UNIVERSIDADE FEDERAL DO PARANÁ

JORGE LUIZ DOS SANTOS CANUTO

RECONHECIMENTO DE EXPRESSÕES FACIAIS DE ALUNOS COM CNN

JANDAIA DO SUL
2023

JORGE LUIZ DOS SANTOS CANUTO

RECONHECIMENTO DE EXPRESSÕES FACIAIS DE ALUNOS COM CNN

Trabalho de Conclusão de Curso apresentado ao Curso de Licenciatura em Computação da Universidade Federal do Paraná como requisito para a obtenção do título de Licenciado em Computação.

Orientador: Prof. Dr. Rodrigo Clemente Thom de Souza

JANDAIA DO SUL
2023

UNIVERSIDADE FEDERAL DO PARANÁ

PARECER Nº **1/2023/UFPR/R/JA**
PROCESSO Nº 23075.009134/2023-12
INTERESSADO: @INTERESSADOS_VIRGULA_ESPACO@

Título: Reconhecimento de expressões faciais de alunos com CNN

Autor: Jorge Luiz dos Santos Canuto

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do grau no curso de Licenciatura em Ciência da Computação, aprovado pela seguinte banca examinadora.

RODRIGO CLEMENTE THOM DE SOUZA (Orientador)

HELENA MACEDO REIS

ROGÉRIO FERREIRA DA SILVA

Jandaia do Sul, 23 de fevereiro de 2023.

Documento assinado eletronicamente por **RODRIGO CLEMENTE THOM DE SOUZA**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 23/02/2023, às 14:02, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **HELENA MACEDO REIS**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 23/02/2023, às 14:03, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **ROGERIO FERREIRA DA SILVA**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 23/02/2023, às 14:04, conforme art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida aqui informando o código verificador **5312591** e o código CRC **7B00B234**.

Referência: Processo nº 23075.009134/2023-12 SEI nº 5312591

AGRADECIMENTOS

A Deus, por me permitir ultrapassar todos os obstáculos encontrados ao longo da minha vida acadêmica.

À minha mãe e meus irmãos, que me ajudaram a superar momentos difíceis e compreenderam a minha ausência durante todos esses anos longe de casa.

À minha namorada e companheira, Daniele Oliveira dos Santos, que esteve ao meu lado me apoiando, incentivando e acreditando em mim durante a minha jornada acadêmica.

Ao meu orientador, Rodrigo Clemente Thom de Souza, por todos os conselhos, pela ajuda, pela paciência e por toda confiança depositada em mim durante todos esses anos.

À professora Lucilene Lusia Adorno de Oliveira, por ser uma das pessoas mais especiais com quem convivi durante a graduação.

Aos meus professores, por todos os seus ensinamentos e pela paciência com a qual compartilharam seus conhecimentos e guiaram meu aprendizado.

Aos meus amigos, pela amizade e assistência demonstrada ao longo de todo o período da graduação.

Obrigado a todos aqueles que participaram, de forma direta ou indireta, da minha formação acadêmica.

RESUMO

Reconhecer as emoções dos alunos é fundamental para qualquer sistema educacional, pois permite projetar melhores estratégias de ensino e aprendizagem. Identificar as emoções pode ser uma tarefa simples para os humanos, entretanto, se torna uma tarefa complexa para as máquinas. Avanços recentes em computação afetiva, em inglês *Affective Computing* (*AC*), e visão computacional estão permitindo que máquinas percebam as expressões faciais humanas e consigam reconhecer as emoções a partir delas. A *AC* visa tornar as máquinas sensíveis às emoções humanas, utilizando, para isso, o reconhecimento de emoções. O reconhecimento de emoções visa identificar as emoções humanas, principalmente, por meio das expressões faciais. O reconhecimento das emoções baseado em expressões faciais pode ser realizado através de algoritmos de aprendizagem profunda conhecidos como Redes Neurais Convolucionais, em inglês *Convolutional Neural Network* (*CNN*). As *CNNs* são modelos de aprendizagem profunda aplicáveis a diversos problemas de visão computacional, dentre eles, temos o reconhecimento de expressões faciais baseado em classificação de imagens. Recentemente, uma técnica de aprendizagem de máquina conhecida como transferência de aprendizagem, em inglês *Transfer Learning* (*TL*), têm prometido ajudar no desempenho dos modelos de *CNN* em diversas tarefas. Diante disso, o presente trabalho visa avaliar o desempenho de modelos de *CNN* com *TL* aplicado ao problema de reconhecimento de expressões faciais. Como melhor resultado, obtivemos o desempenho de 65% de acurácia com o modelo *VGG16*.

Palavras-chave: *Facial Expression Recognition* (*FER*). *Affective Computing* (*AC*). *Convolutional Neural Network* (*CNN*). *Transfer Learning* (*TL*).

ABSTRACT

Recognizing students' emotions is fundamental for any educational system, as it allows designing better teaching and learning strategies. Identifying emotions may be a simple task for humans, however, it becomes a complex task for machines. Recent advances in affective computing (AC), and computer vision are allowing machines to perceive human facial expressions and be able to recognize emotions from them. AC aims to make machines sensitive to human emotions, using emotion recognition for that. Emotion recognition aims to identify human emotions, mainly through facial expressions. The recognition of emotions based on facial expressions can be performed through deep learning algorithms known as Convolutional Neural Networks (CNN). CNNs are deep learning models applicable to several computer vision problems, among them we have the recognition of facial expressions based on image classification. Recently, a machine learning technique known as transfer learning (TL) promises to help CNN models perform in various tasks. Therefore, the present work aims to evaluate the performance of CNN models with TL applied to the problem recognizing facial expressions. As a best result, we obtained a performance of 65% accuracy with the VGG16 model.

Keywords: Facial Expression Recognition (FER). Affective Computing (AC). Convolutional Neural Network (CNN). Transfer Learning (TL).

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AC	*Affective Computing*
Adaboost	*Adaptive Boosting*
CNN	*Convolutional Neural Network*
CV	*Computer Vision*
DL	*Deep Learning*
DBM	*Deep Boltzmann Machines*
DBN	*Deep Belief Networks*
DSC	*Depthwise Separable Convolutions*
FACS	*Facial Action Coding System*
FER	*Facial Expression Recognition*
FN	*False Negative*
FP	*False Positive*
GPU	*Graphics Processing Unit*
KNN	*K-Nearest Neighbors*
ML	*Machine Learning*
NAS	*Network Architecture Search*
PNN	*Probabilistic Neural Network*
RBM	*Restricted Boltzmann Machine*
SRC	*Sparse Representation-based Classifier*
SVM	*Support Vector Machine*
TL	*Transfer Learning*
TN	*True Negative*
TP	*True Positive*

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, vimos um crescimento exponencial de modelos de aprendizagem baseados na internet (*e-Learning*). A transição para as tecnologias online na educação oferece oportunidades para usar novas metodologias de aprendizagem e métodos de ensino mais eficazes (BAYLARI; MONTAZER, 2009). Além disso, a substituição dos sistemas tradicionais de educação pelo *e-Learning* oferece vários benefícios, como melhoria de desempenho e redução de custos (IMANI; MONTAZER, 2019).

Entretanto, com o aumento da adoção de sistemas de aprendizagem baseados na internet, gerou-se a necessidade de melhorar os métodos de acompanhamento dos alunos à distância. Uma das formas de acompanhamento dos alunos é o reconhecimento de emoções durante o processo de aprendizagem. Além disso, reconhecer as emoções dos alunos é importante, pois, na literatura, vários trabalhos comprovaram a interdependência entre desempenho de aprendizagem e as emoções (KOUAHLA *et al.*, 2022).

Nesse sentido, considerar as emoções do aluno é essencial para qualquer sistema de aprendizagem, principalmente em *e-Learning* (IMANI; MONTAZER, 2019). Entender as emoções humanas por meio das máquinas é papel da computação afetiva. A computação afetiva visa entender e simular as emoções humanas por meio do computador. Para isso, este campo de pesquisa utiliza-se principalmente do reconhecimento das emoções humanas (WANG *et al.*, 2022).

Dentre os diferentes meios para reconhecimento das emoções humanas, temos aqueles que são baseados em reconhecimento de expressões faciais e que utilizam métodos de visão computacional. A visão computacional, em inglês *Computer Vision* (*CV*), é um campo de pesquisa que visa replicar a visão humana em máquinas. Entre as diversas aplicações da *CV*, o reconhecimento de expressões faciais, em inglês *Facial Expression Recognition* (*FER*), vem ganhando destaque nos últimos anos. O *FER* visa analisar expressões da face humana para classificação em rótulos predeterminados (LI; LIMA, 2021; YANG *et al.*, 2018).

O *FER* baseado em aprendizagem profunda, em inglês *Deep Learning* (*DL*), é implementado, principalmente, usando imagens ou vídeos que contenham pistas faciais das emoções. Recentemente, as *CNNs* se tornaram uma das abordagens de *DL* mais notáveis, principalmente, para tarefas relacionadas à classificação e reconhecimento de imagens (WANG *et al.*, 2022; ZHU; CHEN, 2020).

Um dos grandes desafios dos modelos de *CNN* é a dependência de uma grande quantidade de dados rotulados. Visando mitigar a dependência de grandes conjuntos de dados e melhorar a eficiência dos modelos de *DL*, a transferência de aprendizagem, em inglês *Transfer Learning* (*TL*), é um método usado para transferir o conhecimento adquirido em uma tarefa para resolver outra (RIBANI; MARENGONI, 2019). No entanto, nem sempre a *TL* apresenta efeitos positivos, a depender do contexto de utilização, a *TL* pode, inclusive, piorar o desempenho dos modelos de *DL*, esse fenômeno é conhecido como transferência negativa.

Diante disso, esse trabalho tem como objetivo avaliar a técnica de *TL* em algortimos de *DL* para aplicação em *FER*, visando uma posterior utilização em imagens faciais de alunos. Reconhecer as emoções por meio de expressões faciais é fundamental para qualquer modelo de ensino, pois permite projetar melhores estratégias, modelos de aprendizagem adequados e, consequentemente, melhorar os ganhos na realização do processo de ensino e aprendizagem (IMANI; MONTAZER, 2019).

Para facilitar a compreensão, o presente trabalho está organizado em seis capítulos. O segundo capítulo revisa, na literatura, os conceitos de computação afetiva, reconhecimento de expressões faciais, visão computacional, redes neurais convolucionais e transferência de aprendizagem. O capítulo três apresenta os trabalhos relacionados à aplicação de *FER* no contexto educacional utilizando *CNNs*. A metodologia é apresentada no capítulo quatro. No capítulo cinco tratamos dos resultados. Por fim, no capítulo seis, apresentamos a conclusão e finalizamos o trabalho.

1.1 PROBLEMA DE PESQUISA

É notável que obter o *feedback* emocional dos alunos é importante para qualquer sistema de ensino. Nesse sentido, o *FER* pode utilizar algoritmos de *DL* em conjunto com a técnica de *TL* para melhorar seu desempenho e eficiência. Nesse contexto, o presente trabalho visa responder a seguinte questão de pesquisa: "A *TL* em modelos de *DL* trará melhores resultados em termos de acurácia, precisão, revocação e pontuação-F1, quando aplicada ao problema de *FER* em imagens?".

1.2 HIPÓTESE

Partimos da hipótese de que a técnica de *TL* em modelos de *DL* apresenta melhores resultados para o *FER* do que o treinamento desses modelos do zero. Fundamentamos nossa hipótese na ideia de que os parâmetros transferidos para os modelos de *DL* possuem uma melhor representação inicial dos recursos visuais das imagens faciais do que a inicialização aleatória dos pesos dos modelos.

1.3 ESCOPO

O escopo do presente trabalho consiste na utilização de métodos de aprendizagem profunda para *FER*. Neste trabalho, não serão utilizados métodos rasos, ou seja, métodos baseados em aprendizagem de máquina convencional. Além disso, devido ao alto custo computacional, não utilizaremos dados de vídeo, apenas dados de imagem da face humana.

1.4 JUSTIFICATIVA

É consenso que as emoções são de extrema importância para o ser humano, pois afetam todos os aspectos das nossas vidas, em particular, aspectos relacionados à aprendizagem. Há várias formas de reconhecer as emoções, dentre elas, temos as expressões faciais. As expressões faciais humanas são o principal canal para expressar emoções (ZHU; CHEN, 2020).

Reconhecer as expressões faciais é uma tarefa simples para os humanos, entretanto, se torna uma tarefa complexa para as máquinas. Avanços recentes no campo da computação afetiva e da visão computacional estão permitindo que máquinas consigam perceber as expressões faciais humanas e consigam inferir emoções a partir delas.

Atualmente, técnicas de *TL* prometem melhorar significativamente o desempenho de modelos de *DL*, o que possibilita a aplicação do *FER* no contexto educacional com alto desempenho. Reconhecer as expressões faciais para prever as emoções que elas representam é importante no cenário educacional. Portanto, os aspectos citados anteriormente justificam a implementação de um reconhecimento automático de emoções que seja eficiente e independente da observação face a face humana.

