UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE GRADUAÇÃO EM ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Luiza Amador Pozzobon

UM ESTUDO DA CLASSIFICAÇÃO DE EMOÇÕES EM FACES PARCIALMENTE OCLUSAS

Santa Maria, RS
2019

Luiza Amador Pozzobon

UM ESTUDO DA CLASSIFICAÇÃO DE EMOÇÕES EM FACES PARCIALMENTE OCLUSAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Controle e Automação, Área de Concentração em Ciência da Computação, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **em Engenharia de Controle e Automação**.

ORIENTADOR: Prof. Dr. Rodrigo da Silva Guerra

Santa Maria, RS
2019

End. Eletr.: luiza.pozzobon@gmail.com

Luiza Amador Pozzobon

UM ESTUDO DA CLASSIFICAÇÃO DE EMOÇÕES EM FACES PARCIALMENTE OCLUSAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Controle e Automação, Área de Concentração em Ciência da Computação, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **em Engenharia de Controle e Automação**.

Aprovado em 18 de dezembro de 2019:

_____

Rodrigo da Silva Guerra, Dr. (UFSM)
(Presidente/Orientador)

_____

Daniel Fernando Tello Gamarra, Dr. (UFSM)

_____

Daniel Welfer, Dr. (UFSM)

_____

Paulo Lilles Jorge Drews Junior, Dr. (FURG)

Santa Maria, RS
2019

AGRADECIMENTOS

Agradeço a meus pais, por sempre proverem a mim todas as condições para o estudo e para a vida. A meu pai, por me amparar, amar e respeitar.

À Mônica, minha eterna amiga e parceira de (praticamente todos!) trabalhos da graduação. Ao Gustavo e ao Vinícius pelo companheirismo e piadas nesses anos. Independente de onde estivermos no futuro, saibam que podem sempre contar comigo.

Aos xerosos por prontamente me acolherem ao grupo e aos rolês. À Carol e à Thais pela amizade, conselhos e risadas.

À Alice por ser a melhor amiga de todas as melhores amigas, além da mulher mais forte e batalhadora que já conheci. Você me inspira todos os dias a ser maior que as dificuldades que possamos encontrar nesse mundão. Ainda vamos ter nosso apzinho lotado de plantas e gatos.

À Carol pelo carinho, ternura e preocupação. Por zelar pelo meu cuidado e sempre estar presente. Por segurar as minhas pontas quando nem eu consigo.

A todas as outras amizades que construí nesses quase 7 anos e meio de UFSM. Embora por muitas vezes distante, saibam que meu sentimento por vocês é imenso. Chuck, pelos conselhos e conversas, pelo calor que tu traz contigo. Bianca, por estar presente em algumas das minhas memórias mais queridas. Thayse, pelas melhores e mais engraçadas conversas, até altas horas da madrugada, nos tempos da civil. Pedro e Vidor, por me presentearem com suas amizades e carinho. Érico e Jesus pela troca de experiências e apoio. David, meu guru dos hardware, pela tua capacidade de fazer qualquer um se sentir em casa. Valéria e Carol pela conexão e pela afeição tão fortes que criamos.

Ao professor Guerra por não deixar que os 19169,93km que separam Santa Maria de Taiwan impedissem nossos encontros.

A todos os professores que passaram a mim não só conhecimento, mas também humanidade.

À Universidade Federal de Santa Maria por fornecer um ensino público, gratuito e de qualidade. Que muitos outros ainda possam usufruir de algo tão grandioso e bonito.

To the engineers and scientists
who will one day build minds;
from whatever material,
in whatever form.

Hello from a time when
we thought it was all magic.

(Exurb1a - The Fifth Science)

RESUMO

UM ESTUDO DA CLASSIFICAÇÃO DE EMOÇÕES EM FACES PARCIALMENTE OCLUSAS

AUTORA: Luiza Amador Pozzobon
ORIENTADOR: Rodrigo da Silva Guerra

Com o aumento do uso de tecnologias ou máquinas inteligentes pelo usuário comum, a necessidade de criação de algoritmos e computadores humano-centrados urge atenção. Nesse sentido, a inferência do estado emocional do ser humano com o qual uma máquina interage é fundamental para o estabelecimento e regulação de conexões afetivas, possuindo aplicações em áreas que vão desde a educação até a de tecnologias assistivas. O uso de algoritmos inteligentes, entretanto, começa a ser regularizado e projetos como a *General Data Protection Regulation* implicam, para o usuário, o direito à explicação de inferências automatizadas. Este trabalho é enraizado nos dois tópicos citados: estimação de emoção a partir de imagens faciais e explicabilidade do algoritmo inteligente em questão. Para isso, foram comparadas arquiteturas de redes neurais com e sem mecanismos de atenção, treinadas separadamente em dois conjuntos de dados, com e sem a presença de oclusões faciais. Ao fim do estudo, constata-se que a adição de mecanismos de atenção contribui, na maioria dos casos, positivamente para a classificação de emoções. Ainda, o treino com presença de imagens com oclusões parece expandir o foco de uma rede neural em outras imagens de entrada. Para o conjunto CK+, o melhor modelo obteve 97,77% de acurácia quando treinado e validado em imagens sem tarjas e 90,01% quando os dados continham tarjas faciais. No teste de cruzamento de *datasets*, o melhor modelo de rede neural obteve 44,13% de acurácia para o conjunto JAFFE.

Palavras-chave: Emoção facial. Classificação. Explicabilidade. Humano-centrado. Oclusão facial.

ABSTRACT

A STUDY ON EMOTION CLASSIFICATION OF PARTIALLY OCCLUDED FACES

AUTHOR: Luiza Amador Pozzobon
ADVISOR: Rodrigo da Silva Guerra

With the increased use of intelligent technologies or machines by the average user, the need for the creation of human-centered computers and algorithms urges attention. In this sense, the inference of the emotional state of the human being with which a machine interacts with is fundamental for the establishment and regulation of affective connections, having applications in areas ranging from education to assistive technologies. The usage of intelligent algorithms, however, is starting to be regularized, and projects such as the General Data Protection Regulation guarantees for the user the right to an explanation of such automated inferences. This work is rooted in both topics mentioned: emotion estimation from facial images and the explicability of the intelligent algorithm in question. For this, we compared neural network architectures with and without attention mechanisms, trained separately in two datasets, with and without the presence of facial occlusions. At the end of the study, it is found that the addition of attention mechanisms contributes positively to emotion classification in most cases. Also, training with occluded images seems to expand the focus of a neural network on other input images. Regarding the CK+ dataset, the best model obtained an accuracy of 97.77% when trained and validated on images without occlusions and 90.01% when the data contained facial occlusions. In the cross-dataset evaluation, the best neural network model obtained 44.13% of accuracy for the JAFFE dataset.

Keywords: Facial Emotion. Classification. Explicability. Human-centered. Facial Occlusion.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	Artificial Neural Network, ou rede neural artificial
AU	Action Unit
$AvgPool$	Average Pooling
$CBAM$	Convolutional Block Attention Module
$CK+$	Conjunto de dados Cohn-Kanade Extentido
$Conv$	Convolução, ou camada convolucional
$FACS$	Facial Action Coding System
fc	Fully Connected, ou camada totalmente conectada
FN	Falso Negativo
FP	Falso Positivo
$GDPR$	General Data Protection Regulation
GPU	Graphics Processing Unit, ou Unidade de Processamento Gráfico
$Grad - CAM$	Gradient-weighted Class Activation Mapping
$LGPD$	Lei Geral de Proteção de Dados
$MaxPool$	Max Pooling
MLP	Multilayer Perceptron
$ReLU$	Rectified Linear Unit
$Resnet$	Residual Network ou rede residual
$Resnet50$	Rede residual de 50 camadas
SWM	Spatial Weights Mechanism
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

LISTA DE SÍMBOLOS

ρ	Função de Ativação
σ	Função Sigmoid
μ	Média
$\boldsymbol{\sigma}$	Desvio Padrão
I_c^u	Canal uniformizado de uma imagem
$\odot$	Multiplicação elemento a elemento
$f^{7\times7}$	Filtro convolucional de tamanho 7×7
$\mathbb{R}$	Conjunto dos números reais
$L^c_{Grad-CAM}$	Mapa de localização discriminativo da classe c por Grad-CAM
α_k^c	Coeficiente de importância do neurônio k para a classificação c
ϵ	*Learning rate* ou taxa de aprendizagem

SUMÁRIO

1 INTRODUÇÃO

A linguagem corporal é, por vezes, tão ou mais crucial para o desenvolvimento de relações pessoais do que outros tipos de comunicação. A habilidade de predizer a emoção ou dividir o estado emocional de outra pessoa é chamada de empatia emocional, aptidão que estimula comportamentos pró-sociais (GONZALEZ-LIENCRES; SHAMAY-TSOORY; BRÜNE, 2013) e que é altamente dependente da percepção emocional (MITCHELL; PHILLIPS, 2015). Nesse sentido, a expressividade emocional através da face é apontada por Ekman (1992) como um fator de peso para a regulação e criação de conexões afetivas. Desde a infância, emoções expressas pela face tem o poder de formar laços e também de acelerar ou reduzir agressões (EKMAN, 1992).

Tendo em vista que a percepção emocional e, mais pontualmente, as expressões faciais podem ser interpretadas como recompensas (MATTHEWS; WELLS, 1999) ou punições (BLAIR, 1995) que regulam o comportamento de outros indivíduos (NIEDENTHAL; RIC, 2017), é natural a transferência dessa aptidão para tecnologias humano-centradas. A habilidade de inferir o estado emocional do ser humano com o qual uma máquina interage é fundamental para aplicações humano-centradas em educação (LIN *et al.*, 2013) e assistência a crianças com autismo (BHARATHARAJ *et al.*, 2017), por exemplo.

Uma vez que tais tecnologias de predição se fazem presentes no cotidiano do usuário comum, é de responsabilidade ética dos inventores possibilitarem que esses sistemas sejam explicáveis, de fácil entendimento até mesmo para leigos. Essa necessidade é ampliada com a implementação de pacotes como o *General Data Protection Regulation (GDPR)*, em vigor na União Europeia desde 2018, por exemplo, que garante o direito, por parte do usuário, à requisição de explicações das inferências automatizadas (GOODMAN; FLAXMAN, 2017).

Dessa forma, a construção de sistemas preditivos explicáveis torna-se fundamental para não só garantir a conformidade às leis da atualidade, como também para melhor engajar com o usuário. Ainda, genuinamente entender como humanos se sentem deve ser um dos focos da automatização da percepção emocional, conforme bem elucidado por Riedl (2016), quando falava sobre inteligência narrativa computacional. Aplicações desse viés estão enraizadas tanto na interação humano-computador, quanto no próprio problema de pesquisa de inteligência artificial (RIEDL, 2016).

Visto que a face humana possui 17 conjuntos musculares[1], há uma infinidade de representações emocionais possíveis, por vezes minimamente diferentes umas às outras. Por isso, busca-se não só a habilidade de observação de regiões não óbvias da face pelo algoritmo de aprendizagem, como também a capacidade de análise pon-

[1] https://www.anatomynext.com/muscles-facial-muscles/

tual de regiões de maior interesse. Essa competência pode ser adicionada a modelos de redes neurais através de mecanismos de atenção (ZHANG *et al.*, 2017; WOO *et al.*, 2018). Com esses módulos, objetiva-se a atenuação de regiões não importantes de uma face para a classificação correta da emoção, bem como a intensificação de localidades que viabilizam o acerto. O uso de ferramentas semelhantes foi bem sucedido em classificações de diferentes espécies de pássaros (FU; ZHENG; MEI, 2017), que diferem em detalhes mínimos, e em tarefas de respostas a partir de imagens (CHEN *et al.*, 2015).

Sendo assim, o presente projeto espera investigar a explicabilidade de algoritmos de inferência de emoções faciais baseados em técnicas de *deep learning*. Comparar-se-ão os efeitos de interpretabilidade de dois mecanismos de atenção adicionados a um modelo genérico de rede neural de forma quantitativa - através de métricas como a acurácia - e de forma qualitativa, pela inspeção visual de mapas de saliência gerados.

Ainda, busca-se compreender se a aprendizagem a partir de faces parcialmente oclusas intensifica a capacidade de generalização dos modelos preditivos. Ou seja, se, ao cobrirmos uma região significativa da face observada, o algoritmo direciona sua atenção a outras regiões da face que poderiam transmitir a informação necessária para a classificação de uma determinada emoção facial. Em outras palavras, se a adição de mecanismos de atenção, o treino com adição de oclusões, ou a união de ambos, permite que redes “enxerguem” emoções de forma mais acurada.

Definem-se, portanto, como objetivos do projeto: (1) estipular três arquiteturas para comparação do impacto da adição de mecanismos de atenção aos resultados de classificação de emoção facial; (2) treinar os modelos preditivos em dois conjuntos de dados, um com e outro sem oclusões faciais; (3) elucidar a performance quantitativa das arquiteturas para as distribuições de dados disponíveis; e (4) inspecionar visualmente a interpretação de emoções de cada arquitetura.

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo serão abordados não só os conceitos técnicos de redes neurais, instrumento principal deste trabalho, como também questões teóricas da psicologia e da expressividade facial humana. Na Seção 2.1, estão as ideias chave de aprendizagem supervisionada por redes neurais, a Seção 2.2 trata da importância de construção de redes explicáveis e interpretáveis, na Seção 2.3 está o conhecimento necessário de expressões faciais para entendimento do projeto, e na Seção 2.4 trabalhos que buscam análises semelhantes às desse.

2.1 REDES NEURAIS E O MODELO HUMANO

No cérebro humano, a informação é transmitida como sinais elétricos através de sinapses nos terminais sinápticos entre os neurônios. Um neurônio, então, libera os sinais elétricos no terminal sináptico, que é capturado por outro neurônio através dos dendritos (HAYKIN *et al.*, 2009).

Salientam-se, então, três pontos principais para observação:

- O neurônio apenas transmite o sinal se um número suficiente de sinais elétricos é capturado pelos dendritos;
- Neurônios podem receber entradas de diversos outros neurônios adjacentes, assim como podem transmitir sinais para diversos outros neurônios adjacentes. As entradas são cumulativas;
- Cada neurônio possui um *threshold* para sua ativação, representado pelo peso sináptico.

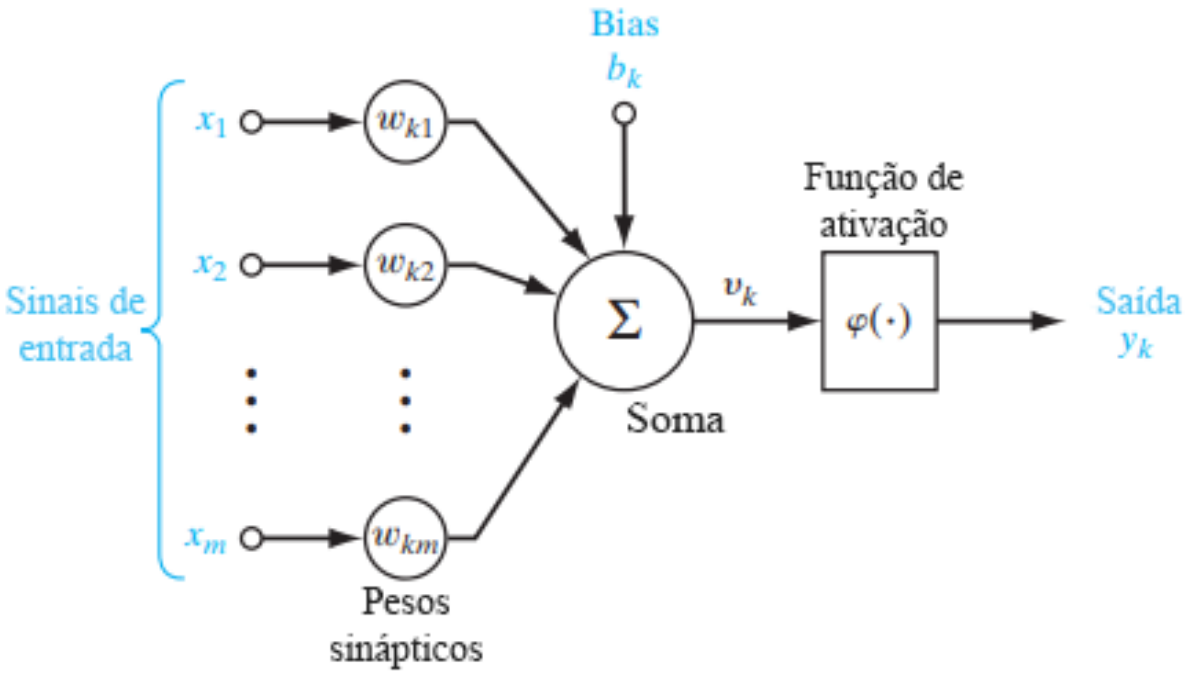
No ramo computacional, existem as redes chamadas de Redes Neurais Artificiais (*ANN*), algoritmos baseados superficialmente no funcionamento de neurônios humanos apresentado. Esses ganham conhecimento através da extração de informações de dados, atuando como aproximadores de funções. Mapeiam entradas para saídas, e são compostas por unidades computacionais interconectadas, conhecidas como neurônios, de forma análoga ao cérebro humano (BOEHMKE, 2018). Esses neurônios possuem pesos individuais que variam de acordo com a tarefa a ser desempenhada. Embora individualmente cada neurônio possua pouca capacidade computacional, ao serem combinados, essa performance aumenta consideravelmente.

De acordo com Haykin *et al.* (2009),

> Uma rede neural é um massivo processador distribuído paralelo composto por unidades simples de processamento que tem a propensão natural de armazenar conhecimento experiencial e disponibilizá-lo para uso. Redes neurais lembram o cérebro em dois aspectos: (1) o conhecimento é adquirido pela rede a partir do ambiente através de um processo de aprendizagem; (2) as conexões interneuronais, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

Na Figura 2.1 está o modelo matemático de um neurônio k, que é composto por entradas, uma saída e, assim como o modelo humano, pesos sinápticos. As entradas x_m são multiplicadas pelo peso de cada sinapse, w_{km}, onde k se refere ao neurônio em questão e m à saída sináptica daquele peso. Os pesos resultantes de cada entrada são somados, de forma que as operações citadas resultem em um combinador linear. Ou seja, o modelo matemático de um neurônio é equivalente a uma função linear. Por fim, o valor obtido é multiplicado por uma "função de ativação", que tem como objetivo não só limitar a amplitude da saída dos neurônios, mas também introduzir não linearidades ao conjunto (HAYKIN *et al.*, 2009).

Figura 2.1 – Modelo matemático de um neurônio.



Fonte: Tradução livre da autora a partir de Haykin *et al.* (2009).

O modelo apresentado na Figura 2.1 pode ser descrito para o neurônio k pelas Equações 2.1 e 2.2 (HAYKIN *et al.*, 2009), onde $x_1, x_2, ..., x_m$ são os sinais de entrada; $w_{k1}, w_{k2}, ..., w_{km}$ são os pesos sinápticos do neurônio k; b_k é o *bias* do neurônio k; v_k é o combinador linear resultante devido à multiplicação dos sinais de entrada pelos pesos do neurônio com adição do *bias*; y_k é a saída esperada do neurônio e φ é a

função de ativação utilizada.

$$v_k = \sum_{m=1}^{m} w_{km} x_m + b_k \tag{2.1}$$

$$y_k = \varphi(v_k) \tag{2.2}$$

Salienta-se a possibilidade de introduzir um valor de *bias*, b_k, que permite aumentar ou diminuir o valor de saída da rede na função de ativação. A inclusão dessa grandeza gera uma transformação afim a partir dos valores anteriores da rede. A transformação afim, cujo exemplo pode ser observado na Figura 2.2, tem como características não só a preservação da colinearidade entre os pontos da reta, mas também das relações de distâncias ao longo de uma linha.

Figura 2.2 – Transformação afim, realizada a partir da adição do *bias* ao neurônio.

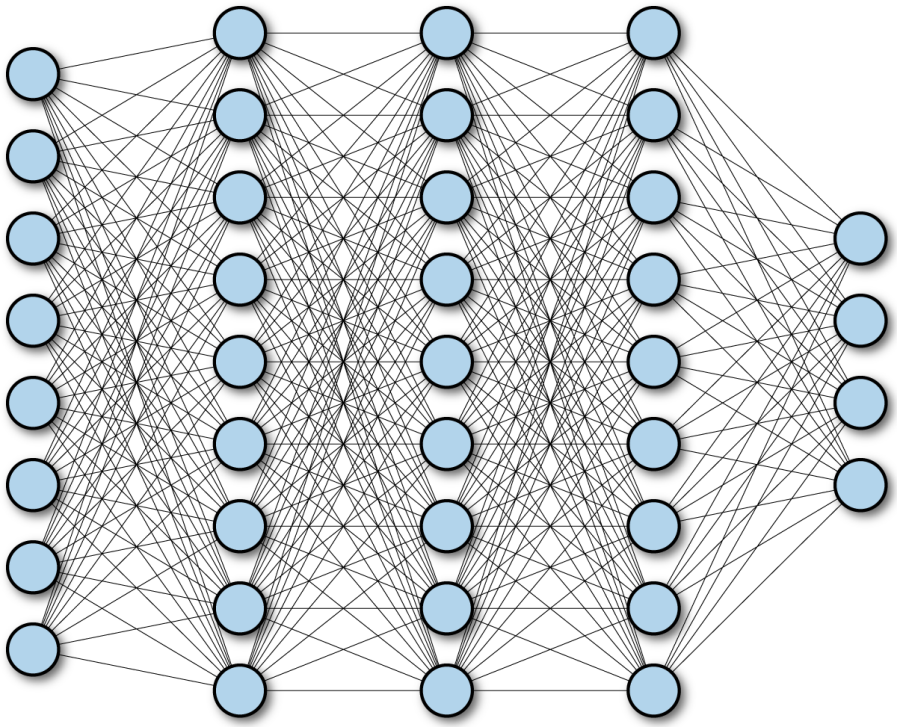
Fonte: Tradução livre da autora a partir de https://www.graphicsmill.com/docs/gm/affine-and-projective-transformations.htm

2.1.1 Redes Neurais Totalmente Conectadas

As redes neurais totalmente conectadas são uma das principais topologias nesse tipo de aprendizagem, visto que não é necessário qualquer pressuposto para com o tipo de entrada na rede (se são imagens, vídeos, dados distribuídos no tempo, etc.) (RAMSUNDAR; ZADEH, 2018). Nessa arquitetura, todos os nós ou neurônios de uma camada estão conectados a todos os outros das camadas imediatamente anterior e posterior. Na Figura 2.3 observa-se a topologia totalmente conectada.

A descrição matemática desse tipo de rede segue o modelo apresentado anteriormente para os neurônios únicos, nas Equações 2.1 e 2.2. Também, com a apresentação dessa topologia, nota-se a possibilidade da incrementação das camadas da rede. Além das entradas e saídas, as redes neurais podem ter quantas camadas escondidas de neurônios desejar-se, aquelas que não são nem entradas nem saídas. Na Figura 2.3, a rede em questão possui três camadas escondidas totalmente conectadas uma a outra.

Figura 2.3 – Topologia de rede neural totalmente conectada.



Fonte: Ramsundar e Zadeh (2018).

2.1.2 Funções de Ativação

As funções de ativação tem como objetivo limitar a amplitude do valor retornado por um neurônio, de forma a comprimir o resultado dentro de uma determinada amplitude de valor finito (HAYKIN *et al.*, 2009). De acordo com Trask (2019), idealmente:

1. as funções de ativação devem ser contínuas;
2. preferencialmente monotônicas (sua direção não deve mudar);
3. são não lineares e introduzem as não linearidades ao modelo de rede neural;
4. são eficientes computacionalmente.

A seguir, algumas funções de ativação relevantes ao presente projeto são exploradas com maior atenção.

2.1.2.1 Função Sigmoid

A função de ativação *sigmoid* exibe equilibradamente tanto comportamento de função linear, quanto de não linear (HAYKIN *et al.*, 2009) e pode ser expressada pela

Equação 2.3 de uma função logística (BISHOP, 2006). Na Figura 2.4a é possível observar o comportamento da *sigmoid*.

$$\varphi(x) = \frac{1}{1 + \exp^{-x}} \tag{2.3}$$

A *sigmoid* apresenta tanto variação contínua de 0 a 1, como também é possível derivá-la. A possibilidade de derivação é de suma importância para a aprendizagem com redes neurais, conforme será explicitado na Seção 2.1.3, que trata do processo de treino.

2.1.2.2 Função ReLU

A função de ativação *Rectified Linear Unit* (*ReLU*), cujo comportamento é observado pela Figura 2.4b e pela Equação 2.4, assemelha-se muito a uma função linear, entretanto resulta em valores não lineares quando aplicada à saída de uma camada, visto que é composta por duas seções lineares distintas (GOODFELLOW; BENGIO; COURVILLE, 2016). A *ReLU* é diferenciável em todos os pontos com exceção de $x = 0$ e a proximidade à linearidade a torna computacionalmente leve, além de favorecer o processo de aprendizagem e intensificar a capacidade de generalização de uma rede neural, característica presente em funções lineares (GOODFELLOW; BENGIO; COURVILLE, 2016). Quando os neurônios possuem valores diferentes de zero, ou seja, quando estão ativos, problemas como o desaparecimento de gradientes, tratado futuramente no texto, são fortemente atenuados (GLOROT; BORDES; BENGIO, 2011), visto que o gradiente será unitário. É a função de ativação mais indicada atualmente (GOODFELLOW; BENGIO; COURVILLE, 2016).

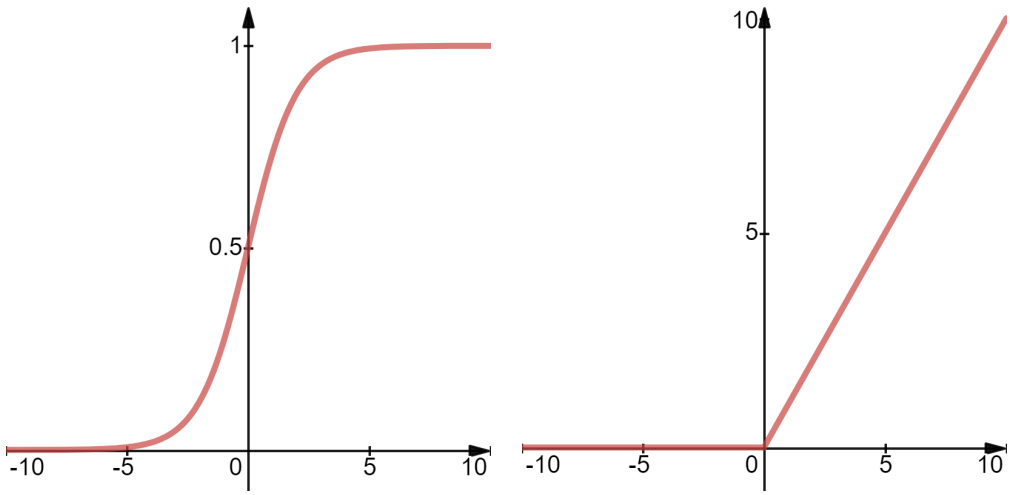
$$\varphi(x) = \begin{cases} 0, & se \quad x < 0 \\ x, & se \quad x \geq 0 \end{cases} \tag{2.4}$$

2.1.2.3 Função Softmax

A função *Softmax* depende de um conjunto de entrada, não sendo possível aplicá-la a apenas um elemento, conforme visualizado pela Equação 2.5. Ela eleva cada entrada exponencialmente e depois a divide pela soma das exponenciais das entradas. Resulta sempre em números positivos cuja soma é 1 (TRASK, 2019).

$$Softmax(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{2.5}$$

Figura 2.4 – Funções de ativação utilizadas em canais de redes neurais.



(a) Sigmoid (b) ReLU

Fonte: autora.

É amplamente utilizada na última camada de redes neurais classificatórias, pois retorna uma estimativa da probabilidade das classes para uma entrada: aquela de maior valor é a classe de maior probabilidade de ser a correta. Essa função tem a característica da "nitidez de atenuação", que é uma forma de encorajamento para a predição de uma saída com alta probabilidade, quão maior for a probabilidade de uma classe, menor serão as das outras (TRASK, 2019).

2.1.3 O Processo de Treino

O processo de treino supervisionado de uma rede neural se dá por três passos: predição, comparação e aprendizagem (TRASK, 2019).

