.6.2 Redes Neurais Recorrentes

Redes Neurais Recorrentes (do inglês *Recurrent Neural Networks* - RNN) [80] são um tipo de rede neural artificial que, além das conexões entre as camadas, incluem também ligações entre neurônios adjacentes, introduzindo uma noção de tempo [60]. Esse tipo de rede neural foi projetado para processar sequências de valores, como texto, genomas, séries numéricas, entre outros dados.

Uma RNN é ilustrada na Figura 2.4. As redes recorrentes recebem como entrada não apenas um valor, mas também valores anteriores no tempo. Um laço permite que informações sejam passadas e armazenadas de uma etapa para a outra da rede.

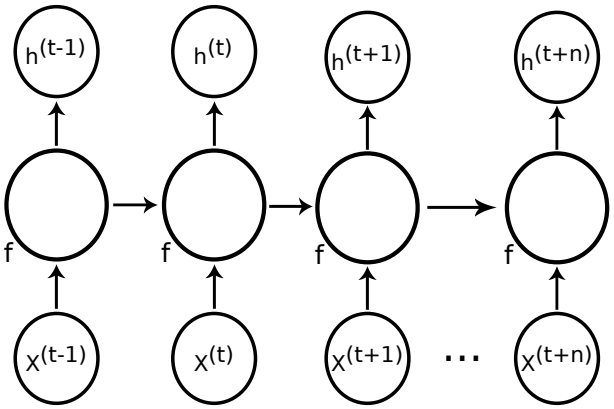


Figura 2.4: Ilustração de uma rede recorrente. A rede processa as informações da entrada x no tempo t incorporando-as ao estado que é transmitido ao longo do tempo.

A redes recorrentes possuem um estado interno (memória) para processar sequências de entradas de tamanhos variáveis. Assim, as informações sequenciais são preservadas

nos estados da rede à medida que elas avançam na rede e são processadas pelos outros estados.

Um dos principais problemas de redes recorrentes é o desaparecimento do gradiente, que ocorre quando o gradiente se torna cada vez menor e os parâmetros da rede não sofrem alterações significativas, dificultando assim o aprendizado de longas sequências de dados. Há também o problema de explosão do gradiente, que ocorre quando o valor do gradiente aumenta exponencialmente. Em ambos os casos, nenhum aprendizado real é feito.

As redes do tipo memória de curto e longo prazo (*Long Short-Term Memory* - LSTM) [36] surgiram para lidar com os problemas comuns que ocorrem com as redes neurais recorrentes. As redes LSTM possuem um mecanismo de memória capaz de "lembrar" valores em determinados intervalos, sendo adequadas para tarefas de classificação e predição baseadas em dados de séries temporais com intervalos de tempo e duração não definidos.

Este mecanismo de memória é útil, por exemplo, na tarefa de legendagem de imagens, em que as células de memórias da rede aprendem, em cada intervalo de tempo, quais informações de entrada já foram observadas até aquele momento [99].

2.7 Transformers

Os *transformers* [97] são uma arquitetura baseada em mecanismos de atenção, responsáveis por, dada uma sequencia de entrada, decidir em cada etapa quais partes da sequência são importantes e que devem ser focadas para realizar determinada tarefa.

Os *transformers* surgiram como uma alternativa às redes recorrentes, pois, assim como as redes neurais recorrentes, os *transformers* vêm sendo utilizados principalmente no campo de processamento de linguagem natural, projetados para lidar com dados sequenciais, tais como textos, em tarefas de classificação, tradução e sumarização.

Enquanto as rede neurais convolucionais permitem ser paralelizáveis, as redes recorrentes não possuem essa característica, entretanto, têm como vantagem o fato de serem capazes de lidar com entradas de tamanho variáveis, conseguindo manter informações de dependência entre uma sequência, característica que não está presente nas redes neurais convolucionais. Já os *transformers* buscam suprir essas deficiências, sendo paralelizáveis como as redes neurais convolucionais, bem como são capazes de lidar com entradas de tamanhos variáveis [1].

A ideia por trás dos *transformers* é lidar com as dependências de informações sequenciais da entrada por meio de mecanismos de atenção e funções de recorrência. A arquitetura de um *transformer* é composta de um codificador e um decodificador. O bloco do codificador possui uma camada de Atenção Multi-Cabeça (do inglês *Multi-Head Attention*) [97], seguida por uma camada simples de rede neural, ambas possuem conexões residuais e uma normalização. O bloco decodificador, é composto por duas camada de Atenção Multi-Cabeça, sendo uma delas uma camada Mascarada de Atenção Multi-Cabeça (do inglês,*Masked Multi-Head Attention*), e recebe uma representação contínua do codificador e gera uma única saída, ao passo que é retroalimentado com a saída anterior.

Uma outra estrutura importante presente na arquitetura dos *transformers* é o módulo

de codificação posicional, responsável por manter as informações sobre as posições dos *tokens* na sequência de entrada. Assim, para cada *token*, um valor é atribuído baseado na sua posição dentro da sequência, no tamanho do *embedding* e na profundidade do modelo.

Apesar do alto custo de tempo e de memória, os *transformers* vêm apresentando bons desempenhos em várias tarefas a que são submetidos e representam o estado da arte em diversos problemas de processamento de linguagem natural [13, 18, 21].

2.7.1 BERT

O BERT (*Bidirectional Encoder Representations from Transformers*) [21] é uma arquitetura que utiliza *transformers*, em especial, utiliza apenas o mecanismo codificador dos *transformers* para gerar um modelo de linguagem. O termo bidirecional é relativo à arquitetura do BERT, que utiliza modelos de linguagem mascarada para permitir que este seja pré-treinado utilizando representações bidirecionais, ou seja, é capaz de ler toda a sequência de palavras de uma vez, ao invés de ler a entrada de texto de maneira sequencial – tanto da esquerda para a direita quanto da direita para a esquerda. O BERT recebe como entrada uma soma de três *embeddings*: o *embedding* dos *tokens* da sentença, o de segmentação e outro de codificação posicional. O *token* `[CLS]` é o *token* especial utilizado em tarefas de classificação, enquanto o *token* `[SEP]` é utilizado para indicar uma separação de sentenças, por exemplo, para tarefas do tipo *Question-Answering* [1]. A Figura 2.5 ilustra o formato de entrada do modelo BERT.

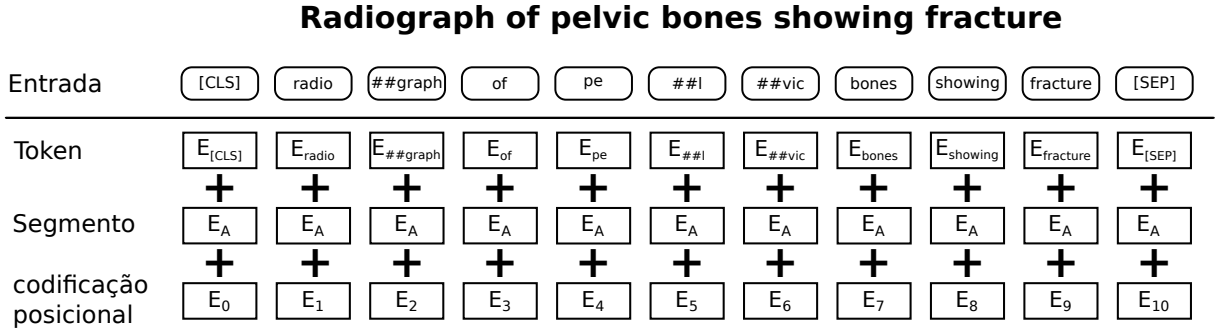


Figura 2.5: Ilustração do formato da entrada do modelo BERT [21]. Acima, em negrito, o texto original a ser formatado. A entrada é formada pela concatenação dos *embeddings* (representados pelos blocos E) dos *tokens*, da segmentação e do codificador posicional.

O modelo BERT é pré-treinado de maneira não-supervisionada, por meio de duas tarefas: modelo de linguagem mascarada (do inglês *Masked Language Model*), que consiste em predizer os *tokens* que estão mascarados com o *token* especial `[MASK]`; e predição de próxima sentença (do inglês *Next Sentence Prediction*), que consiste em uma classificação binária, para verificar se uma sentença é continuação da sentença passada como entrada para o modelo.

O BERT possui como proposta poder ser utilizado como um modelo base já treinado, sendo apenas necessária a realização de ajuste fino para tarefas específicas, gerando modelos que representam o estado da arte em diversos problemas de processamento de linguagem natural e impulsionando seu uso na literatura.

O BERT apresenta duas versões: (i) uma base (BERT-base) com 12 camadas de atenção de 12 cabeças e uma camada oculta com 768 neurônios, totalizando $110M$ de parâmetros e (ii) uma profunda (BERT-large) com 24 camadas de atenção de 16 cabeças e uma camada oculta de 1024, totalizando $340M$ de parâmetros. Neste trabalho de pesquisa, utilizamos o BERT em sua versão base.

2.8 Aumentação de Dados

Aumentação de Dados (*Data Augmentation*) é uma técnica, ou um conjunto de técnicas, para gerar novas amostras a partir de outras presentes no conjunto de dados, a fim de elevar a generalidade do modelo. Esta técnica visa melhorar a eficácia de um modelo na etapa de treinamento dos dados, pois uma das grandes preocupações durante o processo de aprendizagem é o problema de sobreajuste (*overfitting*).

O sobreajuste ocorre tipicamente quando o conjunto de dados utilizado é pequeno ou não representativo o suficiente e consiste no fato de a rede memorizar os dados de treinamento, aprendendo padrões específicos das amostras do conjunto. Dessa forma, quando a rede é submetida a um novo conjunto, ela não é capaz de generalizar satisfatoriamente seu desempenho, comprometendo a tarefa de classificação.

Há várias estratégias possíveis para realizar a aumentação de dados. Métodos comuns aplicam combinações de transformações geométricas e radiométricas da área de processamento de imagens, tais como translação, rotação, perspectiva, cisalhamento, adição ou remoção de ruído, borramento, entre outras operações.

Na Figura 2.6, apresentamos o resultado do processo de aumentação de dados, em que foram geradas 4 novas imagens a partir de uma amostra original (Figura 2.6a) por meio da aplicação de operações como rotações, translações e cisalhamentos em determinados ângulos. O principal objetivo é fazer com que, na etapa de treinamento, o modelo receba imagens com maior variabilidade, favorecendo a generalização dos resultados pela rede.

2.9 Transferência de Aprendizado

A técnica de Transferência de Aprendizado (do inglês *Transfer Learning*) é utilizada com o objetivo de reusar parte do "aprendizado" já adquirido por um modelo, aproveitando-o com o intuito de resolver novas tarefas, reduzindo assim o tempo necessário para treinar um novo modelo de rede neural profunda [15]. Esta técnica visa reutilizar informações aprendidas por um modelo anterior, em que essas informações são consideradas adequadas para uma nova tarefa. Isto traz benefícios no sentido de que pode suprir a falta de dados na etapa de treinamento dos modelos [67].

Como os modelos de aprendizado profundo normalmente requerem uma grande quantidade de dados para serem treinados e nem sempre essa quantidade de dados se encontra disponível ou acessível, a realização desta tarefa torna-se complexa. Dessa forma, uma prática comum é treinar uma rede convolucional com uma grande base de dados pública, tal como a ImageNet [81], que contém 1,2 milhões de imagens divididas em 1000 classes, e utilizar essa rede como uma inicialização dos pesos ou extrator de características para

(a) Imagem original (b) Imagens geradas a partir de aumentação de dados.

Figura 2.6: Exemplo de aumentação de dados em que 4 novas imagens (b) foram geradas a partir de uma imagem original (a). Os métodos aplicados para aumentação foram, da esquerda para a direita, rotação, cisalhamento, translação e ampliação.

a nova tarefa.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, descrevemos alguns trabalhos e estudos existentes na literatura para a classificação de imagens e textos, bem como abordagens multimodais aplicadas em diferentes domínios, principalmente no contexto de dados médicos.

Métodos de aprendizado profundo são atualmente o estado da arte na resolução de várias tarefas nos campos de visão computacional e processamento de linguagem natural em problemas relacionados à classificação de imagens e textos, bem como envolvendo uma abordagem multimodal desses dados.

3.1 Classificação de Imagens e Textos Médicos

A tarefa de classificação automática de imagens médicas vem sendo amplamente explorada para diversos tipos de imagens em diferentes áreas da medicina. Consequentemente, inúmeras abordagens de aprendizado de máquina surgiram na literatura para a resolução desses problemas. Essas abordagens consistem normalmente de algumas etapas, tais como pré-processamento, extração de características e, por fim, a etapa de classificação [40].

Estudos como os realizados por Kawahara et al. [46] e Menegola et al. [67] investigaram o problema de classificação de imagens de lesões de pele. Kawahara et al. [46] propuseram o treinamento de um classificador linear baseado em características extraídas de imagens por meio do uso de redes neurais convolucionais pré-treinadas. Os autores utilizaram a arquitetura AlexNet [53] pré-treinada com o conjunto ImageNet [81], convertendo as camadas totalmente conectadas em camadas convolucionais para que estas atuassem como filtros convolucionais, extraindo características da imagem.

De maneira semelhante, Menegola et al. [67] propuseram um conjunto de experimentos com diferentes configurações de um modelo e analisaram os efeitos do uso da transferência de aprendizado e do ajuste fino. Para o treinamento e teste dos modelos propostos, os autores utilizaram as bases *Interactive Atlas of Dermoscopy* (Atlas) [6] e *ISBI Challenge 2016 / ISIC Skin Lesion Analysis Towards Melanoma Detection — Part 3: Disease Classification* (ISIC) [30], compostas de imagens de lesões de pele com diagnósticos que incluem lesões benignas, melanomas e carcinomas, entre outros tipos de lesões. Os autores realizaram experimentos para os problemas de classificação de lesões entre Maligno × Benigno, Melanoma × Benigno e Melanoma × Carcinoma × Benigno. Por meio do

uso de transferência de aprendizado a partir do conjunto ImageNet [81] e da técnica de ajuste fino, os autores obtiveram valores de área sob a curva (do inglês, *area under the curve*- AUC) iguais 82,5%, 80,9% e 83,6, respectivamente, para os experimentos citados anteriormente.

O uso de redes neurais convolucionais pré-treinadas para o problema relacionado a imagens de pele foi capaz de obter bons resultados [46] e, de forma análoga, o uso de ajuste fino também propiciou uma melhora nos resultados da classificação [67].

Embora utilizem dados de imagens diferentes do que este trabalho propõe, os estudos mencionados anteriormente abordam o mesmo conceito de sistemas de auxílio ao diagnóstico médico, assim como os estudos de Gale et al. [25] e Rajpurkar et al. [75], os quais utilizam imagens radiográficas de fraturas ósseas, mostrando o quão generalizáveis esses modelos podem ser.