1.5 OBJETIVOS

1.5.1 OBJETIVO GERAL

O presente trabalho tem como objetivo geral avaliar a técnica de *TL*, em termos de acurácia, precisão, revocação e pontuação-F1, quando aplicada em algoritmos de *DL* para *FER* em imagens.

1.5.2 OBJETIVOS ESPECÍFICOS

- Obter conjuntos de dados de imagens de expressões faciais disponíveis na internet;
- Implementar algoritmos para reconhecimento de faces;
- Entender sobre técnicas de *DL* para *FER*;
- Implementar técnicas de *DL* para *FER*;
- Entender sobre a técnica de *TL*;
- Aplicar a técnica de *TL*.

2 REVISÃO DE LITERATURA

2.1 COMPUTAÇÃO AFETIVA

A computação afetiva, em inglês *Affective Computing* (*AC*), é um campo multidisciplinar que se relaciona e deliberadamente influencia as emoções, tornando as máquinas emocionalmente inteligentes (PICARD, 1997). Em outras palavras, a *AC* visar capacitar as máquinas para que elas sejam sensíveis às emoções humanas e também consigam reproduzi-las e regulá-las. Diante disso, o entendimento sobre as emoções humanas é fundamental para a *AC*.

As emoções são difíceis de reconhecer até para os próprios seres humanos, pois podem ser expressadas de diversas formas. Porém, para entendermos a *AC* é essencial compreender as emoções humanas e os principais aspectos envolvidos na sua manifestação. Nesse sentido, existem dois modelos de emoção na computação afetiva, a saber, o modelo de emoção categórica e o modelo de emoção dimensional (WANG *et al.*, 2022).

O modelo de emoção categórica parte da ideia de que todos os seres humanos possuem um conjunto inato de emoções básicas que são culturalmente reconhecíveis (WIJASENA; FERDIANA; WIBIRAMA, 2021). Um dos pioneiros do modelo de emoção categórica, Ekman, desenvolveu um paradigma usando seis emoções básicas: alegria, nojo, raiva, surpresa, tristeza e medo (SHARMA; DIWAKAR; ARYA, 2022). O modelo de emoção categórica proposto por Ekman é constantemente utilizado no campo da *AC*.

Por outro lado, no modelo de emoção dimensional as emoções são ilustradas em diferentes espaços dimensionais, como o espaço bidimensional que inclui recursos de valência e excitação, já no espaço tridimensional, inclui recursos de valência, excitação e poder (SHARMA; DIWAKAR; ARYA, 2022). Em outras palavras, no modelo dimensional uma emoção é definida com base em algumas dimensões previamente estabelecidas.

Além dos modelos de emoções, a *AC* envolve dois tópicos distintos: reconhecimento de emoções e análise de sentimentos (WANG *et al.*, 2022). A Análise de Sentimentos é uma tarefa do Processamento de Linguagem Natural que visa extrair sentimentos e opiniões de textos (BIRJALI; KASRI; BENI-HSSANE, 2021). Geralmente, a análise de sentimento tem como resultado final a polaridade de um texto.

Diferentemente da análise de sentimento, o reconhecimento de emoções visa identificar o estado emocional dos seres humanos (WANG *et al.*, 2022). Nesse sentido, o modelo de emoção categórica de Ekman é amplamente aceito e utilizado para estabelecer as seis classes de emoções básicas. Normalmente, o resultado final do reconhecimento de emoções é a classificação, ou seja, determinar a qual classe de emoção uma determinada instância pertence.

Para alcançar os objetivos estabelecidos, existem vários métodos para o reconhecimento de emoções utilizando computadores, dentre eles, os principais são: reconhecimento de voz, reconhecimento de expressão facial, reconhecimento de gestos e reconhecimento por meio de

sinais vitais (IMANI; MONTAZER, 2019). Vale ressaltar que as emoções humanas são expressas principalmente por meio de expressões faciais (55%), voz (38%) e linguagem (7%) (WANG *et al.*, 2022).

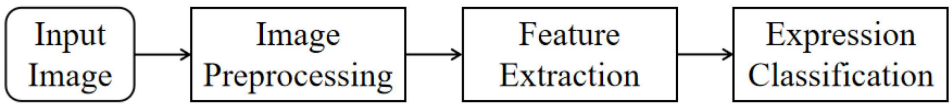
2.2 RECONHECIMENTO DE EXPRESSÕES FACIAIS

Há várias formas de comunicação social que podem envolver aspectos verbais e não verbais. Durante a comunicação, os seres humanos expressam seus desejos e emoções também por meio de expressões faciais. As expressões faciais se caracterizam como um dos sinais mais poderosos, naturais e universais para o ser humano transmitir seus estados emocionais e intenções (LI; DENG, 2022). Diante disso, reconhecer as expressões faciais humanas é uma característica desejada por diversos sistemas computacionais.

O reconhecimento de expressões faciais, em inglês *Facial Expression Recognition* (*FER*), refere-se principalmente à identificação de expressões que transmitem emoções básicas, como medo, felicidade, nojo e outros (KHAIREDDIN; CHEN, 2021). Além disso, para alcançar o objetivo de reconhecer expressões da face humana, o *FER* pode utilizar a abordagem convencional ou profunda.

A abordagem *FER* convencional é composta de três etapas principais, a saber, pré-processamento de imagem, extração de características e classificação de expressão (HUANG *et al.*, 2019; KOUAHLA *et al.*, 2022; REVINA; EMMANUEL, 2021). Cada etapa da abordagem convencional possui uma função específica e pode ser feita de forma isolada. A Figura 1 demonstra as etapas para realização do *FER* baseado na abordagem convencional.

Figura 1 – Etapas para abordagens *FER* convencionais.



Fonte: (HUANG *et al.*, 2019).

O pré-processamento é realizado visando melhorar o desempenho do sistema de *FER* (REVINA; EMMANUEL, 2021). Para isso, o pré-processamento elimina informações irrelevantes das imagens de entrada e aumenta a capacidade de detecção de informações relevantes (HUANG *et al.*, 2019). Geralmente, a etapa de pré-processamento deve ser realizada antes da extração de características.

A extração de características das imagens consiste em encontrar e representar características de interesse dentro de uma imagem, para processamento posterior (REVINA; EMMANUEL, 2021). Em visão computacional, a etapa de extração de características é uma das fases fundamentais para que o modelo de classificação tenha um bom desempenho.

A classificação é a etapa final de um sistema de *FER*, é nela que o classificador categoriza as imagens de expressões faciais que foram repassadas como entrada na etapa de

pré-processamento (LI; DENG, 2022; REVINA; EMMANUEL, 2021). Além disso, o desempenho de um sistema *FER* está associado à escolha de um modelo classificador. Os classificadores comumente usados em sistemas *FER* incluem *k-Nearest Neighbors* (*KNN*), *Support Vector Machine* (*SVM*), *Adaptive Boosting* (*Adaboost*), *Bayesian*, *Sparse Representation-based Classifier* (*SRC*) e *Probabilistic Neural Network* (*PNN*) (HUANG *et al.*, 2019).

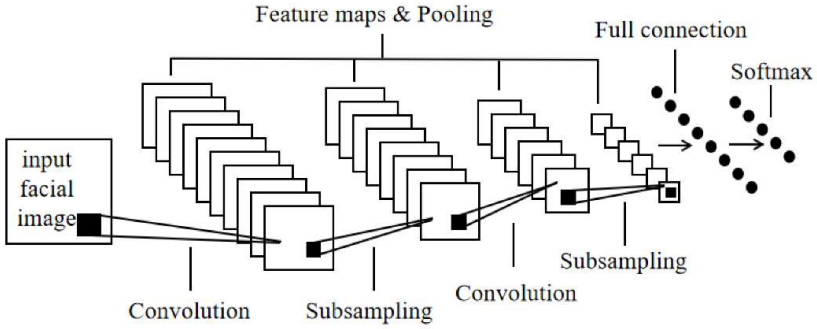
A abordagem convencional não necessita de grande poder computacional e tem a vantagem de ser um método interessante para pequenas amostras de dados. Entretanto, essa abordagem é altamente dependente da engenharia manual de recursos (HUANG *et al.*, 2019). Além disso, uma competição em 2013 mostrou que o desempenho da abordagem convencional é inferior ao desempenho da abordagem profunda (ZHU; CHEN, 2020).

A abordagem profunda do *FER* também possui três etapas principais para sua realização, a saber: pré-processamento, extração profunda de recursos de imagem e classificação. Na primeira etapa, o pré-processamento dos dados é realizado, o que geralmente é necessário para alinhar e normalizar as informações semânticas visuais transmitidas pela face humana (LI; DENG, 2022).

Na segunda etapa, é realizada a extração profunda de recursos dos dados de entrada, onde são capturadas as abstrações de alto nível, por meio de múltiplas transformações e representações não lineares. O resultado da segunda etapa é passado como entrada para a próxima etapa, ou seja, a classificação (LI; DENG, 2022).

A etapa de classificação recebe os dados de entrada e os classifica em uma das categorias previamente estabelecidas (LI; DENG, 2022). Geralmente, as etapas de classificação e extração profunda de recursos são realizadas por meio das redes neurais convolucionais, pois, recentemente, essa abordagem alcançou resultados impressionantes (ZHU; CHEN, 2020). Um exemplo de *FER* baseado em *CNN* pode ser visto na Figura 2.

Figura 2 – Abordagem *FER* profunda baseada em *CNN*.



Fonte: (HUANG *et al.*, 2019).

Diferentemente da abordagem convencional, a abordagem profunda realiza as etapas de extração profunda de recursos e classificação simultaneamente, por meio da atualização de peso iterativa, propagação reversa e otimização de erros (IMANI; MONTAZER, 2019; LI; DENG, 2022; ZHU; CHEN, 2020). Vale ressaltar que, para que a abordagem profunda tenha

sucesso, um grande número de amostras de dados de treinamento é necessário.

A abordagem profunda para *FER* pode ser dividida em *FER* profundo estático e *FER* profundo dinâmico, a depender do tipo de dado utilizado para o treinamento, podendo ser imagem estática ou sequência dinâmica de quadros. Nos métodos de *FER* profundo estático a representação de recursos é codificada apenas com informações espaciais de uma única imagem por vez. Por outro lado, no *FER* profundo dinâmico é considerada a relação temporal entre quadros contíguos na sequência de expressões faciais de entrada (LI; DENG, 2022).

Os métodos de *FER* profundo estático são vantajosos em comparação com os métodos de *FER* profundo dinâmico, pois são menos custosos computacionalmente. Além disso, a quantidade e disponibilidade de conjuntos de dados relacionados ao *FER* profundo estático é maior em comparação com o *FER* profundo dinâmico (LI; DENG, 2022).

As abordagens *FER* baseadas em aprendizado profundo são vantajosas em comparação com a abordagem convencional, pois reduzem a dependência da extração de características, empregando o aprendizado diretamente dos dados de entrada para o resultado da classificação. Além disso, esse tipo de abordagem é mais robusta, conseguindo superar desafios como diferenças de iluminação e oclusão (HUANG *et al.*, 2019).

2.3 VISÃO COMPUTACIONAL

A Visão Computacional, em inglês *Computer Vision* (*CV*), compreende métodos e técnicas através dos quais sistemas de visão artificial podem ser construídos e empregados de forma razoável em aplicações práticas. Esta área da ciência da computação inclui o *software*, *hardware* e técnicas de imagem necessárias para esses métodos (PATRÍCIO; RIEDER, 2018). A *CV* possui diversas áreas de aplicação que incluem reconhecimento facial, estimativa de pose, reconhecimento de atividade, vigilância por vídeo, biometria, produção de filmes, medicina, jogos de realidade aumentada, novas interfaces de usuário e muito mais (KHAN *et al.*, 2020; PULLI *et al.*, 2012).

No estágio inicial do desenvolvimento da *CV*, a construção desses sistemas concentrava-se em uma abordagem baseada em aprendizagem de máquina, em inglês *Machine Leaning* (*ML*), pois o aprendizado profundo, em inglês *Deep Learning* (*DL*), enfrentava dificuldades devido às limitações relacionadas a recursos computacionais (CHAI *et al.*, 2021). Posteriormente, com a disseminação das *Graphics Processing Unit* (*GPUs*), iniciou-se a *CV* baseada em *DL* (VOULODIMOS *et al.*, 2018).

Nos últimos anos, os métodos de *DL* demonstraram superar as técnicas anteriores de *ML* de última geração em vários campos, um dos casos mais notáveis do avanço das abordagens baseadas em *DL* é a *CV*. Os principais modelos de aprendizagem profunda para *CV* são *Deep Belief Networks* (*DBNs*), *Deep Boltzmann Machines* (*DBMs*), *Autoencoders Stacked* e *Convolutional Neural Networks* (*CNNs*) (VOULODIMOS *et al.*, 2018).