A propagação do sinal de entrada para os neurônios e sua posterior ativação são denominadas como a propagação *forward*, por seguir a ordem natural de entrada a saída da rede. Supondo a ativação randomizada dos pesos, até essa etapa obtemos a saída, ou predição, de uma rede neural. Os elementos apresentados até a seção anterior são os constituintes da propagação *forward*.

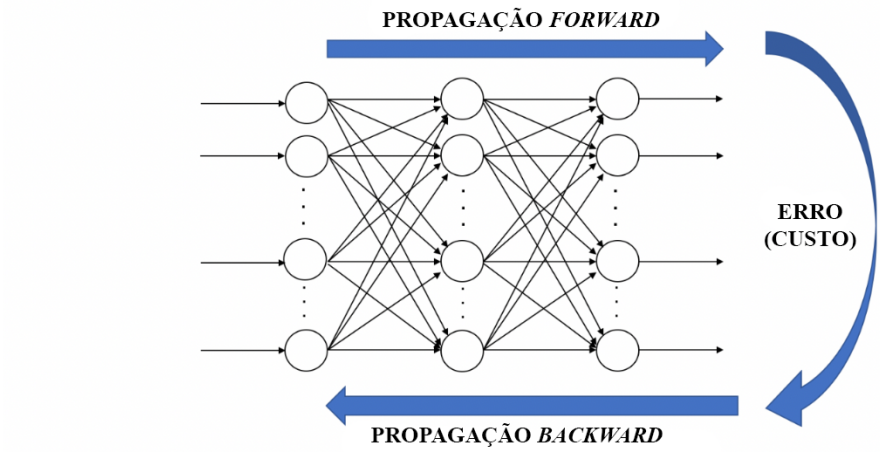
A etapa de comparação se dá com auxílio das funções de perda. Para problemas de classificação, por exemplo, pode-se comparar o rótulo esperado (chamado de *ground-truth* ou *label*, em problemas de classificação) para uma determinada entrada com a predição gerada pela rede. Assim, tem-se uma métrica quantitativa do quão preciso foi o resultado da rede neural naquele momento para aquele dado de entrada.

Por fim, a etapa de aprendizagem consiste, no geral, em minimizar o valor retornado pela função de perda, ao mesmo tempo em que se atualiza os valores dos

neurônios para atingir esse objetivo, chamada *backpropagation*. Deseja-se minimizar a função de perda, de forma a maximizar a taxa de acertos da rede que está sendo treinada. A atualização dos pesos e minimização da função de perda é guiada por um método de otimização. Na Figura 2.5 estão os fluxos *forward* e *backward* de uma rede neural. O processo de aprendizagem para problemas de classificação, portanto, pode ser exemplificado pelo conjunto de passos:

1. Inicializar pesos, tipicamente, de forma aleatória e *biases* dos neurônios;
2. Obter a **predição** gerada pela rede a partir de um conjunto de entradas;
3. **Comparar** a predição com o valor de saída esperado (*label* ou rótulo) e calcular a perda, ou custo;
4. Realizar o *backpropagation* para atualizar os parâmetros dos neurônios de forma a minimizar as perdas, de acordo com o método de otimização escolhido. Esse é o momento de **aprendizagem** da rede;
5. Repetir a partir do passo 2 até atingir um modelo adequado de acordo com o critério estabelecido.

Figura 2.5 – Propagações *forward* e *backward*.



Fonte: Tradução livre da autora a partir de https://towardsdatascience.com/how-do-artificial-neural-networks-learn-773e46399fc7

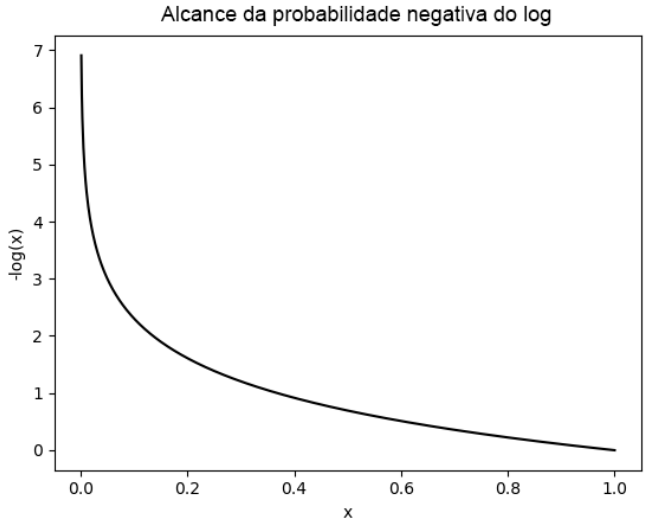
A seguir será exposta uma função de perda, também chamada de “critério”, e métodos de otimização que constituem o processo de aprendizagem para um problema de classificação não-binária ou multi-classe.

2.1.3.1 Entropia Cruzada

A perda, ou erro, para problemas multi-classe pode ser calculada com o uso da função *Softmax* e da probabilidade negativa do log (KARPATHY, 2018). O interesse de utilizar tal união, chamada de entropia cruzada (*Cross Entropy Loss*), para classificação tem em vista as propriedades das duas funções. Conforme observado na Figura 2.6, a probabilidade negativa do log resulta em saídas maiores quando os valores de entrada são pequenos. Sendo assim, como o negativo do log sempre decresce, maximizar a probabilidade nesse caso é equivalente à redução do erro (BISHOP, 2006). De acordo com o explicitado anteriormente, a função *Softmax* retorna um conjunto de C valores (para C classes) cuja soma resulta em 1. Quando há probabilidade elevada para uma das classes, todas as outras tem seus valores reduzidos. Por outro lado, quando não há probabilidade elevada, os valores de probabilidade retornados são baixos e semelhantes (TRASK, 2019).

Assim, ao aplicar-se a probabilidade negativa do log na saída da função *Softmax*, obteremos valores elevados quando a maior probabilidade for baixa (quando a rede ainda está "indecisa"), e, consequentemente, valores baixos quando a probabilidade para uma das classes for alta. Na Equação 2.6 está o processo completo da *Cross Entropy Loss*, onde y corresponde ao rótulo esperado e $\hat{y}$ ao resultado previsto após o *Softmax* (KARPATHY, 2018).

Figura 2.6 – Probabilidade negativa do log, função de perda utilizada em problemas de classificação.



Fonte: Tradução livre da autora a partir de https://ljvmiranda921.github.io/notebook/2017/08/13/softmax-and-the-negative-log-likelihood/

$$CrossEntropy(x_i) = -y \log \hat{y} \tag{2.6}$$

2.1.3.2 Otimização por gradiente descendente estocástico

Bem como outros algoritmos de aprendizagem de máquina, redes neurais fazem uso de métodos de otimização para a evolução dos resultados. A otimização tem como objetivo a minimização ou maximização de um dada função $f(x)$ a partir de um *critério* ou *função de perda* que quantifica o desempenho atual do modelo (GOODFELLOW; BENGIO; COURVILLE, 2016).

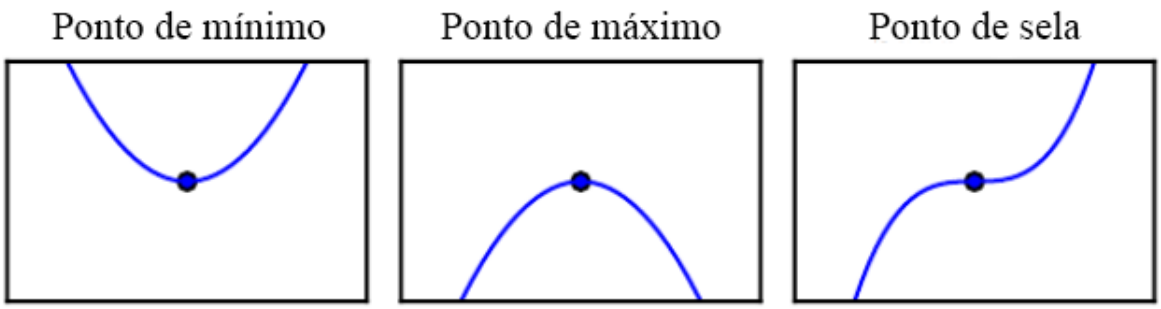
O gradiente descendente (ou, em inglês, *gradient descent*) é um método de otimização que atualiza os pesos dos neurônios da rede de acordo com a predição obtida e o erro calculado pela função de perda utilizada da última camada até a primeira, através de derivações em cadeia. O termo "estocástico" se refere à quantia de amostras em que a otimização é realizada por iteração, que é unitária. Supondo uma função $y = f(x)$, onde tanto x quanto y são números reais, sua derivada é representada por $f' = \frac{dy}{dx}$. Conforme a teoria matemática por trás dessa operação, a derivada de uma função retorna a inclinação de $f(x)$ no ponto x (GOODFELLOW; BENGIO; COURVILLE, 2016). Ou seja, como visualizado na Equação 2.7, fornece a informação de como incrementar x minimamente para obtenção do resultado y correspondente, levando em consideração o uso de uma taxa de crescimento ϵ (em inglês, *learning rate*) que controla essa atualização (GOODFELLOW; BENGIO; COURVILLE, 2016).

$$f(x + \epsilon) \approx f(x) + \epsilon f'(x) \tag{2.7}$$

Quando minimizando funções, tem-se quatro pontos chave na otimização por gradientes: (1) pontos críticos; (2) mínimos locais; (3) mínimos globais e (4) pontos de sela. Os pontos críticos são identificados quando $f'(x) = 0$, resultado que não fornece a direção para onde x deve ser modificado para também alterar y. Pontos de mínimos locais, por sua vez, são regiões onde $f(x)$ é menor que todos os pontos em seu redor, sendo impraticável a redução do valor de $f(x)$ a partir de passos infinitesimais (GOODFELLOW; BENGIO; COURVILLE, 2016). Mínimos globais são os pontos de menor valor na função como um todo. E, por fim, pontos de sela são pontos críticos que não são pontos de mínimo (nem de máximo), onde não é possível realizar adições infinitesimais para modificar $f(x)$ (GOODFELLOW; BENGIO; COURVILLE, 2016). Na Figura 2.7 os pontos de interesse mencionados estão ilustrados, com adição do ponto de máximo.

A presença de pontos de mínimos locais e de sela em excesso prejudica a convergência de redes neurais, dificultando o processo de otimização. Embora o objetivo seja encontrar o ponto de mínimo global de $f(x)$, por vezes um ponto de mínimo local resulta em um modelo bom o suficiente (GOODFELLOW; BENGIO; COURVILLE, 2016), sendo critério do programador o momento de interrupção do treino.

Figura 2.7 – Pontos de interesse para a otimização por gradientes de uma função.



Fonte: Tradução livre da autora a partir de Goodfellow, Bengio e Courville (2016).

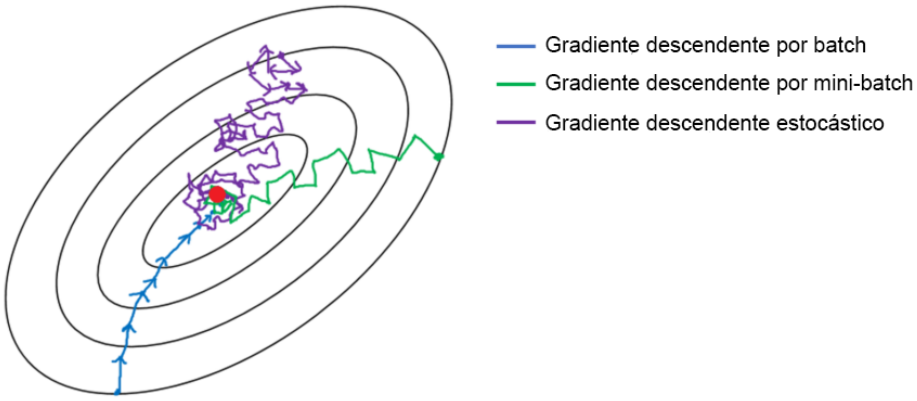
2.1.3.3 Gradiente descendente por mini-batches

O algoritmo de gradiente descendente por *mini-batches* difere do anterior, o estocástico, apenas pelo número de amostras computadas. Enquanto o primeiro realiza a otimização individual de cada entrada, o segundo a realiza para um conjunto m de entradas, chamado de *mini-batch*. O uso desse algoritmo em detrimento do primeiro ocorre devido à maior agilidade para convergência (ou para um desempenho bom o suficiente), por obtermos uma estimação sem *bias* do gradiente médio a cada *mini batch* (GOODFELLOW; BENGIO; COURVILLE, 2016). Ainda, é indicado reduzir a taxa de aprendizagem ϵ gradualmente para melhor convergência (GOODFELLOW; BENGIO; COURVILLE, 2016). Na Figura 2.8 observa-se a diferença de comportamentos entre um algoritmo de gradiente descendente por *batches*, que faz uso da totalidade do conjunto de dados, por *mini-batch*, que usa uma pequena amostra m dos dados (entre 8 e 128 unidades, geralmente) e estocástico, que otimiza amostras individuais.

Outros parâmetros adicionados a esse algoritmo são o momento (POLYAK, 1964) e a queda de pesos (*weight decay*). O primeiro busca a aceleração da otimização através de um acúmulo exponencial da média móvel descendente dos gradientes passados, de forma a incentivar a movimentação em sua direção (GOODFELLOW; BENGIO; COURVILLE, 2016). A queda dos pesos, por sua vez, é um conhecido método de regularização, também chamado de penalidade L^2. O uso de técnicas de regularização objetiva aumentar o poder de generalização de uma rede neural[1]. Esse método direciona os pesos dos neurônios às proximidades da origem através da adição de um fator de regularização $\Omega(\theta) = \frac{1}{2} \|w\|_2^2$ (GOODFELLOW; BENGIO; COURVILLE, 2016).

[1] https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/

Figura 2.8 – Comparação entre otimização de gradiente descendente por *batch*, *mini-batch* e estocástico.



Fonte: Tradução livre da autora a partir de https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3

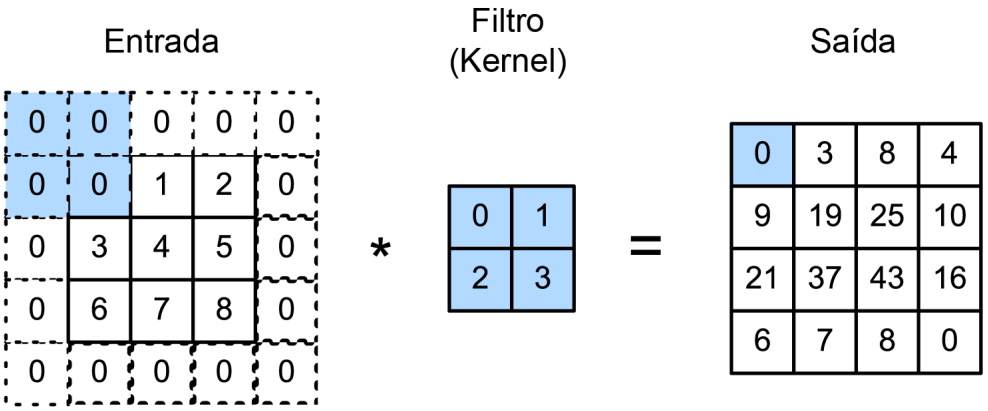
2.1.4 Redes Neurais Convolucionais

Conforme Goodfellow, Bengio e Courville (2016), redes neurais convolucionais são um tipo especializado de redes neurais para o processamento de dados tabulares, como imagens (dados tabulares de duas dimensões) e séries distribuídas no tempo (dados tabulares de uma dimensão). Esse tipo de rede faz uso de convoluções em pelo menos uma de suas camadas, conforme sugerido pelo seu nome, ao invés das multiplicações matriciais usuais como nas redes totalmente conectadas.

A operação de convolução 2D consiste no deslocamento de um filtro, ou *kernel*, que é uma pequena matriz de pesos $n \times n$, através do dado tabular a ser convolucionado (GOODFELLOW; BENGIO; COURVILLE, 2016). O filtro realiza operações de multiplicação número a número, cuja soma resulta em um único valor de uma nova matriz gerada. Essa nova matriz pode ser chamada de mapa de *features*. Na Figura 2.9 observa-se a aplicação da operação de convolução a um dado tabular. O presente projeto fará uso de imagens como dados de entrada, portanto, as entradas de convoluções serão referenciadas no texto como tais.

Quando utiliza-se imagens como dados de entrada, redes convolucionais tem vantagens em relação às totalmente conectadas, conforme explicitado por Karpathy (2018). Supondo uma imagem de três canais, de $32 \times 32 \times 3$ *pixels*, uma rede totalmente conectada teria cerca de 3072 pesos de neurônios. Entretanto, se a imagem for maior, de $200 \times 200 \times 3$ *pixels*, há um total de 120.000 pesos, o que pode levar à dificuldade de convergência devido ao alto número de parâmetros otimizáveis (KARPATHY, 2018). Sendo assim, o uso de operações convolucionais restringe a rede à uma pequena região da camada anterior, ao invés da totalidade dos pesos como a rede totalmente

Figura 2.9 – Operação de convolução a um dado tabular com um filtro 2×2.

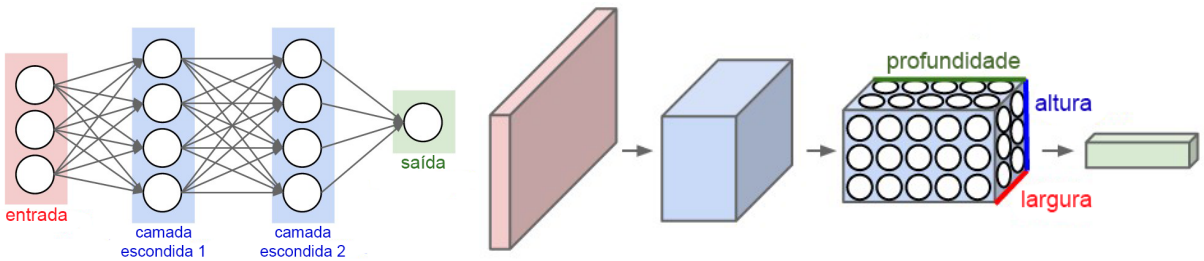


Fonte: Tradução livre da autora a partir de http://deeplearning.net/software/theano/tutorial/conv_arithmetic.html

conectada (KARPATHY, 2018).

O número de filtros utilizados na operação de convolução define a quantia de canais de *features* (representado pela grandeza de profundidade). Por outro lado, os valores de altura ou largura das *features* de saída são estipulados tanto pelo tamanho do filtro, quanto pelo passo de convolução (ou *stride*), que é o salto de *pixels* realizado pelos filtros entre cada computação de valores. Na Figura 2.10, observam-se as diferenças de processamento dos dados entre uma rede neural totalmente conectada em relação a uma rede convolucional. A última é capaz de rearranjar as dimensões do dado de entrada.

Figura 2.10 – Comparação entre redes totalmente conectadas e convolucionais. Observa-se que a segunda é capaz de redimensionar o dado de entrada em três dimensões.



(a) Rede totalmente conectada. (b) Rede convolucional.

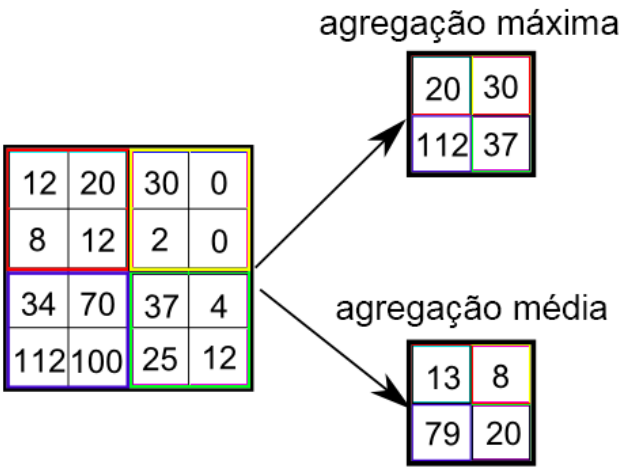
Fonte: Tradução livre da autora a partir de http://cs231n.github.io/convolutional-networks/

2.1.4.1 Agregação

Tipicamente, uma rede convolucional é composta por três etapas: (1) as camadas convolucionais, que geram pesos lineares, que são entregues a (2) funções de ativação, responsáveis pela adição de não-linearidades a cada camada; e (3) uma etapa de agregação ou *pooling* ao fim do processo para modificação dos resultados obtidos em (2) (GOODFELLOW; BENGIO; COURVILLE, 2016). O uso de camadas de agregação objetiva o aumento da robustez da rede, de forma a tornar sua representação quase que invariante a pequenas modificações na entrada. Intensifica a *existência* de determinada característica, ao invés de *onde* ela está localizada (GOODFELLOW; BENGIO; COURVILLE, 2016).

Dois dos modos de agregação amplamente utilizados são a agregação máxima (*Max Pooling*) e a agregação média (*Average Pooling*). O primeiro retorna o valor máximo presente na região de interesse de um filtro $n \times n$, enquanto o segundo retorna a média de tais valores, conforme a Figura 2.11.

Figura 2.11 – Operações de agregação máxima (*Max Pooling*) e média (*Average Pooling*).



Fonte: Tradução livre da autora a partir de https://www.quora.com/What-is-max-pooling-in-convolutional-neural-networks

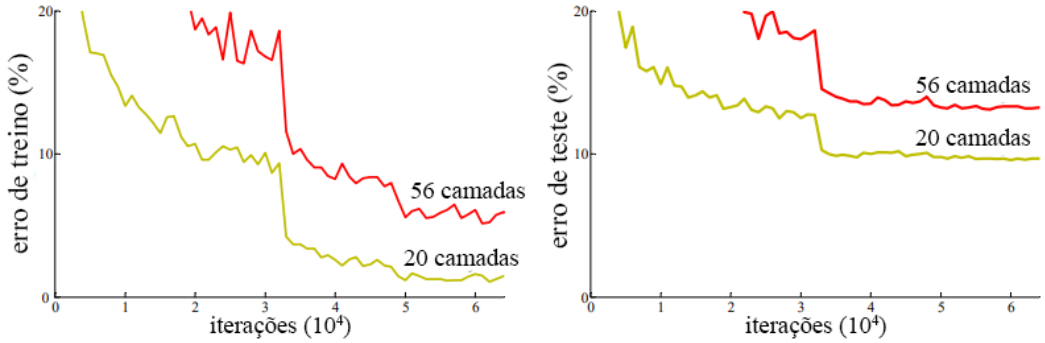
2.1.5 Redes Neurais Convolucionais Residuais

Com o aumento do poder computacional dos últimos anos através do uso de unidades de processamento gráfico (*GPUs*), cresceu a possibilidade de criação de redes neurais profundas, com uma quantidade elevada de camadas escondidas. Entretanto, ao mesmo tempo em que redes mais profundas tem maior poder descritivo,

seu treino é conturbado e a rede torna-se de difícil otimização, sofrendo com problemas como gradientes que somem (*vanishing gradients*) antes de percorrer todas as camadas (HOCHREITER, 1998; WANG, 2019).

O problema dos gradientes que somem ocorre devido aos valores de pesos serem muito pequenos (quando utiliza-se funções de ativação como a *Sigmoid*, que limitam os valores em no máximo 1, por exemplo). No processo de *backpropagation*, os gradientes desses valores reduzidos multiplicam-se e tornam-se cada vez menores. Ao final do percurso, o gradiente é ínfimo o suficiente para não gerar alterações significativas na atualização dos pesos dos neurônios (WANG, 2019). Outro problema presente com o aumento da profundidade das redes neurais é a degradação, ou a redução da acurácia em comparação com arquiteturas menores (HE *et al.*, 2016), conforme observado na Figura 2.12. Como solução dessas questões, surgem as Redes Residuais (HE *et al.*, 2016).

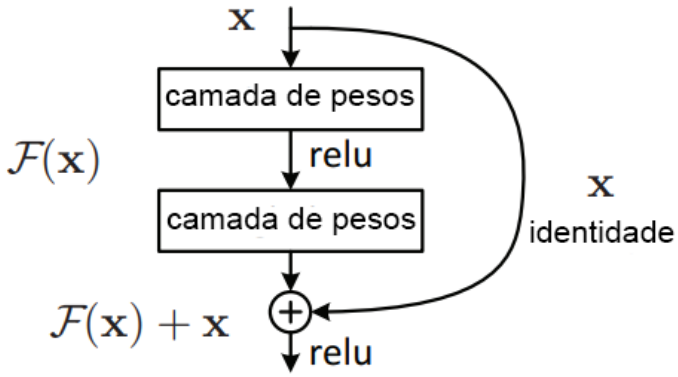
Figura 2.12 – A degradação da acurácia de redes profundas.



Fonte: Tradução livre da autora a partir de He *et al.* (2016).

As redes residuais vão de encontro a arquiteturas tradicionais e puramente sequenciais, trazendo o conceito de *conexões de atalho* (ou "*shortcut connections*") (HE *et al.*, 2016), visualizadas na Figura 2.13, que permitem a propagação do sinal por todo o modelo. O valor obtido no bloco anterior nunca passa pela função de ativação do bloco atual e, portanto, não sofre a redução comentada anteriormente (WANG, 2019). Ainda, essa solução tem fundamentação no *mapeamento de identidade*, quando $f(x) = x$. A adição das conexões de atalho permitem que seja verdade a afirmação de que "um modelo mais profundo não deve produzir um erro de treino superior que seu modelo mais superficial de contrapartida" (HE *et al.*, 2016).

Figura 2.13 – Bloco da rede residual com a conexão de atalho que propaga o sinal de identidade x ao longo da arquitetura.



Fonte: Tradução livre da autora a partir de He *et al.* (2016).

2.2 A NECESSIDADE DE REDES INTERPRETÁVEIS E EXPLICÁVEIS

Redes neurais são tidas como métodos “caixa preta” por produzirem pouca explicação de como cada unidade de processamento interfere em um processo de predição (OLDEN; JACKSON, 2002). Assim, a necessidade de visualização do que ocorre entre suas camadas é de suma importância para a interpretabilidade de uma arquitetura, e soluções são exploradas há algum tempo (OLDEN; JACKSON, 2002; TZENG; MA, 2005). Estudam-se, portanto, maneiras de visualizar as camadas interiores de uma rede neural, de forma a compreender como essa chegou a uma determinada classificação, ou o porquê de ter gerado uma classificação errônea, por exemplo.

Ao relacionar a interpretabilidade de redes com sua profundidade, Bau *et al.* (2017) constataram que redes mais profundas são mais interpretáveis que menores, quando define-se interpretabilidade de forma análoga à interpretação visual humana, pela quantificação do alinhamento das respostas dos neurônios de uma rede em relação à mesma imagem rotulada por humanos (BAU *et al.*, 2017). Zeiler e Fergus (2014) observam que, com o aumento da profundidade de uma rede, suas camadas adquirem maior poder de representação (e consequentemente maior acurácia de classificação).

A investigação de algoritmos inteligentes tem sua necessidade intensificada gradualmente, até mesmo em termos legais. Em 2018, na União Europeia, entrou em vigor a nova “Regulamentação Geral de Proteção de Dados” (*General Data Protection Regulation*[2] ou GDPR). Essa regulamentação exige que modelos de inferência automatizados e individuais (serviços de sugestões e reconhecimento facial, por exemplo) venham acompanhados de explicações acerca de seu funcionamento e resultados, livres para requisição por qualquer usuário. Tal medida, portanto, estabelece o “direito à explicação” de um algoritmo automatizado para o usuário cotidiano (GOODMAN;

[2]Regulamentação (EU) 2016/679

FLAXMAN, 2017). No Brasil, tramita a nova Lei Geral de Proteção de Dados Pessoais (LGPD)[3] [4], inspirada na regulamentação europeia, que estabelece normas não só sobre o uso, mas também sobre o tratamento e armazenamento dos dados do usuário, objetivando o domínio por parte do usuário sobre seus dados pessoais. Entretanto, o direito à explicação ainda não se faz presente no texto proposto.

Tais medidas protetivas tem como um de seus objetivos a redução da discriminação que os algoritmos podem produzir, definida como o tratamento injusto de um indivíduo devido a seu pertencimento a um grupo específico em termos de raça, gênero, religião, etc. (ALTMAN, 2016). Ainda, almeja-se o aumento da segurança desses sistemas, visto que cotidianamente uma grande parcela da população utiliza-os, mesmo sem notar, em plataformas de distribuição de música e nas próprias redes sociais. Citam-se como maus exemplos os casos de algoritmos racistas e sexistas ao atribuir possíveis profissões[5] ou que tem tendência a violar os direitos humanos em sistemas de imigração[6].

