Aplicação de Redes Neurais na Detecção de Símbolos da Língua Brasileira de Sinais (Libras)

Ana Carolina Zhang. Heloisa Mariani Rodrigues.
Rooney Ribeiro Albuquerque Coelho.

Faculdade de Estudos Interdiciplinares da PUC-SP
Rua Monte Alegre, 984, Perdizes - São Paulo - SP CEP: 05014-901. Edifício Cardeal Mota – Sala P 68.
(e-mail: rracoelho@pucsp.br)

Abstract: In this study, we explored strategies to detect Brazilian Sign Language (Libras) signs, facing significant challenges. Initially, using a Convolutional Neural Network (CNN) with the UEFS dataset, we encountered inaccuracies due to the low resolution of the images. When applying the CNN to the ROBOFLOW dataset, the results were still unsatisfactory, indicating the complexity of the problem and the presence of overfitting. We chose to integrate MediaPipe with the ROBOFLOW data, achieving an accuracy of 93.66%. The results underscore the importance of technological innovation in promoting inclusion and accessibility for the deaf community, with the potential to significantly improve the quality of life and inclusion of deaf individuals.

Resumo: Neste estudo, exploramos estratégias para detectar sinais da Língua Brasileira de Sinais (Libras), enfrentando desafios significativos. Inicialmente, usando uma Rede Neural Convolucional (CNN) com o conjunto de dados da UEFS, enfrentamos imprecisões devido à baixa resolução das imagens. Ao aplicar a CNN ao conjunto de dados do ROBOFLOW, os resultados ainda não foram satisfatórios, indicando a complexidade do problema e a presença de overfitting. Optamos por integrar a MediaPipe com os dados do ROBOFLOW, alcançando uma acurácia de 93,66%. Os resultados destacam a importância da inovação tecnológica na promoção da inclusão e acessibilidade para a comunidade surda, com potencial para melhorar significativamente a qualidade de vida e a inclusão dos surdos.

Keywords: Brazilian Sign Language (Libras); Sign detection; Convolutional Neural Network (CNN); MediaPipe; Accessibility; Assistive technology; Deaf community; Inclusion; Gesture recognition; ROBOFLOW; MLP (Multilayer Perceptron).

Palavras-chaves: Língua Brasileira de Sinais (Libras); Detecção de sinais; Rede Neural Convolucional (CNN); MediaPipe; Acessibilidade; Tecnologia assistive; Comunidade surda; Inclusão; Reconhecimento de gestos; ROBOFLOW; MLP (Multilayer Perceptron).

1. INTRODUÇÃO

Em 2002, com a Lei Federal nº 10.436 de 24 de abril de 2002, a Língua Brasileira de Sinais, foi oficialmente reconhecida, como meio legal de comunicação e expressão das comunidades surdas brasileiras e foi regulamentada, em 2005, por meio do Decreto Federal nº 5626 promulgado no dia 22 de dezembro de 2005.

De acordo com o livro “Se Liga nos Sinais” de Jadson Nunes, foi o pedagogo espanhol Juan Pablo Bonet (1579- 1629), quem propôs o alfabeto manual. Já o francês Jacob Rodrigues Pereira (1715-1780) modificou e aprimorou o método proposto por Bonet (1620) e introduzindo o alfabeto manual: pontuação, acentuação e números. Pode-se concluir que a língua brasileira de sinais tem sua origem a partir da língua de sinais francesa, entretanto, por ser uma língua viva e em uso, no decorrer dos tempos sofreu alterações naturais e resultou no que hoje se conhece como Libras.

Segundo o artigo “O alfabeto manual como recurso para a incorporação de elementos do português na formação de sinais em libras”, UFRGS, a educação bilíngue para surdos prioriza a língua de sinais (Libras) como língua materna e o Português como segunda língua. Incentivada desde a educação infantil, essa abordagem visa à aquisição natural da Libras em escolas específicas para surdos, seguida pelo ensino do Português como segunda língua. O alfabeto manual da Libras é fundamental para soletrar palavras, identificar nomes próprios, citar obras artísticas e endereços, além de ser usado quando um sinal específico é desconhecido.

