UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA
DCET - DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

NATHAN FERRAZ DA SILVA

VIRE
SISTEMA DE RECONHECIMENTO DE EMOÇÕES PARA ACOMPANHAMENTO EMOCIONAL HUMANO

VITÓRIA DA CONQUISTA – BA
2024

NATHAN FERRAZ DA SILVA

VIRE
SISTEMA DE RECONHECIMENTO DE EMOÇÕES PARA ACOMPANHAMENTO EMOCIONAL HUMANO

Pesquisa entregue à disciplina Trabalho Supervisionado II como requisito para obtenção do Grau de Bacharel em Ciência da Computação pela Universidade Estadual do Sudoeste da Bahia.

Orientador: Prof. Dr. Geraldo Pereira Rocha Filho

VITÓRIA DA CONQUISTA – BA
2024

DEDICATÓRIA

Aos meus pais, que, sob muito sol, me trouxeram até aqui, na sombra.

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Este trabalho tem como objetivo propor um sistema de acompanhamento emocional baseado em redes neurais convolucionais para o reconhecimento de emoções humanas em rostos. A pesquisa baseia-se no princípio da existência de seis emoções básicas propostas por Paul Ekman, segundo o qual é possível identificar cada uma delas a partir da composição do estado de diversos músculos da face. Seguindo essa perspectiva, foi utilizado redes neurais convolucionais, uma vez que esses padrões podem ser observados em imagens e esse tipo de inteligência artificial é especialista em identificar padrões complexos em dados visuais. Para atingir o objetivo, a pesquisa testou o desempenho de três arquiteturas diferentes: AlexNet, DenseNet e uma CNN personalizada. A metodologia envolveu a coleta de dados da base FER2013, o pré-processamento das imagens, busca por hiperparâmetros e o treinamento dos modelos. Com isso, a pesquisa pretende sugerir o modelo com melhor desempenho para compor o sistema VIRE, um sistema para identificação de emoções para pessoas que necessitam de acompanhamento constante. O reconhecimento de emoções tem aplicações importantes em diversas áreas, como segurança e interações humano-computador, e esta pesquisa pretende demonstrar uma aplicação na saúde.

Palavras-chave: redes neurais convolucionais, reconhecimento de emoções, inteligência artificial, redes profundas.

ABSTRACT

This study aims to propose an emotional monitoring system based on convolutional neural networks for recognizing human emotions in faces. The research is based on the principle of the existence of six basic emotions proposed by Paul Ekman, according to which each emotion can be identified from the state of various facial muscles. Following this perspective, convolutional neural networks were used, as these patterns can be observed in images and this type of artificial intelligence specializes in identifying complex patterns in visual data. To achieve the objective, the study tested the performance of three different architectures: AlexNet, DenseNet, and a custom CNN. The methodology involved collecting data from the FER2013 dataset, preprocessing the images, searching for hyperparameters, and training the models. With this, the research aims to suggest the best-performing model to compose the VIRE system, a system for emotion recognition for people who require constant monitoring. Emotion recognition has important applications in various areas, such as security and human-computer interactions, and this research aims to demonstrate an application in healthcare.

Keywords: convolutional neural networks, emotion recognition, artificial intelligence, deep networks.

SUMÁRIO

1. Introdução

Uma emoção pode ser expressas de diversas maneiras. Numa conversa, por exemplo, a elevação no tom ou na altura da voz pode ser interpretada como agressividade e o contrário como calma, a gesticulação do corpo também pode complementar a emoção que está sendo expressa e até mesmo as vestimentas de uma pessoa podem carregar informações sobre seu estado emocional. Contudo, a maioria das pesquisas foca nas expressões faciais para reconhecer emoções (EKMAN, 1971). Isso se deve ao fato de o rosto possuir uma grande quantidade de músculos, especialmente por causa da necessidade de mastigação, o que oferece uma vasta mobilidade em diversas direções. Paul Ekman (1971), defende a existência de emoções básicas que podem ser identificadas a partir da composição dos estados dos músculos do rosto. A expressão da alegria, por exemplo, é comumente caracterizada pela elevação do músculo zigomático maior, que eleva a área entre os lábios e a bochecha, resultando no sorriso. Além disso, as expressões faciais são frequentemente utilizadas como indicadores de estados emocionais em pesquisas sobre saúde mental.

É possível afirmar que há uma predominância do sentimento de tristeza nos estados depressivos uma vez que, “na depressão, a falta de significado de cada empreendimento e de cada emoção, e da própria vida se tornam evidentes e o único sentimento que resta nesse estado despido de amor é a insignificância” (SOLOMON, 2001). Esses sentimentos intensos e persistentes podem fazer parte de um conjunto mais amplo de sintomas de algum transtorno mental. Por sua vez, um transtorno mental “é uma síndrome caracterizada por perturbações significativas na cognição, na regulação emocional ou no comportamento do indivíduo, frequentemente associadas ao sofrimento ou incapacidade que afetam as atividades sociais e profissionais” (AMERICAN PSYCHIATRIC ASSOCIATION et al., 2014).

Diante da complexidade de um diagnóstico, é fundamental dispor de ferramentas que aumentem a confiabilidade no processo diagnóstico. Ao contrário de outras condições em que é possível diagnosticar uma doença através de marcadores ou exames de imagem, a abordagem diagnóstica de transtornos mentais baseia-se principalmente na interpretação clínica de fenômenos humanos (DALGALARRONDO, 2008), além de critérios diagnósticos muitas vezes limitados. Logo um sistema de reconhecimento de emoções apresenta-se então como uma alternativa interessante que pode favorecer o processo diagnóstico.

Nesse sentido, a inteligência artificial é extremamente útil, especialmente as redes neurais convolucionais (CNNs), que são altamente competentes em captar características espaciais em dados visuais. As CNNs funcionam aplicando filtros que detectam bordas,

texturas e outros padrões básicos, que são então combinados em camadas subsequentes para reconhecer características mais complexas (CHOLLET, 2018). A composição de padrões visuais no rosto, como a elevação da sobrancelha ou a posição da boca, pode ser reconhecida por essas redes, já que elas são especializadas em identificar padrões hierárquicos em imagens. Dessa forma, é possível combinar tecnologia e saúde, utilizando o reconhecimento de emoções para identificar estados emocionais que possam indicar condições de saúde que merecem atenção.

Apesar dos avanços significativos alcançados pelas redes neurais convolucionais, elas ainda estão sujeitas a fatores que podem afetar seu treinamento e sua capacidade de interpretação. Modelos treinados para reconhecimento emocional necessitam de um grande volume de dados rotulados, que podem estar sujeitos a erros e vieses, uma vez que essa rotulagem é realizada por humanos (HOSSEIN et al. 2013). Além disso, a grande demanda por dados e a complexidade computacional das redes convolucionais exigem o uso de hardware de alto desempenho. Por fim, entender os processos que levam uma rede neural a determinado resultado é uma atividade extremamente complexa, especialmente quando se trata de redes profundas. Ainda que as CNN's possua essas peculiaridades, elas podem se tornar ferramentas poderosas desde que configuradas adequadamente.

Este trabalho tem como objetivo geral propor um sistema de acompanhamento emocional baseado em redes neurais convolucionais, capaz de analisar expressões faciais. Para alcançar esse objetivo, dois objetivos específicos foram definidos: (1) encontrar um modelo que seja capaz de classificar uma captura facial em uma das seis emoções básicas propostas por Ekman — alegria, raiva, tristeza, medo, nojo e surpresa — além de uma emoção neutra, garantindo bom desempenho na tarefa; (2) analisar o desempenho de três modelos distintos (DenseNet, AlexNet e uma CNN) na classificação correta das imagens, com base em suas métricas de acurácia, loss e matriz de confusão alcançado através dos hiperparâmetros sugeridos pelo Hyperopt.

Esses tópicos serão abordados com mais detalhes nas seções posteriores. Assim, o trabalho está organizado da seguinte maneira: A seção 2 apresenta uma breve revisão bibliográfica com os principais conceitos necessários para o entendimento do estudo. Em seguida a seção 3 descreve a metodologia do estudo, detalhando o sistema desenvolvido e as etapas seguidas para sua conclusão. A seção 4 apresenta os resultados obtidos, suas implicações e as discussões pertinentes. Por fim, a seção 5 oferece as conclusões do estudo.

2. Revisão Bibliográfica

Este capítulo apresenta uma revisão dos conceitos fundamentais abordados nesta pesquisa, dividida em dois tópicos principais: teorias sobre emoções e expressões faciais e redes neurais artificiais. Cada um desses tópicos é essencial para entender o contexto e a metodologia utilizados no desenvolvimento do sistema de reconhecimento de emoções proposto.

2.1. Teorias sobre emoções e expressões faciais

Reconhecer emoções pode parecer um exercício simples, dado que as pessoas frequentemente lidam com as emoções umas das outras no cotidiano. A alegria, por exemplo, pode ser identificada por um sorriso, e a tristeza por lágrimas, mas essas associações nem sempre são precisas. Na teoria James-Lange as emoções são mudanças fisiológicas resultantes da interpretação do cérebro a estímulos externos (JAMES, 1890). Isso ocorre quando, por exemplo, uma pessoa se depara com um cão raivoso e imediatamente seus batimentos cardíacos aceleram.

A abordagem psicoevolucionista propõe que as emoções surgiram através da evolução das espécies, funcionando como respostas adaptativas ao ambiente em que determinada população se encontrava (DARWIN, 1872). Dessa forma, embora a manifestação de algumas emoções possa ser aprendida, existem outras, especialmente as expressões faciais, que são inatas ao ser humano. A noção de emoções inatas e comuns a todas as culturas humanas foi uma observação crucial relatada na pesquisa de Darwin, que posteriormente dialogaria com as pesquisas de Paul Ekman, reforçando a ideia de universalidade das expressões emocionais.

A literatura das ciências psicológicas apresenta duas abordagens amplamente utilizadas para o reconhecimento de emoções: a categórica e a contínua. A representação categórica é conhecida por classificar as emoções em diferentes categorias distintas. Em contrapartida, a abordagem contínua considera as emoções como pontos em um espaço multidimensional, criando uma relação espacial entre elas.

2.1.1. Representação Categórica

Em 1972, Paul Ekman publicou o artigo "*Constants across cultures in the face and emotion*", onde apresentou suas ideias revolucionárias sobre as emoções básicas. Através de um conjunto de estudos, ele demonstrou que a felicidade, a raiva, o nojo, o medo, a tristeza e

a surpresa têm expressões faciais específicas que são reconhecíveis em todas as culturas. No livro "*Emotions Revealed*", ele argumenta que essa universalidade se deve à complexidade muscular do rosto, que permite com que uma ampla gama de movimentos e combinações sejam possíveis, tornando as expressões faciais um meio de comunicação emocionalmente carregada.

A teoria das emoções básicas também foi abordada por Robert Plutchik (2002), que adiciona duas emoções extras, aceitação e expectativa, à lista de emoções primárias. Além disso, ele propôs que as emoções podem se agrupar para formar emoções complexas e que a maioria dos estados emocionais é, na verdade, uma composição de emoções básicas. A exemplo da saudade pode ser vista como uma mistura entre alegria e tristeza, destacando assim a interconexão e a complexidade das experiências emocionais humanas.

2.1.2. Representação Contínua

Para Russel (1980), as emoções podem ser organizadas em um espaço bidimensional circular, conforme ilustrado na Figura 1, definido por duas dimensões que mapeiam os níveis de valência e ativação. Na dimensão da valência, o extremo positivo indica sensações de prazer, enquanto o extremo negativo representa o desprazer. No eixo da ativação, um extremo indica alta ativação, enquanto o oposto denota calmaria. Nessa forma de representação, as emoções são determinadas pela combinação única entre os níveis de valência e ativação. Com esse modelo, é possível observar que emoções dispostas próximas umas das outras são expressas de maneira semelhante, tornando-se mais difíceis de serem distinguidas entre si do que emoções localizadas em quadrantes diferentes.