2.2.1 A Explicabilidade e a Interpretabilidade de Redes Neurais

Conforme descrito por Gilpin *et al.* (2018),

> O objetivo da interpretabilidade é descrever o interior de um sistema de uma forma compreensível para os humanos. O sucesso desse objetivo é ligado à cognição, conhecimento e tendências (*biases*) do usuário: para um sistema ser interpretável, ele deve produzir descrições que são simples o suficiente para uma pessoa entender utilizando um vocabulário que é significativo para o usuário.

Por isso, a clareza de como uma determinada rede neural construiu sua resposta é de alta relevância não só para permitir a verificação mais intuitiva de seu desempenho, mas também para aprofundar o entendimento de seu funcionamento. Entretanto, a interpretabilidade de um rede neural é sempre balanceada em termos de sua completude. Métodos de explicação completos, como os matemáticos, que descrevem o sistema de uma forma acurada, geralmente são menos simples e menos interpretáveis. Por outro lado, os menos completos e mais interpretáveis são melhores aceitos pelas pessoas, por serem facilmente compreensíveis (GILPIN *et al.*, 2018).

Existem três vieses para a elucidação da explicabilidade de uma rede neural. O primeiro, focado na explicação do processamento, busca solucionar a questão: “Por

[3] Lei 13.709 de 14 de agosto de 2018

[4] https://politica.estadao.com.br/blogs/fausto-macedo/a-protecao-dos-dados-pessoais-chega-ao-brasil/

[5] https://www.siliconrepublic.com/machines/imagenet-roulette-ai-racism

[6] https://ihrp.law.utoronto.ca/news/canadas-adoption-ai-immigration-raises-serious-rights-implications

que uma determinada entrada leva a uma determinada saída?". O segundo é focado na representação, ou no funcionamento interno de um programa, e objetiva responder a questão "Qual informação essa rede contem?" (GILPIN *et al.*, 2018). Em um terceiro âmbito há redes auto-explicáveis, construídas de forma a simplificar a interpretabilidade de seu próprio funcionamento, representação ou operação (GILPIN *et al.*, 2018).

Este projeto objetiva o estudo da explicabilidade de redes neurais em duas das esferas citadas: (1) através de redes auto-explicáveis, construídas com os ditos "mecanismos de atenção" que serão melhor detalhados na próxima seção e no capítulo de metodologia e (2) entender por que uma entrada leva a uma saída. Busca-se a aplicação do segundo método junto ao primeiro. Deseja-se compreender como diferentes mecanismos de atenção influenciam a saída de uma rede neural.

2.2.2 Mecanismos de Atenção em Redes Convolucionais

Por vezes, redes neurais geram classificações ou estimações errôneas pelo excesso de informação em uma imagem, ou pela confusão em relação a qual região da imagem de entrada carrega a informação mais relevante para uma determinada classificação. Para atenuar esse problema, os mecanismos de atenção, quando aplicados em redes convolucionais, tem como objetivo a intensificação das regiões mais relevantes de uma imagem de entrada para uma certa classe de saída. Com esses mecanismos, espera-se tanto o aumento da acurácia de uma rede classificatória, quanto a atenuação dos pesos atribuídos às regiões não relevantes à classificação correta (WOO *et al.*, 2018). O uso dessas ferramentas foi bem sucedido não só em tarefas de geração de legendas sobre imagens (XU *et al.*, 2015), como também na resposta de perguntas sobre imagens (*Visual Question Answering*) (CHEN *et al.*, 2015) e o reconhecimento de diferentes espécies de pássaros, onde os detalhes que os diferem são mínimos (FU; ZHENG; MEI, 2017).

Esses instrumentos se inspiram no sistema visual humano. Humanos, quando observando uma cena, tendem a focar em seus aspectos mais salientes, de forma a não só reduzir a necessidade de "processamento dos pixels", bem como prover a habilidade de enxergar detalhes mínimos (LAROCHELLE; HINTON, 2010).

As ferramentas de atenção, contudo, podem tomar diversas formas e arranjos. No presente projeto, serão priorizados mecanismos de atenção que podem ser facilmente integrados a redes neurais existentes e que, portanto, não interferem significativamente na arquitetura dessas. Nesse contexto de aplicabilidade, existem mecanismos de atenção focados na extração de informações em diferentes dimensões das camadas convolucionais, realizando operações em termos (1) de profundidade, (2) de

espaço ou (3) de ambos, que serão aprofundados no capítulo de metodologia.

2.2.3 Visualização por Gradientes

Quando estudam-se redes convolucionais, uma análise qualitativa para a questão da interpretabilidade de redes profundas é atingida através do cálculo de gradientes (ERHAN *et al.*, 2009; SIMONYAN; VEDALDI; ZISSERMAN, 2013). Essa técnica é amplamente utilizada para a resolução da primeira questão citada anteriormente: “Por que uma determinada entrada leva a uma determinada saída?”.

Simonyan, Vedaldi e Zisserman (2013) apresentaram duas abordagens para a visualização de redes neurais, cujas representações são visualizadas na Figura 2.14: a primeira, baseada no trabalho de Erhan *et al.* (2009), consiste em gerar imagens que maximizam a pontuação para uma determinada classe, de forma a compreender qual a noção de representatividade que a rede neural convolucional tem para uma classe específica.

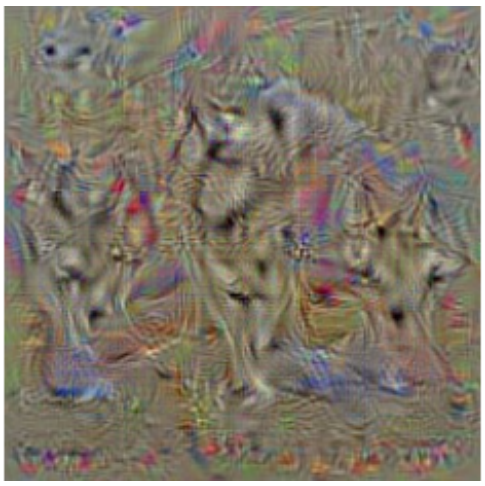
A segunda técnica é a de mapas de saliências. Computam-se os gradientes para uma camada convolucional com uma iteração de *backpropagation* e, a partir de uma imagem de entrada e uma classe, tem-se a importância de cada pixel da imagem para aquela classe. Formalmente, dada uma imagem I_0, uma classe c e uma rede convolucional de classificação com a função de pontuação de classes $S_c(I)$ (a função de perda), almeja-se classificar os *pixels* de I_0 de acordo com sua influência na pontuação $S_c(I_0)$. Com isso, espera-se que os *pixels* mais ativados correspondam à localização da classe na imagem. Ainda, quão maior a magnitude do gradiente de um *pixel*, menor a mudança necessária nesse para impactar a classificação resultante da rede convolucional (SIMONYAN; VEDALDI; ZISSERMAN, 2013).

De encontro aos métodos propostos anteriormente, a técnica de visualização por mapas de saliência é mais rápida, por requerer apenas uma iteração de *backpropagation*. Além disso, necessita-se apenas da imagem de entrada e de uma classe alvo para computação dos valores (SIMONYAN; VEDALDI; ZISSERMAN, 2013).

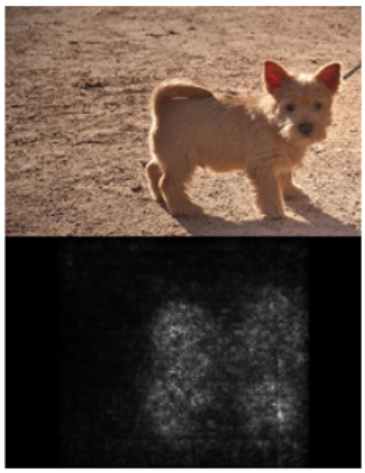
2.2.3.1 A confiabilidade da inspeção visual em métodos geradores de mapas de saliência

Recentemente, Adebayo *et al.* (2018) investigaram a eficácia das descrições geradas a partir de diferentes métodos de visualização por mapas de saliências. Nesse trabalho, os autores compararam métodos baseados em aprendizagem de máquina como a *Backpropagation* Guiada (ZEILER; FERGUS, 2014), o Gradiente $\odot$ Entrada

Figura 2.14 – Os dois métodos de explicação estudados por Simonyan, Vedaldi e Zisserman (2013).



(a) Mapa de ativação máximo da classe “Husky”, técnica apresentada por Erhan *et al.* (2009).



(b) Mapa de saliência gerado a partir da imagem do cachorro.

Fonte: Simonyan, Vedaldi e Zisserman (2013).

(ou *Gradient* $\odot$ *Input*) (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017) e o *Gradient-weighted Class Activation Mapping* (*Grad-CAM*) (SELVARAJU *et al.*, 2017) com um método puramente descritivo, uma função fixa e invariável da imagem de entrada, que é um detector de bordas. Objetivaram, portanto, compreender que tipo de visualizações os métodos geram, como seus resultados diferem entre si e quais suas limitações tanto em relação aos dados de treino, quanto em relação ao modelo analisado.

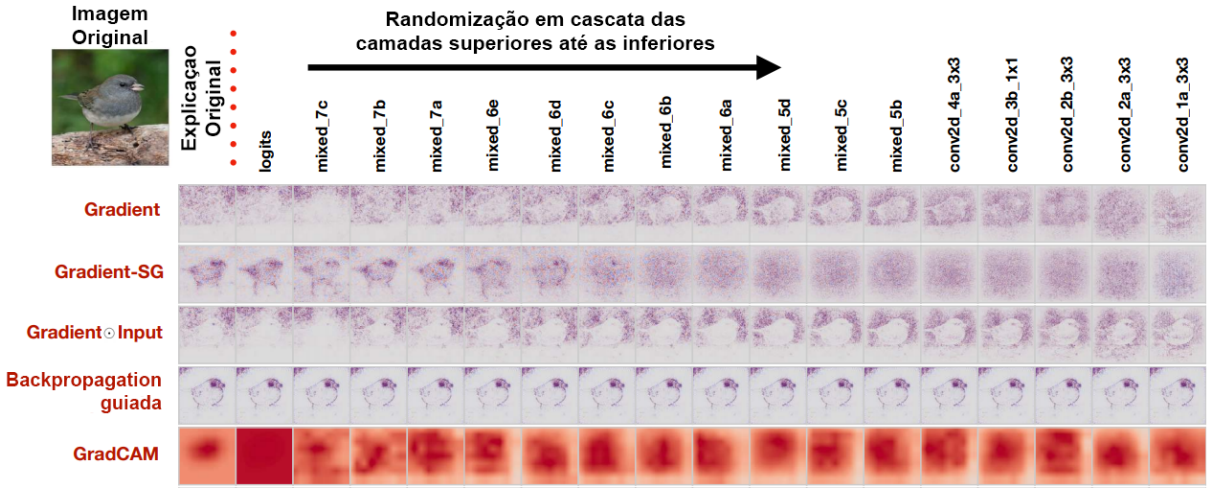
Tal inspeção é obtida a partir de dois ensaios: o primeiro é o treino de um modelo com dados ruidosos; e o segundo, a degradação dos modelos treinados a partir da randomização dos seus parâmetros. Espera-se que, com modificações em quaisquer uma dessas variáveis, as visualizações geradas pelos algoritmos também sejam alteradas.

O método de *Backpropagation* Guiada (SPRINGENBERG *et al.*, 2014) se baseia nas operações de deconvolução (ZEILER; FERGUS, 2014; ZEILER *et al.*, 2010) e, enquanto os gradientes negativos são levados a zero, os positivos passam por uma função de ativação *ReLU*. O trabalho de Shrikumar, Greenside e Kundaje (2017), por outro lado, realiza a multiplicação elemento a elemento do gradiente pelo dado de entrada. O último método citado, *Grad-CAM* (SELVARAJU *et al.*, 2017), será explicitado em maior detalhes no próximo capítulo, mas tem como objetivo a geração de uma mapa de gradiente (ou de saliência) das regiões da imagem de entrada de maior influencia para uma determinada predição da rede convolucional.

Como conclusão da investigação, Adebayo *et al.* (2018) constataram que os métodos baseados em *Backpropagation* Guiada são invariantes a modificações não

só dos dados em que o modelo foi treinado, como também dos modelos em si, tendo comportamento semelhante a métodos descritivos, como o detector de bordas. Por outro lado, os métodos baseados apenas em gradientes, como o Gradiente ⊙ Entrada e o *Grad-CAM*, foram bem sucedidos nos testes de sanidade, tendo resultados variáveis em ambas situações citadas. Por isso, o último método será utilizado para visualização dos resultados obtidos neste trabalho. Na Figura 2.15 estão as comparações dos resultados obtidos para os algoritmos quando utilizam-se dados de modelos degradados. Observa-se a invariabilidade de métodos guiados, enquanto que métodos de gradientes tem sua interpretação modificada.

Figura 2.15 – Explicações geradas a partir da degradação em cascata de modelos convolucionais.



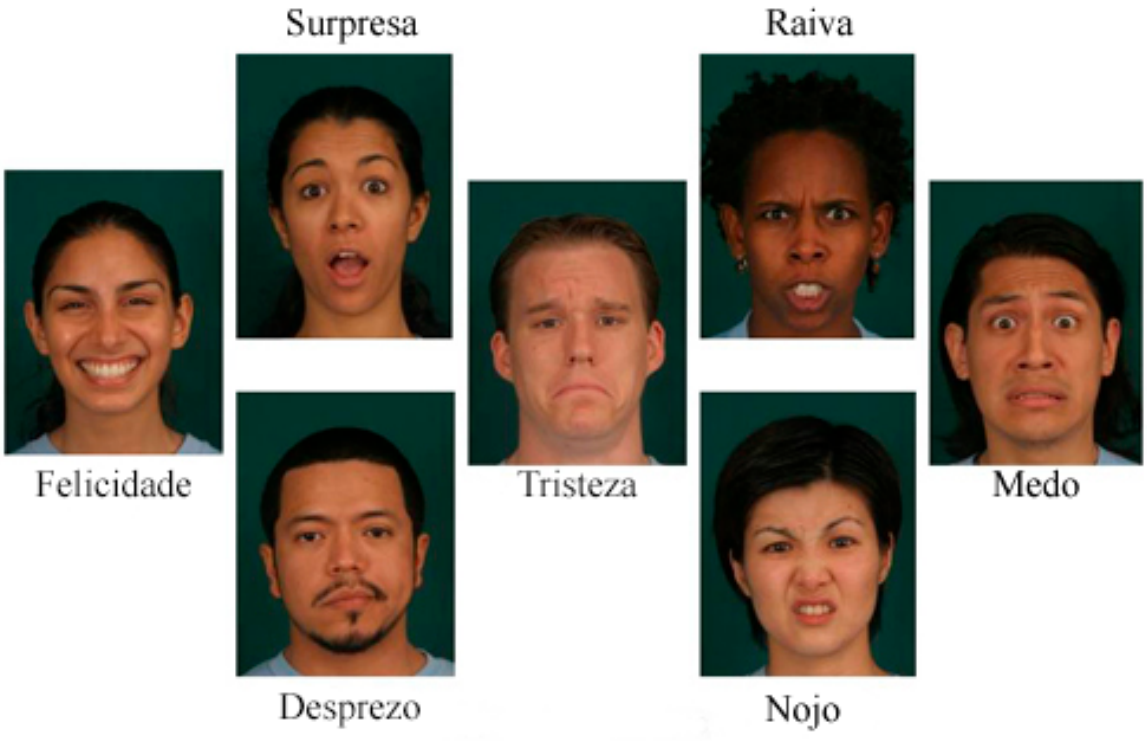
Fonte: Tradução livre da autora a partir de Adebayo *et al.* (2018).

2.3 EXPRESSÕES E EMOÇÕES FACIAIS

Darwin (1872) foi o precursor no estudo que conecta os sentimentos e suas representações por meio da expressão facial. Nesse momento, atribuiu à evolução as emoções transmitidas pelas expressões faciais e também julgou imprescindível o estudo das expressões de animais para melhor compreendermos a dos humanos. Desde então, Ekman (1989) desenvolveu ainda mais o estudo desse tópico, o que o levou à conceitualização das seis emoções básicas representadas pela face que sobrepõem barreiras culturais, sendo elas: a raiva, o medo, a felicidade, a tristeza, o nojo e a surpresa (EKMAN, 1989), cujos exemplos são visualizados na Figura 2.16. Conforme Ekman (1992) afirma, a ampla pesquisa em emoções faciais, expressadas tanto espontânea quanto deliberadamente, fornece evidências substanciais à existência e à distinção das emoções citadas.

Ekman (1992) aponta a existência de nove características para distinção não só entre as seis emoções básicas, mas também em relação a outros fenômenos afetivos, conforme observa-se na Tabela 2.1, dentre as quais cita-se como exemplo a coerência com a resposta emocional. As características 1, 3, e 4 permitem a distinção de uma emoção a outra, enquanto as outras são eficazes na distinção a outros estados afetivos, como temperamentos. Temperamentos diferem de emoções no seu tempo de duração, que é superior (EKMAN; DAVIDSON, 1994), podendo prolongar-se por horas ou dias em comparação a emoções, que perduram por segundos a minutos.

Figura 2.16 – As seis emoções básicas de Ekman, mais o desprezo, expressão algumas vezes adicionada ao conjunto.



Fonte: Tradução livre da autora a partir de https://www.apa.org/science/about/psa/2011/05/facial-expressions.

Ekman (1992) propõe, ainda, a existência de “famílias de emoções”, onde estariam englobadas não só variações das seis emoções básicas, mas também diversas outras emoções não completamente diferenciáveis de acordo com seu estudo. Essa divisão leva em consideração o compartilhamento de certas configurações da face, padrões musculares característicos de uma única família e, portanto, suficientes para a diferenciação dessas.

Existem cerca de 60 diferentes expressões correlacionadas à raiva, por exemplo, mas todas elas apresentam estímulos musculares em comum: a ativação dos músculos das sobrancelhas, rebaixados e aproximados, a elevação da pálpebra supe-

Tabela 2.1 – As nove características diferenciadoras de emoções e estados afetivos.

	Característica	Distinção entre	
		Emoções	**Estados afetivos**
1.	Sinais universais distintos	x	x
2.	Presença em outros primatas		x
3.	Fisiologia distintiva	x	x
4.	Universais distintos em eventos antecedentes	x	x
5.	Coerência com a resposta emocional		x
6.	Início rápido		x
7.	Breve duração		x
8.	Avaliação automática		x
9.	Ocorrência espontânea		x

Fonte: Tradução livre da autora a partir de Ekman (1992).

rior e o aperto dos músculos dos lábios. O autor atribui como fatores diferenciadores das expressões faciais de emoções internas de cada família (1) as diferentes experiências de aprendizado, (2) diferentes ocasiões ou contextos onde as emoções são apresentadas, e (3) diferentes constituições biológicas, essa última corroborada por Öhman (1986).

2.3.1 As Unidades de Ação Facial

Em outro âmbito, a análise da expressividade facial levou ao estudo das unidades de ação facial por Ekman e Friesen (1976), padronizadas pelo Sistema de Codificação de Ação Facial, que será referenciado no texto por *FACS*, sigla correspondente na língua inglesa (EKMAN; FRIESEN, 1978). Tal trabalho surgiu em contraponto ao que ocorria na época: realizavam-se estudos de comportamento facial de acordo com a percepção subjetiva da face por terceiros. Esperava-se entender se observadores conseguiriam identificar emoções de acordo com a face ou se a interpretação das emoções diferia de acordo com a cultura, por exemplo. Assim, o *FACS* emerge como uma técnica analítica de observação da face, que permite a descrição de qualquer movimento facial (EKMAN; FRIESEN, 1976) e se estabelece como um guia para a descrição padronizada desses a partir dos músculos da face.

Ekman e Friesen (1976) salientam que uma das barreiras do *FACS* é a possibilidade de análise apenas de ativações musculares claramente visíveis na face, que podem trazer consequências sociais, descartando-se aquelas de difícil distinção. A escolha da análise apenas do que pode ser visto também tem como objetivo sua aplicabilidade em quaisquer registros video ou fotográficos, não sendo necessário o uso

de equipamentos como detectores de sinais eletromiográficos (EMG).

É possível correlacionar a ativação de determinadas unidades de ação com as seis emoções faciais, conforme observado na Tabela 2.2. A análise das unidades de ação pelo *FACS* permite uma interpretação muito mais abrangente, de forma que mesmo emoções específicas ou representações faciais não distinguíveis pelas nove características de Ekman (1992), apresentadas anteriormente, possam ser estudadas. Na Figura 2.17 algumas das unidades de ação faciais estão exemplificadas.

Tabela 2.2 – Unidades de ação facial presentes em cada uma das seis emoções básicas e o desprezo.

Emoção	**Critério**
Raiva	AU23 e AU24 devem estar presentes
Nojo	AU9 ou AU10 devem estar presentes
Medo	AU1, AU2 e AU4 devem estar presentes
Felicidade	AU12 deve estar presente
Tristeza	AU1, AU4 e AU15 devem estar presentes.
Surpresa	AU1 e AU2 devem estar presentes
Desprezo	AU14 deve estar presente (unilateral ou bilateral)

Fonte: Lucey *et al.* (2010).

Figura 2.17 – Algumas unidades de ação faciais.

(a) AU1 - Levantador de Sobrancelha Externa

(b) AU2 - Levantador de Sobrancelha Interna

(c) AU4 - Abaixador de Sobrancelha

(d) AU9 - Enrugador de Nariz

(e) AU12 - Puxador de Canto do Lábio

(f) AU15 - Depressor de Canto do Lábio

Fonte: Nijs (2016).

3 TRABALHOS RELACIONADOS

Nesse capítulo estão expostos trabalhos relacionados ao presente projeto. Aqui, revisam-se métodos de estudo de classificação de emoções faciais, bem como o enfrentamento a imagens com presença de oclusões. Ainda, projetos com análises através de métodos qualitativos são expostos.

3.1 CLASSIFICAÇÃO DE EMOÇÃO EM FACES SEM OCLUSÃO

Com o crescente uso de máquinas no cotidiano humano, a necessidade de empatia ou compreensão para com quem interagem torna-se fundamental, de forma a produzir sistemas humano-centrados. Nesse sentido, a área de Análise de Expressão Facial (AEF) encarrega-se de automaticamente reconhecer qual a provável emoção transmitida pela face das pessoas. A capacidade de compreender emoções abre possibilidades não só em setores de interação humano-computador, com robôs sensíveis, como também no monitoramento de fatiga de motoristas ou mesmo jogos interativos (ZHANG *et al.*, 2018b). Por isso, a detecção de emoções em faces é amplamente investigada.

Uma parcela dos trabalhos, entretanto, mantém-se estrito ao mesmo conjunto de dados tanto para treino, quanto para teste. Nesse contexto, Liu *et al.* (2013), que fazem uso do mesmo método de validação de modelos (validação em 10 *folds*) e quantia de dados utilizada nesse trabalho (1308 imagens distribuídas em 8 classes), obtiveram 92,05% de acurácia para o conjunto de dados CK+. Esse resultado foi alcançado com o modelo *AUDN* (*Action Unit Aware Deep Networks*), que leva em consideração as unidades de ação facial para estimação das emoções. Ding, Zhou e Chellappa (2017), por outro lado, atingem 96,8% de acurácia para as mesmas 8 classes e distribuição de dados utilizadas através de uma rede de duas etapas de treino, a primeira em relação a atributos da face, e a segunda das expressões faciais, chamada *FaceNet2ExpNet*.

3.1.1 Cruzamento de *datasets*

A classificação de emoções entre diferentes conjuntos de dados (cruzamento de *datasets*) foi investigada, porém em menor intensidade. Shan, Gong e McOwan (2009), por exemplo, fazem uso de *Support Vector Machines* junto a outros métodos de aprendizagem de máquina para obter a classificação de emoções entre diferentes

bancos de imagens. Quando seu modelo foi treinado no conjunto CK+ e testado no JAFFE obtiveram 41,3% de acurácia.

Zavarez, Berriel e Oliveira-Santos (2017), por outro lado, fizeram uso de diversos conjuntos de dados, como o CK+ (LUCEY *et al.*, 2010), o JAFFE (LYONS *et al.*, 1998), o MMI (PANTIC *et al.*, 2005), o RaFD (LANGNER *et al.*, 2010) e o KDEF (LUNDQVIST; FLYKT; ÖHMAN, 1998). Os autores treinaram modelos da rede neural *VGG*, pré-treinada com o conjunto de imagens *VGG-Face* (PARKHI; VEDALDI; ZISSERMAN, 2015), em todos menos um dos *datasets* escolhidos, sendo que o *dataset* não utilizado para treino é o utilizado para teste. Para o conjunto de teste CK+, obtiveram o resultado de 88,58% de acurácia, enquanto que a acurácia para o JAFFE foi de 44,32% aproximadamente. O melhor resultado para o teste no JAFFE (com cruzamento de *datasets*) encontrado foi o de Ali, Iqbal e Choi (2016). Ali, Iqbal e Choi (2016) treinaram seu modelo, chamado *boosted Neural Network Ensemble* (ou *boosted NNE*), no conjunto RaFD, que possui 8040 imagens[1], e atingiram 48,67% de acurácia.

3.2 CLASSIFICAÇÃO DE EMOÇÃO EM FACES PARCIALMENTE OCLUSAS

Boa parte dos conjuntos de dados disponíveis foram fabricados em ambientes controlados e questões como variedade de gênero, raça e idade dos sujeitos; mudanças de iluminação; ou oclusões faciais são negligenciadas (ZHANG *et al.*, 2018b). Essa rigidez laboratorial prejudica o uso dos sistemas em ambientes e aplicações reais, que podem trazer benefícios imediatos à humanidade.

Conforme Zhang *et al.* (2018b) constatam, uma variedade das limitações citadas recebem atenção dos pesquisadores, entretanto poucos trabalhos foram eficientes o suficiente para sanar a questão das oclusões faciais. A presença de oclusões dificulta - e por vezes impossibilita - a extração de características fundamentais da face. Assim, o objetivo de um sistema AEF projetado com consciência da existência de oclusões é identificar corretamente emoções quando uma porção da face está escondida (ZHANG *et al.*, 2018b).

Dentre as abordagens ao lidar com problemas de oclusão, segundo apontado por Zhang *et al.* (2018b), estão aquelas baseadas em:

1. reconstrução das regiões oclusas da imagem através de métodos geométricos;

2. representações esparsas, encontrando imagens com e sem oclusão da mesma classe e as combinando linearmente;

[1] Blog oficial sobre o desenvolvimento do *dataset* RaFD: http://facedb.blogspot.com/

3. divisão em sub-regiões, utilizando apenas regiões não oclusas para classificação de emoção;

4. modelos estatísticos, que inferem as *features* das regiões oclusas;

5. dados em 3D que, com adição de elementos de profundidade, esperam criar modelos robustos à oclusão causada por mudanças de poses da cabeça (DRIRA *et al.*, 2013);

6. métodos baseados em *Deep Learning*, que não necessitam nem de extração manual de *features* nem da identificação da oclusão facial, e fazem uso dos *pixels* brutos das imagens.

3.2.1 Tipos de Oclusão

Towner e Slater (2007) definiram dois modos de oclusão facial: temporários e sistemáticos. As oclusões temporárias ocorrem quando uma região da face é obstruída devido à movimentação ou posicionamento da cabeça, que momentaneamente a coloca fora da região da câmera. Por outro lado, a oclusão sistemática surge devido à adição de acessórios (como máscaras e óculos) ou, mais naturalmente, devido a elementos faciais como cabelo e barba. Os autores apontam, ainda, que a segunda categoria de oclusão é a possivelmente mais devastadora para algoritmos de classificação por suas capacidades de oclusão total de elementos da face.

Embora existam conjuntos de dados que tragam imagens com oclusões naturais (MAHMOUD *et al.*, 2011; COLOMBO; CUSANO; SCHETTINI, 2011), a maior parte dos estudos faz uso de bancos clássicos como o Cohn-Kanade Extendido (CK+) (LUCEY *et al.*, 2010) e o JAFFE (LYONS *et al.*, 1998). Por esses dois conjuntos serem compostos de imagens sem quaisquer oclusões, as obstruções nas faces são inseridas computacionalmente. Alguns exemplos de oclusões já realizadas nesses conjuntos são expostos na Figura 3.1.

