Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Abordagens de classificação binária no auxílio do problema de reconhecimento de expressões faciais

Tiago Luiz Ferreira de Carvalho

JUIZ DE FORA
AGOSTO, 2022

Abordagens de classificação binária no auxílio do problema de reconhecimento de expressões faciais

Tiago Luiz Ferreira de Carvalho

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Saulo Moraes Villela

JUIZ DE FORA

AGOSTO, 2022

Abordagens de classificação binária no auxílio do problema de reconhecimento de expressões faciais

Tiago Luiz Ferreira de Carvalho

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Saulo Moraes Villela
D.Sc. em Engenharia de Sistemas e Computação

Marcelo Bernardes Vieira
D.Sc. em Ciência da Computação

Luiz Maurílio da Silva Maciel
D.Sc. em Engenharia de Sistemas e Computação

JUIZ DE FORA
12 DE AGOSTO, 2022

Aos meus pais e à minha irmã.

Resumo

Algoritmos de aprendizado profundo têm sido cada vez mais utilizados para tarefas relacionadas à Visão Computacional, dada sua vasta aplicabilidade prática. Uma das aplicações eficazes dessa ferramenta é no problema de reconhecimento de expressões faciais, que consiste em, a partir de imagens de rostos humanos, classificar as expressões em categorias como felicidade, medo ou surpresa. Diversos métodos já foram empregados em tentativas de obter o modelo computacional mais eficaz possível nessa tarefa. Neste trabalho, é apresentado um estudo sobre o estado da arte desse problema e sobre os principais métodos utilizados, além de ser proposto o uso de redes neurais especialistas em classificações binárias para mitigar erros específicos, visando maior acurácia. Obteve-se resultados discretos, com ganho ou perda de acurácia, a depender do limite de confiança pré-estabelecido.

Palavras-chave: Aprendizado de máquina, aprendizado profundo, visão computacional, reconhecimento de expressões faciais.

Abstract

Deep learning algorithms have been increasingly used for tasks related to Computer Vision, given their wide practical applicability. One of the effective applications of this tool is in the facial expression recognition problem, which is the task of classifying the expressions on face images into various categories, such as happiness, fear or surprise. Several methods have been employed in attempts to obtain the most efficient computational model possible for this task. In this work, a study is presented on the state of the art of this problem and on the methods used, and it is proposed the use of binary classification neural networks for avoiding specific mistakes, aiming at greater accuracy in the tests.

Keywords: Machine learning, deep learning, computer vision, facial expression recognition.

Agradecimentos

Aos meus pais, Áulus e Josiane, e à minha irmã, Marina, pelo apoio e amor incondicionais. Serei eternamente grato por todo o esforço e sacrifícios pessoais que vocês fizeram para que eu pudesse chegar até aqui.

Aos amigos e colegas que fiz ao longo dessa jornada, com quem compartilhei momentos em salas de aula, empresa júnior, estágios e inúmeras experiências enriquecedoras que foram essenciais para a minha formação.

À minha namorada, Letícia, por todo amor, suporte, carinho e companheirismo ao longo desses anos, mesmo com a distância.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso que contribuíram de algum modo para o meu enriquecimento pessoal e profissional. Em especial, ao professor Saulo, pela orientação e paciência.

À Tula.

"Though the course may change sometimes, rivers always reach the sea".

Led Zeppelin (Ten Years Gone)

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Apresentação e contextualização do problema

Um algoritmo de aprendizado de máquina (*machine learning*), conforme definido por Goodfellow, Bengio e Courville (2016), é um algoritmo que é capaz de aprender a partir de dados. Esse aprendizado pode se dar através do reconhecimento de padrões, de forma que o algoritmo consiga realizar tarefas como classificação de dados ou tomada de decisões condicionadas ao aprendizado obtido. Nesse contexto, aprendizado profundo (*deep learning*) é uma abordagem derivada do aprendizado de máquina, que nos permite lidar com problemas mais específicos e diferentes tipos de abstrações.

Nas últimas décadas, o uso de modelos de aprendizado profundo para atacar diversos problemas em diferentes áreas popularizou-se, se tornando uma ferramenta eficiente para tarefas relacionadas a Visão Computacional, o que foi favorecido também pelo recente avanço em termos de poder computacional.

Visão Computacional é um campo de pesquisa que busca desenvolver modelos que possibilitem aos computadores interpretarem dados de entrada em formato de imagens ou vídeos, obterem informações a partir desses dados e então tomarem ações a partir do conhecimento obtido, como a classificação de imagens e reconhecimento de objetos. Dentre os diversos problemas na área de Visão Computacional, um dos que tiveram significativo avanço nos últimos anos é o de reconhecimento de expressões faciais (*Facial Expression Recognition* – FER), que consiste em, a partir de imagens de rostos humanos, classificar as emoções expressas por eles em algumas categorias, como felicidade, medo ou surpresa.

1.2 Motivação

As expressões faciais desempenham um papel de grande importância na comunicação humana e através delas consegue-se enfatizar trechos de frases e dar outros sentidos e conotações ao que é falado, além de poder reconhecer e interpretar as intenções e emoções

das pessoas com quem se interage (STUART; BYRNE, 2004). Nesse contexto, a tarefa de reconhecer automaticamente expressões faciais em imagens e vídeos através de modelos computacionais se torna interessante e com grande aplicabilidade nas áreas de pesquisa de Visão Computacional, Interação Humano-Computador, dentre outras.

1.3 Objetivos

O presente trabalho tem como objetivo geral o estudo e a aplicação de técnicas de aprendizado de máquina e aprendizado profundo para o problema de reconhecimento de expressões faciais, com foco em classificar imagens de rostos humanos em 7 categorias: raiva, nojo, medo, felicidade, tristeza, surpresa e neutro. Ainda, é proposta aqui uma estratégia que consiste em definir redes neurais especialistas em classificações binárias, que vão ser invocadas pela rede neural principal para auxiliar na redução de confusões feitas por esta. Essa abordagem visa mitigar erros cometidos pela rede na classificação das imagens.