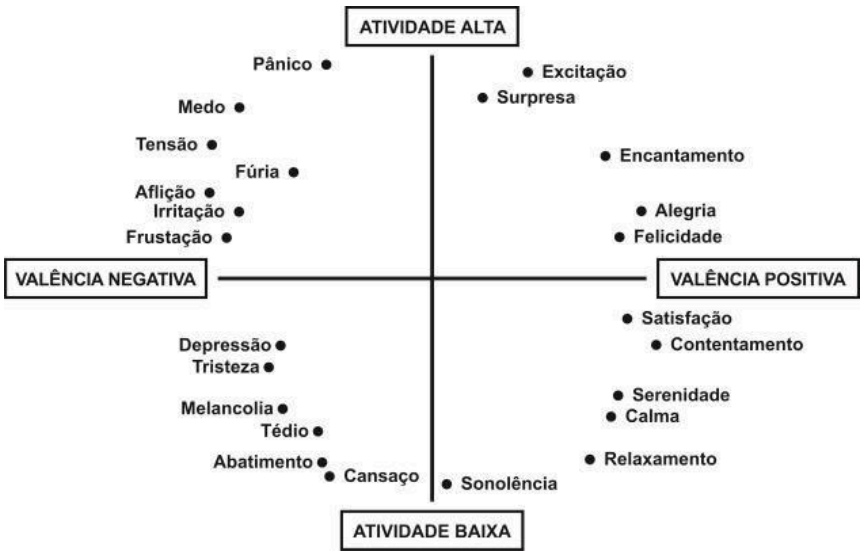

Figura 1: Modelo Circumplexo de Russel (NOGUEIRA, 2018)

2.2. Redes Neurais Artificiais

Inspiradas no comportamento dos neurônios humanos, as redes neurais artificiais são estruturas fundamentais na Inteligência Artificial Conexionista. Esta área da IA postula que, ao construir um sistema que simule a estrutura do cérebro, esse sistema apresentará inteligência, será capaz de aprender, assimilar, cometer erros e aprender com eles (FURTADO, 2019). Essa característica tem destacado as redes neurais como uma poderosa ferramenta no campo da IA e do aprendizado de máquina, especialmente porque a computação tradicional não tem sido capaz de resolver problemas complexos e não lineares, como o reconhecimento de padrões em imagens.

2.2.1. Neurônio Artificial

Em 1943, McCulloch e Pitts propuseram um modelo matemático de neurônio, conforme ilustrado na Figura 2. Simplificadamente, este modelo "dispara" quando a combinação linear de suas entradas excede um determinado limiar (Norvig, 1962). Essa arquitetura pioneira estabeleceu as bases para o desenvolvimento das redes neurais modernas, pois apesar de que cada neurônio tenha a função de realizar um processamento simples, uma rede com múltiplos neurônios é capaz de realizar operações bastante complexas (Furtado, 2019).

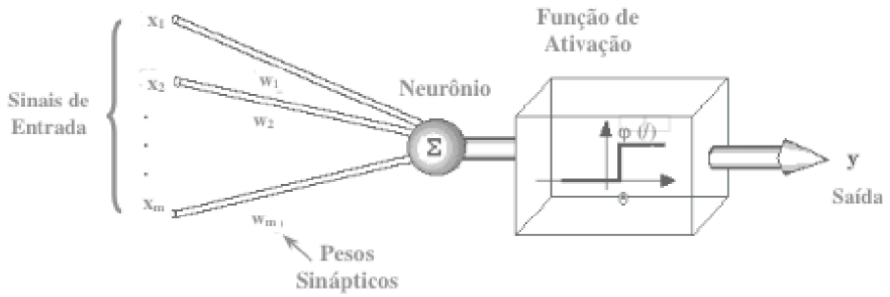

Figura 2: Unidade Processadora de McCullock e Pitts (FURTADO, 2019).

A atividade de um neurônio consiste em receber sinais de entrada, representados na imagem pela variável "x", e multiplicá-los pelos respectivos pesos sinápticos "w". Além disso, ele também é polarizado por uma entrada especial "w0" conhecida como Bias, cujo valor normalmente é igual a 1. Em seguida, uma soma ponderada desses valores é calculada para determinar um nível de atividade e então esse nível é avaliado por uma função de

ativação que, se exceder um determinado limite, fará com que o neurônio produza uma saída. Esse comportamento é diretamente inspirado no neurônio real, onde a célula neural recebe um impulso eletromagnético através dos dendritos, reage no corpo celular e, em seguida, a propaga ou não através do axônio para outros neurônios, como mostra a figura 3.

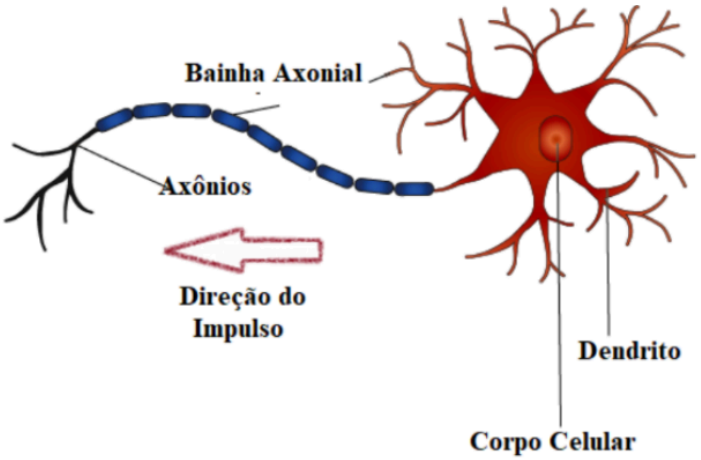

Figura 3: Neurônio biológico. (FURTADO, 2019)

Esse modelo é amplamente utilizado na maioria das redes neurais, com variações apenas na função de ativação. A função de ativação determina a atividade ou ausência de atividade do neurônio com base no resultado da soma ponderada dos pesos. Normalmente, os valores possíveis podem ser definidos em um intervalo fechado [0,1], em [-1,1], ou até mesmo em um intervalo $(-\infty, +\infty)$. Entretanto as funções mais utilizadas são a Threeshold, a Linear e a Sigmóide mostradas na Figura 4.

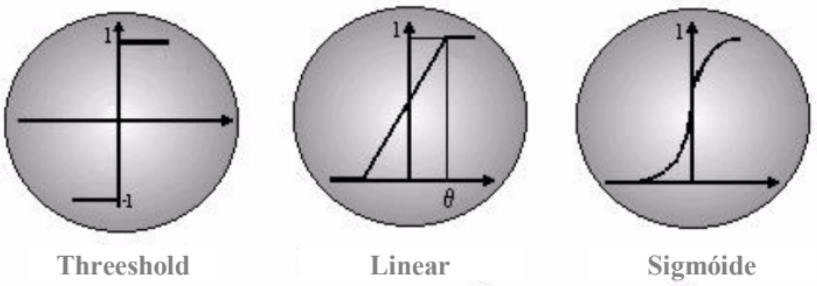

Figura 4: Função de Ativação. (FURTADO, 2019)

Um neurônio pode conectar sua saída à entrada de vários outros neurônios, formando uma extensa rede de conexões e, consequentemente, uma grande variedade de combinações. Essas conexões, em comparação com o sistema biológico, representam o contato dos dendritos com os outros neurônios, formando assim as sinapses (FURTADO, 2019). Esse tipo de arquitetura, conforme ilustrado na Figura 5, pode ser configurado variando o número de

camadas intermediárias, as conexões, a quantidade de neurônios, a função de ativação ou o método de aprendizado. A configuração especificada é determinada pelo tipo de problema que a rede é projetada para resolver.

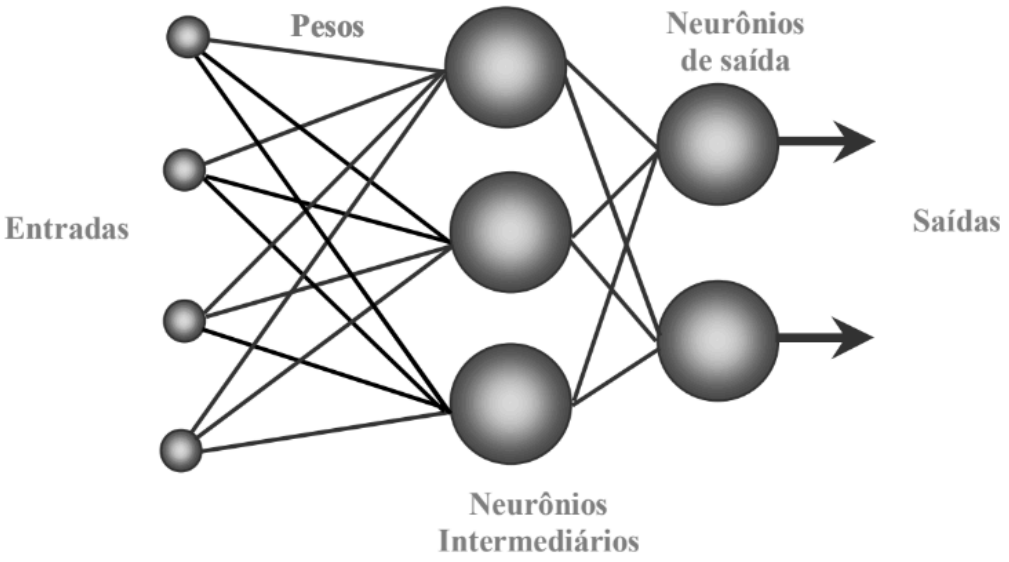

Figura 5: Rede Neural Artificial de Múltiplas Camadas. (FURTADO, 2019)

2.2.2. Redes Profundas

Os grupos de neurônios posicionados entre as camadas de entrada e saída são conhecidos como neurônios intermediários, camada oculta ou camada intermediária. Abordagens que utilizam um pequeno número de camadas são normalmente referidas como aprendizado superficial (*shallow learning*). Essa abordagem possui limitações em aprender padrões complexos nos dados; no entanto, é bastante utilizada em cenários onde os dados possuem padrões simples e a relação entre os dados e os rótulos é direta. Alguns exemplos de algoritmos que utilizam aprendizado superficial são redes neurais de camada única, máquinas de vetores de suporte (SVM) e k-vizinhos mais próximos (k-NN).

Em contrapartida, no aprendizado profundo (*deep learning*), a profundidade se refere à utilização de múltiplas camadas intermediárias conectadas que contribuem significativamente para o desempenho do modelo. Arquiteturas modernas de aprendizado profundo podem ter dezenas a centenas de camadas interligadas, cada uma aprendendo a partir dos dados a que são expostas. Essas camadas sucessivas elevaram o nível da inteligência artificial, permitindo alcançar resultados próximos aos do nível humano em tarefas como classificação de imagens, processamento de linguagem natural e transcrição de escrita manual.

2.2.3. Redes Neurais Convolucionais

As redes neurais convolucionais (CNNs) são amplamente utilizadas para resolver problemas de visão computacional, devido à sua especialização na classificação de imagens. Essas redes são capazes de identificar padrões em imagens e retornar rótulos correspondentes, tornando-as ideais para tarefas como reconhecimento de objetos, detecção de rostos e segmentação de imagens. As CNNs fazem parte do grupo de redes profundas e recebem esse nome porque a "convolução" é a operação principal realizada dentro da rede.

A operação de convolução envolve a aplicação de um filtro sobre uma imagem, resultando em um mapa de características. Na Figura 6, por exemplo, a imagem de entrada representa a escrita manual do número zero. Quando um filtro de três pixels brancos alinhados diagonalmente é aplicado sobre essa imagem, o mapa de características retornado destaca as regiões onde esse padrão específico está presente. Esse processo permite que a operação de convolução extraia e realce diferentes padrões na imagem, como bordas, texturas e outros detalhes importantes, transmitindo essa informação para as camadas subsequentes da rede neural.