As *DBNs* são modelos generativos probabilísticos que fornecem uma distribuição de probabilidade conjunta sobre dados e rótulos observáveis. A *DBN* inicialmente emprega uma

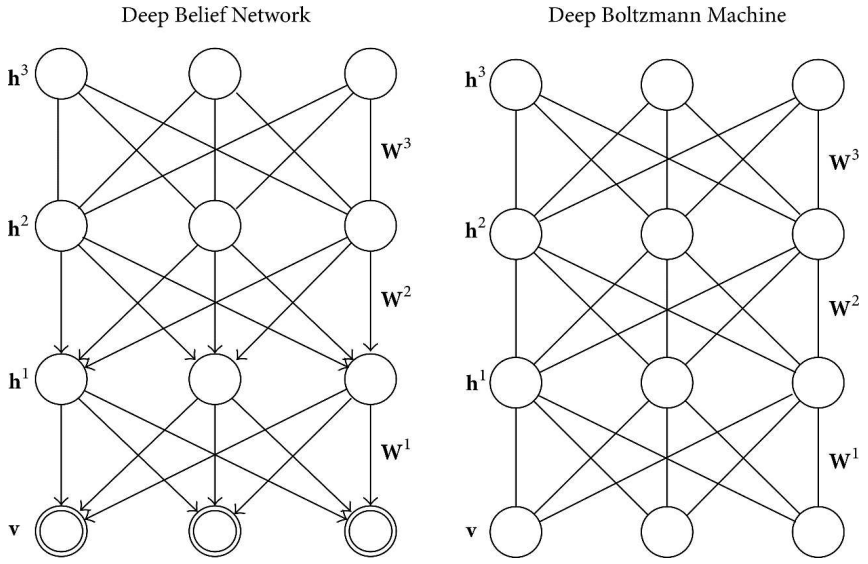
estratégia de aprendizado guloso, camada por camada, para inicializar a rede profunda e, na sequência, ajusta todos os pesos em conjunto com as saídas desejadas (HINTON, 2009).

Além disso, as *DBNs* possuem duas vantagens principais. Primeiro, aborda o desafio da seleção apropriada de parâmetros, garantindo assim que a rede seja inicializada adequadamente. Em segundo lugar, não há necessidade de dados rotulados, pois o processo não é supervisionado. Em contrapartida, as *DBNs* não levam em conta a estrutura bidimensional de uma imagem de entrada, o que pode afetar significativamente seu desempenho e aplicabilidade em problemas de *CV* (VOULODIMOS *et al.*, 2018).

As *DBMs* são outro tipo de modelo profundo que utiliza as *Restricted Boltzmann Machine* (*RBM*) como bloco de construção. A diferença entre as arquiteturas *DBN* e *DBM* é que nas *DBNs* as duas camadas superiores formam um grafo não direcionado e as camadas inferiores formam um grafo direcionado, por outro lado, nas *DBMs*, todas as conexões são não direcionadas (MELCHIOR; FISCHER; WISKOTT, 2016).

Como vantagens, as *DBMs* apresentam a captura de muitas camadas de representações complexas de dados de entrada e são apropriadas para aprendizado não supervisionado, mas também podem ser ajustadas para uma tarefa específica de maneira supervisionada. Entretanto, no que diz respeito às desvantagens das *DBMs*, uma das mais importantes é o alto custo computacional de inferência (VOULODIMOS *et al.*, 2018). Podemos observar as arquiteturas *DBN* e *DBM* na Figura 3.

Figura 3 – Arquitetura *DBN* e *DBM*.

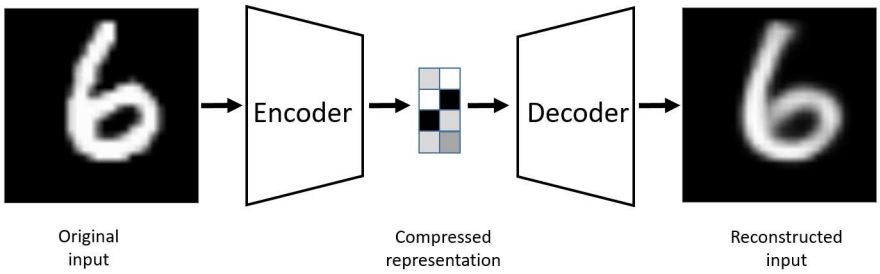


Fonte: (VOULODIMOS *et al.*, 2018).

O *autoencoder* de redução de ruído é uma versão estocástica do *autoencoder* onde a entrada é estocasticamente corrompida, mas a entrada não corrompida ainda é usada como alvo para a reconstrução. Em outras palavras, existem dois aspectos principais na função de um *autoencoder* de redução de ruído: primeiro, ele tenta codificar a entrada e, em seguida, tenta desfazer o efeito de um processo de corrupção aplicado estocasticamente à entrada do

autoencoder. É possível empilhar *autoencoders* de redução de ruído para formar um modelo profundo, alimentando o *autoencoder* de redução de ruído da camada atual com a saída do *autoencoder* da camada anterior (BANK; KOENIGSTEIN; GIRYES, 2020). A Figura 4 ilustra um exemplo de *autoencoder*.

Figura 4 – Exemplo de *autoencoder*.

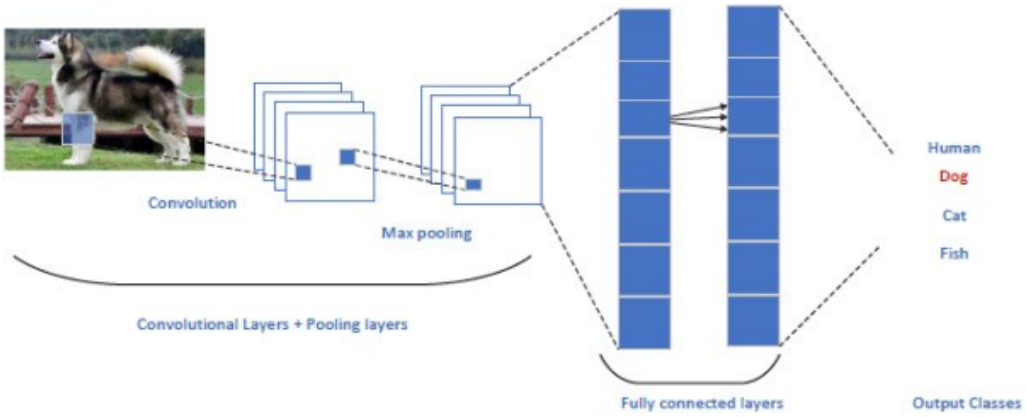


Fonte: (BANK; KOENIGSTEIN; GIRYES, 2020).

Um dos benefícios dos *autoencoders*, como um componente básico não supervisionado de uma arquitetura profunda é que, ao contrário dos *RBMs*, eles permitem praticamente qualquer parametrização das camadas, desde que o critério de treinamento seja contínuo nos parâmetros. Em contraste, uma das deficiências dos *autoencoders* é que eles não correspondem a um modelo generativo. Em modelos generativos, como *RBMs* e *DBNs*, as amostras podem ser extraídas para verificar as saídas do processo de aprendizagem (VOULODIMOS *et al.*, 2018).

Por fim, as *CNNs* consistem em camadas convolucionais, camadas de agrupamento e camadas totalmente conectadas. Nas camadas convolucionais, uma *CNN* usa vários *kernels* para convoluir toda a imagem e os mapas de recursos intermediários, gerando vários mapas de recursos finais. As camadas de agrupamento são utilizadas para reduzir as dimensões dos mapas de recursos e dos parâmetros do modelo. As camadas totalmente conectadas, normalmente, estão no final de cada arquitetura de *CNN* e funcionam como um classificador que atribui um rótulo pré-definido à uma entrada (CHAI *et al.*, 2021). A Figura 5 apresenta um exemplo de arquitetura de *CNN* para classificação de imagens.

Figura 5 – Arquitetura de *CNN* para classificação de imagens.



Fonte: (CHAI *et al.*, 2021).

Uma vantagem do uso das *CNNs* é que ela aproveita o uso da coerência espacial local nas imagens de entrada, o que permite que elas tenham menos pesos, pois alguns parâmetros são compartilhados (TAMMINA, 2019). No entanto, uma das desvantagens que podem surgir com o treinamento das *CNNs* está relacionada com o grande número de parâmetros que precisam ser aprendidos, o que pode levar ao problema de *overfitting* (VOULODIMOS *et al.*, 2018).

Em resumo, os principais modelos de *DL* para *CV* possuem vantagens e desvantagens, porém, as *CNNs* estão cada vez mais se destacando tanto na academia como na indústria. Desde o excelente desempenho na competição de classificação de imagem *ImageNet*, as *CNNs* se tornaram uma das abordagens de *DL* mais notáveis. A classificação de imagens é um dos campos mais importantes e básicos para aplicações da *CV* (CHAI *et al.*, 2021).

2.4 REDES NEURAIS CONVOLUCIONAIS

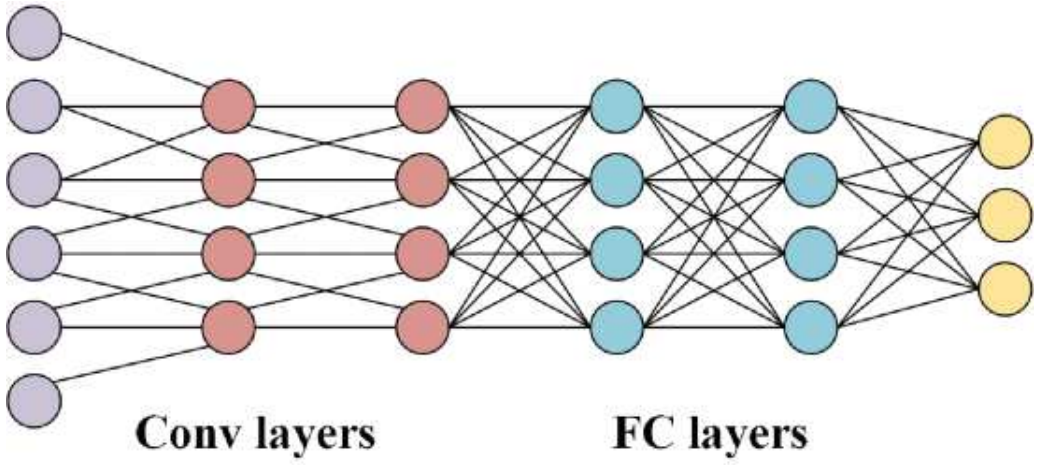
Em 1959, Hubel e Wiesel descobriram que as células do córtex visual animal são responsáveis pela detecção de luz em campos receptivos. Inspirado por essa descoberta, Kunihiko Fukushima propôs o *Neocognitron* em 1980, que pode ser considerado o antecessor da Rede Neural Convolucional, em inglês, *Convolutional Neural Network* (*CNN*) (GU *et al.*, 2018). Posteriormente, o modelo apresentado por Fukushima seria a inspiração utilizada para o desenvolvimento das primeiras *CNNs*.

Anos depois, LeCuN *et al.* propuseram a primeira *CNN* multicamadas, chamada *ConvNet*. A *ConvNet* possuía um treinamento supervisionado usando o algoritmo de retropropagação, diferentemente de sua antecessora *Neocognitron*, que utilizava um esquema de aprendizado por reforço não supervisionado (LINDSAY, 2021). Em 1998, LeCuN propôs uma versão aprimorada do *ConvNet*, conhecida como *LeNet-5*, e iniciou o uso da *CNN* na classificação de caracteres em aplicativos relacionados ao reconhecimento de documentos (KHAN *et al.*, 2020). A partir desse momento, criou-se as bases para os modelos de *CNNs* atuais.

As *CNNs* são modelos de *DL* inspirados no mecanismo de percepção visual natural dos seres vivos (GU *et al.*, 2018). Os modelos de *CNNs* são aplicados a diversos problemas relacionados à *CV*, dentre eles, a classificação de imagens é uma das tarefas mais importantes e disseminadas. Além do mais, modelos de classificação de imagens frequentemente são a base para modelos de detecção de objetos, localização de objetos e segmentação semântica (WANG; YANG, 2019).

Existem inúmeras variantes de arquiteturas *CNN* na literatura. No entanto, seus componentes básicos são muito semelhantes (GU *et al.*, 2018). Nesse sentido, uma arquitetura de *CNN* típica compreende camadas alternadas de convolução e agrupamento seguidas por uma ou mais camadas totalmente conectadas no final (KHAN *et al.*, 2020). A Figura 6 apresenta um diagrama das camadas de uma *CNN*.

Figura 6 – Diagrama das camadas de uma *CNN*.