3.2.2 Investigação do efeito de oclusões na classificação de emoções faciais

A influência das oclusões são investigadas primariamente de forma quantitativa, através das acurácias de classificação dos algoritmos. Nesse sentido, são esperados efeitos determinados de acordo com cada modo de oclusão, como já condensado por Zhang *et al.* (2018b):

Figura 3.1 – Oclusões faciais fabricadas encontradas na literatura.

(a) Oclusões na região dos olhos.

(b) Oclusões na região da boca.

Fonte: Zhang *et al.* (2018b).

- Sentimentos de raiva, medo, tristeza e felicidade sofrem maior redução de acurácia com a oclusão da boca do que com a oclusão ocular;
- A oclusão ocular atinge a performance de classificação para o sentimento de nojo (mais para o conjunto CK+ do que para o JAFFE);
- A oclusão da região inferior da face reduz significativamente a performance de classificação para as seis emoções básicas e neutra;
- Raiva e felicidade são mais atingidos pela oclusão da região superior da face do que faces que expressam medo ou neutralidade;
- Não há grande perda de acurácia quando ocorre oclusão no nariz;
- Quando a oclusão é aleatória, a redução da acurácia global é elevada.

Devido a grande quantia de diferentes tipos de oclusões presentes na literatura, a comparação exata torna-se um desafio. Até o melhor conhecimento da autora, não foram encontrados trabalhos que utilizassem técnicas de *deep learning*, oclusões similares e os conjuntos de dados utilizados. Entretanto, algoritmos de aprendizagem de máquina usuais foram utilizados de forma semelhante. Cornejo e Pedrini (2018) utiliza filtros de Weber para tal, atingindo 93,13% de acurácia quando as imagens do conjunto CK+ estão limpas e 91,04% quando estão com algum tipo de oclusão facial, por exemplo. Trabalhos cujos modelos foram treinados com tarjas no CK+ e testados no JAFFE não foram encontrados, então a análise não se fez possível para esses casos.

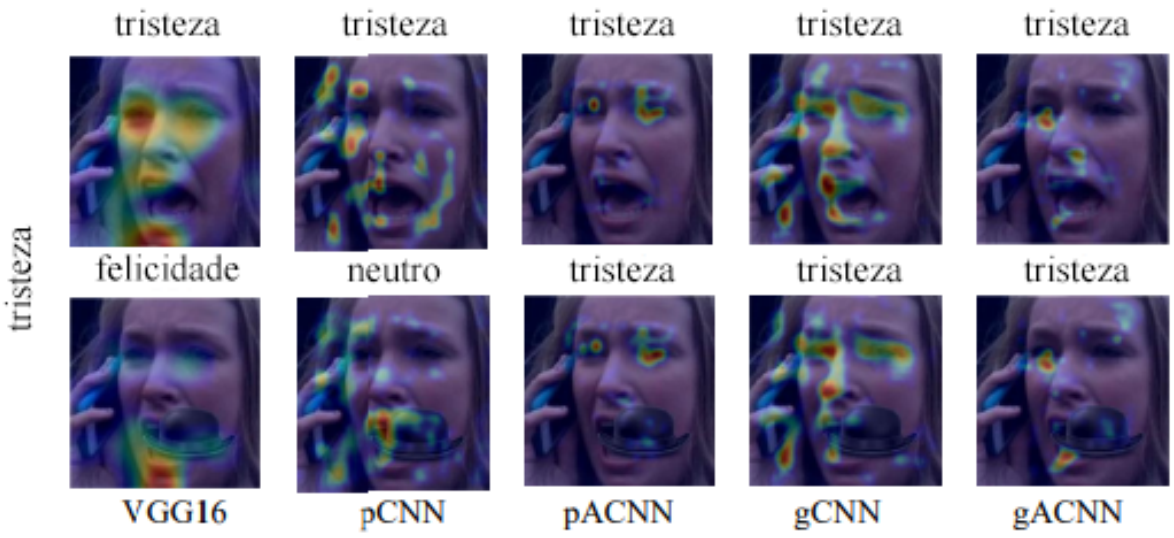
No tópico de interpretabilidade, algumas arquiteturas apresentam a adição de "máscaras" que atuam como mecanismos de atenção espacial e buscam a atenuação ou intensificação de regiões da imagem de entrada. Wang, Yuan e Feng (2019) combinam máscaras de atenção espacial com um mecanismo de auto-atenção que

“computa a correlação de regiões faciais locais em expressões faciais sutis” (WANG; YUAN; FENG, 2019). Similarmente, na tarefa de detecção facial, Wan e Chen (2017) utilizam máscaras para localizar obstruções e oclusões faciais.

Outro viés de análise seria de maneira qualitativa e visual, de forma a inspecionar o comportamento interno de um modelo de rede neural através de mapas de saliência. Tal análise objetiva a apresentação de uma explicação de qual região da imagem foi crucial para geração de uma classificação, conforme já mencionado anteriormente no texto. Essa abordagem é interessante para verificar se um modelo consegue distinguir locais da face que indicam determinada emoção, mesmo se a região com maior informação está oclusa. É uma abordagem que pode revelar a eficiência de mecanismos de atenção e máscaras, abundantemente utilizados (JUEFEI-XU *et al.*, 2016; ZHANG *et al.*, 2018a) para tarefas de classificação de elementos faciais, sendo esses emoções ou não. Ainda, a investigação da interpretabilidade de um modelo tem sua importância intensificada após a aprovação de leis como a *GDPR* da União Europeia.

No tema de classificação de emoções em faces, foram encontrados poucos trabalhos que realizaram diagnósticos semelhantes. Li *et al.* (2018) desenvolveram uma arquitetura que é capaz de detectar regiões oclusas da imagem com auxílio de redes de atenção que observam regiões específicas das imagens de entrada. Por fim, comparam os mapas de saliência gerados para entradas com e sem oclusão, conforme observado na Figura 3.2.

Figura 3.2 – Li *et al.* (2018) exibem o resultado de classificação e o mapa de saliências obtidos para diversas arquiteturas de redes neurais. Acima de cada imagem está a predição obtida. Na primeira linha, imagens sem oclusão e na segunda com adição de oclusão.



Fonte: Tradução livre da autora a partir de Li *et al.* (2018).

4 METODOLOGIA

Objetiva-se compreender como redes neurais entendem emoções faciais quando existem oclusões que impedem a extração de dados possivelmente relevantes para uma determinada emoção, situados principalmente em regiões oculares e mandibulares. Ainda, se modelos que apresentam mecanismos de atenção apresentam resultados superiores, tanto em termos quantitativos, quanto qualitativos (embora a análise qualitativa seja de alto teor subjetivo).

Para isso, serão comparados três modelos em dois conjuntos de dados distintos, com e sem oclusões faciais. O primeiro modelo não tem qualquer mecanismo de atenção; o segundo possui um mecanismo de atenção espacial; e o terceiro detém de um mecanismo de atenção em termos de profundidade e de espaço. Os três modelos serão treinados tanto em dados limpos, quanto em dados com oclusões faciais, totalizando seis modelos. A seguir os materiais e métodos utilizados nesse estudo serão melhor delimitados e aprofundados.

4.1 FERRAMENTAS

Todos os códigos foram realizados na linguagem de programação *Python*. Para manuseio, edição e adiação de tarjas nas imagens, conforme será apresentado, fez-se uso de bibliotecas como o *OpenCV* e o *PIL*.

Os algoritmos de inteligência artificial foram construídos com auxílio do *PyTorch* (PASZKE *et al.*, 2019), e os resultados expostos com a biblioteca *matplotlib*. Para observação do mapa de gradientes das imagens de entrada, modificaram-se algoritmos de *GradCAM* disponíveis na plataforma *GitHub* [1].

Realizou-se o treino dos modelos de redes neurais no *Google Colaboratory*[2], que dispõe de *GPUs* Tesla P100 e Tesla K80. A *GPU* Tesla P100 é equipada com 16GB de memória RAM, enquanto que a K80 possui 12GB. Os ambientes de desenvolvimento com uso de *GPUs* disponibilizados possuem 68GB de capacidade de armazenamento com sessões de até 12 horas de uso contínuo.

[1] O algoritmo de *GradCAM* utilizado neste trabalho foi construído a partir de exemplos encontrados em https://github.com/jacobgil/pytorch-grad-cam/blob/master/grad-cam.py e https://github.com/1Konny/gradcam_plus_plus-pytorch

[2] *Google Colaboratory* disponível em https://colab.research.google.com/

4.2 CONJUNTO DE DADOS DE TREINO

O *dataset* utilizado para os treinos é o Cohn Kanade Extendido (CK+) (LUCEY *et al.*, 2010). Esse conjunto de dados possui 593 sequências de imagens de 123 sujeitos, que partem da expressão neutra até o pico de uma emoção. Dessas sequências, apenas 327 estão rotuladas com emoções faciais e unidades de ação facial. Os rótulos incluem as seis emoções faciais básicas - felicidade, tristeza, raiva, surpresa, nojo e medo - além de uma não básica, o desprezo (ZHAO *et al.*, 2016). Ainda, o conjunto de dados fornece os 68 marcos faciais para cada um dos *frames* que o compõem.

Conforme Khorrami, Paine e Huang (2015), Liu *et al.* (2014) e Liu *et al.* (2013), para construção do conjunto de dados de treino, tomou-se a primeira imagem de cada sequência, rotulada como emoção neutra, e as três últimas imagens da sequência, que correspondem ao rótulo de uma das emoções citadas. Nesses trabalhos, os modelos foram treinados com as 1308 imagens resultantes, distribuídas em oito classes, e validados com validação cruzada de 10 *folds*. Na Tabela 4.1 observa-se a distribuição dos rótulos no conjunto de dados final. Na Figura 4.2 estão exemplos de imagens desse conjunto de dados e seus respectivos rótulos.

Tabela 4.1 – Distribuição das classes no conjunto de treino composto da primeira e das três últimas imagens de cada indivíduo presente no CK+.

Emoção	**Raiva**	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**	**Total**
Quantia	135	54	177	75	207	84	249	327	**1308**

Fonte: autora.

4.2.1 Oclusão facial

O *dataset* CK+ apresenta imagens sem quaisquer oclusões faciais. Por isso, para este estudo, oclusões nas regiões oculares e mandibulares foram fabricadas, de forma a representar, por exemplo, o uso de óculos solares ou máscaras nas faces dos indivíduos. O formato das oclusões fabricadas neste conjunto de dados foi realizada de forma similar por Kotsia, Buciu e Pitas (2008).

A partir do conjunto de 1308 imagens, a presença ou não de oclusões foi randomizada em terços. Na Tabela 4.2 observa-se a quantia de dados de cada classe para cada tipo de imagem: original, com oclusão ocular ou com oclusão mandibular.

As oclusões são particulares para cada indivíduo a partir dos 68 marcos faciais fornecidos no conjunto de dados. Na Figura 4.1 está um exemplo dos 68 marcos faciais e onde cada oclusão está localizada.

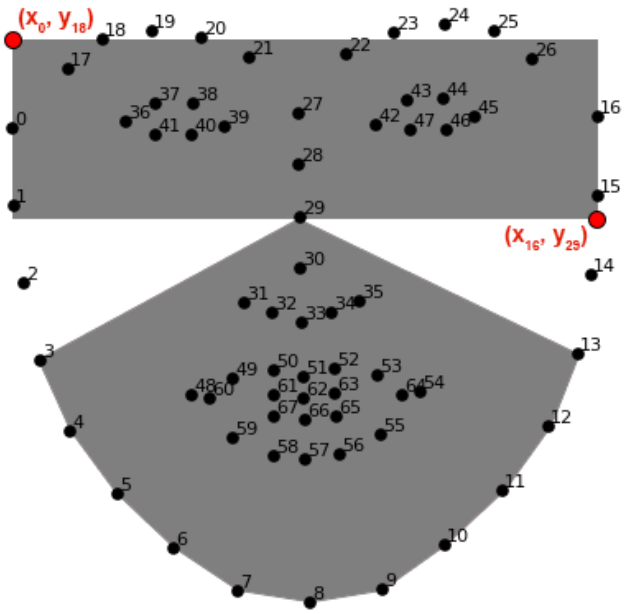
Tabela 4.2 – Quantia de amostras por método de oclusão para cada emoção do conjunto de dados. As amostras foram randomizadas em aproximadamente 33% para cada forma de oclusão e se mantém as mesmas durante toda a análise.

Emoção	Sem Oclusão	Oclusão Ocular	Oclusão Mandibular	Total
Raiva	56	41	38	135
Desprezo	17	18	19	54
Nojo	62	59	56	177
Medo	24	22	29	75
Felicidade	58	66	83	207
Tristeza	25	32	27	84
Surpresa	90	83	76	249
Neutro	110	110	107	327
Total	442	435	431	1308
Percentagem	33,26%	33,79%	32,95%	100%

Fonte: autora.

Para a oclusão ocular, um retângulo de cor sólida foi adicionado à imagem com seu canto superior esquerdo definido pela posição x do marco de índice 0 e posição y do marco de índice 18. O canto direito inferior tem as coordenadas x, y dos marcos 16 e 29, respectivamente. A oclusão mandibular, por sua vez, simula uma máscara que percorre os marcos de contorno facial 3 a 13, tem seu pico no marco 29 do nariz e retorna para o marco 3. Ainda, na Figura 4.2b apresentam-se exemplos das imagens do conjunto de dados fabricado.

Figura 4.1 – Localização, de acordo com os 68 marcos faciais, das oclusões fabricadas.



Fonte: Modificado pela autora a partir de https://medium.com/@suzana.svm/reconhecendo-landmarks-em-faces-com-dlib-python-7bfb094e1bb4.

Figura 4.2 – Amostras de ambos conjuntos CK+.

(a) Conjunto sem oclusões, original.

(b) Conjunto com amostragem de 33% de imagens sem oclusões, com oclusões oculares e oclusões mandibulares.

Fonte: Lucey *et al.* (2010) e autora.

Nesse documento estão apenas imagens cuja publicação foi autorizada pelos indivíduos. Não houve nenhum indivíduo que demonstrou a emoção de desprezo e

autorizou publicação, não sendo possível demonstrar exemplos dessa no trabalho. Entretanto, todas as imagens do conjunto de dados foram incluídas na análise e os resultados estatísticos não excluem imagens não publicáveis.

4.3 CONJUNTO DE DADOS DE TESTE

Além do CK+, para uma parcela dos testes realizados foi utilizado o *dataset* JAFFE (LYONS *et al.*, 1998). Esse conjunto de dados possui 213 imagens em tons de cinza de sete expressões faciais (as 6 básicas e a de neutralidade) de 60 mulheres japonesas.

4.3.1 Oclusão facial

Bem como para o conjunto CK+, um conjunto extra com oclusões faciais foi fabricado. Dessa vez, entretanto, na taxa de aproximadamente 50/50% de oclusões oculares e mandibulares, ou seja, sem incluir faces sem oclusões nesse conjunto. Na Tabela 4.3 estão as distribuições das classes e oclusões dentre as imagens. Ainda, na Figura 4.3 é possível observar amostras de imagens originais e oclusas para esse conjunto de dados.

Tabela 4.3 – Distribuição por classe das oclusões fabricadas no conjunto JAFFE.

Oclusão	**Raiva**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**	**%**
Ocular	14	15	13	16	18	14	19	51,17%
Mandibular	16	14	19	15	13	16	11	48,83%
Total	30	29	32	31	31	30	30	**213 - 100%**

Figura 4.3 – Amostras de ambos conjuntos JAFFE.

(a) Conjunto sem oclusões, original.

(b) Conjunto com amostragem de aproximadamente 50% de imagens com oclusões oculares e oclusões mandibulares.

Fonte: Lyons *et al.* (1998) e autora.

4.4 MECANISMOS DE ATENÇÃO

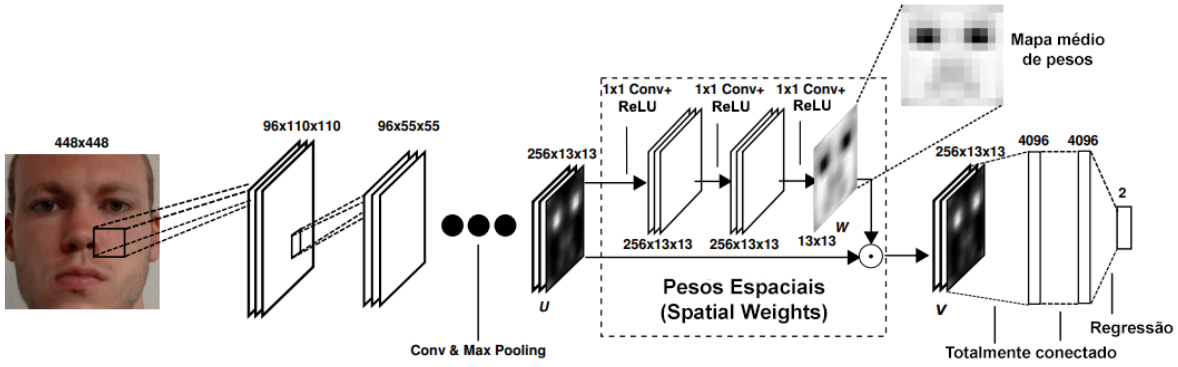
Este trabalho compara dois mecanismos de atenção aplicados a redes convolucionais. O método proposto por Zhang *et al.* (2017), *Spatial Weights Mechanism*, referenciado no texto por *SWM*, que apresenta um dos melhores resultados para a tarefa de estimação de direção do olhar pela aparência (onde as entradas da rede são imagens faciais); e o *Convolutional Block Attention Mechanism*, ou *CBAM* proposto por Woo *et al.* (2018) para tarefas gerais de classificação e respostas sobre imagens.

O primeiro método, escolhido devido ao seu sucesso com entradas semelhantes ao do presente projeto, é um mecanismo de atenção puramente espacial. Por outro lado, o segundo faz uso conjunto de atenção espacial e por profundidade, embora o método espacial desse seja diferente do primeiro. Ambos são integrados diretamente a arquiteturas já existentes de redes neurais, conforme será explicitado a seguir.

4.4.1 *Spatial Weights Mechanism* (SWM)

Este mecanismo de atenção espacial é composto por três camadas convolucionais com filtros de tamanho 1×1, seguidos de uma função de não linearidade *ReLU*, resultando em uma matriz de pesos espaciais. Sua implementação ocorre a partir da última camada convolucional da arquitetura convencional da rede neural, onde a matriz de pesos gerada é multiplicada por essa mesma camada. Assim, todas as camadas do último bloco convolucional original são multiplicadas pela mesma matriz de pesos, que corresponde diretamente à região facial da imagem de entrada, resultando em mapas de ativação ponderados (ZHANG *et al.*, 2017). Na Figura 4.4, observa-se a implementação do *SWM* na arquitetura escolhida por Zhang *et al.* (2017), uma AlexNet[3], para a estimação de direção do olhar.

Figura 4.4 – Arquitetura AlexNet com adição do mecanismo de atenção *SWM*.



Fonte: Tradução livre da autora a partir de Zhang *et al.* (2017).

Formalmente, considere um tensor de ativação U, de dimensões $N \times H \times W$ como saída da última camada convolucional da rede, onde N é o número de canais convolucionais e H e W são a altura e a largura desse bloco. O mecanismo de pesos espaciais (*SWM*) gera uma matriz de pesos espaciais $\boldsymbol{W}$ de dimensões $H \times W$ após a passagem *forward* de U por três camadas convolucionais com filtros 1×1, sendo que as duas primeiras geram 256 e a última apenas 1 canal. O mapa de ativação ponderado $\boldsymbol{W}$ é então multiplicado em termos de elementos com as ativações originais U (ZHANG *et al.*, 2017)

$$V_c = \boldsymbol{W} \odot U_c \tag{4.1}$$

onde U_c corresponde ao c-ésimo canal de U e V_c é o mapa de ativação ponderado do mesmo canal. Todos os mapas são então empilhados para formar o tensor de ativação V para entrada no próximo canal (ZHANG *et al.*, 2017).

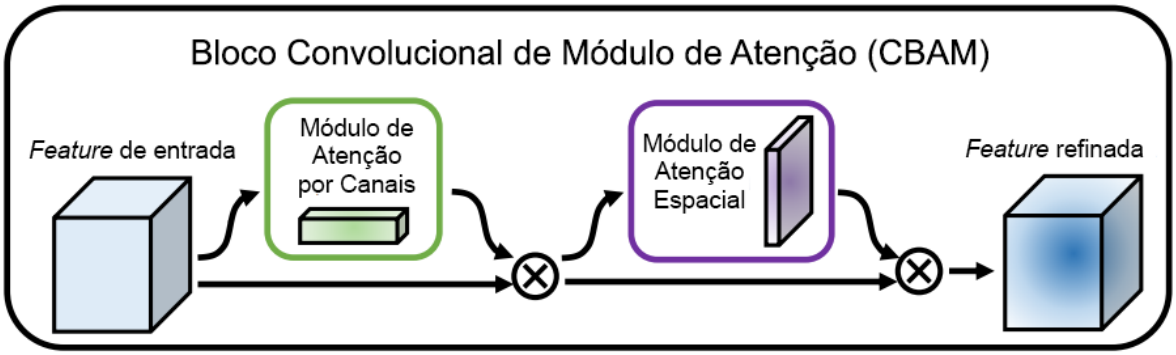
[3]Arquitetura de rede neural vencedora do desafio de classificação *ImageNet* de 2012, amplamente utilizada como parâmetro em estudos de visão computacional com aprendizagem profunda.

Para o treinamento dos modelos, as duas primeiras camadas convolucionais são iniciadas com pesos aleatórios da distribuição Gaussiana, com média 0 e 0,01 de variância, além de um *bias* constante de 0,1. A última camada convolucional é inicializada da mesma forma, porém com pesos de média 0 e variância 0,001, além do *bias* de 1 (ZHANG *et al.*, 2017).

4.4.2 Convolutional Block Attention Mechanism (CBAM)

Esse mecanismo, conforme já citado, possui elementos que garantem atenção tanto em termos de profundidade (nos canais das camadas convolucionais), quanto em termos de espaço, conforme observado na Figura 4.5. A separação dos módulos de atenção por canal e por espaço torna esse mecanismo não só eficiente, mas também de rápida conectividade a redes existentes (*plug-and-play*) (WOO *et al.*, 2018).

Figura 4.5 – Bloco convolucional de módulo de atenção, composto por módulos de atenção por canal e espacial.



Fonte: Tradução livre da autora a partir de Woo *et al.* (2018).

Ressalta-se dentre outros mecanismos por impactar significativamente na performance de redes residuais. A adição do mecanismo ao fim de todos os blocos residuais (repetidos ao longo da arquitetura *Resnet*) gera um aumento de 2% de acurácia na classificação do banco de dados *ImageNet*[4] por uma rede *Resnet50* (WOO *et al.*, 2018). Entretanto, para fins comparativos com o método anterior, *SWM*, o *CBAM* também será adicionado apenas após a última camada convolucional da rede.

[4]Conjunto de dados ImageNet disponível em http://www.image-net.org/

4.4.2.1 *Módulo de Atenção por Canais*

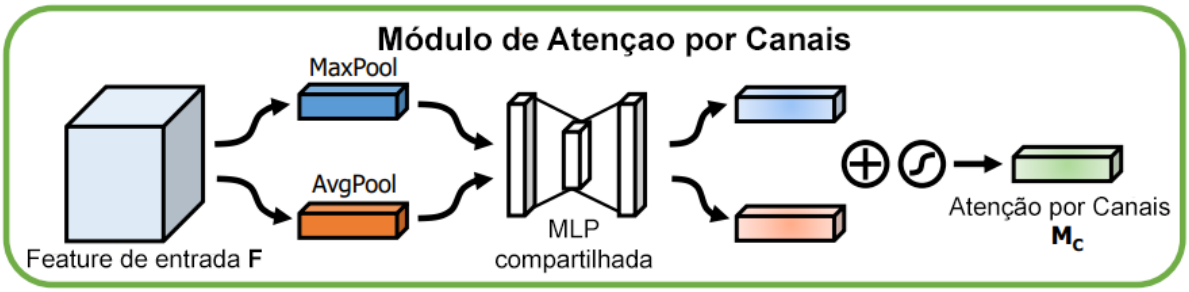
O módulo de atenção por canais busca intensificar “o que” é importante na imagem de entrada (WOO *et al.*, 2018), visto que os canais convolucionais podem ser interpretados como detectores de *features* (ZEILER; FERGUS, 2014). Portanto, de forma a agregar o máximo de informações possíveis dentre os canais convolucionais, esse módulo faz uso tanto de *Average Pooling*, para computação de estatísticas espaciais (HU; SHEN; SUN, 2018; WOO *et al.*, 2018), quanto de *Max Pooling*, para realce de características distintivas da imagem (WOO *et al.*, 2018).

As *features* captadas pelos dois métodos de agregação são então enviadas individualmente a uma rede compartilhada de camada escondida única (um *perceptron* multi-camadas, ou *MLP*). Para extração do mapa de atenção por canais, os dois resultados são somados elemento a elemento e então passam pela função de ativação *Sigmoid*. Na Equação 4.2, estão descritas essas operações, onde M_c é a saída do bloco de atenção por canais, σ representa a função *Sigmoid*, e F a camada convolucional de entrada (WOO *et al.*, 2018). A Figura 4.6 apresenta as mesmas operações. Em seguida, o mapa de *features* resultante, M_c, é multiplicado elemento a elemento com a camada convolucional original para gerar F', conforme a Equação 4.3.

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{4.2}$$

$$F' = F \odot M_c(F) \tag{4.3}$$

Figura 4.6 – Módulo de atenção por canais.



Fonte: Tradução livre da autora a partir de Woo *et al.* (2018).

4.4.2.2 *Módulo de Atenção por Espaço*

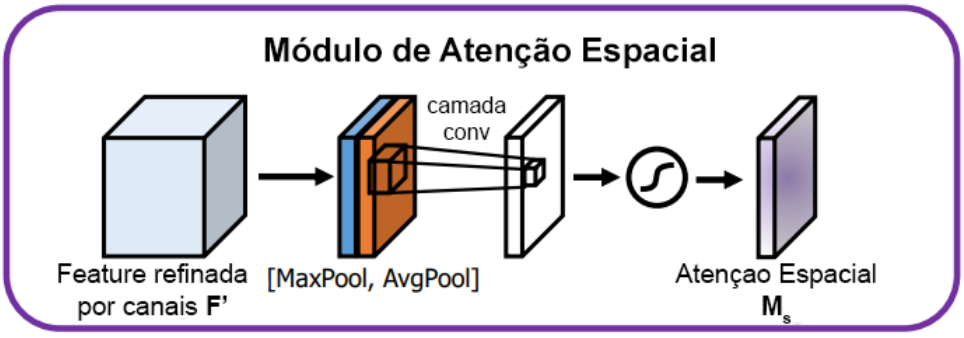
Por outro lado, o módulo de atenção por espaço busca a representação de “onde” está a informação relevante para classificação na imagem de entrada, de forma

a complementar o que já foi obtido no módulo anterior (WOO *et al.*, 2018). Aqui, as operações de *Average Pooling* e *Max Pooling* também são utilizadas, concatenadas, submetidas à convolução por um filtro convolucional de tamanho 7×7 e, finalmente, o resultado final F'' é obtido após à computação da função *Sigmoid* σ, como na Figura 4.7. As operações são visualizadas na Equação 4.4, onde F' é a saída do bloco de atenção por canais. O resultado, F'', é multiplicado elemento a elemento com F' e retornado ao fluxo principal da rede, conforme a Equação 4.5 (WOO *et al.*, 2018).

$$M_s(F') = \sigma(f^{7x7}([AvgPool(F'), MaxPool(F')])) \tag{4.4}$$

$$F'' = F' \odot M_s(F') \tag{4.5}$$

Figura 4.7 – Módulo de atenção espacial.