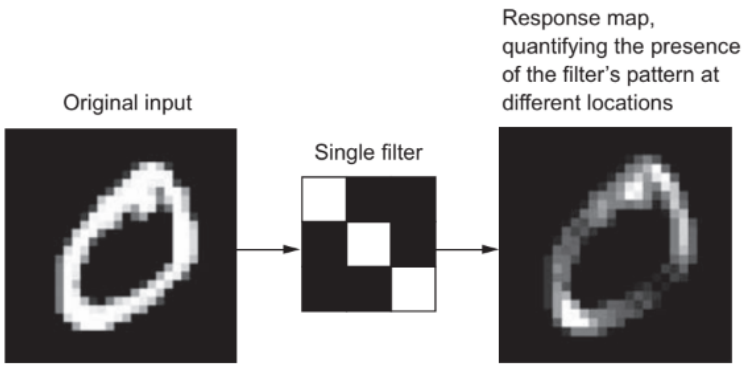

Figura 6: Aplicação de um filtro sobre uma imagem. (CHOLLET, 2018)

Segundo François Chollet, a principal diferença entre uma camada completamente conectada e uma camada convolucional é que a primeira aprende padrões globais a partir de sua entrada, enquanto a camada convolucional foca em padrões locais, como mostra a Figura 7. Em uma imagem, esses padrões locais podem ser bordas, texturas e outros pequenos detalhes que, quando combinados, formam a compreensão global da imagem.

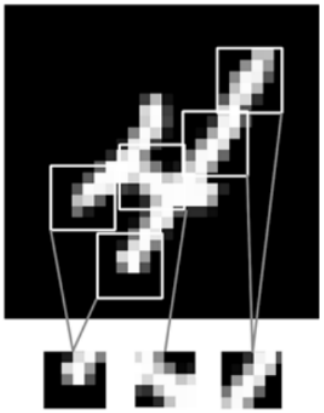

Figura 7: Uma imagem pode ser quebrada em padrões locais como borda e textura. (CHOLLET, 2018)

A operação de convolução ao longo das camadas revela um recurso fundamental que está diretamente relacionado ao potencial das CNNs em reconhecer padrões em imagens: a capacidade de aprender padrões hierárquicos. Na Figura 8, a primeira camada convolucional identifica pequenos padrões na imagem, como bordas e texturas. A segunda camada observa padrões a partir dos recursos extraídos pela primeira, e assim sucessivamente. Esse processo hierárquico permite que as camadas mais profundas reconheçam padrões complexos e abstratos, culminando na classificação precisa da imagem.

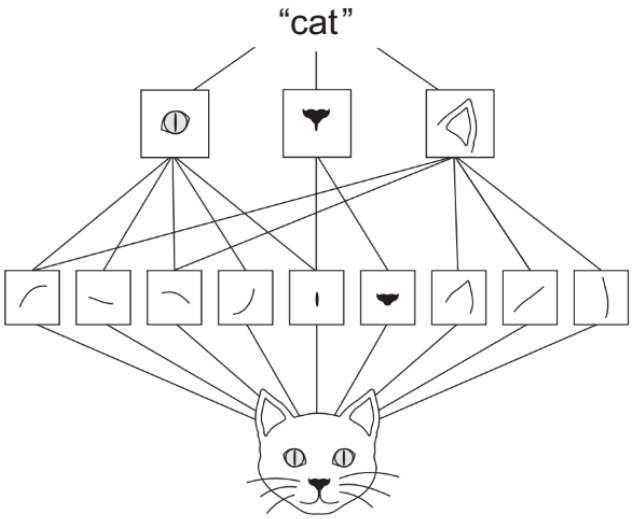

Figura 8: Padrões hierárquicos em imagem que levam a classificação de um gato. (CHOLLET, 2018)

A função de ativação mais utilizada nas Redes Neurais Convolucionais (CNNs) é a Rectifier Linear Unit, ou ReLU. Esta função é eficaz no tratamento do problema do gradiente desvanecente, no qual, durante a retropropagação do erro, as atualizações dos pesos nas camadas iniciais se tornam insignificantes, prejudicando a eficiência do modelo. A função ReLU aborda essa questão ao anular os valores negativos e manter os valores positivos inalterados, como demonstrado no gráfico da Figura 9.

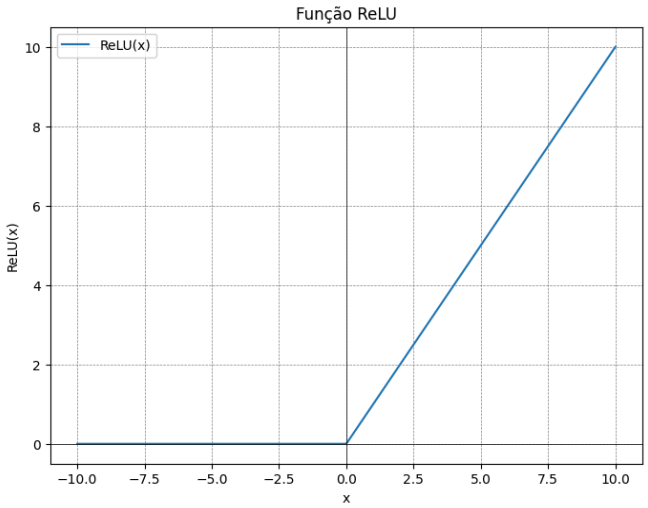

Figura 9: Gráfico da função ReLU.

Outra operação fundamental nas redes CNNs é o Pooling. Essa função é aplicada ao mapa de ativação com o objetivo de reduzir a dimensionalidade da imagem e destacar os valores com maior ou menor probabilidade. O Pooling ajuda a condensar a informação, tornando o processamento mais eficiente e robusto a variações menores na posição dos objetos na imagem. A Figura 10 exemplifica a execução da operação de max pooling, onde uma janela de 2x2 percorre a matriz de ativação e retorna o maior valor presente em cada sub-região para o mapa de ativação.

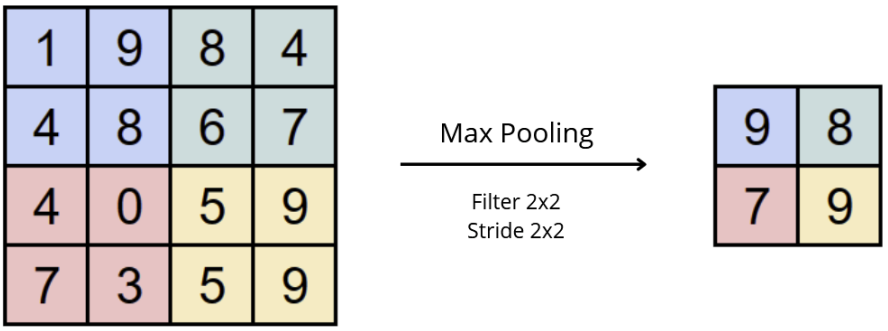

Figura 10: Max Pooling.

3. Metodologia

Este capítulo apresenta o *Visual Identification of Recognition of Emotions* (VIRE)[1], uma solução de e-health para identificar emoções através da análise de imagens. Para tanto, o VIRE foi modelado a partir da análise de três redes neurais convolucionais projetadas para

[1] Github

extrair características visuais profundas das imagens. Portanto, o VIRE aproveita a capacidade de processamento de imagem da CNN para detectar as sutilezas nas expressões faciais que correspondam a estados emocionais específicos.

3.1. VIRE - Visual Identification of Recognition of Emotions

O VIRE opera em ambiente domiciliar, utilizando o monitoramento de câmeras para identificar emoções e recomendar cuidados a pessoas como idosos, crianças e pessoas que necessitam de acompanhamento constante, conforme ilustrado na Figura 11. O processo ocorre em três etapas principais: A primeira, é a captura de imagens dos rostos (Seção A) das pessoas sendo monitoradas. A etapa seguinte trata-se da análise e identificação de emoções (Seção B), onde os dados das imagens são processados pela rede neural convolucional que identifica as emoções expressas pelos rostos capturados. Por fim, o responsável recebe notificações importantes sobre o estado emocional dos indivíduos monitorados em seu dispositivo móvel (Seção C).

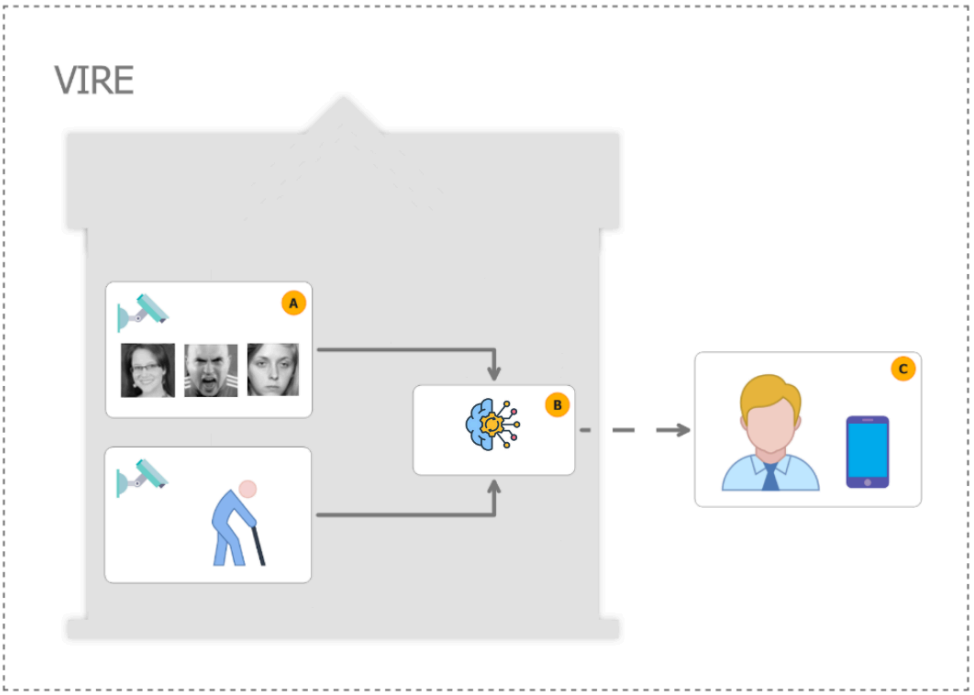

Figura 11: Visão geral do funcionamento do VIRE.

3.2. Coleta dos dados

A detecção de emoções pelo VIRE depende da identificação de padrões nos rostos das pessoas observadas. Ekman, com base em suas pesquisas sobre expressões faciais, propôs que

algumas emoções básicas, como alegria, raiva, tristeza, medo, nojo e surpresa, são universalmente reconhecidas por meio dessas expressões. Portanto, esta pesquisa utilizou o conjunto de imagens FER2013, que classifica essas seis emoções básicas conforme proposto por Ekman além de uma neutra.

O FER2013 (Facial Expression Recognition 2013) foi apresentado pela primeira vez em 2013 como um desafio de classificação de emoções no Kaggle, sendo um conjunto de dados totalmente novo até aquele momento. Ele foi compilado por Pierre-Luc Carrier e Aaron Courville com imagens encontradas na internet como parte do projeto de pesquisa deles. O dataset é composto por imagens de faces com dimensões 48 x 48 pixels, em preto e branco organizadas em sete diferentes emoções, sendo elas raiva (*angry*), nojo (*disgust*), medo (*fearful*), felicidade (*happy)*, tristeza (*sad*), surpresa (*surprise*), e neutro (*neutral*) como visto na Figura 12.

Figura 12: Amostras do FER2013.

Para o treinamento do modelo foram submetidas às 28.709 amostras de treino do FER2013, o que equivale a cerca de 80% da base de dados. Essas imagens, classificadas em sete emoções foram distribuídas em cada uma delas conforme mostra a Figura 13.