De forma específica, tem-se os seguintes objetivos:

- Desenvolver uma rede neural especialista na classificação binária de imagens nas classes medo e raiva;
- Incorporar a rede desenvolvida à implementação VGGNet, desenvolvida por Khaireddin e Chen (2021);
- Avaliar a contribuição da rede especialista para o desempenho do modelo proposto.

1.4 Organização

Este trabalho está dividido em 6 capítulos. No segundo, é apresentado o referencial teórico que fundamentou o desenvolvimento deste trabalho. Ainda, conceitos importantes sobre Inteligência Artificial e aprendizado profundo são definidos e brevemente discorridos.

No Capítulo 3, o problema é definido com mais detalhes, e são apresentadas as principais abordagens para atacá-lo. Este capítulo discorre sobre o funcionamento de sistemas modernos que realizam reconhecimento de expressões faciais, bem como suas principais técnicas e bases de dados usadas.

O Capítulo 4 apresenta a proposta desenvolvida neste trabalho, que consiste na chamada de redes especializadas na classificação binária das classes “medo” e “tristeza” para mitigar os erros do modelo na classificação das imagens dessas categorias. Também apresenta-se a análise que motivou essa proposta, além de detalhes de sua implementação.

No quinto capítulo, são expostos os experimentos realizados, tanto para reproduzir os dados tomados como base para comparação,apresentados por Khaireddin e Chen (2021), quanto para validar a hipótese formulada no Capítulo 4.

O último capítulo apresenta as considerações finais do trabalho, bem como algumas perspectivas para trabalhos futuros.

2 Fundamentação teórica

2.1 Expressões faciais

Ao estudarem expressões faciais e seu papel na comunicação, Ekman e Friesen (1978) definiram o Sistema de Codificação de Ação Facial (*Facial Action Coding System* – FACS), que descreve e categoriza expressões faciais de acordo com movimentos dos músculos faciais. Com o FACS, foram definidas seis emoções básicas que se pode interpretar a partir de expressões faciais: surpresa, raiva, medo, nojo, tristeza e felicidade. Considera-se também a expressão facial "neutra" como uma outra categoria, formando as sete expressões que os estudos sobre sistemas de reconhecimento de expressões faciais convencionalmente usam.

2.2 Aprendizado Profundo

2.2.1 Aprendizado de Máquina

Apesar de representarem um grande avanço e serem capazes de tratar vários problemas importantes da Ciência da Computação, os modelos tradicionais de Aprendizado de Máquina não são eficazes para lidar com problemas de natureza mais prática, como a classificação de objetos em imagens e vídeos ou o reconhecimento de voz. Aliado a isso, identificou-se a necessidade de métodos capazes de trabalhar de forma eficiente com grandes quantidades de dados complexos.

Nesse contexto, redes neurais com múltiplas camadas ocultas se mostraram eficientes, caracterizando uma abordagem chamada aprendizado profundo. Recentemente, aprendizado profundo se tornou o estado da arte para uma grande variedade de problemas ligados à Inteligência Artificial e é um tema cada vez mais pesquisado. Essa abordagem faz com que os modelos computacionais consigam aprender conceitos complexos com base em conceitos mais simples e se propõe a trabalhar com abstrações de alto nível através de estruturas hierárquicas com transformações e representações não-lineares.

Para que redes com várias camadas ocultas aprendam, entretanto, é necessário uma quantidade significativa de dados de treino, já que a qualidade de um sistema de aprendizagem de máquina depende diretamente da quantidade e da qualidade do conjunto de dados de treinamento (SIMARD et al., 2003). Nesse sentido, a falta de base de dados de qualidade para treinar os modelos é um dos principais obstáculos no desenvolvimento de novas redes neurais profundas, pois pode comprometer sua generalização para situações do mundo real, que podem diferir muito dos dados usados no treinamento das redes.

2.2.2 Redes Neurais Convolucionais

Como Goodfellow, Bengio e Courville (2016) definem, redes neurais convolucionais são aquelas que utilizam a operação de convolução em vez da multiplicação de matrizes em pelo menos uma camada, e sua eficiência se apresenta especialmente para tratar dados com uma topologia em grade, como imagens.

Matematicamente, a convolução é definida como uma operação linear entre duas funções que produz uma função resultante que expressa a soma do produto das funções de entrada, ao longo da região originada pela superposição delas, em função do deslocamento entre elas, conforme a Eq. (2.1):

$$s(t) = (x * w)(t) = \int x(a)w(t-a)da. \tag{2.1}$$

Nas redes neurais convolucionais, os argumentos são matrizes multidimensionais que representam, respectivamente, os dados de entrada da rede e o núcleo (*kernel*), e a convolução ocorre através de sucessivas operações de soma dos produtos com cada região da matriz de entrada que está sobreposta pelo núcleo, até que todas as regiões de dimensão igual a do núcleo sejam cobertas, como é ilustrado na Figura 2.1. O resultado da convolução é chamado de mapa de características. No caso de imagens coloridas, tem-se um núcleo correspondente a cada canal, geralmente de acordo com o padrão RGB, ou seja, com 3 canais. Os mapas de características servem como entrada para funções de ativação, e para as redes neurais modernas recomenda-se usar a função ReLu ou unidade linear retificadora (JARRETT et al., 2009), que anula os valores menores que zero e retorna os

valores maiores ou iguais a zero, como mostra a Eq. (2.2). O uso de funções de ativação não lineares como a ReLu permite que situações práticas sejam modeladas pelas redes neurais, uma vez que normalmente os problemas do mundo real não são lineares.