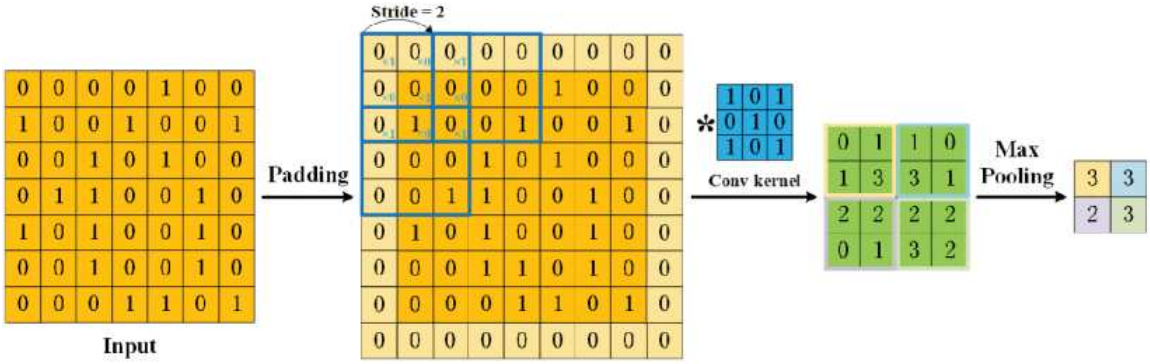
Fonte: (LI *et al.*, 2022).

A camada convolucional é composta por um conjunto de *kernels* convolucionais, onde cada neurônio atua como um *kernel*. Devido à capacidade de compartilhamento de peso da operação convolucional, diferentes conjuntos de recursos dentro de uma imagem podem ser extraídos pelo *kernel* (LI *et al.*, 2022). Geralmente, uma *CNN* possui algumas camadas convolucionais, assim, cada camada consegue extrair recursos deferentes da entrada. As primeiras camadas extraem recursos de nível inferior, como arestas, pontos finais e cantos. As camadas de nível superior extraem recursos mais complexos processando os recursos de nível inferior (WANG; YANG, 2019). As camadas convolucionais possuem dois hiperparâmetros principais, o preenchimento, em inglês *Padding*, e o passo, em inglês *Stride*. O preenchimento pode ser utilizado para aumentar a entrada adicionando zeros. Já o passo, controla a distância ao mover o *kernel* sobre uma imagem.

Após o processamento da camada convolucional, temos como saída um mapa de recursos, que é passado como entrada para uma camada de agrupamento. A camada de agrupamento, em inglês *Pooling*, tem como função reduzir gradativamente o tamanho espacial

dos dados, minimizando o número de parâmetros da rede e proporcionando a diminuição do consumo de recursos computacionais. Além disso, a camada de agrupamento também pode aprender alguns recursos invariáveis da entrada (QIN *et al.*, 2018; WANG; YANG, 2019). O agrupamento máximo é o tipo de agrupamento mais aplicado, ele recupera os maiores valores em um mapa de recursos. A Figura 7 demonstra o procedimento de uma *CNN*.

Figura 7 – Procedimento de uma *CNN*.



Fonte: (LI *et al.*, 2022).

No final de uma *CNN*, a saída da última camada de agrupamento atua como entrada para a camada totalmente conectada, em inglês *Fully Connected*. Totalmente conectada significa que todos os neurônios de uma camada estão conectados a todos os neurônios da próxima camada (GU *et al.*, 2018; TAMMINA, 2019). A camada totalmente conectada recebe informações da etapa de extração de recursos, analisa globalmente a saída de todas as camadas anteriores e produz uma classificação (KHAN *et al.*, 2020). Na Figura 8 podemos observar a arquitetura simplificada de uma *CNN*.

Figura 8 – Arquitetura de uma *CNN*.

Fonte: (LASRI; SOLH; BELKACEMI, 2019).

De maneira geral, as *CNNs* começam envolvendo um conjunto de filtros com uma imagem de entrada, gerando, com isso, mapas de recursos semelhantes. Logo depois, o agrupamento é aplicado, criando respostas mais complexas e aprendendo recursos invariáveis da imagem de entrada. Após várias iterações desse padrão, camadas totalmente conectadas são adicionadas e a última camada contém tantos neurônios quanto o número de categorias para classificação, por fim, gera-se um rótulo de categoria para a imagem (LINDSAY, 2021).

A *CNN* é uma das redes mais importantes no campo de aprendizado profundo. Recentemente, a *CNN* fez conquistas impressionantes, principalmente, no campo da *CV* e,

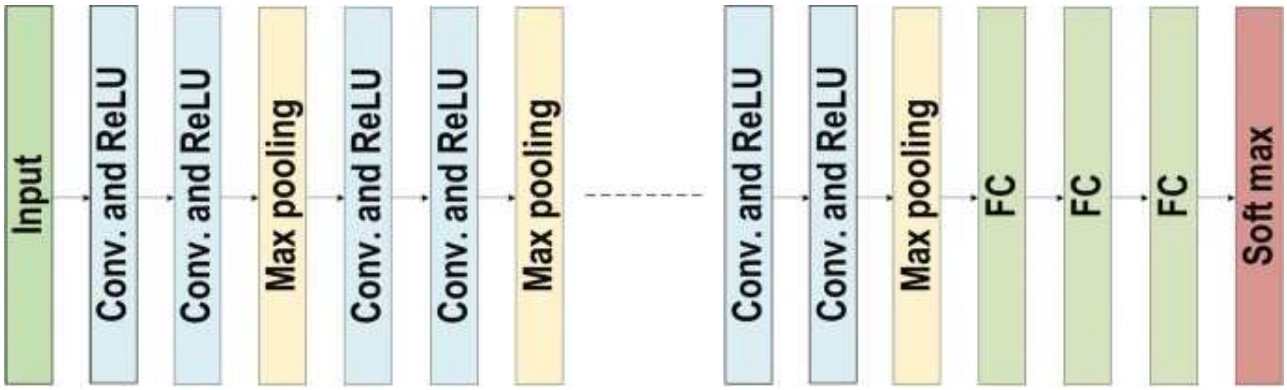
diante disso, atraiu muita atenção da indústria e da academia nos últimos anos. A visão computacional baseada em *CNN* permitiu que fosse possível a realização de tarefas que antes eram consideradas impossíveis de serem realizadas por máquinas nos últimos séculos, como reconhecimento facial, veículos autônomos e tratamentos médicos inteligentes (LI *et al.*, 2022).

2.4.1 MODELOS DE REDES NEURAIS CONVOLUCIONAIS

2.4.1.1 *VISUAL GEOMETRY GROUP* (*VGG*)

O modelo *VGG* é uma *CNN* que utiliza uma estratégia de aplicação de filtros convolucionais muito pequenos, de 3x3. Essa estratégia permitiu aumentar a profundidade do modelo entre 16 e 19 camadas de pesos e, além disso, melhorou o desempenho da *VGG* em comparação com modelos anteriores de *CNN* (WANG; YANG, 2019). Na Figura 9 podemos observar a arquitetura do modelo *VGG*.

Figura 9 – Arquitetura do modelo *VGG*.



Fonte: (ALZUBAIDI *et al.*, 2021).

2.4.1.2 *RESIDUAL NETWORK* (*RESNET*)

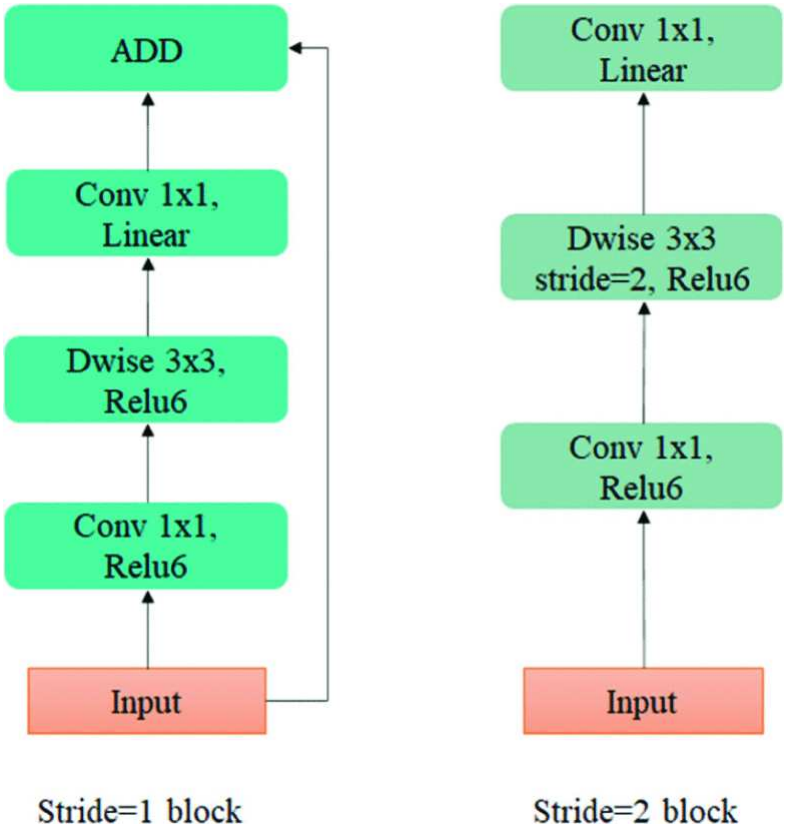
Os modelos de *CNN* do tipo *ResNet* foram desenvolvidos visando solucionar o problema do gradiente de fuga em modelos de *CNN* muito profundos. Diante disso, vários tipos de *ResNet* foram desenvolvidos com diferentes números de camadas, começando com 34 camadas e indo até 1202 camadas. O modelo mais comum de *ResNet* é o *ResNet50*, que compreende 49 camadas convolucionais mais uma única camada totalmente conectada (ALZUBAIDI *et al.*, 2021; WANG; YANG, 2019).

O modelo *ResNet* é construído, principalmente, por blocos de aprendizado residual, o que permite a conectividade entre camadas. Os modelos *ResNet* tem o potencial de evitar problemas de diminuição de gradiente, pois os blocos de aprendizagem residual aceleram a convergência profunda da rede. A *ResNet* foi a rede vencedora do campeonato 2015-ILSVRC com 152 camadas de profundidade, isso representa 8 vezes a profundidade do modelo *VGG* (ALZUBAIDI *et al.*, 2021).

2.4.1.3 *MOBILENETV2*

O modelo *MobileNetV2* aplica a técnica de Convoluções Separáveis em Profundidade, em inglês *Depthwise Separable Convolutions* (*DSC*), para introduzir portabilidade à rede. As *DSCs* mitigaram o problema de destruição de informações em camadas não lineares, em blocos de convolução, usando gargalos lineares, mas também estabeleceram uma nova estrutura chamada Resíduos invertidos, que conseguem preservar as informações no modelo (DONG *et al.*, 2020). A Figura 10 apresenta a estrutura do bloco convolucional do modelo *MobileNetV2*.

Figura 10 – Bloco convolucional do modelo *MobileNetV2*.



Fonte: (DONG *et al.*, 2020).

2.4.1.4 *REGNET*

A *RegNet* é um modelo que combina as vantagens do design manual e da pesquisa de arquitetura de rede, em inglês *Network Architecture Search* (*NAS*). O espaço de design *RegNet* pode funcionar perfeitamente em uma ampla gama de regimes de operações de ponto flutuante por segundo, pois é uma rede simples e rápida. A depender das configurações de treinamento, o modelo *RegNet* supera o popular modelo *EfficientNet*, em até 5 vezes na *GPU* (CHAI *et al.*, 2021). Neste trabalho, utilizamos o modelo *RegNetX002*.

2.4.2 MÉTRICAS DE AVALIAÇÃO

Há várias formas de avaliar modelos de *CNN* durante a etapa de treinamento e teste. No entanto, as métricas de desempenho mais comuns na área de *DL*, para a tarefa de classificação, são as seguintes: acurácia, precisão, revocação, pontuação-F1. Para as definições de métricas, usaremos as seguintes abreviações: o número de verdadeiros positivos, em inglês *True Positive* (*TP*), o número de falsos positivos, em inglês *False Positive* (*FP*), o número de verdadeiros negativos, em inglês *True Negative* (*TN*), e o número de falsos negativos, em inglês *False Negative* (*FN*) (KOROTCOV *et al.*, 2017). Podemos observar a definição de cada uma das métricas nas equações a seguir.

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precisão = \frac{TP}{TP + FP} \tag{2}$$

$$Revocação = \frac{TP}{TP + FN} \tag{3}$$

$$\textit{Pontuação-F1} = \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{4}$$

A acurácia (Equação 1) é a proporção de classes preditas corretas para o número total de amostras avaliadas. De outro modo, a precisão (Equação 2) é utilizada para calcular os padrões positivos que são corretamente previstos por todos os padrões previstos em uma classe positiva. A Revocação (Equação 3) calcula a fração de padrões positivos que são classificados corretamente. Por fim, a pontuação-F1 (Equação 4) calcula a média harmônica entre as taxas de revocação e precisão (ALZUBAIDI *et al.*, 2021). Outra forma para observar o desempenho dos modelos de classificação é a matriz de confusão, apresentada na Tabela 1.

Tabela 1 – Exemplo de matriz de confusão.