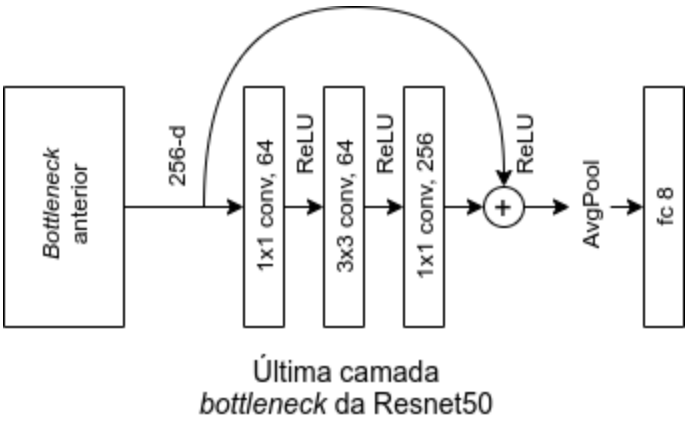


Fonte: Tradução livre da autora a partir de Woo *et al.* (2018).

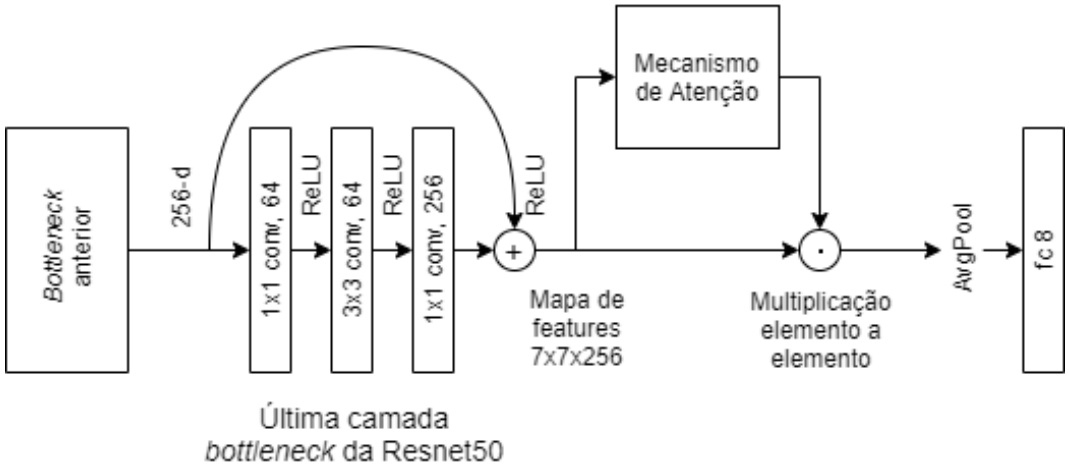
4.5 ARQUITETURAS COMPARADAS

Neste trabalho, utiliza-se como arquitetura base uma *Resnet* de 50 camadas (HE *et al.*, 2016), por ser uma construção eficiente e de poder computacional alto, conforme já comentado no capítulo anterior. Além da arquitetura original, dois outros modelos são construídos, um com o módulo de atenção espacial *SWM* e outro com o módulo *CBAM*. Ambos blocos de atenção são adicionados ao fim da última camada base (*bottleneck layer*, nome dado à camada que se repete ao longo da arquitetura *Resnet*) do modelo residual. Na Figura 4.8 estão os diagramas de saída da arquitetura original e da que possui algum dos mecanismos de atenção. Ainda, no Apêndice A estão os hiperparâmetros utilizados nos treinos de cada um dos modelos.

Figura 4.8 – Diagrama das terminações das arquiteturas comparadas.



(a) *Resnet50* original, sem mecanismo de atenção.



(b) *Resnet50* com atenção. O mecanismo de atenção pode ser tanto o *SWM*, quanto o *CBAM*.

Fonte: autora.

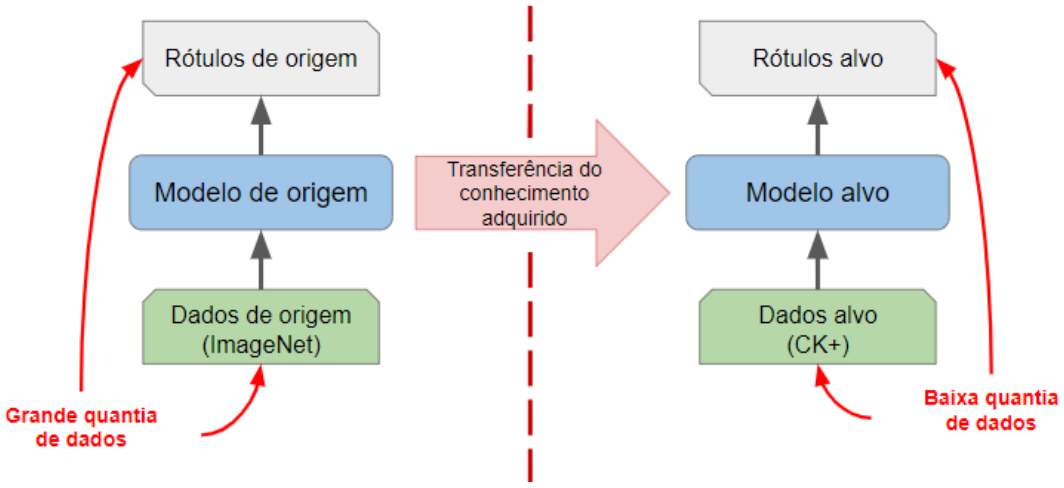
4.5.1 Pré-treinamento e inicialização dos pesos

De acordo com o apresentado por Razavian *et al.* (2014) e ressaltado por Gilpin *et al.* (2018), camadas de redes neurais treinadas em uma tarefa classificatória x podem ser utilizadas para facilitar uma segunda tarefa y, de domínio diferente de x. Esse método de aprendizagem é especialmente útil quando não possuímos uma quantia elevada de dados para treino e queremos fazer uso das características já extraídas a partir de dados massivos[5]. Na Figura 4.9 está o diagrama de inicialização

[5]https://medium.com/the-official-integrate-ai-blog/transfer-learning-explained-7d275c1e34e2

de pesos utilizado no presente projeto. As redes *Resnet50* foram pré-treinadas com o *dataset ImageNet* para a classificação de imagens dentre 1000 classes. Portanto, substituiu-se a última camada linear de 1000 neurônios por uma de oito para predição das oito classes de sentimentos do conjunto CK+.

Figura 4.9 – Funcionamento do *transfer learning* neste projeto.



Fonte: Modificado pela autora a partir de https://miro.medium.com/max/2626/1*Z11P-CjNYWBofEbmGQrptA.png.

4.6 PROCESSO DE TREINO E VALIDAÇÃO

A seguir, são explicitados os procedimentos de treino para as arquiteturas analisadas. Todos os modelos foram treinados na plataforma *Google Colaboratory*, conforme mencionado anteriormente.

4.6.1 Pré-processamento das imagens

O pré-processamento das imagens constitui-se de três etapas: (1) redução das imagens para a região de interesse - as faces; (2) redimensionamento para o formato de $224 \times 224 \times 3$; e (3) normalização e uniformização dos dados para média zero e variância unitária.

Captura-se a região de interesse de cada imagem individualmente, a partir dos marcos faciais presentes no conjunto de dados. Identifica-se quais os marcos de coordenadas mais exteriores na imagem em todos os lados (em cima, embaixo, na esquerda e na direita), adiciona-se uma pequena margem para cada coordenada e, então, realiza-se a exclusão do que não está dentro da região.

As duas operações seguintes, redimensionamento e normalização, são realizadas com auxílio da biblioteca de aprendizado profundo utilizada, *PyTorch*. Fornece-se, então, o tamanho desejado da imagem, 224 *pixels*, e os valores de média e desvio padrão para cada uma das camadas de cor de todas as 1308 imagens. Na Equação 4.6, está disposto o processo de centralização, ou uniformização, dos dados de um canal I_c da imagem I, pertencente ao conjunto de dados d, que possui média do canal c igual a μ_d e desvio padrão $\boldsymbol{\sigma_d}$. Tal operação é realizada individualmente para os três canais das imagens.

$$I_c^u = \frac{I_c - \mu_d}{\boldsymbol{\sigma_d}} \tag{4.6}$$

Salienta-se que, como os dados do conjunto CK+ apresentam tanto imagens coloridas quanto imagens em escala de cinza, todas foram convertidas para escala de cinza com auxílio da biblioteca *PIL*. Essa biblioteca aplica a transformação ITU-R 601-2[6] da Equação 4.7, sendo R, G e B as matrizes dos três canais de cores vermelho, verde e azul, respectivamente, e I a imagem resultante.

$$I = R \odot \frac{299}{1000} + G \odot \frac{587}{1000} + B \odot \frac{114}{1000} \tag{4.7}$$

Assim, as imagens de entrada da rede estão nas dimensões $224 \times 224 \times 3$, sendo que todas as camadas de cor são iguais e correspondem à versão em escala de cinza dessas. Faz-se uso das três camadas iguais para que seja possível inserir a imagem diretamente na rede *Resnet* pré-treinada, sem necessidade de troca de sua camada de entrada. Na Tabela 4.4 estão os valores de média e desvio padrão utilizados na normalização para os dois conjuntos de treino: original e com oclusão.

Tabela 4.4 – Média e desvio padrão das imagens em tons de cinza dos conjuntos de treino com e sem oclusão.

	Média (μ)	**Desvio Padrão (σ)**
Original (Sem oclusão)	0,5293	0,2763
Com oclusão	0,4842	0,2694

Fonte: autora.

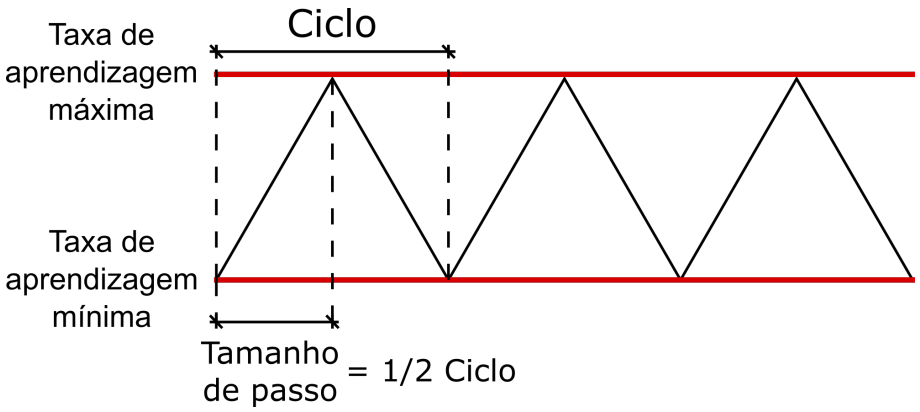
4.6.2 Taxas de Aprendizagem Cíclicas

De acordo com o comentado na subseção de otimização do capítulo anterior, alguns métodos de otimização demandam a redução gradual da taxa de aprendiza-

[6] https://pillow.readthedocs.io/en/latest/reference/Image.html#PIL.Image.Image.convert

gem (ou, em inglês, *learning rate*) para atingirem a convergência com maior facilidade. De encontro a esse padrão, Smith (2017) propôs o método de Taxas de Aprendizagem Cíclicas, e, ao invés do contínuo decrescimento do valor dessa variável, ela é livre para transitar linearmente, na forma de uma onda triangular, com intervalos de variação definidos pelo usuário (em x, o tamanho de passo, e, em y, as taxas de aprendizagem máxima e mínima), conforme a Figura 4.10 (SMITH, 2017).

Figura 4.10 – Variação da taxa de aprendizagem de acordo com o método proposto por Smith (2017).



Fonte: autora.

Para a definição dos intervalos ótimos de taxas de aprendizagem, Smith (2017) também apresentou um método de busca:

1. Define-se um intervalo de variação de taxa de aprendizagem extenso, por exemplo, taxas de aprendizagem mínima e máxima de valores, respectivamente, 10^{-7} e 10^{2};

2. O tamanho de passo de variação é definido de acordo com o número de iterações por época, que, por sua vez, é computado pela divisão do número total de amostras de treino pelo tamanho do *mini-batch*. Recomenda-se, então, definir o passo de 2 a 10 vezes o valor de iterações calculado;

3. O modelo é exposto a algumas épocas de treino, enquanto a taxa de aprendizagem é modificada linearmente de acordo com a onda triangular da Figura 4.10;

4. Ao fim do processo, gera-se um gráfico de valor de perda versus acurácia de cada época de treino;

5. Identifica-se a curva mais íngreme e decrescente desse gráfico, onde os valores mínimo e máximo de taxa de aprendizagem são, respectivamente, os pontos de início e de fim de decrescimento;

6. Uma regra aproximada, indicada pelo autor, é a definição do valor mínimo de taxa de aprendizagem como um terço ou um quarto do valor máximo escolhido (SMITH, 2017), método adotado neste trabalho.

De acordo com Smith (2017), o uso de taxas de aprendizagem cíclicas permite que, durante a otimização de um modelo, este consiga sair de possíveis pontos de sela ou mínimos locais, visto que o passo de atualização de gradientes, definido pela variável manipulada, pode ser menor ou maior. Ainda, constata-se que o tempo de treino quando faz-se uso da política de taxas de aprendizagem cíclicas é inferior para valores iguais ou superiores de acurácia quando comparado a métodos usuais. Tal resultado é apresentado por Smith (2017) para o conjunto de dados de classificação CIFAR-10[7], com o qual o autor atinge 81,4% de acurácia em aproximadamente 25.000 iterações, ao invés das aproximadas 70.000 iterações de métodos usuais.

Durante a etapa de treino do presente trabalho, as redes neurais treinadas em imagens com tarjas faciais encontraram dificuldades de convergência quando utilizou-se taxas de aprendizagem continuamente decrescentes. Nesse cenário, o uso de taxas de aprendizagem cíclicas intensificou os resultados obtidos, bem como aumentou a velocidade para o alcance de boas métricas.

4.7 MAPA DE SALIÊNCIAS GERADO POR GRAD-CAM

O método de visualização de mapas de saliências por *Grad-CAM* não necessita de quaisquer modificações nos modelos de redes neurais. A importância de cada neurônio da rede para uma determinada classificação é obtida pela informação de gradiente que flui desde a última camada convolucional durante a etapa de *backpropagation* (SELVARAJU *et al.*, 2017).

De acordo com Selvaraju *et al.* (2017), para obtermos o mapa de localização discriminativo de classes, $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$, onde u e v correspondem à altura e largura do mapa para uma classe c, primeiro calcula-se, anteriormente ao *Softmax*, o gradiente do resultado y^c para a classe c, com respeito aos mapas de *features* A^k de uma camada convolucional. Toma-se a agregação média (*Average Pool*) desses gradientes que fluem reversamente na rede para obtermos a importância de cada neurônio α^c_k, conforme a Equação 4.8. O coeficiente α^c_k captura, portanto, a importância do

[7]Conjunto de dados para classificação de imagens entre 10 classes https://www.cs.toronto.edu/ kriz/cifar.html

mapa de *features* k para a classificação da classe c (SELVARAJU *et al.*, 2017).

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{Agregação Média}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{Gradiente}} \tag{4.8}$$

A seguir, computa-se a combinação ponderada de todos os k mapas de ativação do sentido direto da rede (*forward activation maps*) junto a uma função *ReLU*, como observado na Equação 4.9. Com o uso da função *ReLU*, objetiva-se a obtenção apenas dos neurônios que representaram uma influência positiva no resultado previsto pela rede, que correspondem aos *pixels* de maior importância para a classificação y^c, aqueles que devemos aumentar para também aumentar y^c (SELVARAJU *et al.*, 2017).

$$L_{Grad-CAM}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{Combinação Linear}} \tag{4.9}$$

O mapa de gradientes, $L_{Grad-CAM}^c$, tem altura e largura iguais à camada convolucional à que foi aplicada, sendo de 7×7 para a última camada convolucional de uma *Resnet50*. Por fim, para visualização, redimensiona-se $L_{Grad-CAM}^c$ para o tamanho da imagem de entrada e sobrepõe-se a primeira à segunda.

5 RESULTADOS

No processo de criação de um modelo de rede neural, três etapas são cruciais para legitimar sua eficácia: o treino, a validação e o teste, sendo que tanto os dados de validação quanto os de teste não podem ser vistos pela rede durante a etapa de treino. A etapa de treino e validação utilizam, geralmente, dados com distribuições semelhantes, e as métricas de validação são as bases para definição dos hiperparâmetros escolhidos para o treino.

Entretanto, como os hiperparâmetros de treino são ajustados de acordo com os dados de validação, há a incidência de um *bias* por parte do programador, que ajusta-os para atingir a performance esperada e selecionar o modelo desejado. Por isso, os dados de validação não são mais "intocados" e desconhecidos totalmente. Para sanar a dúvida da integridade dos modelos finais, realizam-se testes para confirmação da performance em dados que a rede com certeza nunca viu nem foi ajustada para classificar corretamente.

Sendo assim, os resultados estão divididos em duas seções primárias: validação e teste, esse último que subdivide-se em análises quantitativas e qualitativas. A validação é em relação a dados similares aos presentes no treino de cada modelo, do conjunto CK+, enquanto que os testes foram realizados, principalmente, para o conjunto JAFFE. A seguir, as métricas utilizadas para a seleção de modelos e explicitação dos resultados são expostas.

5.1 MÉTRICAS

Ao longo do processo de treino, é essencial observar a evolução dos modelos. Nesta seção são apresentadas métricas para avaliação da performance de modelos de classificação não-binária.

5.1.1 Acurácia e Pontuação F1

A acurácia é a pontuação atribuída à comparação direta entre resultado previsto pelo modelo e rótulo esperado para uma determinada entrada. Sendo assim, quão maior for a taxa de acerto entre predição e rótulo, maior é a acurácia[1]. Por outro lado, a pontuação F1 é a média harmônica de dados chamados de precisão e de *recall*[2],

[1] Métricas de avaliação de modelos classificatórios https://bit.ly/2OzukB3
[2] Pontuação F1 https://deepai.org/machine-learning-glossary-and-terms/f-score

conforme as Equações 5.1, 5.2 e 5.3.

A precisão mede a quantia de predições verdadeiras e positivas (VP) dentre o universo de predições positivas (quando é apontado pelo modelo que há ocorrência da classe, sendo o resultado verdadeiro ou falso). *Recall*, ou sensibilidade, é a quantia de predições verdadeiras e positivas dividida pelo conjunto de predições esperadas positivas (verdadeiros positivos, quando é previsto resultado 1 e esperado 1 ou falsos negativos (FN), quando é previsto resultado 0 e esperado 1). A Tabela 5.1 esclarece os conceitos utilizados[3].

$$F1 = \frac{2 \cdot precisão \cdot recall}{precisão + recall} \tag{5.1}$$

$$Precisão = \frac{VP}{VP + FP} \tag{5.2}$$

$$Recall = \frac{VP}{VP + FN} \tag{5.3}$$

Tabela 5.1 – Matriz de confusão binária.

		Previsto	
		Negativo	**Positivo**
Esperado	**Negativo**	Verdadeiro Negativo (VN)	Falso Positivo (FP)
	Positivo	Falso Negativo (FN)	Verdadeiro Positivo (VP)

5.1.2 Matriz de Confusão

A matriz de confusão é outro método de visualização do desempenho de modelos, conforme exemplificado pela Tabela 5.1, onde cada linha representa as instâncias de rótulos reais e cada coluna representa as instâncias de predições de rótulos. Sendo assim, os valores expostos na diagonal principal da matriz correspondem às classificações corretas realizadas pelo modelo. As instâncias não presentes na diagonal principal são os erros, ou confusões, das predições geradas pelo modelo.

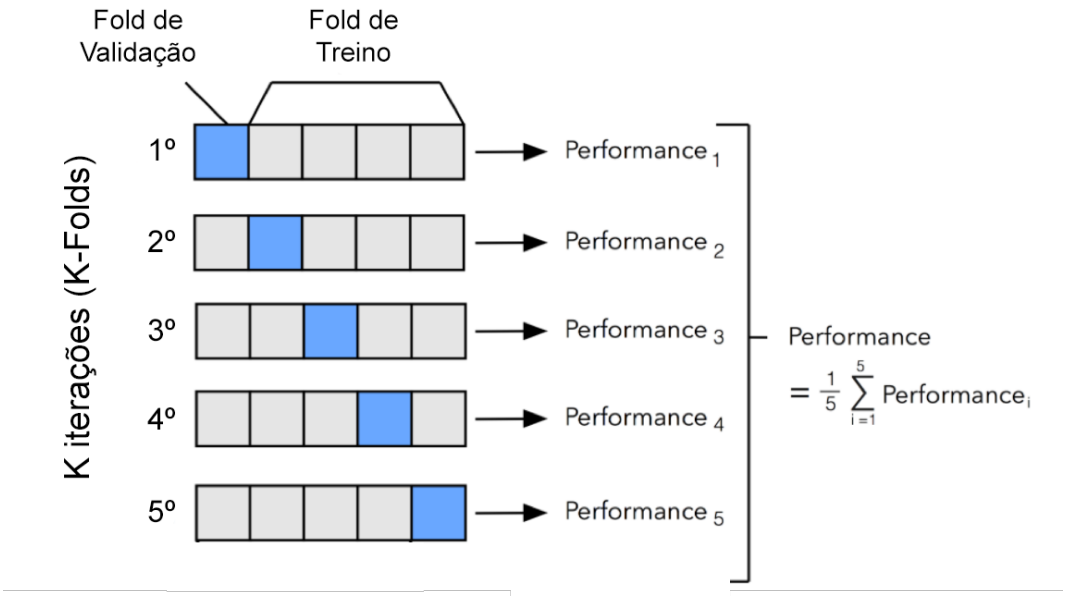
5.1.3 Validação Cruzada

Uma métrica utilizada para a seleção de modelos quando se possui poucos dados é a validação cruzada. Na validação cruzada, os dados de treino são divididos

[3] https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

em K amostras (ou *folds*), a seguir seleciona-se a amostra $k \in 1, .., K$ de teste e treinamos o modelo em todas as amostras menos k, até que $k = K$ (MURPHY, 2012), de acordo com a Figura 5.1. Salienta-se que embora a Figura 5.1 apresente $K = 5$, no presente projeto fez-se uso de $K = 10$ de acordo com a literatura, conforme mencionado no capítulo de metodologia.

Figura 5.1 – Processo de validação cruzada por K-Folds. Nesse caso, K = 5.



Fonte: Tradução livre da autora a partir de http://ethen8181.github.io/machine-learning/model_selection/model_selection.html.

Após a validação nos K-Folds, anotando as métricas relevantes, como acurácia e pontuação F1, tem-se a estimativa da performance do modelo para todo o conjunto de dados. Sendo assim, ao fim do processo de validação cruzada, o modelo final é retreinado, dessa vez com todo o conjunto de dados disponível.

5.2 RESULTADOS DE VALIDAÇÃO

Três modelos de rede neural, cuja diferença está apenas no mecanismo de atenção presente (ou na falta dele) foram treinados tanto para dados sem, quanto para dados com oclusões. Assim, tem-se um total de seis modelos: três treinados em dados sem oclusões e três treinados em dados com oclusões fabricadas.

Nessa subseção encontram-se, além dos dados de validação cruzada dos modelos treinados com a totalidade dos dados, também os resultados de validação para modelos treinados com 80% dos dados disponíveis (e, consequentemente, validados em 20% desses). Os modelos resultantes da validação cruzada são utilizados nos

testes com o conjunto de dados JAFFE, enquanto que utilizam-se os treinados com 80% dos dados para teste quando invertemos os conjuntos de validação: modelos treinados sem oclusões expostos a faces oclusas e o contrário, para compreensão de como a adição de dados com oclusões impacta nos resultados de cada mecanismo de atenção. As matrizes de confusão para ambos os modos de validação estão dispostas no Apêndice B.

5.2.1 Validação cruzada

Os modelos treinados e metrificados por validação cruzada desta seção são utilizados para a etapa de testes com o conjunto de dados JAFFE. Espera-se que o treino com todas as imagens conceda a melhor performance possível para a distribuição de dados disponível. Na Tabela 5.2 estão os resultados de validação cruzada de 10-*folds* para os três modelos (sem atenção, atenção *SWM* e atenção *CBAM*) treinados em dados do conjunto CK+ com ou sem oclusões.

Tabela 5.2 – Valores de acurácia e pontuação F1 obtidos na validação cruzada em 10-*folds* para modelos treinados em dados com e sem oclusões.

Métricas	**Treino sem oclusões**			**Treino com oclusões**		
	Sem atenção	**CBAM**	**SWM**	**Sem atenção**	**CBAM**	**SWM**
Acurácia	97,83%	97,33%	97,27%	91,75%	91,3%	88,47%
Pontuação F1	97,40%	96,44%	96,71%	89,65%	88,78%	85,11%
Épocas	27	29	26	39	39	42

Conforme esperava-se, o conjunto CK+ com oclusões teve sua convergência dificultada quando comparada ao treino sem oclusões. Por isso, estendeu-se o treinamento dos modelos por uma quantia maior de épocas, além de ser necessário utilizar o método de Taxas de Aprendizagem Cíclicas para obtenção dos resultados expostos, ao invés de taxas de aprendizagem continuamente decrescentes, como realizado no treino sem oclusões.

Observa-se uma queda de aproximadamente 7% nas duas métricas, acurácia e pontuação F1, entre os modelos correspondentes treinados em imagens com e sem oclusões.

5.2.2 Treino e validação usuais

Para inferência do impacto do treino com dados oclusos, será apresentado futuramente no texto o teste com a troca do conjunto de validação entre os modelos.

Assim, modelos treinados em imagens sem oclusões serão testados nas mesmas imagens utilizadas em suas validações, mas dessa vez, com presença de oclusões, sendo realizado também o inverso.

Para isso, outros seis modelos foram treinados com separação entre dados de treino e validação na relação de 80/20%, respectivamente. Na Tabela 5.3 estão as métricas obtidas para validação dos modelos treinados com 80% dos dados. Salienta-se que as imagens oclusas mantém-se as mesmas da subseção anterior, de validação cruzada, inclusive em relação ao tipo de oclusão (ocular ou mandibular).

Tabela 5.3 – Valores de acurácia e pontuação F1 obtidos no conjunto de validação para as duas modalidades de treino: dados originais sem oclusões e dados com 33% de oclusões oculares e mandibulares.

Métricas	**Treino sem oclusões**			**Treino com oclusões**		
	Sem atenção	**CBAM**	**SWM**	**Sem atenção**	**CBAM**	**SWM**
Acurácia	96,18%	96,56%	97,77%	87,02%	90,01%	87,02%
Pontuação F1	95,30%	95,79%	96,77%	81,55%	86,61%	84,49%
Épocas	28	24	21	39	34	28

Atenta-se para as quedas de 4% e 8% de acurácia e pontuação F1 para o modelo sem atenção treinado com oclusões, quando comparado aos resultados da subseção anterior, de validação cruzada. Esse decréscimo, acentuado quando se faz a mesma comparação para os modelos com mecanismos de atenção, é um possível indicativo do impacto dos mecanismos de atenção quando há uma redução no conjunto de dados de treino.

5.3 ANÁLISE QUANTITATIVA

5.3.1 Inversão dos dados de validações do conjunto CK+

Com o intuito de se observar o impacto da presença de imagens com oclusão no conjunto de treino, na Tabela 5.4 estão os valores de acurácia e pontuação F1 ao inverterem-se os conjuntos de validação dos modelos treinados com e sem oclusões. No Apêndice C estão as matrizes de confusão obtidas no teste de inversão dos conjuntos de validação. Conforme o esperado, houve uma queda acentuada, de aproximadamente 16%, nas métricas dos modelos treinados sem oclusões quando expostos a dados com oclusões. Pelo contrário, as métricas dos modelos treinados em dados com oclusões obtiveram um aumento de cerca de 6% na validação com as imagens originais.

Tabela 5.4 – Valores de acurácia e pontuação F1 obtidos ao inverter os conjuntos de validação entre os modelos treinados com e sem oclusão.

	Modelos sem oclusão em dados com oclusão			**Modelos com oclusão em dados sem oclusão**		
Métricas	**Sem atenção**	**CBAM**	**SWM**	**Sem atenção**	**CBAM**	**SWM**
Acurácia	80,15%	81,68%	81,68%	94,66%	95,42%	94,28%
Pontuação F1	77,55%	79,52%	79,62%	90,85%	93,17%	91,91%

Tais resultados comprovam a dificuldade de classificação de sentimentos em imagens com oclusões e expõe que, para os modelos treinados com oclusão, mesmo reduzindo-se as amostras de faces limpas (para aproximadamente 33%, devido a presença de 66% de faces oclusas) a alta taxa de acertos é mantida, sendo apenas de 1 a 3% inferior quando comparada ao modelo correspondente treinado com os dados originais, sem oclusões.