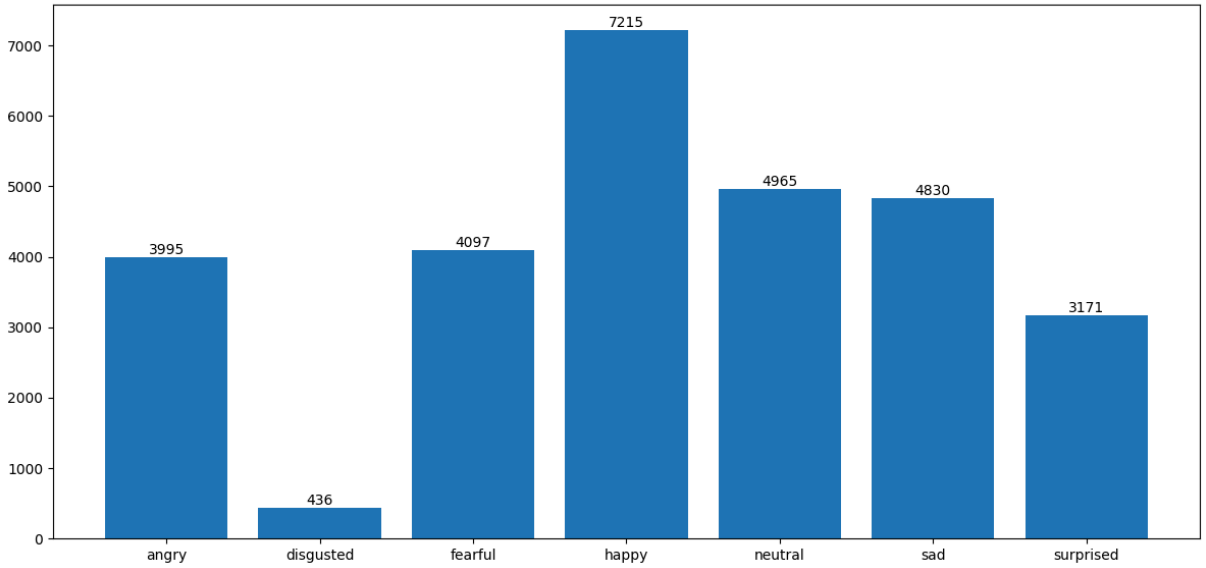

Figura 13: Distribuição de amostras entre as classes.

Na avaliação de desempenho do modelo, foram utilizadas as 7.178 imagens previamente separadas no FER2013, correspondendo aproximadamente a 20% da base de dados, conforme sugere a técnica Hold-out na Figura 14. Para avaliar o desempenho do modelo, foram utilizadas as seguintes métricas: (i) Acurácia: essa métrica permite avaliar a capacidade do modelo em classificar corretamente um rosto em uma emoção; (ii) Matriz de confusão: em caso de desbalanceamento entre as classes, a acurácia pode oferecer resultados tendenciosos. A matriz de confusão, por sua vez, detalha a precisão para cada emoção, mostrando a quantidade de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos para cada classe. Com essas duas métricas, é possível obter uma visão mais abrangente e precisa do desempenho do modelo.

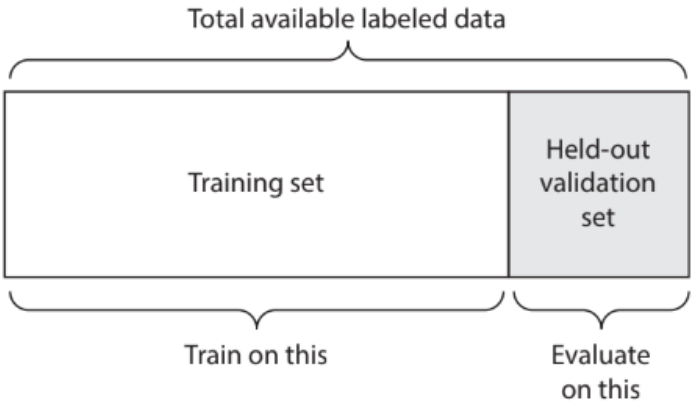

Figura 14: Hold out (CHOLLET, 2017).

Para melhorar o desempenho da rede neural durante o treinamento, é essencial que os dados estejam adequados para alimentar a rede. Isso requer a aplicação de pré-processamento no conjunto de dados. A normalização, técnica de pré-processamento aplicada, apresentada na Figura 15, envolve a transformação de uma imagem em um tensor de ponto flutuante. Inicialmente, na Seção A, um arquivo de imagem é decodificado em matrizes de pixels RGB. Em seguida, na Seção B, todos os valores são convertidos para ponto flutuante e redimensionados para uma escala de 0 a 1. Dessa forma, a rede neural pode trabalhar com valores menores, evitando problemas de estouro de memória.

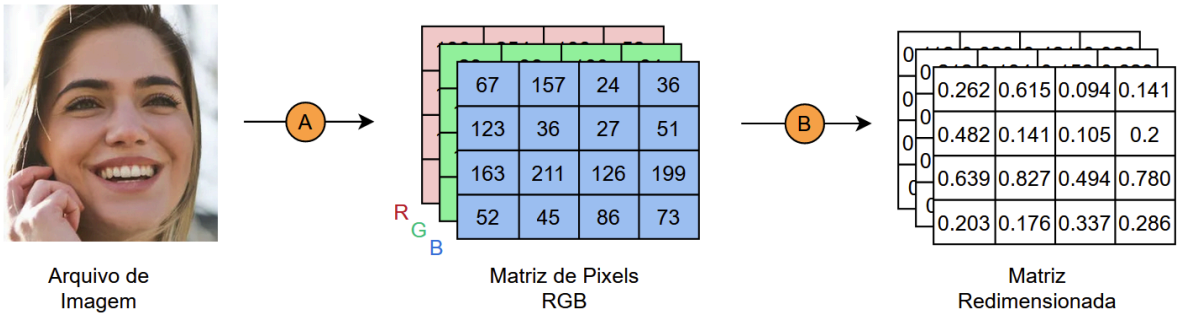

Figura 15: Normalização dos dados.

3.3. Mecanismos de reconhecimento

Nesta seção serão apresentados os modelos de reconhecimento de emoções com base em visão computacional que foram implementadas no VIRE. Os modelos propostos a serem analisados foram uma CNN, a rede Alexnet e a Densenet.

3.3.1. CNN

A CNN modelada é composta por doze camadas, como apresentada na Figura 16. As seis primeiras consistem em dois conjuntos que seguem a sequência de uma camada convolucional, uma de max pooling e uma de dropout[2]. Essas camadas são responsáveis por identificar pequenos padrões locais nas imagens, como bordas e texturas.

Após a camada de achatamento[3] (*Flatten*), o modelo é composto por dois conjuntos de camadas densamente conectadas cuja função de ativação é a ReLU e uma camada de dropout, finalizando com uma camada densa de sete saídas. Essa última sequência é responsável por identificar padrões globais e retornar uma classificação para a imagem submetida ao modelo.

[2] A camada de droupout é uma camada em que durante o treinamento alguns neurônios tão temporariamente desligados forçando a rede a aprender representações mais robustas e diminuindo a dependência entre neurônios para reconhecer padrões.
[3] A camada Flatten transforma o formato dos dados de entrada em um vetor unidimensional frequentemente utilizado após uma série de camadas convolucionais e antes das camadas densamente conectadas.

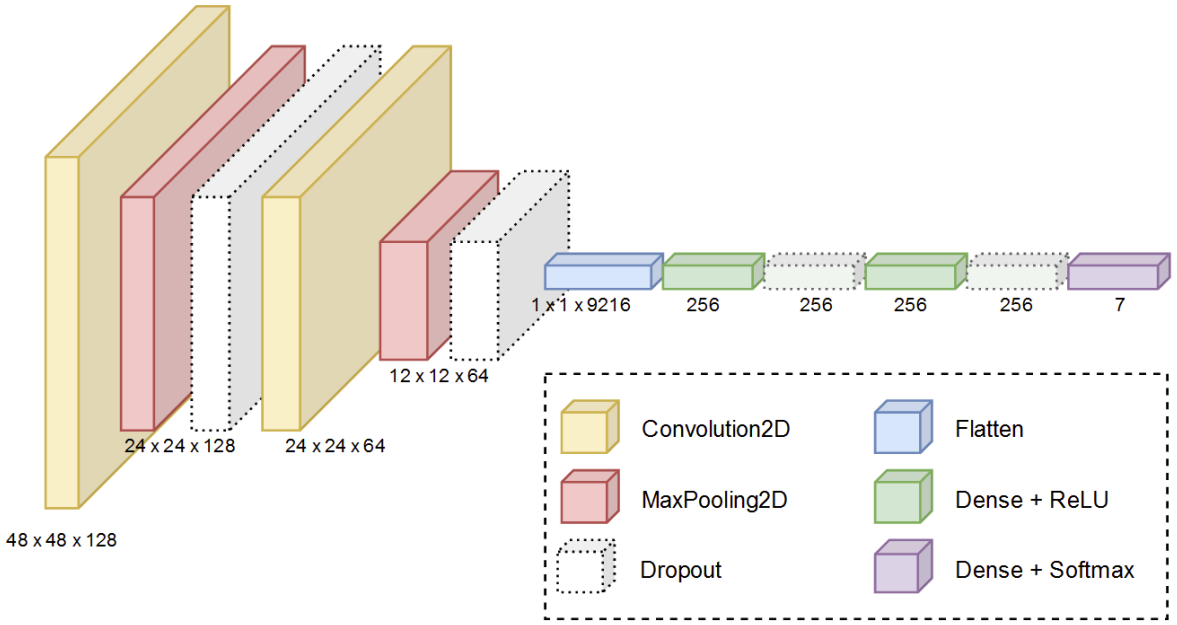

Figura 16: Arquitetura da CNN.

3.3.2. Densenet

DenseNet é uma arquitetura proposta por Gao Huang, Zhuang Liu, Laurens van der Maaten e Kilian Q. Weinberger no artigo "Densely Connected Convolutional Networks", que também dá nome à rede. Ela é uma de rede neural convolucional composta por sequências de camadas de convolução, pooling e funções de ativação ReLU. No entanto, o que a diferencia das demais redes convolucionais é o conceito de blocos densos, que se situam entre as camadas de convolução e pooling, conforme exemplificado na Figura 17.

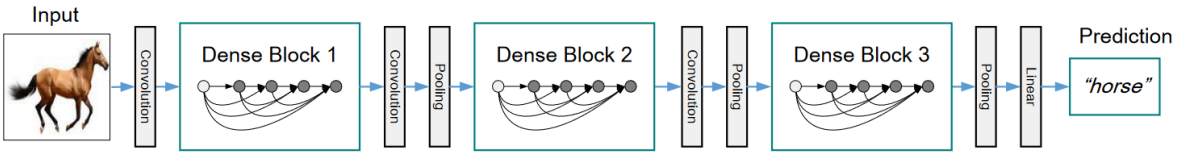

Figura 17: Arquitetura Densenet com três blocos densos. (HUANG ET AL., 2017)

Um bloco denso é um conjunto de camadas onde há uma conexão direta entre uma camada e todas as subsequentes, como mostrado na Figura 18. Essa característica difere de uma CNN tradicional, na qual uma camada recebe apenas a saída da camada anterior. Ao fazer isso, os blocos densos mitigam o problema do desvanecimento do gradiente, que ocorre em redes muito profundas quando a propagação do erro ao longo das camadas se reduz a zero.