Gale et al. [25] investigaram o problema de detecção de fraturas no quadril por meio de radiografias pélvicas. Em sua base de dados, os autores utilizaram arquivos clínicos de radiologia do *Royal Adelaide Hospital*. Primeiramente, os autores propuseram uma arquitetura de rede convolucional, denominada *CNN-Frontal*, e a treinaram para identificar radiografias pélvicas frontais, pois a base também incluía imagens laterais do quadril, tórax e coluna vertebral. Em seguida, eles propuseram outra arquitetura de rede convolucional, denominada *CNN-bounding*, para realizar uma regressão, que foi treinada para localizar o colo do fêmur que, segundo os autores, seria o local mais relevante em que ocorre a fratura. Por fim, uma terceira arquitetura foi proposta, chamada de *CNN-metal*, cujo objetivo foi excluir casos com metais implantados nas fraturas e outras operações semelhantes. Para a classificação, Gale et al. [25] aplicaram então um modelo de rede convolucional conhecido como DenseNet [38] com 172 camadas de profundidade e utilizaram um estratégia de busca em grade (*grid search*) para determinar os melhores hiperparâmetros do modelo. Os autores obtiveram um valor de acurácia igual a 97%.

Rajpurkar et al. [75] propuseram uma base de dados de radiografias musculoesqueléticas com imagens classificadas como normal ou anormal, validadas por radiologistas certificados pelo conselho do *Stanford Hospital*. A base contém radiografias das extremidades superiores do corpo, sendo elas ombro, úmero, cotovelo, antebraço, punho, mão e dedo. Rajpurkar et al. [75] propuseram também um modelo para predizer a probabilidade de anormalidade em um estudo radiográfico pertencente à base proposta. Para isso, de forma similar ao trabalho de Gale et al. [25], eles utilizaram um modelo baseado na arquitetura DenseNet [38], entretanto, com 169 camadas de profundidade pré-treinada com o conjunto ImageNet. Eles modificaram a camada final totalmente conectada para possuir uma saída única. Os autores avaliaram a eficácia do modelo para cada região do corpo pertencente à base de dados e obtiveram uma média do coeficiente kappa, para o conjunto de teste, igual a 70,5%. Os autores também reportaram a eficácia de 3 radiologistas, em que cada radiologista obteve uma média do coeficiente kappa igual a 73,1%, 76,3% e 77,8% respectivamente. O conjunto de teste utilizado pelos autores, contendo 556 imagens, não foi disponibilizado publicamente.

Wang et al. [101] propuseram a base ChestX-ray8, contendo 108948 imagens de raio-X do tórax e exploraram o uso de vários modelos de DCNN pré-treinados para realizar uma classificação multirrótulo das patologias presentes nas imagens. Além disso, os autores

propuseram uma técnica para localização de patologias combinando as saídas das camadas de predição e *pooling* da rede para gerar um mapa de calor para cada patologia.

Chen et al. [16] também exploraram a classificação de doenças torácicas em imagens de raios-X de tórax. Os autores propuseram uma rede de aprendizado de características assimétricas, chamada DualCheXNet, para explorar a complementaridade e cooperação entre duas redes no aprendizado de características discriminativas. Os autores utilizaram a base ChestX-ray14 [101] e propuseram uma estratégia de treinamento iterativo entre as redes em que, ao final, alcançaram uma média do valor de AUC de $82,3\%$ alcançada pelo modelo.

Smit et al. [87] propuseram um método, denominado CheXbert, para classificação de laudos de radiologia por meio da combinação de classificadores existentes com anotações manuais para obter resultados precisos. Os autores utilizaram um modelo BERT [21] e relataram que um modelo pré-treinado em dados biomédicos é capaz de superar o modelo original pré-treinado com as bases de dados *BookCorpus* e *English Wikipedia*. Os autores também investigaram o uso de uma estratégia chamada de retrotradução (*backtranslation*) para melhorar o desempenho dos modelos e relataram que o CheXbert supera os modelos anteriores treinados apenas em laudos rotulados por radiologistas ou apenas nas saídas de rotuladores existentes, obtendo uma média da medida F1 de 79,8% no conjunto MIMIC-CXR [42], representando o estado da arte para rotulação de relatórios em conjuntos de dados de raios-X de tórax.

3.2 Classificação de Dados Multimodais

Muitos trabalhos recentes exploraram problemas envolvendo dados multimodais, que requerem o aprendizado de informações de muitos domínios diferentes utilizando arquiteturas e estratégias de redes neurais e redes neurais profundas [74].

Atualmente, há muitas pesquisas em aplicações multimodais, por exemplo, legendagem de imagens [10], tradução automática [23], *Visual Question-Answering* [4], entre outras [24, 102].

Os trabalhos de Arevalo et al. [5], Kiela et al. [47, 48], Vielzeuf et al. [98] e Perez-Rua et al. [74] exploraram abordagens multimodais para classificação de gêneros de filmes por meio das imagens dos pôsteres e textos das sinopses.

Arevalo et al. [5] apresentaram o modelo chamado *Gated Multimodal Unit* (GMU) para aprendizado multimodal baseado em redes neurais recorrentes. O objetivo do modelo GMU é encontrar uma representação intermediária com base em uma combinação de dados de diferentes modalidades e aprender como estas modalidades influenciam a ativação das unidades recorrentes da rede.

Kiela et al. [48] investigaram vários métodos para classificação multimodal em larga escala, explorando métodos para combinar o que os autores chamaram de elementos discretos (por exemplo, texto) e elementos contínuos (por exemplo, representações visuais) em uma estrutura multimodal. Os autores se concentraram no desenvolvimento de métodos apropriados para classificar grandes quantidades de dados de forma rápida.

Vielzeuf et al. [98] propuseram uma abordagem de fusão multimodal que integra infor-

mações provenientes de dados de múltiplas mídias. Os autores propuseram um modelo, denominado *CentralNet*, que consiste em uma rede central para realizar uma representação conjunta conectando as diferentes camadas de redes neurais específicas de cada modalidade.

Perez-Rua et al. [74] apresentaram um método de busca para encontrar de forma precisa arquiteturas de fusão para classificação multimodal. Os autores propuseram uma busca por uma arquitetura neural para refinar a fusão de informações multimodais com base em um esquema de otimização.

Kiela et al. [47] propuseram um modelo multimodal que funde informações de codificadores de texto e imagem. Segundo os autores, a combinação de representações textuais utilizando métodos de processamento de linguagem natural, com abordagens de visão computacional utilizando arquiteturas de redes neurais convolucionais, pode fornecer desempenho que representa o estado da arte em várias tarefas de classificação multimodal.

Os autores reportaram os resultados para cada abordagem individual (imagem e texto) e multimodal, bem como apresentaram o benefício dessa abordagem. Na Tabela 3.1, apresentamos os resultados de micro F1-*score* reportado pelos autores em cada trabalho. Perez-Rua et al. [74] não utilizaram a métrica micro F1-*score*, mas reportaram resultados de macro F1-*score* de 33,5%, 45,9% e 55,6% para abordagens de imagem, texto e multimodal, respectivamente.

Método	**Imagem**	**Texto**	**Multimodal**
Arevalo et al. [5]	43,7%	59,5%	63,0%
Kiela et al. [48]	49,3%	58,8%	62,3%
Vielzeuf et al. [98]	47,8%	60,2%	63,9%
Kiela et al. [47]	44,7%	65,2%	66,4%

Tabela 3.1: Resultados de micro F1-*score* para diferentes trabalhos apresentando a melhora que a abordagem multimodal produz em relação à abordagem individual.

Pelka et al. [73] propuseram uma abordagem para combinar palavras-chave geradas automaticamente a partir de radiografias, com as próprias imagens radiográficas e, em seguida, realizar uma etapa de classificação. Eles apresentaram um método que permite representações de imagens multimodais, fundindo informações textuais que foram geradas, com as imagens radiográficas. Os autores utilizaram a base de dados MURA com o objetivo de realizar o reconhecimento das partes do corpo e identificar anormalidades nas imagens. Os autores reportaram valores de acurácia de 79,8% e 54,2% obtidas com as abordagem individuais de imagem e texto, respectivamente, e 81,5% utilizando a abordagem multimodal.

3.3 Considerações Finais

Neste capítulo, apresentamos alguns trabalhos correlatos para as tarefas de classificação de imagens e textos envolvendo dados médicos, além de trabalhos que abordam estas tarefas com o uso de dados multimodais.

Em nossa busca por trabalhos relacionados ao tema investigado em nossa pesquisa, identificamos um vasto conjunto de métodos e técnicas, tanto envolvendo a classificação de imagens (o tópico principal deste trabalho) quanto envolvendo abordagens de processamento de linguagem natural para problemas relacionados a texto.

A principal motivação para o uso de abordagens multimodais é extrair e combinar informações relevantes das diferentes modalidades e, portanto, superar os resultados obtidos individualmente. Este trabalho está relacionado a abordagens que investigam a integração de informações de múltiplas modalidades em redes neurais, bem como explorar métodos de fusão. Similar a alguns trabalhos mencionados, nossa abordagem explora a cooperação entre redes de imagem e texto para, em contraste com a maioria desses, a classificação de anormalidades em dados médicos.

Capítulo 4

Conjuntos de Dados e Métricas de Avaliação

Neste capítulo, apresentamos as bases de dados utilizadas paras as diferentes tarefas a serem realizadas neste trabalho, em que descrevemos os tamanhos, os formatos e as características dos seus dados. Descrevemos também as métricas utilizadas para avaliar a qualidade dos métodos propostos em cada etapa do trabalho.

4.1 Conjuntos de Dados

Há muitos conjuntos de dados que utilizam imagens e tecnologias de exames médicos (tais como raios-X, ressonância magnética, tomografia computadorizada, ultrassonografia e mamografia) de partes do corpo (tais como pulmão, abdômen, tórax e cérebro) que podem ser empregados em muitas tarefas, por exemplo, classificação, detecção e segmentação [51, 61]. Nesta seção, apresentamos as bases de dados utilizadas no processo de desenvolvimento do método proposto.

Musculoskeletal Radiographs (MURA)

O conjunto de dados que escolhemos para conduzir os experimentos para a tarefa de classificação é conhecido como *Musculoskeletal Radiographs* (MURA) [75], que é um grande conjunto de dados de radiografias ósseas formado por imagens de raios-X de estudos musculoesqueléticos de extremidades superiores do corpo. Cada estudo radiográfico pertence a uma dentre sete extremidades: cotovelo, dedo, antebraço, mão, úmero, ombro e pulso. A Figura 4.1 apresenta algumas radiografias para ilustrar cada uma das sete classes de anatomia.

Cada estudo na base MURA contém uma ou mais imagens radiográficas e foram rotulados como "normais" ou "anormais" por radiologistas. A Figura 4.2 apresenta algumas imagens contidas em um mesmo estudo radiográfico pertencente à região do corpo úmero, cujo rótulo desse estudo foi definido como "anormal".

O conjunto de dados MURA possui 40561 imagens radiográficas musculoesqueléticas de 14863 casos de estudo a partir de 12173 pacientes e é composto de três subconjuntos,

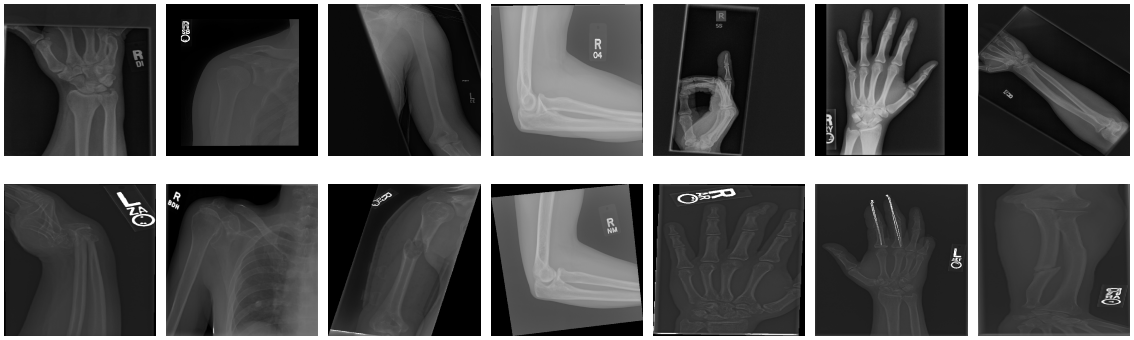


Figura 4.1: Exemplos de radiografias presentes no conjunto de dados MURA. As imagens da primeira fileira pertencem à classe negativa de anormalidade, enquanto as imagens da segunda fileira pertencem à classe positiva de anormalidade.

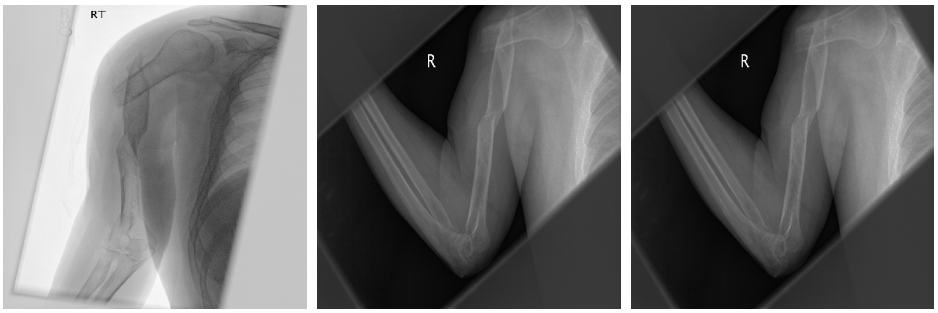


Figura 4.2: Exemplos de imagens de radiografia presentes em um mesmo estudo de caso da região do úmero rotulado como "anormal" contido no conjunto de dados MURA.

treinamento, validação e teste, contendo um total de 36808, 3197 e 556 imagens, respectivamente. No entanto, o conjunto de teste não está disponível publicamente. Dessa forma, para fins de avaliação e reprodutibilidade, consideramos o conjunto de validação como sendo o de teste e um novo conjunto de validação foi criado utilizando 20% do conjunto de treinamento[1].

A Tabela 4.1 mostra o arranjo das imagens em cada região e classe nos conjuntos de treinamento, validação e teste. Pode-se observar um desbalanceamento na quantidade de imagens em determinadas classes para cada região do corpo. Apenas a região "Ombro" apresenta uma distribuição mais igualitária entre as classes, tanto no conjunto de treinamento quanto nos conjuntos de validação e teste.

As imagens contidas nesta base de dados são monocromáticas e estão no formato *Portable Network Graphics* (PNG). Algumas imagens apresentam uma certa rotação e variam em termos de razão de aspecto e de resolução, geralmente entre 132×512 e 512×512 pixels.

[1]Não há sobreposição de pacientes entre qualquer um dos conjuntos.

