UNIVERSIDADE FEDERAL DE GOIÁS
INSITUTO DE INFORMÁTICA

JONES JOSÉ DA SILVA JÚNIOR

Redes Neurais Profundas para Reconhecimento Facial no Contexto de Segurança Pública

Goiânia
2020

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[x] Dissertação [ ] Tese

2. Nome completo do autor

Jones José da Silva Júnior

3. Título do trabalho

Redes Neurais Profundas para Reconhecimento Facial no Contexto de Segurança Pública

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **JONES JOSE DA SILVA JUNIOR**, **Discente**, em 11/08/2020, às 09:34, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Anderson Da Silva Soares**, **Professor do Magistério Superior**, em 11/08/2020, às 09:37, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1480788** e o código CRC **AE0C1300**.

Referência: Processo nº 23070.031507/2020-39 SEI nº 1480788

JONES JOSÉ DA SILVA JÚNIOR

Redes Neurais Profundas para Reconhecimento Facial no Contexto de Segurança Pública

Dissertação apresentada ao Programa de Pós–Graduação do Insituto de Informática da Universidade Federal de Goiás, como requisito para Qualificação do Mestrado em Ciência da Computação.

Área de concentração: Ciência da Computação.

Orientador: Prof. Dr. Anderson da Silva Soares

Co-Orientador: Prof. Dr. Gustavo Teodoro Laureano

Goiânia
2020

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Da Silva Júnior, Jones José
Redes Neurais Profundas para Reconhecimento Facial no Contexto de Segurança Pública [manuscrito] / Jones José Da Silva Júnior. - 2020.
LXXXV, 85 f.

Orientador: Prof. Dr. Anderson Da Silva Soares; co-orientador Dr. Gustavo Teodoro Laureano.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2020.
Bibliografia.
Inclui siglas, gráfico, tabelas, algoritmos, lista de figuras, lista de tabelas.

1. Visão computacional. 2. Reconhecimento Facial. 3. Redes Neurais. I. Da Silva Soares, Anderson, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **17/2020** da sessão de Defesa de Dissertação de **Jones José da Silva Júnior**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos vinte e nove dias do mês de julho de dois mil e vinte, a partir das nove horas, via sistema de webconferência da RNP, realizou-se a sessão pública de Defesa de Dissertação intitulada **"**Redes Neurais Profundas para Reconhecimento Facial no Contexto de Segurança Pública**"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Anderson da Silva Soares (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Wesley Pacheco Calixto (Departamento de Eletrotécnica/IFG), membro titular externo; Professor Doutor Ronaldo Martins da Costa (INF/UFG), membro titular interno. A realização da banca ocorreu per meio de videoconferência, em atendimento à recomendação de suspensão das atividades presenciais na UFG emitida pelo Comitê UFG para o Gerenciamento da Crise COVID-19, bem como à recomendação de isolamento social da Organização Mundial de Saúde e do Ministério da Saúde para enfrentamento da emergência de saúde pública decorrente do novo coronavírus. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Anderson da Silva Soares, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos vinte e nove dias do mês de julho de dois mil e vinte.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Anderson Da Silva Soares**, **Professor do Magistério Superior**, em 29/07/2020, às 11:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Wesley Pacheco Calixto**, **Usuário Externo**, em 29/07/2020, às 11:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Ronaldo Martins Da Costa**, **Professor do Magistério Superior**, em 29/07/2020, às 11:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **JONES JOSE DA SILVA JUNIOR**, **Discente**, em 29/07/2020, às 11:45, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?

Esse trabalho é o resultado de muito esforço, abdicação e dedicação à minha família, assim como a vontade de gerar retorno à sociedade e colaborar de alguma forma para a ciência e ao bem comum.

Dedico este trabalho à minha esposa Leila, companheira de todas as horas e com paciência e sabedoria de sobras pra me ajudar neste período que exigiu tanto de nós.

Dedico também aos meus pais, Maria e Jones, que são os maiores exemplos de dedicação e esforço na educação dos filhos que eu tenho, assim como às minhas irmãs Roberta e Patrícia que me deram toda a força do universo.

Dedico ainda ao meu enteado Guilherme aos meus sobrinhos Lays, Lyan, Pedro, Benício, Felipe, Rafael, Benício, Lys e em especial ao Bernardo, que com apenas 4 anos de idade e na véspera da qualificação deste trabalho demonstrou força imensa para superar um obstáculo que nenhuma criança deveria passar, mas que encarou o desafio de frente como guerreiro que é e graças a ciência e a Deus está cheio de saúde e forte como um dinossauro.

Agradecimentos

Gostaria de agradecer à Universidade Federal de Goiás e ao Instituto de Informática pela oportunidade de estudar uma instituição de tão elevado nível acadêmico e científico. Gostaria de agradecer imensamente ao professor Anderson Soares que me orientou e me guiou neste trabalho. Voltar ao ambiente acadêmico após 11 anos da conclusão da graduação foi um obstáculo gigantesco que foi amenizado com a sua compreensão, conhecimento e motivação na busca de melhores resultados. A sua dedicação tanto no aprimoramento do conhecimento científico quanto na transformação deste conhecimento em resultados para a UFG e para a sociedade é inspirador. Agradeço também o professor Gustavo Teodoro tanto pelos conhecimentos transmitidos de visão computacional quanto pelas explicações e correções visando um melhor rigor de escrita formal do texto.

Gostaria de agradecer também à Polícia Civil do Estado de Goiás, bem como à direção e aos amigos que me apoiaram e me proveram condições para a realização deste trabalho. Nesta instituição conheci pessoas valorosas e honradas que se dedicam com risco à própria vida em nome do bom combate.

Gostaria de agradecer também ao meu grande amigo e Pedro Vitor Quinta de Cadastro. Em vários momentos que imaginei que não conseguiria concluir ele me ajudou, motivou e inspirou a continuar em frente.

Gostaria de agradecer também a todos os colegas do LAB de IA. Poder desfrutar do conhecimento compartilhado por todos durante os seminários foi um privilégio para mim.

Resumo

Silva Junior, Jones J.. **Redes Neurais Profundas para Reconhecimento Facial no Contexto de Segurança Pública**. Goiânia, 2020. 85p. Dissertação de Mestrado. Insituto de Informática, Universidade Federal de Goiás.

O reconhecimento facial é uma importante ferramenta para a segurança pública na repressão ao crime. A capacidade de comparar uma imagem facial de um suspeito filmado em local de crime com um banco de dados de milhões de fotos e assim encontrar a sua verdadeira identidade representa um ganho significativo nas taxas de elucidação de delitos. Mesmo alcançado o estado da arte em *benchmarks* públicos, estes modelos podem não atingir os mesmos resultados no mundo real. O principal motivo é a falta de representatividade na distribuição de dados destas coleção públicas de imagens de treinamento e testes, que resulta em modelos com maior capacidade de reconhecimento em determinados subgrupos demográficos e significativa queda na acurácia em outros sub-grupos, como por exemplo, mulheres afrodescendentes. Apesar do estado da arte apontar para soluções que mapeiam um espaço de características com distância angular esta estratégia desconsidera a hipótese da existência de grupos minoritários importantes no conjunto de dados. Este trabalho visa investigar estratégias de treinamento de ajuste fino para arquiteturas de redes neurais profundas no contexto de reconhecimento facial em segurança pública. O conjunto de dados utilizado inclui uma coleção de imagens de treinamento com o padrão do biotipo facial brasileiro com o objetivo de gerar um modelo com maior acurácia para a investigação policial. Utilizando conjunto de dados privado foi possível aprimorar a acurácia na coleção de teste com amostras representativas do contexto deste trabalho. Além disso, foi possível demonstrar que funções de custo com maior capacidade exploratória dos dados são mais aderentes ao problema de viés de dados no reconhecimento facial. Os resultados demonstram inclusive, que foi possível reduzir substancialmente a diferença das taxas de erro entre os gêneros masculino e feminino, além dos aspectos raciais.

Palavras–chave

Reconhecimento Facial, Redes Neurais, Visão Computacional, Segurança Pública.

Abstract

Silva Junior, Jones J.. **<Deep Neural Networks for Face Recognition in the Context of Public Security>**. Goiânia, 2020. 85p. MSc. Dissertation. Insituto de Informática, Universidade Federal de Goiás.

Face recognition is an important tool for law enforcement. Bein able to compare a face image of a suspect filmed at a crime scene with a database of millions of photos and thus find his true identity represents a significant increase in crime resolution rates. Although this task has been researched since the 1970s, it was with the use of Convolutional Neural Networks (RNCs) from 2014 that a relevant advance was achieved that allowed some to reach 99.63% accuracy in the benchmark *Labeled Faces in the Wild (LFW)*. Despite different architectures and cost functions, a common feature of the papers published since then is the fact that they are trained in a supervised manner, thus requiring large collections of facial images previously labeled. Even state of arts models in public *benchmarks*, they may not achieve the same results in the real world. The main reason is the lack of demographic data distribuition of these public datasets, which results in models with greater accuracy in specific demographic subgroups and worst accuracy in other subgroups, such as afro-descendant women. This work aims to investigate the fine tuning training strategies of deep neural network architectures for facial recognition in public safety context, using a dataset with the Brazilian faces in order to generate a more accurate model for a investigations police department. We managed to improve accuracy on test set with samples representative of the context of this work training a model with private dataset with a very small number of samples compared to the public ones.

Keywords

<Face Recognition, Neural Network, Computer Vision.>

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

1.1 Justificativa

Biometria é uma ciência que estuda as medidas de características físicas ou comportamentais dos seres vivos com o objetivo de identificá-los e individualizá-los [43]. Reconhecimento facial é uma técnica de identificação biométrica que utiliza informações morfológicas e antropométricas extraídas da face humana com o objetivo de estabelecer o grau de similaridade entre imagens faciais diferentes. De acordo com [51], a posição dos olhos, do nariz e da boca assim como as distâncias entre essas características são alguns dos fatores mais comumente considerados.

O reconhecimento facial é uma importante técnica de identificação na segurança pública em geral e especificamente no campo da investigação policial, que é o foco deste trabalho. De acordo com o relatório produzido pelo Instituto IJIS[1] e pela Associação Internacional de Chefes de Polícia[2] [20], o reconhecimento facial é uma ferramenta que ajuda a inocentar pessoas não culpadas e identificar possíveis suspeitos. De acordo com o departamento de polícia *Pinellas*, no estado da Flórida, EUA, desde 2014 foram resolvidos mais de 400 crimes com a utilização de sistema de reconhecimento facial [55]. Isso porque nestas atividades de investigação os policiais rotineiramente se deparam com a necessidade de identificar uma pessoa, suspeito, vítima ou testemunha de algum crime, a partir de uma imagem digital facial que pode ser oriunda de diferentes fontes, como câmeras de circuito fechado de TV, perfis de redes sociais e fotos tiradas por equipes de investigação de campo.

Para atender à essa necessidade dos órgãos policiais o Grupo de Trabalho para Identificação Facial Científica (*Facial Identification Scientific Working Group*) [12] criou e organizou padrões e técnicas para a realização de identificação facial humana forense, incluindo a definição dos pontos característicos faciais necessários para individualização,

[1] Organização não lucrativa para promover a utilização de tecnologia na segurança pública - https://ijis.site-ym.com/default.aspx

[2] International Association of Chiefs of Police - IACP

assim como os cálculos de distâncias proporcionais e ângulos utilizados entre estes pontos. Este trabalho é realizado pelo profissional da área sem a utilização de algoritmos específicos, sendo a marcação dos pontos realizada de forma manual, com apoio de computação gráfica e os cálculos geralmente realizados em planilha. No Brasil este tipo essa análise recebe o nome de Exame Prosopográfico e é realizado pelo Papiloscopista Policial, quadro pertencente geralmente às policiais civis e federal.

Porém, por ser tratar de um conjunto de técnicas com elevado trabalho humano é inviável utilizá-la para comparar uma imagem com um banco de dados com milhares de outras imagens, sendo então utilizada exclusivamente na tarefa de verificação facial, onde o objetivo é analisar e gerar um indicador de similaridade facial entre duas imagens diferentes. Ou seja, aplica-se em situações em que a polícia já dispõe de um suspeito.

Portanto, quando a investigação policial possui apenas a imagem com identidade desconhecida, torna-se necessário compará-la à milhões de fotos de pessoas com identificação conhecida, listando aquelas com maior similaridade facial. Assim, verifica-se a necessidade de automação desse tipo de tarefa, sendo este um problema investigado cientificamente desde os anos 70. Conforme [56] as primeiras pesquisas foram as teses de doutorado de [27] em 1970 e [26] em 1973. Ambos propuseram detectores de borda e contorno para localizar um conjunto de pontos específicos da face e então medir posições relativas e distâncias entre eles. Desde então, diversas técnicas foram propostas tendo como objetivo principal aumentar a acurácia e a robustez dos algoritmos, dado que vários fatores tornam o problema mais complexo, como etnia, iluminação, ângulo, expressão facial, idade, oclusão e acessórios como óculos e chapéus.

Além da necessidade de desenvolver algoritmo de reconhecimento facial para suprir a necessidade demonstrada é importante também investigar o problema do viés de dados no contexto da segurança pública e aplicado à realidade brasileira. As técnicas propostas desde 2015 até as mais recentes, que alcançaram estado da arte neste tipo de tarefa, tem como característica em comum serem baseadas em Redes Neurais treinadas de forma supervisionada em coleções de dados públicas [61]. Como estas coleções possuem distribuição demográfica de dados desigual [40], quando analisados em relação a etnia, idade e gênero, o resultado práticos destes modelos vem apresentando alguns problemas relacionados ao viés de dados, algum com grandes repercussões [2, 40, 71].

No trabalho [2], os autores estudaram duas coleções de dados de teste utilizadas em diversos artigos para comparação de resultados, a *IJB-A* [29] e a *Adience* [9], e descobriram que os dados eram predominantemente de indivíduos de pele clara, 79,6% para o *IJB-A* e 86,2% para *Adience*, de forma que os resultados reportadas nestas coleções não são suficientes para uma avaliação abrangente da acurácia dos modelos. Os

autores avaliaram ainda produtos comerciais da empresas *IBM*[3], *Microsoft*[4] e *Face++*[5], e concluíram enormes disparidades nas taxas de erro quando analisadas por etnia e gênero. Os pesquisadores constataram que para as mulheres afrodescendentes, por exemplo, os sistemas apresentaram taxas de erro de até 35% enquanto que homens com características caucasianas apresentaram menos de 1% de erro.

Considerando o problema de reconhecimento facial aplicado à segurança pública no Brasil, a etnia pode ser uma variável ainda com maior impacto na acurácia dos modelos do que o já demonstrado em [2]. Uma vez que a miscigenação é uma das principais características da população brasileira, é bastante complexo e subjetivo realizar qualquer classificação seguindo esse critério, além de ser um tema bastante polêmico. O próprio IBGE[6] adotou a autoclassificação como critério no censo realizado em 2010 [7]. Nesta pesquisa 47,73% se declararam como pardos, critério que abrangem mulatos, cablocos, cafuzos, mamelucos ou mestiços de negro com pessoa de outra raça.

Conforme exposto, existe forte sugestão de que a avaliação de desempenho de algoritmos de reconhecimento facial sem considerar a particularidade racial brasileira pode ser considerada insuficiente. Não foram encontrados trabalhos publicados analisando os resultados alcançados tanto por soluções comerciais quanto por modelos treinados com coleções de dados públicos em coleções de dados com amostras específicas da população brasileira.

Diante dessas evidências fica demonstrada a necessidade de avaliar e principalmente melhorar o desempenho de modelos treinados em coleções de dados públicas para a realidade brasileira e no contexto de segurança pública. Conforme [61] a forma mais direta de resolver esse problema é a coleta massiva de imagens faciais rotuladas com melhor distribuição demográficas, porém trata-se de uma atividade bastante onerosa de termos financeiros e de tempo. A técnica de transferência de aprendizado [70], ou *Transfer Learning*, tem sido investigada e utilizada para aproveitar as características aprendidas por modelos treinados em um determinado domínio com coleções de dados com grande quantidade de amostras e adaptá-lo para outros domínios quando se dispõe de conjuntos de treinamento com número limitado de amostras.

Especificamente no campo do reconhecimento facial a transferência de conhecimento foi utilizada em [62], onde foi proposta uma arquitetura baseada em adaptação de domínio não supervisionado, *Unsupervised Domain Adaptation - UDA*, para aumentar a capacidade discriminativa de uma rede treinada com coleções de imagens faciais caucasianas quando utilizada em pessoas de diferentes etnias. Para a adaptação do domínio

[3]International Business Machines Corporation
[4]Microsoft Corporation
[5]Face++ AI Open Platform
[6]Instituto Brasileiro de Geografia e Estatística

foi utilizada conjunto de imagens não rotuladas de pessoas de origem africana, indiana e asiática totalizando 500 mil amostras destas etnias.

Porém foi possível estruturar uma coleção de imagens faciais brasileiras com um número de amostras bem menor, totalizando 61 mil imagens, destas apenas 45.000 para treinamento, número 10 vezes menor que o utilizado na proposta anterior e também insuficiente para o treinamento de modelos a partir do zero. Por este motivo este trabalho se propôs a investigar a utilização da função custo *Triplet Loss* proposta no trabalho conhecido como *FaceNet* [50], como método para mitigação do viés de dados, através do treinamento de ajuste fino a partir de um modelo pré-treinado com coleção de dados pública e majoritariamente caucasiana.

Essa escolha se deve pelo fato do treinamento ser realizado através de trios de imagem, os *triplets*, que são compostos por duas imagens de um mesmo indivíduo e a terceira de um outro indivíduo, e que são utilizados pela *Triplet Loss* no sentido de diminuir a distância intraclasse (imagens de mesma identidade) e aumentar a interclasse (imagens de identidade diferentes). A hipótese de trabalho é que esse método de treinamento intrinsecamente explora de maneira combinatória as amostras de forma que ocorra um melhor aproveitamento deste limitado número de imagens no aprendizado de características mais discriminativas do padrão facial brasileiro. Esse fator, somado à geração *on-line* dos *triplets* orientado para a redução do viés de dados hipoteticamente supervisiona o modelo de forma mais eficaz nesta tarefa que as muitas das funções custo propostas posteriormente, que embora tenham alcançado o estado da arte são variações modernas da *Softmax Loss*[61] e portanto treinadas como um classificador.

1.2 Objetivos

Os objetivos deste trabalho são avaliar as arquiteturas baseadas em Redes Neurais Convolucionais para aprimorar o desempenho do reconhecimento facial em coleção de dados de fotos de brasileiros em um contexto de segurança pública. Nas seções a seguir serão definidos os objetivos geral e específicos deste trabalho.

1.2.1 Objetivo Geral

O objetivo principal desta dissertação é verificar, através de treinamento em modo ajuste fino, que funções de custo que visam a exploração dos dados, como é o caso da *Triplet Loss*, são mais adequadas para tratar o problema de viés de dados para reconhecimento facial. Pretende-se mostrar que a hipótese é valida a partir de experimentos com Redes Neurais Convolucionais pré-treinadas em coleções de dados

públicas e que posteriormente são re-treinadas com ajuste fino em um banco de dados com baixo número de exemplos.

1.2.2 Objetivos Específicos

- Coletar, estruturar e limpar coleção de dados de faces brasileiras para avaliar e treinar arquiteturas de Redes Neurais Profundas.
- Realizar estudo acerca das arquiteturas de Redes Neurais Profundas para reconhecimento facial, com o objetivo de entender as principais características das funções custo utilizadas e que alcançaram o estado da arte.
- Identificar as funções custo com melhor desempenho e que sejam conceitualmente mais diferentes entre si para utilização neste trabalho.
- Realizar o treinamento de ajuste fino (*fine tuning*) a partir de modelos pré-treinados que já tenham atingido acurácia compatível com o estado da arte, utilizando a coleção de dados de imagens brasileiras.
- Utilizando o modelo pré-treinado, identificar os indivíduos que possuem a maior chance de erro na comparação facial no conjunto de imagens brasileiras. Estes serão utilizados para comparação dos resultados antes e após o treinamento e irão compor a coleção de dados de teste.
- Avaliar os resultados antes e após o treinamento de ajuste fino, com foco em verificar se realmente ocorreu aumento na acurácia no conjunto de testes de imagens brasileiras, evidenciando a existência de viés de dados relacionada às faces brasileiras.
- Analisar os resultados obtidos, antes e após o ajuste fino, considerando os gêneros feminino e masculino.

1.3 Hipóteses de Pesquisa

As Redes Neurais Profundas, treinadas de forma supervisionada, alcançaram o estado da arte em reconhecimento facial, inclusive superando a capacidade humana [61]. Porém o viés de dados, que contribui para a depreciação da acurácia em pessoas com características faciais não representadas de forma igualitária, vem sendo constantemente analisado e demonstrado [61, 42, 46, 38, 40]. Devido a este fator os modelos utilizando essa abordagem vem apresentando taxas de erro maiores quando avaliadas em determinados subgrupos, como mulheres negras.

Com o intuito de contribuir com as pesquisas de avaliação e diminuição do viés de dados em Reconhecimento Facial, este trabalho busca entender como este fator pode afetar o desempenho de modelos treinados em coleções de dados públicos no contexto da

segurança pública brasileira, além de propor o treinamento de ajuste fino utilizando *Triplet Loss* como forma de melhor aproveitar a coleção de treino estruturada neste trabalho, que possui como característica limitado número de amostras e também pequena quantidade de amostras por classe (ou indivíduo).

Hipótese Primária: É possível melhorar a acurácia de modelos de Reconhecimento Facial para a aplicação em imagens faciais brasileiras realizando treinamento de ajuste fino em modelos pré-treinados em coleções de dados públicas mesmo que já tenham alcançado o estado da arte.

Hipótese Secundária: O número total de amostras e a quantidade de amostras por identidade afeta o desempenho de modelos treinados com as diferentes funções custo que alcançaram resultados no estado da arte. No cenário deste trabalho, onde ambos quantitativos são bem menores em comparação às coleções de dados públicas, a *Triplet Loss*, por sua característica combinatória de amostras, em conjunto com a geração *on-line* de tiplets, possui maior capacidade de mitigação do viés de dados.

1.4 Experimentação Prática

Durante todo o período deste trabalho modelos foram treinados com diferentes funções custo e coleções de dados com o objetivo de adquirir conhecimento das ferramentas utilizadas, principalmente o *Tensorflow*, aprimorar os conhecimentos teóricos sobre inteligência artificial e visão computacional adquiridos nas disciplinas e entender bem a influência dos hiper parâmetros tanto na convergência do modelo quanto na mitigação do *overfitting*.