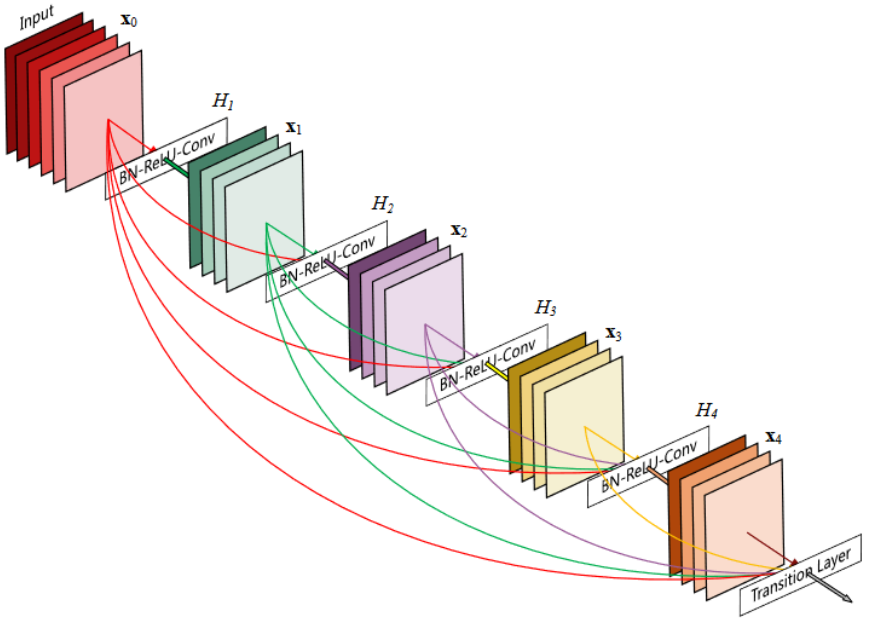

Figura 18: Um bloco denso de 5 camadas (HUANG ET AL., 2017)

A arquitetura da DenseNet utilizada nesta pesquisa é ilustrada na Figura 19. Ela começa com uma camada de entrada para imagens de dimensão 48 x 48 com 1 canal (preto e branco). Seguindo, há uma camada convolucional inicial que aplica um filtro 3 x 3, reduzindo a dimensão das imagens para 24 x 24 x 64, seguida por uma camada de normalização em lote e ativação ReLU. Finalizando essa primeira parte, há uma camada de pooling médio que reduz a dimensão para 12 x 12 x 64.

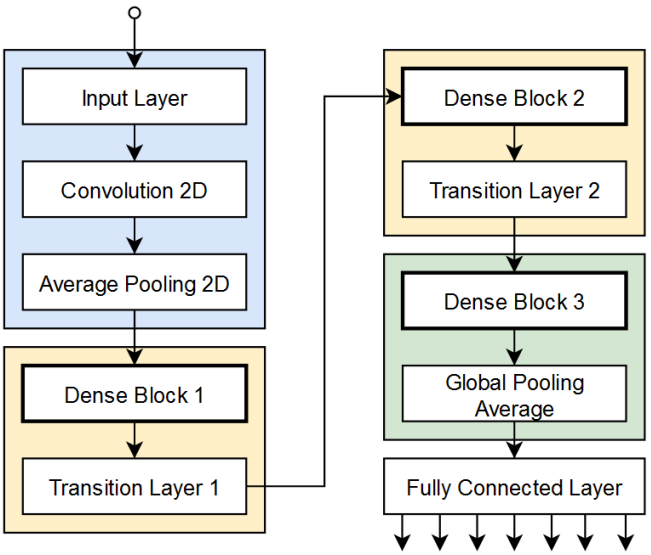

Figura 19: Arquitetura da Densenet implementada.

A camada seguinte é um bloco denso que contém repetidas camadas de normalização em lote, ativação ReLU e convolução com 32 filtros 3 x 3. Após cada convolução, as saídas

são concatenadas, resultando em uma dimensão de 12 x 12 x 192. Na camada de transição, há uma normalização em lote, convolução que reduz o número de filtros para 96, e pooling médio, resultando em dimensões de 6 x 6 x 96.

A próxima camada de bloco denso funciona de maneira semelhante à primeira, aumentando a dimensão para 6 x 6 x 224, enquanto a camada de transição subsequente reduz novamente para 3 x 3 x 112, realizando as mesmas operações. Em seguida, um bloco denso aumenta as dimensões para 3 x 3 x 124, e por fim, a camada de pooling médio reduz as dimensões a um vetor de 240 unidades, alimentando a última camada da rede, uma camada densa que retorna a classificação do modelo.

3.3.3. Alexnet

AlexNet é uma arquitetura de CNN que surgiu em 2012 como uma proposta para o desafio ImageNet Large Scale Visual Recognition Challenge (ILSVRC). A competição exigia que o modelo fosse treinado com um banco de dados contendo 1000 categorias, cada uma com 1000 amostras. AlexNet conseguiu obter resultados significativamente superiores aos de seus concorrentes, estabelecendo um novo padrão para o desempenho em tarefas de classificação de imagens.

A estrutura da AlexNet é composta por 5 camadas convolucionais, 3 camadas de pooling e, por fim, 3 camadas totalmente conectadas, conforme mostrado na Figura 20. A primeira camada de convolução aplica 96 filtros de dimensão 11 x 11 com stride[4] 4, seguida por uma camada de max pooling de 3 x 3 com stride 2. A janela de convolução da camada seguinte é reduzida para 5 x 5, com 256 filtros, stride 1 e padding 2, seguida por outra camada de max pooling de 3 x 3 com stride 2.

A terceira, quarta e quinta camadas utilizam 384, 384 e 256 filtros de 3 x 3, todas com stride e padding 1. Após a quinta camada convolucional, há uma camada de max pooling de 3 x 3 com stride 2. Após a última camada convolucional, o modelo é seguido por duas camadas densas de 4096 neurônios cada, com função de ativação ReLU e dropout de 50%. Finalmente, há uma camada de classificação com 1000 saídas.

[4] Número de pixels que uma janela convolucional salta ao se mover

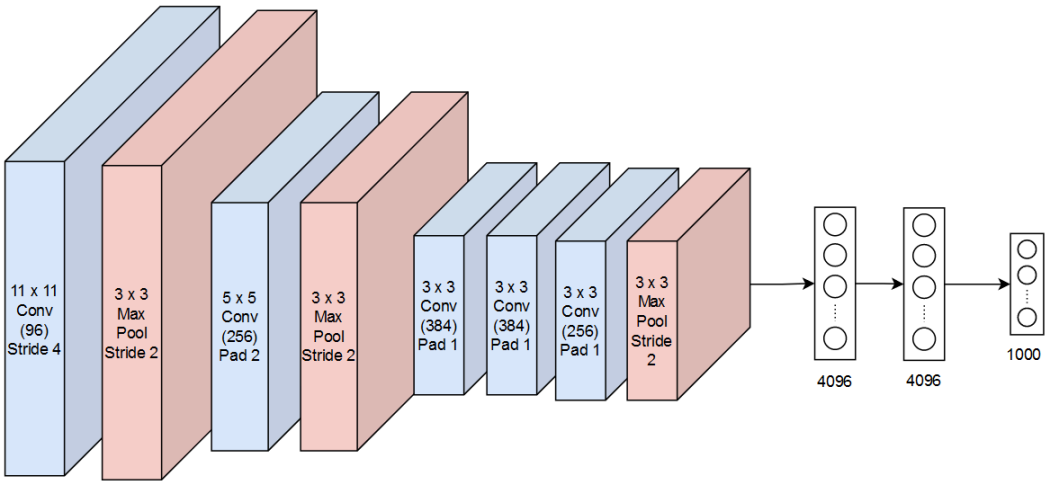

Figura 20: Arquitetura da Alexnet.

3.4. Otimização da busca pelos melhores hiperparâmetros

Cada modelo possui parâmetros e hiperparâmetros que definem seu comportamento durante o treinamento e, consequentemente, influenciam o resultado final. Os parâmetros são valores ajustados iterativamente durante o processo de treinamento do modelo, como os pesos de cada neurônio. Em contraste, os hiperparâmetros são valores configurados pelo desenvolvedor antes do treinamento, como exemplo incluem a taxa de aprendizado, número de épocas e tamanho do batch, cada um influenciando a complexidade e eficiência do modelo. A escolha correta dos hiperparâmetros é crucial para evitar o overfitting, permitindo que o modelo se generalize bem em dados não vistos.

Esta pesquisa utilizou o Hyperopt, um framework de busca para otimizar a configuração de hiperparâmetros a partir de uma lista de combinações possíveis. O algoritmo utiliza uma abordagem bayesiana, onde, dado um modelo, ele atribui uma pontuação a cada configuração de hiperparâmetros. Essas configurações e pontuações são atualizadas iterativamente, com o objetivo de maximizar a pontuação com base nos resultados anteriores. Dessa forma, o Hyperopt busca encontrar a melhor combinação de hiperparâmetros de forma eficiente e precisa.

Ao Hyperopt é necessário definir alguns conceitos importantes como a função objetivo, o espaço de configuração e o algoritmo de pesquisa. A função objetivo será responsável por medir o desempenho de uma determinada configuração de hiperparâmetros, retornando uma métrica que o algoritmo de pesquisa tentará minimizar ou maximizar. O

espaço de configuração define o conjunto de todos os possíveis valores que os hiperparâmetros podem assumir. Enquanto que o algoritmo de pesquisa, por sua vez, determina a estratégia para explorar o espaço de configuração.

Para comparar de forma justa o desempenho de cada um dos três modelos propostos na pesquisa, o Hyperopt foi utilizado para encontrar os melhores hiperparâmetros de cada um deles. Dessa forma, foram utilizados os seguintes conjuntos de valores para o espaço de configuração: a taxa de aprendizado variou entre 0.0001 e 0.01; o tamanho do lote (batch) incluiu 8, 16, 32, 64 e 128; o número de épocas foi testado com 25, 50, 100 e 150; e os algoritmos de inicialização dos pesos considerados foram Glorot Uniform, He Normal e Lecun Normal. O algoritmo TPE foi utilizado como algoritmo de pesquisa. Esta abordagem garantiu a otimização dos modelos, maximizando seu desempenho na tarefa de classificação de imagens.

4. Resultados e Discussão

Este capítulo apresenta os resultados obtidos para as diversas configurações de hiperparâmetros e o desempenho dos modelos treinados com esses parâmetros, seguindo a metodologia aplicada nas duas fases do treinamento. Primeiramente, os melhores hiperparâmetros são encontrados através do Hyperopt. Em seguida, são apresentados os resultados do treinamento de cada modelo para cada configuração encontrada e são analisadas as métricas de loss e acurácia para os conjuntos de dados de treinamento e validação. Por fim, as métricas de performance de cada arquitetura são comparadas com o objetivo de identificar tendências, vieses ou problemas, destacando o modelo que obteve o melhor desempenho entre as métricas avaliadas.

4.1. Fase de Treinamento I

O Hyperopt foi executado ao longo de dez iterações, também conhecidas como trials. Durante cada trial, o Hyperopt selecionou uma configuração específica de hiperparâmetros, treinou a rede neural com esses parâmetros e avaliou o desempenho do modelo. Esse ciclo se repetiu até que o número total de trials fosse atingido. Na Figura 21, é apresentada uma comparação dos valores de loss em cada configuração ao longo dos dez trials para cada uma das arquiteturas de rede neural propostas. Dado que o loss indica a discrepância entre o

resultado esperado e o resultado obtido, a arquitetura cuja pontuação dessa métrica se destacou entre as demais foi a CNN.

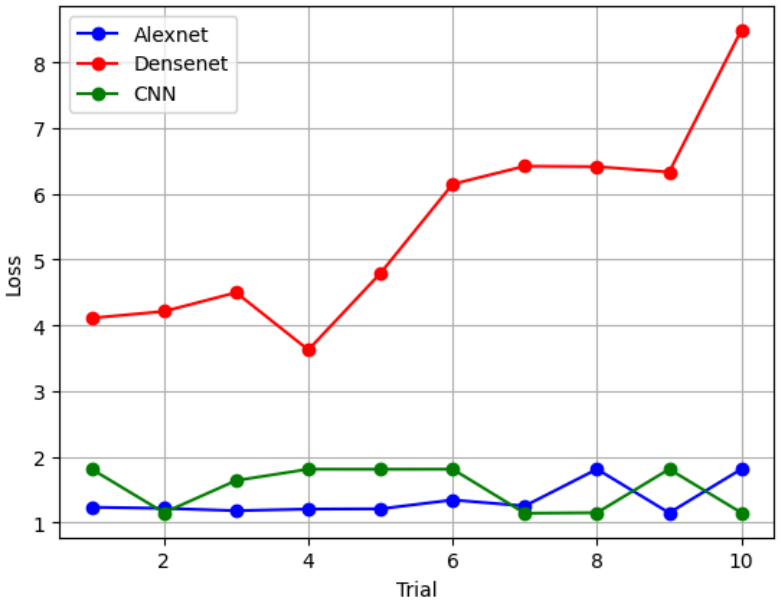

Figura 21: Desempenho de cada configuração na busca dos hiperparâmetros da fase I.