$$f(x) = max(0, x). \tag{2.2}$$

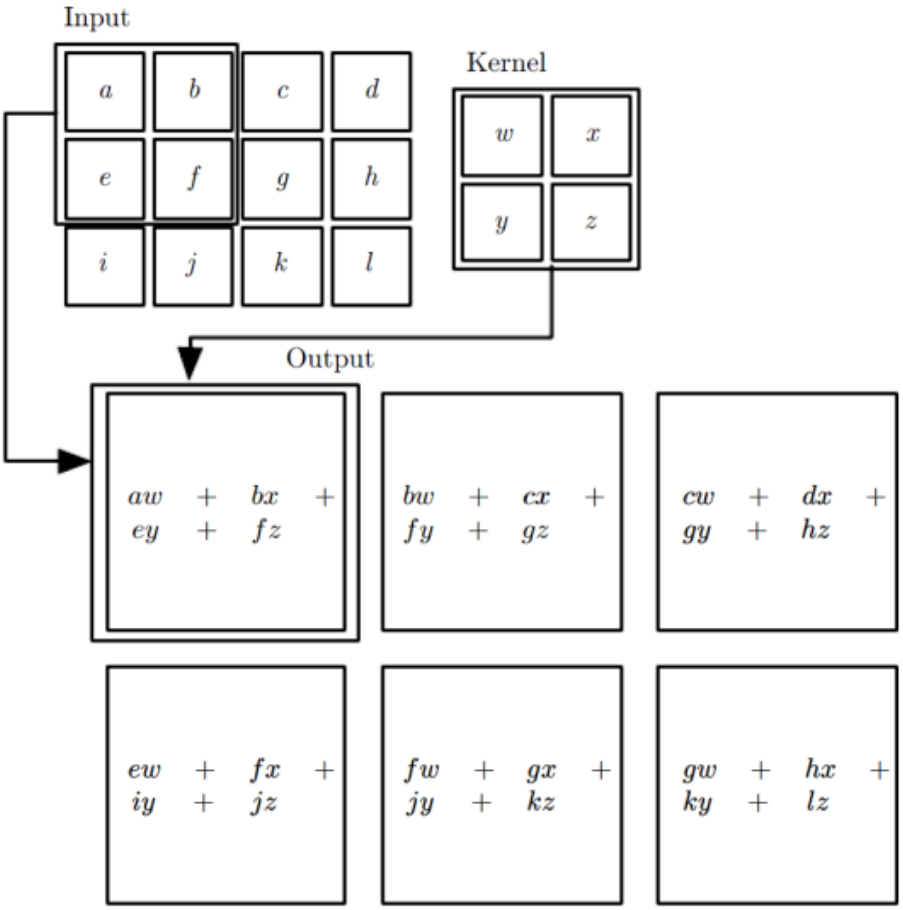


Figura 2.1: Operação de convolução 2D. Fonte: Goodfellow, Bengio e Courville (2016).

Nas redes neurais totalmente conectadas, cada elemento de uma camada se conecta a todos os outros elementos da camada seguinte. Dessa forma, o processamento se torna custoso para tratar dados de entrada das dimensões de imagens em aplicações práticas. Por exemplo, uma imagem de dimensões 48×48 píxeis em apenas um canal de cor (como escala de cinza), representada por uma matriz, teria 2304 neurônios em sua camada de entrada. À medida que todos esses neurônios se conectam com outras camadas da rede, computar e armazenar as operações e os parâmetros da rede neural se torna pouco eficiente.

Nas redes neurais convolucionais, este problema é mitigado devido a uma propriedade dessas redes chamada por Goodfellow, Bengio e Courville (2016) de interações esparsas. Ao fazer com que o núcleo (*kernel*) tenha dimensões menores que as dos dados de entrada, é possível fazer a rede detectar pequenos padrões e aprender características importantes, como bordas ou curvas relevantes, que ocupam apenas frações do tamanho original da imagem, e que servirão de insumos para que as camadas subsequentes aprendam a reconhecer padrões cada vez mais complexos. Dessa forma, o número de parâmetros armazenados e o número de operações realizadas são reduzidos, tornando o processamento mais rápido e consumindo menos memória.

Outra característica que faz com que as redes convolucionais sejam ainda mais eficientes é a aplicação das operações de agrupamento (*pooling*), que além de diminuir progressivamente o tamanho das entradas (e, consequentemente, o número de parâmetros armazenados), provê uma invariância a pequenos deslocamentos na imagem. Essa operação normalmente é feita através de uma das duas principais estratégias, *max-pooling* e *average-pooling*. Na primeira, uma área quadrada dos dados é resumida pelo maior elemento contido na área; na segunda, a área é resumida pela média aritmética de todos os valores ali contidos.

As arquiteturas de redes convolucionais apresentam camadas totalmente conectadas no final da rede para que os mapas de características sejam linearizados e utilizados para finalmente a rede classificar os dados de entrada em uma das categorias determinadas pelo problema modelado.

3 Reconhecimento de Expressões Faciais

O problema de reconhecimento de expressões faciais pode ser dividido em duas categorias: estático e dinâmico. Na primeira, a análise se dá exclusivamente sobre as informações espaciais das características da imagem de entrada, enquanto na segunda considera-se as informações temporais dos *frames* da sequência de imagens de entrada (um vídeo), indicando uma direção do movimento dos músculos faciais.

Com o progresso das pesquisas e aplicações de modelos de aprendizado profundo em várias áreas, o problema de reconhecimento de expressões faciais passou a ser mais explorado em base de dados (*datasets*) com imagens que não são adquiridas em ambientes controlados. Por exemplo a base de dados FER2013, introduzida por Goodfellow et al. (2013) e constituída a partir da *API* de busca de imagens da Google[1], fornece 28709 imagens de expressões faciais de pessoas fora de ambientes controlados, ou seja, em condições na natureza (*in the wild*), caracterizando uma vertente do problema de FER ainda mais desafiadora. Isso se dá devido à diversidade das imagens, como variações da pose dos rostos, de idade e etnia das pessoas, qualidade e iluminação nas imagens.