Classes	Treinamento		Validação		Teste		Total
	Normal	Anormal	Normal	Anormal	Normal	Anormal	
Pulso	4612	3189	1153	798	364	295	**10411**
Ombro	3368	3334	843	834	285	278	**8942**
Mão	3247	1187	812	297	271	189	**6003**
Dedo	2510	1574	628	394	214	247	**5567**
Cotovelo	2340	1604	585	402	235	230	**5396**
Antebraço	931	528	233	133	150	151	**2126**
Úmero	538	479	135	120	148	140	**1560**
Total	17546	11895	4389	2978	1667	1530	**40005**

Tabela 4.1: Distribuição das imagens em cada região do corpo e classe nos conjuntos de treinamento, validação e teste do conjunto de dados MURA.

Radiology Objects in COntext (ROCO)

Radiology Objects in COntext (ROCO) [72] é um conjunto de dados composto de 81825 imagens de radiologia oriundas de diferentes modalidades de imagens médicas, incluindo Tomografia Computadorizada (CT), Ultrassonografia, Raios-X, Fluoroscopia, Tomografia por Emissão de Pósitrons (PET), Mamografia, Imagem por Ressonância Magnética (MRI), Angiografia e PET-CT. A Figura 4.3 apresenta exemplos das várias modalidades de imagens presentes na base ROCO.

Figura 4.3: Exemplos de imagens de radiologia contidas no conjunto de dados ROCO, ilustrando a variedade de modalidades de imagens médicas.

As imagens da base ROCO têm associados um *caption*, *keywords*, *Concept Unique Identifiers* (CUIs) e *Semantic Types* (semType). A Figura 4.4 ilustra um exemplo de imagem de radiologia contendo informações textuais.

O conjunto de dados ROCO também possui um subconjunto, denominado "fora da classe" (*Out-Of-Class*), de imagens radiológicas. Este subconjunto inclui (Figura 4.5) 6127 imagens de radiologia sintética, fotografias clínicas, retratos, bem como arte digital. Este subconjunto não foi considerado neste trabalho.

A Tabela 4.2 apresenta a distribuição das imagens radiológicas e "fora de classe" presentes na base de dados ROCO nos conjuntos de treinamento validação e teste.

CheXpert

A base CheXpert [39] é um grande conjunto de dados de raios-X de tórax contendo 224316 radiografias a partir de 65240 pacientes do *Stanford University Medical Center*. O conjunto CheXpert foi rotulado com 14 observações sendo elas: cardiomiopatia aumentada,

Caption: Radiograph of a shoulder in lateral scapular view showing breakage of synthesis material (arrow).

Keywords: shoulder, scapular, lateral, material, radiograph, arrow, breakage, view, synthesi

CUI:

C1306645	Plain x-ray
C0524468	Structure of right shoulder region
C1548003	Radiograph
C0036277	Bone structure of scapula
C0043299	Diagnostic radiologic examination
C0037004	Shoulder

SemTypes:

T060	Diagnostic Procedure
T170	Intellectual Product
T121	Pharmacologic Substance
T033	Finding
T023	Body Part, Organ, or Organ Component
T109	Organic Chemical
T029	Body Location or Region

Figura 4.4: Exemplo de uma amostra de imagem de radiologia com legenda, palavras-chave, conceitos e tipos semânticos correspondentes pertencente ao conjunto de dados ROCO.

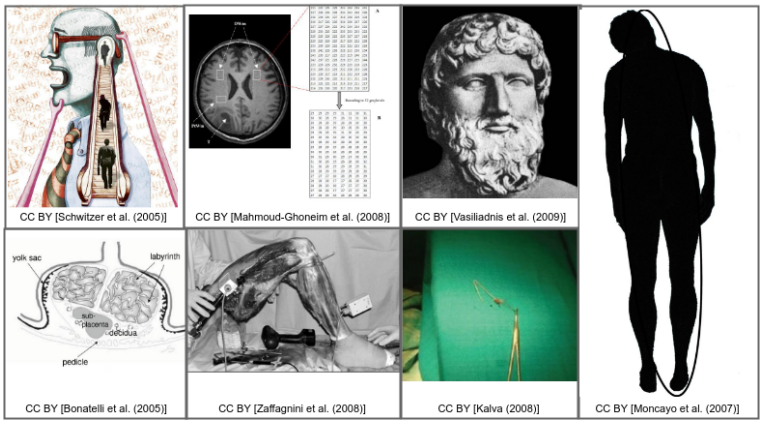


Figura 4.5: Exemplos de imagens contidas no subconjunto *Out-Of-Class* do conjunto de dados ROCO [72].

Tipo de imagem	**Treinamento**	**Validação**	**Teste**	**Total**
Radiológicas	65460	8183	8182	81825
Out-of-Class	4902	612	613	6127
Total	70362	8795	8795	87952

Tabela 4.2: Distribuição das imagens presentes no conjunto de dados ROCO nos conjuntos de treinamento, validação e teste.

cardiomegalia, lesão pulmonar, opacidade pulmonar, edema pulmonar, consolidação pulmonar, pneumonia, atelectasia, pneumotórax, derrame pleural, outros problemas pleurais, fratura, presença de dispositivos de suporte e uma classe de nenhum achado (*no finding*). Na Figura 4.6, apresentamos algumas radiografias presentes no conjunto CheXpert para ilustrar cada uma das quatorze classes de observações.

O conjunto de dados é dividido em subconjuntos de treinamento, validação e teste, contendo 223616, 200 e 500 estudos, respectivamente. Para cada estudo, um valor foi atribuído a respeito das patologias como positiva (1), negativa (0) ou incerta (−1). Na

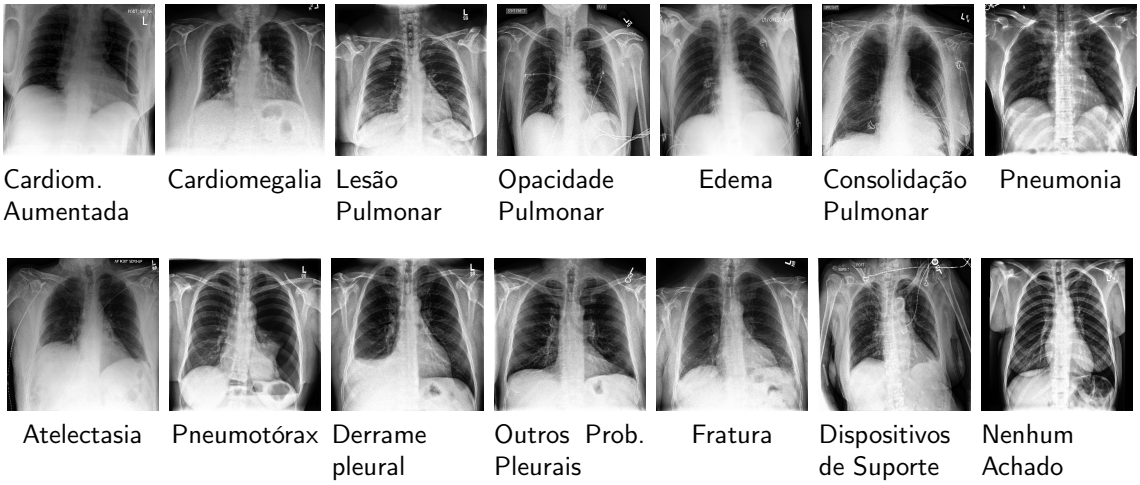


Figura 4.6: Exemplos de radiografias presentes no conjunto de dados CheXpert. As imagens podem possuir mais de uma classe de observação. Imagens que não possuem observações de patologias pertencem à classe "nenhum achado".

Tabela 4.3, apresentamos a distribuição de estudos para cada observação no conjunto de treinamento.

Patologias	**Positiva**	**Incerta**	**Negativa**
Cardiomiopatia Aumentada	9020	10148	168473
Cardiomegalia	23002	6597	158042
Lesão Pulmonar	6856	1071	179714
Opacidade Pulmonar	92669	4341	90631
Edema	48905	11571	127165
Consolidação Pulmonar	12730	23976	150935
Pneumonia	4576	15658	167407
Atelectasia	29333	29377	128931
Pneumotórax	17313	2663	167665
Derrame pleural	75696	9419	102526
Outros Problemas Pleurais	2441	1771	183429
Fratura	7270	484	179887
Dispositivos de Suporte	105831	898	80912
Nenhum Achado	16627	0	171014

Tabela 4.3: Distribuição das observações de patologias em cada estudo radiográfico presentes no conjunto de treinamento da base CheXpert [39].

4.2 Métricas de Avaliação

Um conjunto de métricas é normalmente empregado para avaliar a eficácia de um modelo de classificação. Essas métricas baseiam-se no conceito de matriz de confusão, que oferece uma representação para mensurar o desempenho do modelo de classificação ao apresentar as predições efetuadas para cada classe.

Uma ilustração da matriz de confusão é apresentada na Tabela 4.4. As métricas que serão utilizadas neste trabalho para avaliar os resultados obtidos com os métodos propostos são definidas a seguir.

		Classe real	
		Positivo	**Negativo**
Resultado predito	**Positivo**	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Tabela 4.4: Matriz de confusão para um problema de classificação.

A acurácia (ACC), definida na Equação 4.1, consiste na quantidade de elementos que foram preditos corretamente, positiva ou negativamente, dividida pela quantidade total de amostras. A acurácia visa indicar o quão frequente o classificador está correto e é idealmente utilizada quando as quantidades de amostras que pertencem a cada classe na base de dados são balanceadas.

$$\text{ACC} = \frac{\text{VP+VN}}{\text{Total}} \tag{4.1}$$

Quando a base é desbalanceada, idealmente utiliza-se a acurácia balanceada (ACC_b) [12], que consiste na média da acurácia obtida em cada classe. Assim, pode-se evitar valores elevados de acurácia apenas por conta do desbalanceamento da base. A acurácia balanceada para a classificação binária, por exemplo, é expressa na Equação 4.2.

$$\text{ACC}_b = \frac{1}{2}\left(\frac{\text{VP}}{\text{VP+FN}} + \frac{\text{VN}}{\text{VN+FP}}\right) \tag{4.2}$$

A área sob a curva ROC (*Area Under the Receiver Operating Characteristic* (ROC) *Curve* - AUC) é uma medida da área sob a curva formada pela Taxa de Verdadeiros Positivos (TVP ou revocação) e a Taxa de Falsos Positivos (TFP), expressas nas Equações 4.3 e 4.4, respectivamente. A métrica fornece a medida da qualidade das predições de um modelo.

$$\text{TVP} = \frac{\text{VP}}{\text{VP + FN}} \tag{4.3}$$

$$\text{TFP} = \frac{\text{FP}}{\text{FP + VN}} \tag{4.4}$$

Um exemplo de gráfico contendo a curva ROC e o valor da área pode ser visto na Figura 4.7. Uma curva próxima da linha tracejada (área = 0.5) representa um desempenho inadequado do modelo, enquanto valores mais próximos a 1 são considerados adequados.

O coeficiente *Cohen's kappa* [19] é uma medida estatística para avaliar o nível de concordância entre dois conjuntos de dados em um problema de classificação. Ele é uma medida mais robusta do que um simples cálculo de concordância percentual, já que este coeficiente leva em consideração a possibilidade desta concordância ocorrer por acaso. O

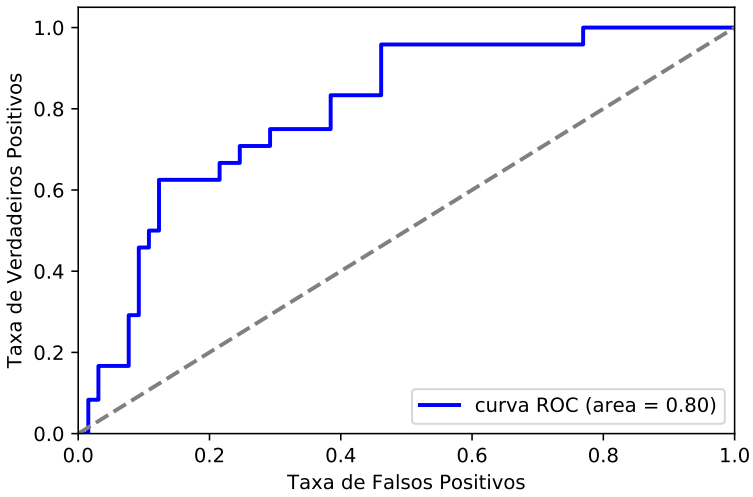


Figura 4.7: A área sob a curva ROC mede toda a área do gráfico, do ponto $(0, 0)$ a $(1, 1)$.

coeficiente kappa pode ser expresso pela Equação 4.5:

$$\text{kappa} = \frac{P_o - P_e}{1 - P_e} \tag{4.5}$$

em que P_o é a concordância observada, cuja fórmula é análoga à da acurácia (Equação 4.1), e P_e é a concordância esperada, que está relacionada ao número de instâncias de cada classe, juntamente com o número de instâncias que o classificador predisse corretamente. Sua fórmula é apresentada na Equação 4.6.

$$P_e = \frac{(\text{FP}+\text{VN}) \times (\text{FN}+\text{VN}) + (\text{VP}+\text{FN}) \times (\text{VP}+\text{FP})}{\text{Total}^2} \tag{4.6}$$

O valor do coeficiente kappa é sempre menor ou igual a 1. Um valor igual a 1 implica concordância perfeita e valores menores do que 1 implicam menos concordâncias entre as anotações. Possíveis interpretações do coeficiente kappa, segundo Landis e Koch [54], estão sumarizadas na Tabela 4.5.

Valor do coeficiente kappa	**Nível de concordância**
<0,0	Não há concordância
[0,0 - 0,2]	Concordância mínima
(0,2 - 0,4]	Concordância razoável
(0,4 - 0,6]	Concordância moderada
(0,6 - 0,8]	Concordância substancial
(0,8 - 1,0]	Concordância perfeita

Tabela 4.5: Classificação dos diferentes níveis de concordância do coeficiente kappa.

Capítulo 5

Método Proposto

Neste capítulo, apresentamos o método proposto para abordar o problema de classificação de anormalidades musculoesqueléticas com uma estratégia multimodal com imagens e textos médicos, a fim de realizar uma classificação binária na base de dados utilizada.

A Figura 5.1 apresenta uma visão geral do fluxo de execução do método proposto. Abordamos, primeiramente, o problema de legendagem de imagens médicas como um método prévio para auxiliar a abordagem de classificação multimodal proposta. Em seguida, a rede DenseNet-169 e o modelo BERT são usados para aprender a partir dos dados de imagem e texto, respectivamente. As características extraídas de cada rede são concatenadas e utilizadas como entrada para o classificador de fusão. A Seção 5.1 descreve cada um desses modelos.