A Libras, sendo a língua da comunidade surda brasileira, incorpora elementos do Português devido ao constante contato entre as duas línguas. O alfabeto manual é um dos mecanismos pelos quais os surdos sinalizantes "tomam emprestados" itens lexicais da língua oral, representando manualmente a forma escrita das palavras.

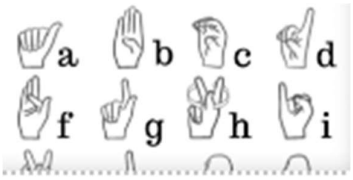
A falta de acessibilidade em diversas mídias, como televisão, internet e mídias sociais, representa um obstáculo significativo para as pessoas surdas no acesso à informação e comunicação. A ausência de legendas, intérpretes de língua de sinais em programas de TV e conteúdos online não adaptados dificulta

sua plena integração na sociedade. Além disso, os serviços de saúde muitas vezes não estão preparados para atender às necessidades específicas da comunidade surda. A escassez de intérpretes de língua de sinais em consultórios médicos e hospitais pode criar barreiras na comunicação entre profissionais de saúde e pacientes surdos. Essas lacunas na acessibilidade prejudicam o direito fundamental das pessoas surdas de receberem informações e serviços de saúde de forma adequada e compreensível.

A criação de um intérprete automático para auxiliar a comunidade surda é de extrema importância, pois oferece acesso imediato à comunicação e informação em tempo real. Essa tecnologia proporcionaria maior independência aos surdos em várias situações cotidianas, incluindo saúde, educação e interações sociais. Além disso, um intérprete automático poderia ajudar a reduzir as disparidades sociais ao garantir uma participação mais igualitária na sociedade. Em resumo, essa inovação tem o potencial de melhorar significativamente a qualidade de vida e a inclusão dos surdos, promovendo uma sociedade mais acessível e equitativa.

O Brasil emergiu como protagonista no campo das traduções ao anunciar a disponibilidade da Tradução do Novo Mundo da Bíblia Sagrada em língua brasileira de sinais (Libras) no jw.org e no aplicativo JW Library Sign Language. Hamilton Vieira, membro da Comissão de Filial do Brasil, revelou essa conquista durante um programa transmitido ao vivo do auditório da filial brasileira, alcançando uma audiência de mais de 36.300 pessoas. Essa é a primeira vez que uma tradução completa da Bíblia em Libras é disponibilizada, tornando-se a terceira Bíblia completa em língua de sinais no mundo.

O projeto de tradução teve início em 2006, com a tradução do Evangelho de Mateus, seguido pelo progressivo lançamento dos demais livros da Bíblia à medida que eram traduzidos. Essa realização representa não apenas um marco significativo para a comunidade surda brasileira, mas também destaca a expertise e o compromisso do Brasil em promover a acessibilidade e a inclusão por meio das traduções.



Fomos impulsionados por uma necessidade urgente de abordar a escassez de intérpretes de Libras na sociedade, reconhecendo a importância da inclusão e acessibilidade para a comunidade surda. Inspirados por uma pesquisa anterior realizada pela Universidade Estadual de Feira de Santana (UEFS), que iniciou investigações sobre o reconhecimento de gestos da Libras, sentimos a motivação para aprimorar ainda mais o modelo existente.

Conscientes dos desafios enfrentados por indivíduos surdos na comunicação cotidiana e na acessibilidade a serviços essenciais, nosso objetivo era desenvolver uma solução mais eficaz e abrangente para facilitar a compreensão e a comunicação entre pessoas surdas e ouvintes. Essa motivação intrínseca impulsionou nosso compromisso em melhorar continuamente o sistema de reconhecimento de gestos, buscando proporcionar um impacto positivo tangível na vida da comunidade surda.

2. MULTILAYER PERCEPTRON (MLP)

MLP (Multilayer Perceptron) é uma arquitetura de rede neural composta por várias camadas de neurônios totalmente conectadas, onde cada neurônio em uma camada está conectado a todos os neurônios na camada subsequente. No entanto, as MLPs enfrentam limitações significativas quando lidam com imagens de entrada de baixa resolução. A explosão do número de parâmetros, a incapacidade de considerar a estrutura espacial dos dados e a exigência de memória e poder computacional são desafios críticos.