A execução do algoritmo resultou na identificação dos hiperparâmetros otimizados apresentados na Tabela 1. Ao analisar esses valores, destacam-se algumas tendências relevantes. Por exemplo, observa-se que a taxa de aprendizado para a CNN é consideravelmente maior em comparação com os modelos AlexNet e DenseNet. Esse padrão pode ser atribuído à capacidade das CNNs de capturar eficientemente características espaciais complexas dos dados, permitindo ajustes mais agressivos na taxa de aprendizado. Logo, essa observação ressalta a importância de inicializar os modelos com hiperparâmetros ideais para alcançar o melhor desempenho possível na tarefa proposta.

Modelo	**Taxa de Aprendizado**	**Batch**	**Épocas**	**Inicializador**
Alexnet	0.00014143674456470649	64	50	Lecun Normal
Densenet	0.00013955748563558045	8	25	Lecun Normal
CNN	0.000828395792147589	8	50	He Normal

Tabela 1: Hiperparâmetros sugeridos pelo Hyperopt no primeiro treinamento.

Conforme a Tabela 1, novos treinamentos foram realizados utilizando os hiperparâmetros sugeridos pelo Hyperopt. Os resultados serão detalhados individualmente,

com base nas métricas de acurácia e loss para os datasets de treinamento e validação, nas seções a seguir.

Alexnet

O treinamento da rede AlexNet foi realizado ao longo de 50 épocas e, ao analisar os gráficos da Figura 22, algumas conclusões podem ser extraídas. O valor do loss para o conjunto de dados de treinamento diminui progressivamente até se aproximar de zero, indicando que o modelo está se ajustando bem aos dados de treinamento. Em contrapartida, o loss para os dados de validação aumenta consideravelmente, sugerindo que o modelo não é capaz de generalizar para novos dados de entrada. Esse comportamento é um forte indicativo de overfitting.

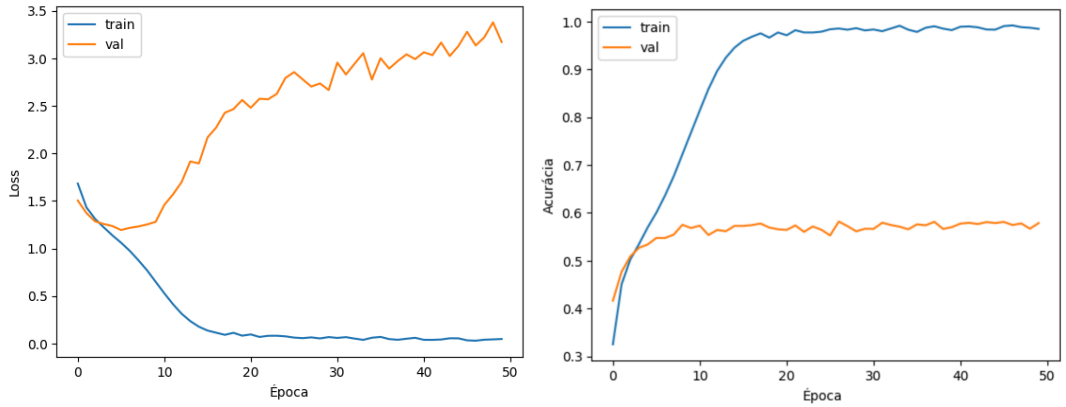

Figura 22: Desempenho da Alexnet no treinamento I.

Ao observar o gráfico da acurácia, é possível confirmar que o modelo está bem ajustado aos dados de treinamento. No entanto, a discrepância significativa entre a acurácia do conjunto de treinamento e a acurácia do conjunto de validação reforça a evidência de overfitting. Isso mostra que, embora o modelo tenha aprendido os padrões dos dados de treinamento, ele não consegue aplicar esse aprendizado de forma eficaz a dados não vistos. Os resultados finais das métricas de desempenho para este modelo para a validação foram: Acurácia = 0.5786 e Loss = 3.1725; Para treinamento: Acurácia = 0.9846 e Loss = 0.0466.

Densenet

Para a DenseNet, o número sugerido de épocas pelo Hyperopt foi de 25. Ao observar o gráfico na Figura 23, é possível chegar a conclusões semelhantes às da AlexNet. O gráfico de loss mostra que, durante o treinamento, o valor diminui rapidamente até se aproximar de zero, indicando uma boa adaptação aos dados de treinamento. No entanto, para o conjunto de validação, o loss aumenta continuamente.

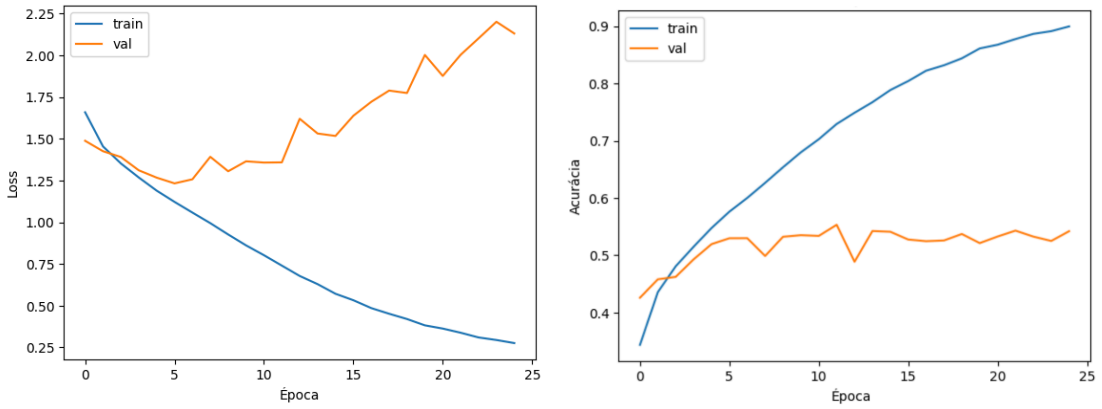

Figura 23: Desempenho da Densenet no treinamento I.

Analisando a acurácia, o modelo demonstra um bom ajuste aos dados de treinamento, mas uma diferença substancial é observada em relação aos dados de validação. Isso sugere que a rede também está sofrendo de overfitting. Os resultados finais das métricas de desempenho para este modelo para a validação foram: Acurácia = 0.5421 e Loss = 2.1306; Para treinamento: Acurácia = 0.8996 e Loss = 0.2761.

CNN

O modelo também foi treinado por 50 épocas, e o comportamento de seus gráficos de desempenho na Figura 24 difere consideravelmente dos modelos anteriores. Uma observação inicial é que os valores de loss e acurácia seguem trajetórias semelhantes tanto para o conjunto de treinamento quanto para o de validação, o que sugere uma generalização satisfatória e exclui parcialmente a possibilidade de overfitting.

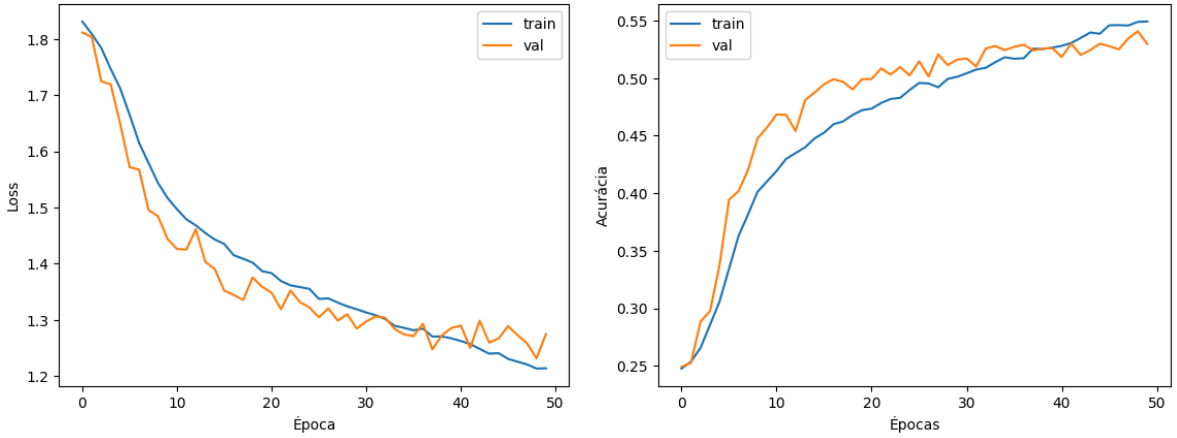

Figura 24: Desempenho da CNN no treinamento I.

No entanto, ao final do treinamento, as taxas de acurácia para ambos os conjuntos permaneceram relativamente baixas, indicando que o modelo pode precisar de mais ajustes, como um número maior de épocas de treinamento. Os resultados finais das métricas de

desempenho para este modelo para a validação foram: Acurácia = 0.5297 e Loss = 1.2749; Para treinamento: Acurácia = 0.5493 e Loss = 1.2139.

Dado que havia a possibilidade de realizar mais treinamentos para esse modelo, ele foi treinado novamente, desta vez com 300 épocas. Observando os resultados na Figura 25, é possível notar que, entre a 40ª e a 60° época, os valores de loss e acurácia para o conjunto de treinamento e validação começam a divergir. Enquanto o modelo continua a se ajustar aos dados de treinamento, os valores de loss para o conjunto de validação permanecem relativamente estáveis, sem mostrar melhoria significativa. Essa divergência crescente indica que o modelo está sofrendo de overfitting antes de alcançar resultados satisfatórios.

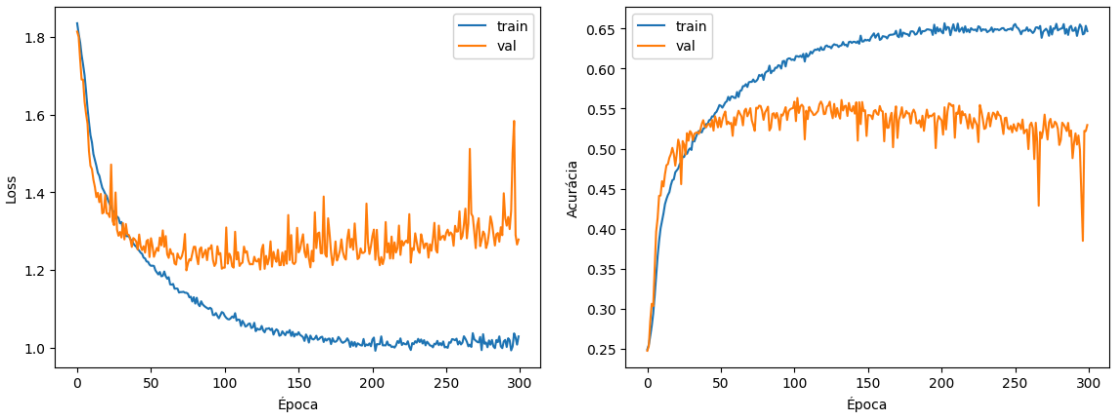

Figura 25: Desempenho da CNN com 300 épocas no treinamento I.

As métricas de desempenho ao final desse novo treinamento alcançaram os seguintes resultados: para validação loss = 1.2782, acurácia = 0.5295; para treinamento loss = 1.0284 e acurácia = 0.6467.

Resultados

Mesmo utilizando os parâmetros propostos pelo Hyperopt, todos os modelos enfrentaram problemas de generalização, resultando em overfitting em algum momento. A tabela 2 apresenta um resumo dos resultados obtidos para cada treinamento com suas respectivas configurações sugeridas. As redes AlexNet e DenseNet alcançaram um loss próximo de zero e uma acurácia de treinamento próxima de um, indicando uma boa capacidade de ajuste sobre os dados de treinamento. No entanto, ambas apresentaram um loss muito alto nos dados de validação, resultando em uma acurácia baixa e, consequentemente, uma capacidade de generalização insatisfatória.

Modelo	Loss Treino	Loss Val	Acc Treino	Acc Val
Alexnet	0.0466	3.1725	0.9846	0.5786
Densenet	0.2761	2.1306	0.8996	0.5421
CNN	1.2139	1.2749	0.5493	0.5297
CNN 300e	1.0284	1.2782	0.6467	0.5295

Tabela 2: Resultado dos treinamentos na fase I.