	Classe Predita	
Classe Real	*TP*	*FN*
	FP	*TN*

A matriz de confusão é amplamente utilizada em modelos de aprendizagem de máquina para classificação supervisionada, ajudando na compreensão do comportamento dos modelos de classificação. A estrutura quadrada de uma matriz de confusão é representada por linhas e colunas, onde as linhas são as classes reais das instâncias e as colunas são as classes preditas (HASNAIN *et al.*, 2020).

2.5 TRANSFERÊNCIA DE APRENDIZAGEM

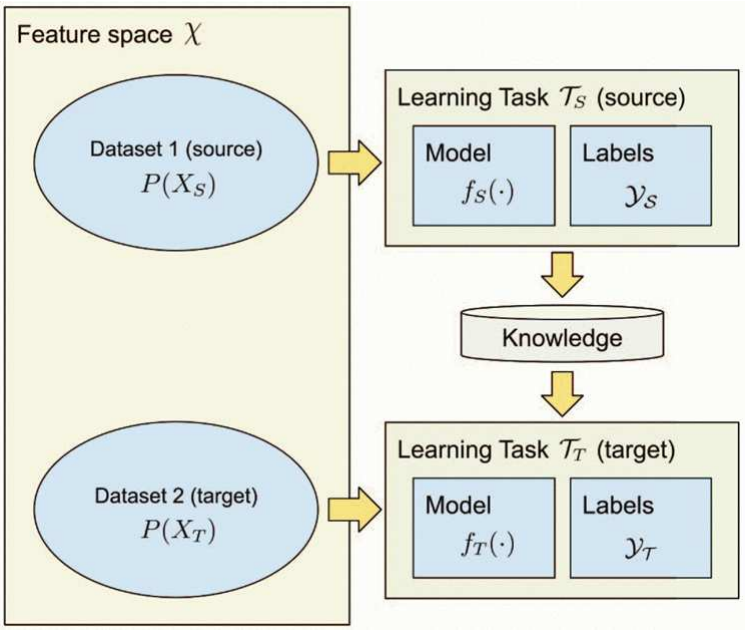
Segundo a teoria da generalização da transferência, proposta pelo psicólogo CH Judd, aprender a transferir é o resultado da generalização da experiência. É possível realizar a transferência de uma situação para outra, desde que a pessoa generalize sua experiência. Inspirado nas capacidades dos seres humanos para transferir conhecimento entre domínios, a transferência de aprendizagem, em inglês *Transfer Learning* (*TL*), visa alavancar o conhecimento de um domínio de origem para melhorar o desempenho do aprendizado em um domínio de destino (ZHUANG *et al.*, 2021).

A *TL* é um tópico emergente que pode impulsionar o sucesso da aprendizagem de máquina na pesquisa e na indústria. A falta de dados em tarefas específicas é um dos principais motivos para a aplicação da *TL*, tendo em vista que coletar e rotular dados pode ser muito caro e demorado e, além disso, recentes preocupações com a privacidade dificultam o uso de dados reais dos usuários (RIBANI; MARENGONI, 2019).

A aprendizagem por transferência é um método usado para transferir o conhecimento adquirido de uma tarefa para resolver outra. Este procedimento pode ajudar a melhorar a precisão ou reduzir o tempo de treinamento. No entanto, aplicar a *TL* de forma errada trará efeitos indesejados, podendo, inclusive, penalizar a precisão de um modelo, este efeito negativo da *TL* é conhecido como transferência negativa (RIBANI; MARENGONI, 2019). Diante disso, é evidente que a *TL* é um campo que pode ser explorado em diversas aplicações onde buscam-se melhores resultados de forma eficiente.

Formalmente, temos que um domínio pode ser representado como D = {X, P(X)}, onde X denota um espaço de características e P(X) denota uma distribuição marginal para $X = [x^1, x^2, ..., x^n] \in \mathbb{R}^{m \times n}$. Para um domínio específico D = {X , P(X)}, uma tarefa pode ser representada formalmente como T = {Y, f(·)}, onde Y denota um espaço de rótulos e f(·) denota uma função de decisão. Portanto, a *TL* pode ser definida da seguinte forma: dado um domínio de origem D_S e tarefa de aprendizado T_S e um domínio de destino D_T e tarefa de aprendizado T_T, a *TL* visa ajudar a melhorar o aprendizado da função de decisão f(·) em D_T (YU; XIU; LI, 2022). A Figura 11 ilustra a transferência de aprendizagem.

Figura 11 – Representação da *TL*.

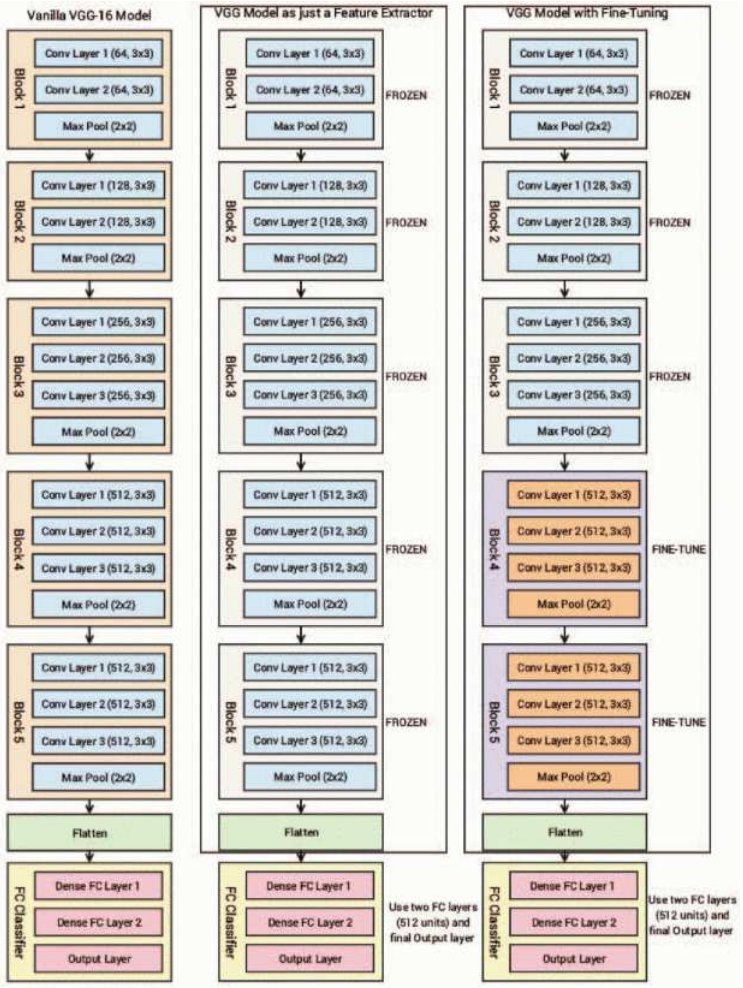


Fonte: (RIBANI; MARENGONI, 2019).

Em *DL*, geralmente em tarefas de classificação, a *TL* pode ser aplicada por meio de duas abordagens principais. A primeira, conhecida como classificação de prateleira, simplesmente recupera características aprendidas pelos modelos de *DL* em outros conjuntos de dados e as utiliza como entrada para um novo classificador. Neste cenário, os pesos dos modelos são congelados, ou seja, não mudam durante a fase de treinamento, e o classificador final da camada superior é o único componente do modelo que é realmente treinado (SABATELLI *et al.*, 2019).

Em contrapartida, na segunda abordagem para aplicação de *TL* em modelos de classificação de *DL*, conhecida como ajuste fino, os pesos de uma ou mais camadas dos modelos de *DL* originais são descongelados, ou seja, são ajustados durante a fase de treinamento, e as arquiteturas neurais são treinadas junto com o classificador final (SABATELLI *et al.*, 2019). Na prática, ao invés de inicializar aleatoriamente os parâmetros de um modelo de *DL*, tais modelos utilizam parâmetros que foram aprendidos em algum conjunto de dados de grande escala, sendo frequentemente utilizado o conjunto de dados *ImageNet*, pois contém 1.4 milhão de imagens com 1.000 classes distintas (KENSERT; HARRISON; SPJUTH, 2019; RIBANI; MARENGONI, 2019). Na Figura 12, temos como exemplo o modelo de *CNN VGG-16* ilustrando a *TL* com ajuste fino e congelamento de pesos.

Figura 12 – *VGG-16* com ajuste fino e congelamento de pesos.



Fonte: (RIBANI; MARENGONI, 2019).

Para decidir entre ajuste fino ou congelamento, é importante considerar a distribuição dos dados. Nesse sentido, caso o domínio de origem e o domínio de destino tenham as mesmas classes, mas a tarefa a ser resolvida seja diferente, provavelmente, o congelamento de camadas trará melhores resultados devido às representações de características comuns contidas nos domínios de origem e destino. No entanto, se o domínio de origem e de destino forem diferentes, provavelmente será necessário descongelar e ajustar algumas ou todas as camadas do modelo para fazer com que o modelo se adapte à nova distribuição contida no domínio de destino (RIBANI; MARENGONI, 2019).

A *TL*, em conjunto com modelos de *DL*, tem alcançado um excelente desempenho em visão computacional, classificação de texto, reconhecimento de comportamento e processamento de linguagem natural. A *TL* aliada à *DL* aplica o aprendizado de ponta a ponta e supera a desvantagem da *ML* tradicional, que considera cada conjunto de dados individualmente. Embora existam algumas pesquisas sobre *TL* em modelos de *ML*, falta atenção especial aos avanços recentes em *TL* para modelos de *DL* (YU; XIU; LI, 2022).

3 TRABALHOS RELACIONADOS

Este capítulo apresenta trabalhos relacionados ao tema central desta pesquisa. Para isso, buscamos, nas plataformas científicas Google Scholar e Science Direct, alguns trabalhos que tratem da aplicação de *FER* no contexto educacional utilizando *CNNs*, filtrando os resultados da busca para trabalhos publicados nos últimos quatro anos, e descrevemos os principais aspectos de cada trabalho a seguir.

O trabalho de (DUKIĆ; Sovic Krzic, 2022) realizou um experimento com *FER*, em sala de aula, que consistiu em 40 participantes que foram solicitados a resolver 8 tarefas de programação visual usando um robô educacional. No experimento *FER*, os autores utilizaram as *CNNs Inception-v3* e *ResNet-34* para prever as emoções em tempo real utilizando vídeos dos participantes de um workshop envolvendo um robô educacional.

Em (TONGUÇ; Ozaydın Ozkara, 2020) foram examinadas mudanças nas emoções de 67 alunos durante as aulas de Tecnologias de Informação Básica em uma universidade. As expressões faciais dos alunos foram analisadas quanto aos sentimentos de nojo, tristeza, alegria, medo, desprezo, raiva e surpresa, utilizando o sistema de codificação de movimentos faciais, em inglês *Facial Action Coding System* (*FACS*), por meio de um *software* desenvolvido com a ajuda da *API Microsoft Emotion Recognition* e linguagem de programação C#.

No trabalho de (LASRI; SOLH; BELKACEMI, 2019) foi desenvolvido um sistema que reconhece as emoções dos alunos em seus rostos. O sistema construído possui três etapas: detecção de face usando *Haar Cascades*, normalização e reconhecimento de emoção usando *CNN* no banco de dados *FER* 2013 com sete tipos de expressões faciais.

Em (PABBA; KUMAR, 2022) foi proposto um sistema em tempo real para monitorar o envolvimento de um grupo de alunos, analisando suas expressões faciais e reconhecendo estados afetivos acadêmicos: tédio, confusão, foco, frustração, bocejo e sono. Para isso, os autores realizaram etapas de pré-processamento, reconhecimento de expressão facial baseado em *CNN* e etapas de pós-processamento. Neste trabalho, os autores relataram precisão de treinamento de 78,70% e de teste de 76,90%.

Na abordagem apresentada em (SHEN *et al.*, 2022), foi proposta uma estrutura de avaliação de engajamento de aprendizagem que introduz o reconhecimento de expressões faciais para obter as mudanças emocionais dos alunos. Além do mais, um novo método de reconhecimento de expressões faciais é proposto com base na adaptação de domínio. Como resultados para *FER*, os autores obtiveram 51% de acurácia no conjunto de dados JAFFE e 54% de acurácia no conjunto de dados ck+.

No trabalho de (ZHANG *et al.*, 2020) foi proposto um novo algoritmo de detecção do engajamento de aprendizagem, com base em dados de comportamento dos alunos, obtidos a partir de câmeras e do mouse no ambiente de aprendizagem online. Como resultado dos experimentos realizados, os autores apresentaram 94,60% de taxa de reconhecimento como o maior resultado obtido durante os experimentos realizados.

(Zatarain Cabada *et al.*, 2020) propuseram uma metodologia utilizando algoritmos genéticos para otimização de hiperparâmetros de uma *CNN* utilizada para identificar o estado afetivo de uma pessoa. A *CNN* otimizada foi embutida em um sistema de tutoria inteligente executado em um celular. O processo de treinamento da *CNN* foi realizado em um computador com *GPU* e a rede neural treinada foi embarcada em um ambiente móvel. Os autores relataram uma melhoria de 8% com algoritmos genéticos em comparação com o método de tentativa e erro.