Para mitigar essas limitações, é comum aplicar pré-processamento externo, como redimensionamento ou técnicas de redução de dimensionalidade, antes de alimentar a rede. No entanto, para tarefas de processamento de imagem, arquiteturas mais avançadas, como CNNs (Convolutional Neural Networks), são preferíveis, pois são projetadas para capturar estruturas espaciais e padrões locais de forma mais eficaz.

O artigo Reconhecimento de Gestos Estáticos aplicados à Língua de Sinais Brasileira, UEFS inclui a criação de um conjunto de dados de imagens, tal como a aplicação de redes neurais junto a uma estratégia de segmentação de pele para a entrada do modelo. A abordagem do artigo representou um avanço significativo, visando reconhecer uma ampla gama de gestos em Libras e disponibilizar publicamente o conjunto de dados levantado.

Observou-se que os sinais com as taxas mais baixas de reconhecimento foram aqueles em que a primeira etapa não alcançou os melhores resultados, sugerindo que uma distribuição mais eficiente dos grupos poderia melhorar os resultados.

Com isso, Utilizamos o conjunto de dados fornecido pela UEFS, composto por imagens de gestos em Libras. O conjunto de dados é dividido em várias partes, isto é, diferentes mãos em arquivos diferentes denominadas "fold1", "fold2", e assim por diante. Cada uma dessas partes representa uma divisão dos dados para fins de treinamento, validação e teste.

Dentro de cada "fold", encontramos subpastas adicionais, denominadas "alfabetos" e "números". Essas subpastas representam as diferentes categorias de gestos da Língua Brasileira de Sinais (Libras) presentes no conjunto de dados.

3. CONVOLUTIONAL NEURAL NETWORKS (CNN)

As Redes Neurais Convolucionais (CNNs) são uma categoria especializada de redes neurais profundas amplamente empregadas em tarefas de visão computacional, especialmente

em análise de imagens. As CNNs assumem que os dados de entrada são imagens e apresentam camadas convolucionais, de pooling e funções de ativação que introduzem não-linearidades.

A estrutura básica de uma CNN inclui camadas convolucionais que aplicam filtros sobre a imagem de entrada para produzir imagens filtradas com valores modificados. As camadas de pooling reduzem a amostragem, diminuindo o tamanho da imagem através da seleção do valor máximo ou mínimo em uma determinada região. Além disso, as CNNs podem incluir camadas totalmente conectadas, seguindo o mesmo princípio das Redes Neurais convencionais.

A arquitetura das CNNs permite uma redução no número de pesos necessários para processar imagens de entrada, tornando-as eficientes para tarefas de reconhecimento de padrões em imagens. Essa capacidade de redução de parâmetros é uma das razões pelas quais as CNNs se destacam em comparação com outros modelos de redes neurais em problemas de visão computacional.

Embora uma alta acurácia seja geralmente desejável em modelos de CNN, é importante reconhecer que uma acurácia alta nem sempre indica um desempenho ideal. Uma das principais razões para uma alta acurácia com previsões incorretas pode ser devido ao desequilíbrio entre as classes no conjunto de dados. Se uma classe é significativamente mais representada do que outras, o modelo pode aprender a prever predominantemente a classe majoritária, resultando em uma acurácia global alta, mas com previsões incorretas para as classes minoritárias.

Além disso, problemas como overfitting, ambiguidade nos dados e características enganosas podem contribuir para previsões incorretas, mesmo em modelos com alta acurácia. Portanto, é crucial realizar uma análise abrangente dos resultados, considerando outras métricas de desempenho e investigando casos específicos de previsões incorretas para identificar e corrigir possíveis falhas no modelo.

4. METODOLOGIA

A jornada para desenvolver um sistema eficaz de detecção de sinais da Libras foi marcada por uma série de desafios e iterações. Inicialmente, ao adotar a abordagem baseada em MLP proposta pelo projeto da Universidade Estadual de Feira de Santana (UEFS), esperávamos obter resultados satisfatórios.