Além disso foi desenvolvida uma interface *web* e uma camada de *backend* para permitir a utilização de alguns destes modelos em situações reais. O resultado dessa aplicação, embora não utilizando o modelo com maior acurácia resultado deste trabalho, gerou resultados muito importantes para a segurança pública. Atualmente são realizadas em média 40 pesquisas por dia com imagens oriundas de diversas fontes e situações, como redes sociais com perfis falso e imagens de câmeras de local de crime. Diversos foragidos foram recapturados tentando enganar as autoridades com uso de documentos falsos, muitos deles com mandado de prisão em aberto em outros estados. Foi possível contabilizar no mínimo 170 casos onde a utilização do reconhecimento facial foi primordial para o trabalho policial, sendo que atualmente órgãos policiais de todo o Brasil tem enviado solicitações de pesquisa muitos vezes com resultados positivos, evidenciando a importância deste trabalho para o meio da segurança pública.

1.5 Organização da Dissertação

O restante da dissertação será organizado da seguinte forma:

- o Capítulo 2 trata da fundamentação teórica do trabalho, abordando as etapas de um sistema de reconhecimento facial; conceitos teóricos e fundamentos das redes neurais; trabalhos que investigam e quantificam o viés de dados
- o Capítulo 3 descreve os métodos utilizados nas etapas anteriores à comparação facial que não o foco deste trabalho; os conceitos e definições matemáticas das funções custo empregados; estrutura e informações sobre as coleções de dados; métricas de avaliação;
- o Capítulo 4 apresenta os resultados obtidos;

CAPÍTULO 2

Fundamentação Teórica

Este capítulo apresenta a fundamentação teórica utilizada neste trabalho. Inicialmente descreve-se com maior detalhamento a técnica chamada *eigenfaces* já mencionada na Seção 2.1. Então, são apresentadas as etapas de reconhecimento facial que geralmente são utilizadas na maioria dos trabalhos. Posteriormente descreve-se a fundamentação de redes neurais convolucionais profundas. Por fim, descreve-se o problema do viés de dados para o problema de reconhecimento biométrico facial.

2.1 Trabalhos relacionados

Dentre as técnicas de algoritmos de reconhecimento facial podemos citar aquelas que usam *eigenfaces* e redes neurais profundas. As *eigenfaces* foram inicialmente descritas em [57] e posteriormente evoluídas em outros trabalhos. Conforme [61] esta técnica representou um marco histórico nessa área em função dos resultados alcançados. Posteriormente surgiram técnicas apoiadas nas redes neurais profundas que alcançaram o estado da arte atingindo a performance humana com a *DeepFace* [68], posteriormente sendo superada pela Facenet [50].

Em [57] a proposta se baseia na ideia de que existem características discriminativas que podem, ou não, serem diretamente relacionadas àquelas que intuitivamente os indivíduos utilizam para identificar e comparar pessoas, como nariz, olhos e boca. Tais características relevantes, e por vezes latentes, podem ser extraídas em um processo onde uma dada face é codificada da forma mais eficiente possível e comparada com todo o conjunto de faces codificadas exatamente da mesma maneira. Essa codificação tem como objetivo revelar apenas as variações entre as faces e que, portanto, diferenciam um indivíduo de todos os outros.

Conforme [57], em termos matemáticos, as características que representam essas variações podem ser capturadas através dos autovetores da matriz de covariância de uma coleção de imagens. Assim, cada autovetor é considerado uma característica diferente, que representa uma determinada quantidade de variação. Cada face contribui de forma diferente para cada autovetor, de forma que cada um pode ser visualizada como uma

face genérica, denominada *eigenface* e cada indivíduo pode ser representado como uma combinação linear das *eigenfaces*.

Autovetores são tipos especiais de vetores que são associados a uma transformação linear de uma matriz A, de tal forma que:

$$Ax = \lambda x \tag{2-1}$$

onde x é um autovetor da matriz A e o λ é um escalar não nulo denominado autovalor. Os autovetores são ortogonais entre si e possuem, como uma das suas propriedades, a capacidade de mensurar a variância de um conjunto de dados representado pela matriz A. O autovalor associado ao autovetor pode ser interpretado como a quantidade de variância que este autovetor captura.

Para encontrar as *eigenfaces* é necessário encontrar a matriz de covariância C. Por sua vez, para calculá-la, é necessário calcular a face média de um conjunto de treinamento e a diferença entre cada face e a face média. Seja $\Gamma_1, \Gamma_2, ...\Gamma_m$ as M amostras que compõem o conjunto de treinamento, então a face média é calculada através do somatório destas amostras divida pela quantidade 2-2. A diferença Φ entre cada face e a média é obtida através de subtração de cada amostra pela média Ψ 2-3:

$$\Psi = \frac{1}{M}\sum_{n=1}^{M}\Gamma_n \tag{2-2}$$

$$\Phi = \Gamma_n - \Psi \tag{2-3}$$

Em sequência é aplicada a Análise dos Componentes Principais[1] (*Pincipal Component Analysis*), ou ACP, de forma que sejam obtidos os vetores ortogonais, ou *eigenfaces* (autovetores), a partir da matriz de covariância C. Em 2-4 essa matriz é obtida pelo somatório da multiplicação entre a matriz Φ da diferença entre cada face média e a sua transposta Φ^T . Então, os M *eigenfaces* u_n, que melhor descrevem a distribuição dos dados, são obtidos em 2-5:

$$C = \frac{1}{M}\sum_{n=1}^{M}\Phi_n\Phi_n^T \tag{2-4}$$

$$\lambda_k = \frac{1}{M}\sum_{n=1}^{M}(u_k^T\phi_n)^2 \tag{2-5}$$

sendo λ_k o autovalor correspondente do autovetor u_k.

[1] Método matemático que, utilizando transformação linear ortogonal, realiza a redução de dimensionalidade de dados complexos e assim permite identificar padrões nos dados e expressá-los de forma que suas semelhanças e diferenças sejam destacadas.

A aplicação das *eigenfaces* para reconhecimento facial pode ser divido em duas fases distintas: o treinamento e a pesquisa.

A fase de treinamento é definida pela geração dos *eigenfaces* com base em uma coleção de imagens rotuladas por indivíduo. Pode ser resumida da seguinte forma:

1. Calcular as *eigenfaces* a partir da coleção de imagens de treinamento e manter aquelas M características que possuem os maiores autovalores. Essa limitação tem como objetivo garantir que apenas as características com maior variância, e portanto mais discriminativas, serão utilizadas. Esse conjunto de *eigenfaces* é denominado espaço de faces, ou *face space*.
2. Calcular, para cada indivíduo, através de suas imagens de amostra, a sua melhor representação no espaço de faces. Ou seja, encontrar a melhor combinação linear entre as *eigenfaces* selecionadas no passo anterior, gerando ao final um vetor de pesos W que melhor representa o indivíduo. Este cálculo é feito através da projeção de cada amostra no referido espaço de faces.

Uma vez realizado o treinamento, imagens faciais de identidade desconhecida podem ser pesquisadas de forma que o indivíduo com maior similaridade facial pode ser encontrado. Esta fase pode ser resumida conforme segue:

1. Para a face a ser pesquisada, encontrar o vetor de pesos W_x no espaço de faces.
2. Comparar os pesos W_x da face ignorada a cada vetor de pesos W de cada uma das amostras utilizadas na fase de treinamento através de cálculo de distância vetorial. Aquele indivíduo cujo vetor W tiver a menor distância em relação à W_x é considerado com maior semelhança, sendo possível atribuir a sua identidade à imagem desconhecida. O cálculo pode ser realizado através de métodos conhecidos de distância entre vetores, dentre eles a distância euclidiana[2]

Este método representou um grande passo nas tarefas de reconhecimento facial. Entretanto, de acordo com [61], os trabalhos baseados neste conceito não ultrapassaram 60% de acurácia. Isso se deve principalmente ao baixo desempenho das *eigenfaces* quando empregadas em imagens com maior grau de variações como iluminação, escala, pose, expressões faciais e oclusão.

Com os expressivos resultados alcançados pelas Redes Neurais Convolucionais, ou RNC, na área de visão computacional, as pesquisas de reconhecimento facial se voltaram para essa técnica, em especial após o lançamento da AlexNet, que venceu a competição ImageNet em 2012 [61] com diminuição da taxa de erro em 10,8%. Através

[2]Distância geométrica entre dois pontos em um espaço multidimensional dada pelo segmento de reta que os liga.

de uma sequência de camadas contendo filtros convolucionais que aprendem a realizar detecção e reconhecimento de características em níveis cada vez mais abstratos, essa rede alcançou o estado da arte nas tarefas de detecção de objetos e classificação de imagens.

Ainda de acordo com [61], essas arquiteturas baseadas em RNC também demonstraram capacidade de atingir o estado da arte em reconhecimento facial. Porém, estes resultados foram alcançados com funções custo específicas para esta tarefa. A utilização da função custo *softmax loss*, tipicamente utilizada nas tarefas de classificação, não apresentou resultados satisfatórios no protocolo de testes *open-set* [61, 35], uma vez que os vetores de características faciais extraídos não são suficientemente discriminativos.

Como exemplo do problema anteriormente relatado, a Figura 2.1 exibe os pontos, em um sistema de coordenadas, correspondentes aos vetores de características faciais de 2 dimensões, extraídos a partir de amostras de dois indivíduos diferentes, identificados nas cores amarelo e violeta. Cada eixo mapeia os valores de cada uma das duas dimensões destes vetores. Nesta figura pode-se observar que o vetor facial referente a amostra **A2**, do indivíduo representado pela cor violeta, está espacialmente mais próximo da amostra **A1**, da pessoa em amarelo, que outra amostra **D1** desta mesma pessoa. Portanto, embora as arquiteturas baseadas em RNC utilizadas em problemas gerais de visão computacional tenham mostrado bons resultados ficou demonstrado a necessidade de especializar a forma de supervisão para o Reconhecimento Facial.

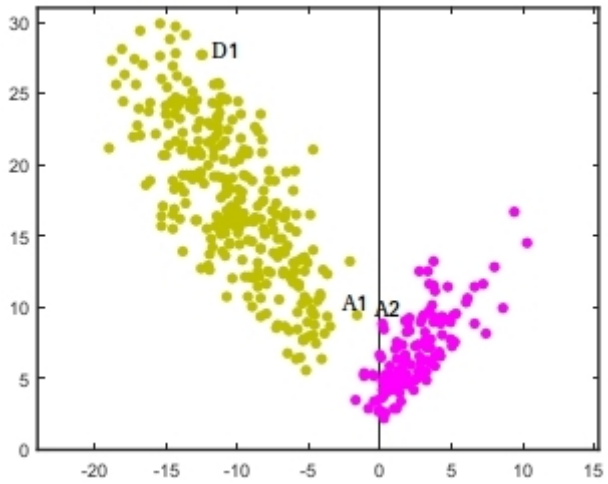


Figura 2.1: Imagem traduzida e adaptada de [35]. Cada ponto é a localização de um vetor de características de 2 dimensões em um sistema de coordenadas de dois indivíduos simbolizados pela cor amarela e violeta. Extraídas de uma RNC treinada com supervisão de *softmax loss* em um experimento de testes.

Assim, o desenvolvimento e o aprimoramento funções custo ganhou bastante importância e atenção, resultando em diversos trabalhos publicados [50, 35, 68, 52, 8, 47, 59, 58, 60, 36] . Novas técnicas foram propostas com o objetivo de aprimorar a capacidade de generalização dos modelos e o estado da arte foi aperfeiçoado em alguns trabalhos [61].

Em outros, a acurácia se manteve no mesmo patamar, porém com menor complexidade de treinamento e utilizando coleções de dados menores.

Uma dessas técnicas é denominada FaceNet [50], trabalho publicado em 2014 que atingiu o estado da arte, com 99,63% no dataset LFW [19] (Labeled Faces in the Wild). Trata-se de uma proposta baseada em um modelo que aprende a mapear diretamente imagens faciais para um espaço euclidiano, onde as distâncias correspondem diretamente ao grau de similaridade entre as faces.

O método consiste em treinar um modelo com arquitetura baseada em RNC de forma supervisionada com o objetivo de que a sua saída seja uma representação vetorial compacta, a partir de uma imagem facial de entrada de alta dimensionalidade. Esta representação compacta, chamada de vetor de características faciais (*face embedding*), é gerada de tal forma que os vetores representativos das imagens faciais da mesma pessoa tenham distâncias pequenas entre si e de pessoas distintas tenham distâncias maiores, dentro de um espaço euclidiano de dimensão n. Uma vez que o vetor facial tenha sido produzido, as tarefas posteriores se tornam mais objetivas: a verificação facial consiste em encontrar o limiar que separa as distâncias intraclasse das interclasses e a tarefa de identificação se torna um problema de vizinhos mais próximos.

Os resultados obtidos pela *FaceNet* foram alcançados posteriormente com a utilização de novas arquiteturas, funções perda e coleções de dados para treinamento. A Tabela 2.1 contém os dados compilados em [61]. Essa tabela relaciona o nome do método, o ano de publicação, a arquitetura, o número de redes, a coleção de dados de treinamento e acurácia na coleção *LFW*.

Tabela 2.1: Resultados alcançados por RNCs utilizando as principais funções perdas de acordo com [61].

Método	Ano	Perda	Arquitetura	N. de Redes	Coleção de Treinamento	Acurácia±Std(%)
DeepFace	2014	softmax	Alexnet	3	Facebook (4.4M,4k)	97,35±0,25
DeepID2	2014	constrative loss	Alexnet	25	CelebFace+ (0.2M,10k)	99,15±0,13
DeepID3	2015	constrative loss	VGGNet-10	50	CelebFace+ (0.2M,10k)	99,53±0,10
FaceNet	2015	triplet loss	GoogleNet-24	1	Google (200M,8M)	99,63±0,09
Baidu	2015	triplet loss	CNN-9	10	Baidu (1.2M,18K)	99,77
VGGFace	2015	triplet loss	VGGNet-16	1	VGGface (2.6M,2.6K)	98,95
light-CNN	2015	softmax	light CNN	1	MS-Celeb-1M (8.4M,100K)	98,8
Center Loss	2016	center loss	Lenet+-7	1	CASIA-WebFace, CACD2000, Celebrity+ (0.7M,17K)	99,28
L-softmax	2016	L-softmax	VGGNet-18	1	CASIA-WebFace (0.49M,10K)	98,71
Range Loss	2016	range loss	VGGNet-16	1	MS-Celeb-1M ,CASIA-WebFace (5M,100K)	99,52
L2-softmax	2017	L2-softmax	ResNet-101	1	MS-Celeb-1M (3,7M,58K)	99,78
Normface	2017	constrative loss	ResNet-28	1	CASIA-WebFace (0.49M,10K)	99,19
CoCo loss	2017	CoCo loss	-	1	MS-Celeb-1M (3M,80K)	99,86
vMF loss	2017	vMV loss	ResNet-27	1	MS-Celeb-1M (4.6M,60K)	99,58
Marginal Loss	2017	marginal loss	ResNet-27	1	MS-Celeb-1M (4.6M,60K)	99,48
SphereFace	2017	A-softmax	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99,42
CCL	2018	center invariant loss	ResNet-27	1	CASIA-WebFace (0.49M,10K)	99,12
AMS loss	2018	AMS loss	ResNet-20	1	CASIA-WebFace (0.49M,10K)	99,12
Cosface	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99,33
Arcface	2018	arcface	ResNet-100	1	MS-Celeb-1M (3.8M,85K)	99,83
Ring loss	2018	Ring loss	ResNet-64	1	MS-Celeb-1M (3.5M,31K)	99,50
AM-Softmax	2018	AM-Softmax	ResNet-20	1	CASIA-WebFace (0.49M,10K)	99,12

Conforme pode ser visto na Tabela 2.1 outros trabalhos atingiram resultados similares, ou até melhores, em comparação à *Facenet* [50] utilizando coleções de treina-

mento bem menores.

2.2 Etapas para Reconhecimento Facial

De acordo com [56], sistemas de reconhecimento facial geralmente são divididos nas etapas: detecção e alinhamento facial, extração de características e comparação, conforme segue:

$$C[R(A(D(I_i))), R(A(D(I_j)))] \tag{2-6}$$

onde I_i e I_j são duas imagens sendo comparadas; D simboliza a etapa de detecção facial, onde a localização exata da face na imagem é realizada; A significa o alinhamento facial, onde pontos característicos da face são encontrados para ajustes posteriores; R é a etapa de extração de representação facial, responsável por codificar as imagens faciais alinhadas na etapa A em um vetor de menor dimensão, compacto, mas que possui características únicas de cada indivíduo; C é a etapa onde alguma técnica de comparação das faces codificadas em vetores é aplicada com o objetivo de medir o grau de similaridade facial entre I_i e I_j.

A etapa A se inicia com a detecção facial, D, que tem como objetivo encontrar em uma imagem qualquer a localização de uma ou mais faces. As saídas dessa etapa são as coordenadas exatas, considerando a imagem como uma matriz de duas dimensões, de um retângulo delimitador que abrange exatamente a região de cada uma das faces na imagem, conforme observado na Figura 2.2(a).

Após a detecção da face é realizado o alinhamento facial, etapa em que são realizados o recorte das faces encontradas, bem como o redimensionamento de forma que todas fiquem na mesma escala. Essas tarefas tipicamente são realizadas utilizando como referência pontos faciais fixos, ou *landmarks*, localizados pelo sistema usando um detector próprio para essa finalidade. Em alguns casos também é realizada nessa etapa alguma correção de ângulo de forma que os olhos fiquem horizontalmente alinhados. Na Figura 2.2(b) podem ser observados os *landmarks* encontrados e na Figura 2.2(c) é exibido um exemplo de face já recortada e redimensionada.

(a) *Retângulo delimitador* (b) *Landmarks* (c) *Alinhamento*

Figura 2.2: (a) Retângulo delimitador encontrado por um detector facial. (b) Landmarks utilizados para alinhamento. (c) Face alinhada

Na etapa de extração da representação, a imagem facial, alinhada e recortada, é transformada em um vetor de características compacto e discriminativo. O termo compacto significa que o vetor possui dimensionalidade muito menor do que a entrada conforme Figura 2.3 . Ou seja, se a face recortada e alinhada possui dimensão $\mathbb{R}^{n \times m}$, então o vetor facial possui dimensão $\mathbb{R}^{d}$ tal que $n \times m >> d$. E o termo discriminativo significa que a relação entre estes vetores mantêm forte correspondência com a similaridade facial das imagens faciais de entrada.

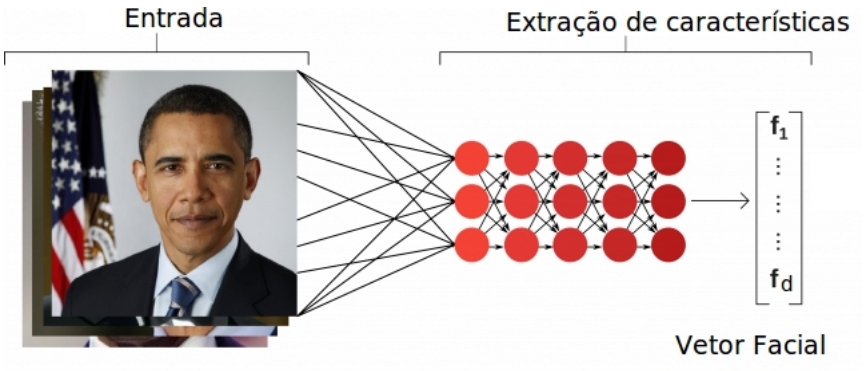


Figura 2.3: Extração de representação facial traduzido e adaptado de [39]

Após a extração dos vetores de características representativos das imagens I_i e I_j o cálculo de correspondência entre os mesmos é realizado com o objetivo de encontrar alguma pontuação, indicador ou probabilidade que indiquem se estas imagens são do mesmo indivíduo ou de indivíduos diferentes. A métrica de correspondência utilizada deve ser congruente com a técnica empregada na etapa de extração dos vetores faciais, considerando os critérios que foram utilizados para extrair as características discriminativas.

2.3 Redes Neurais

2.3.1 Neurônio Artificial

O neurônio artificial utilizado hoje na maioria das arquiteturas de redes neurais é baseado nos modelos idealizados em [63] e em [48], sendo popularmente conhecido como *Perceptron* [44]. A principal característica do neurônio artificial é a sua capacidade de ser treinado, a partir de um conjunto de amostras previamente rotuladas, de forma que a sua configuração interna ao final seja capaz de gerar determinada resposta consistente com o objetivo atribuído a ele. Cada amostra é composta por conjunto de sinais, ou características, que são informações representativas e relevantes do comportamento do processo a ser mapeado.

Trata-se de um modelo computacional inspirado no neurônio biológico, onde uma unidade de processamento irá computar uma saída em função de uma ou mais entradas. A Figura 2.4 apresenta esquematicamente este modelo em que uma entrada x_n, constituída por n características, será ponderada pelo conjunto de pesos w_i e o resultado será somado de forma a produzir a saída u.

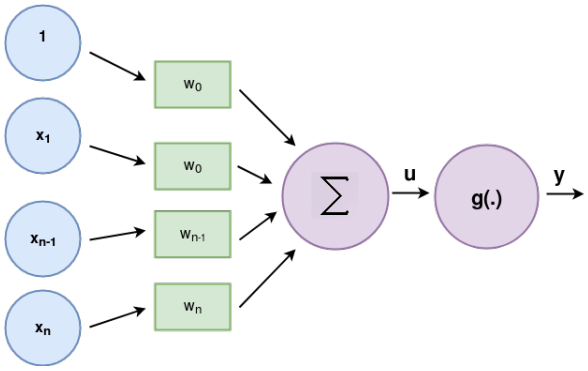


Figura 2.4: Representação gráfica do modelo do neurônio artificial. Cada característica x_n da entrada é multiplicada por um peso w_n, que tem como objetivo ponderar a influência da característica no cálculo da resposta final .

Em seguida o valor de u é utilizado como entrada em uma função de ativação $g(.)$, cujo resultado é a saída y. Em definição matemática, este modelo do *Perceptron* pode ser expressado da seguinte forma:

$$\begin{aligned} u &= \sum_{t=0}^{n} w_i \cdot x_i \\ y &= g(u) \end{aligned} \tag{2-7}$$

Neste ponto é importante destacar que os valores dos pesos w_i ponderam as características de cada uma das entradas x_n refletindo a relevância e a importância de cada uma destas características para a composição da saída.

A utilização de um único neurônio possui como principal limitação a capacidade de modelar, no caso de classificação, apenas problemas em que as classes são apenas duas e linearmente separáveis, conforme [41] e [13]. Por isso os neurônios são utilizados agrupados em camadas e organizados em uma rede denominada *Redes Perceptron de Multicamadas*, ou RPM. Outro benefício dessa abordagem é a capacidade de aprender um espaço de características diferentes, com maior grau abstração [13]. Uma RMP é constituída em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída e seu esquema pode ser visualizado na Figura 2.5.