Entretanto, *datasets* com imagens em ambientes controlados continuam relevantes e são constantemente usados por pesquisadores para estudar e testar técnicas modernas de aprendizado profundo. Destaca-se a base de dados CK+ (LUCEY et al., 2010), que apresenta 593 vídeos de 123 pessoas diferentes esboçando as 7 emoções básicas. Apesar de conter apenas vídeos, esta base também é usada para estudar o problema de FER estático, de forma que apenas um dos últimos *frames* de cada vídeo é tomado como *input* para se fazer a análise. Destaca-se também a base JAFFE (LYONS et al., 1998), que contém imagens de 10 mulheres japonesas, também expressando as 7 emoções básicas.

Um modelo computacional para reconhecimento de expressões faciais geralmente é dividido em três etapas: pré-processamento das imagens, extração de características faciais e classificação de expressão faciais. A primeira etapa envolve o tratamento das imagens para lidar com variações no *dataset*, corrigir possíveis desalinhamentos e iluminação,

[1] https://www.google.com/

aplicação de técnicas de aumento de dados, dentre outras ações. A segunda etapa, como o nome sugere, extrai as características dos dados de entrada para a interpretação das expressões, e é feita através de métodos que podem ser divididos em duas categorias: aqueles para a análise de imagens estáticas e aqueles para a análise de sequência de imagens dinâmicas. A terceira etapa é a classificação das expressões faciais, que define um mecanismo para identificar e categorizar as expressões faciais, o que também pode ser feito através de diferentes métodos.

3.1 Pré-processamento

Nas bases de dados usadas para o estudo do problema de reconhecimento de expressões faciais, especialmente naquelas do tipo *in-the-wild*, as imagens apresentam variações que não interessam ao problema, como o fundo da imagem, detalhes de iluminação e posicionamento do rosto na imagem. Por isso, antes de começar o processo de aprendizagem da rede neural, convém realizar algumas padronizações.

3.1.1 Aumento de dados

Redes neurais que possuem várias camadas precisam de uma grande quantidade de dados para que a rede seja treinada o suficiente a ponto de conseguir generalizar os padrões reconhecidos e as características aprendidas. Caso não haja dados suficientes para um treino eficaz, o modelo não irá conseguir generalizar o suficiente para ter um desempenho satisfatório quando lidar com dados novos, o que faz com que a diferença entre a acurácia dos dados de treino e de teste seja expressiva, fenômeno conhecido como *overfitting*. Por isso, um dos principais desafios ao trabalhar com redes neurais profundas é a falta de bases de dados robustas para treinar os modelos.

Nesse cenário, a técnica de aumento de dados se apresenta como uma abordagem interessante para reduzir a ocorrência de *overfitting*. Aumento de dados (*data augmentation*) consiste em fazer uso de alguns métodos simples que fazem transformações geométricas e fotométricas nas imagens de um *dataset* de forma a aumentá-lo e fornecer mais dados de treino para uma rede neural, além de torná-lo mais invariante a mudanças

sutis nos dados de entrada.

3.1.2 Detecção facial

Dada uma imagem para servir de insumo de treino de uma rede neural convolucional, é conveniente delimitar a área de interesse da imagem, a partir da qual a rede irá, de fato, extrair características relevantes. Para isso, durante o pré-processamento é interessante detectar e remover as áreas que não são de interesse. No caso do problema de reconhecimento de expressões faciais, interessa à rede apenas o rosto na imagem. Portanto, nessa etapa é removido o fundo da imagem e qualquer outra região que não seja a face em questão. Para realizar essa tarefa, o método de Viola-Jones para detecção de faces é um dos mais usados, já que é eficaz, robusto e rápido (DABHI; PANCHOLI, 2016).

3.1.3 Normalizações

As variações das imagens em aspectos como iluminação e posição do rosto podem prejudicar o desempenho dos modelos de reconhecimento de expressões faciais. Normalizações de posicionamento e de iluminação são pertinentes nesse cenário, visando mitigar o impacto dessas variações.

Variações no posicionamento do rosto na imagem é um obstáculo esperado de se encontrar em um base de imagens *in the wild*, já que irá conter imagens de vários ângulos diferentes. Assim, o uso de técnicas de frontalização facial, como a apresentada por Hassner et al. (2015), são frequentemente empregadas para estimar uma imagem correspondente com o rosto da perspectiva frontal, com suas coordenadas 3D, para cada imagem tratada.

A iluminação nas imagens também tende a variar de forma significativa, dados os diferentes cenários de cada imagem, o que impacta diretamente no desempenho da rede neural. Por isso, diversas técnicas de tratamento da iluminação das imagens já foram estudadas e aplicadas no problema de FER. Por exemplo, Shin, Kim e Kwon (2016) avaliaram algoritmos como normalização baseada em difusão isotrópica, diferença de gaussianas e transformação discreta de cosseno na tarefa de reconhecimento de expressões faciais. Já Pitaloka et al. (2017) e Yu e Zhang (2015) usam equalização de histograma, método que

ajusta as intensidades das imagens para realçar seu contraste, acentuando alguns detalhes pouco visíveis anteriormente.

3.2 Extração de características

Além da definição da caixa delimitadora mínima da face na imagem, realizada na etapa de pré-processamento, alguns trabalhos também realizam a definição de pontos de referência facial (*facial landmarks*), que são pontos de interesse no rosto para a extração de características, possibilitando o processo de reconhecimento da expressão facial. Esse processo leva em conta a distância entre os pontos de referência facial e também os ângulos entre eles. Existem alguns algoritmos eficientes para a determinação dos pontos de referência, cuja análise não está no escopo deste trabalho, como *Active Appearance Model* (AAM) e *Multi-Task CNN* (MTCNN).