No entanto, logo ficou claro que essa abordagem apresentava limitações significativas em termos de tempo de processamento e sensibilidade à resolução da imagem e posição da mão. Em busca de uma solução mais robusta, voltamos nossa atenção para as Redes Neurais Convolucionais (CNNs), reconhecidas por sua eficácia na classificação de imagens.

Embora a CNN tenha proporcionado uma melhoria na velocidade e independência em relação à resolução da imagem, ainda enfrentamos desafios, principalmente devido à baixa resolução das imagens e às previsões incorretas.

Isso nos levou a explorar a base de dados do ROBOFLOW e, posteriormente, a adotar a MediaPipe como uma solução abrangente para a detecção de gestos da Libras. Ao combinar a robustez da base de dados do ROBOFLOW com a precisão da MediaPipe, conseguimos alcançar resultados promissores, destacando a importância de abordagens mais sofisticadas e adaptáveis para resolver problemas complexos de detecção de gestos.

A capacidade da MediaPipe em compreender e interpretar as características da mão foi fundamental para o aumento significativo na precisão das detecções, demonstrando sua versatilidade e eficácia em aplicações de reconhecimento de gestos. Essa jornada de aprendizado e aperfeiçoamento refletem o compromisso em superar obstáculos e encontrar soluções inovadoras para atender às demandas específicas do reconhecimento de sinais da Libras.

5. RESULTADOS

Para avaliar a eficácia da abordagem proposta que utiliza a CNN em conjunto com os datasets da plataforma ROBOFLOW e do projeto da UEFS, conduzimos uma série de testes e análises detalhadas das métricas de acurácia. Além disso, exploramos a geração de características da mão obtidas através das imagens da ROBOFLOW para uma compreensão mais aprofundada do desempenho do modelo. A seguir, apresentamos os resultados obtidos e discutimos suas implicações.

5.1 Resultados com a Abordagem da CNN e os Datasets

Ao aplicar a abordagem da CNN aos datasets da plataforma ROBOFLOW e do projeto da UEFS, observamos resultados com acurácia correta, porém com previsões incorretas. Uma análise detalhada revelou que nos datasets do projeto da UEFS, as imagens são de baixa resolução, o que influencia a detecção da imagem. Além disso, durante o treinamento no dataset da plataforma ROBOFLOW, embora tenha sido utilizado um arquivo em formato XLSX, as previsões foram incorretas, devido ao overfitting.

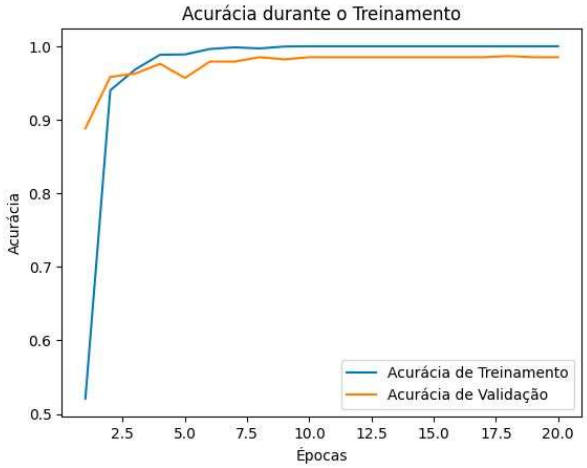
5.2 Análise de métricas de acurácia

Durante o treinamento dos modelos utilizando os dataframes de ROBOFLOW e UEFS, observamos padrões distintos de acurácia ao longo das épocas.

No caso dos dataframes da UEFS, observamos uma tendência diferente. Desde o início do treinamento, o modelo demonstrou uma acurácia substancialmente mais alta, começando em torno de 41,37% na primeira época e alcançando uma taxa de acurácia de 100% nas épocas subsequentes. No entanto, mesmo com acurácia correta, observamos previsões incorretas. Uma análise mais detalhada revelou que nos datasets do projeto da UEFS, as imagens são de baixa resolução, o que pode influenciar negativamente na detecção e interpretação dos sinais da Libras.