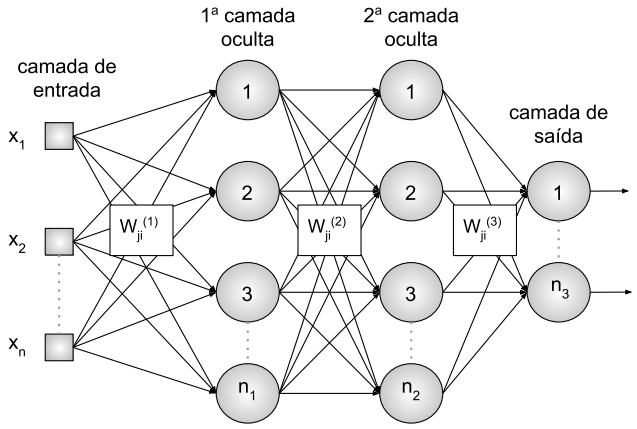


Figura 2.5: A RPM possui as seguintes camada de entrada, uma ou mais camadas ocultas e a saída. As camadas ocultas e a de saída possuem como entrada a saída da anterior.

2.3.2 Redes Neurais Convolucionais

Conforme relatado em [10], as Redes Neurais Convolucionais, também conhecidas por RNC (CNN - Convolutional Neural Networks), foram inicialmente propostas por [32] em um trabalho sobre reconhecimento de códigos de endereço ZIP escritos a mão. Posteriormente outros artigos utilizando RNC foram publicados com objetivos diferentes, mas ainda dentro do campo da visão computacional, como o reconhecimento de dígitos também escritos a mão em 1998 [33] utilizando uma coleção de dados denominada MNIST [31].

Ainda de acordo com [10] as RNCs foram aplicadas e avaliadas em trabalhos científicos nos anos 1990's e 2000's, principalmente em tarefas de classificação, porém sem grande sucesso. Nesse período outras técnicas como Máquina de Vetores de Suporte (*Support Vector Machine* - SVM) e modelos Bayesianos apresentaram melhores resultados nestes tipos de tarefas. Um motivo importante para isso foi o fato de que nesse período as coleções de dados eram insuficientes para treinar redes neurais com camadas profundas com milhões de parâmetros, enquanto estas outras técnicas, com muito menos parâmetros, apresentaram melhores resultados com conjuntos reduzidos de dados.

Entretanto, com o surgimento de coleções de dados muito maiores, tornou-se possível o treinamento de RNCs com maior acurácia, sendo considerado como marco o trabalho de [30], em que a arquitetura proposta superou substancialmente os resultados dos concorrentes na Competição de Reconhecimento Visual de Grande Escala - ImageNet (ImageNet Large Scale Visual Recognition Competition - ILSVRC) [1].

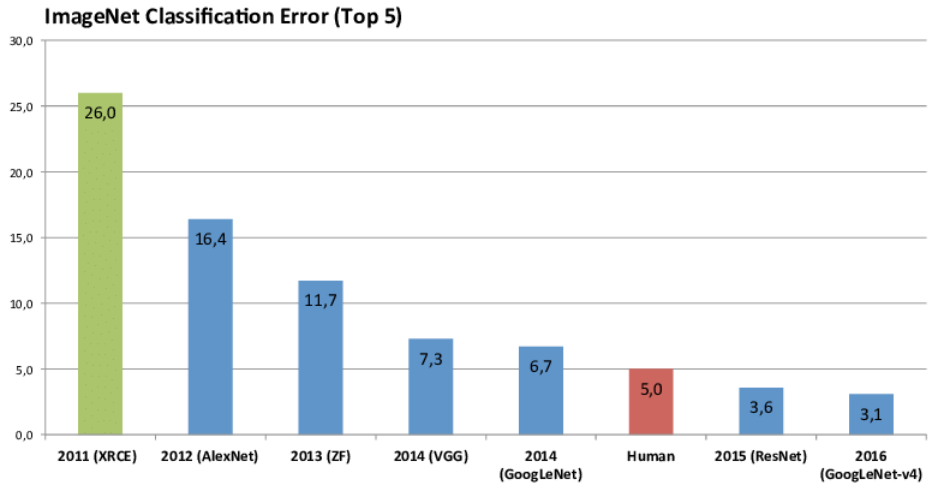


Figura 2.6: Evolução dos resultados no desafio ImageNet [22] onde as Redes Neurais Convolucionais foram usadas pela primeira vez em 2012 com a arquitetura denominada AlexNet e apresentaram uma grande evolução na diminuição do erro.

Além das coleções coleções de dados com maior número de amostras no total e por classe, também surgiram processadores com maior capacidade computacional e memória, em destaque as Unidades de Processamento Gráficos, ou GPUs, que geraram significativo avanço no desempenho de treinamento de RNCs.

Conforme detalhado em [13] as RNCs são tipos especializados de Redes Neurais para processamento de dados com topologia matricial, como imagens, que podem ser pensadas como matrizes de pixels com duas dimensões. O termo "convolucional" indica que este tipo de rede emprega a operação matemática denominada convolução, simbolizada por um asterisco conforme 2-8 . Pode ser compreendida como uma operação entre duas funções $x(t)$ e $h(t)$ resultando uma terceira $y(t)$, que por sua expressa como a função $h(t)$, também conhecida como *filtro*, ou *kernel*, afeta a entrada $x(t)$.

$$x(t) * h(t) = y(t) \tag{2-8}$$

O *kernel* $h(t)$ exerce o papel de filtro, tipicamente possui dimensões menores que a entrada $x(t)$ e é deslizado sobre esta, de forma que a saída $y(t)$ possui valores de maior magnitude nas regiões onde as características representadas pelo filtro estão presentes.

O deslocamento do filtro sobre a matriz de entrada é realizado no eixo horizontal e o tamanho do passo é determinado por um parâmetro chamado *stride*, como pode ser

observado na 2.7. Para cada posição do filtro a soma dos produtos entre o filtro e a matriz correspondente de entrada é obtido e armazenado na saída, que também é denominada mapa de características (*feature map*).

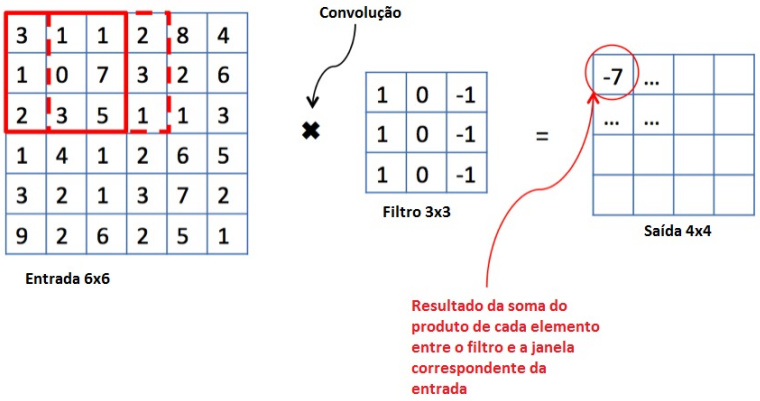


Figura 2.7: Exemplo de convolução adaptado de [5]. O deslocamento do filtro utilizando *stride* igual a um pode ser visualizado na matriz de entrada, demonstrado pelo retângulo sólido vermelho, que representa o filtro na posição inicial. O retângulo vermelho tracejado representa a posição seguinte.

Assim como em 2-7 após a operação de convolução também é utiliza função de ativação, que tem como objetivo adicionar não-linearidade à saída, de forma que problemas mais complexos possam ser modelados. A função de ativação mais utilizada em CNNs é denominada Unidade de Retificação Linear, ou *ReLU* (*Rectfied Linear Unit* e é definida pela fórmula 2-9. A Figura 2.8 retrata o comportamento desta função. Segundo [13], embora a aplicação da *ReLU* produza uma transformação não linear, a função permanece muito próximo da linearidade, preservando algumas características que tornam modelos lineares mais fáceis de otimizar.

$$g(t) = \max(0, t) \tag{2-9}$$

Em arquiteturas de RNCs, tipicamente após a camada convolucional e a linearidade, uma camada de *pooling* é utilizada. Esta camada tem como objetivo substituir a informação, em um certo local do mapa de ativação, por um resumo estatístico considerando informações arredores, propagando então a saída resumida para as próximas camadas. Os benefícios são:

- Maior desempenho computacional, uma vez que o *pooling* ocasiona em redução da dimensão espacial e portanto menos informações para o processamento das camadas posteriores.
- Invariância a translação, uma vez que pequenas mudanças na disposição espacial dos dados não resulta em mudança na saída do *pooling*.

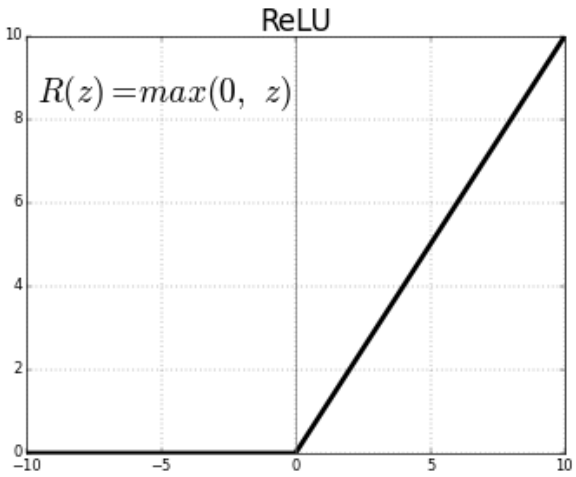


Figura 2.8: Imagem exemplificativa da função *ReLU* obtida e adaptada de [34].

- Preserva as características dominantes das entradas, eliminando assim as informações menos relevantes.

A camada de *pooling*, assim como na convolução, também é aplicada através de uma janela que se desloca por toda a entrada, sendo a operação realizada na seção correspondente conforme 2.9.

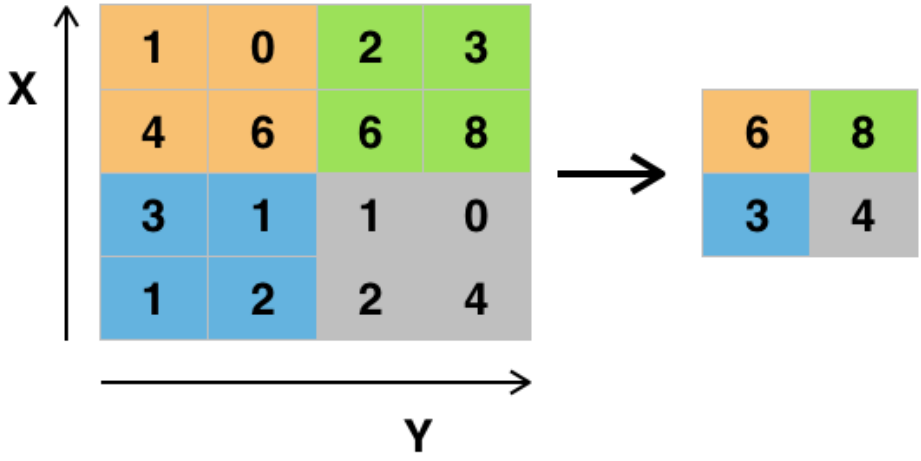


Figura 2.9: Neste exemplo adaptado de [24] a operação MAX é utilizada. Um filtro de tamanho 2x2 é aplicado a toda a entrada preservando apenas o maior valor.

O tipo mais mais comum de operação é o *max-pooling*, onde, a cada posição da janela é preservado o maior valor da seção correspondente.

2.3.3 Treinamento

Para que o neurônio artificial seja capaz de modelar um determinado comportamento ele precisa ser treinado de acordo com o objetivo funcional. O treinamento consiste

em ajustar os pesos w_i até que a saída estimada seja próxima à saída real relacionada à entrada correspondente. Existem diversas técnicas de treinamento de redes de neurônios artificiais, porém a descrita nesta dissertação é conhecida por treinamento supervisionado com retropropagação, ou *backpropagation*, por ser esta a técnica utilizada neste trabalho.

Conforme [44] o processo de treinamento de RMPs é realizado através da execução de duas etapas consecutivas por um número determinado de iterações. A primeira é denominada propagação adiante (*feedforward*) e nesta as n características de cada amostra são aplicadas à camada de entrada x_n e propagadas através das camadas ocultas até a camada de saída. As saídas de cada uma das camadas, que são utilizadas como entradas da camada posterior, são obtidas através das equações definidas em 2-7. O objetivo dessa etapa é calcular os valores de saída da rede dados os valores dos pesos sinápticos w_i.

Após a obtenção da resposta da rede as saídas computadas são comparadas às respectivas respostas esperadas e então a diferença entre elas, denominada erro, é calculada. Este erro então será utilizado para ajustar os pesos sinápticos w_i na etapa posterior, denominada propagação reversa (*backward*). Estas duas etapas sequenciais se repetem até que que o erro seja mínimo. Á cada iteração, com a execução completa da propagação adiante e da propagação reversa dá-se o nome de época.

O erro é calculado através de uma função que seja consistente com o objetivo funcional do treinamento e é denominada função custo. A incumbência dela é medir o desvio entre as respostas computadas pelo modelo em relação às respostas esperadas. Modelos treinados com objetivos diferentes utilizarão funções diferentes. Uma função custo muito utilizada é a função erro quadrático e é dada pela expressão 2-10:

$$E(k) = \frac{1}{2}\sum_{j=1}^{n}(d(k) - y(k))^2 \tag{2-10}$$

onde $y(k)$ é o valor produzido por cada um dos n neurônios da camada de saída ao aplicar a k-ésima amostra à camada de entrada, enquanto $d(k)$ é o valor esperado.

Assumindo que o conjunto de treinamento é constituído de p amostras, o erro total de cada época é calculado através do erro quadrático médio definido em 2-11:

$$E(m) = \frac{1}{p}\sum_{k=1}^{p} E(k) \tag{2-11}$$

onde $E(m)$ é o erro quadrático obtido.

Computado o erro gerado pelo modelo é realizado então o ajuste dos pesos, processo que é realizado por um algoritmo otimizador e que por sua vez pode ser dividido em duas etapas: aplicação do operador gradiente e o ajuste propriamente dito dos pesos. O gradiente é um vetor que indica a direção e o sentido onde a taxa de mudança de uma

função aumenta mais rapidamente a partir de um determinado ponto, possui valor zero nos locais máximos ou mínimos e é simbolizado pelo operador ∇. Como o objetivo do treinamento é minimizar o erro é utilizado então o negativo do vetor gradiente, uma vez que ele aponta para o local onde este erro é decrementado mais rapidamente e por este motivo esse método também é chamado de Gradiente Descendente. O gradiente é obtido através do cálculo das derivadas parciais da função erro em relação a cada um dos pesos, ou seja:

$$\nabla E(w) = \frac{\partial E(w)}{\partial w} \tag{2-12}$$

A Figura 2.10 exibe um exemplo de gráfico de uma função $E(w)$ composta por dois pesos, ou variáveis, w_1 e w_2. Foram plotados os valores de saída da função considerando os valores possíveis que w_1 e w_2 podem assumir. Considerando o ponto inicial A, o objetivo é caminhar em direção ao ponto B, onde o erro dado pela função custo $E(w)$ possui o menor valor, encontrando assim os valores ótimos das variáveis que satisfazem o objetivo. Na imagem também é possível visualizar a direção do vetor gradiente, demonstrado pela seta vermelha, que possui direção e sentido orientado para o crescimento mais rápido da função e é calculado pelas derivadas parciais de $E(w)$ em relação a w_1 e w_2 no ponto A.

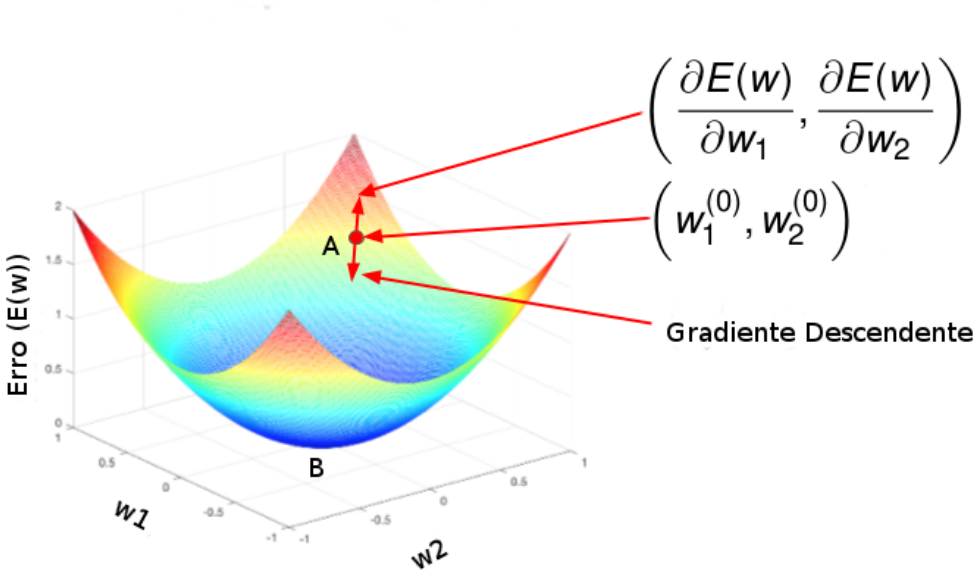


Figura 2.10: Gráfico exemplificativo de uma função $E(w)$ de duas variáveis, w_1 e w_2. Para que o ponto B, onde a função assume menor valor, seja alcançado é utilizado o vetor gradiente para indicar a direção e o sentido onde o erro é minimizado e os valores ótimos para w_1 e w_2 são encontrados.

Calculados os gradientes em relação a cada um dos pesos w_i o ajuste dos mesmos é realizado de forma proporcional à colaboração de cada um para o erro. O valor deste ajuste, para cada um dos pesos, é dado por:

$$\Delta w = -\eta \cdot \frac{\partial E(w)}{\partial w} \tag{2-13}$$

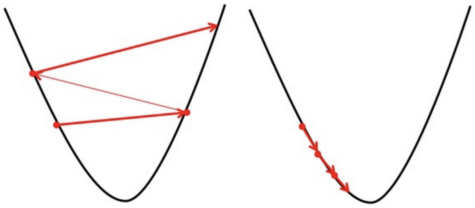
onde η é um parâmetro, chamado taxa de aprendizagem, a ser definido durante o treinamento e é utilizado para ponderar o tamanho do ajuste baseado no gradiente. Considerando uma época t, os pesos da próxima época $t+1$ irão possuir valores tais que:

$$w(t+1) = w(t) - \eta \cdot \frac{\partial E(w)}{\partial w} \tag{2-14}$$

ou seja, ao final de uma época os pesos w_i são atualizados através de uma subtração entre os valores atuais e uma fração do gradiente de cada um dos pesos em relação ao erro. O otimizador Gradiente Descendente, tal como apresentado, possui alguns alguns

problemas, dentre eles:

1. Para conjuntos grandes de treinamentos se torna impossível carregar todas as características de cada amostra na memória para cálculo do erro e gradientes.
2. O ajuste dos pesos é realizado apenas uma vez a cada época. Por isso pode ser necessário um elevado número de iterações para que os sucessivos ajustes conduzam o modelo até o ponto de menor erro e atinja a convergência, consumindo assim muito tempo e esforço computacional.
3. Taxa de aprendizado fixa pode gerar dois tipos de situações indesejadas: caso seja grande, pela equação 2-13, o ajustes dos peso também será grande fazendo que o local de mínimo de erro não seja atingido devido a essas grandes variações, conforme ilustrado na Figura 2.11. Caso essa taxa seja muito pequena poderão ser necessários um número muito grande de épocas, dado que a convergência do modelo no sentido do local mínimo será também muito lenta.



(a) *Taxa de aprendizado grande* (b) *Taxa de aprendizado pequena*

Figura 2.11: Gráfico exemplificativo de uma função custo de apenas uma variável. Na Figura 2.11(a) as setas vermelhas longas representam os saltos maiores que ocorrem durante o treinamento devido aos ajustes dos pesos mais drásticos, impedindo o modelo de alcançar o erro mínimo. Na Figura 2.11(b) as setas vermelhas mais curtas simbolizam os pequenos avanços que são realizados quando a taxa de aprendizado é pequena, resultado na necessidade de um grande número de épocas para que o erro mínimo seja alcançado.

Para solucionar estas limitações algumas evoluções do Gradiente Descendente foram propostas, dentre elas o otimizador *Adam*, ou Estimação do Momento Adaptativo (*Adaptive Moment Estimation*). Os problemas 1 e 2 são atenuados através da divisão do conjunto de treinamento em *mini-batch*, que é um sub-conjunto com número de reduzido de amostras e que seja suportado pela memória disponível. Assim é possível executar todo o algoritmo de retropropagação neste sub-conjunto, melhorando assim tanto o problema de falta de espaço em memória quanto de velocidade de convergência, uma vez que os pesos serão atualizados várias vezes dentro da mesma época permitindo que a trajetória em direção ao erro mínimo seja mais rapidamente corrigida e portando, alcançada.

O otimizador *Adam* utiliza também taxas de aprendizagens diferentes para cada um dos pesos, variável e calculada através da média móvel exponencial dos gradientes anteriores, o que contribuiu para atenuar o problema 3. O objetivo da taxa de erro adaptativa é que cada peso seja atualizado em magnitude proporcional à sua influência para o cálculo do erro na saída da rede.

2.4 Viés de dados

O viés também pode ser entendido como a tendência do algoritmo em não modelar o problema de forma correta, ou não generalizar o suficiente, por não considerar o conjunto dos dados necessário. Neste caso, pode-se ter dois cenários: 1) as características de cada amostra não são suficientemente capturadas pelo modelo durante o treinamento (ou não estão disponíveis) e 2) o conjunto de dados disponível para treinamento não contém exemplos representativos do problema a ser modelado.

Considerando este último cenário, de insuficiência de dados representativos, em [61] este problema é conhecido como **viés de dados** (*data bias*). Artigos científicos publicados recentemente, 2018 e 2019, evidenciam tanto o interesse da comunidade científica nesta questão quanto a necessidade de investigação sobre o quanto esse fenômeno afeta os modelos de reconhecimento facial, assim como os métodos que podem ser aplicados para mitigá-los [61, 42, 46, 38, 40].

Conforme [40] o problema de viés não é causado pela Inteligência Artificial em si, mas a forma como os modelos são treinados. Para que o reconhecimento facial funcione como o esperado, com elevada acurácia e imparcialidade, o conjunto de dados de treinamento deve possuir equilíbrio e abrangência em suas amostras, com diversidade representativa o suficiente para que reflita todas as formas possíveis em que as faces podem inerentemente se diferenciar.