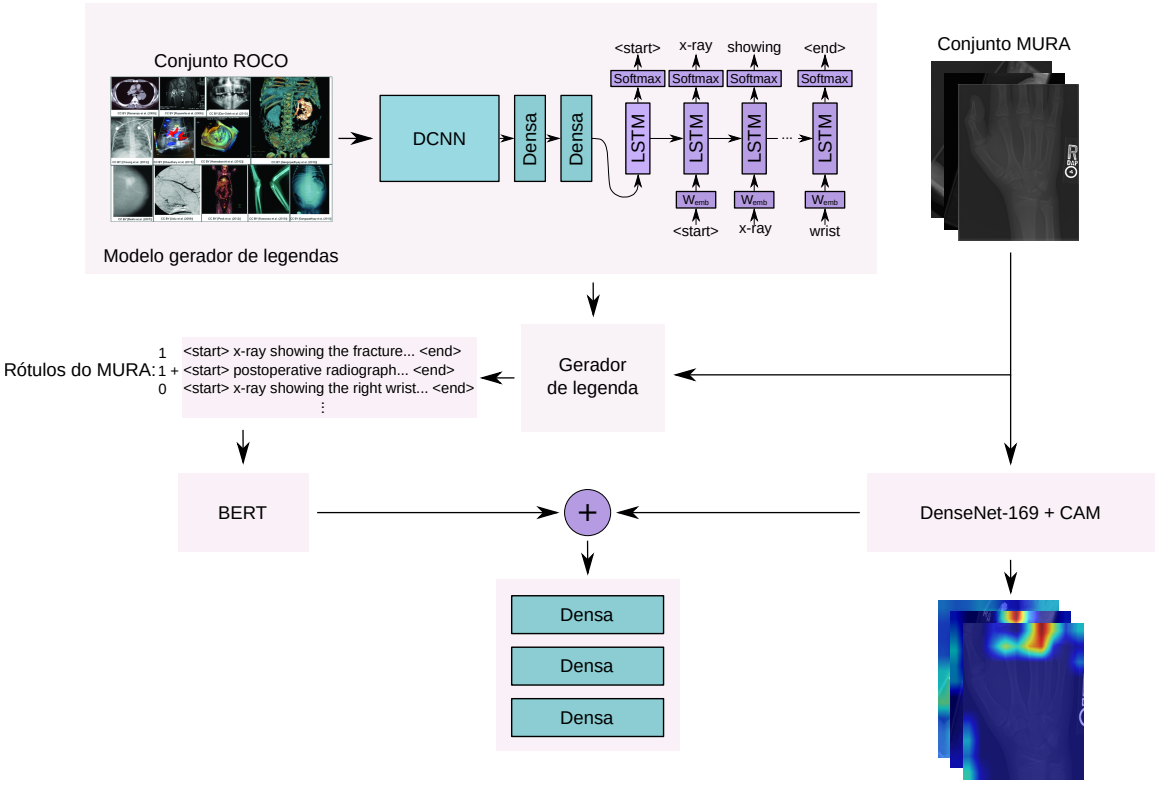


Figura 5.1: Representação do fluxo de execução do método proposto, com os principais passos e modelos utilizados em cada etapa.

5.1 Modelos Implementados

Nesta seção, descrevemos os modelos que foram implementados e utilizados em cada etapa do método proposto, apresentando seus formatos e suas características.

5.1.1 Modelo Gerador de Legendas

O modelo projetado para gerar texto a partir de imagens é uma arquitetura de Codificador-Decodificador (*Encoder-Decoder*) baseada no modelo *Show and Tell* desenvolvido por Vinyals et al. [99].

Como codificador, utilizamos uma Rede Neural Convolucional Profunda (DCNN), chamada ResNet-152 [34], pré-treinada com o conjunto ImageNet [81]. O codificador é responsável por extrair um vetor de características de uma imagem de entrada, que é linearmente transformado para ser usado como entrada para o decodificador. O decodificador é uma rede baseada em uma LSTM, que recebe como entrada o vetor de características do codificador e produz a legenda da imagem como saída.

Para a construção do vocabulário, realizamos o processo de *tokenização*, isto é, particionamos o texto em *tokens*. Os *tokens* foram criados para cada palavra presente nas sentenças. Além disso, foram adicionados ao vocabulário os *tokens* especiais: `<start>` e `<end>`, colocados no começo e no final da sentença de entrada; `<unk>`, para substituir eventuais palavras não pertencentes ao vocabulário; e `<pad>`, para preencher sentenças com um tamanho menor que o definido. O vocabulário construído possui 16,380 palavras.

A arquitetura do modelo de geração de legenda é ilustrada na Figura 5.2. O modelo foi treinado utilizando o par de imagens e legendas do conjunto de dados ROCO.

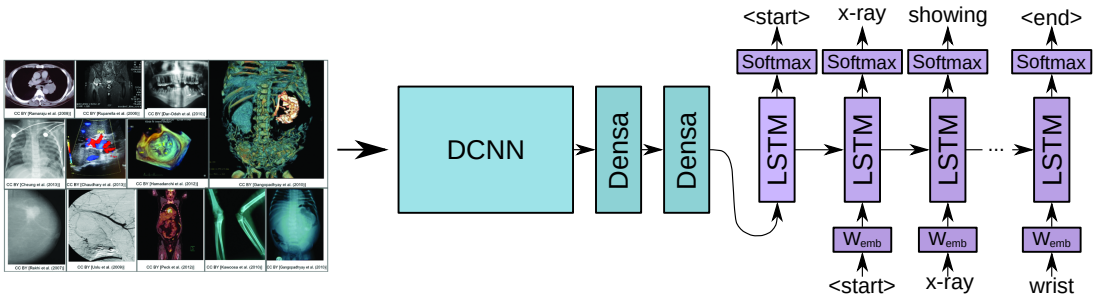


Figura 5.2: Representação do modelo Codificador-Decodificador para a geração de legendas a partir de imagens radiográficas.

Como representações textuais não estão presentes no conjunto de dados MURA, este modelo de geração de legendas foi utilizado para criar artificialmente essas representações.

5.1.2 DenseNet-169

O modelo que empregamos para classificar as imagens radiográficas foi a DenseNet-169. Inicialmente, utilizamos a DenseNet-169 pré-treinada com o conjunto ImageNet [20] e treinamos a rede em nosso conjunto de dados de imagens médicas. Este modelo também foi usado posteriormente como um extrator de características.

Neste modelo, adicionamos um módulo de atenção denominado *Class Activation Mapping* (CAM) [107], empregado para indicar uma região relevante e discriminativa detectada pelo modelo DCNN para identificar a classe correta.

Após este módulo de atenção, aplicamos uma camada de *pooling* Média-Máxima (AVG-MAX) para reduzir a complexidade computacional e extrair características de baixo e alto níveis da vizinhança. Extraímos os *embeddings* das imagens a partir da camada de AVG-MAX *pooling*, a última camada antes do classificador. Na Figura 5.3, apresentamos a arquitetura do modelo da DenseNet-169 utilizada para a classificação e a extração de características das imagens.

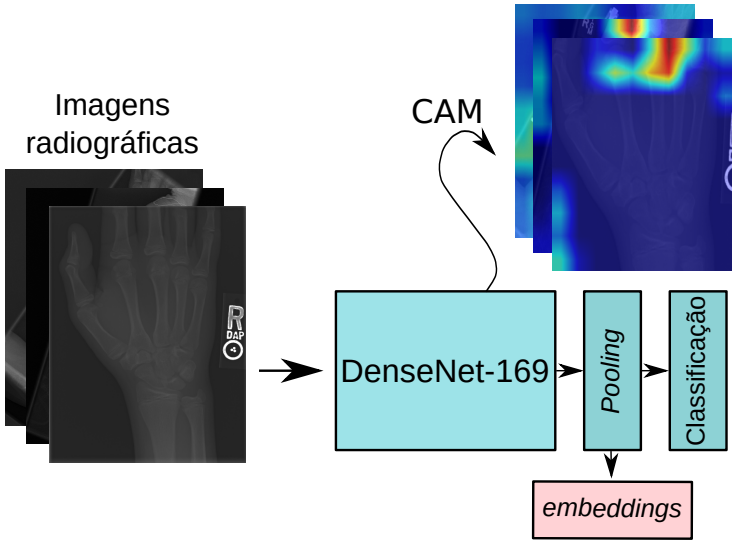


Figura 5.3: Representação do modelo DenseNet-169 utilizado para a classificação e a extração de características de imagens radiográficas.

Inicialmente, as imagens são redimensionadas para 328×328 pixels e normalizadas com a média e o desvio padrão da base ImageNet. Após o treinamento deste modelo, realizamos uma etapa em que extraímos os *embeddings* de todas as imagens de treinamento e validação, que serão utilizadas em etapas posteriores. Ao final, na fase de teste, o mesmo processo é realizado para cada imagem do conjunto de teste.

5.1.3 BERT

Utilizamos uma abordagem de Processamento de Linguagem Natural para extrair rótulos estruturados a partir de texto de legendas de imagens. O modelo usado para classificar esses textos foi o BERT [21] para realizar um ajuste fino.

A arquitetura do modelo BERT é baseada em transformador bidirecional de multi-camadas [97] e pré-treinado de forma não supervisionada em duas tarefas: modelo de linguagem mascarada e predição da próxima sentença [9].

O modelo segue a mesma arquitetura do BERT. Cada texto de legenda de imagem é *tokenizado* e o número máximo de *tokens* em cada sequência de entrada é limitado a 64. O estado da camada oculta final é então utilizado como entrada para cada uma das cabeças lineares do modelo BERT. Alteramos a dimensão de saída do BERT para 1 para contemplar o nosso problema de classificação binária.

Aplicamos uma operação de *pooling* Média-Máxima (AVG-MAX) antes da camada do classificador e também utilizamos este modelo como um extrator de características para as próximas etapas.

5.1.4 Fusão

O modelo utilizado como módulo de fusão multimodal foi um Perceptron de Multicamadas (*Multilayer Perceptron* - MLP) formado por três camadas densas (ou camadas totalmente conectadas). Este modelo recebe como entrada a concatenação das características extraídas das imagens pela rede DenseNet-169 e as características extraídas das legendas das imagens pelo modelo BERT.

5.2 Geração Automática de Legendas

Um dos primeiros passos para a interpretação de imagens médicas é identificar as anormalidades que podem ser visualizadas em cada imagem [51]. Dadas as imagens de exames médicos (por exemplo, raios-X, PET/CT), procuramos identificar as características relevantes que descrevem achados e identificam anormalidades associadas à imagem.

Ao identificar anormalidades presentes em uma imagem, uma das tarefas que podem ser realizadas em seguida é a de produzir um documento, similar a um laudo médico, em que são descritas essas anormalidades, que é uma tarefa semelhante à de legendagem de imagens [99, 104].

A tarefa de legendagem de imagens consiste em gerar legendas que descrevem objetos ou características presentes em uma imagem (Figura 5.4a) e as relações entre os mesmos. Esta tarefa vem sendo bastante explorada, uma vez que, recentemente, muitos métodos surgiram e foram aplicados com sucesso obtendo bons resultados [2, 63]. As técnicas de legendagem de imagens têm sido aplicadas a imagens médicas [41] com o objetivo de produzir um relatório ou diagnóstico que descreve a condição de um paciente [51]. Na Figura 5.4b, apresentamos um exemplo de legenda gerada para uma imagem médica.

(a) Exemplo de legendagem em imagem

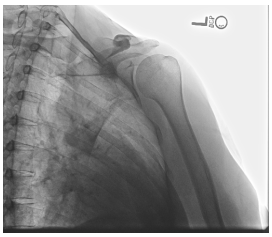
No acute cardiopulmonary abnormality. Stable appearance of the thoracic aorta. The right lateral lower lobe is noted in the right lower right midlung. No large pleural effusion or focal airspace disease. Mild interstitial opacities. Atherosclerotic calcifications bony structures bilaterally. There is no pleural effusion or pneumothorax developed in the right lower lobe.

(b) Exemplo de legendagem em imagem médica

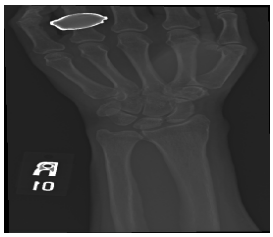
Figura 5.4: Exemplos de legendas geradas para (a) uma imagem geral e para (b) uma imagem médica. Em (a), a legenda é gerada pelo trabalho de Vinyals et al. [99] com o modelo *Show and Tell*. Em (b), a legenda foi gerada pelo modelo proposto por Jing et al. [41].

Como representações textuais não estão presentes no conjunto de dados MURA [75], usamos o modelo Codificador-Decodificador (descrito na Seção 5.1.1) como gerador de legendas. O modelo foi treinado no conjunto de dados ROCO [72] para gerar relatórios médicos automáticos para as imagens no conjunto de dados MURA.

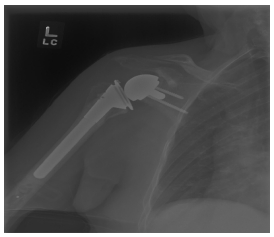
Treinamos o modelo usando pares de imagens e legendas do conjunto de dados ROCO. Após a etapa de treinamento do modelo, um conjunto de dados contendo as imagens originais do conjunto MURA e as legendas artificiais geradas foi construído. Ou seja, ao final desta etapa, temos uma combinação de imagem e classes originais da base MURA e uma legenda gerada automaticamente associada. Na Figura 5.5, apresentamos exemplos de legendas geradas para quatro radiografias do conjunto MURA.



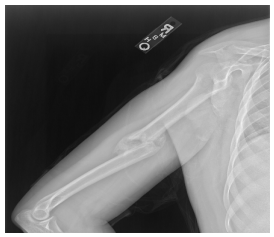
Conjunto: Ombro
Classe: Normal
Legenda: the right shoulder is normal



Conjunto: Pulso
Classe: Normal
Legenda: X-ray showing the of the right wrist



Conjunto: Ombro
Classe: Anormal
Legenda: X-ray of the shoulder showing the acromioclavicular joint



Conjunto: Úmero
Classe: Anormal
Legenda: X-ray showing the fracture of the right humerus

Figura 5.5: Exemplos de legendas geradas. Todas essas informações estão presentes no conjunto de dados construído. O modelo do gerador de legenda foi treinado usando todas as imagens de radiologia do conjunto de dados ROCO.

5.3 Considerações Finais

O objetivo deste capítulo foi o de propor uma abordagem utilizando representações de dados multimodais para detectar anormalidades musculoesqueléticas em estudos radiográficos.

Utilizamos um gerador automático de legendas para imagens de radiografias. Em seguida, um método foi proposto para combinar essas legendas geradas artificialmente (Seção 5.2) e as radiografias, com o objetivo de obter uma classificação precisa das anormalidades musculoesqueléticas.