Desde o início dos anos 2000, sabe-se, à luz dos trabalhos (FASEL, 2002b) e (FASEL, 2002a), que redes neurais convolucionais são eficazes para trabalhar com o problema de FER, e até hoje são tidas como a melhor abordagem nesse contexto. Por isso, diversas abordagens para o uso de CNNs na extração de características faciais foram propostas. Em (SUN et al., 2015), foi sugerido o uso de CNNs baseadas em região (R-CNN) para o aprendizado das características. Para o problema de FER dinâmico, Fan et al. (2016) e Nguyen et al. (2017) utilizaram arquiteturas baseadas em convoluções 3D para extrair características que levam em consideração o aspecto temporal.

Além de CNNs, outros tipos de redes neurais já foram implementadas para aprender características no problema de reconhecimento de expressões faciais. Cai et al. (2021) usou redes generativas adversárias (GANs) em alguns *datasets* e obtiveram resultados compatíveis com o estado da arte. Redes neurais recorrentes (RNNs) foram usadas em um *ensemble* com redes convolucionais por Mao, Fan e Peng (2021).

3.3 Classificação

Essa é a etapa do modelo que, após ter aprendido as características faciais das imagens do *dataset*, define a categoria a qual cada imagem pertence, entre as 7 emoções básicas. Nesta

etapa, normalmente aplica-se a função *softmax* para obter as probabilidades finais de cada classe. Entretanto, alguns trabalhos sugeriram substituir essa função: máquinas de vetores suporte (SVMs) foram propostas por Tang (2013), enquanto Dapogny e Bailly (2018) propuseram o uso de *deep neural forests*, e ambos conseguiram resultados competitivos.

4 Abordagem proposta

O primeiro passo da pesquisa que originou este trabalho foi a investigação do estado da arte do problema de reconhecimento de expressões faciais. Como este tópico, assim como diversos outros ligados à Visão Computacional e Inteligência Artificial em geral, tem chamado cada vez mais atenção e sido mais pesquisado por diversas instituições de ensino e organizações privadas ao redor do mundo, seu estado da arte está em constante atualização. Dessa forma, a cada poucos meses surgem diversas novas propostas para otimizar os algoritmos usados, aproveitar melhor os dados disponíveis e, consequentemente, aprimorar a acurácia dos modelos. Nesse cenário, para este trabalho foi utilizada a plataforma *Papers With Code*[2] como principal ferramenta para pesquisar sobre os trabalhos mais recentes sobre o tema e os melhores algoritmos disponíveis. Após o estudo sobre as principais técnicas utilizadas para atacar o problema de FER, bases de dados mais populares e redes com melhores desempenhos, este trabalho teve seu foco voltado para o estudo da FER2013 (GOODFELLOW et al., 2013), que é um dos *datasets* mais robustos e mais estudados dentro do tópico, cuja divisão dos dados pode ser observada na Tabela 4.1. Mais especificamente, foi estudada uma implementação da rede VGG-Net, posicionada como um dos modelos de maior acurácia na base de dados FER2013 em 2021, de acordo com a plataforma *Papers With Code*, e proposta por Khaireddin e Chen (2021). Essa rede utiliza a arquitetura VGG, proposta por Simonyan e Zisserman (2014) e que se tornou uma das mais populares para tarefas de Visão Computacional. A escolha desse trabalho passou também pelo fato de este ter seu código original em um repositório público e disponível para acesso.

O código disponibilizado por Khaireddin e Chen (2021) foi executado em um ambiente *Anaconda*[3], em um notebook com um processador AMD Ryzen 5 5600H, com 16GB de memória RAM e placa de vídeo NVIDIA GeForce RTX 3050 com 4GB de memória dedicada, no sistema operacional Windows 11. Entretanto, o resultado obtido não foi o

[2]https://paperswithcode.com/
[3]https://www.anaconda.com/products/distribution

Tabela 4.1: Base de dados FER2013, apresentada em Goodfellow et al. (2013).

Emoção	Treino	Validação	Teste	Total
0 - Raiva	3995	491	467	4953
1 - Nojo	436	55	56	547
2 - Medo	4097	528	496	5121
3 - Felicidade	7215	879	895	8989
4 - Tristeza	4830	594	653	6077
5 - Surpresa	3171	416	415	4002
6 - Neutro	4965	626	607	6192
Total	28709	3589	3589	35887

mesmo dos autores originais, pois foi obtida uma acurácia ligeiramente menor, como é exposto no Capítulo 5. Essa discrepância pode ser explicada pela possível diferença de *hardware* entre os experimentos, porém não foi possível obter a informação de qual foi exatamente a configuração das máquinas em que foram executados os testes em Khaireddin e Chen (2021). De qualquer forma, neste trabalho será considerado o resultado obtido nos experimentos realizados, e não o resultado relatado pelos autores, a fim de validar futuras comparações com outros resultados obtidos no mesmo ambiente de execução.