Muitos conjuntos de dados públicos disponíveis para treinamento, como *Casia-Webface*2 [21], MS-celeb-1M [16] e *VGGFace2*[3] foram coletados na internet e são constituídos por fotos de celebridades e pessoas famosas onde estão, inclusive, muitas vezes sorrindo, com maquiagem e em lugares com boa iluminação, conforme exemplo da Figura 2.12. Estes bancos de imagens faciais possuem distribuição de dados bastante irregular quanto à etnia, com forte predominância de indivíduos caucasianos. Conforme [62], os percentuais de pessoas desta etnia, nas coleções dados anteriormente citadas, são respectivamente 84,4%, 76,3% and 74,2%. Inclusive o benchmark *LFW* também apresenta desbalanceamento neste aspecto, uma vez que possui 69,9%[62] de seus indivíduos classificados como caucasianos.

Em [40] a distribuição desigual de grupos demográficos, como etnia, gênero e idade nestas coleções de dados públicas é analisada e quantificada. Essa falta de

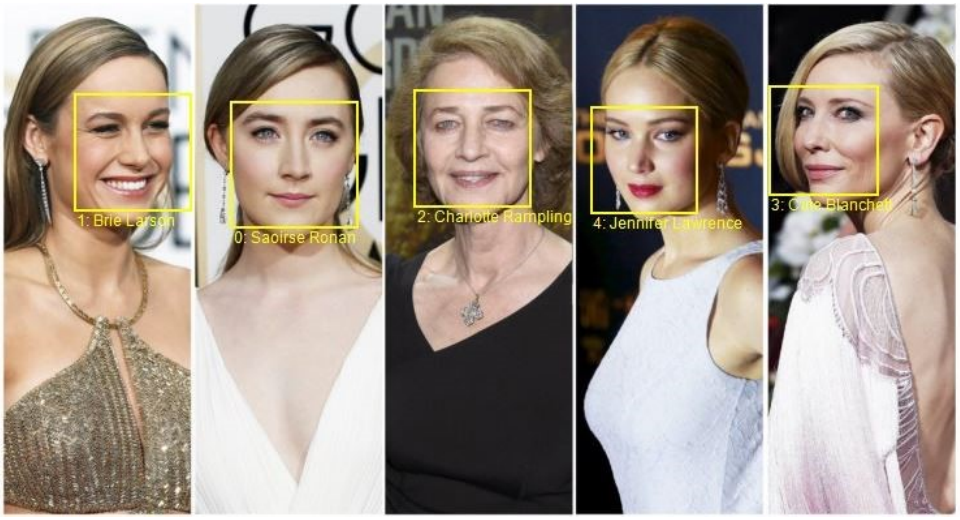


Figura 2.12: Exemplos de imagem de faces retirada na base [16]. É possível notar que as imagens possuem um mesmo padrão de pessoas com estereótipo caucasiano, sorrindo e no geral com boa iluminação.

conjuntos de dados de treinamento mais representativos tem levado a diversos problemas de classificação errônea, principalmente falso positivo, que é quando o algoritmo aponta que uma determinada foto é de uma pessoa, sendo que verdadeiramente trata-se de outra.

Tabela 2.2: Distribuição das amostras nas coleções de dados públicas de treinamento, agrupadas por gênero e cor de pele [40]. O conjunto LFW, que é amplamente utilizado como referência para comparação de resultados, possui elevado desequilíbrio na distribuição dos dados, tanto em gênero quanto em cor de pele.

	Gênero		Cor de pele	
	Feminino	Masculino	Escuro	Claro
LFW	22,5%	77,4%	18,8%	81,2%
IJB-C	37,4%	62,7%	18,0%	82,0%
CelebA	58,1%	42,0%	14,2%	85,8%

Conforme relatado no artigo [42], em 2018 uma entidade de direitos civis americana chamada *American Civil Liberties Union* (ACLU) fez o seguite teste com a ferramenta de reconhecimento facial da empresa *Amazon*[3], conhecida como *Rekognition*: constituindo um banco de dados de criminosos procurados, a entidade aplicou a ferramenta em fotos de deputados americanos e 28 deles foram falsamente reconhecidos como sendo um dos criminosos. Neste artigo também verificou-se que a taxa de erro foi 100% maior em deputados afrodescendentes, expondo então além do problema de falso positivo uma tendência maior em errar em grupos étnicos específicos. Esta situação inclusive gerou protestos e pedidos de diversos grupos organizados de forma que a Amazon anunciou

[3] Amazon.com, Inc

que temporariamente estão suspensas as vendas do *Rekognition* para órgãos de segurança pública.

Outro artigo com resultados concretos em relação ao viés de dados é o [46]. O trabalho consistiu em analisar e verificar os resultados de um algoritmo denominado *Gender Shades*, considerado o primeiro algoritmo de auditoria de gênero e cor de pele, desenvolvido especificamente para validação de soluções de reconhecimento facial.

O *Gender Shades* utiliza imagens categorizadas quanto ao gênero e a cor da pele para testar os algoritmos de reconhecimento facial, sendo o conjunto de dados denominado *Pilot Parliaments Benchmark*. Quanto ao gênero as imagens são classificadas de forma binária em masculino e feminino. Quanto a cor da pele, as categorias são de peles escura e clara. Os resultados foram produzidos também analisando intersecções dessas categorias, como feminino de pele escura. A coleção de dados desse algoritmo possui igual representação de todos os subgrupos, de forma que uma avaliação igualitária possa ser realizada.

Os testes foram realizados contratando os serviços de reconhecimento facial *Face++*, *MSFT* , *IBM*, *Amazon* e *Kairos*. Os resultados foram coletados e os erros apresentados conforme Tabela 2.3:

Tabela 2.3: Erro apontado pela coleção de dados *Pilot Parliaments Benchmark*.

Empresa	Total	Feminino	Masculino	Pele escura	Pele clara	FE	ME	FC	MC
Face++	1,6	2,5	0,9	2,6	0,7	4,1	1,3	1,0	0,5
MSFT	0,48	0,90	0,15	0,89	0,15	1,52	0,33	0,34	0,0
IBM	4,41	9,36	0,43	8,16	1,17	16,97	0,63	2,37	0,26
Amazon	8,86	18,73	0,57	15,11	3,08	31,37	1,26	7,12	0,00
Kairos	6,60	14,10	0,60	11,10	2,80	22,50	1,30	6,40	0,00

Conforme resultados apresentados na tabela 2.3 pode ser verificado de uma forma geral que os algoritmos possuem acurácia muito maior no grupo de homens. Quanto a cor da pele a categoria com menor erro é a de pele clara. De forma oposta, as mulheres de pele escura são as que foram erroneamente classificadas em maior número. A *Amazon* apresentou taxa de erro bastante expressiva no grupo de mulheres negras (31,37% de erro), seguida pela *Kairos* com 22,5%, sendo valores bastante altos quando comparados com homens de pele clara, que a próprioa *Amazon* acertou 100%, indicando de forma bastante eloqüênte que o viés é algo real, concreto e precisa ser melhor estudado e superado. Interessante notar que em [38] é descrito que esse viés existe também em humanos. Segundo os autores existe um viés *in-group* nas pessoas, de forma que a capacidade de reconhecer indivíduos da mesma idade, etnia e gênero é maior enquanto comparado com indivíduos de outros grupos (*out-group*).

Em [42] os autores mergulharam no problema de análise do viés de dados e criaram experimentos para tentar responder as seguintes perguntas:

- As CNNs aprendem a codificar informação específica de raça?
- As CNNs aprendem a codificar informação específica de idade?

Para responder a essas perguntas foram utilizadas 4 redes:

- *LightCNN-9* [66] treinada do zero .
- *LightCNN-29* [66] treinada com o *MS-Celeb-1M* [16].
- *ResNet50* [17] treinada com o *MS-Celeb-1M* [16] e *VGG-Face2* [4], totalizando mais de 13 milhões de imagens.
- *SENet50* [18] treinada com o *MS-Celeb-1M* [16] e *VGG-Face2* [4], totalizando mais de 13 milhões de imagens.

Para o estudo de caso de etnia foram utilizados conjuntos de dados divididos em 2 subgrupos: Caucasoide (branca) e Negroide (negra). Foram utilizados os conjuntos de dados de teste *CMU Multi-PIE* [14] para o primeiro e Craniofacial Longitudinal *Morphological (MORPH) Album-2* [11] para o segundo.

Para o estudo de caso de idade foi utilizado um conjunto de dados composto por 3 subgrupos, sendo (i)(0-14) anos, (ii)(15-32) anos e (iii)(+33) anos, sendo as amostras combinadas de 2 coleções: *Adience* [9] e *Cross-Age Celebrity* [6].

Vários experimentos foram realizados. A rede que foi treinada do zero aprendeu com conjunto de dados contendo apenas 1 subgrupo de cada vez, etnia ou idade, para observar as regiões discriminativas que são aprendidas para reconhecimento facial. Experimentos cruzados também foram realizados para tentar entender quais regiões foram ativadas em cada caso durante a classificação. Por exemplo, uma rede treinada com imagens apenas caucasianas foi utilizada para classificar faces do conjunto de dados negroide.

Foi realizado também treinado de ajuste fino, *fine-tuning*, em redes pré-treinadas para verificar como as regiões discriminativas evoluem para diferentes subgrupos. A Figura 2.13 apresenta os resultados da rede treinada em um subgrupo e testada em outro, para o estudo de caso das etnias, e os resultados sugerem que existe viés nesse contexto. A rede treinada no subgrupo caucasoide atingiu acurácia de 79,23% no conjunto de teste da mesma etnia mas apenas 34,31% no conjunto negroide.

Além da acurácia, outra técnica utilizada em [42] para investigação do viés de dados é a visualização dos mapas de características[4]. A Figura 2.14 exibe as regiões faciais que mais foram utilizadas para a extração de características. Essas imagens foram obtidas pela interpolação das respostas dos filtros da última camada convolucional e a sobreposição na imagem de entrada. Pode-se verificar que as RNCs utilizam regiões diferentes da face para extrair características de acordo com o a etnia.

[4]Essa técnica permite visualizar quais as características são detectadas pelo modelo, considerando uma determinada entrada.

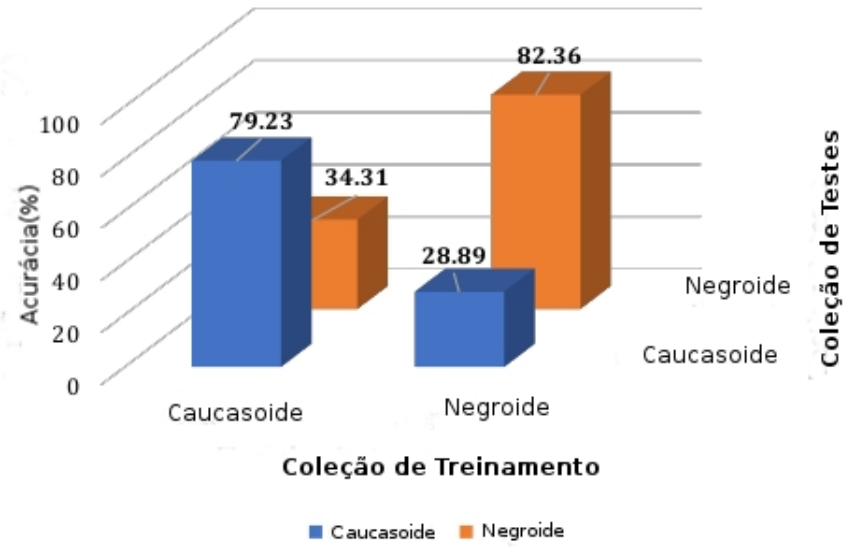
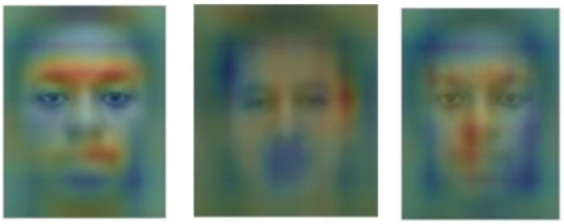


Figura 2.13: Figura extraída de [42]. A Acurácia da rede treinada em um subgrupo e testada em outro, para a análise do viés com etnias. O modelo treinado em coleção de dados negroide atingiu acurácia de 82,36% de acurácia em conjunto de testes também negroide e apenas 34,31% em caucasoide, sugerindo forte viés étnico.



(a) *Negroide* (b) *Caucasoide* (c) *Combinadas*

Figura 2.14: Imagem extraída de [42]. Visualização das características dominantes obtidas em redes treinadas a partir do zero usando as coleções de dados (a) negroide, (b) caucasoide e (c) combinadas.

Em pessoas de etnia negroide as regiões mais utilizadas são a parte acima dos olhos e os lábios. Em pessoas caucasianas percebe-se uma visão mais holística, sendo utilizado a morfologia do rosto um todo. Considerando as diferenças na acurácia já demonstradas, somadas à visualização das características faciais dominantes, agrupadas por etnia, têm-se uma forte sugestão do viés de dados como fator importante e impactante no desempenho dos algoritmos de reconhecimento facial.

Redes pré-treinadas também foram utilizadas para comparar as acurácias com coleções de testes dos dois subgrupos e os resultados também demonstram o viés étnico: *ResNEt50* obteve 98,7% no subgrupo caucasóide e 96,3% na negróide. *SENet50* atingiu 98,8% e 96,5% respectivamente. O modelo *LightCNN-29* alcançou 96,47% em caucasói-

des e 78,12% em negróides. A redução do desempenho das redes sugerem que a capacidade de generalização dos modelos dependem em muito da variabilidade dos dados de treinamento.

O treinamento em modo ajuste fino também foi utilizado em [42]. Redes pré-treinadas em conjuntos de dados públicos foram submetidas ao ajuste fino utilizando conjuntos de dados específicos dos subgrupos analisados. O que se observou foi uma mudança nas características predominantes, sendo que inicialmente apresentaram padrões gerais e posteriormente, específicos. Por exemplo, após ajuste fino com o subgrupo negróide esses mapas apresentaram características de redes treinadas exclusivamente com o conjunto de dados deste subgrupo, conforme Figura 2.15. Portanto, verifica-se que existem evidências de que o ajuste fino realmente pode contribuir quando se procura um modelo com maior acurácia para aplicação em subgrupos específicos .

(a) *SENet50 pré-treinada* (b) *ResNet50 pré-treinada* (c) *LightCNN-29 pré-treinada* (d) *LightCNN-9 treinada* (e) *SENet50 com ajuste fino* (f) *ResNet50 com ajuste fino*

Figura 2.15: As imgens 2.15(a),2.15(b) e 2.15(c), extraías de [42], exibem características de redes pré-treinadas, onde são observadas regiões faciais que ativaram mais neurônios nas CNNs, sendo observado nessas 3 primeiras uma fusão das regiões utilizadas em ambos subgrupos. Nas imagens 2.15(d),2.15(e) e 2.15(f) observa-se o deslocamento dessas regiões para os padrões similares ao da rede treinada do zero com o conjunto de dados negróide..

Em [42] foram realizados experimentos para verificar a existência de viés relacionado à idade. Os modelos pré-treinados *ResNet50*, *SENet50* e *LightCNN-29* foram avaliados em 3 grupos diferentes: (i) 0-14 anos, (ii) 15-32 anos e (iii)+33 anos. As redes pré-treinadas nas coleções de dados mais amplas, *ResNet40* e *SENet50*, alcançaram acurácias de 77,9% e 75,5% respectivamente no grupo de +33 anos. No grupo de idade entre 15-32 ambos os modelos alcançaram melhores resultados, 81 e 83% e menos de 45% para o grupo de até 15 anos.

Como pode ser percebido na Figura 2.16 esse mesmo viés se mostrou na rede *LightCNN-29*, sendo que este modelo, assim como no caso de etnias, demonstrou piores resultados. Portanto também existem evidências que o viés dos dados no contexto de idade pode ser acentuado caso o conjunto de treinamento não contenha uma distribuição

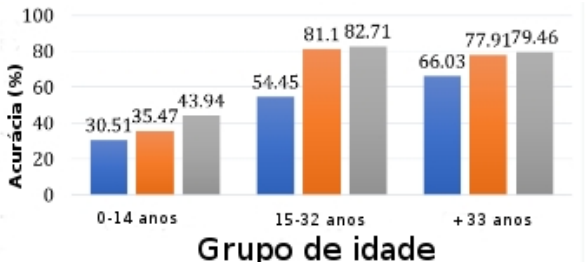


Figura 2.16: Imagem extraída de [42], exibe a acurácia de modelos pré-treinados para os 3 diferentes grupos de idade.

de dados entre os grupos suficientemente abrangente, dado que este modelo foi treinado apenas no conjunto *MS-Celeb-1M*, enquanto os modelos anteriormente citados foram treinados também no *VGG-Face2*.

A visualização das características predominantes da rede *LightCNN-9*, treinada do zero em cada um dos subgrupos de idade, também fornece evidências de que regiões faciais diferentes são utilizadas. A Figura 2.17 permite analisar as diferenças para cada agrupamento de idade. Verifica-se uma diferença mais acentuada entre as idades com maior distância. Em crianças até quinze anos a região superior da cabeça e a morfologia do rosto parecem ser as mais discriminativas. Em adultos com mais de 33 anos a região dos olhos se apresenta como a de maior influência.

Figura 2.17: No grupo de crianças até 15 anos as regiões mais utilizadas para reconhecimento são o cabelo e a morfologia da face. No grupo entre 15-32 anos os lábios e a parte inferior do rosto que parecem ser mais influentes. Acima de 33 anos nota-se que os olhos parecem ter predominância discriminativa para as CNNs

CAPÍTULO 3

Trabalho Proposto

Neste capítulo, discute-se os dados utilizados no desenvolvimento deste trabalho, a técnica de pré processamento das coleções de dados de imagens faciais e a proposta de treinamento das Redes Neurais Convolucionais para aumentar o desempenho do reconhecimento facial no contexto da segurança pública brasileira.

Considerando os objetivos relacionados na Seção 1.2 e a justificativa descrita na Seção 1.1 este trabalho investiga o problema de viés de dados considerando imagens faciais em coleção de dados brasileiro. Também propõe a utilização da função custo *Triplet Loss* em treinamento de ajuste fino como o método apropriado para para mitigar este problema partindo de um modelo pré-treinado de referência quando se dispõe de limitado número de amostras. Como forma de demonstrar o desempenho superior desta função custo os resultados foram comparados com os resultados alcançados após o treinamento também em ajuste fino do mesmo modelo de referência, utilizando o mesmo conjunto de treinamento e supervisionado pela função custo *AM-Softmax*[58]. Na Seção 3.1 a coleção de dados utilizada é explicada, assim como os critérios para a sua divisão em conjunto de treinamento, validação e testes. Na Seção 3.2.2 são descritas os modelos de representação facial que atingiram o estado da arte e que por isso foram **experimentados e avaliados**. Eles foram escolhidas por utilizarem funções de custo com diferenças conceituais significativas e portanto espera-se que gerem resultados baseados em princípios distintos de forma que possa ser encontrado aquele com maior desempenho.

Na seção 3.2.1 é descrito o método de alinhamento e detecção facial. Embora não façam parte do escopo desse projeto, conforme 2.2, são etapas fundamentais para o reconhecimento facial. O *framework* utilizado foi proposto por [69] e alcançou o estado da arte nestas tarefas. Na seção 3.2.2, descreve-se as técnicas da etapa de reconhecimento facial que serão investigadas e experimentadas.

3.1 Coleção de Dados

A coleção de dados de faces brasileiras utilizada nos experimentos dessa dissertação foi construída a partir de bancos de dados de fotos privados da Polícia Civil do

Estado de Goiás. Foi realizado um trabalho de intersecção entre as diversas bases com o objetivo de encontrar todos os indivíduos com mais de uma imagem, considerando apenas fotos diferentes de cada identidade. Essa restrição é necessária devido ao fato que tanto para treinar quanto para testar é necessário que existem ao menos duas amostras por classe, ou identidade.

Dessa forma foi estruturada uma coleção de dados inicialmente composta por 61.221 imagens de 27.653 identidades, ou classes, diferentes, conforme exibido na Tabela 3.1.

Total classes	Total de amostras	Média	Desvio Padrão	Classe com mais amostras	Classe com menos amostras
27.653	61.221	2,21	1,23	32	2

Embora o número de amostras por identidade apresente variação, tendo como mínimo 2 e máximo 32, a distribuição dessa proporção se concentra no primeiro caso, conforme histograma da Figura 3.1. O número de identidades com apenas duas amostras é de 25.737, seguido de 923 classes com 3 amostras. Apenas 177 classes possuem mais de 10 amostras.

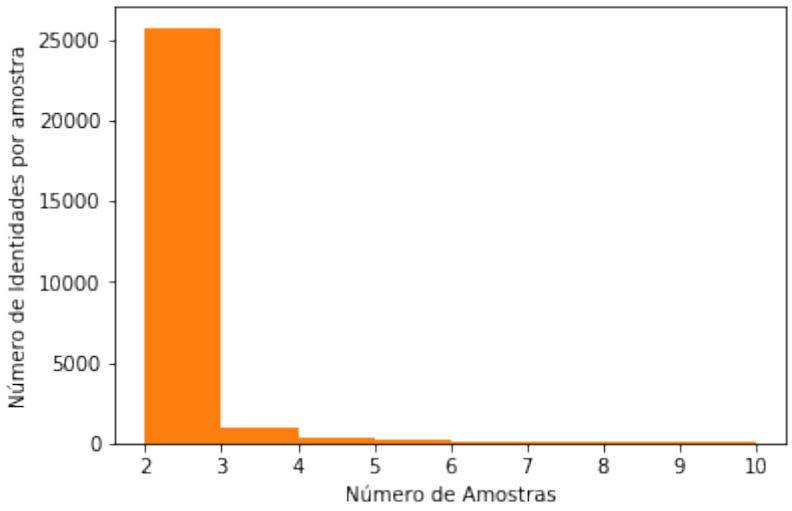


Figura 3.1: Histograma do número de amostras por identidade.

Conforme [3] as coleções de dados podem ser classificadas como larga ou profunda, sendo o primeiro caso caracterizado por um grande número de identidades com poucas amostras de cada. Os conjuntos de dados profundos são aquelas com reduzido número classes, porém elevado número de amostras para cada identidade. Portanto a coleção de dados utilizada neste trabalho se caracteriza pela largura.

Uma vez consolidada, a coleção de dados foi dividida de acordo com os objetivos do projeto, que são tanto o treinamento quanto a avaliação de modelos. Para o treinamento é necessário um numero número maior de identidades, pois segundo [28] existe uma relação direta entre a quantidade de amostras de treino e o desempenho do modelo.

Portanto, foram removidas, do conjunto de dados inicial, classes e suas respectivas amostras para avaliação e o restante foi utilizado para treinamento. A remoção completa de todas as amostras de cada umas identidades utilizadas para avaliação visou garantir não haver sobreposição com a coleção de imagens de treino.