Inicialmente, a CNN mostrou resultados que sugerem um potencial de melhoria com mais épocas de treinamento. No entanto, ao treinar a rede novamente com 300 épocas, observou-se pouca variância e indicadores que não foram melhores do que os obtidos com 50 épocas, além de evidenciar o problema do overfitting.

4.2. Fase de Treinamento II

Nessa etapa o objetivo é utilizar *data augmentation* para mitigar o overfitting presente nos resultados anteriores. A partir do conjunto de dados de teste, essa técnica irá gerar novas imagens variando as que já existem com rotação, translação, distorção, zoom e até recorte. As transformações aplicadas estão listadas na tabela 3 com seus respectivos níveis de intensidade.

Atributo	rotation_range	width_shift_range	height_shift_range	shear_range	zoom_range	horizontal_flip	fill_mode
Valor	10%	10%	10%	10%	10%	True	nearest

Tabela 3: Atributos de data augmentation

Iniciando o treinamento da mesma forma que na etapa anterior, o Hyperopt foi executado novamente. No gráfico da Figura 26, nota-se que, nesta execução, o pior valor de loss é significativamente menor que o da execução anterior. A rede AlexNet e a CNN apresentaram grande variação ao longo dos trials, obtendo, respectivamente, 1.0942326784133911 e 1.039408802986145 como melhor valor de loss. Já a DenseNet manteve-se com muito menos variância, apresentando o melhor loss entre as três: 1.0275566577911377.

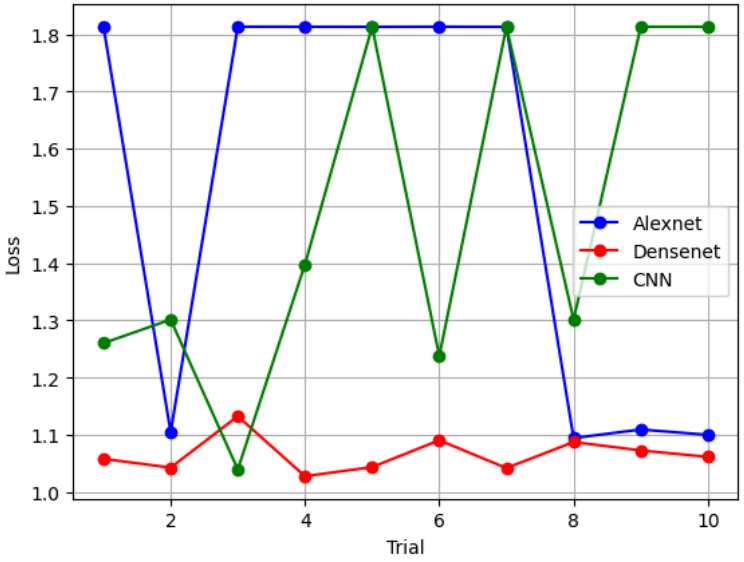

Figura 26: Desempenho de cada configuração na busca dos hiperparâmetros da fase II.

A nova execução resultou nos hiperparâmetros listados na tabela 4. Desta vez, cada modelo recebeu uma função inicializadora de pesos diferente, e todos treinaram pelo mesmo número de épocas. As redes AlexNet e DenseNet receberam valores de taxa de aprendizado relativamente maiores que na execução anterior, além do menor valor possível para o batch entre os listados. Em contrapartida, a CNN recebeu uma taxa de aprendizado menor, enquanto o batch aumentou.

Modelo	**Taxa de Aprendizado**	**Batch**	**Épocas**	**Inicializador**
Alexnet	0.000248422068699183	8	150	Lecun Normal
Densenet	0.0004337574711035279	8	150	He Normal
CNN	0.00016716393601209385	16	150	Glorot Uniform

Tabela 4: Hiperparâmetros sugeridos pelo Hyperopt no segundo treinamento.

Alexnet

Apesar de treinar com os dados transformados, a AlexNet começou a apresentar um desempenho ruim logo após a vigésima época, similar ao comportamento observado no treinamento anterior. Nota-se na Figura 27 que a rede consegue se ajustar bem aos dados de treinamento, mas falha em generalizar, resultando em valores muito altos para o loss no dataset de validação. A acurácia manteve-se estável e não superou substancialmente os valores do treinamento anterior. Especificamente, a rede obteve um loss de 0,2570 e uma

acurácia de 0,9115 no treinamento, enquanto na validação o loss foi de 2,1214 e a acurácia de 0,6092.

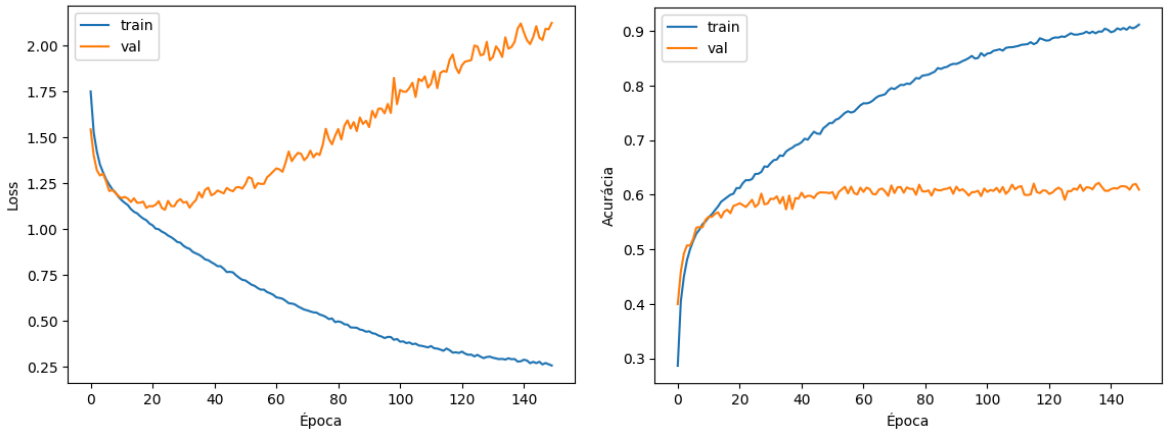

Figura 27: Desempenho da Alexnet no treinamento II.

Densenet

A DenseNet, por sua vez, obteve resultados melhores do que seu treinamento sem data augmentation como observado na Figura 28. O loss, por exemplo, em nenhum momento alcançou os picos observados no treinamento anterior, mantendo-se relativamente estável ao longo das 150 épocas. Apesar da melhora, o valor do loss para validação não conseguiu cair significativamente, evidenciando dificuldades na generalização. Esse comportamento refletiu-se na acurácia de validação, que, embora tenha mostrado um desempenho superior, ainda não foi muito melhor que a alexnet. Os resultados obtidos foram: Treinamento loss 0.4944 e acurácia 0.8163; Validação loss 1.2913 e acurácia 0.6443.

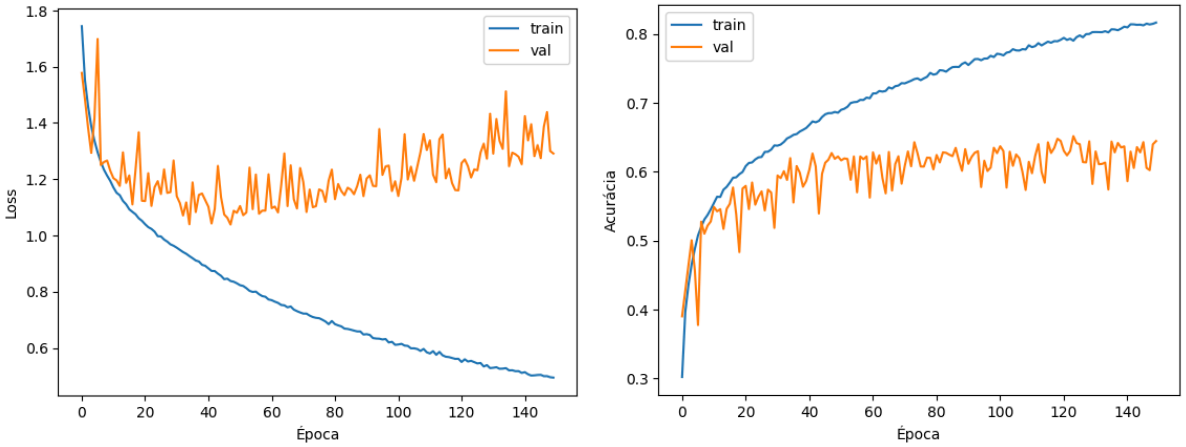

Figura 28: Desempenho da Densenet no treinamento II.

CNN

A CNN apresentou uma resposta interessante ao treinamento com data augmentation. Observando os gráficos da Figura 29, nota-se que a acurácia e o loss mantiveram certa

regularidade até o final. Essa estabilidade, juntamente com a pequena diferença entre as métricas de treino e validação, revela que o modelo conseguiu lidar bem com o overfitting. Ao final do treinamento, ela alcançou um loss de 1.2052 e uma acurácia de 0.5399 para o conjunto de treinamento, enquanto para o conjunto de validação os valores foram 1.0547 e 0.6013, respectivamente.

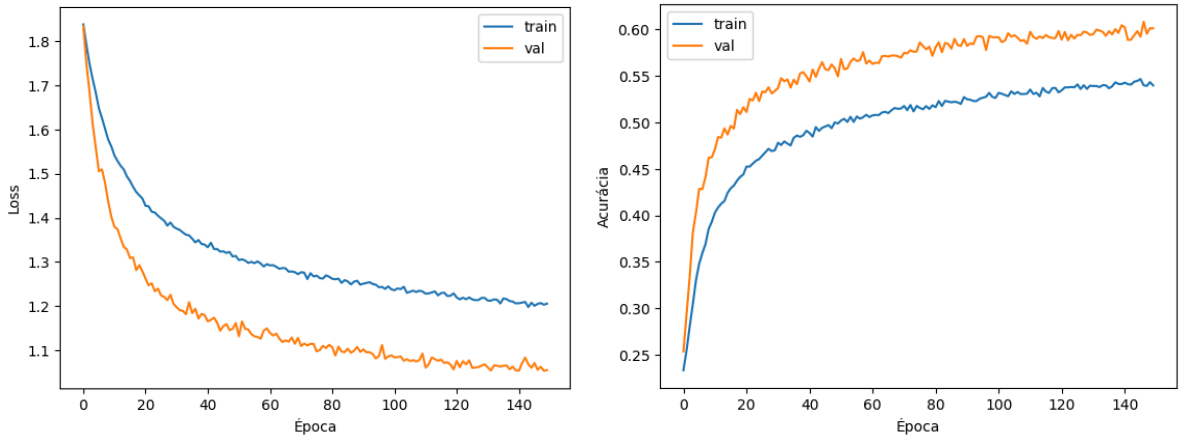

Figura 29: Desempenho da CNN no treinamento II.

Esses resultados indicam que o modelo conseguiu se ajustar aos dados de treinamento e generalizar para novos dados de entrada. A melhoria no loss e na acurácia de validação em comparação com o treinamento anterior demonstra que a data augmentation ajudou a criar um modelo mais robusto, capaz de capturar as variações presentes no conjunto de dados, aumentando assim a sua capacidade de generalização.

Resultados

A Tabela 5 apresenta um resumo dos resultados obtidos por cada modelo nesta nova etapa de treinamento. Observa-se que, embora a acurácia de validação não tenha crescido significativamente em relação ao treinamento anterior, todos os modelos se beneficiaram do uso de data augmentation, especialmente por este ter ajudado a reduzir os valores do loss de validação. Entre todos os modelos, o que mais se beneficiou com este método foi a CNN, que apresentou uma melhora substancial em comparação com o treinamento inicial. Esta melhoria indica que o data augmentation foi eficaz em aumentar a capacidade da CNN de generalizar para novos dados, tornando-o mais robusto e eficiente.