Um modelo de DCNN foi treinado em imagens de radiografias musculoesqueléticas e, paralelamente, foi realizado um ajuste fino em um modelo BERT com as legendas geradas automaticamente de cada imagem radiográfica. Por fim, utilizamos uma abordagem de fusão de dados por meio da combinação de características oriundas das diferentes modalidades. Este processo permitiu uma representação de dados multimodais utilizados como entrada para uma rede neural simples, empregada como módulo de fusão. Esta representação de dados multimodais apresentou os resultados mais elevados (descritos no Capítulo 6) nas métricas analisadas.

Capítulo 6

Resultados Experimentais

Neste capítulo, descrevemos os recursos computacionais utilizados ao longo do desenvolvimento deste trabalho, a linguagem de programação e as bibliotecas empregadas na implementação dos métodos propostos.

Avaliamos o uso de redes neurais convolucionais, em conjunto com as técnicas de aprendizado de máquina e visão computacional, para detectar anormalidades em imagens de raio-X. Realizamos uma série de experimentações de forma incremental, em que, para cada técnica aplicada, avaliamos a contribuição dos modelos implementados.

Por fim, os resultados obtidos são apresentados com base nos modelos descritos no Capítulo 5, em que examinamos o desempenho de cada modelo individualmente utilizado na tarefa de classificação, bem como o desempenho da abordagem multimodal proposta.

6.1 Recursos Computacionais

Os métodos propostos foram implementados utilizando a linguagem Python, que provê uma gama de bibliotecas e recursos para Visão Computacional, Processamento de Linguagem Natural e Aprendizado de Máquina, tais como `NumPy`[1], `Matplotlib`[2], `OpenCV`[3], `scikit-learn`[4], `scikit-image`[5] e `PyTorch`[6].

Na construção do modelo BERT para a tarefa de classificação de textos, utilizamos a implementação da biblioteca `Hugging Face`[7], que contém códigos abertos e materiais de referência para processamento de linguagem natural.

Os experimentos deste trabalho foram conduzidos no Laboratório de Informática Visual (LIV) do Instituto de Computação da Unicamp. As máquinas que foram utilizadas para a realização dos experimentos seguem as especificações descritas na Tabela 6.1.

[1] `https://numpy.org/`
[2] `https://matplotlib.org/`
[3] `https://opencv.org/`
[4] `https://scikit-learn.org/`
[5] `https://scikit-image.org/`
[6] `https://pytorch.org/`
[7] `https://huggingface.co/`

Processador	Intel i7-3770 3.5GHz
Memória RAM	32GB
Placa de vídeo	NVidia GeForce GTX 1080
Memória	11GB
Sistema Operacional	Ubuntu 16.04 LTS

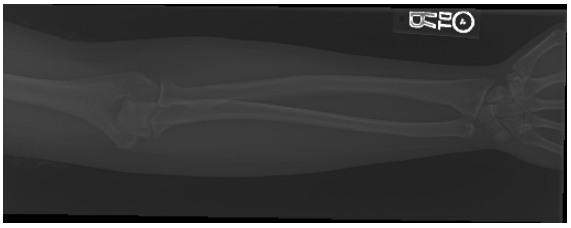
Tabela 6.1: Especificações das máquinas utilizadas nos experimentos.

6.2 Experimentos Utilizando Redes Neurais Convolucionais

Nesta seção, descrevemos os experimentos que foram realizados, as técnicas aplicadas e os processos e modelos utilizados. Os resultados destes experimentos foram obtidos utilizando o conjunto de validação da base de dados MURA. O conjunto de teste foi utilizado apenas para a avaliação final da metodologia proposta.

Experimento I: Analisando Transformações

Nosso foco inicial foi usar redes pré-treinadas como referência (*baseline*). No entanto, as redes pré-treinadas geralmente utilizam imagens de dimensões quadráticas como entradas, enquanto as imagens da base MURA possuem dimensões variáveis. Dessa forma, antes de avaliar uma variedade de redes, experimentamos três tipos de transformações para aplicar nas imagens: *fit*, *pad*, ou *stretch*. Três transformações são ilustradas na Figura 6.1.



(a) Imagem Original

(b) *Fit* (c) *Pad* (d) *Stretch*

Figura 6.1: Três transformações aplicadas em uma imagem retangular. O *fit* não deforma a imagem, mas causa perda de informações. A transformação *pad* também não deforma a imagem, mas adiciona informações irrelevantes. A transformação *stretch* inclui todos os pixels, mas deforma a imagem.

Fit realiza um recorte na maior dimensão da imagem para corresponder à menor, mantendo o quadrado central. Esta transformação preserva as proporções da imagem,

entretanto, causa perda na quantidade de informação, que pode ser considerável se a imagem for muito alta ou larga.

Pad adiciona uma borda preta no maior lado, transformando a imagem em um quadrado sem perder nenhuma região. Esta transformação também preserva as proporções da imagem, entretanto, preenche parte da imagem com informações irrelevantes.

Stretch altera diretamente a proporção da imagem para um quadrado, estendendo o lado mais curto até que corresponda ao mais longo. Esta transformação causa uma distorção na imagem, entretanto, inclui a imagem inteira.

Para decidir o melhor método a ser aplicado nas imagens, treinamos uma rede ResNet utilizando as três transformações descritas e comparamos os desempenhos resultantes. Na Tabela 6.2, apresentamos os resultados obtidos.

	Acurácia	**Acurácia Balanceada**	**AUC ROC**	**Kappa**
Fit	0,8232	0,8140	0,8720	0,6373
Pad	0,8165	0,8054	**0,8723**	0,6222
Stretch	**0,8274**	**0,8175**	0,8617	**0,6453**

Tabela 6.2: Desempenho do modelo ResNet utilizando as três transformações propostas. Destacamos (em **negrito**) os melhores valores para cada métrica.

Os desempenhos obtidos utilizando os três métodos foram bastante similares, entretanto, ao utilizar a transformação *stretch*, obteve-se os melhores valores na métricas acurácia, acurácia balanceada e coeficiente kappa. Além disso, *stretch* foi a operação mais rápida nos experimentos, carregando, por exemplo, o conjunto de treinamento 10% mais rápido do que a operação *pad* e 31% mais rápido do que a operação *fit*. Portanto, escolhemos a transformação *stretch* para ser usada nos experimentos subsequentes.

Experimento II: Utilizando Redes Pré-Treinadas

Modelos pré-treinados têm demonstrado bons resultados em tarefas de classificação de imagens [31] e seu uso proporciona uma arquitetura já consolidada e validada. Além disso, o uso desses modelos reduz o tempo de treinamento em comparação ao treinamento necessário para alcançar resultados semelhantes sem o pré-treinamento.

Modelos pré-treinados podem, entretanto, ter algumas ressalvas. Os modelos foram treinados usando o conjunto de dados ImageNet [81], contendo imagens coloridas, diferentemente das imagens médicas aqui utilizadas. Além disso, as imagens da ImageNet contêm três canais de cores (RGB) e, consequentemente, nossas imagens que estão em tons de cinza devem ser remodeladas para estar de acordo com esta restrição [103].

Utilizamos uma variedade de modelos de Redes Neurais Convolucionais já existentes na literatura e que apresentaram eficácias satisfatórias para tarefas de classificação similares [31, 46]. Os modelos experimentados foram DenseNet [38], EfficientNet [93], Inception [90, 91], ResNet [34] e VGG [85].

Na Tabela 6.3, apresentamos os resultados obtidos em cada modelo testado. Todos os modelos obtiveram eficácia similar, apresentando valores de coeficiente kappa entre 0,6450

e 0,6671.

Modelos	Acurácia	Acurácia Balanceada	AUC ROC	Kappa
DenseNet-121	0,8365	0,8279	0,8874	0,6649
DenseNet-161	**0,8374**	**0,8293**	**0,8892**	**0,6671**
EfficientNet-B7	0,8274	0,8182	0,8873	0,6458
Inception-v3	0,8290	0,8197	0,8769	0,6491
Inception-v4	0,8315	0,8226	0,8783	0,6546
Inception-ResNet-v2	0,8324	0,8229	0,8864	0,6559
ResNet-18	0,8274	0,8175	0,8617	0,6453
ResNet-152	0,8299	0,8220	0,8780	0,6519
VGG-16	0,8357	0,8254	0,8877	0,6621
VGG-19	0,8282	0,8155	0,8840	0,6450

Tabela 6.3: Desempenho obtido em cada modelo experimentado. A rede DenseNet-161 apresentou os melhores resultados nas métricas avaliadas.

Experimento III: Transferência de Aprendizado a Partir de Outra Tarefa com Dados Médicos

Um possível problema com a abordagem do experimento anterior de utilizar apenas redes pré-treinadas na base ImageNet é a otimização dos hiper-parâmetros da rede para realizar a classificação, pois, como mencionamos anteriormente, o conjunto de dados ImageNet é substancialmente diferente do conjunto de dados MURA.

Para reduzir a discrepância entre as amostras da ImageNet (que contêm, por exemplo, veículos, objetos e animais) e MURA (compostas por imagens de raio-X), experimentamos treinar um modelo usando o conjunto de dados CheXpert [39] como um caminho intermediário e, em seguida, ajustar esse modelo para o conjunto de dados MURA.

Como o objetivo neste experimento foi apenas atualizar os parâmetros de uma rede com imagens mais próximas daquelas da base MURA e não tínhamos o propósito de classificar as patologias torácicas presentes na base CheXpert, adaptamos e simplificamos a base para uma tarefa cujo objetivo é classificar apenas o tipo de visualização da amostra (se lateral ou frontal) e se há alguma anormalidade presente.

Utilizamos a rede DenseNet-161, que obteve os melhores resultados no experimento anterior (Experimento II), para este experimento. Na Tabela 6.4, apresentamos os resultados obtidos neste experimento. O pré-treinamento usando um conjunto intermediário com imagens de raio-X não melhorou o desempenho do modelo.

Experimento IV: Aumentação de Dados

A técnica de aumentação de dados eleva a capacidade de generalização de um modelo, expondo-o a um grande número e a uma variedade de cenários. Neste experimento,

Pré-treinamento	Acurácia	Acurácia Balanceada	AUC ROC	Kappa
ImageNet	**0,8374**	**0,8293**	0,8892	**0,6671**
mageNet+CheXpert	0,8332	0,8235	**0,8906**	0,6574

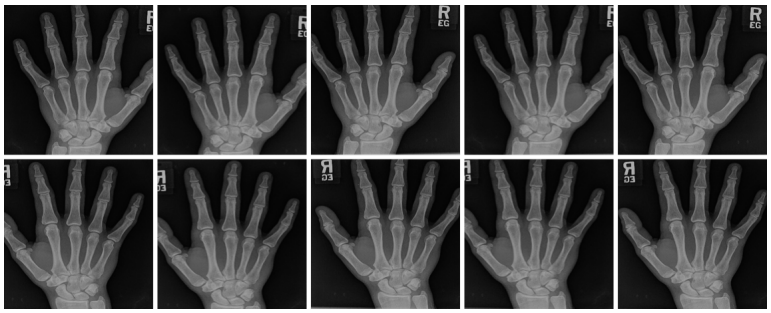
Tabela 6.4: Desempenho da rede DenseNet-161. O pré-treinamento na base ImageNet utiliza os pesos padrões da ImageNet fornecidos pela biblioteca PyTorch. O pré-treinamento na base CheXpert utiliza inicialmente os pesos da ImageNet e, então, o modelo é treinado usando o conjunto de dados CheXpert. Destacamos (em **negrito**) os melhores valores para cada métrica.

aplicamos duas técnicas de aumentação de dados: recortes e inversões horizontais.

Para verificar o efeito dessas aumentações, treinamos a rede DenseNet-161 usando as imagens originais da base MURA (Figura 6.2a), empregando apenas a técnica de inversões horizontais, dobrando o número de amostras de imagens (Aumentação A na Figura 6.2b) e utilizando aumentação com cinco recortes aleatórios nas imagens e suas inversões, resultando em um aumento de 12 vezes no tamanho da base (Aumentação B na Figura 6.2c), em um processo similar à aumentação proposta por Krizhevsky et al. [53].

(a) Imagem original

(b) Aumentação A. Inversão horizontal da imagem original.



(c) Aumentação B. Na primeira linha os recortes feitos na imagem original, na segunda linha as inversões horizontais destes.

Figura 6.2: Exemplo da aumentação de dados realizada em (a) uma amostra da base MURA; (b) na Aumentação A, apenas uma amostra foi criada, dobrando o número de imagens; (c) na Aumentação B, cinco recortes e suas respectivas inversões foram feitas e criadas, aumentando em 12 vezes o tamanho da base.

Na Tabela 6.5, apresentamos um comparativo dos resultados obtidos nos experimentos realizados.

Base	Acurácia	Acurácia Balanceada	AUC ROC	Kappa
Original	0,8374	0,8293	0,8892	0,6671
Aumentação A	**0,8465**	**0,8370**	**0,8929**	**0,6848**
Aumentação B	0,8432	0,8355	0,8922	0,6792

Tabela 6.5: Comparação da eficácia da rede para cada aumentação de dados proposta e para as imagens originais da base MURA, em que Aumentação A é o conjunto de dados usando apenas a técnica de inversão horizontal e Aumentação B é o conjunto de dados usando tanto a técnica de inversão horizontal quanto recorte. Destacamos (em **negrito**) os melhores valores para cada métrica.

A partir desses resultados, concluímos que, apesar de possuir mais dados, a base Aumentação B apresenta uma eficácia pior do que a Aumentação A, ainda sendo melhor do que a original. Além disso, como o tempo de treinamento é aproximadamente linear no número de amostras, o uso da base Aumentação B resulta em um tempo de treinamento maior. Portanto, treinamos outras redes usando a base Aumentação A. É importante destacar que a técnica de Parada Antecipada (do inglês, *Early Stopping*) foi empregada durante os treinamentos. Dessa forma, comparamos os melhores desempenhos obtidos para as duas aumentações.

Na Tabela 6.6, apresentamos os resultados obtidos por outras cinco redes treinadas nesta base. Essas redes foram escolhidas com base nos resultados apresentados na Tabela 6.3, em que optamos por evitar mais de uma rede da mesma categoria.

Modelo	Acurácia	Acurácia Balanceada	AUC ROC	Kappa
DenseNet-161	**0,8465**	**0,8370**	**0,8929**	**0,6848**
EfficientNet-B7	0,8399	0,8314	0,8913	0,6719
Inception-ResNet-v2	0,8457	0,8369	0,8892	0,6836
VGG-16	0,8374	0,8302	0,8967	0,6676
ResNet-152	0,8349	0,8243	0,8931	0,6602

Tabela 6.6: Eficácia de cada modelo treinado utilizando o conjunto de dados Aumentação A.

Experimento V: Modelos de *Ensemble*

Um modelo de *ensemble* combina vários modelos para realizar uma predição final, como uma forma de consultar opiniões variadas para uma amostra. Isso pode melhorar os resultados do modelo, pois, se um único modelo tiver uma eficácia ruim para uma amostra, sua

predição pode ser substituída pela predição de outros modelos, melhorando a estabilidade de uma forma geral.