A configuração da coleção de imagens de avaliação pode ser visualizada na Figura 3.2. De acordo com as métricas estabelecidas em 3.2.5, cada uma dessas amostras é comparada a uma galeria de imagens, que é composta tanto pelos pares correspondentes das identidades do conjunto de avaliação quanto por imagens diversas. Os conjuntos de teste e validação são compostos por imagens que serão comparadas, uma a uma, a todas as imagens da galeria. As amostras das mesmas identidades compõem o conjunto de imagens correspondentes da galeria.

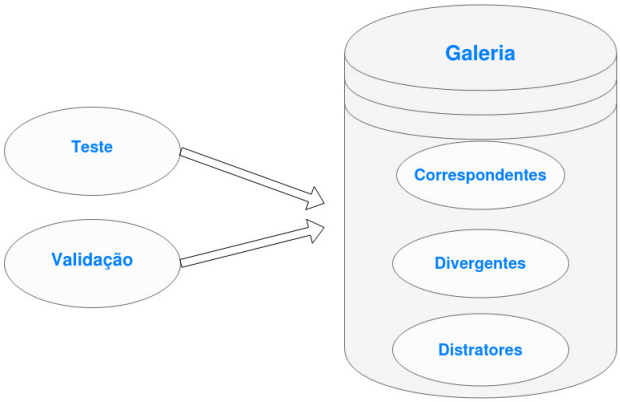


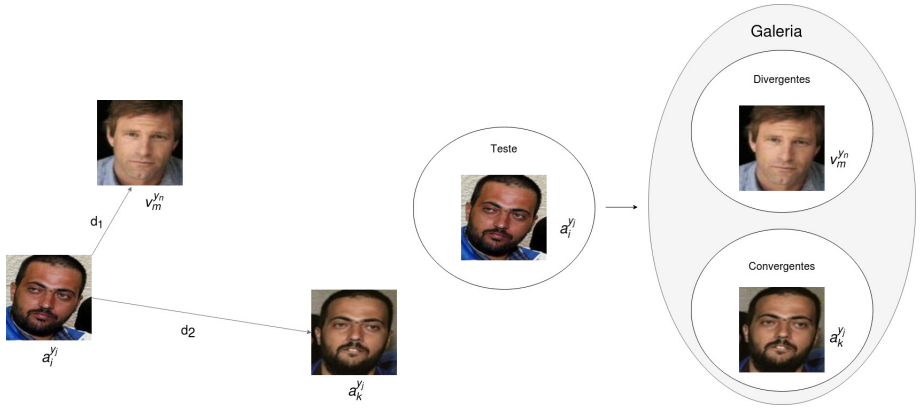
Figura 3.2: Estrutura das imagens de avaliação, dividas em: teste, validação e galeria de imagens.

O conjunto de testes é composto por amostras de identidades consideradas como os exemplos mais difíceis, que são aquelas em que, utilizando um dos modelos de referência, a distância interclasse é maior que a intraclasse. Ou seja, situações em que houve um erro de identificação facial. Encontrar estas situações de erro está coerente com os objetivos da dissertação, que são investigar a existência de viés de dados no modelo e propor um método de mitigação deste problema, conforme 1.2.1. Para analisar o viés de dados é necessário entender e quantificar os erros.

Para identificar esses exemplos mais difíceis foi utilizado o modelo de referência com extração de vetores de características faciais em 128 dimensões (128d), conforme seção 3.2.3. Os seguintes passos foram executados:

1. Foram gerados os vetores de características faciais de todas as imagens.
2. Dado um i-ésimo vetor $a_i^{y_j}$, pertencente à classe y_j, foi identificado o seu vizinho mais próximo, v, utilizando distância quadrática. Esse passo foi aplicado em todos i vetores da coleção de dados inicial.

3. Caso o vetor mais próximo v não tenha sido da mesma identidade y_j do vetor a_i, conforme Imagem 3.3(a), então todas as amostras desta identidade são removidas da coleção de dados. A amostra a_i é adicionada ao conjunto de testes e as imagens restantes da mesma classe são adicionadas ao conjunto de imagens correspondentes.
4. Todas as amostras da identidade da imagem relativa ao vetor v, caso a condição de erro do passo anterior tenha ocorrido, são movidas para a coleção de dados de divergência, como exemplo da Figura 3.3(b).



(a) *Erro de identificação facial* (b) *Divisão das amostras que geraram erro*

Figura 3.3: (a) Erro de identificação facial, uma vez que a distância d_1 entre exemplos de classes diferentes é menor que a distância d_2 entre amostras da mesma identidade. (b) As amostras das classes com erro de identificação são movidas para as coleções de teste e galeria, de forma a manter o mesmo cenário em avaliações posteriores de modelos.

A coleção de imagens divergentes faz parte da galeria e, portanto, não são utilizadas para treinamento dos modelos. Elas foram movidas para a galeria por violarem a restrição de distanciamento interclasses definida na função custo utilizada no treinamento dos modelos de referência, de forma a garantir que estes mesmos cenário dos exemplos mais difíceis também existam na avaliação dos modelos treinados durante a fase de experimentos desta dissertação.

Além da coleção de testes também foi gerada uma coleção de imagens de validação. Foram selecionadas, de forma aleatória, 20% das identidades restantes, totalizando 5.000. Para cada classe uma amostra foi, também, aleatoriamente selecionada para compor o conjunto de validação, que será confrontada com a galeria. As outras imagens de cada classe foram movidas para a galeria, na coleção de imagens correspondentes. Esse conjunto de dados tem como objetivo avaliar a generalização dos modelos, de forma que seja possível garantir que estes, após os treinamentos, tenham melhorado a acurácia nos exemplos mais difíceis, coleção de testes, mas também mantenham a taxa de acertos nos demais casos.

Após a reestruturação da coleção de dados inicial, com as sucessivas divisões, os conjuntos de imagens e suas respectivas quantidades podem ser visualizada na Tabela 3.1. Por esta tabela é possível verificar que o conjunto completo de imagens é composto, aproximadamente, em 70% de pessoas do sexo masculino e apenas 30% do sexo feminino.

Porém, após utilização do modelo de referência para encontrar os erros de identificação facial e extração das amostras de teste, verifica-se que este conjunto possui quantidade de imagens quase balanceada quanto ao gênero, sendo 56% do sexo masculino e 43% do feminino. Portanto, é possível afirmar que existem indícios de que o modelo de referência apresente viés quanto ao gênero, uma vez que, proporcionalmente, errou mais em pessoas do sexo feminino[1].

	Total (classes/amostras)	Masculino (classes/amostras)	Feminino (classes/amostras)
Treinamento	20.672 / 45.698	14.268 (69,1%) / 32.666 (71,4%)	6.404 (30,9%) / 13.032 (28,6%)
Teste	838 / 1.252	477 (56,9%) / 706 (56,3%)	361 (43,0%) / 546 (43,6%)
Validação	5.000 / 5.000	3.464 (69,2%) / 3.464 (69,2%)	1.536 (30,7%) / 1.536 (30,7%)
Divergentes	1.143 / 2.782	669 (58,5%) / 1808 (64,9%)	474 (41,4%) / 974 (35,0%)
Correspondentes	5.838 / 7.340	3.941 (67,5%) / 5.185 (70.6%)	1.897 (32,5%) / 2155 (29,3%)
Total	27.653 / 61.221	18.878 (68,2%) / 43.359 (70,8%)	8.887 (31,7%) / 17.862 (29,2%)

O somatório dos sub-totais das amostras das coleções de dados ultrapassa o número total de amostras e classes exibido na última linha da Tabela 3.1. Isso se deve ao fato de haver sobreposição de amostras entre as coleções de dados de teste e de correspondentes, uma vez que uma determinada amostra pode estar em ambas coleções.

Além das amostras correspondentes e divergentes a galeria também é composta por um conjunto de distratores. Inspirado na proposta de avaliação em [28], esses distratores, que são imagem únicas de pessoas sem nenhuma correspondência com as outras coleções de dados, tem como objetivo avaliar a escalabilidade dos modelos em tarefas de identificação em que uma imagem facial é comparada à um banco de dados com um elevado número de fotos de diferentes identidades, de forma a mensurar o desempenho neste cenário. Essa opção foi adotada por estar coerente com os objetivos do projeto, uma vez que é bem próximo da circunstância da operação de identificação facial em Segurança Pública.

Na Tabela 3.1 são apresentados o total de imagens, bem como os quantidades por gênero.

	Total	Masculino	Feminino
Distratores	208.187	132.576 (63,67%)	75.618 (36,3%)

[1] orientador: fundamental discutir esse cenário, assim como essa sugestão

3.2 Métodos Utilizados

3.2.1 Alinhamento Facial

Conforme descrito na seção 2.2, o alinhamento das imagens é uma etapa importante para o reconhecimento facial. Geralmente a operação é realizada após a detecção facial e consiste em detectar *landmarks* nas imagens para que sejam realizadas outras tarefas como o recorte e o redimensionamento das faces para que todas fiquem na mesma escala. Geralmente outros ajustes também podem ser aplicados, como por exemplo a utilização de transformadas *affine* para correção de ângulos e alinhamento horizontal dos olhos.

Neste projeto será utilizado o método de alinhamento desenvolvido por [69], abreviado como MT-CCN, que propõe uma metodologia baseada em RNCs para realizar detecção facial e alinhamento. Nesse método são utilizados 3 estágios em cascata, utilizando 3 RNCs projetadas para realizar detecção de face e *landmarks* de uma maneira inicialmente mais grosseira, no primeiro estágio, até uma forma mais refinada, no último estágio.

Conforme a Figura 3.4, obtida de [69], inicialmente a imagem é redimensionada em diferentes escalas para construção de uma pirâmide. No primeiro estágio uma Rede Neural Convolucional chamada *Proposal Network* (P-Net) é usada para obter regiões que podem conter face, assim como os respectivos retângulos delimitadores. Posteriormente aplica-se a supressão dos não máximos, *non-maxima supression* (NMS), para realizar a fusão dos retângulos delimitadores e diminuir os candidatos que se sobrepõem.

No estágio 2, outra RNC chamada *Refine Network* (R-Net) é usada para rejeitar os falsos candidatos gerados pela *P-Net*. Essa mesma rede realiza a calibração dos retângulos delimitadores e aplica NMS para nova fusão dos candidatos restantes. No estágio 3, a rede chamada *Output Network* (O-Net) realiza tarefa similar à anterior, porém com mais exatidão. Além do retângulo delimitador essa rede produz como saída cinco landmarks. No momento da publicação do trabalho, o MT-CCN superou o estado da arte nas coleções de dados *Face Detection Data Set and Benchmark* (FDDB) [23] [37], *WIDER FACE* e *Annotated Facial Landmarks in the Wild* (AFLW) [37] benchmark, tanto nas tarefas de detecção quanto de marcação de *landmarks*.

A escolha desse método conjunto de detecção e alinhamento se deu pelo fato de que, além dos excelentes resultados obtidos, os próprios autores disponibilizaram no repositório de códigos GitHub [54] uma implementação, que será utilizada para realizar essas etapas na fase de experimentos.

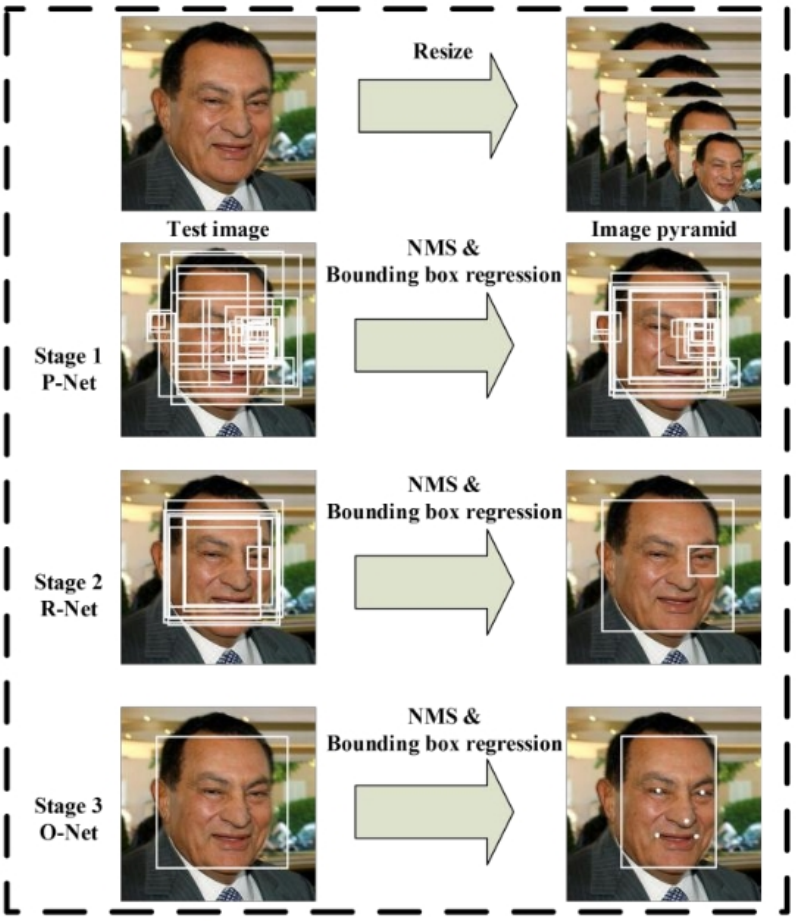


Figura 3.4: Fluxo em cascata que representa a sequência de tarefas executadas. Primeiramente é construída uma pirâmide de imagens. Então cada uma é aplicada a rede *Proposal Network* (P-Net) onde a saída são as imagens candidatas. As imagens são submetidas ao próximo estágio que utiliza também uma CNN chamada *Refinement Network* (R-Net). Por último, a rede chamada *Output Network* (O-Net) produz o retângulo delimitador e a posição dos *landmarks*.

3.2.2 Modelos de Representação Facial

A etapa de reconhecimento facial é onde a face previamente alinhada é submetida em um modelo para que seja extraída a representação facial compacta e discriminativa em formato vetorial, o vetor de características faciais. Os vetores de características são comparados de acordo com suas particularidades de forma que seja gerado algum indicador, ou pontuação, que indique a probabilidade de serem da mesma pessoa.

No âmbito do reconhecimento facial o componente mais relevante é a função custo, dado que ela supervisiona o treinamento com objetivo de que, na saída das arquiteturas baseadas em *RNC*, sejam gerados vetores faciais compatados mapeados em um determinado espaço de características, onde alguma métrica distância entre estes vetores seja correspondente à similaridade facial das imagens de entrada. As funções custo utilizadas nesta dissertação são a *FaceNet* [50] e a *AM-Softmax* [58].

A *FaceNet*, publicada 2014, propõe um modelo que aprende a mapear imagens faciais diretamente para um espaço euclidiano[2] compacto, onde as distâncias correspondem a uma medida de similaridade facial. O método é baseado no treinamento de uma RNC para que seja extraído um vetor de características faciais euclidiano por imagem de maneira que a distância quadrática entre estes vetores de características seja diretamente relacionada à semelhança facial entre os indivíduos que os originou. Vetores de características faciais da mesma pessoa possuem distância pequena entre eles e de pessoas diferentes possuem distâncias maiores.

O vetor de características faciais, ou *face embedding*, representado por $f(x) \in \mathbb{R}^d$, codifica uma imagem x em um espaço euclidiano de d dimensões. Ou seja, é uma representação vetorial da imagem. Estes vetores são normalizados tais que $\|f(x)\|_2 = 1$.

Uma vez que os vetores faciais tenham sido produzidos, as duas tarefas mais comuns do reconhecimento facial podem ser resolvidas com técnicas conhecidas de reconhecimento de padrões. A verificação facial envolve encontrar o limiar de forma que distâncias menores significam a mesma identidade e maiores, identidades diferentes. E o reconhecimento pode ser resolvido, por exemplo, pelo algoritmo dos K vizinhos mais próximos (*K-NN*, ou K - Nearest Neighbors).

A função custo empregada na *FaceNet* para o treinamento da RNC é denominada *Triplet Loss*, inspirada em [25]. Os vetores de características faciais são gerados a partir de uma imagem x, tal que a distância quadrática entre todas as faces, independente das condições da imagem, seja pequena para uma mesma identidade e grande para imagens de identidades diferentes. A função *Triplet Loss* supervisiona o treinamento do modelo de forma que o vetor facial gerado a partir da imagem âncora x_i^a de uma dada pessoa deve ser espacialmente mais próximo ao vetor extraído de outra imagem da mesma pessoa, exemplo positivo x_i^p, do que de o vetor facial originado de imagem de outra pessoa, exemplo negativo x_i^n, em um espaço de características $\mathbb{R}^d$.

A Figura 3.5 é uma representação da transformação pretendida pela utilização da *Triplet Loss*, onde a distância entre a imagem âncora, ponto em azul, e a imagem negativa, ponto em vermelho e de identidade diferente, através do treinamento do modelo, aumenta de tal forma que se torne maior do que a distância entre a âncora e a imagem positiva, de mesma identidade e presentada pelo ponto verde.

A formulação matemática do objetivo é apresentada na Equação 3-1. A distância quadrática entre os vetores faciais normalizados da amostra âncora x_i^a e positiva x_p^a, definida por $\left\|f(x_i^a) - f(x_i^p)\right\|_2^2$, somada a uma margem α, deve ser menor que a distância entre os vetores faciais dos exemplos âncora e negativo x_n^a, definida em $\|f(x_i^a) - f(x_i^n)\|_2^2$,

[2]Espaço vetorial de dimensão finita n que obedece axiomas e postulados da geometria euclidiana.

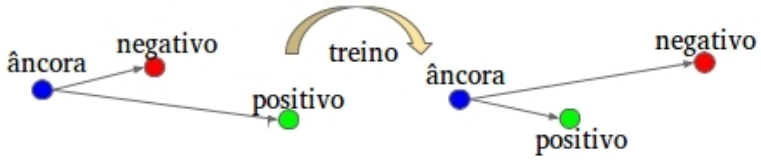


Figura 3.5: Figura traduzida e adaptada de [50]. A função *Triplet Loss* tem como objetivo diminuir as distâncias entre imagens da mesma pessoa, ou seja, entre a imagem *âncora* e a imagem *positiva* e também aumentar a distância entre imagens de pessoas diferentes, nesse caso a imagem *âncora* e a imagem *negativa*.

$$\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 + \alpha < -\| f(x_i^a) - f(x_i^n) \|_2^2, \tag{3-1}$$

$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \tau \tag{3-2}$$

onde τ é o conjunto de todos os trios de imagens possíveis na coleção de dados.

Portanto o erro a ser minimizado é dado por:

$$L = \sum_{n=1}^{n} [\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \alpha] \tag{3-3}$$

Além da função custo este trabalho também propôs uma forma de escolher os trios de imagens para treinamento. Isso porque os autores, como pesquisadores da *Google*, dispunham de algo em torno de 200 milhões de imagens, tornando o treinamento quase impraticável considerando todas as combinações possíveis. Além disso, boa parte dos trios agregam pouco à evolução do modelo. Por exemplo, a utilização de uma imagem facial de pessoa muito diferente, como exemplo negativo, pode não contribuir muito para o aprendizado. Portanto, na *FaceNet* foi proposto um algoritmo para seleção dos trios, com o objetivo de encontrar aqueles que tem maior chance de violar a restrição imposta em 3-1.

Os autores então definiram que a geração de trios é realizada durante o treinamento, dentro de cada *mini-batch*, tal que $\arg\min_{x_i^n} \| f(x_i^a) - f(x_i^n) \|_2^2$. Isto é, seleciona-se as amostras em que a distância entre os vetores faciais do exemplo âncora x_i^a e negativo x_i^n são menores, indicando que existe maior similaridade facial entre pessoas diferentes. Estes pares são denominados *negativos difíceis* e contribuem para que o modelo aprenda a extrair características faciais mais discriminativas.

Os resultados alcançados pela *FaceNet*[50] demonstram que o estado da arte foi superado à época. Considerando a atualização do *LFW* [19] foi utilizado o protocolo chamado *Unrestricted with labeled outside data*, que permite a utilização de outras coleções de dados para treinamento, sendo então a própria coleção *LFW* para teste. A

acurácia reportada foi de 99,63%, sendo este o melhor resultado já atingido até então, reduzindo o erro reportado pela *DeepFace* [68] por fator maior que 7 e o valor anterior do até então estado-da-arte da *DeepId2+* [52] em 30%.

Outros trabalhos posteriormente alcançaram resultados muito próximos ao *FaceNet* utilizando coleções de dados de treinamento com um número reduzido de amostras, na ordem de 500 vezes menor, sendo um deles denominado *SphereFace* [35], propôs uma nova função perda que avançou o estado da arte. O ponto de partida dos autores foi o questionamento "A distância euclidiana é sempre a mais adequada para o aprendizado de características faciais discriminativas?". Neste trabalho foi proposto uma versão modificada da função custo *Softmax loss* de forma que a separação entre os vetores faciais de mesma identidade e identidades diferentes seja angular.

Outro trabalho que também utilizou a distância angular como critério para a delimitação de fronteira de decisão e garantir a separabilidade entre os vetores de características faciais de indivíduos diferentes foi [36]. Conforme observado neste trabalho, modelos supervisionados pela função *Softmax* geram fronteiras de decisão que são fortemente determinadas pela similaridade angular entre cada classe. Isso porque, segundo [36], esta função utiliza distância do cosseno como critério para classificação. Por isso os autores também propuseram uma adaptação da função *Softmax loss*, denominada *L-Softmax*, de forma a garantir uma margem angular maior de separação entre os vectores de características faciais de indivíduos diferentes.

Com base na *Sphereface* [35] e *L-Softmax* [36], em [58] também foi proposta uma versão modificada da *Softmax loss* para supervisionar o treinamento de modelos baseados em RNC com o objetivo de gerar vetores de características faciais com ênfase na separação angular entre aqueles do mesmo indivíduo e indivíduos diferentes. Este trabalho alcançou o estado da arte, tendo resultados muito próximos àqueles do *SphereFace* e da *L-Softmax* e com menor complexidade de implementação. Por esta razão é o modelo utilizado nesta dissertação com o objetivo de comparação com a *FaceNet* para definição de qual abordagem apresenta melhores resultados na diminuição do viés de dados para faces brasileiras a partir de um modelo pré-treinado.