Modelo	Loss Treino	Loss Val	Acc Treino	Acc Val
Alexnet	0.2570	2.1214	0.9115	0.6092

Modelo	Loss Treino	Loss Val	Acc Treino	Acc Val
Densenet	0.4944	1.2913	0.8163	0.6443
CNN	1.2052	1.0547	0.5399	0.6013

Tabela 5: Resultado dos treinamentos na fase II.

Analisando mais profundamente o desempenho da CNN, a Figura 30 apresenta a matriz de confusão para o modelo. A primeira observação é que nenhum dado correspondente ao nojo foi corretamente classificado. Isso provavelmente se deve ao fato de que essa emoção possui poucas imagens em comparação com as outras classes. Em contrapartida, as imagens referentes à felicidade se confundem bastante com as de outras emoções, o que pode levar o modelo a classificações erradas da emoção. Por exemplo, 439 imagens de rostos felizes foram classificadas como tristeza, um número quase igual ao de previsões corretas para felicidade.

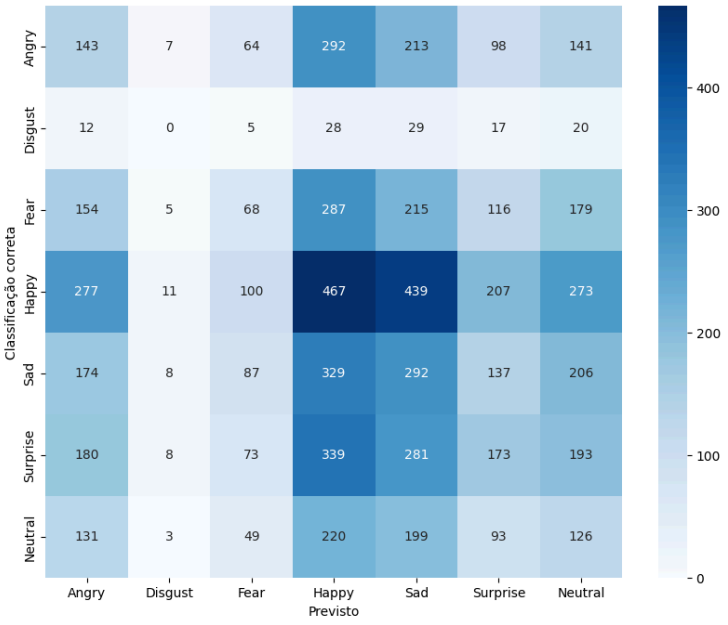

Figura 30: Matriz de Confusão da CNN.

De modo geral, para que o modelo tivesse um desempenho melhor, o número de predições na diagonal principal deveria ser maior em relação às outras células do gráfico. Mas como a matriz de confusão revela, mesmo com a CNN tendo o melhor desempenho em relação aos outros, ela apresenta dificuldades em distinguir entre algumas classes específicas. A Figura 31 também exibe a matriz de confusão para os outros modelos, cujos resultados podem ser interpretados de maneira similar aos da CNN.

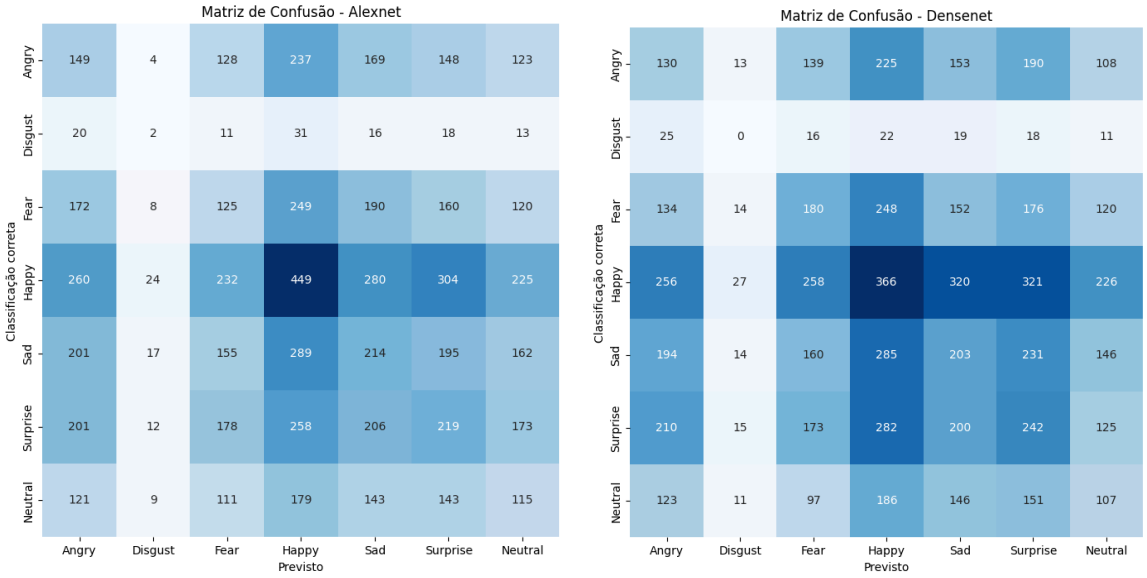

Figura 31: Matriz de Confusão da Alexnet e da Densenet.

5. Conclusão

Esta dissertação apresentou o VIRE, um sistema de análise de imagens projetado para o reconhecimento de emoções humanas e auxílio na saúde mental. A pesquisa explorou os principais conceitos nas teorias de classificação de emoções e utilizou técnicas avançadas de aprendizado de máquina, visão computacional e aprendizado profundo para implementar o módulo de reconhecimento, que é baseado em uma rede neural convolucional.

Para que o núcleo do sistema fosse capaz de classificar imagens de forma satisfatória, a pesquisa propôs o treinamento e a análise de três arquiteturas diferentes: AlexNet, DenseNet e uma CNN. O treinamento de todos os modelos foi realizado utilizando o conjunto de dados FER2013, que contém imagens em preto e branco classificadas em sete emoções básicas.

O treinamento foi realizado em duas fases complementares, em que cada fase visava mitigar problemas identificados na fase anterior. Ambas as fases seguiram a mesma sequência de etapas: busca dos melhores hiperparâmetros utilizando o algoritmo Hyperopt; treinamento de cada modelo com os hiperparâmetros encontrados; análise dos modelos com base nas métricas de loss e acurácia para os conjuntos de dados de treinamento e validação; e, finalmente, resumo e conclusão da etapa.

Esta pesquisa tem uma natureza prática, ou seja busca demonstrar a viabilidade de ideias por meio da aplicação prática (DEMO, 1985). Dessa forma foram utilizadas diversas tecnologias e bibliotecas, para propor o núcleo reconhecedor do sistema VIRE, sendo que as principais foram: (i) a linguagem de programação Python, devido à sua vasta gama de

bibliotecas voltadas para o desenvolvimento de inteligência artificial; (ii) TensorFlow 2.10 e Keras, bibliotecas utilizadas para a construção e treinamento de redes neurais; (iii) NumPy, para manipulação e análise dos dados; (iv) Matplotlib, para plotagem dos resultados; e (v) Hyperopt, para otimização dos hiperparâmetros. Além disso, ela foi conduzida em um ambiente de desenvolvimento com a seguinte configuração: processador Intel Core i5-7300HQ, GPU NVIDIA GeForce GTX 1050, 16GB de RAM com velocidade de 2400 MHz, sistema operacional Windows 10, utilizando VSCode e Jupyter Notebook para experimentação e execução.

Assim foi possível fazer com que todos os modelos fossem capazes de reconhecer emoções, embora com variações de desempenho entre eles. Todos alcançaram uma acurácia de cerca de 60% e demonstraram uma capacidade razoável de generalização para novos dados. No entanto, os resultados indicam que o melhor dos modelos avaliados ainda enfrenta desafios com falsos positivos e classificações equivocadas, o que pode comprometer a viabilidade do sistema proposto. Ainda assim, ao final dos treinamentos, a CNN mostrou uma tendência a superar o overfitting ao longo das épocas, mantendo bons indicadores de desempenho e se destacando como o melhor candidato para compor o núcleo reconhecedor do VIRE.

Como trabalho futuro para esta pesquisa, é possível citar os seguintes tópicos: (i) aplicação de outros métodos de aumento de dados no conjunto de treinamento; (ii) realizar o treinamento utilizando outros datasets de reconhecimento de emoções; e (iii) comparar o desempenho com outras arquiteturas de CNN. Essas iniciativas poderão oferecer novas perspectivas sobre o reconhecimento de emoções e possibilitar a proposição de um módulo reconhecedor com maior capacidade de reconhecimento em comparação com os analisados aqui e tornar a implementação do sistema possível para uma aplicação real de acompanhamento emocional.

REFERÊNCIAS BIBLIOGRÁFICAS

AMERICAN PSYCHIATRIC ASSOCIATION et al. DSM-5: Manual diagnóstico e estatístico de transtornos mentais. Artmed Editora, 2014.

CHOLLET, François. Deep Learning with Python. Shelter Island: Manning Publications, 2018.

DALGALARRONDO, Paulo. Psicopatologia e semiologia dos transtornos mentais. 3. ed. Porto Alegre: Artmed, 2019.

DARWIN, Charles. A expressão das emoções no homem e nos animais. São Paulo: Companhia das Letras, 2000 [1872].

DEL PORTO, J. A. Conceito e diagnóstico. Revista Brasileira de Psiquiatria, v. 21, supl. 1, p. 5, 1999. Disponível em: https://www.scielo.br/j/rbp/a/dwLyt3cv3ZKmKMLXv75Tbxn/. Acesso em: 31 de maio de 2024.

DEMO, Pedro. Introdução à metodologia da ciência. 2. ed. São Paulo: Atlas, 1985.

DUMITRU, Ian; GOODFELLOW, Will; CUKIERSKI, Yoshua; BENGIO. Challenges in representation learning: facial expression recognition challenge. Kaggle, 2013. Disponível em: https://kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge. Acesso em: 05 de março de 2024.

EKMAN, P.; FRIESEN, W. V. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, v. 17, n. 2, p. 124-129, 1971.

EKMAN, Paul. Emotions Revealed. New York: Times Books, 2003.

FURTADO, Maria Inês Vasconcellos. Redes neurais artificiais. Ponta Grossa (PR): Atena Editora, 2019.

HOSSEINI, S. S.; YAMAGHANI, M. R.; ARABANI, S. P. A review of the methods of recognition multimodal emotions in sound, image and text. International Journal of Applied Operational Research, v. 12, n. 1, p. 29-41, Winter 2024.

HUANG, Gao; LIU, Zhuang; VAN DER MAATEN, Laurens; WEINBERGER, Kilian Q. Densely Connected Convolutional Networks. In: IEEE Conference on Computer Vision and

Pattern Recognition (CVPR), 2017, Honolulu, HI, USA. Proceedings... Los Alamitos, CA: IEEE Computer Society, 2017. p. 2261-2269.

JAMES, W. The principles of psychology. New York: Holt, 1890.

MIGUEL, Fabiano Koich. Psicologia das emoções: uma proposta integrativa para compreender a expressão emocional. Psico-USF, São Paulo, v. 20, n. 1, p. 114-126, jan.-abr. 2015. Disponível em: https://www.scielo.br/j/pusf/a/FKG4fvfsYGHwtn8C9QnDM4n/. Acesso em: 31 de maio de 2024.

NOGUEIRA, K. A. Estudo de respostas emocionais às cores no contexto de cartazes de cinema. Design e Tecnologia, v. 8, n. 15, p. 1-11, 30 jun. 2018.

PLUTCHIK, Robert. Emotions and life: Perspectives from psychology, biology and evolution. Washington, DC: American Psychological Association, 2002.

RUSSELL, J. A. A circumplex model of affect. Journal of Personality and Social Psychology, v. 39, n. 6, p. 1161–1178, 1980. DOI: 10.1037/h0077714.

RUSSELL, Stuart J.; NORVIG, Peter. Inteligência artificial. Tradução de Regina Célia Simille. Rio de Janeiro: Elsevier, 2013.

SOLOMON, Andrew. O demônio do meio-dia: uma anatomia da depressão. Tradução de Myriam Campello. 2. ed. São Paulo: Companhia das Letras, 2014.