Ao considerar as métricas obtidas a partir do conjunto de treino com oclusões, separadas por tipo de oclusão, encontra-se um efeito inusitado. De acordo com a Tabela 5.5, quando comparam-se ambos os modelos sem quaisquer mecanismos de atenção, o treino a partir de dados com oclusões gerou um resultado levemente inferior para imagens com oclusão ocular. Isso indica que, quando não se utiliza mecanismos de atenção na arquitetura estudada, *Resnet50*, não há melhora significativa na classificação de emoções em faces com oclusão ocular.

Ainda, confirmam-se os resultados esperados de acordo com a literatura que antecipavam a maior influência negativa de oclusões na região da boca em detrimento de oclusões nas regiões dos olhos. Entretanto, o treino com oclusões presentes diminui a diferença entre os valores de acurácia para oclusões oculares e mandibulares de, em média, 20,31% para apenas 9,32% dentre todos os mecanismos de atenção. Por fim, constata-se que a adição de mecanismos de atenção, como o *CBAM* ou o *SWM*, aprimora as acurácias e pontuação F1 para oclusões oculares em, respectivamente, 8 e 6% em comparação ao modelo *Resnet50* puro quando o conjunto de treino apresenta oclusões. Entretanto, a adição desses mecanismos não mostra aumento tão significativo para oclusões mandibulares, na verdade apresentando redução para o caso da atenção *SWM*.

5.3.2 Testes com o conjunto de dados JAFFE

Os resultados na subseção anterior mostraram a performance esperada da atuação dos modelos comparados quando o conjunto de teste/validação é semelhante ao de treino. A seguir, apresentam-se os resultados de testes realizados em um conjunto

Tabela 5.5 – Valores de acurácia e pontuação F1 para modelos treinados com e sem oclusão no teste de cruzamento dos conjuntos de validação.

Dados de treino	Atenção	Oclusão	Métricas	
			Acurácia	Pontuação F1
Sem oclusão	Sem atenção	Ocular	85,71%	83,72%
		Mandibular	60,22%	53,90%
	CBAM	Ocular	81,82%	78,32%
		Mandibular	59,14%	50,22%
	SWM	Ocular	80,52%	77,84%
		Mandibular	67,74%	56,79%
Com oclusão	Sem atenção	Ocular	84,42%	83,09%
		Mandibular	80,65%	77,02%
	CBAM	Ocular	92,21%	86,75%
		Mandibular	82,80%	75,20%
	SWM	Ocular	92,21%	89,65%
		Mandibular	77,42%	72,00%

de imagens totalmente diferente dos de treino: o JAFFE.

Semelhantemente ao conjunto CK+, executaram-se testes em dois conjuntos JAFFE: um original, sem oclusões, e outro com distribuição de 50% de oclusões oculares e mandibulares, cujo arranjo entre classes foi exposta na Tabela 4.3 do capítulo anterior. Na Tabela 5.6 estão os resultados obtidos para as três arquiteturas treinadas no conjunto CK+ com e sem oclusões. No Apêndice D estão as matrizes de confusão obtidas para os testes nesse conjunto de dados.

Relembra-se que os modelos testados nessa subseção foram treinados com a totalidade dos dados do conjunto CK+. Assim, atribui-se parcela da queda de performances obtidas às diferenças étnico-raciais dos conjuntos de imagens (CK+ é constituído de 69% de Euro-Americanos, 13% Afro-Americanos e 6% de outras etnias e raças (LUCEY *et al.*, 2010), enquanto o JAFFE é composto exclusivamente por faces de mulheres asiáticas).

Da Tabela 5.6, quando observam-se os testes realizados em dados sem oclusão, a arquitetura de melhor acurácia para os dois modos de treino (imagens originais e imagens com oclusões) foi a de atenção *CBAM*. Esta foi única arquitetura a apresentar resultados melhores nesse teste quando treinada com dados com oclusão facial.

Em segunda instância, ao analisar os testes realizados em imagens com tarjas fabricadas, constata-se a ineficiência da adição de mecanismos de atenção quando os modelos são treinados apenas com dados limpos e originais. Nesse teste, a arquitetura sem atenção foi a única a não apresentar queda tão acentuada nas métricas de acurácia e pontuação F1. Por outro lado, ao observarmos os modelos treinados a

Tabela 5.6 – Valores de acurácia e pontuação F1 obtidos para os modelos treinados com a totalidade do conjunto CK+ com e sem oclusões faciais. Os testes são realizados no conjunto JAFFE com e sem oclusões.

Métricas	**Treino com conjunto CK+ sem oclusões**			**Treino com conjunto CK+ com oclusões**		
	Testes no conjunto JAFFE sem oclusão					
	Sem atenção	**CBAM**	**SWM**	**Sem atenção**	**CBAM**	**SWM**
Acurácia	41,32%	41,78%	37,09%	36,15%	44,13%	37,09%
Pontuação F1	35,81%	32,82%	28,68%	34,04%	38,02%	34,62%
	Testes no conjunto JAFFE com oclusão					
Acurácia	30,99%	14,09%	21,60%	27,23%	35,68%	28,17%
Pontuação F1	34,99%	12,96%	20,01%	25,17%	27,47%	24,53%

partir de imagens com oclusões, o cenário muda. Nesse caso, a arquitetura de melhor acurácia é, novamente, a que conta com atenção *CBAM*, mas ainda assim possui pontuação F1 inferior ao modelo sem atenção treinado em dados sem oclusões.

Similarmente ao apresentado na subseção anterior, na Tabela 5.7 estão as métricas obtidas para modelos treinados com e sem oclusões, mas testados em faces oclusas. Ainda, no Apêndice E estão as matrizes de confusão obtidas, separadas por tipo de oclusão presente, tanto para os testes no conjunto JAFFE quanto para a inversão dos conjuntos de validação CK+. Aqui, repetem-se alguns dos padrões encontrados anteriormente:

1. Conforme o esperado, no geral, a presença de oclusões nos dados de treino incrementou as métricas obtidas;
2. Para a oclusão ocular, a arquitetura sem atenção treinada em imagens limpas performou melhor que a treinada em imagens com tarjas;
3. A arquitetura treinada em dados oclusos e com atenção *CBAM* apresenta os valores mais consistentes, tanto para oclusões oculares, quanto para mandibulares.

5.4 ANÁLISE QUALITATIVA

A análise qualitativa dos dados consiste na inspeção visual dos mapas de saliência gerados a partir do *Grad-CAM*. De acordo com o explicitado no capítulo anterior, o *Grad-CAM* indica quais regiões da imagem de entrada carregam maior relevância

Tabela 5.7 – Valores de acurácia e pontuação F1 separados pelo tipo de oclusão facial. Os modelos foram treinados com o conjunto CK+, com ou sem oclusões, e testados no conjunto JAFFE *com oclusões*.

Dados de treino	Atenção	Oclusão	Métricas	
			Acurácia	Pontuação F1
CK+ sem oclusões	**Sem atenção**	**Ocular**	42,20%	50,75%
		Mandibular	19,23%	23,05%
	CBAM	**Ocular**	12,85%	13,98%
		Mandibular	15,39%	13,44%
	SWM	**Ocular**	29,36%	29,77%
		Mandibular	13,46%	11,64%
CK+ com oclusões	**Sem atenção**	**Ocular**	31,19%	33,91%
		Mandibular	23,08%	21,93%
	CBAM	**Ocular**	38,53%	33,47%
		Mandibular	32,69%	23,27%
	SWM	**Ocular**	33,03%	31,05%
		Mandibular	23,08%	25,03%

para uma determinada classificação quando analisa-se uma camada convolucional específica. Nos mapas de saliência, a intensidade é representada por cores, onde vermelho carrega maior valor e azul menor. Em todas as análises, as imagens foram obtidas a partir da observação da última camada convolucional da rede *Resnet50*.

Salienta-se que, devido ao alto teor de subjetividade da inspeção visual, não é de intenção da autora explicar exacerbadamente ou desvendar minuciosamente como cada arquitetura opera, bem como as interpretações aqui expostas não são universais. Espera-se, apenas, encontrar alguns possíveis padrões nos comportamentos de mecanismos de atenção e também verificar se há diferença entre teores de importância em uma imagem quando há treino em dados distintos (nesse caso, com e sem oclusões).

Sendo assim, definiu-se dois eixos principais de análise: (1) como cada mecanismo de atenção interpreta as emoções; e (2) como o treino com e sem oclusões influencia na interpretação dos modelos.

5.4.1 Como os mecanismos de atenção influenciam na interpretação das emoções?

Para observação de como cada uma das três arquiteturas interpretam uma determinada emoção, selecionam-se as imagens cujas inferências de classe foram cor-

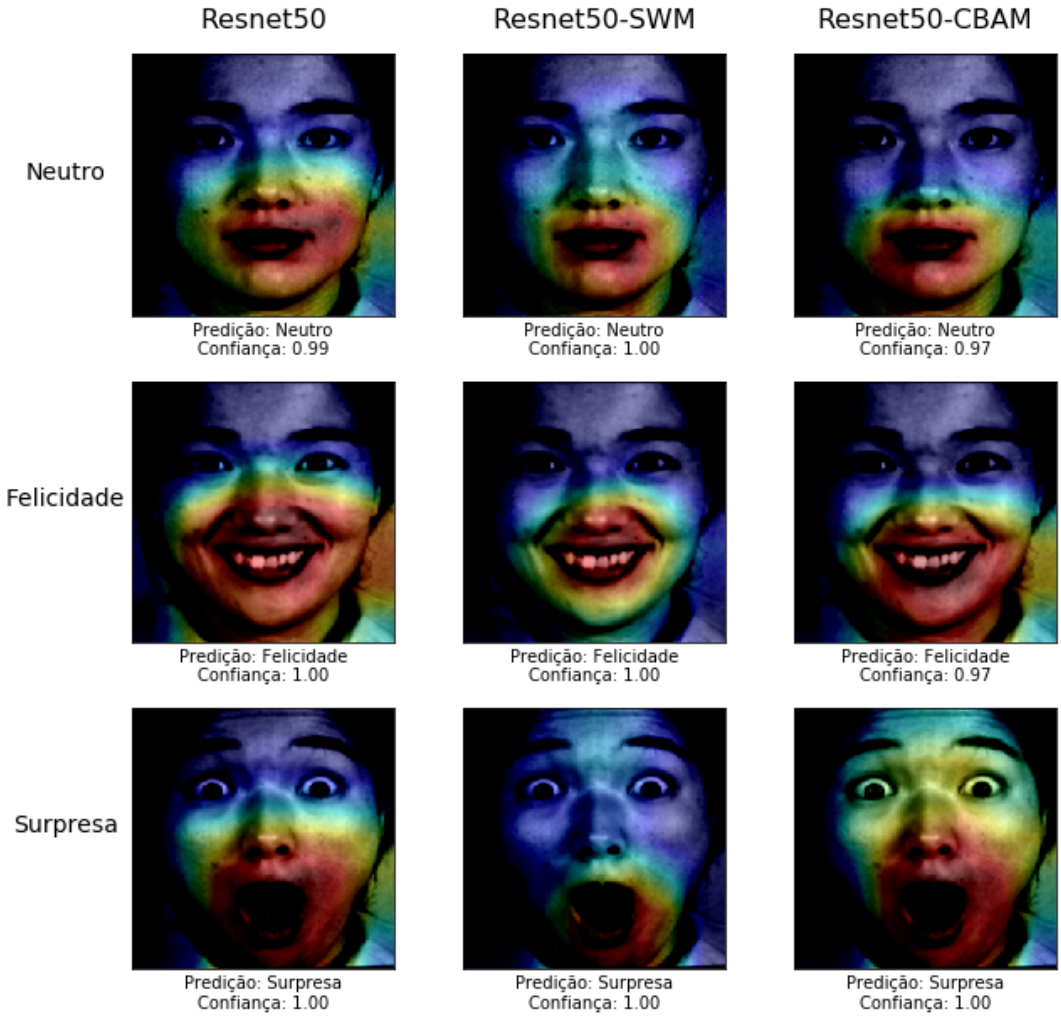
retas. Nesse momento, os modelos treinados em faces sem oclusões são expostos ao mesmo tipo de imagem, bem como os modelos treinados com oclusões são expostos a imagens oclusas. Primeiramente, serão analisadas amostras com predições corretas pelas redes, e, a seguir, imagens onde houve confusão, ou erro de inferência.

Na Figura 5.2 estão as imagens selecionadas para os modelos treinados sem oclusões, bem como os resultados de predição e confiança de predição. À primeira vista, observa-se que os *pixels* salientes nas faces diferenciam-se entre os modelos.

Para a face neutra, todos os três modelos dão maior importância para a região bucal, mas o módulo *SWM* provê alguma importância, mesmo que de baixa magnitude, até o topo do nariz, perto das sobrancelhas. Para a emoção de felicidade, a rede que abrange a maior parte da face e que dá maior importância para boca e nariz, é a que não contém nenhum mecanismo de atenção. Nesse caso, o modelo *SWM* não engloba todo o sorriso como os outros dois.

Por fim, para a emoção de surpresa, observa-se que, novamente, as redes que mais distribuem pesos pela face são a *Resnet50* pura e a com módulo de atenção *CBAM*. A segunda atribui valores quase que à totalidade da face observada, o que é interessante para a emoção em questão, que apresenta alta deformação facial não só na região bucal, mas também na região das sobrancelhas. A arquitetura que faz uso do *SWM*, mais uma vez, tem uma região de saliência menor, em termos espaciais, que suas “adversárias”.

Figura 5.2 – Amostras corretas pelos três modelos treinados em dados sem oclusão.



Fonte: autora.

Para a análise de imagens oclusas, presentes na Figura 5.3, retomam-se os elementos fornecidos pela literatura:

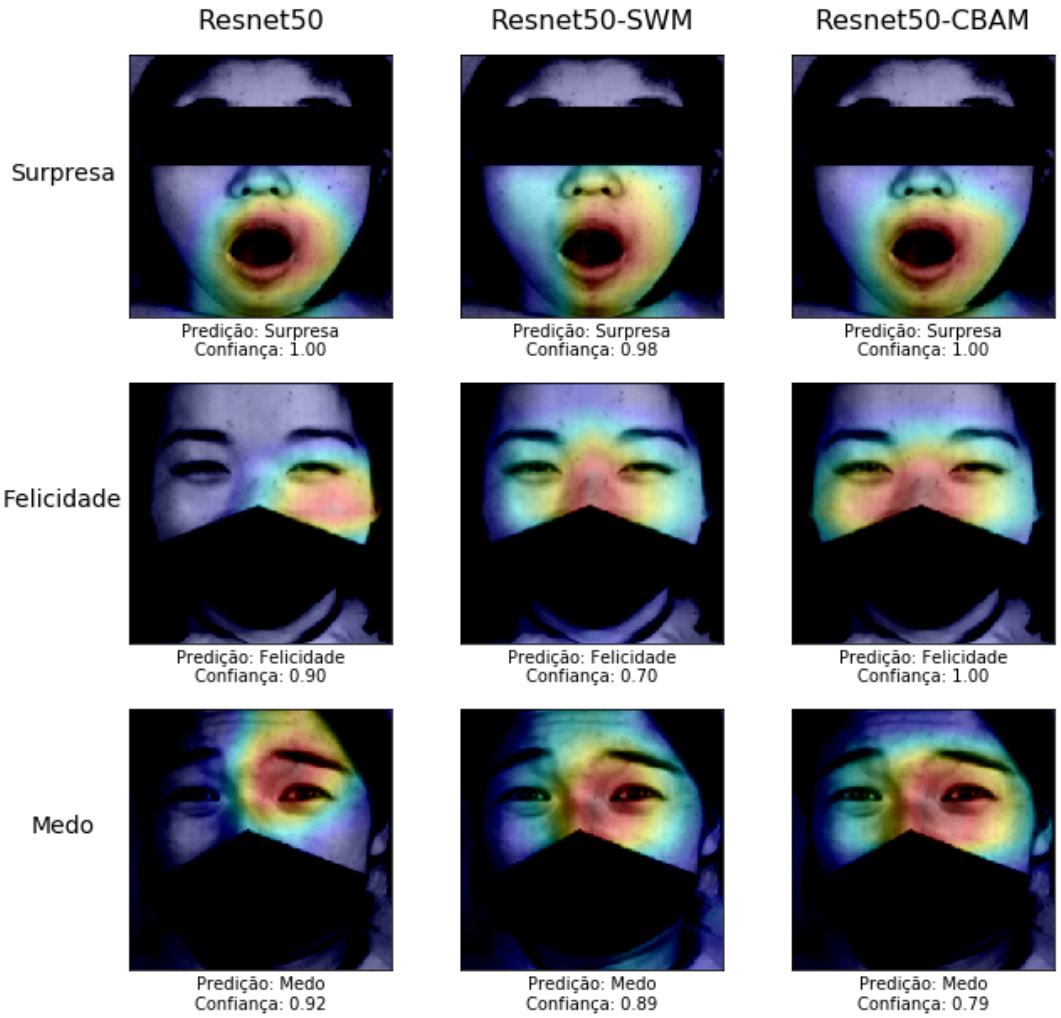
- a oclusão na região bucal dificulta a classificação de emoções como felicidade, raiva, medo e tristeza, mas também reduz a acurácia de todas as emoções;
- a oclusão ocular atinge o sentimento de nojo, principalmente.

Quando não há oclusão em regiões cruciais da imagem, como a de sentimento de surpresa, onde a boca, que carrega mais informação, está completamente exposta, as redes atribuem confiança quase totalmente unitária à classificação. Entretanto,

quando a oclusão é mandibular, os graus de confiança decrescem, como o esperado a partir dos dados citados anteriormente.

Em relação às diferentes arquiteturas analisadas, observa-se que nessas amostras aquelas que possuem mecanismos de atenção são mais abrangentes que o modelo *Resnet50* puro. Para as três emoções da Figura 5.3, o mapa de saliência é mais extenso nos modelos com *SWM* e *CBAM*, que fornecem relevância a ambos os olhos, por exemplo, nos sentimentos de medo e felicidade. Pelo contrário, o modelo puro baseia sua resposta em apenas um dos lados da face.

Figura 5.3 – Amostras *corretas* pelos três modelos treinados em dados *com oclusão*.

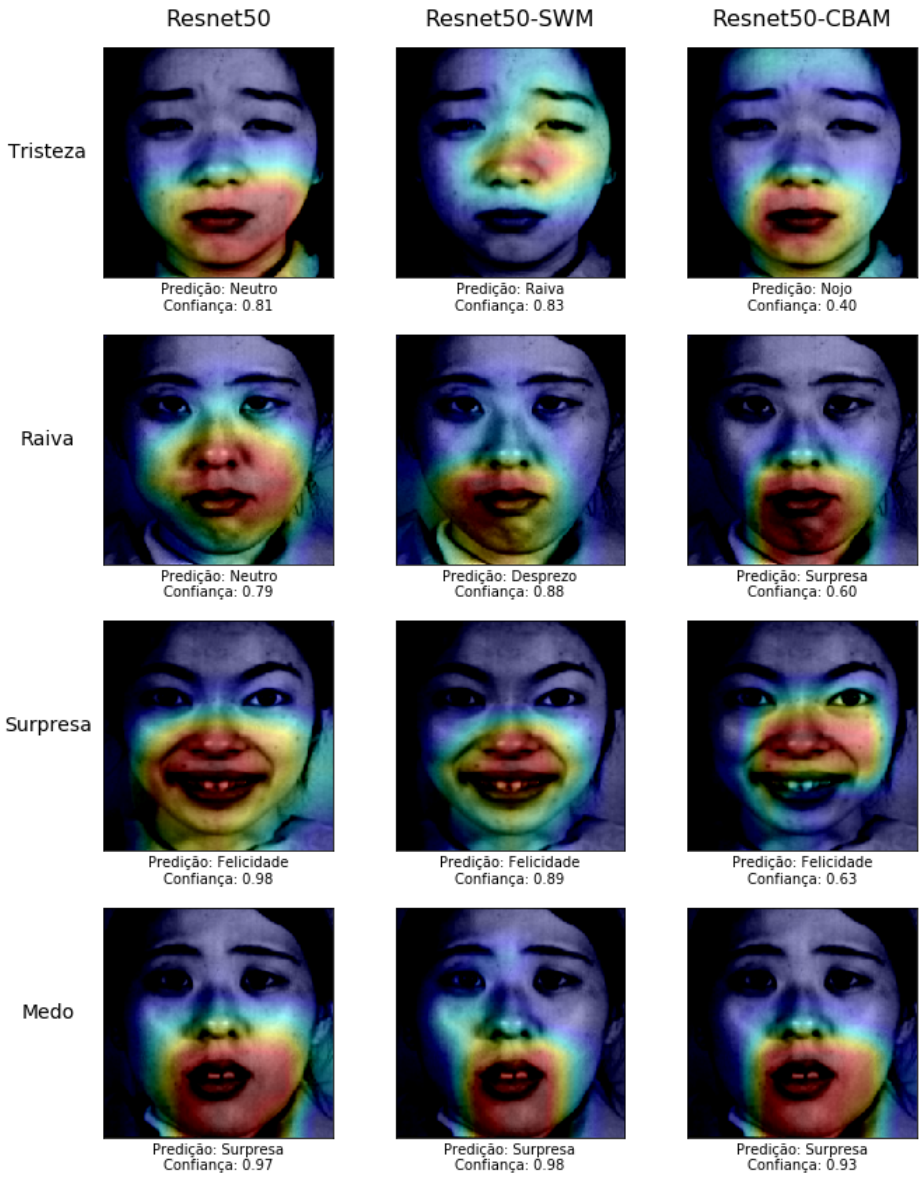


Fonte: autora.

5.4.1.1 Confusões

Quando erros são tomados pelos modelos comparados, observa-se que suas atenções focam-se em regiões semelhantes, conforme a Figura 5.4. Se as três redes predizem o mesmo, as regiões de maior saliência pouco diferem, de acordo com as imagens dos sentimentos de surpresa e medo. Pelo contrário, quando as redes chegam a resultados distintos, as regiões observadas nas imagens de entrada são, também, levemente diferentes. Nesse sentido, ressalta-se a emoção de tristeza que, nos modelos com atenção, teve a região superior da face (testa) apontada como um possível local de interesse, ao contrário do modelo sem atenção.

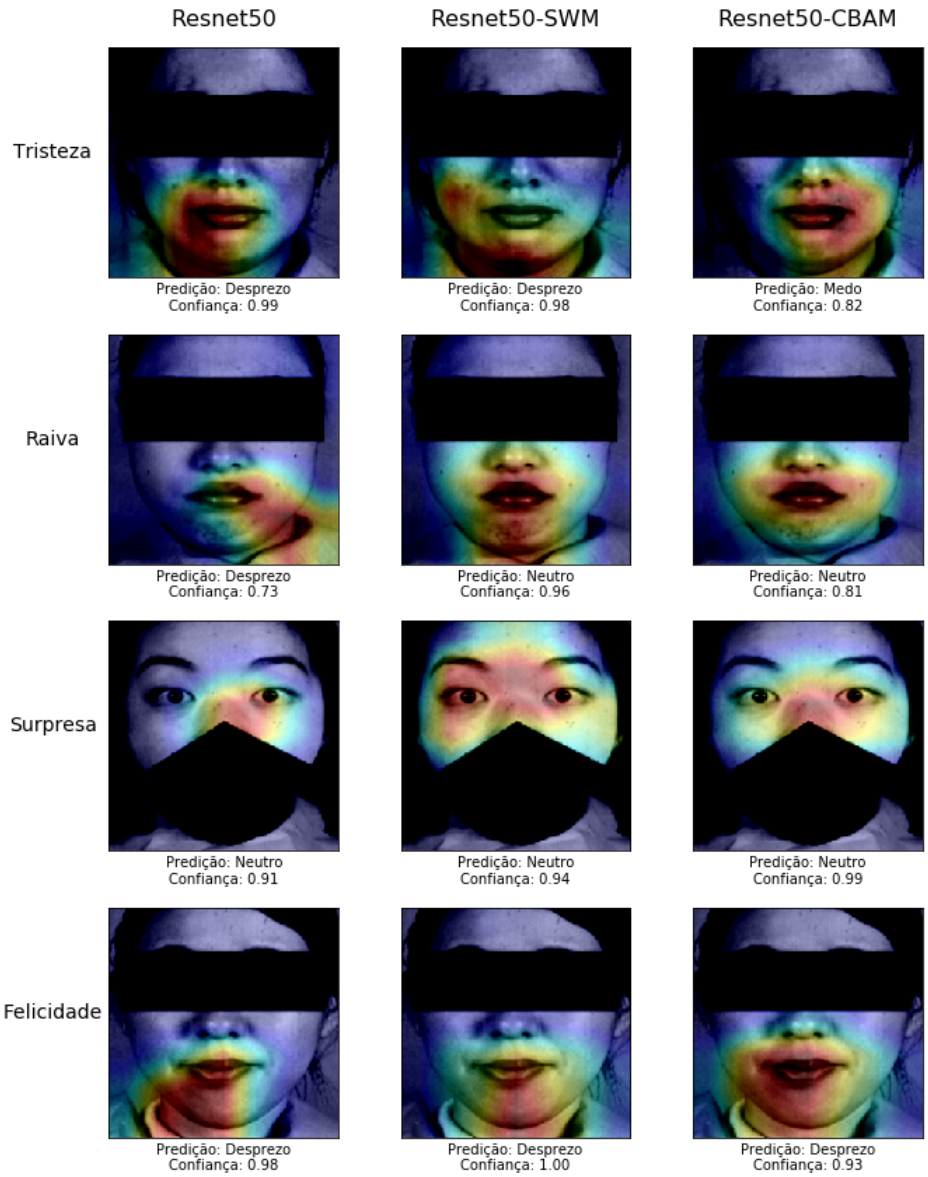
Figura 5.4 – Amostras *incorretas* pelos três modelos treinados em dados *sem oclusão*.



Fonte: autora.

O comportamento descrito se repete para o par modelo e imagens oclusas, pelo visto na Figura 5.5. Entretanto, embora a primeira imagem, correspondente à emoção de tristeza, apresente alguma movimentação na região superior à oclusão ocular, a rede falha em “ultrapassar” a tarja, apresentando saliências apenas para a região mandibular. Aqui, observa-se mais uma vez que os mecanismos de atenção tendem a ser mais generalistas, atribuindo importância a regiões maiores das imagens, como é possível enxergar nos exemplos de surpresa e raiva. Observa-se que o modelo treinado sem atenção ocasionalmente dá importância a regiões de fundo de imagem, como aconteceu nas amostras de felicidade e raiva.

Figura 5.5 – Amostras *incorretas* pelos três modelos treinados em dados *com oclusão*.



Fonte: autora.

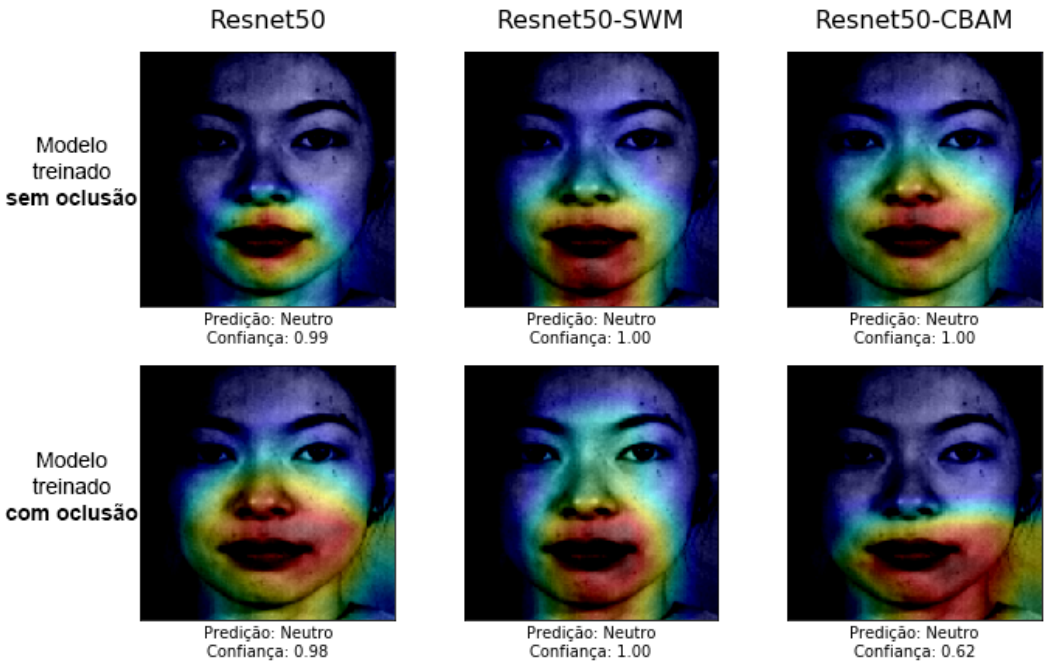
5.4.2 Qual o impacto dos treinos com e sem oclusão?