A função custo proposta em [58] é denominada *AM-Softmax*. O trabalho parte da formulação original da função custo *Softmax loss* dada por:

$$L_s = \frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{W_{y_i}^T f_i}}{\sum_{j=1}^{c} e^{W_j^T f_i}} \tag{3-4}$$

onde f_i é a entrada da última camada totalmente conectada da i-ésima amostra, $W_{y_i}^T$ é o vetor de pesos associados à classe y e W_j é o j-ésimo vetor de pesos desta última camada totalmente conectada. O termo $W_{y_i} f_i$, *logit* alvo da i-ésima amostra, é o produto

interno entre o vetor de pesos e a sua entrada e, pela similaridade do cosseno[3], este termo pode ser formulado como $W_{y_i} f_i = \|W_{y_i}\| \|f_i\| \cos(\theta_{y_i})$, de forma que a função *softmax loss* também pode ser expressa como:

$$L_s = \frac{1}{n} \sum_{i=1}^{n} \log \frac{e^{\|W_{y_i}\| \|f_i\| \cos(\theta_{y_i})}}{\sum_{j=1}^{c} e^{\|W_j\| \|f_i\| \cos(\theta_j}} \tag{3-5}$$

onde θ_{y_i} é o ângulo entre o vetor de pesos W_{y_i}, associado à classe y, e o vetor de características de entrada da função.

Então os autores propuseram uma função $\psi(\theta)$ que introduziu uma margem aditiva à função *Softmax* conforme Equação 3-6:

$$\psi(\theta) = \cos(\theta) - m. \tag{3-6}$$

Essa margem aditiva garante que os vetores de características faciais de uma mesma identidade sejam separados por uma larga fronteira angular daqueles de identidades diferentes. A função *Softmax* passa a ser definida como na Equação 3-7:

$$L_s = \frac{1}{n} \sum_{i=1}^{n} \log \frac{e^{\|W_{y_i}\| \|f_i\| \psi(\theta_{y_i})}}{e^{\|W_{y_i}\| \|f_i\| \psi(\theta_{y_i})} + \sum_{j=1, j \neq y_i}^{c} e^{\|W_j\| \|f_i\| \cos(\theta_j)}} \tag{3-7}$$

Os autores também propuseram que tanto o vetor de pesos W quanto a entrada f sejam normalizados. Neste caso a norma dos vetores é igual a um e, portanto, $\|W_{y_i}^T\| \|f_{y_i}\| = 1$. Assim, a entrada da função $\psi(x)$, na Equação 3-9 é tal que 3-8:

$$x = \cos(\theta_{y_i}) = W_{y_i}^T f_i. \tag{3-8}$$

$$\psi(x) = x - m. \tag{3-9}$$

Como foi aplicada normalização aos pesos e vetores de características faciais, um hiper parâmetro s foi utilizado para escalar os valores dos cossenos. Ao final, a função custo, denominada *AM-Softmax*, foi definida como na Equação 3-10:

$$L_{AMS} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{e^{s.(\cos(\theta_{y_i}) - m}}{e^{s.(\cos(\theta_{y_i}) - m)} + \sum_{j=1, j \neq y_i}^{c} e^{s.\cos(\theta_j)}} \tag{3-10}$$

Aplicando 3-8 em 3-10 a função *AM-Softmax* pode ser escrita conforme Equação 3-11 :

[3]Medida de similaridade entre dois vetores que mede o cosseno do ângulo entre eles [65]

$$L_s = \frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s.(W_{y_i}^T f_i - m)}}{e^{s.(W_{y_i}^T f_i - m)} + \sum_{j=1, j\neq y_i}^{c} e^{sW_j^T f_i}} \tag{3-11}$$

De forma experimental foram definidos os valores ótimos dos hiper parâmetros, sendo $s = 30$ e $m = 0,35$. Para melhor entender e demonstrar como a *AM-Softmax* consegue promover a separação angular dos vectores de características, mantendo as amostras da mesma classe angularmente próximas e as amostras de classes diferentes, distantes, foi realizado um experimento de teste utilizando a coleção de dados *Fashion MNIST* [67]. Esta coleção contém 10.000 amostras de imagens de peças de vestuário de 10 classes diferentes como camiseta, calça, vestido, sandália e etc.

Este experimento consistiu no treinamento de modelos baseados em RNC de 7 camadas para o aprendizado de extração de vetores de características de 3 dimensões, utilizando as funções custo *Softmax* e *AM-Softmax*. Na Figura 3.6 cada cor representa uma classe, sendo cada ponto relativo a uma amostra. Neste imagem pode ser verificado que os vetores intraclasse estão próximos e que existe uma margem angular maior entre as amostras de cores diferentes, interclasse, na visualização produzida pela *AM-Softmax* em relação à *Softmax*.

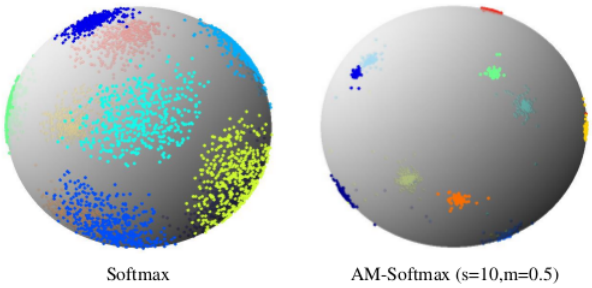


Figura 3.6: Figura traduzida e adaptada de [58]. A margem de separação angular entre as amostras de classes diferentes, identificadas por cores diferentes, é maior na visualização gerada com vetores de características extraídos de modelo treinados com *AM-Softmax*.

3.2.3 Modelo de Referência

Foram utilizados dois modelos de referência disponíveis em [49], treinados com a mesma arquitetura e com a coleção de dados *CASIA-Webface* [21], sendo a diferença entre eles a dimensão dos vetores de características faciais, 128d e 512d. A utilização destes dois modelos para fins de comparação é necessária uma vez que as funções

custo utilizadas nos experimentos publicaram seus resultados com dimensões diferentes. A *FaceNet* apresentou melhor acurácia com os vetores de características faciais de dimensão 128d. Já a *AM-Softmax* publicou os resultados utilizando 512d, sem mencionar resultados utilizando outras dimensões. Por isso a necessidade de investigar os resultados do treinamento de ajuste fino em ambas dimensões.

O modelo de referência foi treinado utilizando uma função custo chamada *Center Loss* proposta em [64] para supervisão conjunta com a *Softmax*. Esta função tem um objetivo duplo: aprender um centro, em um espaço euclidiano $\mathbb{R}^d$, para os vetores de características faciais de dimensão d para cada indivíduo e penalizar o modelo quando os vetores intraclasse ficarem distantes do centro, mesmo que a classificação tenha sido correta.

Os benefícios desta supervisão conjunta são uma maior compactação intraclasse e consequentemente maior distanciamento entre vetores faciais interclasse, tornando-os então não apenas separáveis mas também mais discriminativos, conforme ilustrado na Imagem 3.7. Com esta abordagem o modelo alcançou resultados do estado da arte em coleções de dados de teste de referência, como 99,2% de acurácia no *LFW*, tenso sido treinado com a coleção de dados *CASIA*.

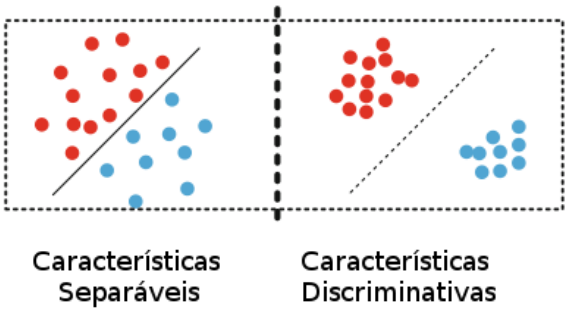


Figura 3.7: Imagem adaptada e traduzida de [64]. São exibidos pontos à partir de um conjunto de vetores de características de 2 dimensões extraídos de amostras de 2 classes diferentes, simbolizadas pelas cores vermelha e azul. A imagem da direita ilustra uma situação em que as características são separáveis, arranjo tipicamente resultado de modelos de classificação binária treinados com *Softmax* conforme [64]. A imagem da esquerda ilustra o objetivo da função *Center Loss*, que é gerar vetores de características discriminativos, onde a maior distância intraclasse ainda é menor que a menor distância extraclasse.

Para atingir estes objetivos os autores propuseram a função custo *center loss* formulada conforme a Equação 3-12 :

$$L_c = \frac{1}{2}\sum_{i=1}^{m}\|x_i - c_{y_i}\|_2^2 \tag{3-12}$$

onde x_i é o vetor de características da i-ésima amostra, y_i é a classe, ou identidade, desta amostra e c_{y_i} é o centro dos vetores de características faciais da classe y_i, com $\{c_{y_i}, x_i\} \in \mathbb{R}^d$. A expressão $\|x_i - c_{y_i}\|_2^2$ calcula a distância entre o vetor de características x_i e o seu respectivo centro c_{y_i}, penalizando o modelo quando esta distância for maior e contribuindo assim para forçar uma maior compactação dos vetores intraclasse. Os centros são computados pela média dos vetores de características de cada amostra e atualizado a cada iteração do treinamento.

Por se tratar de supervisão conjunta, o erro total calculado pelas duas funções custo, *Softmax Loss* e *Center Loss*, é definido na Equação 3-13:

$$L = L_s + \lambda L_c \tag{3-13}$$

sendo λ um parâmetro escalar para balancear a influência da *Center Loss* no erro total. Portanto a definição completa do custo no modelo proposto é dado pela Equação 3-14:

$$L = \frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{W_{y_i}^T f_i}}{\sum_{j=1}^{c} e^{W_j^T f_i}} + \frac{\lambda}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2 \tag{3-14}$$

Para ilustrar a capacidade da função *Center Loss* em promover a compactação intraclasse dos vetores de características, tornando-os mais discriminativos, foi realizado um experimento teste utilizando utilizando a coleção de dados MNIST [31] e com diferentes valores para o parâmetro λ. Os vetores de características, gerados a partir das amostras, possuem dimensão igual a 2 e cada cor simboliza uma das 10 classes desta coleção de dados. A Figura 3.8 apresenta a distribuição espacial dos pontos correspondentes aos vetores de características, evidenciando que os vetores das amostras de mesma classe permanecem mais próximos com o aumento do valor do λ.

Estes modelos utilizam a arquitetura de Redes Neurais Convolucionais denominada *Inception-ResNet-v1* [53] e foram implementados no *framework TensorFlow*[4].

3.2.4 Treinamento dos Modelos

Os modelos foram treinados em modo ajuste fino, de forma que todos os pesos de modelo de referência foram carregados antes do início do treinamento. Assim, o desempenho inicial do treino já é próximo ao estado da arte, com as diversas camadas da arquitetura com elevada capacidade de extração de características faciais úteis à identificação facial.

[4] https://www.tensorflow.org/

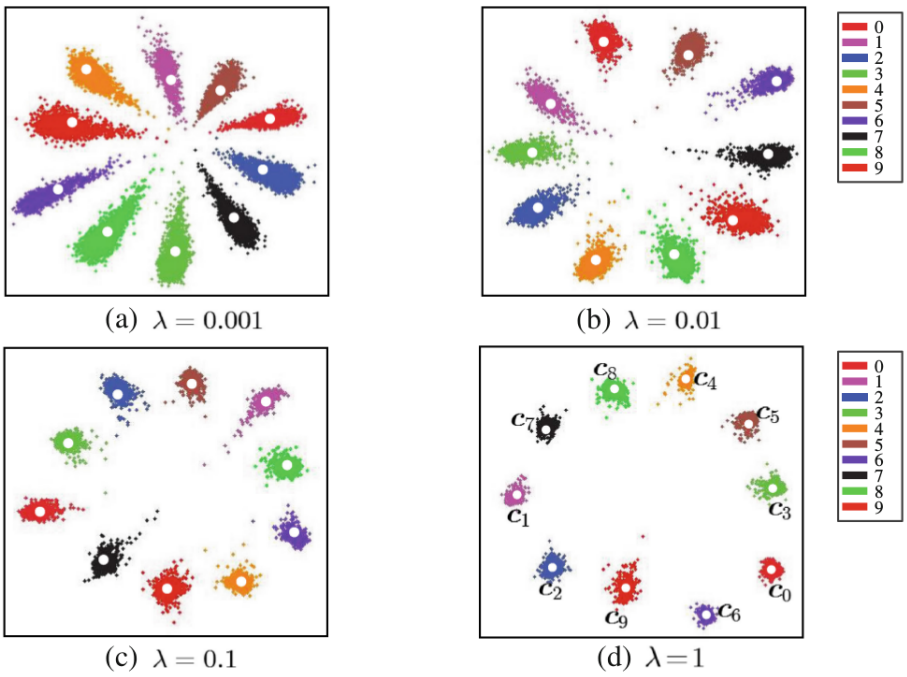


Figura 3.8: Imagem adaptada e traduzida de [64]. São exibidos pontos à partir do conjunto de vetores de características de 2 dimensões extraídos de amostras de 10 classes diferentes, simbolizadas por cores. Em (a) é possível observar que, embora sejam separáveis, as distâncias intraclasse são maiores e algumas inclusive excedem amostras de outras classes. Isso se deve ao fato do parâmetro λ possuir valor de apenas 0,0001, de forma que a *Center Loss* influenciou pouco no erro e portanto o resultado tendo sido gerado em grande parte pela supervisão da *Softmax*. Em (d) o λ possui valor 1, equilibrando então a influência das 2 funções custo e gerando assim a imagem em que as amostras apresentam reduzida variação intraclasse.

Com o objetivo de comparar o desempenho das funções custo da forma mais isolada foram utilizados os mesmos valores para todos os hiper-parâmetros comuns da arquitetura. Todos os modelos foram treinados no mesmo computador, equipado com Unidade de Processamento Gráfico *GTX-1070* com 8GB de *RAM* e 1920 núcleos *CUDA*. O tamanho do *batch* utilizado é de 60 amostras e empregando Gradiente Descendente Estocástico com *Momentum* [45] como otimizador, sendo o termo de momento γ igual a 0,9.

A taxa de aprendizagem η utilizada foi de 0,01 e como técnicas de regularização foram mantidas aquelas já empregadas no modelo de referência, sendo a regularização **L2** com taxa de decaimento de pesos com valor de 0,0001 e o *Dropout* na camada totalmente conectada, com probabilidade de manutenção de cada neurônio de 0,8.

Com a finalidade de determinar o momento correto de parada do treinamento foram adotados, como critérios de convergência, a curva de erro gerado pela função custo, a acurácia do conjuntos de teste e de validação.

Embora o número de distratores na galeria utilizado para avaliação dos resultados no Capítulo 4 seja de mais de 208.000 indivíduos, durante o treinamento foram utilizados 5.400. Essa redução foi utilizada para reduzir o custo computacional e tempo utilizado.

Os modelos supervisionados pela *Triplet Loss* foram treinados utilizando a estratégia de mineração de trios descrita em [50]. Seguindo esta estratégia foram selecionados, na coleção de imagens de treinamentos, trios de imagens compostos por duas amostras da mesma identidade, x_i^a e x_i^p e uma terceira de identidade diferente x_i^n que violaram a restrição imposta definida em 3-1 pela função custo . Essa seleção foi feita em cada época, durante o treinamento, de forma refletisse o desempenho atual do modelo e conforme os autores, atingisse a convergência mais rapidamente, com o erro estabilizando em torno do valor mais baixo.

O esquema de treinamento e testes pode ser visualizado na Figura 3.9. A camada totalmente conectada (*FC - Fully Conected*), situada após o último módulo convolucional, é utilizada para a extração dos vetores de características faciais. Como a função *Triplet Loss* dirige o treinamento para obtenção de vetores com separação interclasse por meio de distância quadrática, essa é a métrica utilizada durante os testes.

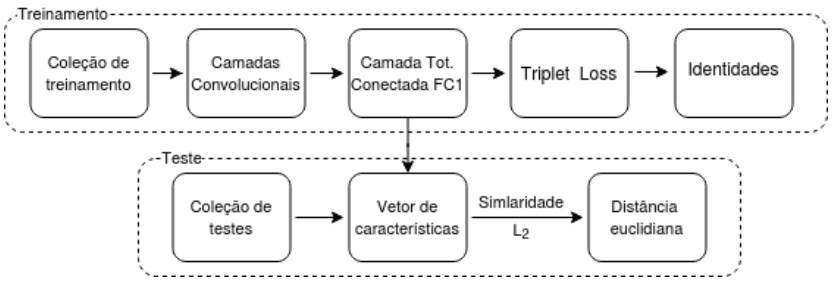
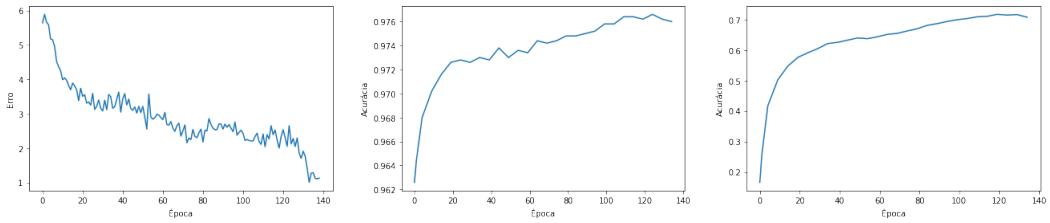


Figura 3.9: Esquema de treinamento e testes do modelo com supervisão *Triplet Loss*. Foram treinados 2 modelos, com as dimensões da camada *FC* 128d e 512d. Uma vez treinado, a camada posterior, com a função custo é dispensada e a saída da camada *FC* produz os vetores de características faciais.

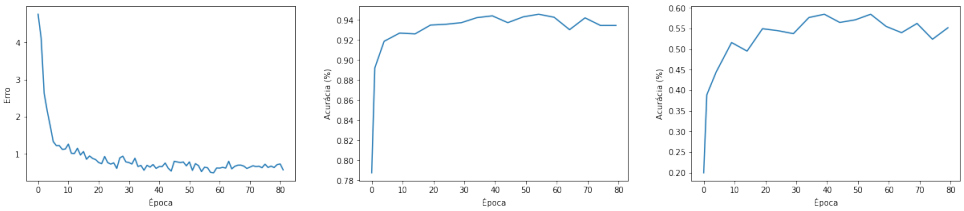
O modelo utilizando *Triplet Loss* com dimensão 128d dos vetores de características faciais foi treinado por 19 horas, totalizando 140 épocas. Conforme pode ser visualizado na Figura 3.10(a) o erro gerado pelo modelo alcançou menor valor após a época 135. Porém a acurácia do conjunto de validação em 3.10(b) se estabilizou em torno de 97,6% entre as épocas 100 e 135 chegando a cair um pouco após a época 138. A acurácia da coleção de testes, conforme Figura 3.10(c) ultrapassou 70% na época 100 e após a época 100 oscilou em torno de 71%, tendo tido como pico, na época 119, o valor de 77,7%.



(a) *Erro da função Triplet Loss* (b) *Acurácia no conjunto de validação* (c) *Acurácia no conjunto de testes*

Figura 3.10: (a) Erro calculado pela função *Triplet Loss*. (b) Acurácia do modelo no conjunto de validação. Após a época 15 percebe-se oscilação em torno do valor 97,2%. (c) Acurácia do modelo na coleção de testes, com valor oscilando em torno de 65%.

O modelo utilizando *Triplet Loss* com dimensão 512d dos vetores de características faciais foi treinado por 16 horas, totalizando 80 épocas. Conforme pode ser visualizado na Figura 3.11(a) o erro gerado pelo modelo estabilizou a partir da época 37, permanecendo em próximo ao valor 0,65. A acurácia do conjunto de validação em 3.11(b) atingiu o valor de 93% na época 24 com variações de $\pm 1\%$ até a época 80, tendo como valor máximo 94,5%. A acurácia da coleção de testes, conforme exibido na Figura 3.11(c) ficou acima de 57% entre as épocas 34 e 60, chegando ao valor máximo de 58,4%. Após a época 60 este indicador sofreu uma queda e permaneceu em torno de 55%.



(a) *Erro da função Triplet Loss* (b) *Acurácia no conjunto de validação* (c) *Acurácia no conjunto de testes*

Figura 3.11: (a) Erro calculado pela função *Triplet Loss* com dimensão de vetores de características faciais de 512d. (b) Acurácia do modelo no conjunto de validação. (c) Acurácia na coleção de testes, com valor oscilando em torno de 55% nas últimas épocas

Os modelos treinados com supervisão da função custo *AM-Softmax* seguiram esquema de treinamento e testes que pode ser visualizado na Imagem 3.12. A camada totalmente conectada (*FC - Fully Conected*), situada após o último módulo convolucional, é utilizada para a extração dos vetores de características faciais. Como a função *AM-Softmax* dirige o treinamento para obtenção de vetores com separação interclasse por meio de similaridade do cosseno, promovendo assim distanciamento angular, essa é a métrica utilizada durante os testes.

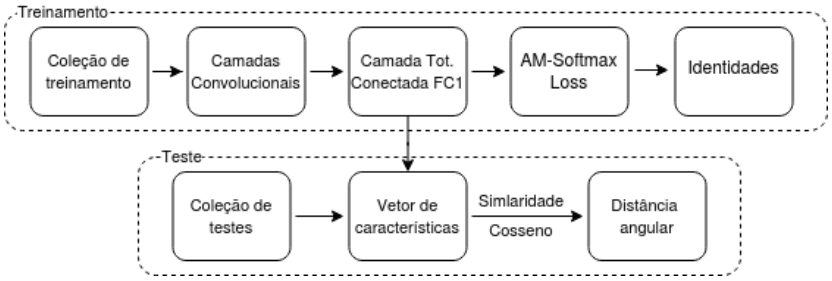
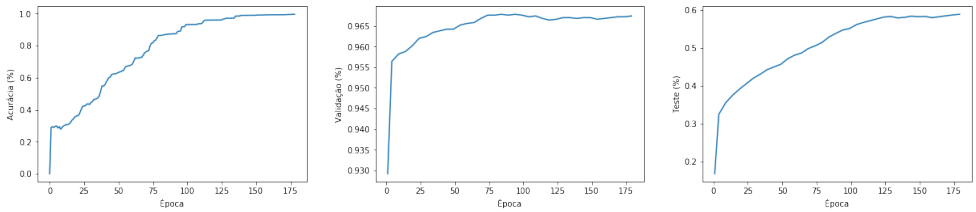


Figura 3.12: Esquema de treinamento e testes do modelo com supervisão *AM-Softmax*. Foram treinados 2 modelos, com as dimensões da camada *FC* 128d e 512d. Uma vez treinado, a camada posterior, com a função custo é dispensada e a saída da camada *FC* produz os vetores de características faciais. A similaridade do cosseno é utilizada para comparar as amostras da galeria e identificar quais pertencem à mesma identidade.