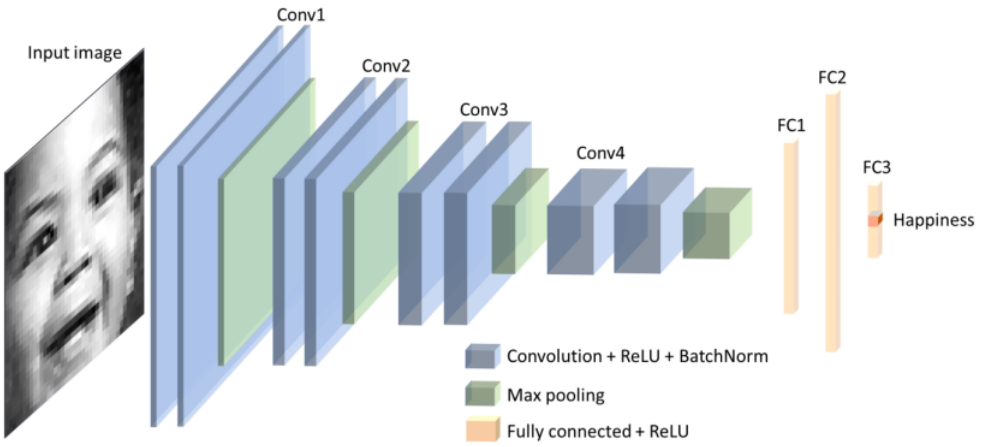


Figura 4.1: Arquitetura VGGNet. Uma imagem contendo uma expressão facial é dada como entrada para a rede neural, que possui quatro blocos de operações de convolução, que extrairão características da imagem. Por fim, camadas totalmente conectadas irão classificar a imagem em uma das sete emoções básicas. Fonte: Khaireddin e Chen (2021).

Apesar da pequena discrepância entre os resultados, em ambos é notória a quantidade de erros cometidos pela rede ao categorizar erroneamente imagens na classe 4 quando deveriam ser da classe 2, ou seja, com grande frequência a rede classifica como “triste” imagens rotuladas como “medo”.

Para analisar o comportamento da rede nesses erros específicos da classe 4 com a classe 2, foram analisadas as probabilidades resultantes da última camada da rede neural convolucional, através da função *softmax*, imediatamente antes de ser realizada a etapa de classificação. Com essa análise, foi investigada a natureza dessas confusões entre as classes 2 e 4: se são frutos de uma incerteza razoavelmente balanceada nas probabilidades ou se a rede de fato tinha uma alta confiança da resposta que estava gerando.

Diante desse cenário, este trabalho propõe uma estratégia para mitigar os impactos dessa confusão, com intuito de recuperar parte dessas classificações erradas entre as classes 2 e 4, através de redes especializadas em classificações binárias. A proposta consiste em definir uma rede neural convolucional com a mesma arquitetura da VGGNet usada no modelo original, porém que fosse especializada apenas em diferenciar imagens rotuladas como "medo" ou "tristeza". Assim, seria possível invocar essa rede especialista quando o modelo houvesse definido que as duas classes com maiores probabilidades fossem essas duas, porém com uma baixa confiança na resposta. Com a chamada para a rede especialista, que deverá ter uma maior capacidade de distinguir entre essas duas categorias, a rede principal pode delegar a tarefa de classificar as imagens que não foram categorizadas com precisão, com intuito de recuperar parte desses erros. Com essa estratégia, espera-se aumentar a acurácia total do modelo, tornando-o mais competente em predizer a expressão facial correta a partir de imagens.

A rede especialista em classificações binárias foi implementada a partir do projeto original de Khaireddin e Chen (2021), a fim de aprimorar seu desempenho na base de dados FER2013, utilizando a linguagem Python[4] e sua biblioteca PyTorch[5] como principais tecnologias. A tarefa de implementação dessa rede não foi um desafio, uma vez que foi aproveitada toda a arquitetura da rede VGGNet, porém com as alterações pontuais para que fossem consumidas apenas as imagens da base FER2013 das classes desejadas e que a classificação ao final da rede fosse feita de forma binária.

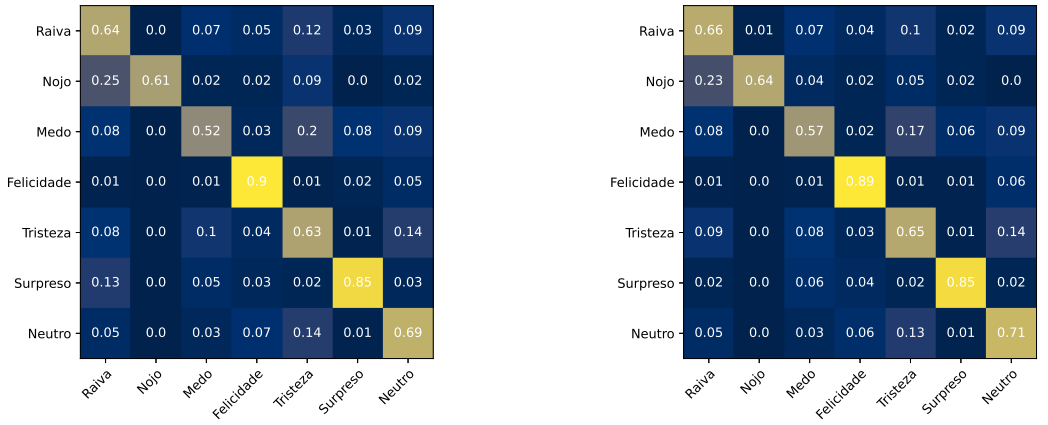
Além da implementação da rede especialista, também foram pontuais os ajustes na rede principal para que ela pudesse chamar a nova rede auxiliar. Estruturas condicionais foram adicionadas na última camada da rede neural para que, para uma dada

[4]https://www.python.org/
[5]https://pytorch.org/

imagem, quando as duas maiores probabilidades fossem das classes 2 e 4, em qualquer ordem, e a maior confiança delas fosse abaixo de um limite pré-estabelecido, a rede especialista fosse invocada para classificar a imagem em questão. Assim, apesar de ter sido considerada a opção de implementar uma estrutura de voto ponderado, com votos das duas redes, optou-se por tomar apenas o voto da rede especialista nos casos em que ela for invocada.

5 Experimentos e resultados

Ao reproduzir o experimento feito por Khaireddin e Chen (2021), foram encontradas discrepâncias entre o resultado obtido, representado pela Figura 5.1a, e aquele reportado pelos autores, representado pela Figura 5.1b. Ao final da execução, com os mesmos parâmetros usados no projeto original, o desempenho da VGGNet na classificação das expressões faciais atingiu uma acurácia de 71,75%, enquanto os autores relataram uma acurácia de 73,28%.