Para este experimento, utilizamos os cinco modelos de redes neurais convolucionais do experimento anterior (Experimento IV), combinando suas predições utilizando diferentes métodos e estratégias de *ensemble* para melhorar os resultados da tarefa de classificação para as imagens da base MURA.

Um dos métodos de *ensemble* que experimentamos foi baseado em um consenso entre as predições dos modelos utilizados. Na Figura 6.3, ilustramos esse método, em que a predição final baseia-se na moda das predições dos modelos, ignorando a distribuição probabilística.

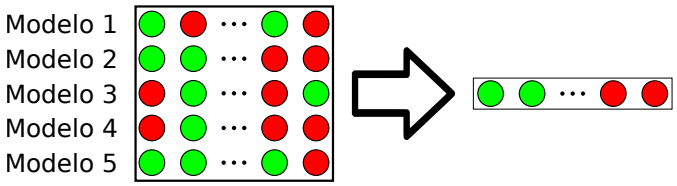


Figura 6.3: Ilustração da predição final do modelo de *ensemble* baseado em consenso. Cada modelo fornece uma predição e a moda dessas predições é considerada a predição final.

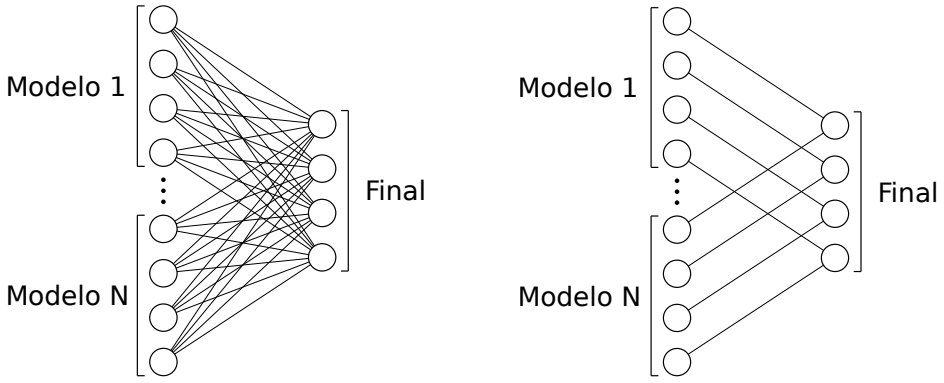
Em outro método de *ensemble* experimentado, combinamos as saídas de probabilidade de cada modelo, usando uma média ponderada em que definimos um peso para cada um dos cinco modelos utilizados.

Também experimentamos um método de *ensemble* baseado em rede neural. Utilizamos duas redes neurais simples formadas por duas camadas lineares, cuja diferença entre ambas é apenas o formato da conectividade entre as camadas. Em um dos modelos (Figura 6.4a), utilizamos a ideia de camadas totalmente conectadas, ou seja, todos os neurônios de uma camada estão conectados a todos os neurônios da camada seguinte, enquanto no outro modelo (Figura 6.4b), utilizamos uma conectividade esparsa entre as camadas. Ambos os modelos recebem como entrada a concatenação das saídas de probabilidades das redes convolucionais utilizadas.

Por fim, experimentamos um método de *ensemble* utilizando as Máquinas de Vetores de Suporte, um modelo clássico de aprendizado de máquina, que também recebe como entrada a concatenação das saídas de probabilidades das redes convolucionais utilizadas para realizar a predição final.

Os resultados obtidos em cada experimento estão apresentados na Tabela 6.7. Comparando os resultados desses experimentos com a média dos resultados obtidos pelos modelos individualmente, pode-se observar que todos os modelos de *ensemble* conseguiram uma eficácia maior do que a obtida pelos modelos individualmente.

Os modelos utilizados no Experimento IV apresentam o problema de sobreajuste (*overfitting*), levando a quase 100% de acurácia no conjunto de treinamento em todos os modelos, tornando impraticável encontrar a melhor maneira de combinar essas predições usando o conjunto de treinamento, já que qualquer combinação levaria a uma acurácia próxima de 100%. Dessa forma, devemos então ajustar os parâmetros dos modelos de *ensemble* utilizando o conjunto de validação, o que deve, inevitavelmente, levar a uma



(a) Arquitetura de rede neural totalmente conectada

(b) Arquitetura de rede neural com conectividade esparsa

Figura 6.4: Representação das duas arquiteturas de rede neural usadas neste experimento de *ensemble*.

Modelo	**Acurácia**	**Acurácia Balanceada**	**AUC ROC**	**Kappa**
Modelo Individual (média)	0,8409	0,8319	0,8926	0,6736
Modelo Individual (máximo)	0,8465	0,8370	0,8929	0,6848
Consenso	0,8449	0,8349	0,8753	0,6811
Média Ponderada	0,8641	0,8572	0,9086	0,7222
Esparsamente Conectada	0,8590	0,8504	0,9039	0,7110
Totalmente Conectada	0,8540	0,8467	0,9102	0,7015
SVM	**0,8699**	**0,8638**	**0,9208**	**0,7345**

Tabela 6.7: Comparação da eficácia dos modelos utilizados. *Modelo Individual* (média) é o desempenho médio dos modelos no Experimento IV, usado como referência. O modelo *Consenso* não produz uma distribuição de probabilidade, apenas uma predição final, o que penaliza seu resultado na métrica AUC ROC. Destacamos (em **negrito**) os melhores valores para cada métrica.

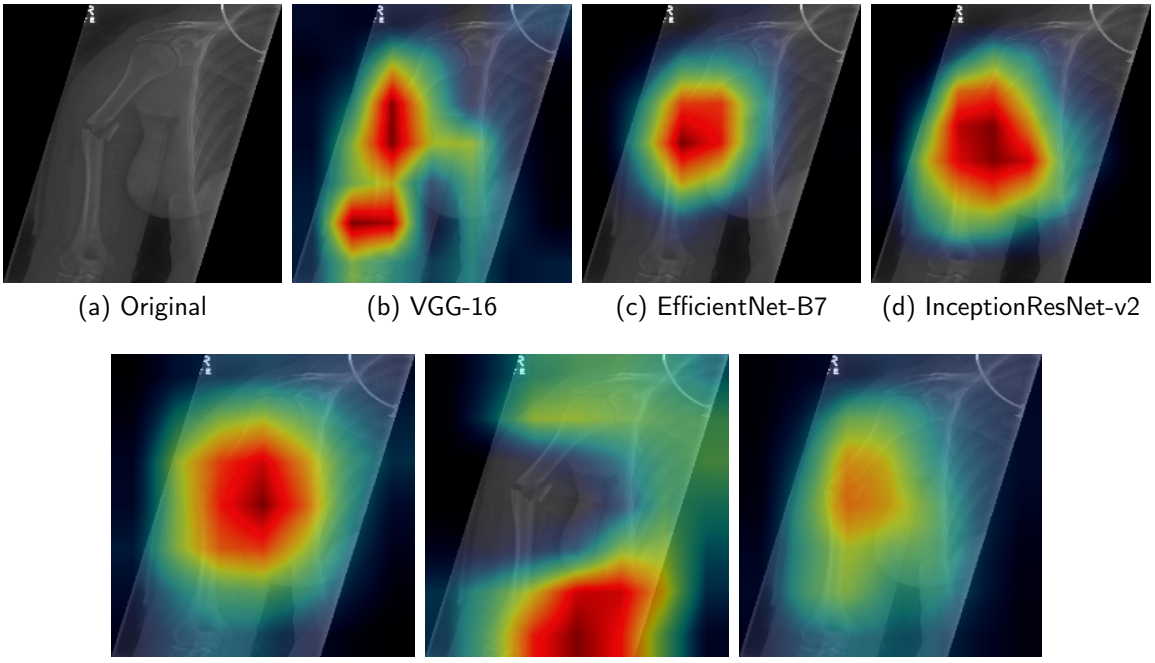
queda no desempenho ao avaliarmos o conjunto de teste.

6.2.1 Mapas de Ativação dos Modelos

O *Gradient-weighted Class Activation Mapping* (Grad-CAM) [68, 82] é um método para visualizar áreas da imagem de entrada que são relevantes para o modelo decidir uma predição de saída. O funcionamento do Grad-CAM ocorre pelo processamento da entrada por meio das camadas da rede em que os gradientes de cada neurônio são definidos como zero, exceto para a classe predita, e retropropagados até a última camada convolucional para obter as regiões dessa camada que mais influenciaram a predição. Considerando que as camadas convolucionais preservam as informações espaciais, essas regiões podem ser extrapoladas para a camada de entrada e convertidas em um mapa de calor.

Os mapas de calor gerados pelo Grad-CAM para cada modelo individualmente e o

modelo de *ensemble* são apresentados na Figura 6.5, o que demonstra a melhoria proporcionada por ele. Neste exemplo, as redes VGG-16, EfficientNet-B7 e InceptionResNet-v2 predisseram "Úmero anormal", enquanto as redes DenseNet-161 e ResNet-152 predisseram "Úmero normal" e a predição do modelo de *ensemble* SVM foi, corretamente, "Úmero anormal". Pode-se observar que os modelos com predições incorretas focalizaram uma parte diferente da imagem, entretanto, o foco no *ensemble* ocorre na parte mais relevante da imagem.



(a) Original (b) VGG-16 (c) EfficientNet-B7 (d) InceptionResNet-v2

(e) DenseNet-161 (f) ResNet-152 (g) *Ensemble*

Figura 6.5: Mapas de calor Grad-CAM para cada modelo individualmente e o modelo de *ensemble* para uma amostra de úmero anormal. Os modelos VGG-16, EfficientNet-B7 e InceptionResNet-v2 predisseram o resultado corretamente, enquanto a DenseNet-161 e ResNet-152 predisseram o resultado incorretamente. O modelo de *ensemble* SVM leva em consideração todos os modelos e produz uma predição final correta.

Como o Grad-CAM utiliza as informações das camadas de características para construir uma visualização e os modelos de *ensemble* combinam os modelos individuais no nível da camada de saída, construímos uma visualização de *ensemble* realizando a média da matriz gerada pelo Grad-CAM (usada para criar o mapa de calor) para os cinco modelos. Dessa forma, o mapa de calor gerado para o modelo de *ensemble* é uma aproximação baseada nos mapas de calor dos modelos individuais.

6.2.2 Resultados

Para realizar uma classificação final, utilizamos o modelo que obteve os melhores resultados no conjunto de validação nos experimentos descritos anteriormente para avaliar a sua qualidade no conjunto de teste.

Na Tabela 6.8, apresentamos os resultados obtidos utilizando o modelo de *ensemble* SVM do Experimento V, que obteve a melhor eficácia nos dados de validação. Pode-

se observar que o desempenho do modelo diminuiu devido ao sobreajuste. O modelo apresentou a pior eficácia no conjunto "Mão", obtendo um valor de kappa de 0,4717, enquanto a melhor eficácia ocorreu no conjunto "Cotovelo", com um valor de kappa de 0,7921. O valor do coeficiente kappa geral foi de 0,6724.

	Acurácia	**Acurácia Balanceada**	**AUC ROC**	**Kappa**
Cotovelo	0,9038	0,8895	0,9023	0,7921
Dedo	0,8563	0,8558	0,8901	0,6817
Antebraço	0,7922	0,7755	0,8142	0,5578
Mão	0,8090	0,7225	0,7759	0,4717
Úmero	0,8077	0,7976	0,8274	0,6061
Ombro	0,8266	0,8270	0,8778	0,6535
Pulso	0,8835	0,8633	0,9142	0,7393
Média	0,8484	0,8319	0,8791	0,6724

Tabela 6.8: Eficácia do modelo de *ensemble* SVM em cada região do corpo. Apresentamos os valores de acurácia, acurácia balanceada, AUC ROC e coeficiente kappa.

A diferença de eficácia do modelo entre os conjuntos de teste e validação se dá devido ao ajuste de parâmetros do modelo de *ensemble* ser feito utilizando o próprio conjunto de validação, como descrito no Experimento V. Pode-se verificar esse fato executando o conjunto de teste no modelo de *ensemble* baseado em consenso, já que este não possui nenhum parâmetro extra para ser ajustado.

Na Tabela 6.9, apresentamos os resultados deste experimento. Os resultados são semelhantes aos obtidos no conjunto de validação, em que, apesar do modelo de *ensemble* baseado em consenso ter apresentado a pior eficácia (Tabela 6.7), obteve uma eficácia melhor nesta avaliação final do que o modelo de *ensemble* SVM, que alcançou a melhor eficácia nos experimentos (Experimento V).

Acurácia	**Acurácia Balanceada**	**AUC ROC**	**Kappa**
0,8593	0,8339	0,8652	0,6899

Tabela 6.9: Eficácia do modelo de *ensemble* baseado em consenso. O modelo apresenta eficácia bastante similar à da Tabela 6.7.

Portanto, podemos supor que, sendo possível treinar os modelos citados no Experimento IV de uma forma que reduza o sobreajuste, os modelos de *ensemble* do Experimento V teriam uma eficácia mais próxima entre a validação e o teste.

6.2.3 Considerações sobre os Experimentos

A tarefa de criar modelos de aprendizado de máquina para auxílio ao diagnóstico médico apresenta certas particularidades e o objetivo deste capítulo foi o de propor e explorar essas dificuldades que ocorrem na tarefa de classificação com aprendizado de máquina utilizando redes neurais convolucionais.

Experimentos indicaram que muitas das técnicas que deveriam melhorar a eficácia na classificação acabaram prejudicando-a. A melhor configuração encontrada foi aplicar a técnica de *stretch* nas imagens de entrada para que tivessem dimensões quadradas, empregar a aumentação de dados com inversões horizontais e utilizar um modelo de *ensemble* para agrupar uma variedade de arquiteturas de rede, atingindo valores de AUC ROC de 0,8791 e coeficiente kappa de 0,6724. A transferência de aprendizado de uma tarefa similar não melhorou os resultados em nossos experimentos.

6.3 Resultados Utilizando o Método Multimodal Proposto

Para avaliar o método proposto na tarefa de classificação, utilizamos os modelos descritos no Capítulo 5 (DenseNet-169, BERT, Fusão) nas imagens da base de dados MURA e nas correspondentes legendas artificiais que foram geradas (Seção 5.2). A avaliação dos modelos foi realizada computando as métricas acurácia balanceada e coeficiente kappa em cada amostra de teste.

Como um estudo de caso na base MURA pode possuir uma ou mais amostras de imagens, os modelos propostos recebem como entrada um estudo de caso e levam em consideração cada amostra do mesmo. Em cada amostra, os modelos predizem a probabilidade de anormalidade, resultando na predição binária "anormal" se a probabilidade de anormalidade para a amostra for maior do que $0,5$.