O modelo utilizando *AM-Softmax* como função custo e com dimensão 128d dos vetores de características faciais foi treinado por 5 horas, totalizando 178 épocas. Como critério para verificação da convergência também foi utilizado a acurácia no conjunto de treinamento. Conforme pode ser visualizado na Figura 3.13(a) a acurácia no conjunto de treinamento atingiu valor acima de 99% a partir da época 150, permanecendo neste patamar até o final. Porém a acurácia do conjunto de validação em 3.13(b) se estabilizou em torno de 96,5% entre as épocas 60 e 179, não ultrapassando 96,7%. Na coleção de testes, conforme Figura 3.13(c) atingiu 58% na época 120, com valor máximo de 58,7%, oscilando em torno de 58,5% até o final do treinamento.

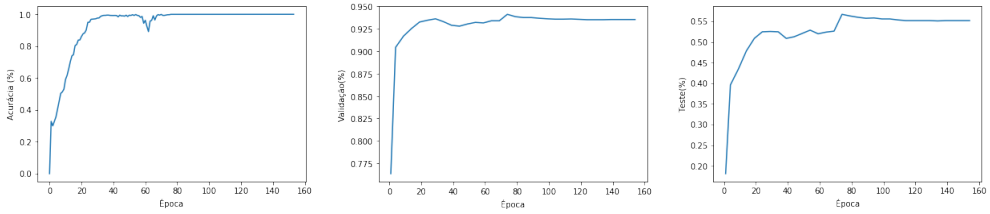


(a) *Acurácia na coleção de treino* (b) *Acurácia no conjunto de validação* (c) *Acurácia no conjunto de testes*

Figura 3.13: (a) Acurácia na classificação do conjunto de treino da função custo *AM-Softmax*. (b) Acurácia no conjunto de validação, apresentando oscilação em torno do valor 96,5% e valor máximo de 96,7%. (c) Acertos na coleção de testes, com valor oscilando em torno de 58% a partir da época 120.

O modelo utilizando *AM-Softmax* como função custo e com dimensão 512d dos vetores de características faciais foi treinado por cerca de 4 horas, totalizando 153 épocas. A acurácia no conjunto de treinamento atingiu valores acima de 99% após a época 34, com uma pequena oscilação entre as épocas 60 e 64, conforme exibido na Figura 3.14(a). O treino não foi interrompido porque, embora essa acurácia média acima

de 99% o desempenho nas coleções de teste e validação ainda estava aumentando, sendo que ambos indicadores atingiram os valores máximos após a época 90. A acurácia na coleção de validação atingiu o valor máximo de 94% na época 74 e após a época 90 se manteve em torno do valor 93,5% conforme Figura 3.14(b). A acurácia na coleção de testes, conforme Figura 3.14(c), atingiu 55% na época 70, com valor máximo de 56,7% na época 74 e oscilando em torno de 55% até o final do treinamento.



(a) *Acurácia na coleção de treino* (b) *Acurácia no conjunto de validação* (c) *Acurácia no conjunto de testes*

Figura 3.14: (a) Acurácia na classificação do conjunto de treino da função custo *AM-Softmax* com dimensão 512d na última camada totalmente conectada. Apesar de atingir estabilidade a partir da época 34 o treinamento não foi interrompido uma vez que o desempenho nas coleções de teste e validação ainda estavam subindo, indicando assim que o modelo ainda não estava em *overfitting*, o que, considerando as curvas de validação e testes, começou a ocorrer após a época 90.

3.2.5 Métricas de Avaliação

As tarefas de reconhecimento facial podem ser agrupadas em duas de acordo com o tipo de problema que destinam-se resolver. Cada uma dessas tarefas possui um conjunto próprio de métricas de avaliação de desempenho.

A verificação facial é a tarefa em que o sistema confronta diretamente duas faces com o objetivo de afirmar se pertencem à mesma pessoa. Este tipo de tarefa é mais útil em cenários de controle de acesso e não fará parte do escopo deste trabalho.

A identificação facial é o problema em que, dada uma imagem de um indivíduo até então desconhecido, o sistema deverá compará-la a todas imagens de uma galeria e apontar aquela como maior similaridade facial, identificando então o indivíduo. Esta é tarefa que será pesquisada neste trabalho.

Os protocolos de teste podem ser divididos em dois grupos: *closed-set* e *open-set* de acordo com [35]. O primeiro significa que todos os indivíduos utilizados no conjunto de teste também estão presentes no conjunto de treinamento, tornando assim o reconhecimento facial uma tarefa de classificação. O segundo significa que os indivíduos do conjunto de teste não fazem parte do treinamento. Para este trabalho será adotado

esse último protocolo, *open-set*, por este ser o mais adequado de acordo com os objetivos definidos em 1.2.

3.2.5.1 Rank-N

A métrica *Rank-N* é diretamente relacionada com a tarefa de identificação facial: dado uma imagem teste e considerando a galeria com ao menos uma foto diferente desta mesma pessoa, denominada imagem correspondente, o algoritmo ordena todas imagens da galeria com base na similaridade facial com esta imagem teste. Caso a imagem correspondente esteja entre as N primeiras imagens, considera-se um acerto. Formalmente, seja $C(n)$ o conjunto de amostras de teste em que a amostra correspondente está entre as n mais próximas:

$$C(n) = \{p_j : indice(p_j) \leq n\} \forall p_j \in P_g \tag{3-15}$$

onde $indice(p_j)$ é a função que retorna o índice da imagem correspondente da amostra de teste p_j na galeria de imagens ordenada pela similaridade facial em relação a p_j. O conjunto $C(r)$ contém apenas as amostras em que a posição da imagem correspondente é menor ou igual a n, dada por $indice(p_j) \leq n$. Todas as imagens da coleção de testes P_g são avaliadas.

Como os modelos treinados utilizados foram supervisionados por funções custos diferentes, a ordenação pela similaridade facial com cada imagem teste é realizada com base na métrica de distância reforçada por cada função custo. Portanto, a distância quadrática é utilizada para ordenação dos vetores de características faciais das imagens da galeria na avaliação dos modelos treinados pela *Triplet Loss*. E a similaridade do cosseno é utilizada quando os modelos em avaliação foram supervisionados pela função *AM-Softmax*.

Uma vez calculado $C(n)$, então é calculado o percentual de amostras do conjunto de teste P_g em que as amostras correspondentes da galeria foram posicionados entre as r primeiras:

$$P_I(n) = \frac{|C(n)|}{|P_g|} \tag{3-16}$$

Assim, o *Rank-N* pode ser interpretado como a probabilidade de que a imagem correspondente seja encontrada, a partir de uma imagem teste de mesma identidade, dentro de uma lista ordenada com as N imagens com maior semelhança facial.

3.2.5.2 Curva CMC

A curva CMC, ou Característica de Correspondência Cumulativa (*Cumulative Match Characteristic*), foi proposta em [15] e permite a visualização do desempenho, utilizando a métrica *Rank-N*, de um modelo na tarefa de identificação facial considerando diferentes valores de *N*. Portanto, são calculados os valores de P_I em 3-16 para todos os valores n até que atinja um valor máximo k. A curva *CMC* então é plotada na forma x P_I x Rank-N, sendo uma curva sempre crescente com valor máximo 1.

CAPÍTULO 4

Resultados

Os quatro modelos descritos no Capítulo 3 foram treinados, juntamente com o modelo de referência. Foram utilizados na aplicação das métricas definidas na Seção 3.2.5. Os resultados foram coletados considerando também diferenças entre os gêneros masculino e feminino com o objetivo de avaliar a existência de viés de dados por esta categorização.

Os testes foram executados utilizando diferentes números de distratores na galeria, com o objetivo de avaliar a escalabilidade dos modelos conforme [28, 35]. Foram utilizados subconjuntos, a partir do conjunto total de distratores composto por 208.187 amostras, selecionados de forma aleatória conforme e agrupados em repositórios numerados conforme Tabela 4.

	Repostório 1	Repositório 2	Repositório 3	Repositório 4	Repositório 5	Repositório 6	Repositório 7
Distratores	2.082	4.164	12.492	20.820	41.640	124.920	208.187

4.1 Rank-1

A Tabela 4.1 contém o percentual de acerto dos modelos nesta métrica, considerando o conjunto completo de distratores, Repositório 7, na galeria. Na primeira linha de resultados estão relacionadas as taxas de acerto do modelo supervisionado pela *Triplet Loss*, que obteve maiores taxas de acerto tanto no conjunto de validação quanto de testes. A primeira coluna contém a taxa de acerto na coleção de validação com os vetores de características de dimensionalidade 128, com 96,06% de acerto, e na coluna ao lado os resultado para dimensionalidade 512, por sua vez com acerto de 86,32%, sendo que os resultados obtidos com 128 dimensões superiores em quase 10%. As duas colunas seguintes contém a acurácia deste mesmo modelo na coleção de testes, tendo o modelo com vetor facial de 128d alcançado 52,64% de acerto, também superado em 20,93% o modelo que utiliza dimensionalidade de 512, que acertou 31,71% das amostras.

A segunda linha da Tabela 4.1 apresenta os resultados obtidos pelo modelo treinado com a supervisão da *AM-Softmax*. Nas duas primeiras colunas colunas os resultados

Função Custo	Validação		Teste	
	128d	512d	128d	512d
Triplet Loss	**96,06%**	86,32%	**52,64%**	31,71%
AM-Softmax Loss	93,46%	87,54%	38,98%	35,22%
Referência	91,22%	66,92%	0,16%	8,79%

alcançados na coleção de validação com vetores de características de dimensionalidade 128d e 512d, 93,46% e 87,54% respectivamente, observa-se a superioridade do primeiro em 5,92%. Também na coleção de testes essa função custo com espaço de características de 128d (38,98%) alcançou a maior taxa de acerto em relação ao espaço de 512d (35,22%) em 3,76%. com vetores de características de dimensionalidade.

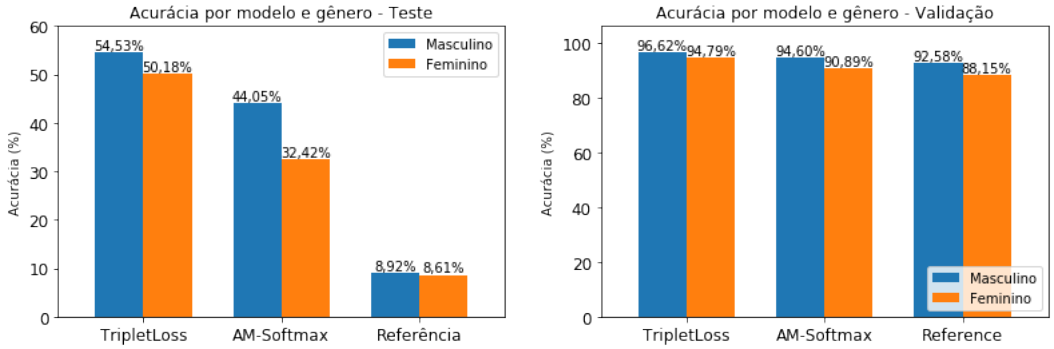
O modelo de referência, cujos resultados estão na última linha da tabela 4.1, obteve maior taxa de acerto na coleção de validação em sua versão de dimensionalidade de 128d com 24,3% mais acertos que o modelo de vetores de 512. Porém, no conjunto de imagens de teste o modelo com saída de 512d (8,79%) superou o modelo com vetores de características de 128d em 8,63%.

Considerando que:

- Os modelos de referência foram treinados na coleção *CASIA-WebFace*, que possui quase 8 vezes mais amostras que a utilizada neste trabalho
- Estes modelos de referência atingiram o estado na arte na coleção de testes *LFW*, tendo atingido 98,5% de taxa de acerto.
- Devido à miscigenação e a resultante dificuldade de se classificar precisamente as faces da coleção de dados brasileira de acordo com as divisão usual de etnias, este trabalho considera essa coleção como sendo de etnia única e própria conforme Seção 1.1.

é possível sugerir que estes modelos de referência possuem viés de dados quanto a etnia brasileira pelo fato do desempenho destes modelos terem melhorado significativamente com o treinamento em modo ajuste fino utilizando imagens faciais brasileiras.

A Figura 4.1 permite a visualização de gráficos de acurácia dos modelos agrupada por gênero nas diferentes coleções. Na Figura 4.1(b), são exibidos os gráficos relativos à coleção de validação. O modelo *Triplet Loss* obteve o melhor resultado para ambos os gêneros. No gênero masculino, representado pela barra de cor azul, esta função custo superou em 2,02% o modelo supervisionado pela *AM-Softmax* e em 4,04% o modelo de referência. O modelo supervisionado pela *Triplet Loss* também superou os outros modelos no gênero feminino, com taxa de acerto 3,9% superior à *AM-Softmax* e 6,64% o modelo de referência. Alcançou também a menor diferença de erro entre os gêneros, 1,83%, contra 3,71% e 4,43% dos modelos *AM-Softmax* e de referência respectivamente.



(a) *Desempenho na coleção de testes* (b) *Desempenho na coleção de validação*

Figura 4.1: A imagem (a) exibe o gráfico com os percentuais de acerto na métrica *Rank-1*, na coleção de testes e agrupado por gênero, de cada um dos modelos avaliados, sendo a função custo *Triplet Loss* alcançou melhor desempenho. Em (b) é exibido o gráfico com o desempenho dos modelos, também por gênero, no conjunto de dados de validação.

Nos gráficos da Imagem 4.1 foram detalhados os percentuais de acertos de acordo com o gênero de cada indivíduo das coleções de teste e de validação, também considerando métrica *Rank-1* e o conjunto completo de distratores, repositório 7. Nesta tabela foram utilizadas as versões dos modelos que apresentaram melhores resultados, considerando a dimensionalidade, de acordo com a Tabela 4.1. Portanto, para a avaliação dos modelos quanto ao gênero foram utilizados, tanto na coleção de teste quanto na de validação, os modelos treinados pelas funções custo *TripletLoss* e *AM-Softmax* com vetores de características faciais de 128 dimensões. No caso do modelo de referência, para a coleção de testes e validação foram utilizadas as versão com vetores de 512d e 128d, respectivamente.

O modelo treinado com a função custo *Triplet Loss* apresentou os melhores resultados para ambos os gêneros na coleção de testes, conforme gráficos da Figura 4.1(a). Este modelo alcançou taxa de acerto de 54,53% nas amostras do gênero masculino, representado pela cor azul na imagem, contra 44,05% de acurácia do modelo com *AM-Softmax* e 8,92% do modelo de referência, superado-os em 10,48% e em 45,61% respectivamente. Em relação ao gênero feminino, representado pela cor laranja, o modelo supervisionado pela *Triplet Loss* atingiu 44,05% de acerto, resultado 17,78% maior do que o modelo treinado com a *AM-Softmax*, que obteve 32,42% de acurácia, e 41,57% superior ao modelo de referência.

Em relação ao resultados da *AM-Softmax*, que apresentou segunda maior taxa de acerto, a *Triplet Loss*, além de alcançar desempenho significativamente superior em ambos os gêneros também foi mais eficaz para diminuir a diferença de erro entre eles, reduzindo assim mais expressivamente este tipo de viés que também foi observado e explicado na Seção 2.4. A diferença na taxa de acerto entre as amostras masculinas e

femininas do modelo *Triplet Loss* foi de 4,35%, enquanto que para o modelo *AM-Softmax* esse desequilíbrio foi de 11,63%, ou seja, $2,67\times$ maior, evidenciando assim a melhor capacidade de extração de características discriminativas independentes de gênero em modelos treinados pela *Triplet Loss* com geração *on-line* de triplets. Embora o conjunto de treinamento seja desbalanceado em relação à essa característica, sendo 71,4% das amostras de gênero masculino e 28,6% feminino conforme Tabela 3.1, a função custo *Triplet Loss* apresentou melhor capacidade de generalização e robustez de resultados em relação ao gênero, tendo a característica de treinamento combinatória empregada neste trabalho mostrado melhores resultados para a diminuição deste viés.

O desempenho de cada um dos modelos, considerando a variação do número de distratores da galeria conforme Tabela 4, é exibido nos gráficos da Figura 4.2. Os modelos utilizados nesta avaliação foram aqueles que apresentaram melhores resultados em cada uma das coleções de dados da Tabela 4.1. O objetivo é observar a variação da acurácia, na métrica *Rank-1*, de acordo com o aumento do número de amostras na galeria e verificar a escalabilidade de cada modelo.

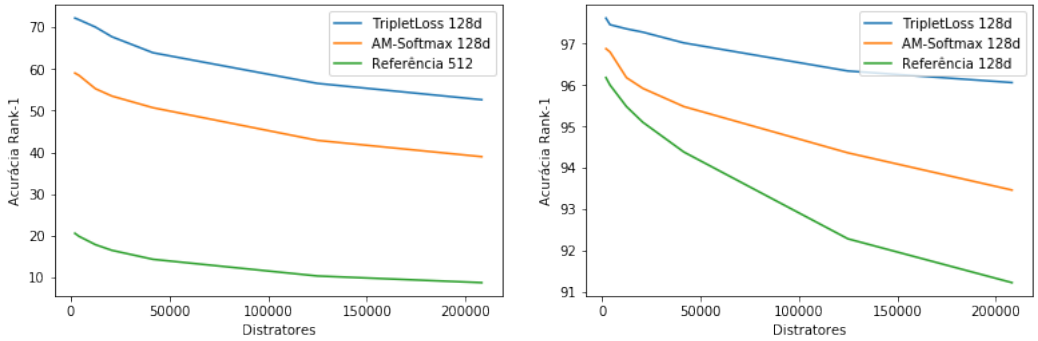
O modelo treinado com supervisão da função *TripletLoss* com vetores de características de 128d alcançou melhores resultados em todas as variações na quantidade de distratores. Na coleção de validação, gráfico da Figura 4.2(b), verifica-se que este modelo atingiu maior escalabilidade com o aumento do número de distratores, apresentando menor decaimento na acurácia, após o número de destes ultrapassar os 125.000, em relação aos outros modelos. O modelo treinado com a *AM-Softmax*, mesmo com desempenho menor, ainda superou substancialmente os resultados alcançados pelo modelo de referência.

Verifica-se também que a degradação do desempenho do modelo de referência aumenta substancialmente na medida que o número de distratores ultrapassa os 125.000 apresentando uma tendência de queda na acurácia mais acentuada. Portanto, é possível verificar que o viés de dados com relação a etnia contribui para uma maior tendência de aumento na taxa de erros na medida que aumenta o número de imagens faciais a serem comparadas.

No gráfico da Figura 4.2(a), que contém os resultados alcançados na coleção de testes, o modelo treinado com a *TripletLoss*, assim como na coleção de validação, superou os outros, com acurácia variando de 72,20% com o menor número de distratores até 52,63% com o número máximo. A função *AM-Softmax* apresentou taxa de decaimento de acurácia similar, porém com resultados menores, variando de 59,02% até 38,98%. O modelo de referência alcançou acurácia inicial de 20,61% e final de 8,78%.

De acordo com os resultados no conjunto de dados de validação, exibidos no gráfico da Figura 4.2(b), o modelo de referência obteve acurácia de 96,20%, apenas 1,44% menor que de melhor *Triplet Loss* (97,62%) no cenário com menor número de distratores. Porém, este modelo apresentou maior queda na acurácia com o aumento de distratores,

alcançado 91,22% de taxa de acerto com a quantidade máxima, com variação total de 4,96%. Este valor é 4,84% menor que o modelo *Triplet Loss*, que teve variação de 1,55%.



(a) *Desempenho na coleção de testes* (b) *Desempenho na coleção de validação*

Figura 4.2: A imagem (a) exibe o gráfico com a variação nos percentuais de acerto na métrica *Rank-1*, na coleção de testes, para cada um dos modelos, de acordo com a variação do número de distratores presentes na galeria. Em (b) é exibido o gráfico com os resultados na coleção de validação. Neste gráfico percebe-se maior degradação do modelo de referência em função do aumento no número de distratores.

4.2 Curva CMC

As curvas CMC da Figura 4.3 foram geradas a partir dos percentuais de acerto de cada um dos modelos com melhor desempenho, quando avaliados nas coleções de validação e testes, conforme Tabela 4.1. Foram coletadas as taxas de acerto do *Rank-1* ao *Rank-30* considerando a galeria com o número máximo de distratores, Repositório 7, conforme 3.1.

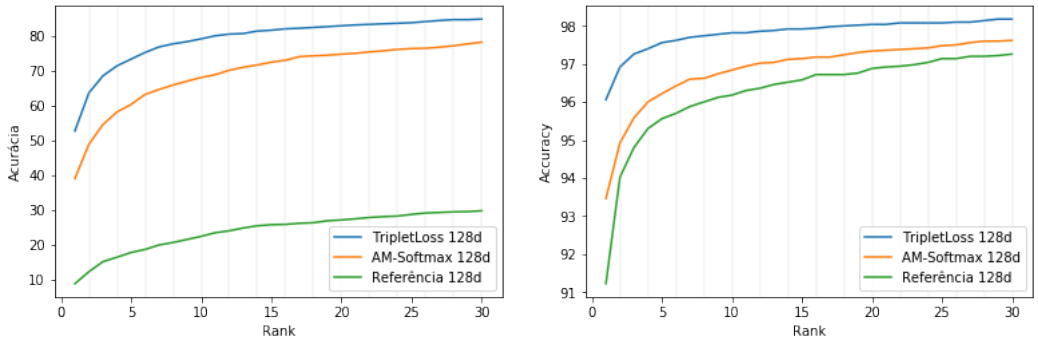
Na avaliação de *Rank-1*, Conforme 4.1, apenas a classe da amostra mais próxima, comparando com toda a galeria, é considerada. Caso seja da mesma identidade da amostra em teste, o acerto é computado. Na avaliação de *Rank-30* é considerado acerto se houver alguma amostra de mesma identidade do exemplo em teste dentre as 30 imagens mais próximas.

A curva CMC visualizada na Figura 4.3(a) foi gerada a partir dos acertos dos modelos utilizando a coleção de testes. Conforme o esperado, todos os modelos alcançaram melhores resultados com o aumento do *Rank*, tendo o modelo de referência apresentado as taxas de acerto mais baixas para todos os diferentes valores do *Rank*.

O modelo treinado com supervisão da função *Triplet Loss* obteve as maiores taxas de acerto em todos os cenários, seguido pelo modelo com função *AM-Softmax*. Ambos apresentaram crescimento na taxa de acerto parecidos com o aumento do *Rank*,

onde o modelo *Triplet Loss* variou de 52,63% no *Rank-1* até 84,74% no *Rank-30*, seguido pelo modelo *AM-Sostmax*, com variação de 38,98% até 78,11%.