Enfatizamos que os trabalhos relacionados que foram descritos neste capítulo não tiveram como foco a investigação da *TL* em modelos de *CNN* para *FER*. Diante disso, o presente trabalho busca atuar nessa lacuna de pesquisa.

4 METODOLOGIA

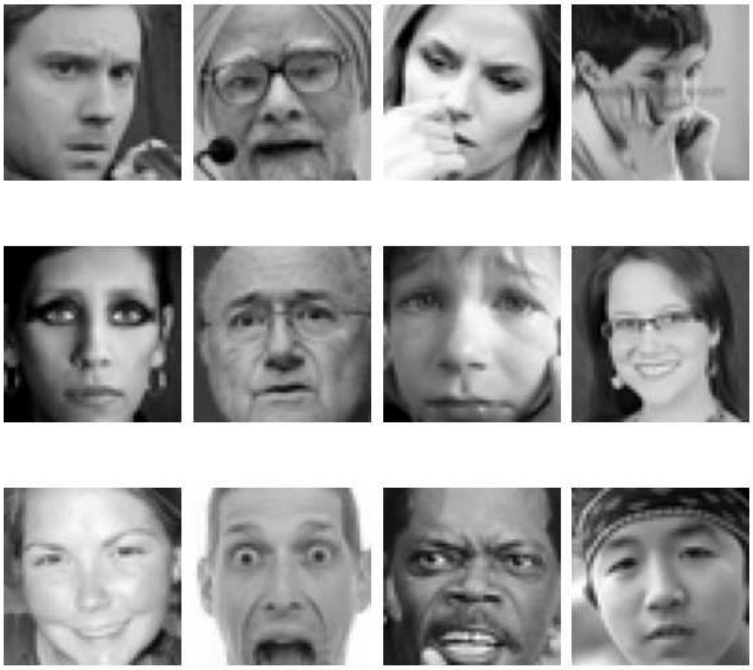
Para a aplicação de *DL* e *TL*, visando realização do *FER* em imagens de alunos, é necessário a execução de algumas etapas, a saber: pré-processamento, extração de recursos de imagem e classificação. As subseções a seguir, descrevem o conjunto de dados utilizado e a realização de cada etapa.

A subseção 4.1 descreve o conjunto de dados utilizado. Na subseção 4.2, apresentamos os recursos computacionais que foram utilizados para realização dos experimentos. O pré-processamento é descrito na subseção 4.3. A extração de recursos e a classificação de imagens são tratados na subseção 4.4.

4.1 CONJUNTO DE DADOS

O conjunto de dados utilizado no presente trabalho é o *FER*2013, que contém imagens em escala de cinza, no formato 48x48, das sete emoções básicas raiva, nojo, medo, alegria, tristeza, surpresa e neutra, totalizando 35887 imagens (GOODFELLOW *et al.*, 2013). O conjunto *FER*2013 foi escolhido para realização dos experimentos, pois é um banco público de imagens e de fácil acesso. A Figura 13 demonstra algumas imagens do conjunto *FER*2013.

Figura 13 – Exemplos de imagens do conjunto *FER*2013.



Fonte: O autor.

4.2 RECURSOS COMPUTACIONAIS

Para realização dos experimentos deste trabalho, utilizamos o Google Colab, uma ferramenta para aprendizagem de máquina, na qual podemos programar em linguagem de programação Python para execução em nuvem. Além disso, a ferramenta possui suporte para utilização de *GPUs*, essencial para algoritmos de *DL* que exigem uma grande capacidade dos recursos computacionais.

4.3 PRÉ-PROCESSAMENTO

4.3.1 AUMENTAÇÃO DE DADOS

A aumentação de dados consiste em aumentar artificialmente o tamanho de um conjunto de dados de treinamento, utilizando, para isso, transformações nos dados e preservando os respectivos rótulos de classe. Essa técnica parte da suposição de que mais informações podem ser extraídas do conjunto de dados original por meio de dados gerados a partir das transformações resultantes da aumentação de dados (SHORTEN; KHOSHGOFTAAR, 2019).

Para realização dos experimentos deste trabalho e visando obter melhores resultados, aplicamos a técnica de aumentação de dados no conjunto de dados *FER*2013 com o auxílio da biblioteca Keras. A Figura 14 demonstra alguns exemplos do resultado obtido ao aplicar a técnica de aumentação no conjunto *FER*2013.

Figura 14 – Exemplo de aplicação da técnica de aumentação de dados.

Fonte: O autor.

Ao final da etapa de aumentação de dados, aplicamos a técnica de normalização, que permite alteração na faixa de valores de pixels das imagens transformando os valores de 0 até

255 para valores entre 0 e 1. Isso é importante, pois permite uma melhor convergência aos modelos de *CNN*.

4.4 EXTRAÇÃO DE RECURSOS E CLASSIFICAÇÃO DE IMAGENS

A extração de recursos e a classificação são duas etapas realizadas em conjunto, pois os modelos de *CNN* são projetados para realizar tanto a extração de recursos quanto a classificação de imagens de forma simultânea. No presente trabalho utilizamos as *CNNs VGG16*, *ResNet50*, *MobileNetV2* e *RegNetX002*. Para todos os modelos, substituímos as camadas totalmente conectadas para permitir a classificação em apenas sete classes. Aplicamos em cada modelos a técnica de *TL* com ajuste fino, permitindo que os modelos refinem seu conhecimento prévio. Na subseção 4.4.1 vamos tratar do treinamento dos modelos de *CNN* e na subseção 4.4.2 vamos apresentar a etapa de teste.

4.4.1 TREINAMENTO DOS MODELOS DE *CNN*

O treinamento das *CNNs* foi realizado com 80% dos dados do conjunto *FER*2013, totalizando 29068 imagens. Paralelamente, a validação acontecia com 10% dos dados, totalizando 3230 imagens. Além disso, o treinamento dos modelos de *CNN* foi realizado com 100 épocas, utilizando uma técnica do Keras conhecida como parada antecipada, em inglês *EarlyStopping*. A parada antecipada recebe uma métrica, que será monitorada, e interrompe o treinamento de um modelo quando essa métrica parar de melhorar. A métrica escolhida para monitoramento da parada antecipada foi a perda de validação. Por fim, a *TL* foi realizada em todos os modelos de *CNN* utilizando ajuste fino e o pré-treinamento foi realizado no conjunto de dados *Imagenet*. Os hiperparâmetros utilizados no otimizador das *CNNs* durante o treinamento são apresentados na Tabela 2.

Tabela 2 – Hiperparâmetros do otimizador.

Hiperparâmetro	Valor
Learning rate	0.001
beta_1	0.9
beta_2	0.999
epsilon	1e-7

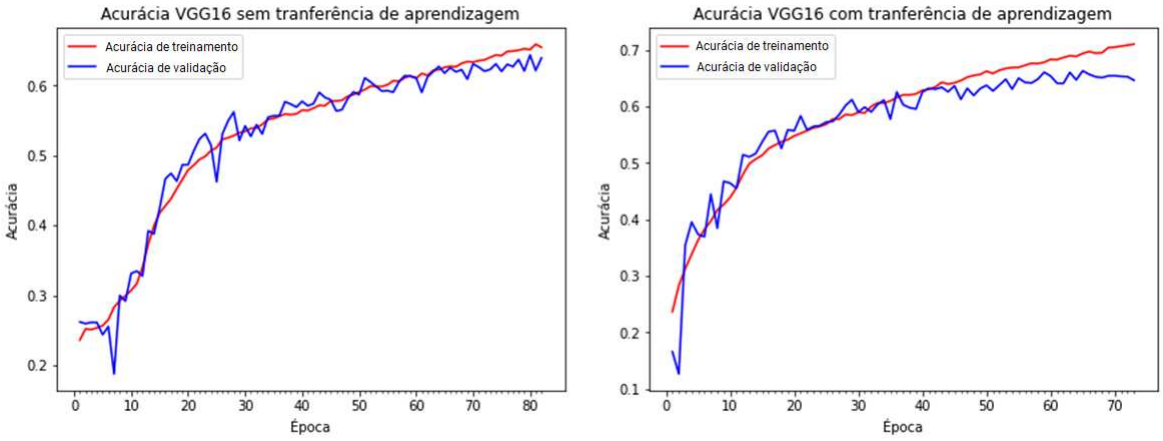
4.4.2 TESTE DOS MODELOS DE *CNN*

Na etapa de teste, 10% dos dados foram utilizados, o que resultou em 3589 imagens. Durante esta etapa, os modelos de *CNN*, anteriormente treinados e com os pesos salvos, são carregados e utilizados para inferência em dados de teste, ou seja, dados que nunca foram apresentados aos modelos. Ainda durante a etapa de teste, as métricas de desempenho são extraídas de cada *CNN*.

5 RESULTADOS E DISCUSSÃO

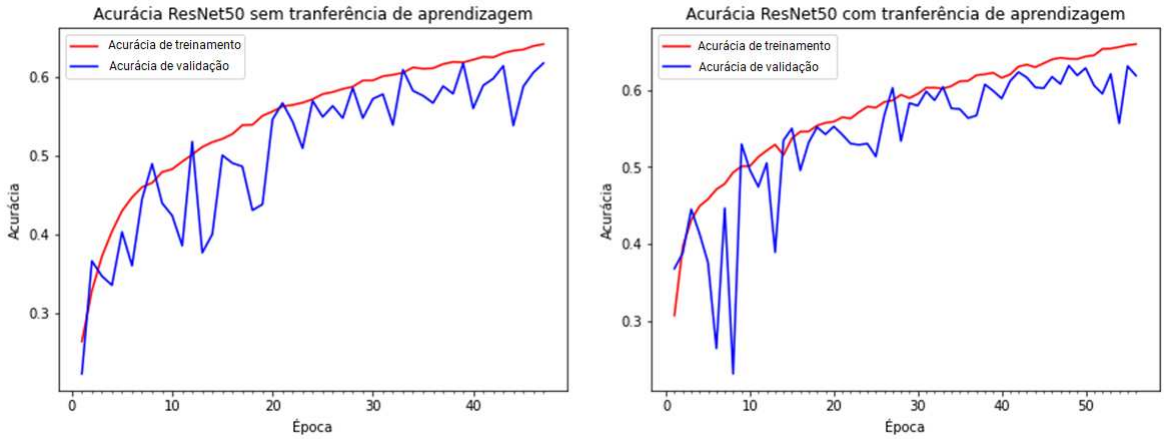
Ao final de cada época, durante a fase de treinamento, são geradas métricas parciais de desempenho dos modelos de *CNN*, tais métricas servem para observarmos a evolução da aprendizagem do modelo, verificando se os resultados melhoram após cada época. Nas Figuras 15, 16, 17 e 18 podemos observar a evolução da acurácia de cada *CNN* com *TL* e sem *TL* durante a fase de treinamento.

Figura 15 – Evolução do modelo *VGG16*.



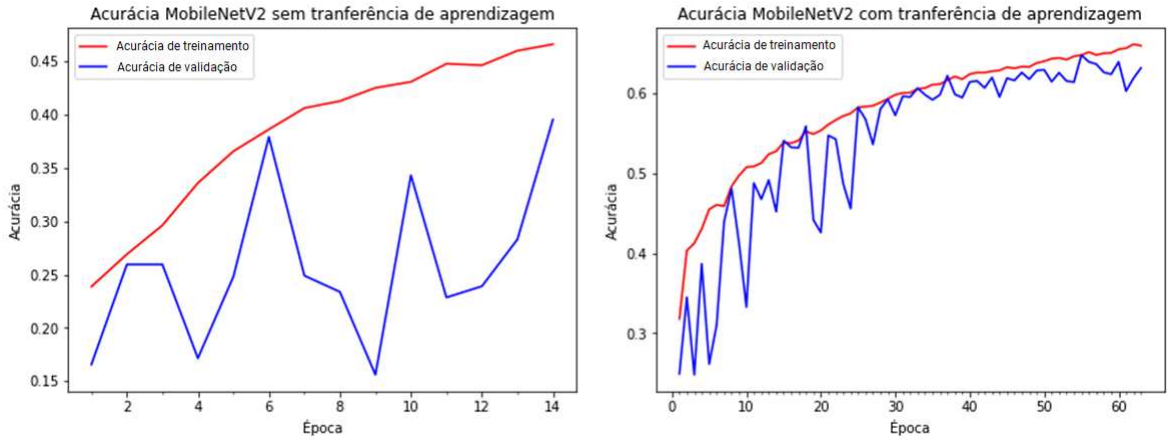
Fonte: O autor.

Figura 16 – Evolução do modelo *ResNet50*.