Ao final, o método proposto realiza um esquema de votação entre os três classificadores mencionados anteriormente (DenseNet-169, BERT e Fusão), em que a predição final é escolhida a partir da classe com maior probabilidade na predição entre os três classificadores. Este processo é ilustrado na Figura 6.6, em que, de forma similar ao modelo de consenso do Experimento V (Seção 6.2), a predição final do método se baseia na classe mais predita entre os classificadores de imagem, texto e de fusão.

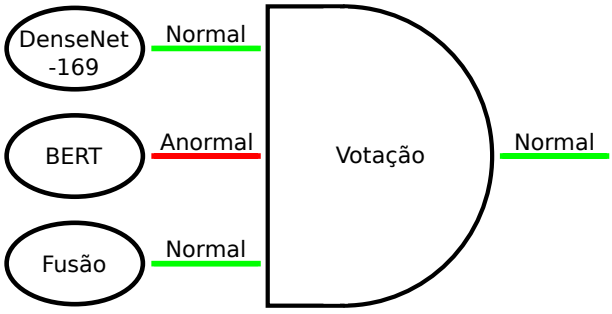


Figura 6.6: Ilustração da predição final do método proposto baseado nas classificações dos modelos DenseNet-169, BERT e Fusão.

Para avaliar a qualidade do método proposto como um todo, comparamos os resultados dos classificadores de imagem, texto e fusão individualmente com o método final proposto. Na Tabela 6.10, apresentamos os resultados de acurácia balanceada e coeficiente kappa obtidos a partir de cada modelo para a classificação de imagens e textos utilizando os modelos com DenseNet-169 e BERT, respectivamente, bem como os resultados obtidos utilizando o modulo de fusão. Finalmente, os resultados obtidos com o método proposto são também apresentados.

	DenseNet-169		**BERT**		**Fusão**		**Método**	
	Acurácia Balanceada	**kappa**	**Acurácia Balanceada**	**kappa**	**Acurácia Balanceada**	**kappa**	**Acurácia Balanceada**	**Kappa**
Cotovelo	0,8537	0,7214	0,6322	0,2562	**0,8731**	**0,7512**	0,8634	0,7364
Dedo	0,8222	0,6521	0,6747	0,3446	**0,8300**	**0,6651**	0,8294	0,6647
Antebraço	0,8292	0,6652	0,5354	0,0705	0,8471	0,6974	**0,8526**	**0,7114**
Mão	0,7680	0,5722	0,4578	-0,0882	**0,7835**	**0,5935**	0,7707	0,5745
Úmero	0,9038	0,8074	0,5868	0,1730	0,9039	0,8075	**0,9113**	**0,8222**
Ombro	0,7724	0,5456	0,5473	0,0934	0,7737	0,5467	**0,7840**	**0,5673**
Pulso	0,8584	0,7390	0,5997	0,1973	**0,8806**	**0,7772**	0,8738	0,7672
Média	0,8296	0,6718	0,5762	0,1495	**0,8417**	0,6912	0,8407	**0,6920**

Tabela 6.10: Eficácia dos modelos DenseNet-169, BERT, Fusão e o método proposto, que aplica uma votação entre os três classificadores. Os valores de acurácia balanceada e coeficiente kappa são apresentados. Destacamos (**negrito**) os melhores e os segundo melhores (sublinhado) valores das métricas em cada tipo de estudo e na média.

Para os estudos "Cotovelo" e "Pulso", a eficácia do método proposto é inferior à do classificador de fusão e ambos as eficácias são melhores do que os resultados da DenseNet-169 e do BERT, individualmente. Para o estudo "Dedo", a eficácia do método final é comparável à eficácia da fusão, que apresenta os melhores resultados.

A eficácia do método proposto apresenta os melhores resultados para os estudos "Ombro", "Úmero" e "Antebraço". Neste último, em particular, o método apresentou o maior ganho em comparação com o modelo DenseNet-169. No estudo "Mão", o modelo BERT apresenta um valor de kappa negativo, indicando que não há concordância em seus resultados e a eficácia do método final é comparável à eficácia da DenseNet-169 individualmente, não apresentando um ganho relevante. Neste estudo, o classificador de fusão apresenta os melhores resultados.

O modelo BERT, conforme esperado, devido ao formato das legendas geradas automaticamente, em que, em sua maioria, contêm textos com baixa qualidade ou sem informações relevantes, apresenta os piores resultados em todos os estudos.

Em comparação com os resultados do modelo DenseNet-169, utilizados como referência (*baseline*), o método proposto apresentou um ganho nos valores das métricas analisadas. Dessa forma, de uma maneira geral, o método foi capaz de extrair e combinar características discriminativas e relevantes de cada modalidade utilizada (imagens e textos). O

método obteve uma média de acurácia balanceada e coeficiente kappa de 0,8407 e 0,6920, respectivamente.

Também comparamos os resultados obtidos por nosso método proposto com outro método existente para classificação de anormalidades musculoesqueléticas [73] utilizando a base de dados MURA sob o mesmo conjunto de teste definido no Capítulo 4.

A Tabela 6.11 reporta os valores de acurácia obtidos em cada método. A coluna "Imagem" está relacionada à tarefa de classificação de imagens, a coluna "Texto" está associada à tarefa de classificação de texto, enquanto a última coluna está relacionada ao método proposto. Nosso método obtém um resultado melhor para a classificação de anormalidades musculoesqueléticas no conjunto de dados MURA, apresentando uma acurácia de 0,8511, enquanto o método de Pelka et al. [73] apresenta um valor de 0,8155.

Tabela 6.11: Comparação com outra abordagem no conjunto de dados MURA. Reportamos o valor de acurácia para cada método. O melhor resultado está destacado (**negrito**).

Method	**Imagem**	**Texto**	**Método**
Pelka et al. [73]	0,7985	0,5424	0,8155
Método proposto	0,8418	0,5807	**0,8511**

Nosso trabalho se difere do método de Pelka et al. [73] ao propormos a utilização de sentenças completas geradas artificialmente para o conjunto de dados MURA. Já Pelka et al. [73] realizaram a geração e a utilização apenas de palavras-chave. Além disso, o nosso método multimodal recebe uma fusão de características extraídas dos modelos de imagens e de textos, enquanto Pelka et al. [73] propuseram um método que insere as palavras-chave geradas nas próprias imagens, como um tipo de ruído.

Trabalhos recentes também apresentam uma metodologia similar à proposta neste trabalho. Um exemplo é o trabalho Zhang et al. [106], em que os autores apresentaram um método que também visa à colaboração entre módulos de imagem e texto, porém utilizando uma abordagem não supervisionada.

6.3.1 Discussão Final

Analisamos amostras específicas do conjunto de teste com a predição de cada modelo utilizado, conforme ilustrado na Figura 6.7. Nos dois primeiros exemplos, todos os modelos (de imagem, texto e fusão) foram capazes de classificar corretamente a imagem e o texto de entrada. No primeiro, a legenda gerada possui uma informação incorreta sobre a parte do corpo apresentada e, no segundo, apesar da legenda inteira não possuir informações relevantes, ela contém a localização correta da parte do corpo e a palavra "normal". Para ambos os casos, o modelo BERT detectou corretamente as informações relevantes e foi capaz de descartar elementos espúrios presentes em ambas as legendas. O modelo DenseNet-169 também conseguiu identificar regiões relevantes nas imagens.

No terceiro exemplo, a legenda gerada não contém absolutamente nenhuma informação semântica ou minimamente relevante. O modelo BERT, como esperado, incorretamente rotulou neste exemplo a anormalidade como negativa. Por outro lado, o modelo

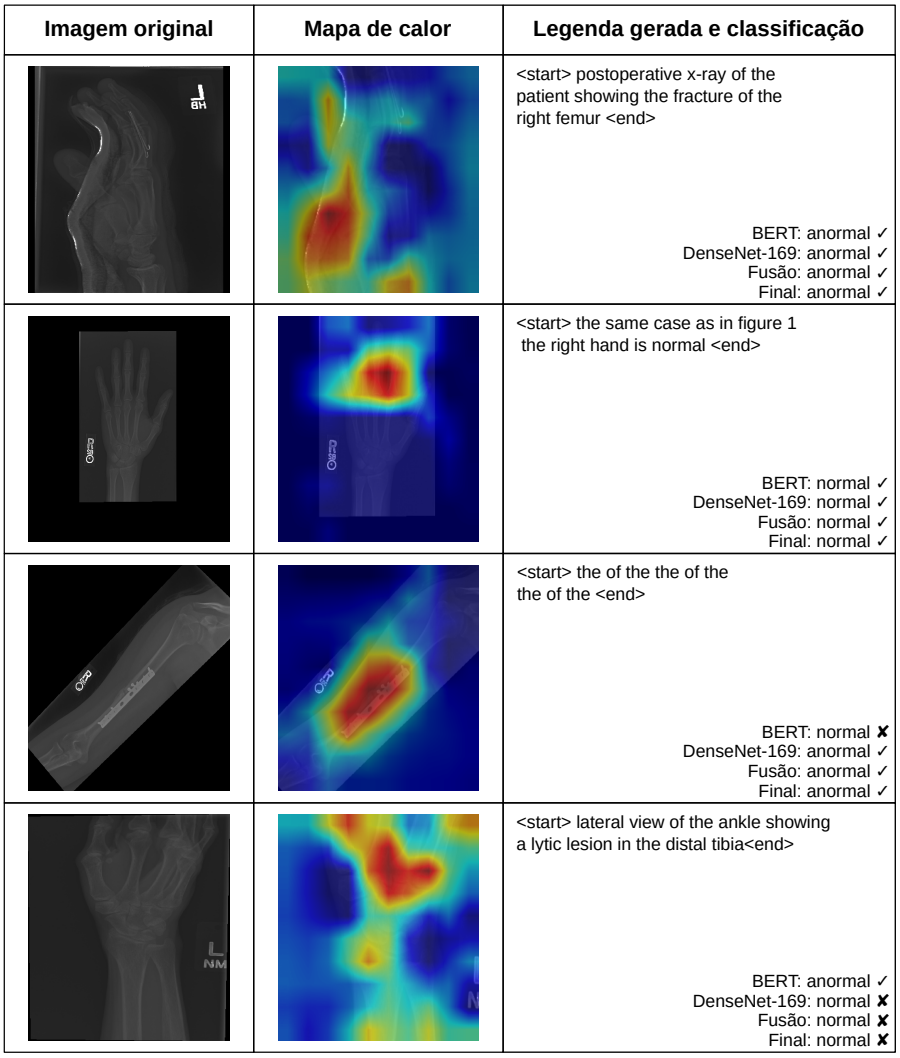


Figura 6.7: Imagem original, mapa de calor da DenseNet-169 e legenda gerada artificialmente em amostras do conjunto de teste. Um marcador (✓) indica a predição correta de cada método.

da DenseNet-169 rotulou corretamente a anormalidade como positiva com uma alta confiança. É possível observar, neste exemplo, uma área bem definida destacada no mapa de calor da radiografia, que corresponde à área com mais ativações na rede, isto é, a área considerada mais importante pela rede para realizar a predição.

No último exemplo, enquanto o modelo DenseNet-169 não foi capaz de identificar a presença de anormalidades na imagem, em que nem mesmo conseguiu produzir um mapa de calor bem definido, o modelo BERT foi capaz de distinguir informações relevantes da legenda e classificar corretamente o exemplo como “anormal”.

A visualização dos mapas de calor gerados pelos modelos preditivos pode auxiliar na tarefa de interpretabilidade desses modelos, fornecendo propriedades visuais que possam ser compreendidas por um especialista que pode tomar decisões a partir destas informações e, assim, fornecer uma confiabilidade maior para as predições.

Capítulo 7

Conclusões e Trabalhos Futuros

Neste capítulo, concluímos esta dissertação apresentando alguns comentários finais sobre o trabalho de pesquisa realizado. Descrevemos também algumas sugestões de trabalhos futuros que podem ser investigados a partir desta pesquisa.

7.1 Conclusões

Neste trabalho, investigamos duas abordagens para a classificação de imagens musculoesqueléticas, cujo objetivo foi detectar anormalidades nas imagens de raio-X, sendo elas com métodos explorando o uso de redes neurais convolucionais e métodos com uma abordagem multimodal com imagens e textos.

Em relação ao método que utiliza redes neurais convolucionais, combinamos técnicas clássicas de aprendizado de máquina e visão computacional, utilizamos etapas de pré-processamento dos dados e integramos tanto classificadores baseados em redes neurais quanto Máquina de Vetores de Suporte e avaliamos a eficácia de cada técnica aplicada, buscando, de forma incremental, construir uma arquitetura sólida que melhorasse os resultados da classificação final.

Para a classificação utilizando uma abordagem multimodal, o primeiro desafio foi a construção de um modelo para gerar legendas a partir dos dados de imagens médicas, em que notamos uma escassez deste tipo de dado, principalmente relacionado a imagens de fraturas ósseas. Após esta etapa, investigamos a utilização de uma abordagem multimodal, em que buscamos combinar as imagens com textos gerados automaticamente. Os nossos resultados mostraram que esta abordagem é capaz de produzir bons resultados, superando aqueles obtidos por meio de imagens e textos individualmente, para a tarefa de classificação proposta. O ponto negativo deste método está relacionado à dependência de um bom gerador de legendas para que haja dados textuais com qualidade para auxiliar o classificador multimodal.

Ao final desta pesquisa, recapitulamos as principais descobertas e procuramos responder às questões de pesquisa introduzidas no Capítulo 1:

- **QP1: Qual é o impacto de diferentes formas de *ensemble* no processo de classificação?** Como apresentado por meio dos resultados da Tabela 6.7, o

ensemble de classificadores utilizando SVM apresentou resultados promissores durante os experimentos realizados, entretanto, esta técnica possui pontos negativos relacionados ao alto sobreajuste do modelo. As Tabelas 6.9 e 6.10 apresentam o uso de consenso (ou votação) como tendo impacto positivo nos resultados.

- **A utilização de uma abordagem multimodal, no contexto de dados médicos, pode produzir resultados melhores aos resultados utilizando apenas imagens e textos?** Como descrito na Seção 6.3 e apresentado nos resultados da Tabela 6.10, a utilização de uma abordagem multimodal apresentou um ganho nos valores das métricas analisadas em comparação à eficácia das abordagens individuais.

- **QP3: É possível gerar, com boa qualidade, dados textuais de legendas artificiais para imagens de radiografias musculoesqueléticas?** Muitos trabalhos demonstraram bons resultados para a tarefa de legendagem de imagens [99,104] em geral. Entretanto, ao analisar alguns trabalhos que abordam esta tarefa [26, 41] no contexto de imagens médicas, constatamos que há uma carência de conjuntos de dados anotados, com os mais variados tipos de dados médicos.