Por outro lado, ao analisar os resultados do treinamento com os dataframes de ROBOFLOW, verificamos um aumento gradual na acurácia durante as épocas de treinamento. No início do processo, a acurácia foi registrada em

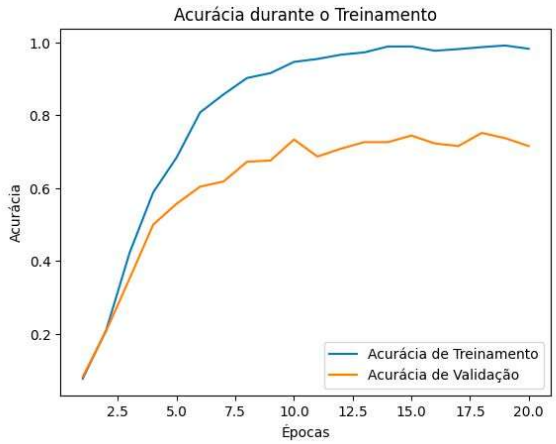
aproximadamente 5,86%, mas rapidamente apresentou uma trajetória ascendente, alcançando uma taxa de acurácia de 95,95% ao final das 20 épocas. Esse crescimento consistente indica uma adaptação eficaz do modelo aos dados, porém observamos uma tendência preocupante de overfitting.



(fig.01 - Treinamento de dataframe UEFS)

O trabalho da UEFS utilizou de validação cruzada com seis dobras, sendo a taxa média de reconhecimento para os 40 sinais de 96,77%, indicando uma eficácia significativa da abordagem. A limitação dessa abordagem é a utilização de imagens em baixa resolução e a necessidade de realizar um pré-processamento nos dados.

Por outro lado, ao utilizar o conjunto de dados da UEFS, nossa abordagem utilizando CNN alcançou uma taxa de acurácia de 98,66%, demonstrando uma performance competitiva em relação ao trabalho da UEFS. Nossa abordagem também enfrentou desafios semelhantes ao lidar com imagens de baixa resolução, exigindo pré-processamento dos dados para garantir resultados satisfatórios.



(fig.02- Treinamento de dataframe ROBOFLOW)

Esses resultados demonstram um aumento progressivo na acurácia durante o treinamento, tanto para os dataframes de ROBOFLOW quanto para os da UEFS. No entanto, é importante notar que a acurácia de validação nem sempre acompanhou a acurácia de treinamento, sugerindo a possibilidade de overfitting em alguns casos. Ainda assim, os resultados finais mostram altas taxas de acurácia tanto no treinamento quanto na validação, indicando a eficácia das abordagens utilizadas.

5.3 Adoção do Média Pipe

Diante desses desafios, optamos por adotar a MediaPipe, uma ferramenta capaz de extrair características da mão com base nos dataframes do ROBOFLOW. Para implementar a detecção de sinais da Libras, construímos um pipeline que integra técnicas de processamento de imagem, redução de dimensionalidade e classificação. Inicialmente, utilizamos o algoritmo de Principal Component Analysis (PCA) para redução da dimensionalidade, seguido de um classificador Random Forest. Durante o treinamento e validação do modelo, ajustamos os hiperparâmetros do classificador usando um Grid Search, a fim de encontrar a combinação ótima de parâmetros que resultasse em melhor desempenho. Após o treinamento, alcançamos uma acurácia de 93,67%. Além disso, observamos que o modelo foi capaz de realizar detecções corretas, demonstrando sua eficácia na interpretação e reconhecimento de gestos.

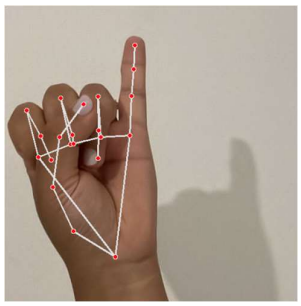
```
# Criar um pipeline com PCA e um classificador
pipeline = Pipeline([('scaler', StandardScaler(with_mean=False)),
                     ('pca', PCA(n_components=50)),
                     ('classifier', RandomForestClassifier(random_state=42))])
```

(fig.03 - algoritmo PCA e classificador Random Forest)

```
Acurácia:  0.9366515837104072
```

(fig.04 - Acuracia de modelo)

Letra classificada: “I”



(fig.05 - detecção correta de imagem inserida)

Carregamos uma nova imagem e extraímos as características da mão usando o MediaPipe. Essas características incluem os pontos de referência da mão e a previsão do gesto ou palavra em língua de sinais. O resultado da previsão demonstra que o MediaPipe conseguiu identificar com sucesso as características da mão e associá-las corretamente com as palavras em língua de sinais correspondentes. Ao desenhar os pontos de referência da mão detectados na imagem e exibir a

previsão da palavra em língua de sinais, conseguimos compreender visualmente o resultado da previsão do modelo.