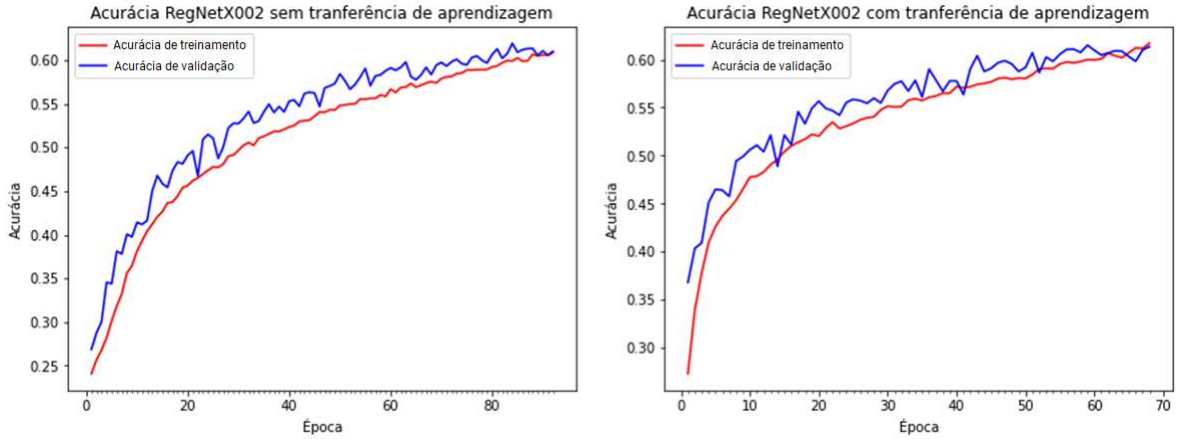
Fonte: O autor.

Figura 17 – Evolução do modelo *MobileNetV2*.



Fonte: O autor.

Figura 18 – Evolução do modelo *RegNetX002*.



Fonte: O autor.

Cada figura apresenta dois gráficos, à esquerda, acurácia por época sem *TL* e, à direita, acurácia por época com *TL*. Além disso, temos valores para o treinamento, em vermelho, e valores para validação, em azul. A partir disso, podemos observar que o modelo *VGG16* com *TL* apresentou o melhor desempenho durante a fase de treinamento. Em contrapartida, o modelo *MobileNetV2* sem *TL* obteve apenas 39% de acurácia de validação durante o treinamento, demonstrando que, esse modelo, quando utilizado sem *TL*, não é robusto o suficiente para dados de *FER*.

Durante a fase de teste, os modelos de *CNN* foram avaliados de acordo com as seguintes métricas de desempenho: acurácia, precisão, revocação e pontuação-F1. Os resultados das métricas de desempenho para cada modelo de *CNN* são apresentados na Tabela 3.

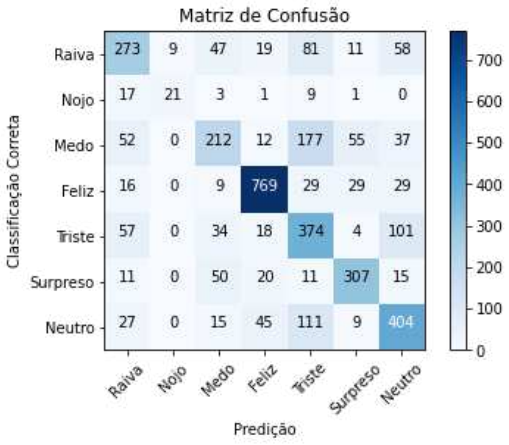
Tabela 3 – Desempenho dos modelos de *CNN*.

CNNs		Métricas			
		Acurácia	Precisão	Revocação	Pontuação-F1
VGG16	Sem *TL*	0.62	0.51	0.54	0.50
	Com *TL*	**0.65**	**0.65**	**0.60**	**0.62**
ResNet50	Sem *TL*	0.60	0.58	0.51	0.50
	Com *TL*	0.62	0.59	0.54	0.54
MobileNetV2	Sem *TL*	0.37	0.24	0.30	0.23
	Com *TL*	0.63	0.62	0.55	0.56
RegNetX002	Sem *TL*	0.61	0.61	0.54	0.55
	Com *TL*	0.60	0.60	0.54	0.55

Por meio da Tabela 3, podemos observar que o modelo *VGG16* com *TL* obteve o melhor resultado em todas as métricas de desempenho durante a fase de teste. Em contrapartida, é possível perceber também que o modelo *MobileNetV2*, sem *TL*, apesentou, assim como demonstrado na fase de treinamento (Figura 17), o pior desempenho em todas as métricas, porém, quando utilizamos o mesmo modelo aplicando a técnica de *TL*, ele obtém um desempenho competitivo com as demais *CNNs*. Isso aponta para uma alta sensibilidade do modelo *MobileNetV2* à técnica de *TL* em dados de *FER*. Além disso, de modo geral, todos os modelos de *CNN* para *FER* aplicados neste trabalho alcançaram melhores resultados quando aliados à técnica de *TL*, validando a hipótese de que a *TL* é benéfica para as *CNNs* quando aplicadas ao problema de *FER*.

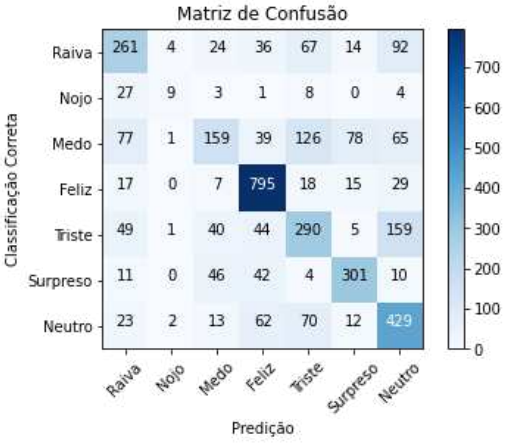
Outra forma de avaliar o comportamento dos modelos na fase de teste é a matriz de confusão, que permite verificar o número absoluto de acertos em cada classe. Como observado na Tabela 3, os modelos com *TL* obtiverem um desempenho superior às suas versões sem *TL*. Diante disso, apresentamos a matriz de confusão de cada modelo de *CNN* com *TL* nas Figuras 19, 20, 21 e 22.

Figura 19 – Matriz de confusão do modelo *VGG16* com *TL*.



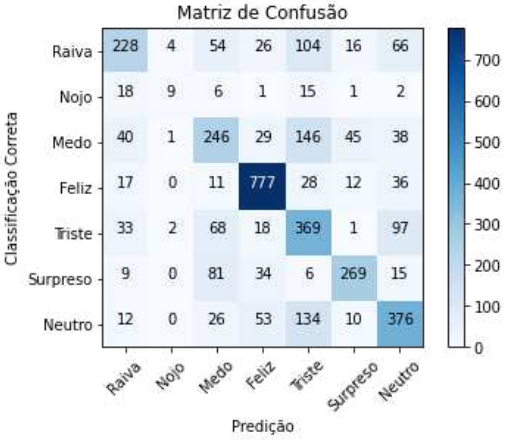
Fonte: O autor.

Figura 20 – Matriz de confusão do modelo *ResNet50* com *TL*.



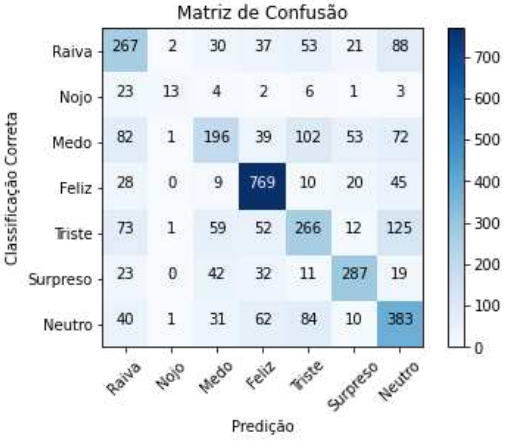
Fonte: O autor.

Figura 21 – Matriz de confusão do modelo *MobileNetV2* com *TL*.



Fonte: O autor.

Figura 22 – Matriz de confusão do modelo *RegNetX002* com *TL*.



Fonte: O autor.

A partir das figuras anteriores, observamos, por meio da matriz de confusão de cada modelo, que as *CNNs* capturaram informações relevantes das imagens e, a partir disso, conseguiram, de maneira geral, diferenciar as classes de imagens contidas no conjunto *FER*2013. Observa-se também que a classe Nojo apresentou maior dificuldade para classificação. Por outro lado, a classe Feliz foi mais facilmente classificada de forma correta.

A classe Nojo foi confundida, na maioria dos casos, com a classe Raiva. Vale ressaltar que isso pode ter relação com o nível de similaridade entre essas classes ou com o desbalanceamento do conjunto de dados, o que não permite que as *CNNs* extraiam todas as características importantes para classificação.

A classe Feliz mostrou-se como o rótulo que mais foi classificado de forma correta pelos modelos de *CNN*, apontando que as *CNNs* compreenderam as características mais importantes para distinguir esta classe. Pode-se atribuir o melhor desempenho das *CNNs*, nesta classe, a qualidade e quantidade de imagens contidas no conjunto *FER*2013.

6 CONCLUSÃO

As emoções são fundamentais em vários aspectos na vida dos seres humanos e na educação não é diferente. Diante disso, o presente trabalho buscou avaliar, por meio de métricas de desempenho, a aplicação da técnica de *TL* em modelos de *CNN* para *FER*, visando, posteriormente, sua utilização no contexto escolar para identificação das emoções dos alunos.

Diante disso, este trabalho comparou as *CNNs VGG16*, *ResNet50*, *MobileNetV2* e *RegNetX002*, com *TL* e sem *TL* para *FER*. Os resultados obtidos foram satisfatórios, com o modelo *VGG16*, na versão com *TL*, apresentando os melhores resultados em todas as métricas de desempenho utilizadas.

Foi visto que quando a técnica de *TL* foi aplicada aos modelos de *CNN*, seus respectivos desempenhos melhoraram significativamente, o que fornece evidências para validar a hipótese levantada neste trabalho, de que a *TL* seria benéfica para as *CNNs* com dados de *FER*. Portanto, concluímos que o desempenho das *CNNs* pode ser melhorado na tarefa de *FER* quando aplicamos a *TL* com ajuste fino.

REFERÊNCIAS

ALZUBAIDI, L.; ZHANG, J.; HUMAIDI, A. J.; AL-DUJAILI, A.; DUAN, Y.; AL-SHAMMA, O.; SANTAMARÍA, J.; FADHEL, M. A.; AL-AMIDIE, M.; FARHAN, L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. **Journal of Big Data**, v. 8, n. 1, p. 53, dec 2021. ISSN 2196-1115. Disponível em: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8>. Citado 2 vezes nas páginas 26 e 28.

BANK, D.; KOENIGSTEIN, N.; GIRYES, R. Autoencoders. **arXiv preprint arXiv:2003.05991**, 2020. Citado na página 22.

BAYLARI, A.; MONTAZER, G. Design a personalized e-learning system based on item response theory and artificial neural network approach. **Expert Systems with Applications**, v. 36, n. 4, p. 8013–8021, may 2009. ISSN 09574174. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S095741740800777X>. Citado na página 14.

BIRJALI, M.; KASRI, M.; BENI-HSSANE, A. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. **Knowledge-Based Systems**, v. 226, p. 107134, aug 2021. ISSN 09507051. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S095070512100397X>. Citado na página 17.

CHAI, J.; ZENG, H.; LI, A.; NGAI, E. W. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. **Machine Learning with Applications**, v. 6, p. 100134, dec 2021. ISSN 26668270. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2666827021000670>. Citado 4 vezes nas páginas 20, 22, 23 e 27.

DONG, K.; ZHOU, C.; RUAN, Y.; LI, Y. MobileNetV2 Model for Image Classification. In: **2020 2nd International Conference on Information Technology and Computer Application (ITCA)**. IEEE, 2020. p. 476–480. ISBN 978-1-6654-0378-8. Disponível em: <https://ieeexplore.ieee.org/document/9422058/>. Citado na página 27.

DUKIĆ, D.; Sovic Krzic, A. Real-Time Facial Expression Recognition Using Deep Learning with Application in the Active Classroom Environment. **Electronics**, v. 11, n. 8, p. 1240, apr 2022. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/11/8/1240>. Citado na página 32.

GOODFELLOW, I. J.; ERHAN, D.; CARRIER, P. L.; COURVILLE, A.; MIRZA, M.; HAMNER, B.; CUKIERSKI, W.; TANG, Y.; THALER, D.; LEE, D.-H.; ZHOU, Y.; RAMAIAH, C.; FENG, F.; LI, R.; WANG, X.; ATHANASAKIS, D.; SHAWE-TAYLOR, J.; MILAKOV, M.; PARK, J.; IONESCU, R.; POPESCU, M.; GROZEA, C.; BERGSTRA, J.; XIE, J.; ROMASZKO, L.; XU, B.; CHUANG, Z.; BENGIO, Y. Challenges in Representation Learning: A Report on Three Machine Learning Contests. In: . [s.n.], 2013. p. 117–124. Disponível em: <http://link.springer.com/10.1007/978-3-642-42051-1_16>. Citado na página 34.