(a) Implementação deste trabalho. (b) Khaireddin e Chen (2021).

Figura 5.1: Diferentes resultados.

Na matriz de confusão dos resultados obtidos, é possível notar a grande quantidade de imagens rotuladas como “medo” (classe 2) que são erroneamente classificadas como “tristeza” (classe 4): das 496 imagens de teste da classe 2, 100 delas sofreram este tipo de erro, representando aproximadamente 20% do total. Há outras confusões relevantes que são apontadas pela matriz, como a das classes “raiva” e “medo” (item [1:0] da matriz), que representa 25% do total de imagens da classe 1. Entretanto, em números absolutos, essa confusão não é tão expressiva: foram 14 imagens erroneamente categorizadas na como “raiva” em vez de “medo”.

Ao analisar como se deram essas confusões, é possível notar que, apesar de a maioria desses erros acontecerem com a rede tendo uma alta confiança na resposta, há algumas ocorrências em que a probabilidade atribuída à categoria certa (classe 2) é ra-

zoavelmente alta, o que caracteriza uma situação de incerteza da rede sobre a resposta correta. Com isso, pode-se interpretar essas ocorrências como erros que são possíveis de serem recuperados, com o auxílio de uma rede especialista na classificação binária entre as classes envolvidas.

Tabela 5.1: Alguns dos erros de classificação entre a classes 2 e 4 em que a rede apresentou uma baixa confiança no resultado.

Classe correta	Classe escolhida	Probab. Classe 2	Probab. Classe 4
2	4	46%	52%
2	4	39%	61%
2	4	38%	62%
2	4	38%	62%
2	4	27%	64%

Após realizar os ajustes detalhados no Capítulo 4 e treinar a rede especializada em classificar corretamente imagens das emoções medo e tristeza, essa rede obteve um resultado final de 77,55% de acurácia, tendo classificado corretamente 68,15% das imagens da classe 2 e 84,69% da classe 4. É interessante notar que, apesar da acurácia em relação às imagens da classe 2 ser menor do que a acurácia total da rede generalista, a acurácia total da rede especialista é significativamente maior.

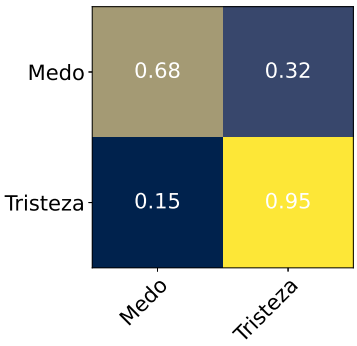


Figura 5.2: Matriz de confusão da rede especialista nas classes 2x4.

Uma vez treinada a rede especialista e realizados os ajustes na rede generalista comentados no Capítulo 4, foi possível fazer o experimento para validar a hipótese formulada de que seu uso para classificar imagens em que a rede original tem baixa confiança traria ganhos de acurácia total, já que, como é visto na Figura 5.2, a acurácia de fato é maior nessa nova rede do que na original. Foram utilizados os hiperparâmetros salvos a

partir das execuções feitas para reproduzir o resultado do trabalho original de Khaireddin e Chen (2021), ou seja, foram feitas execuções com o modelo já pré-treinado.

Como pode ser observado na Tabela 5.2, foram obtidos resultados diferentes de acordo com o limite da confiança para acionar a rede especialista, em alguns o resultado foi positivo e, em outros, negativo. Para todos os limites de confiança acima de 60%, a rede especialista errou mais amostras do que acertou, caracterizando um cenário em que atrapalhou o desempenho da rede original mais do que ajudou, fazendo com que a acurácia caísse em até quase 0,1%, com a diminuição de 3 acertos nas 3589 imagens de teste. Entretanto, nos melhores casos, com o limite de confiança igual ou abaixo de 55%, foram "recuperadas" 2 imagens que estavam erroneamente classificadas, representando um ganho de 0,05% absoluto, o que levou a acurácia para 71,78%.

Tabela 5.2: Resultados dos testes com a chamada à rede especialista com diferentes limites de confiança.

Limite de confiança	Imagens recuperadas (%)	Acurácia resultante
0,50	1	71,77%
0,55	2	**71,80**%
0,60	0	71,75%
0,65	-2	71,69%
0,70	-2	71,69%
0,75	-3	71,66%
0,80	-3	71,66%

Em linhas gerais, o resultado obtido foi discreto, pois o ganho de acurácia (e a perda, nos piores casos) não foi expressivo. Porém, é um resultado válido, uma vez que houve de fato um aumento na acurácia total, ainda que pequeno.

6 Conclusão

Neste trabalho foi desenvolvido um estudo sobre o problema de reconhecimento de expressões faciais (*Facial Expression Recognition* – FER). O problema foi formalizado como a tarefa de classificar imagens ou vídeos de rostos humanos, através de modelos computacionais, entre 7 categorias, representadas pelas emoções básicas introduzidas por Ekman e Friesen (1978): raiva, nojo, medo, felicidade, tristeza, surpresa e a emoção neutra.

Com o recente e expressivo avanço no campo de aprendizado de máquina, diversos métodos em tarefas de Visão Computacional foram propostos e aprimorados. Neste trabalho foram expostas e brevemente comentadas algumas das principais técnicas utilizadas para tratar o problema de reconhecimento de expressões faciais ao longo dos últimos anos.

Modelos computacionais modernos para esse problema normalmente são divididos em três etapas: pré-processamento das imagens que serão tratadas, extração das características faciais e, por fim, classificação das imagens. Essas etapas foram descritas, bem como as principais técnicas utilizadas em cada uma delas, no Capítulo 3.