Esse resultado destaca o MediaPipe como uma ferramenta poderosa, que possui alta confiabilidade no reconhecimento de gestos e tradução de língua de sinais. Isso é significativo para nossa pesquisa e aplicações, pois oferece um método viável para o reconhecimento automático e tradução de língua de sinais, proporcionando melhores meios de comunicação e interação para pessoas com deficiência auditiva.

6. CONCLUSÕES

Nossa pesquisa sobre a detecção de sinais da Língua Brasileira de Sinais (Libras) nos conduziu por uma jornada de exploração e descoberta, enfrentando desafios significativos ao longo do caminho. Inicialmente, ao tentarmos empregar uma Rede Neural Convolucional (CNN) com os dados da UEFS, nos deparamos com previsões imprecisas devido à baixa resolução das imagens. Mesmo ao estendermos o uso da CNN para o conjunto de dados do ROBOFLOW, os resultados ainda não alcançaram a excelência desejada, destacando a complexidade inerente ao problema.

O MediaPipe é uma solução eficiente em termos de custo computacional para o processamento de imagens e vídeos em tempo real. Desenvolvido pelo Google, o MediaPipe oferece uma variedade de modelos pré-treinados e ferramentas para análise de mídia, incluindo detecção de objetos, reconhecimento facial, rastreamento de mãos e muito mais.

Ao otimizar o uso de recursos computacionais e implementar algoritmos eficientes, o MediaPipe pode ser executado em uma ampla gama de dispositivos, desde smartphones até sistemas embarcados, oferecendo um desempenho satisfatório mesmo em hardware com recursos limitados. Isso o torna uma escolha atraente para aplicativos que requerem baixo custo computacional, mas que ainda exigem capacidades avançadas de visão computacional em tempo real.

Nossa decisão de integrar a MediaPipe com os dados do ROBOFLOW nos trouxe um avanço significativo. Alcançamos uma impressionante acurácia de 93,66% com previsões pr ecisas, representando um marco notável no campo da tecnologia assistiva e na promoção da acessibilidade para pessoas surdas ou com deficiência auditiva. Este resultado ressalta não apenas a eficácia dessa abordagem, mas também a importância da inovação tecnológica no fortalecimento e na inclusão desses indivíduos na sociedade.

É crucial reconhecer que o caminho para esses resultados não foi fácil. Enfrentamos desafios técnicos, como a baixa resolução das imagens e o overfitting, que exigiram uma abordagem adaptável e sofisticada. No entanto, nossa determinação em superar esses obstáculos e encontrar soluções inovadoras foi recompensada com sucesso.

Agora, mais do que nunca, é evidente o impacto transformador que a tecnologia pode ter na vida das pessoas com deficiência auditiva. Ao proporcionar meios mais eficazes de comunicação e interação, estamos não apenas promovendo a inclusão, mas também capacitando esses indivíduos a se engajarem plenamente na sociedade.

Este estudo destaca a importância contínua da pesquisa e do desenvolvimento de tecnologias acessíveis e inclusivas. Ao continuarmos avançando nesse campo, estamos construindo um futuro mais igualitário e acessível para todos.

REFERÊNCIAS

Nunes, Jadson. Se Liga nos Sinais. Nogueira, André, and Clovis Batista De Souza. O Alfabeto Manual Como Recurso Para a Incorporação de Elementos Do Português Na Formação de Sinais Em Libras.

Bastos, Igor L. O. Angelo, Michele F. Loula, Angelo C. Reconhecimento de Gestos Estáticos aplicados a Linguagem Brasileira de Sinais (Libras).