GU, J.; WANG, Z.; KUEN, J.; MA, L.; SHAHROUDY, A.; SHUAI, B.; LIU, T.; WANG, X.; WANG, G.; CAI, J.; CHEN, T. Recent advances in convolutional neural networks. **Pattern Recognition**, v. 77, p. 354–377, may 2018. ISSN 00313203. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0031320317304120>. Citado 3 vezes nas páginas 23, 24 e 25.

HASNAIN, M.; PASHA, M. F.; GHANI, I.; IMRAN, M.; ALZAHRANI, M. Y.; BUDIARTO, R. Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking. **IEEE Access**, v. 8, p. 90847–90861, 2020. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/9091880/>. Citado na página 28.

HINTON, G. Deep belief networks. **Scholarpedia**, v. 4, n. 5, p. 5947, 2009. ISSN 1941-6016. Disponível em: <http://www.scholarpedia.org/article/Deep_belief_networks>. Citado na página 21.

HUANG, Y.; CHEN, F.; LV, S.; WANG, X. Facial Expression Recognition: A Survey. **Symmetry**, v. 11, n. 10, p. 1189, sep 2019. ISSN 2073-8994. Disponível em: <https://www.mdpi.com/2073-8994/11/10/1189>. Citado 3 vezes nas páginas 18, 19 e 20.

IMANI, M.; MONTAZER, G. A. A survey of emotion recognition methods with emphasis on E-Learning environments. **Journal of Network and Computer Applications**, v. 147, p. 102423, dec 2019. ISSN 10848045. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1084804519302759>. Citado 4 vezes nas páginas 14, 15, 18 e 19.

KENSERT, A.; HARRISON, P. J.; SPJUTH, O. Transfer Learning with Deep Convolutional Neural Networks for Classifying Cellular Morphological Changes. **SLAS Discovery**, v. 24, n. 4, p. 466–475, apr 2019. ISSN 24725552. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2472555222126304>. Citado na página 30.

KHAIREDDIN, Y.; CHEN, Z. Facial Emotion Recognition: State of the Art Performance on FER2013. may 2021. Disponível em: <http://arxiv.org/abs/2105.03588>. Citado na página 18.

KHAN, A.; SOHAIL, A.; ZAHOORA, U.; QURESHI, A. S. A survey of the recent architectures of deep convolutional neural networks. **Artificial Intelligence Review**, v. 53, n. 8, p. 5455–5516, dec 2020. ISSN 0269-2821. Disponível em: <https://link.springer.com/10.1007/s10462-020-09825-6>. Citado 4 vezes nas páginas 20, 23, 24 e 25.

KOROTCOV, A.; TKACHENKO, V.; RUSSO, D. P.; EKINS, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. **Molecular Pharmaceutics**, v. 14, n. 12, p. 4462–4475, dec 2017. ISSN 1543-8384. Disponível em: <https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.7b00578>. Citado na página 28.

KOUAHLA, M. N.; BOUGHIDA, A.; CHEBATA, I.; MEHENAOUI, Z.; LAFIFI, Y. Emorec: a new approach for detecting and improving the emotional state of learners in an e-learning environment. **Interactive Learning Environments**, p. 1–19, feb 2022. ISSN 1049-4820. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/10494820.2022.2029494>. Citado 2 vezes nas páginas 14 e 18.

LASRI, I.; SOLH, A. R.; BELKACEMI, M. E. Facial Emotion Recognition of Students using Convolutional Neural Network. In: **2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)**. IEEE, 2019. p. 1–6. ISBN 978-1-7281-0003-6. Disponível em: <https://ieeexplore.ieee.org/document/8942386/>. Citado 2 vezes nas páginas 25 e 32.

LI, B.; LIMA, D. Facial expression recognition via ResNet-50. **International Journal of Cognitive Computing in Engineering**, v. 2, p. 57–64, jun 2021. ISSN 26663074. Disponível

em: <https://linkinghub.elsevier.com/retrieve/pii/S2666307421000073>. Citado na página 14.

LI, S.; DENG, W. Deep Facial Expression Recognition: A Survey. **IEEE Transactions on Affective Computing**, v. 13, n. 3, p. 1195–1215, jul 2022. ISSN 1949-3045. Disponível em: <https://ieeexplore.ieee.org/document/9039580/>. Citado 3 vezes nas páginas 18, 19 e 20.

LI, Z.; LIU, F.; YANG, W.; PENG, S.; ZHOU, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. **IEEE Transactions on Neural Networks and Learning Systems**, v. 33, n. 12, p. 6999–7019, dec 2022. ISSN 2162-237X. Disponível em: <https://ieeexplore.ieee.org/document/9451544/>. Citado 3 vezes nas páginas 24, 25 e 26.

LINDSAY, G. W. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. **Journal of Cognitive Neuroscience**, v. 33, n. 10, p. 2017–2031, sep 2021. ISSN 0898-929X. Disponível em: <https://direct.mit.edu/jocn/article/33/10/2017/97402/Convolutional-Neural-Networks-as-a-Model-of-the>. Citado 2 vezes nas páginas 23 e 25.

MELCHIOR, J.; FISCHER, A.; WISKOTT, L. How to Center Deep Boltzmann Machines. **Journal of Machine Learning Research**, v. 17, n. 1, p. 3387–3447, 2016. Citado na página 21.

PABBA, C.; KUMAR, P. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. **Expert Systems**, v. 39, n. 1, jan 2022. ISSN 0266-4720. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/exsy.12839>. Citado na página 32.

PATRÍCIO, D. I.; RIEDER, R. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. **Computers and Electronics in Agriculture**, v. 153, p. 69–81, oct 2018. ISSN 01681699. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0168169918305829>. Citado na página 20.

PICARD, R. W. Affective computing. 1997. Citado na página 17.

PULLI, K.; BAKSHEEV, A.; KORNYAKOV, K.; ERUHIMOV, V. Real-time computer vision with OpenCV. **Communications of the ACM**, v. 55, n. 6, p. 61–69, jun 2012. ISSN 0001-0782. Disponível em: <https://dl.acm.org/doi/10.1145/2184319.2184337>. Citado na página 20.

QIN, Z.; YU, F.; LIU, C.; CHEN, X. How convolutional neural networks see the world — A survey of convolutional neural network visualization methods. **Mathematical Foundations of Computing**, v. 1, n. 2, p. 149–180, 2018. ISSN 2577-8838. Disponível em: <http://aimsciences.org//article/doi/10.3934/mfc.2018008>. Citado na página 25.

REVINA, I.; EMMANUEL, W. S. A Survey on Human Face Expression Recognition Techniques. **Journal of King Saud University - Computer and Information Sciences**, v. 33, n. 6, p. 619–628, jul 2021. ISSN 13191578. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1319157818303379>. Citado 2 vezes nas páginas 18 e 19.

RIBANI, R.; MARENGONI, M. A Survey of Transfer Learning for Convolutional Neural Networks. In: **2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)**. IEEE, 2019. p. 47–57. ISBN 978-1-7281-5270-7. Disponível em: <https://ieeexplore.ieee.org/document/8920338/>. Citado 4 vezes nas páginas 14, 29, 30 e 31.

SABATELLI, M.; KESTEMONT, M.; DAELEMANS, W.; GEURTS, P. Deep Transfer Learning for Art Classification Problems. In: . [s.n.], 2019. p. 631–646. Disponível em: <http://link.springer.com/10.1007/978-3-030-11012-3_48>. Citado na página 30.

SHARMA, T.; DIWAKAR, M.; ARYA, C. A systematic review on emotion recognition by using machine learning approaches. In: . [s.n.], 2022. p. 020045. Disponível em: <http://aip.scitation.org/doi/abs/10.1063/5.0113378>. Citado na página 17.

SHEN, J.; YANG, H.; LI, J.; CHENG, Z. Assessing learning engagement based on facial expression recognition in MOOC's scenario. **Multimedia Systems**, v. 28, n. 2, p. 469–478, apr 2022. ISSN 0942-4962. Disponível em: <https://link.springer.com/10.1007/s00530-021-00854-x>. Citado na página 32.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on Image Data Augmentation for Deep Learning. **Journal of Big Data**, v. 6, n. 1, p. 60, dec 2019. ISSN 2196-1115. Disponível em: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0>. Citado na página 35.

TAMMINA, S. Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. **International Journal of Scientific and Research Publications (IJSRP)**, v. 9, n. 10, p. p9420, oct 2019. ISSN 2250-3153. Disponível em: <http://www.ijsrp.org/research-paper-1019.php?rp=P949194>. Citado 2 vezes nas páginas 23 e 25.

TONGUÇ, G.; Ozaydın Ozkara, B. Automatic recognition of student emotions from facial expressions during a lecture. **Computers Education**, v. 148, p. 103797, apr 2020. ISSN 03601315. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0360131519303471>. Citado na página 32.

VOULODIMOS, A.; DOULAMIS, N.; DOULAMIS, A.; PROTOPAPADAKIS, E. Deep Learning for Computer Vision: A Brief Review. **Computational Intelligence and Neuroscience**, v. 2018, p. 1–13, 2018. ISSN 1687-5265. Disponível em: <https://www.hindawi.com/journals/cin/2018/7068349/>. Citado 4 vezes nas páginas 20, 21, 22 e 23.

WANG, W.; YANG, Y. Development of convolutional neural network and its application in image classification: a survey. **Optical Engineering**, v. 58, n. 04, p. 1, apr 2019. ISSN 0091-3286. Disponível em: <https://www.spiedigitallibrary.org/journals/optical-engineering/volume-58/issue-04/040901/Development-of-convolutional-neural-network-and-its-application-in-image/10.1117/1.OE.58.4.040901.full>. Citado 3 vezes nas páginas 24, 25 e 26.

WANG, Y.; SONG, W.; TAO, W.; LIOTTA, A.; YANG, D.; LI, X.; GAO, S.; SUN, Y.; GE, W.; ZHANG, W.; ZHANG, W. A systematic review on affective computing: emotion models, databases, and recent advances. **Information Fusion**, v. 83-84, p. 19–52, jul 2022. ISSN 15662535. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1566253522000367>. Citado 3 vezes nas páginas 14, 17 e 18.

WIJASENA, H. Z.; FERDIANA, R.; WIBIRAMA, S. A Survey of Emotion Recognition using Physiological Signal in Wearable Devices. In: **2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)**. IEEE, 2021. p. 1–6. ISBN 978-1-6654-2482-0. Disponível em: <https://ieeexplore.ieee.org/document/9466092/>. Citado na página 17.

YANG, B.; CAO, J.; NI, R.; ZHANG, Y. Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images. **IEEE Access**, v. 6, p. 4630–4640, 2018. ISSN 2169-3536. Disponível em: <http://ieeexplore.ieee.org/document/8214102/>. Citado na página 14.

YU, F.; XIU, X.; LI, Y. A Survey on Deep Transfer Learning and Beyond. **Mathematics**, v. 10, n. 19, p. 3619, oct 2022. ISSN 2227-7390. Disponível em: <https://www.mdpi.com/2227-7390/10/19/3619>. Citado 2 vezes nas páginas 29 e 31.

Zatarain Cabada, R.; Rodriguez Rangel, H.; Barron Estrada, M. L.; Cardenas Lopez, H. M. Hyperparameter optimization in CNN for learning-centered emotion recognition for intelligent tutoring systems. **Soft Computing**, v. 24, n. 10, p. 7593–7602, may 2020. ISSN 1432-7643. Disponível em: <http://link.springer.com/10.1007/s00500-019-04387-4>. Citado na página 33.

ZHANG, Z.; LI, Z.; LIU, H.; CAO, T.; LIU, S. Data-driven Online Learning Engagement Detection via Facial Expression and Mouse Behavior Recognition Technology. **Journal of Educational Computing Research**, v. 58, n. 1, p. 63–86, mar 2020. ISSN 0735-6331. Disponível em: <http://journals.sagepub.com/doi/10.1177/0735633119825575>. Citado na página 32.

ZHU, X.; CHEN, Z. Dual-modality spatiotemporal feature learning for spontaneous facial expression recognition in e-learning using hybrid deep neural network. **The Visual Computer**, v. 36, n. 4, p. 743–755, apr 2020. ISSN 0178-2789. Disponível em: <http://link.springer.com/10.1007/s00371-019-01660-3>. Citado 3 vezes nas páginas 14, 16 e 19.

ZHUANG, F.; QI, Z.; DUAN, K.; XI, D.; ZHU, Y.; ZHU, H.; XIONG, H.; HE, Q. A Comprehensive Survey on Transfer Learning. **Proceedings of the IEEE**, v. 109, n. 1, p. 43–76, jan 2021. ISSN 0018-9219. Disponível em: <https://ieeexplore.ieee.org/document/9134370/>. Citado na página 29.