7.2 Trabalhos Futuros

Na abordagem utilizando redes neurais convolucionais, um dos grandes obstáculos observados durante os experimentos realizados foi o sobreajuste (*overfitting*) apresentado pelos modelos. Como próximos passos a serem seguidos, podemos experimentar diferentes abordagens e técnicas para prevenir o sobreajuste no treinamento, como a utilização de regularização, além de analisar a abordagem proposta em conjuntos de dados médicos ainda maiores e avaliar sua eficácia.

Como próximos passos em relação ao método proposto baseado em uma abordagem multimodal, podemos destacar etapas de geração de legendas e fusão. Para a geração de legendas, podemos explorar o uso de representações textuais específicas de dados musculoesqueléticos, além de explorar modelos de geração de legendas baseados em *transformers* [97]. Na etapa de fusão, podemos explorar outros formatos além dos que já foram experimentados neste trabalho, visando combinar características das diferentes modalidades.

Como há diversas tarefas e uma grande variedade de tipos de dados médicos, o trabalho apresentado possui um grande potencial para ser estendido para outras especialidades médicas, combinando os metadados com outras modalidades existentes de imagens.

Referências Bibliográficas

[1] D. G. d. Aguiar Neto. BugBERT: Triagem e Rotulação de Relatórios de Erros Usando Transferência de Aprendizado em Redes Baseadas em Transformer. *Dissertação de Mestrado - Universidade Estadual de Campinas, Instituto de Computação*, 2020.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.

[3] M. Annarumma, S. J. Withey, R. J. Bakewell, E. Pesce, V. Goh, and G. Montana. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology*, 291(1):196–202, 2019.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[5] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González. Gated Multimodal Units for Information Fusion. *arXiv:1702.01992*, 2017.

[6] G. Argenziano, H. P. Soyer, V. De Giorgi, D. Piccolo, P. Carli, and M. Delfino. Interactive Atlas of Dermoscopy. *EDRA Medical Publishing & New Media*, 2000.

[7] Y. Arzhaeva, D. Wang, L. Devnath, S. Amirgholipour, R. McBean, J. Hillhouse, S. Luo, D. Meredith, K. Newbigin, and D. Yates. Development of Automated Diagnostic Tools for Pneumoconiosis Detection from Chest X-Ray Radiographs. *Coal Services Health and Safety Trust Project No. 20647 CSIRO Report No. EP192938*, 2019.

[8] D. H. Ballard and C. M. Brown. *Computer Vision*. Prentice Hall Professional Technical Reference, 1st edition, 1982.

[9] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv preprint arXiv:1903.10676*, 2019.

[10] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank. Automatic Description Generation from Images: A Survey

of Models, Datasets, and Evaluation Measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016.

[11] S. Brahimi, N. B. Aoun, and C. B. Amar. Boosted Convolutional Neural Network for object recognition at large scale. *Neurocomputing*, 330:337–354, 2019.

[12] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The Balanced Accuracy and its Posterior Distribution. In *20th International Conference on Pattern Recognition (ICPR)*, pages 3121–3124, 2010.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*, 2020.

[14] R. Cadene. Pretrained Models for PyTorch. `https://github.com/Cadene/pretrained-models.pytorch`, 2017.

[15] M. C. Carvalho. Esquemas de Transferência para Aprendizado Profundo em Classificação de Imagens. *Dissertação de Mestrado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação*, 2015.

[16] B. Chen, J. Li, X. Guo, and G. Lu. DualCheXNet: Dual Asymmetric Feature Learning for Thoracic Disease Classification in Chest X-rays. *Biomedical Signal Processing and Control*, 53:101554, 2019.

[17] M. Y. M. Chen, T. L. Pope, and D. J. Ott. *Radiologia Básica*. AMGH, 2012.

[18] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating Long Sequences with Sparse Transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[19] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] A. D. Dolatabadi, S. E. Z. Khadem, and B. M. Asl. Automated Diagnosis of Coronary Artery Disease (CAD) Patients using Optimized SVM. *Computer Methods and Programs in Biomedicine*, 138:117–126, 2017.

[23] D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. *arXiv:1710.07177*, 2017.

[24] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2121–2129, 2013.

[25] W. Gale, L. Oakden-Rayner, G. Carneiro, A. Bradley, and L. Palmer. Detecting Hip Fractures with Radiologist-level Performance using Deep Neural Networks. *arXiv:1711.06504*, 2017.

[26] W. Gale, L. Oakden-Rayner, G. Carneiro, L. J. Palmer, and A. P. Bradley. Producing Radiologist-Quality Reports for Interpretable Deep Learning. In *16th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1275–1279, 2019.

[27] H. Garud, S. P. K. Karri, D. Sheet, J. Chatterjee, M. Mahadevappa, A. K. Ray, A. Ghosh, and A. K. Maity. High-magnification Multi-views Based Classification of Breast Fine Needle Aspiration Cytology Cell Samples using Fusion of Decisions from Deep Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 76–81, 2017.

[28] R. Girshick. Fast R-CNN. In *15th International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[29] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[30] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern. Skin Lesion Analysis toward Melanoma Detection: A Challenge. In *International Symposium on Biomedical Imaging (ISBI)*, 2016.

[31] M. Habibzadeh, M. Jannesari, Z. Rezaei, H. Baharvand, and M. Totonchi. Automatic White Blood Cell Classification using Pre-trained Deep Learning Models: ResNet and Inception. In *10th International Conference on Machine Vision (ICMV)*, volume 10696, page 1069612. International Society for Optics and Photonics, 2017.

[32] S. Haykin. *Redes Neurais: Princípios e Prática*. Bookman Editora, 2007.

[33] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In *16th International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[34] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[35] W. Herring. *Learning Radiology: Recognizing the Basics*. Elsevier, 2019.

[36] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[37] J. Howard and S. Ruder. Universal Language Model Fine-tuning for Text Classification. *arXiv preprint arXiv:1801.06146*, 2018.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[39] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, and K. Shpanskaya. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *AAAI Conference on Artificial Intelligence*, 33(1):590–597, 2019.

[40] A. Jiménez-Sánchez, A. Kazi, S. Albarqouni, S. Kirchhoff, A. Sträter, P. Biberthaler, D. Mateus, and N. Navab. Weakly-supervised Localization and Classification of Proximal Femur Fractures. *arXiv:1809.10692*, 2018.

[41] B. Jing, P. Xie, and E. P. Xing. On the Automatic Generation of Medical Imaging Reports. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 2577–2586, 2018.

[42] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. MIMIC-CXR-JPG, A Large Publicly Available Database of Labeled Chest Radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[43] R. Johnson and T. Zhang. Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 919–927, 2015.

[44] R. Johnson and T. Zhang. Deep Pyramid Convolutional Neural Networks for Text Categorization. In *55th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 562–570, 2017.

[45] C. K. R. Natural Language Processing. In *Fundamentals of Artificial Intelligence*, pages 603–649. Springer, 2020.

[46] J. Kawahara, A. BenTaieb, and G. Hamarneh. Deep features to classify skin lesions. In *13th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1397–1400, 2016.

[47] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine. Supervised Multimodal Bitransformers for Classifying Images and Text. *arXiv:1909.02950*, 2019.

[48] D. Kiela, E. Grave, A. Joulin, and T. Mikolov. Efficient Large-Scale Multi-Modal Classification. *arXiv:1802.02892*, 2018.

[49] C. Kokkotis, S. Moustakidis, E. Papageorgiou, G. Giakas, and D. E. Tsaopoulos. Machine Learning in Knee Osteoarthritis: A Review. *Osteoarthritis and Cartilage Open*, page 100069, 2020.

[50] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer. Large Scale Deep Learning for Computer Aided Detection of Mammographic Lesions. *Medical Image Analysis*, 35:303–312, 2017.

[51] V. Kougia. Medical Image Labeling and Report Generation. *Master Thesis - Athens University of Economics and Business (AUEB), Department of Informatics*, 2019.

[52] A. Krizhevsky, G. Hinton, et al. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, Canada, 2009.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105. Curran Associates, Inc., 2012.

[54] R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174, 1977.

[55] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[56] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim. Deep Learning in Medical Imaging: General Overview. *Korean Journal of Radiology*, 18(4):570–584, 2017.

[57] Q. Li, W. Li, J. Zhang, and Z. Xu. An Improved k-Nearest Neighbour Method to Diagnose Breast Cancer. *Analyst*, 143(12):2807–2811, 2018.

[58] E. D. Liddy. Natural Language Processing. In *Encyclopedia of Library and Information Science*. Marcel Decker, Inc., 2001.

[59] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A Structured Self-attentive Sentence Embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[60] Z. C. Lipton, J. Berkowitz, and C. Elkan. A Critical Review of Recurrent Neural Networks for Sequence Learning. *arXiv preprint arXiv:1506.00019*, 2015.

[61] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, , B. v. Van Der Laak, Jeroen A. W. M. van der Laak, and C. I. Sánchez. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88, 2017.

[62] X. Liu, T. Fu, Z. Pan, D. Liu, W. Hu, J. Liu, and K. Zhang. Automated Layer Segmentation of Retinal Optical Coherence Tomography Images Using a Deep Feature Enhanced Structured Random Forests Classifier. *IEEE Journal of Biomedical and Health Informatics*, 23(4):1404–1416, 2018.

[63] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang. Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. In *European Conference on Computer Vision (ECCV)*, pages 353–369, 2018.

[64] J. Ma, F. Wu, J. Zhu, D. Xu, and D. Kong. A Pre-trained Convolutional Neural Network based Method for Thyroid Nodule Diagnosis. *Ultrasonics*, 73:221–230, 2017.

[65] E. Marchiori and M. L. Santos. *Introdução À Radiologia*. Guanabara Koogan, 2009.

[66] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir. Deep Learning in Radiology: An Overview of the Concepts and a Survey of the State of the Art. *arXiv preprint arXiv:1802.08717*, 2018.

[67] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle. Knowledge Transfer for Melanoma Screening with Deep Learning. In *14th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 297–300. IEEE, 2017.

[68] K. Nakashima. Grad-CAM with PyTorch. `https://github.com/kazuto1011/grad-cam-pytorch`, 2017.

[69] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning . In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[70] M. Nishio, O. Sugiyama, M. Yakami, S. Ueno, T. Kubo, T. Kuroda, and K. Togashi. Computer-aided Diagnosis of Lung Nodule Classification Between Benign Nodule, Primary Lung Cancer, and Metastatic Lung Cancer at Different Image Size using Deep Convolutional Neural Network with Transfer Learning. *PLOS ONE*, 13(7), 2018.

[71] J. Olczak, N. Fahlberg, A. Maki, A. S. Razavian, A. Jilert, A. Stark, O. Sköldenberg, and M. Gordon. Artificial Intelligence for Analyzing Orthopedic Trauma Radiographs: Deep Learning Algorithms - Are they on Par with Humans for Diagnosing Fractures? *Acta Orthopaedica*, 88(6):581–586, 2017.

[72] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich. Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 180–189. Springer, 2018.

[73] O. Pelka, F. Nensa, and C. M. Friedrich. Branding - Fusion of Meta Data and Musculoskeletal Radiographs for Multi-modal Diagnostic Recognition. *International Conference on Computer Vision Workshop (ICCV)*, pages 467–475, 2019.

[74] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie. MFAS: Multimodal Fusion Architecture Search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6966–6975, 2019.

[75] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. *arXiv preprint arXiv:1712.06957*, 2017.

[76] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *28th Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015.

[77] M. X. Ribeiro. *Suporte a Sistemas de Auxílio ao Diagnóstico e de Recuperação de Imagens por Conteúdo usando Mineração de Regras de Associação*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2008.

[78] F. Rosenblatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6):386, 1958.

[79] S. Ruder. *Neural Transfer Learning forNatural Language Processing*. PhD thesis, NUI Galway, 2019.

[80] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Representations by Back-propagating Errors. *nature*, 323(6088):533–536, 1986.

[81] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[82] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[83] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 512–519, 2014.

[84] D. Shen, Y. Zhang, R. Henao, Q. Su, and L. Carin. Deconvolutional Latent-Variable Model for Text Sequence Matching. *AAAI Conference on Artificial Intelligence*, 32(1):5438–5445, 2018.

[85] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[86] N. S. Singh, S. Hariharan, and M. Gupta. Facial Recognition Using Deep Learning. In *Advances in Data Sciences, Security and Applications*, pages 375–382. Springer, 2020.

[87] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Ng, and M. Lungren. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. *arXiv preprint arXiv:2004.09167*, 2020.

[88] J. E. Solem. *Programming Computer Vision with Python: Tools and Algorithms for Analyzing Images*. O'Reilly Media, Inc., 2012.

[89] E. Stevens, L. Antiga, and T. Viehmann. Deep Learning with PyTorch. *Shelter Island, NY: Manning Publications*, 2020.

[90] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *AAAI Conference on Artificial Intelligence*, 2017.

[91] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *29th Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[92] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2010.

[93] M. Tan and Q. V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:1905.11946*, 2019.

[94] P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, A. Blum, R. P. Braun, H. Cabo, J.-Y. Gourhant, J. Kreusch, A. Lallas, J. Lapins, A. Marghoob, S. Menzies, N. M. Neuber, J. Paoli, H. S. Rabinovitz, C. Rinner, A. Scope, H. P. Soyer, C. Sinz, L. Thomas, I. Zalaudek, and H. Kittler. Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. *JAMA Dermatology*, 155(1):58–65, 2019.

[95] A. C. G. Vargas, A. Paes, and C. N. Vasconcelos. Um Estudo sobre Redes Neurais Convolucionais e sua Aplicação em Detecção de Pedestres. In *29th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–4, 2016.

[96] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, Ł. Kaiser, N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer, and J. Uszkoreit. Tensor2Tensor for Neural Machine Translation. *arXiv preprint arXiv:1803.07416*, 2018.

[97] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017.

[98] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie. CentralNet: a Multilayer Approach for Multimodal Fusion. In *European Conference on Computer Vision (ECCV) Workshops*, pages 1–15, 2018.

[99] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 7(12):3156–3164, June 2015.

[100] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko. A Support Vector Machine-based Ensemble Algorithm for Breast Cancer Diagnosis. *European Journal of Operational Research*, 267(2):687–699, 2018.

[101] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.

[102] J. Weston, S. Bengio, and N. Usunier. WSABIE: Scaling Up To Large Vocabulary Image Annotation. In *22nd International Joint Conference on Artificial Intelligence (IJCAI)*, volume 1, pages 2764–2770, 2011.

[103] Y. Xie and D. Richmond. Pre-training on Grayscale ImageNet Improves Medical Image Classification. In *European Conference on Computer Vision (ECCV) Workshops*, 2018.

[104] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *32nd International Conference on Machine Learning (ICML)*, pages 2048–2057. PMLR, 2015.

[105] T. Zhang, M. Huang, and L. Zhao. Learning Structured Representation for Text Classification via Reinforcement Learning. In *32nd AAAI Conference on Artificial Intelligence*, pages 6053–6060, 2018.

[106] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text, 2020.

[107] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.