Aqui também foi apresentada uma nova abordagem para aumentar a acurácia total de um modelo de reconhecimento de expressões faciais. Propõe-se tratar confusões específicas feitas pelo modelo através de uma rede neural convolucional especializada na classificação binária entre as classes “medo” e “tristeza”. Desta forma, espera-se que essa rede especialista tenha uma acurácia maior para distinguir entre duas classes específicas do que a rede generalista usada para o problema como um todo. Assim, a rede especialista pode ser invocada quando pertinente e auxiliar na tomada de decisão para classificar imagens que geram incertezas na rede principal.

Foram observados resultados discretos, que podem ser negativos ou positivos, a depender do limite de confiança pré-estabelecido para que a rede especialista seja invocada. No melhor caso, foi possível obter uma acurácia total de 71,80%, o que representa um ganho de 0,05% em relação à acurácia reproduzida pelo trabalho tomado como base ((KHAIREDDIN; CHEN, 2021)).

Para trabalhos futuros, é natural a extensão da ideia de uma rede especialista em uma confusão específica para o uso de várias redes especialistas, uma para cada combinação das emoções. Ou seja, ter-se-ia um número de redes especialistas igual à combinação simples de 7 emoções tomadas 2 a 2, dado pelo seguinte coeficiente binomial:

$$\binom{7}{2} = \frac{7!}{2!(7-2)!} = 21$$

Dessa forma, é possível obter um ganho, mesmo que também discreto, em todos os tipos de confusões feitas pela rede principal, com a expectativa de aumentar de forma mais significativa a acurácia total da rede.

Também é pertinente a investigação de possíveis ajustes na arquitetura da rede especialista para que seja mais assertiva na classificação binária, bem como possíveis algoritmos e otimizadores que possam ser mais adequados para uma classificação entre 2 classes ao invés de 7, como na VGGNet utilizada em Khaireddin e Chen (2021).

Uma abordagem que também pode ser proveitosa é o uso de redes especializadas em classificações “um contra todos”, ao invés da “um contra um” abordada neste trabalho. Dessa forma, seriam necessárias apenas 7 redes especialistas, ao invés de 21, o que reduz drasticamente a complexidade e o esforço do projeto.

Bibliografia

CAI, J. et al. Identity-free facial expression recognition using conditional generative adversarial network. In: IEEE. *2021 IEEE International Conference on Image Processing (ICIP)*. [S.l.], 2021. p. 1344–1348.

DABHI, M. K.; PANCHOLI, B. K. Face detection system based on viola-jones algorithm. *International Journal of Science and Research (IJSR)*, v. 5, n. 4, p. 62–64, 2016.

DAPOGNY, A.; BAILLY, K. Investigating deep neural forests for facial expression recognition. In: IEEE. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. [S.l.], 2018. p. 629–633.

EKMAN, P.; FRIESEN, W. V. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

FAN, Y. et al. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In: *Proceedings of the 18th ACM international conference on multimodal interaction*. [S.l.: s.n.], 2016. p. 445–450.

FASEL, B. Head-pose invariant facial expression recognition using convolutional neural networks. In: IEEE. *Proceedings. Fourth IEEE international conference on multimodal interfaces*. [S.l.], 2002. p. 529–534.

FASEL, B. Robust face analysis using convolutional neural networks. In: IEEE. *Object recognition supported by user interaction for service robots*. [S.l.], 2002. v. 2, p. 40–43.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. ⟨http://www.deeplearningbook.org⟩.

GOODFELLOW, I. J. et al. Challenges in representation learning: A report on three machine learning contests. In: SPRINGER. *International conference on neural information processing*. [S.l.], 2013. p. 117–124.

HASSNER, T. et al. Effective face frontalization in unconstrained images. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 4295–4304.

JARRETT, K. et al. What is the best multi-stage architecture for object recognition? In: *2009 IEEE 12th International Conference on Computer Vision*. [S.l.: s.n.], 2009. p. 2146–2153.

KHAIREDDIN, Y.; CHEN, Z. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint arXiv:2105.03588*, 2021.

LUCEY, P. et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: IEEE. *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. [S.l.], 2010. p. 94–101.

LYONS, M. et al. Coding facial expressions with gabor wavelets. In: IEEE. *Proceedings Third IEEE international conference on automatic face and gesture recognition*. [S.l.], 1998. p. 200–205.

MAO, S.; FAN, X.; PENG, X. Spatial and temporal networks for facial expression recognition in the wild videos. *arXiv preprint arXiv:2107.05160*, 2021.

NGUYEN, D. et al. Deep spatio-temporal features for multimodal emotion recognition. In: IEEE. *2017 IEEE winter conference on applications of computer vision (WACV)*. [S.l.], 2017. p. 1215–1223.

PITALOKA, D. A. et al. Enhancing cnn with preprocessing stage in automatic emotion recognition. *Procedia computer science*, Elsevier, v. 116, p. 523–529, 2017.

SHIN, M.; KIM, M.; KWON, D.-S. Baseline cnn structure analysis for facial expression recognition. In: IEEE. *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. [S.l.], 2016. p. 724–729.

SIMARD, P. Y. et al. Best practices for convolutional neural networks applied to visual document analysis. In: EDINBURGH. *Icdar*. [S.l.], 2003. v. 3, n. 2003.

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

STUART, R. M.; BYRNE, P. J. The importance of facial expression and the management of facial nerve injury. *Neurosurgery Quarterly*, LWW, v. 14, n. 4, p. 239–248, 2004.

SUN, B. et al. Combining multimodal features within a fusion network for emotion recognition in the wild. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. [S.l.: s.n.], 2015. p. 497–502.

TANG, Y. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

YU, Z.; ZHANG, C. Image based static facial expression recognition with multiple deep network learning. In: *Proceedings of the 2015 ACM on international conference on multimodal interaction*. [S.l.: s.n.], 2015. p. 435–442.