Nesta seção objetiva-se compreender se o treino com dados oclusos (ou a falta desses) intensifica a capacidade de generalização da rede. Nesse âmbito, realiza-se a comparação de mapas de saliência de imagens sem oclusão para redes treinadas com e sem oclusões.

5.4.2.1 Comparação dos modelos treinados com e sem oclusões em imagens sem oclusões

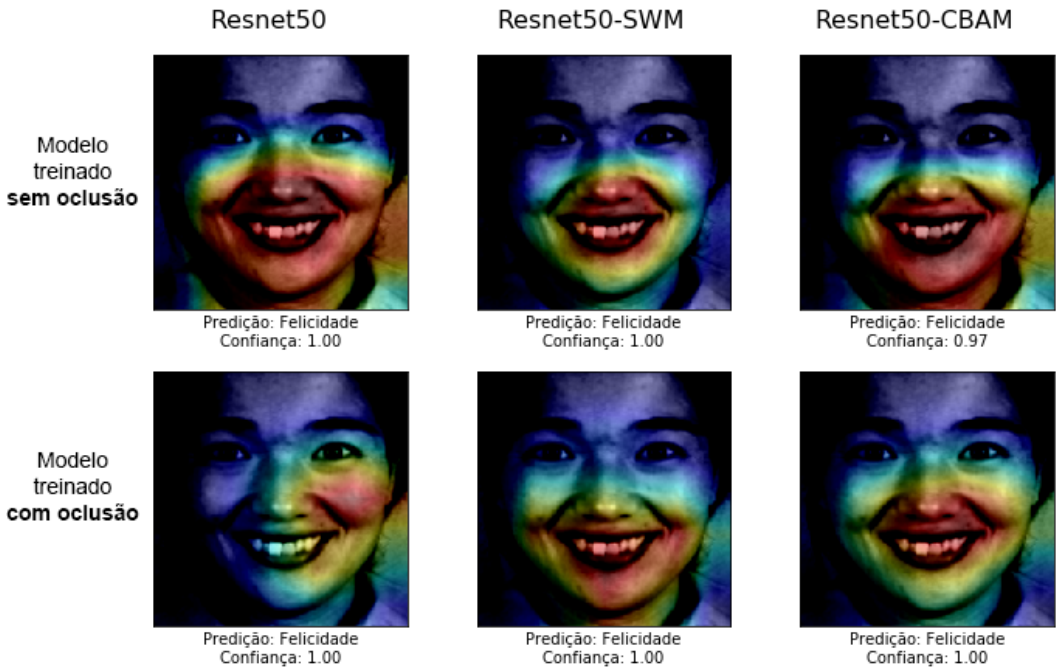
Nas Figuras 5.6, 5.7, 5.8 e 5.9 estão as amostras sem oclusões investigadas nessa subseção, para as quais todas as seis redes obtiveram resultados corretos. Para os sentimentos de neutralidade e felicidade, as importâncias entre as redes treinadas com e sem oclusões não diferiu intensamente: as regiões de maior importância são sempre as da boca, com alguma margem até o nariz. O resultado é esperado quando traz-se a análise por unidades de ação facial, de acordo com a Tabela 2.2, do sentimento de felicidade, para o qual deve existir contração imprescindível da AU12, músculo zigomático maior, da região dos cantos da boca.

Figura 5.6 – Mapa de saliência para predição correta da emoção de *neutralidade* por todas as seis redes (treinadas com e sem oclusões).



Fonte: autora.

Figura 5.7 – Mapa de saliência para predição correta da emoção de *felicidade* por todas as seis redes (treinadas com e sem oclusões).



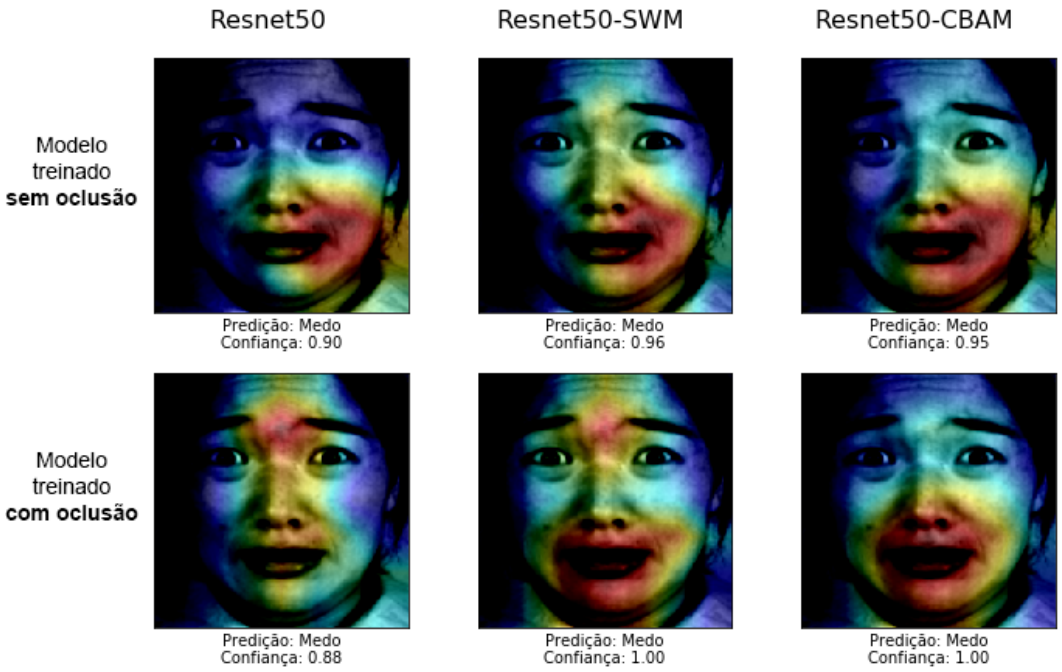
Fonte: autora.

O comportamento descrito difere, entretanto, para os sentimentos de surpresa e medo, onde há maior ativação de outras regiões da face, como sobrancelhas. Constata-se que as redes treinadas com oclusões são, no geral, capazes de atribuir maior significância às regiões superiores da face. Essa conduta, visualizada principalmente para a Figura 5.8 da emoção de medo, é verdade para as três arquiteturas, especialmente para a sem atenção e a com *SWM*: há a criação uma região de atenção intensa entre as sobrancelhas.

Para o sentimento de surpresa, da Figura 5.9, a intensificação de atenção à região superior da face é verdade novamente para o modelo sem atenção e para o com *SWM*. O modelo com atenção *CBAM*, embora atribua importância à região citada, o faz em menor intensidade que quando treinado em dados sem oclusões, que é capaz de considerar quase que a totalidade da face para inferência da classe.

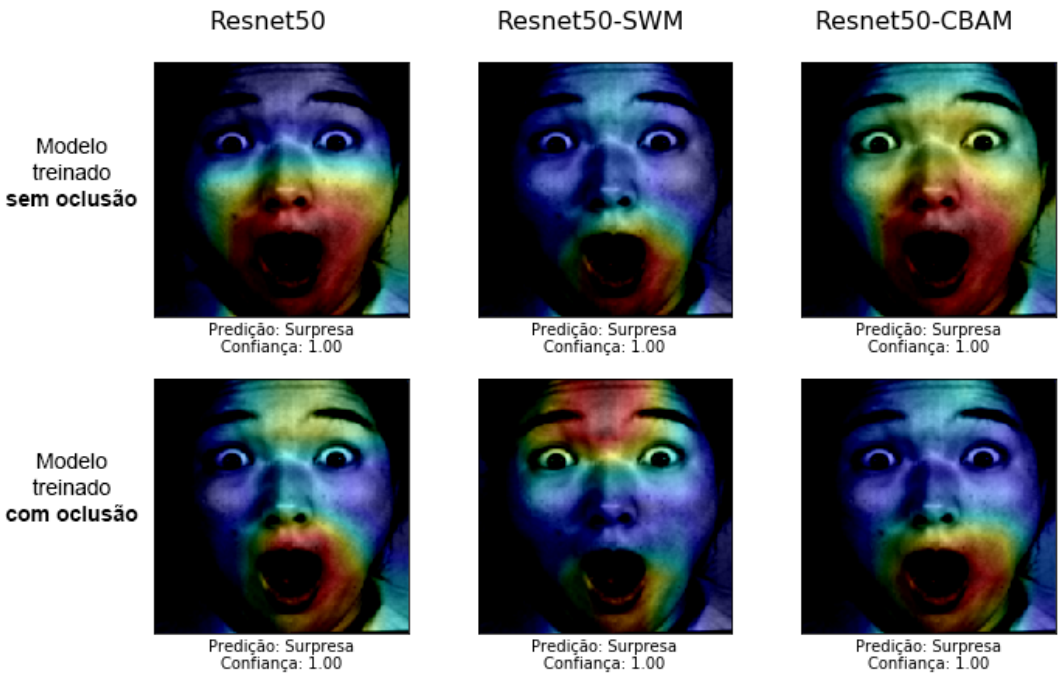
Traz-se, novamente, a análise por unidades de ação facial. Para os sentimentos de surpresa e medo, as unidades AU1 e AU2, correspondentes às regiões da sobrancelha, devem estar presentes. Sendo assim, observa-se que o treino com parcela de dados oclusos pode levar à atribuição de maior intensidade às regiões da face correspondentes às unidades de ação faciais esperadas para cada sentimento, quando comparam-se aos mapas de saliência obtidos para modelos treinados em imagens limpas.

Figura 5.8 – Mapa de saliência para predição correta da emoção de *medo* por todas as seis redes (treinadas com e sem oclusões).



Fonte: autora.

Figura 5.9 – Mapa de saliência para predição correta da emoção de *surpresa* por todas as seis redes (treinadas com e sem oclusões).



Fonte: autora.

6 CONCLUSÃO

Este projeto investigou a classificação de emoções a partir de imagens faciais, com ou sem oclusões parciais, com diferentes arquiteturas de redes neurais, que diferiam no tipo de mecanismos de atenção empregado. Nesse âmbito, realizaram-se testes invertendo os conjuntos de validação para o caso do *dataset* CK+, e no conjunto JAFFE. Ainda, explorou-se a interpretabilidade e explicabilidade dos resultados de forma qualitativa, por inspeção visual dos mapas de saliência gerados a partir do *Grad-CAM*.

Os modelos treinados em imagens sem oclusões atingiram até 97,77% de acurácia de validação com 20% dos dados do CK+, na arquitetura com o mecanismo *SWM*, embora as outras arquiteturas não tenham performado de forma tão inferior. Para o treino com oclusões, o melhor resultado de acurácia foi do modelo com atenção *CBAM*, com 90,01% de acertos. Ambos resultados, sem e com oclusões são comparáveis à literatura, onde encontraram-se acurácias de 96,8% para o conjunto CK+ sem oclusões (DING; ZHOU; CHELLAPPA, 2017) e 91,04% para ele com oclusões (CORNEJO; PEDRINI, 2018).

Ao invertermos os conjuntos de validação para os quais os modelos foram treinados - modelos treinados sem oclusões submetidos ao conjunto com oclusões e vice-versa - observa-se o aumento da acurácia dos modelos treinados com oclusões, enquanto aqueles treinados sem oclusões tem sua classificação deteriorada, conforme o esperado. Nesse experimento também observa-se, inesperadamente, que o treino com dados oclusos não é benéfico para a classificação de imagens com oclusão ocular quando o modelo não conta com nenhum mecanismo de atenção. Para todos os outros casos, os valores de acurácia e pontuação F1 são superiores quando há treino com imagens oclusas.

No teste com as imagens do segundo conjunto, o JAFFE, o modelo com atenção *CBAM* treinado em imagens com oclusões apresenta os resultados mais consistentes tanto quando as imagens estão limpas (44,13% de acurácia), quanto quando possuem algum tipo de tarja facial (35,68% de acurácia). Novamente, ao compararmos à literatura, o resultado obtido para as imagens limpas são competitivas, embora ainda inferior ao melhor resultado encontrado de 48% (ALI; IQBAL; CHOI, 2016).

Ainda, ressalta-se que a adição de mecanismos de atenção, quando o treino é realizado com imagens originais, deterioram a performance dos modelos para imagens com oclusões: enquanto a rede *Resnet50* pura atinge 31% de acertos, os modelos com *CBAM* e *SWM* atingem 14 e 21%, respectivamente. Esse dado indica que a união de mecanismos de atenção ao treino de dados com oclusão facial é recomendando, enquanto que o uso de mecanismos de atenção para treino em imagens

limpas não gera resultados satisfatórios.

No âmbito das análises qualitativas, por outro lado, objetivou-se a resolução de duas questões: (1) como cada mecanismo de atenção interpreta as emoções; e (2) como o treino com e sem oclusões influencia na interpretação dos modelos. Buscou-se resolução da primeira com a observação de amostras corretas para as três arquiteturas nos dois modos de treino (com e sem oclusões). Ainda, apresentou-se exemplos de confusões realizadas pelas redes. No segundo tema, realizou-se a comparação dos mapas de saliências entre modelos treinados com e sem oclusões para predições sempre corretas.

Com a primeira análise qualitativa, conclui-se que as redes que possuem mecanismos de atenção tendem a gerar raios de atenção mais extensos quando há tarja facial. A segunda análise permite inferir que, quando as redes são treinadas com imagens que possuem oclusões, essas tem propensão a distribuir a atenção em diferentes pontos de importância na face analisada. As regiões de atenção nesses casos parecem assemelhar-se às das unidades de ação facial esperadas para as emoções em questão.

Tendo em vista que os resultados mais consistentes de classificação do conjunto de testes nunca visto, JAFFE, foram os da rede com módulo de atenção *CBAM* treinada em dados com oclusão, assume-se que tanto a adição do mecanismo em questão, quanto o treino com presença de oclusões intensifica a possibilidade de acertos do modelo. Ainda, de acordo com o análise qualitativa, o treino com oclusões parece permitir uma maior “exploração” da face pela rede, que atribui maior atenção a regiões não tão óbvias de uma face não oclusa para as arquiteturas puras e com atenção *SWM*, principalmente.

Uma resolução ótima para a questão proposta encontra-se ainda em aberto. Embora seja possível constatar que os modelos com adição de atenção performam de forma superior aos modelos puros na maioria dos casos, nenhuma combinação de atenção e modo de treino foi capaz de gerar predições excelentes quando as regiões fundamentais para uma determinada emoção foram escondidas, ao menos para o conjunto de testes utilizado.

Finalmente, salienta-se a importância de trabalhos que explorem a interpretabilidade de algoritmos inteligentes. Esse campo de estudo deve objetivar não só a melhor adequação às normas vigentes, mas principalmente a democratização do entendimento da inteligência artificial e a transparência desses processos, de forma a viabilizar algoritmos éticos e respeitosos à pluralidade de indivíduos com os quais podem defrontar-se.

REFERÊNCIAS BIBLIOGRÁFICAS

ADEBAYO, J. *et al.* Sanity checks for saliency maps. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2018. p. 9505–9515.

ALI, G.; IQBAL, M. A.; CHOI, T.-S. Boosted nne collections for multicultural facial expression recognition. **Pattern Recognition**, Elsevier, v. 55, p. 14–27, 2016.

ALTMAN, A. Discrimination. In: ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Winter 2016. Metaphysics Research Lab, Stanford University, 2016. Acesso em 02 nov. 2019. Disponível em: <https://plato.stanford.edu/archives/win2016/entries/discrimination/>.

BAU, D. *et al.* Network dissection: Quantifying interpretability of deep visual representations. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2017. p. 6541–6549.

BHARATHARAJ, J. *et al.* Robot-assisted therapy for learning and social interaction of children with autism spectrum disorder. **Robotics**, Multidisciplinary Digital Publishing Institute, v. 6, n. 1, p. 4, 2017.

BISHOP, C. M. **Pattern recognition and machine learning**. [S.l.]: Springer Science+ Business Media, 2006.

BLAIR, R. J. R. A cognitive developmental approach to morality: Investigating the psychopath. **Cognition**, Elsevier, v. 57, n. 1, p. 1–29, 1995.

BOEHMKE, B. **Artificial Neural Network Fundamentals**. 2018. Acesso em 9 set. 2019. Disponível em: <http://uc-r.github.io/ann_fundamentals>.

CHEN, K. *et al.* Abc-cnn: An attention based convolutional neural network for visual question answering. **arXiv preprint arXiv:1511.05960**, 2015.

COLOMBO, A.; CUSANO, C.; SCHETTINI, R. The university of milano bicocca 3d face database. **University of Milano Bicocca**, 2011.

CORNEJO, J. Y. R.; PEDRINI, H. Emotion recognition from occluded facial expressions using weber local descriptor. In: IEEE. **2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP)**. [S.l.], 2018. p. 1–5.

DARWIN, C. **The expression of the emotions in man and animals**. [S.l.]: John Murray, London, 1872.

DING, H.; ZHOU, S. K.; CHELLAPPA, R. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In: IEEE. **2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)**. [S.l.], 2017. p. 118–126.

DRIRA, H. *et al.* 3d face recognition under expressions, occlusions, and pose variations. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 35, n. 9, p. 2270–2283, 2013.

EKMAN, P. The argument and evidence about universals in facial expressions. **Handbook of social psychophysiology**, John Wiley Chichester, England, p. 143–164, 1989.

____. An argument for basic emotions. **Cognition & emotion**, Taylor & Francis, v. 6, n. 3-4, p. 169–200, 1992.

EKMAN, P.; FRIESEN, W. V. Measuring facial movement. **Environmental psychology and nonverbal behavior**, Springer, v. 1, n. 1, p. 56–75, 1976.

____. **Manual for the facial action code. Palo Alto**. [S.l.]: CA: Consulting Psychologists Press, 1978.

EKMAN, P. E.; DAVIDSON, R. J. **The nature of emotion: Fundamental questions.** [S.l.]: Oxford University Press, 1994.

ERHAN, D. *et al.* Visualizing higher-layer features of a deep network. **University of Montreal**, v. 1341, n. 3, p. 1, 2009.

FU, J.; ZHENG, H.; MEI, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 4438–4446.

GILPIN, L. H. *et al.* Explaining explanations: An overview of interpretability of machine learning. In: IEEE. **2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)**. [S.l.], 2018. p. 80–89.

GLOROT, X.; BORDES, A.; BENGIO, Y. Deep sparse rectifier neural networks. In: GORDON, G.; DUNSON, D.; DUDíK, M. (Ed.). **Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics**. Fort Lauderdale, FL, USA: PMLR, 2011. (Proceedings of Machine Learning Research, v. 15), p. 315–323. Disponível em: <http://proceedings.mlr.press/v15/glorot11a.html>.

GONZALEZ-LIENCRES, C.; SHAMAY-TSOORY, S. G.; BRÜNE, M. Towards a neuroscience of empathy: ontogeny, phylogeny, brain mechanisms, context and psychopathology. **Neuroscience & Biobehavioral Reviews**, Elsevier, v. 37, n. 8, p. 1537–1548, 2013.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016. Acesso em 9 set. 2019. Disponível em: <http://www.deeplearningbook.org>.

GOODMAN, B.; FLAXMAN, S. European union regulations on algorithmic decision-making and a right to explanation. **AI Magazine**, v. 38, n. 3, p. 50–57, 2017.

HAYKIN, S. S. *et al.* **Neural networks and learning machines/Simon Haykin.** [S.l.]: New York: Prentice Hall„ 2009.

HE, K. *et al.* Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

HOCHREITER, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, World Scientific, v. 6, n. 02, p. 107–116, 1998.

HU, J.; SHEN, L.; SUN, G. Squeeze-and-excitation networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 7132–7141.

JUEFEI-XU, F. *et al.* Deepgender: Occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention. In: **Proceedings of the IEEE conference on computer vision and pattern recognition workshops**. [S.l.: s.n.], 2016. p. 68–77.

KARPATHY, A. Stanford university cs231n: Convolutional neural networks for visual recognition. 2018. Acesso em 9 set. 2019. Disponível em: <http://cs231n.stanford.edu/syllabus.html>.

KHORRAMI, P.; PAINE, T.; HUANG, T. Do deep neural networks learn facial action units when doing expression recognition? In: **Proceedings of the IEEE International Conference on Computer Vision Workshops**. [S.l.: s.n.], 2015. p. 19–27.

KOTSIA, I.; BUCIU, I.; PITAS, I. An analysis of facial expression recognition under partial facial image occlusion. **Image and Vision Computing**, Elsevier, v. 26, n. 7, p. 1052–1067, 2008.

LANGNER, O. *et al.* Presentation and validation of the radboud faces database. **Cognition and emotion**, Taylor & Francis, v. 24, n. 8, p. 1377–1388, 2010.

LAROCHELLE, H.; HINTON, G. E. Learning to combine foveal glimpses with a third-order boltzmann machine. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2010. p. 1243–1251.

LI, Y. *et al.* Occlusion aware facial expression recognition using cnn with attention mechanism. **IEEE Transactions on Image Processing**, IEEE, v. 28, n. 5, p. 2439–2450, 2018.

LIN, K. C. *et al.* Facial emotion recognition towards affective computing-based learning. **Library Hi Tech**, Emerald Group Publishing Limited, v. 31, n. 2, p. 294–307, 2013.

LIU, M. *et al.* Au-aware deep networks for facial expression recognition. In: IEEE. **2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)**. [S.l.], 2013. p. 1–6.

LIU, P. *et al.* Facial expression recognition via a boosted deep belief network. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2014. p. 1805–1812.

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.

LUNDQVIST, D.; FLYKT, A.; ÖHMAN, A. The karolinska directed emotional faces (kdef). **CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet**, v. 91, p. 630, 1998.

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.

MAHMOUD, M. *et al.* 3d corpus of spontaneous complex mental states. In: **Conference on Affective Computing and Intelligent Interaction**. [S.l.: s.n.], 2011.

MATTHEWS, G.; WELLS, A. The cognitive science of attention and emotion. John Wiley & Sons Ltd, 1999.

MITCHELL, R. L.; PHILLIPS, L. H. The overlapping relationship between emotion perception and theory of mind. **Neuropsychologia**, Elsevier, v. 70, p. 1–10, 2015.

MURPHY, K. P. **Machine learning: a probabilistic perspective**. [S.l.: s.n.], 2012.

NIEDENTHAL, P. M.; RIC, F. **Psychology of emotion**. [S.l.]: Psychology Press, 2017.

NIJS, Y. **Childrens lying behavior towards personified robots: an experimental study**. 03 2016. Tese (Doutorado), 03 2016.

ÖHMAN, A. Face the beast and fear the face: Animal and social fears as prototypes for evolutionary analyses of emotion. **Psychophysiology**, Wiley Online Library, v. 23, n. 2, p. 123–145, 1986.

OLDEN, J. D.; JACKSON, D. A. Illuminating the black box: a randomization approach for understanding variable contributions in artificial neural networks. **Ecological modelling**, Elsevier, v. 154, n. 1-2, p. 135–150, 2002.

PANTIC, M. *et al.* Web-based database for facial expression analysis. In: IEEE. **2005 IEEE international conference on multimedia and Expo**. [S.l.], 2005. p. 5–pp.

PARKHI, O. M.; VEDALDI, A.; ZISSERMAN, A. Deep face recognition. In: **British Machine Vision Conference**. [S.l.: s.n.], 2015.

PASZKE, A. *et al.* **PyTorch: An Imperative Style, High-Performance Deep Learning Library**. 2019.

POLYAK, B. T. Some methods of speeding up the convergence of iteration methods. **USSR Computational Mathematics and Mathematical Physics**, Elsevier, v. 4, n. 5, p. 1–17, 1964.

RAMSUNDAR, B.; ZADEH, R. B. **TensorFlow for deep learning: from linear regression to reinforcement learning**. [S.l.]: "O'Reilly Media, Inc.", 2018.

RAZAVIAN, A. S. *et al.* Cnn features off-the-shelf: an astounding baseline for recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition workshops**. [S.l.: s.n.], 2014. p. 806–813.

RIEDL, M. O. Computational narrative intelligence: A human-centered goal for artificial intelligence. **arXiv preprint arXiv:1602.06484**, 2016.

SELVARAJU, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In: **Proceedings of the IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2017. p. 618–626.

SHAN, C.; GONG, S.; MCOWAN, P. W. Facial expression recognition based on local binary patterns: A comprehensive study. **Image and vision Computing**, Elsevier, v. 27, n. 6, p. 803–816, 2009.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. In: JMLR. ORG. **Proceedings of the 34th International Conference on Machine Learning-Volume 70**. [S.l.], 2017. p. 3145–3153.

SIMONYAN, K.; VEDALDI, A.; ZISSERMAN, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. **arXiv preprint arXiv:1312.6034**, 2013.

SMITH, L. N. Cyclical learning rates for training neural networks. In: IEEE. **2017 IEEE Winter Conference on Applications of Computer Vision (WACV)**. [S.l.], 2017. p. 464–472.

SPRINGENBERG, J. T. *et al.* Striving for simplicity: The all convolutional net. **arXiv preprint arXiv:1412.6806**, 2014.

TOWNER, H.; SLATER, M. Reconstruction and recognition of occluded facial expressions using pca. In: SPRINGER. **International Conference on Affective Computing and Intelligent Interaction**. [S.l.], 2007. p. 36–47.

TRASK, A. **Grokking deep learning**. [S.l.]: Manning Publications Co., 2019.

TZENG, F.-Y.; MA, K.-L. Opening the black box-data driven visualization of neural networks. In: IEEE. **VIS 05. IEEE Visualization, 2005.** [S.l.], 2005. p. 383–390.

WAN, W.; CHEN, J. Occlusion robust face recognition based on mask learning. In: IEEE. **2017 IEEE International Conference on Image Processing (ICIP)**. [S.l.], 2017. p. 3795–3799.

WANG, C.-F. **The Vanishing Gradient Problem**. Towards Data Science, 2019. Acesso em 15 nov. 2019. Disponível em: <https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484>.

WANG, S.; YUAN, Y.; FENG, Y. Local and global feature learning for subtle facial expression recognition from attention perspective. In: SPRINGER. **Chinese Conference on Pattern Recognition and Computer Vision (PRCV)**. [S.l.], 2019. p. 670–681.

WOO, S. *et al.* Cbam: Convolutional block attention module. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. [S.l.: s.n.], 2018. p. 3–19.

XU, K. *et al.* Show, attend and tell: Neural image caption generation with visual attention. In: **International conference on machine learning**. [S.l.: s.n.], 2015. p. 2048–2057.

ZAVAREZ, M. V.; BERRIEL, R. F.; OLIVEIRA-SANTOS, T. Cross-database facial expression recognition based on fine-tuned deep convolutional network. In: IEEE. **2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)**. [S.l.], 2017. p. 405–412.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: SPRINGER. **European conference on computer vision**. [S.l.], 2014. p. 818–833.

ZEILER, M. D. *et al.* Deconvolutional networks. In: IEEE. **2010 IEEE Computer Society Conference on computer vision and pattern recognition**. [S.l.], 2010. p. 2528–2535.

ZHANG, G. *et al.* Generative adversarial network with spatial attention for face attribute editing. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. [S.l.: s.n.], 2018. p. 417–432.

ZHANG, L. *et al.* Facial expression analysis under partial occlusion: A survey. **ACM Computing Surveys (CSUR)**, ACM, v. 51, n. 2, p. 25, 2018.

ZHANG, X. *et al.* It's written all over your face: Full-face appearance-based gaze estimation. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops**. [S.l.: s.n.], 2017. p. 51–60.

ZHAO, X. *et al.* Peak-piloted deep network for facial expression recognition. In: SPRINGER. **European conference on computer vision**. [S.l.], 2016. p. 425–442.