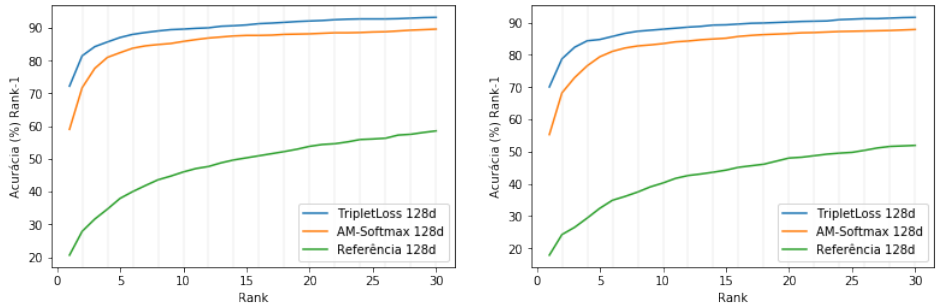
Considerando a mesma avaliação na coleção de validação, a acurácia alcançada pelo modelo treinado com *Triplet Loss* se manteve significativamente mais alta para valores menores do *Rank*, entre 1 e 15, sendo que após o *Rank-25* as acurácias deste modelo e o supervisionado pela *AM-Softmax* se mantiveram praticamente iguais.



(a) *Curva CMC na coleção de testes* (b) *Curva CMC na coleção de validação*

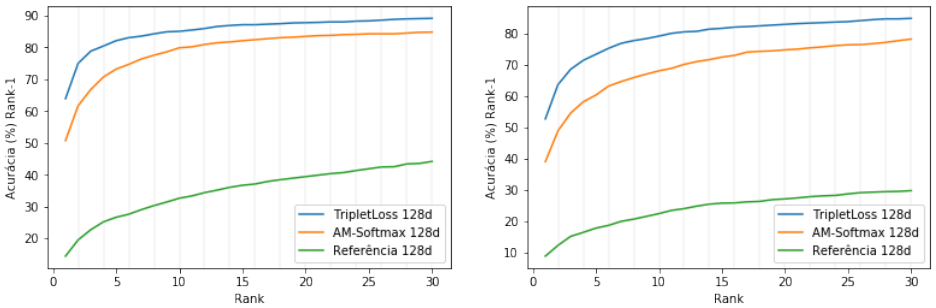
Figura 4.3: A imagem (a) exibe a curva de correspondência cumulativa, CMC, a partir das taxas de acerto obtidas pelos modelos em análise na coleção de testes, compreendendo os acurácias na *Rank-1* até *Rank-30*. Em (b) é exibida curva CMC gerada a partir da coleção de validação.

Com o objetivo de verificar o desempenho dos modelos, na coleção de testes, com a mesma variação do *Rank* mas com diferentes conjuntos de distratores na galeria, foram geradas as curvas CMC exibidas na Figura 4.4 com os totais de distratores sendo 2.082, 12.492, 124.920 e 208.187. De acordo com os gráficos, o modelo supervisionado pela *Triplet Loss* superou os outros modelos, independente da quantidade distratores e do *Rank* considerado.



(a) *Curva CMC na coleção de testes com 2.082 distratores*

(b) *Curva CMC na coleção de testes com 12.492 distratores*



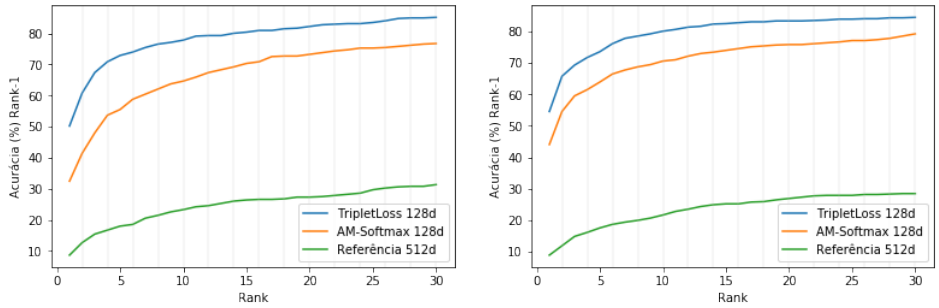
(c) *Curva CMC na coleção de testes com 124.920 distratores*

(d) *Curva CMC na coleção de testes com 208.187 distratores*

Figura 4.4: As imagens exibem as curvas CMC para cada conjunto de distratores diferentes e com variação de *Rank* de 1 a 30.

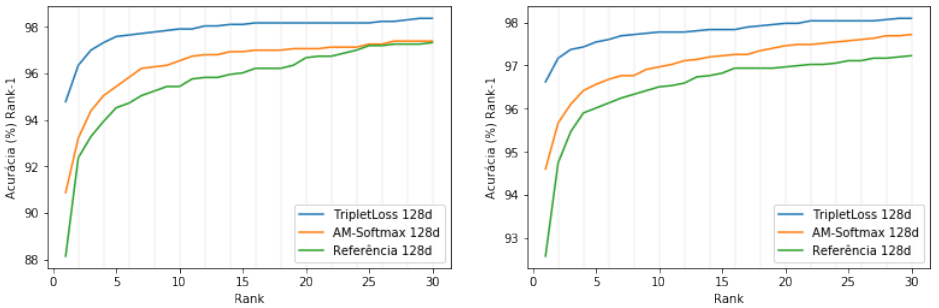
As curvas CMC da Figura 4.5 foram geradas a partir das acurácias dos modelos, nas coleções de validação e teste, agrupados por gênero. O objetivo é avaliar se, na medida que aumenta o *Rank*, os modelos apresentam desempenhos diferentes na tarefa de identificação facial para os gêneros masculino e feminino.

Considerando o desempenho na coleção de testes o modelo treinado com supervisão da função custo *Triplet Loss* apresentou acurácia mais alta em ambos os gêneros e para todos os valores de *Rank* . O modelo supervisionado pela *AM-Softmax* apresentou acurácia similar à *Triplet Loss* até o *Rank* 2, porém, a partir do *Rank* 3 este modelo gerou resultados com acurácia inferior, até o *Rank* 30.



(a) *Curva CMC, gênero feminino, coleção de testes*

(b) *Curva CMC, gênero masculino, coleção de testes*



(c) *Curva CMC, gênero feminino, coleção de validação*

(d) *Curva CMC, gênero masculino, coleção de validação*

Figura 4.5: As imagens exibem as curvas CMC agrupadas por gênero e com variação de *Rank* de 1 a 30.

CAPÍTULO 5

Conclusão

Neste trabalho foi investigada a capacidade da função custo *Triplet Loss* de atenuar o problema de viés de dados com a geração *on-line* de triplets através do treinamento supervisionado em modo de ajuste fino de um modelo pré-treinado baseado em Redes Neurais Profundas com resultados no estado da arte em sua publicação original. Como forma de validar os resultados também foi utilizada uma segunda função custo denominada *AM-Softmax* e que também atingiu acurácia no estado da arte. Também foram avaliadas, para cada uma das funções custo, duas dimensionalidades diferentes dos vetores de características faciais. Os modelos foram treinados até atingir a convergência, onde, a partir de então, não houve mais aumento da acurácia nas coleções de testes e validação.

Os modelos treinados, juntamente com o modelo de referência, foram avaliados nas coleções de teste e validação utilizando as métricas definidas e consistentes com os objetivos do projeto. Essas avaliações buscaram compreender o comportamento do desempenho destes modelos em termos de escalabilidade, com variação do número de amostras na galeria para comparação, e também quanto ao gênero, com o objetivo de identificar qual função custo traria melhores resultados e menor viés tanto étnico quanto de gênero.

Também foi estruturado neste trabalho coleções de dados de imagens faciais brasileiras para treinamento, testes e validação, a partir de bancos de dados privados e anotados por gênero.

5.1 Sumário das Principais Contribuições

Foi apresentada, no início desse trabalho, a hipótese primária elaborada com base na intuição de que a função custo *Triplet Loss* com geração de *triplets on-line* com base na identificação de imagens âncora, positiva e negativa que violam a restrição definida pela própria função custo poderia apresentar resultados mais significativos na diminuição do viés em relação à faces brasileiras existente no modelo de referência. Essa hipótese foi confirmada com base na comparação dos resultados alcançados pela função *AM-Softmax*

e pelo modelo de referência, com acurácia na métrica Rank-1 13,66% superior ao segundo melhor resultado, tendo sido assim o objetivo geral atingido.

A hipótese secundária, diretamente relacionada com a primeira, de que seria possível melhorar a acurácia de modelos de Reconhecimento Facial para a aplicação em imagens faciais brasileiras realizando treinamento de ajuste fino em modelos pré-treinados em coleções de dados públicas mesmo que já tenham alcançado o estado da arte também foi confirmada. Essa hipótese assumiu como o estereótipo facial brasileiro sendo uma etnia específica, não se encaixando nas categorias usuais e com características morfológicas próprias e diferentes de todas as outras. Tanto os resultados da métrica de Rank-1 quanto das curvas CMC demonstram que o treinamento em modo ajuste fino com coleção de dados de faces brasileiras melhorou sensivelmente o desempenho, mesmo com um limitado número de amostras para treinamento. A seguir lista os objetivos e os resultados alcançados, comprovando as hipóteses:

- Foi estruturado banco de dados de imagens faciais brasileiras com todas as identidades categorizadas quanto ao gênero.
- Utilizando o modelo de referência foram encontras as classes em que houve erro de identificação de facial por parte do modelo, onde o vetor de características faciais com menor distância em relação à imagem teste pertencia à outra identidade.
- Foram treinados 4 modelos diferentes, sendo 2 supervisionados pela função custo *Triplet Loss* e 2 supervisionados pela *AM-Softmax*, ambos com dimensionalidade dos vetores de características faciais de 128d e 512d.
- Foi realizado uma extensa coleta dos dados de desempenho dos modelos, com base nas métricas definidas e posterior comparação dos resultados. Nesse ponto verificou-se que o modelo supervisionado pela função custo *Triplet Loss*, com dimensionalidade 128, apresentou resultados melhores na métrica *Rank-1*, sendo esta a que representa maior grau de dificuldade na tarefa de identificação facial, por considerar, após o ordenamento de toda a galeria pela distância em relação à imagem teste, apenas a primeira amostra.
- Verificou-se também que a função custo *Triplet Loss* superou de forma bastante significativa a taxa de acerto para cada um dos gêneros da coleção de validação e testes. Além disso, essa função diminuiu consideravelmente a diferença nas taxas de acerto entre os gêneros, resultando em um erro $2,67\times$ menor que a *AM-softmax* e portanto com melhor de generalização e maior capacidade de extração de características discriminativas independentes de gênero, gerando ao final um modelo com maior equidade mesmo com coleção de dados de treinamento desbalanceada, com quase 72% de identidades masculinas.
- Concluiu-se também que este modelo apresentou maior escalabilidade, uma vez que apresentou melhores taxas de acerto com a variação do número de distratores

presentes na galeria.

- Os desempenhos dos modelos foram comparados também quanto à ordem do *Rank*, através da curva CMC, de forma que a acurácia de cada modelo pudesse fosse obtida considerando um número maior de amostras próximas à imagem teste, tendo a *Triplet Loss* também apresentado melhor desempenho, tanto nos conjuntos de validação quanto de testes.
- Todas as métricas também foram analisadas de forma agrupada por gênero e o modelo supervisionado pela *Triplet Loss* também apresentou taxa de acerto maior em todas as situações para ambos os gêneros.
- Finalmente, observou-se uma significativa melhora na acurácia dos modelos treinados com faces brasileiras, mesmo com reduzida quantidade de amostras. Na coleção de testes o modelo de referência apresentou taxa de acerto de 8,79% na métrica *Rank-1*, enquanto o modelo supervisionado pela *Triplet Loss* alcançou 54,53%. No conjunto de validação este modelo também apresentou evolução no desempenho, atingindo 96,06%, contra 91,22%.

Referências Bibliográficas

[1] BERG, A.; DENG, J.; FEI-FEI, L. **Large scale visual recognition challenge 2010, 2010**. *URL http://www. image-net. org/challenges/LSVRC/2010/index*, 2011.

[2] BUOLAMWINI, J.; GEBRU, T. **Gender shades: Intersectional accuracy disparities in commercial gender classification**. In: *Conference on Fairness, Accountability and Transparency*, p. 77–91, 2018.

[3] CAO, Q.; SHEN, L.; XIE, W.; PARKHI, O. M.; ZISSERMAN, A. **Vggface2: A dataset for recognising faces across pose and age**, 2017.

[4] CAO, Q.; SHEN, L.; XIE, W.; PARKHI, O. M.; ZISSERMAN, A. **Vggface2: A dataset for recognising faces across pose and age**. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, May 2018.

[5] CAVAIONI, M. **Deeplearning series: Convolutional neural networks**. URL: https://medium.com/machine-learning-bites/deeplearning-series-convolutional-neural-networks-a9c2f2ee1524.

[6] CHEN, B.; CHEN, C.; HSU, W. H. **Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset.** *IEEE Transactions on Multimedia*, p. 804–815, 2015.

[7] DE GEOGRAFIA E ESTATÍSTICA, I. B. **Censo demográfico do brasil**. https://sidra.ibge.gov.br/Tabela/3175#resultado, 2010. Acessado em: 23-03-2020.

[8] DENG, J.; ZHOU, Y.; ZAFEIRIOU, S. **Marginal loss for deep face recognition**. In: *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, p. 2006–2014, 07 2017.

[9] E. EIDINGER, R. ENBAR, T. H. **Age and gender estimation of unfiltered faces.** *IEEE Transactions on Information Forensics and Security*, p. 2170–2179, 2014.

[10] FAN, C. **Survey of convolutional neural network**, 2016.

[11] FIERREZ, J.; ORTEGA-GARCIA, J.; ESPOSITO, A.; DRYGAJLO, A.; FAUNDEZ-ZANUY, M. **Morph: Development and optimization of a longitudinal age progression database.** *Biometric ID Management and Multimodal Communication*, p. 17–24, 2009.

[12] FISWG. **Facial Identification Scientific Group**. https://fiswg.org/index.htm, acessado em abril de 2019, 2019.

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

[14] GROSS, R.; MATTHEWS, I.; COHN, J.; KANADE, T.; BAKER, S. **Multi-pie**. *Image and Vision Computing*, p. 807–813, 2010.

[15] GROTHER, P.; MICHEALS, R.; PHILLIPS, J. **Face recognition vendor teste 2002 performance metrics**. URL: https://www.hsdl.org/?view&did=438080, 2002.

[16] GUO, Y.; ZHANG, L.; HU, Y.; HE, X.; GAO, J. **MS-Celeb-1M: A dataset and benchmark for large scale face recognition**. In: *European Conference on Computer Vision*. Springer, 2016.

[17] HE, K.; ZHANG, X.; REN, S.; SUN, J. **Deep residual learning for image recognition**. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[18] HU, J.; SHEN, L.; SUN, G. **Squeeze-and-excitation networks**. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] HUANG, G. B.; LEARNED-MILLER, E. **Labeled faces in the wild: Updates andnew reporting procedures**. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, Maio 2014.

[20] IJIS INSTITUTE. **Law enforcement facial recognition use case catalog**, 2019. [Online; acessado em 04/04/2020].

[21] INSTITUTE OF AUTOMATION, C. A. O. C. **CASIA WebFAce**. http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html, acessado em abril de 2019.

[22] JACOBS, M. **Deep learning in five and a half minutes**. URL: http://www.videantis.com/deep-learning-in-five-and-a-half-minutes.html.

[23] JAIN, V.; LEARNED-MILLER, E. G. **Fddb: A benchmark for face detection in unconstrained settings**. *Technical Report UMCS-2010-009*, 2010.

[24] JOHNSON, J.; LI, F.-F.; KARPATHY, A. **Cs213n convolutional neural networks for visual recognition**. URL: http://cs231n.github.io/convolutional-networks.

[25] K. Q. WEINBERGER, J. B.; SAUL., L. K. **Distance metric learning for large margin nearest neighbor classification**. *MIT Press*, 2011.

[26] KANADE, T. **Picture processing system by computer complex and recognition of human faces**, 1974.

[27] KELLY, M. D. **Visual identification of people by computer**, 1970.

[28] KEMELMACHER-SHLIZERMAN, I.; SEITZ, S.; MILLER, D.; BROSSARD, E. **The megaface benchmark: 1 million faces for recognition at scale**, 2015.

[29] KLAR, B. F.; KLEIN, B.; TABORSKY, E.; BLANTON, A.; CHENEY, J.; ALLEN, K.; GROTHER, P.; MAH, A.; JAIN, A. K. **Pushing the frontiers of unconstrained face detection and recognition: Iarpajanus benchmark a**, 2015.

[30] KRIZHEVSKY, A.; SUTSKEVER, I.; E. HINTON, G. **Imagenetclassification with deep convolutional neural networks**. *Advances in neural information processing systems*, p. 1097–1105, 2012.

[31] LECUN, Y. **The mnist database of handwritten digits**. *http://yann.lecun.com/exdb/mnist/*, 1998.

[32] LECUN, Y.; BOSER, B.; OTHERS. **Backpropagation applied to handwritten zip code recognition.** *Neural Computation*, 1(4):541–551, 1989.

[33] LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. **Gradient-based learning applied to document recognition**. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[34] LIU, D. **A practical guide to relu**. URL: https://medium.com/tinymind/a-practical-guide-to-relu-b83ca804f1f7.

[35] LIU, W.; WEN, Y.; YU, Z.; LI, M.; RAJ, B.; SONG, L. **Sphereface: Deep hypersphere embedding for face recognition**. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[36] LIU, W.; WEN, Y.; YU, Z.; YANG, M. **Large-margin softmax loss for convolutional neural networks**, 2016.

[37] M. KÖSTINGER, P. WOHLHART, P. M. R.; BISCHOF, H. **Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization**. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, p. 2144–2151, 2011.

[38] M. M. NICHOLLS, O. CHURCHES, T. L. **Perception of an ambiguous figure is affected by own-age social biases**. *Scientific Reports*, 2018.

[39] MADIO, P. **A facenet-style approach to facial recognition on the google coral development board**. `https://towardsdatascience.com/a-facenet-style-approach-to-facial-recognition-dc0944efe8d1`. Acessado em: 23-03-2020.

[40] MERLER, M.; RATHA, N.; FERIS, R. S.; SMITH, J. R. **Diversity in faces**, 2019.

[41] MINSKY, M.; PAPERT, S. **Perceptrons: An Introduction to Computational Geometry**. MIT Press, Cambridge, MA, USA, 1969.

[42] NAGPAL, S.; SINGH, M.; SINGH, R.; VATSA, M.; RATHA, N. **Deep learning for face recognition: Pride or prejudiced?**, 2019.

[43] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. **Biometrics**, 2020. [Online; acessado em 01/04/2020].

[44] NUNES DA SILVA, I.; SPATTI, D. H.; FLAUZINO, R. A. **Redes Neurais Artificiais para Engenharia e Ciências Aplicadas**. Artliber Editora, 2010.

[45] QIAN, N. **On the momentum term in gradient descent learning algorithms**. *Neural networks : the official journal of the International Neural Network Society*, 12 1:145–151, 1999.

[46] RAJI, I. D.; BUOLAMWINI, J. **Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products**. *AAAI ACM Conf on AI Ethics and Society*, 2019.

[47] RANJAN, R.; CASTILLO, C. D.; CHELLAPPA, R. **L2-constrained softmax loss for discriminative face verification**, 2017.

[48] ROSENBLATT, F. **The perceptron: A probabilistic model for information storage and organization in the brain**. *Psychological Review*, p. 65–386, 1958.

[49] SANDBERG, D. **Face recognition using tensorflow**. `https://github.com/davidsandberg/facenet`. [Online; acessado em 01/04/2020].

[50] SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. **Facenet: A unified embedding for face recognition and clustering**, 2015.

[51] SRIVASTAVA, H. **A comparison based study on biometrics for human recognition**. *IOSR Journal of Computer Engineering*, 15(1):22–29, 2013.

[52] SUN, Y.; WANG, X.; TANG, X. **Deeply learned face representations are sparse, selective, and robust**. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

[53] SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V.; ALEMI, A. **Inception-v4, inception-resnet and the impact of residual connections on learning**, 2016.

[54] TEAM, G. **Github**. URL: `https://github.com/`.

[55] THE NEW YORK TIMES. **How the police use facial recognition, and where it falls short**, 2020. [Online; acessado em 04/04/2020].

[56] TRIGUEROS, D. S.; MEND, L.; HARTNETT, M. **Face recognition: From traditional to deeplearning methods**. *arXiv preprint arXiv:1811.00116*, 2018.

[57] TURK, M.; PENTLAND., A. **Eigenfaces for recognition**. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[58] WANG, F.; LIU, W.; LIU, H.; CHENG, J. **Additive margin softmax for face verification**, 2018.

[59] WANG, F.; XIANG, X.; CHENG, J.; YUILLE, A. L. **Normface: L2 hypersphere embedding for face verification**. *Proceedings of the 2017 ACM on Multimedia Conference - MM '17*, 2017.

[60] WANG, H.; WANG, Y.; ZHOU, Z.; JI, X.; GONG, D.; ZHOU, J.; LI, Z.; LIU, W. **Cosface: Large margin cosine loss for deep face recognition**, 2018.

[61] WANG, M.; DENG, W. **Deep face recognition: A survey**. *arXiv preprint arXiv:1804.06655*, 2019.

[62] WANG, M.; DENG, W.; HU, J.; TAO, X.; HUANG, Y. **Racial faces in-the-wild: Reducing racial bias by information maximization adaptation network**, 2018.

[63] WARREN S. MCCULLOCH, W. P. **A logical calculus of the ideas immanent in nervous activity**. *Bulletin of Mathematical Biophysics*, 1943.

[64] WEN, Y.; ZHANG, K.; LI, Z.; QIAO, Y. **A discriminative feature learning approach for deep face recognition**. In: Leibe, B.; Matas, J.; Sebe, N.; Welling, M., editors, *Computer Vision – ECCV 2016*, p. 499–515, Cham, 2016. Springer International Publishing.

[65] WIKIPEDIA CONTRIBUTORS. **Cosine similarity Wikipedia, the free encyclopedia**, 2020. [Online; acessado em 01/04/2020].

[66] WU, X.; HE, R.; SUN, Z.; TAN, T. **A light cnn for deep face representation with noisy labels**, 2015.

[67] XIAO, H.; RASUL, K.; VOLLGRAF, R. **Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms**, 2017.

[68] Y. TAIGMAN, M. YANG, M. R.; WOLF, J. **Deepface: Closing the gap to human-level performance in face verification**. *IEEE Conf on CVPR*, p. 815–823, 2014.

[69] ZHANG, K.; ZHANG, Z.; LI, Z.; QIAO, Y. **Joint face detection and alignment using multitask cascaded convolutional networks**. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[70] ZHUANG, F.; QI, Z.; DUAN, K.; XI, D.; ZHU, Y.; ZHU, H.; XIONG, H.; HE, Q. **A comprehensive survey on transfer learning**. *Proceedings of the IEEE*, p. 1–34, 2020.

[71] ZOU, J.; SCHIEBINGER, L. **Ai can be sexist and racist—it's time to make it fair**, 2018.