APÊNDICE A – HIPERPARÂMETROS DE TREINO DE TODOS OS MODELOS

Os valores apresentados nas tabelas referem-se a modelos treinados após validação cruzada. Nos modelos com validações em 20% dos dados, alguns poucos hiperparâmetros foram modificados. Nesses casos, as modificações são explicitadas após cada tabela.

A.1 – TREINO EM DADOS SEM OCLUSÕES

Tabela A.1 – Parâmetros e demais dados do modelo Resnet50 sem atenção.

Parâmetro	**Valor**
Modelo	Resnet50
Atenção	Nenhuma
Oclusões no Treino	Nenhuma
Otimização	Gradiente Descendente
Momentum	0,9
Weight Decay	0,01
Épocas de Treino	27
Tamanho de *Batch*	16
Agendador de *Learning Rate*	Redução de LR na estabilização do valor de perda
***Learning Rate* Base**	0,006667

Para o modelo sem atenção e validado em 20% dos dados, o único parâmetro alterado é o número de épocas de treino, que ao invés de 27 foi 28.

Tabela A.2 – Parâmetros e demais dados do modelo Resnet50 com *CBAM*.

Parâmetro	**Valor**
Modelo	Resnet50
Atenção	CBAM
Oclusões no Treino	Nenhuma
Otimização	Gradiente Descendente
Momentum	0,9
Weight Decay	0,01
Tamanho de *Batch*	16
Agendador de *Learning Rate*	Redução de LR na estabilização do valor de perda
Épocas de Treino (modelo congelado)	5
Épocas de Treino (modelo completo)	24
***Learning Rate* Base (modelo congelado)**	0,01
***Learning Rate* Base (modelo completo)**	0,006667

Para o modelo sem atenção e validado em 20% dos dados nenhum hiperparâmetro foi modificado.

Tabela A.3 – Parâmetros e demais dados do modelo Resnet50 com *SWM*.

Parâmetro	**Valor**
Modelo	Resnet50
Atenção	SWM
Oclusões no Treino	Nenhuma
Otimização	Gradiente Descendente
Momentum	0,9
Weight Decay	0,01
Tamanho de *Batch*	16
Agendador de *Learning Rate*	Redução de LR na estabilização do valor de perda
Épocas de Treino (modelo congelado)	5
Épocas de Treino (modelo completo)	21
***Learning Rate* Base (modelo congelado)**	0,003
***Learning Rate* Base (modelo completo)**	0,0017

Para o modelo com atenção *SWM* e validado em 20% dos dados, o único parâmetro alterado é o valor de *Learning Rate* para o modelo congelado, que ao invés de 0,003 foi de 0,001.

A.2 – TREINO EM DADOS COM OCLUSÕES

Tabela A.4 – Parâmetros e demais dados do modelo Resnet50 sem atenção.

Parâmetro	**Valor**
Modelo	Resnet50
Atenção	SWM
Oclusões no Treino	33% oculares, 33% mandibulares e 33% sem oclusão
Otimização	Gradiente Descendente
Momentum	0,9
Weight Decay	0,0001
Tamanho de *Batch*	16
Agendador de *Learning Rate*	*Learning Rate*s Cíclicas
Épocas de Treino (modelo completo)	39
***Learning Rate* Base (modelo completo)**	0,006667
***Learning Rate* Máxima (modelo completo)**	0,03

Para o modelo sem atenção e validado em 20% dos dados nenhum hiperparâmetro foi modificado.

Tabela A.5 – Parâmetros e demais dados do modelo Resnet50 com *CBAM*.

Parâmetro	**Valor**
Modelo	Resnet50
Atenção	CBAM
Oclusões no Treino	33% oculares, 33% mandibulares e 33% sem oclusão
Otimização	Gradiente Descendente
Momentum	0,9
Weight Decay	0,0001
Tamanho de *Batch*	16
Agendador de *Learning Rate*	*Learning Rate*s Cíclicas
Épocas de Treino (modelo congelado)	5
Épocas de Treino (modelo completo)	34
***Learning Rate* Base (modelo congelado)**	0,001
***Learning Rate* Máxima (modelo congelado)**	0,003
***Learning Rate* Base (modelo completo)**	0,003
***Learning Rate* Máxima (modelo completo)**	0,009

Para o modelo com atenção *CBAM* e validado em 20% dos dados nenhum hiperparâmetro foi modificado.

Tabela A.6 – Parâmetros e demais dados do modelo Resnet50 com *SWM*.

Parâmetro	**Valor**
Modelo	Resnet50
Atenção	SWM
Oclusões no Treino	33% oculares, 33% mandibulares e 33% sem oclusão
Otimização	Gradiente Descendente
Momentum	0,9
Weight Decay	0,0001
Tamanho de *Batch*	16
Agendador de *Learning Rate*	*Learning Rate*s Cíclicas
Épocas de Treino (modelo congelado)	5
Épocas de Treino (modelo completo)	37
***Learning Rate* Base (modelo congelado)**	0,001
***Learning Rate* Máxima (modelo congelado)**	0,003
***Learning Rate* Base (modelo completo)**	0,0006667
***Learning Rate* Máxima (modelo completo)**	0,002

Para o modelo com atenção *SWM* e validado em 20% dos dados, o único parâmetro alterado é o número de épocas de treino com o modelo completo, que ao invés de 37 foi 28.

APÊNDICE B – MATRIZES DE CONFUSÃO DOS CONJUNTOS DE VALIDAÇÃO

B.1 – MODELOS TREINADOS SEM OCLUSÃO

Tabela B.1 – Resnet50 sem atenção

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**1,000**	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Desprezo	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,000	0,000	**0,971**	0,000	0,000	0,000	0,000	0,029
Medo	0,000	0,000	0,000	**0,842**	0,000	0,000	0,053	0,105
Felicidade	0,000	0,000	0,000	0,000	**1,000**	0,000	0,000	0,000
Tristeza	0,000	0,000	0,000	0,000	0,000	**0,929**	0,000	0,071
Surpresa	0,000	0,000	0,000	0,000	0,000	0,000	**0,959**	0,041
Neutro	0,000	0,033	0,000	0,000	0,000	0,017	0,000	**0,950**

Tabela B.2 – Resnet50 com atenção *SWM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**1,000**	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Desprezo	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,000	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000
Medo	0,000	0,000	0,000	**0,895**	0,000	0,000	0,105	0,000
Felicidade	0,000	0,000	0,000	0,000	**1,000**	0,000	0,000	0,000
Tristeza	0,000	0,000	0,000	0,000	0,000	**0,929**	0,000	0,071
Surpresa	0,000	0,020	0,000	0,000	0,000	0,000	**0,959**	0,020
Neutro	0,000	0,017	0,000	0,000	0,000	0,000	0,000	**0,983**

Tabela B.3 – Modelo Resnet50 com atenção *CBAM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,967**	0,000	0,000	0,000	0,000	0,000	0,000	0,033
Desprezo	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,000	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000
Medo	0,000	0,000	0,000	**0,895**	0,000	0,000	0,000	0,105
Felicidade	0,000	0,000	0,000	0,000	**1,000**	0,000	0,000	0,000
Tristeza	0,000	0,000	0,000	0,000	0,000	**0,786**	0,000	0,214
Surpresa	0,000	0,000	0,000	0,000	0,000	0,000	**0,959**	0,041
Neutro	0,000	0,017	0,000	0,000	0,000	0,000	0,000	**0,983**

B.2 – MODELOS TREINADOS COM OCLUSÃO

Tabela B.4 – Resnet50 sem atenção

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,867**	0,000	0,067	0,000	0,000	0,033	0,000	0,033
Desprezo	0,000	**0,444**	0,222	0,000	0,111	0,000	0,000	0,222
Nojo	0,000	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000
Medo	0,053	0,000	0,000	**0,737**	0,053	0,000	0,105	0,053
Felicidade	0,000	0,000	0,022	0,000	**0,978**	0,000	0,000	0,000
Tristeza	0,000	0,000	0,071	0,143	0,000	**0,571**	0,000	0,214
Surpresa	0,000	0,000	0,000	0,020	0,000	0,000	**0,898**	0,082
Neutro	0,033	0,017	0,017	0,017	0,033	0,000	0,017	**0,867**

Tabela B.5 – Resnet50 com atenção *SWM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,900**	0,000	0,033	0,000	0,000	0,000	0,033	0,033
Desprezo	0,111	**0,778**	0,000	0,000	0,000	0,000	0,000	0,111
Nojo	0,029	0,000	**0,914**	0,000	0,029	0,000	0,029	0,000
Medo	0,000	0,000	0,000	**0,632**	0,000	0,000	0,263	0,105
Felicidade	0,000	0,000	0,000	0,000	**0,957**	0,000	0,043	0,000
Tristeza	0,000	0,071	0,000	0,000	0,000	**0,714**	0,000	0,214
Surpresa	0,000	0,020	0,020	0,000	0,000	0,000	**0,898**	0,061
Neutro	0,000	0,033	0,000	0,000	0,000	0,017	0,083	**0,867**

Tabela B.6 – Modelo Resnet50 com atenção *CBAM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,900**	0,000	0,033	0,000	0,000	0,033	0,000	0,033
Desprezo	0,111	**0,778**	0,000	0,000	0,000	0,000	0,000	0,111
Nojo	0,000	0,000	**0,971**	0,000	0,000	0,029	0,000	0,000
Medo	0,158	0,000	0,000	**0,579**	0,053	0,000	0,000	0,211
Felicidade	0,000	0,000	0,000	0,000	**1,000**	0,000	0,000	0,000
Tristeza	0,071	0,000	0,000	0,000	0,000	**0,714**	0,000	0,214
Surpresa	0,000	0,020	0,000	0,000	0,000	0,000	**0,878**	0,102
Neutro	0,017	0,000	0,000	0,000	0,000	0,017	0,000	**0,967**

APÊNDICE C – MATRIZES DE CONFUSÃO COM INVERSÃO DOS CONJUNTOS DE VALIDAÇÃO DO CONJUNTO CK+ COM E SEM OCLUSÕES

C.1 – MODELOS TREINADOS SEM OCLUSÕES E TESTADOS EM IMAGENS COM OCLUSÕES

Tabela C.1 – Resnet50 sem atenção testado em dados oclusos

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,933**	0,000	0,000	0,033	0,000	0,000	0,000	0,033
Desprezo	0,111	**0,556**	0,000	0,222	0,000	0,000	0,000	0,111
Nojo	0,200	0,000	**0,600**	0,029	0,000	0,000	0,000	0,171
Medo	0,000	0,000	0,000	**0,789**	0,000	0,000	0,000	0,211
Felicidade	0,000	0,000	0,000	0,022	**0,913**	0,022	0,000	0,043
Tristeza	0,000	0,000	0,000	0,000	0,000	**0,857**	0,000	0,143
Surpresa	0,000	0,000	0,000	0,102	0,020	0,041	**0,653**	0,184
Neutro	0,017	0,017	0,000	0,000	0,000	0,050	0,000	**0,917**

Tabela C.2 – Resnet50 com atenção *SWM* testado em dados oclusos

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,733**	0,000	0,000	0,000	0,000	0,100	0,033	0,133
Desprezo	0,000	**0,778**	0,000	0,000	0,000	0,000	0,000	0,222
Nojo	0,029	0,000	**0,829**	0,057	0,000	0,000	0,000	0,086
Medo	0,000	0,000	0,053	**0,684**	0,000	0,053	0,211	0,000
Felicidade	0,043	0,000	0,022	0,087	**0,674**	0,000	0,109	0,065
Tristeza	0,000	0,000	0,000	0,000	0,000	**0,786**	0,071	0,143
Surpresa	0,020	0,020	0,020	0,000	0,000	0,000	**0,898**	0,041
Neutro	0,000	0,000	0,000	0,000	0,000	0,033	0,017	**0,950**

Tabela C.3 – Resnet50 com atenção *CBAM* testado em dados oclusos

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,933**	0,000	0,000	0,000	0,000	0,000	0,033	0,033
Desprezo	0,111	**0,556**	0,000	0,000	0,000	0,000	0,000	0,333
Nojo	0,143	0,000	**0,714**	0,029	0,000	0,000	0,000	0,114
Medo	0,158	0,000	0,000	**0,632**	0,000	0,000	0,000	0,211
Felicidade	0,022	0,000	0,000	0,022	**0,804**	0,000	0,000	0,152
Tristeza	0,000	0,000	0,000	0,071	0,000	**0,786**	0,000	0,143
Surpresa	0,020	0,000	0,000	0,061	0,000	0,020	**0,653**	0,245
Neutro	0,017	0,017	0,000	0,000	0,000	0,000	0,000	**0,967**

C.2 – MODELOS TREINADOS COM OCLUSÕES E TESTADOS EM IMAGENS SEM OCLUSÕES

Tabela C.4 – Resnet50 sem atenção testado em dados originais

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,967**	0,000	0,000	0,000	0,000	0,000	0,000	0,033
Desprezo	0,000	**0,556**	0,111	0,000	0,000	0,000	0,000	0,333
Nojo	0,000	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000
Medo	0,000	0,000	0,053	**0,895**	0,000	0,000	0,000	0,053
Felicidade	0,000	0,000	0,000	0,000	**1,000**	0,000	0,000	0,000
Tristeza	0,000	0,000	0,000	0,071	0,000	**0,714**	0,000	0,214
Surpresa	0,000	0,000	0,000	0,000	0,000	0,000	**0,959**	0,041
Neutro	0,000	0,017	0,000	0,000	0,000	0,000	0,000	**0,983**

Tabela C.5 – Resnet50 com atenção *SWM* testado em dados originais

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,933**	0,000	0,033	0,000	0,000	0,000	0,000	0,033
Desprezo	0,000	**0,889**	0,000	0,000	0,000	0,000	0,000	0,111
Nojo	0,000	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000
Medo	0,000	0,000	0,000	**0,895**	0,000	0,000	0,000	0,105
Felicidade	0,000	0,000	0,000	0,022	**0,978**	0,000	0,000	0,000
Tristeza	0,000	0,000	0,000	0,000	0,000	**0,643**	0,000	0,357
Surpresa	0,000	0,02	0,000	0,000	0,000	0,000	**0,959**	0,020
Neutro	0,017	0,017	0,000	0,000	0,000	0,000	0,000	**0,967**

Tabela C.6 – Resnet50 com atenção *CBAM* testado em dados originais

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,967**	0,033	0,000	0,000	0,000	0,000	0,000	0,000
Desprezo	0,000	**0,889**	0,000	0,000	0,000	0,000	0,000	0,111
Nojo	0,000	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000
Medo	0,000	0,000	0,053	**0,842**	0,053	0,000	0,000	0,053
Felicidade	0,000	0,000	0,000	0,000	**1,000**	0,000	0,000	0,000
Tristeza	0,000	0,000	0,000	0,000	0,000	**0,714**	0,000	0,286
Surpresa	0,000	0,020	0,000	0,000	0,000	0,000	**0,939**	0,041
Neutro	0,000	0,000	0,000	0,000	0,000	0,000	0,000	**1,000**

APÊNDICE D – MATRIZES DE CONFUSÃO GERAIS OBTIDAS NOS TESTES COM O CONJUNTO DE DADOS JAFFE

D.1 – MODELOS TREINADOS SEM OCLUSÕES TESTADOS EM DADOS SEM OCLUSÕES

Tabela D.1 – Modelo Resnet50 sem atenção

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,333**	0,000	0,000	0,000	0,067	0,000	0,000	0,600
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,345	0,000	**0,069**	0,207	0,000	0,000	0,103	0,276
Medo	0,000	0,000	0,000	**0,250**	0,000	0,000	0,500	0,250
Felicidade	0,000	0,032	0,000	0,065	**0,581**	0,000	0,000	0,323
Tristeza	0,355	0,000	0,000	0,129	0,000	**0,000**	0,000	0,516
Surpresa	0,033	0,000	0,000	0,000	0,033	0,000	**0,700**	0,233
Neutro	0,033	0,000	0,000	0,000	0,000	0,000	0,000	**0,967**

Tabela D.2 – Modelo Resnet50 com atenção *SWM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,133**	0,133	0,033	0,000	0,033	0,000	0,067	0,600
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,138	0,069	**0,069**	0,138	0,034	0,034	0,069	0,448
Medo	0,000	0,000	0,000	**0,281**	0,000	0,000	0,469	0,250
Felicidade	0,000	0,161	0,000	0,032	**0,419**	0,000	0,000	0,387
Tristeza	0,258	0,000	0,000	0,097	0,000	**0,032**	0,032	0,581
Surpresa	0,000	0,000	0,000	0,000	0,033	0,000	**0,700**	0,267
Neutro	0,033	0,000	0,000	0,000	0,000	0,000	0,000	**0,967**

Tabela D.3 – Modelo Resnet50 com atenção *CBAM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,267**	0,067	0,000	0,000	0,100	0,067	0,200	0,300
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,276	0,000	**0,069**	0,414	0,034	0,034	0,172	0,000
Medo	0,000	0,000	0,000	**0,406**	0,000	0,000	0,375	0,219
Felicidade	0,032	0,097	0,000	0,194	**0,548**	0,000	0,000	0,129
Tristeza	0,355	0,000	0,065	0,032	0,129	**0,097**	0,129	0,194
Surpresa	0,000	0,000	0,000	0,000	0,067	0,000	**0,767**	0,167
Neutro	0,000	0,000	0,000	0,100	0,033	0,000	0,100	**0,767**

D.2 – MODELOS TREINADOS COM OCLUSÕES TESTADOS EM DADOS COM OCLUSÕES

Tabela D.4 – Modelo Resnet50 sem atenção

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,067**	0,600	0,000	0,000	0,000	0,000	0,000	0,333
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,138	0,276	**0,138**	0,069	0,103	0,034	0,069	0,172
Medo	0,031	0,250	0,000	**0,375**	0,000	0,031	0,062	0,250
Felicidade	0,032	0,484	0,000	0,000	**0,194**	0,000	0,032	0,258
Tristeza	0,129	0,290	0,000	0,065	0,000	**0,065**	0,032	0,419
Surpresa	0,000	0,300	0,000	0,167	0,033	0,000	**0,267**	0,233
Neutro	0,067	0,133	0,000	0,000	0,000	0,000	0,000	**0,800**

Tabela D.5 – Modelo Resnet50 com atenção *SWM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,200**	0,267	0,000	0,000	0,000	0,000	0,000	0,533
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,379	0,241	**0,034**	0,000	0,069	0,034	0,000	0,241
Medo	0,250	0,219	0,000	**0,125**	0,000	0,094	0,000	0,312
Felicidade	0,032	0,484	0,000	0,000	**0,387**	0,000	0,000	0,097
Tristeza	0,323	0,290	0,000	0,000	0,032	**0,032**	0,000	0,323
Surpresa	0,133	0,133	0,000	0,000	0,000	0,000	**0,300**	0,433
Neutro	0,033	0,067	0,000	0,000	0,000	0,000	0,000	**0,900**

Tabela D.6 – Modelo Resnet50 com atenção *CBAM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,100**	0,033	0,033	0,000	0,100	0,000	0,033	0,700
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,103	0,069	**0,034**	0,345	0,138	0,069	0,103	0,138
Medo	0,031	0,000	0,000	**0,375**	0,000	0,156	0,219	0,219
Felicidade	0,000	0,065	0,000	0,032	**0,581**	0,000	0,000	0,323
Tristeza	0,194	0,032	0,000	0,194	0,097	**0,065**	0,000	0,419
Surpresa	0,000	0,067	0,000	0,100	0,033	0,000	**0,467**	0,333
Neutro	0,033	0,000	0,000	0,100	0,000	0,000	0,000	**0,867**

APÊNDICE E – MATRIZES DE CONFUSÃO, SEPARADAS POR TIPO DE OCLUSÃO, PARA MODELOS TREINADOS COM E SEM OCLUSÕES, MAS TESTADOS EM DADOS OCLUSOS DO CONJUNTO JAFFE

E.1 – MODELOS TREINADOS SEM OCLUSÕES

E.1.1 – Matrizes de confusão para oclusão ocular

Tabela E.1 – Resnet50 sem atenção

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Supresa**	**Neutro**
Raiva	**0,357**	0,000	0,214	0,000	0,071	0,000	0,000	0,357
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,4	0,000	**0,600**	0,000	0,000	0,000	0,000	0,000
Medo	0,231	0,000	0,462	**0,000**	0,000	0,000	0,000	0,308
Felicidade	0,188	0,000	0,000	0,000	**0,438**	0,000	0,000	0,375
Tristeza	0,389	0,000	0,333	0,000	0,000	**0,000**	0,000	0,278
Surpresa	0,214	0,000	0,143	0,000	0,000	0,000	**0,5**	0,143
Neutro	0,053	0,000	0,000	0,000	0,000	0,000	0,000	**0,947**

Tabela E.2 – Resnet50 com atenção *CBAM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,143**	0,143	0,714	0,000	0,000	0,000	0,000	0,000
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,067	0,400	**0,533**	0,000	0,000	0,000	0,000	0,000
Medo	0,000	0,308	0,615	**0,077**	0,000	0,000	0,000	0,000
Felicidade	0,125	0,375	0,375	0,000	**0,125**	0,000	0,000	0,000
Tristeza	0,222	0,000	0,778	0,000	0,000	**0,000**	0,000	0,000
Surpresa	0,071	0,000	0,857	0,000	0,000	0,000	**0,071**	0,000
Neutro	0,000	0,000	1,000	0,000	0,000	0,000	0,000	**0,000**

E.1.2 – Matrizes de confusão para oclusão mandibular

Tabela E.3 – Resnet50 sem atenção

	Raiva	Desprezo	Nojo	Medo	Felicidade	Tristeza	Supresa	Neutro
Raiva	**0,000**	0,000	0,000	0,000	0,312	0,312	0,000	0,375
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,286	0,000	**0,000**	0,000	0,214	0,429	0,000	0,071
Medo	0,000	0,000	0,000	**0,000**	0,053	0,895	0,000	0,053
Felicidade	0,067	0,000	0,000	0,000	**0,400**	0,467	0,000	0,067
Tristeza	0,000	0,000	0,000	0,000	0,385	**0,538**	0,000	0,077
Surpresa	0,000	0,000	0,000	0,000	0,062	0,625	**0,250**	0,062
Neutro	0,000	0,000	0,000	0,000	0,182	0,545	0,000	**0,273**

Tabela E.4 – Resnet50 com atenção *SWM*

	Raiva	Desprezo	Nojo	Medo	Felicidade	Tristeza	Surpresa	Neutro
Raiva	**0,062**	0,125	0,000	0,000	0,000	0,125	0,000	0,688
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,357	0,071	**0,071**	0,000	0,000	0,357	0,000	0,143
Medo	0,053	0,000	0,000	**0,053**	0,000	0,895	0,000	0,000
Felicidade	0,133	0,000	0,133	0,000	**0,000**	0,267	0,000	0,467
Tristeza	0,385	0,154	0,000	0,000	0,000	**0,308**	0,000	0,154
Surpresa	0,000	0,062	0,000	0,062	0,000	0,500	**0,062**	0,312
Neutro	0,000	0,273	0,000	0,000	0,000	0,182	0,000	**0,545**

Tabela E.5 – Resnet50 com atenção *CBAM*

	Raiva	Desprezo	Nojo	Medo	Felicidade	Tristeza	Surpresa	Neutro
Raiva	**0,000**	0,312	0,000	0,250	0,000	0,375	0,000	0,062
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,000	0,286	**0,000**	0,000	0,000	0,714	0,000	0,000
Medo	0,000	0,000	0,000	**0,263**	0,000	0,737	0,000	0,000
Felicidade	0,000	0,267	0,067	0,200	**0,133**	0,267	0,067	0,000
Tristeza	0,000	0,231	0,000	0,231	0,000	**0,538**	0,000	0,000
Surpresa	0,000	0,000	0,000	0,188	0,000	0,750	**0,062**	0,000
Neutro	0,000	0,182	0,000	0,455	0,000	0,273	0,000	**0,091**

E.2.1 – Matrizes de confusão para oclusão ocular

Tabela E.6 – Resnet50 sem atenção

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,143**	0,500	0,000	0,000	0,000	0,000	0,000	0,357
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,000	0,400	**0,267**	0,000	0,000	0,000	0,133	0,200
Medo	0,000	0,538	0,000	**0,000**	0,000	0,000	0,154	0,308
Felicidade	0,000	0,500	0,000	0,000	**0,188**	0,000	0,000	0,312
Tristeza	0,111	0,222	0,000	0,000	0,000	**0,000**	0,056	0,611
Surpresa	0,000	0,357	0,000	0,000	0,000	0,000	**0,500**	0,143
Neutro	0,000	0,053	0,000	0,000	0,000	0,000	0,000	**0,947**

Tabela E.7 – Resnet50 com atenção *SWM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,286**	0,286	0,000	0,000	0,000	0,000	0,000	0,429
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,133	0,467	**0,067**	0,000	0,000	0,000	0,000	0,333
Medo	0,077	0,154	0,000	**0,077**	0,000	0,000	0,000	0,692
Felicidade	0,000	0,562	0,000	0,000	**0,375**	0,000	0,000	0,062
Tristeza	0,222	0,278	0,000	0,000	0,000	**0,000**	0,000	0,500
Surpresa	0,000	0,214	0,000	0,000	0,000	0,000	**0,357**	0,429
Neutro	0,000	0,000	0,000	0,000	0,000	0,000	0,000	**1,000**

Tabela E.8 – Resnet50 com atenção *CBAM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,214**	0,071	0,000	0,000	0,143	0,000	0,071	0,500
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,067	0,133	**0,067**	0,4	0,000	0,000	0,200	0,133
Medo	0,077	0,000	0,000	**0,154**	0,000	0,000	0,462	0,308
Felicidade	0,000	0,062	0,000	0,062	**0,625**	0,000	0,000	0,250
Tristeza	0,278	0,056	0,000	0,167	0,000	**0,000**	0,000	0,500
Surpresa	0,000	0,143	0,000	0,000	0,000	0,000	**0,786**	0,071
Neutro	0,053	0,000	0,000	0,158	0,000	0,000	0,000	**0,789**

E.2.2 – Matrizes de confusão para oclusão mandibular

Tabela E.9 – Resnet50 sem atenção

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,000**	0,688	0,000	0,000	0,000	0,000	0,000	0,312
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,286	0,143	**0,000**	0,143	0,214	0,071	0,000	0,143
Medo	0,053	0,053	0,000	**0,632**	0,000	0,053	0,000	0,211
Felicidade	0,067	0,467	0,000	0,000	**0,200**	0,000	0,067	0,200
Tristeza	0,154	0,385	0,000	0,154	0,000	**0,154**	0,000	0,154
Surpresa	0,000	0,250	0,000	0,312	0,062	0,000	**0,062**	0,312
Neutro	0,182	0,273	0,000	0,000	0,000	0,000	0,000	**0,545**

Tabela E.10 – Resnet50 com atenção *SWM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,125**	0,250	0,000	0,000	0,000	0,000	0,000	0,625
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,643	0,000	**0,000**	0,000	0,143	0,071	0,000	0,143
Medo	0,368	0,263	0,000	**0,158**	0,000	0,158	0,000	0,053
Felicidade	0,067	0,400	0,000	0,000	**0,400**	0,000	0,000	0,133
Tristeza	0,462	0,308	0,000	0,000	0,077	**0,077**	0,000	0,077
Surpresa	0,250	0,062	0,000	0,000	0,000	0,000	**0,250**	0,438
Neutro	0,091	0,182	0,000	0,000	0,000	0,000	0,000	**0,727**

Tabela E.11 – Resnet50 com atenção *CBAM*

	Raiva	**Desprezo**	**Nojo**	**Medo**	**Felicidade**	**Tristeza**	**Surpresa**	**Neutro**
Raiva	**0,000**	0,000	0,062	0,000	0,062	0,000	0,000	0,875
Desprezo	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000	0,000
Nojo	0,143	0,000	**0,000**	0,286	0,286	0,143	0,000	0,143
Medo	0,000	0,000	0,000	**0,526**	0,000	0,263	0,053	0,158
Felicidade	0,000	0,067	0,000	0,000	**0,533**	0,000	0,000	0,400
Tristeza	0,077	0,000	0,000	0,231	0,231	**0,154**	0,000	0,308
Surpresa	0,000	0,000	0,000	0,188	0,062	0,000	**0,188**	0,562
Neutro	0,000	0,000	0,000	0,000	0,000	0,000	0,000	**1,000**