UNIVERSIDADE FEDERAL DA FRONTEIRA SUL
CAMPUS DE CHAPECÓ
CURSO DE CIÊNCIA DA COMPUTAÇÃO

DIOGO BALTAZAR DO NASCIMENTO

DATASET E MODELOS PARA RECONHECIMENTO DA LÍNGUA LIBRAS ATRAVÉS DE PROCESSAMENTO DE VÍDEOS

CHAPECÓ
2024

DIOGO BALTAZAR DO NASCIMENTO

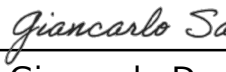
DATASET E MODELOS PARA RECONHECIMENTO DA LÍNGUA LIBRAS ATRAVÉS DE PROCESSAMENTO DE VÍDEOS

Trabalho de conclusão de curso apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal da Fronteira Sul.

Orientador: Prof.Dr. Giancarlo Dondoni Salton

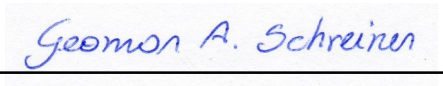
Este trabalho de conclusão de curso foi defendido e aprovado pela banca avaliadora em: 28/07/2024.

BANCA AVALIADORA



Prof.Dr. Giancarlo Dondoni Salton – UFFS

Prof.Dr. Guilherme Dal Bianco - UFFS



Prof.Dr. Geomar André Schreiner - UFFS

Dataset e Modelos para Reconhecimento da Língua Libras Através de Processamento de Vídeos

Diogo Baltazar do Nascimento
Universidade Federal da Fronteira Sul
Campus Chapecó
diogo.nascimento@estudante.uffs.edu.br

Giancarlo D. Salton
Universidade Federal da Fronteira Sul
Campus Chapecó
gian@uffs.edu.br

Abstract

No Brasil, mais de 10,7 milhões de pessoas têm deficiência auditiva, enfrentando desafios significativos para sua inclusão. Desde a infância, enfrentam dificuldades na integração familiar, educacional e profissional. Escolas e profissionais despreparados e uma sociedade que muitas vezes opta pela exclusão contribuem para altas taxas de evasão escolar. Embora centros para surdos sejam importantes, podem isolá-los ainda mais socialmente. Dessa forma, a presente pesquisa tem como objetivo desenvolver um sistema que faz uso de tecnologias e técnicas de processamento de vídeos para o reconhecimento e interpretação da Língua Brasileira de Sinais (Libras), visando assim promover a inclusão e a acessibilidade para a comunidade surda em interação com a população não surda. A abordagem adotada se baseia no uso de técnicas de visão computacional, usando o framework *TensorFlow* para treinamento e inferência de arquiteturas de aprendizado profundo, para que, dessa forma, o algoritmo possa aprender e tentar traduzir o sinal em Libras para texto em português. A criação de um sistema baseado em tecnologias de processamento de vídeos e aprendizado profundo pode representar um avanço na inclusão da comunidade surda no Brasil, integrando a Língua Brasileira de Sinais com a população não surda.

1 Introdução

Desde os primórdios da humanidade, a comunicação foi uma necessidade para estabelecer relações entre os indivíduos de maneira ordenada e, com isso, buscar um entendimento e crescimento da sociedade e da socialização (de Barros et al., 2021), no entanto, para pessoas surdas, a comunicação pode se tornar um desafio adicional. A Língua Brasileira de Sinais (Libras), é uma língua de modalidade gestual-visual, ou seja, tem como meio de comunicação os gestos, expressões faciais e corporais, dessa forma esta pesquisa tem como objetivo analisar e apresentar uma abordagem de tradução desta para texto em português, a fim de justamente promover a inclusão, acessibilidade e de facilitar a comunicação entre a comunidade surda e a não surda.

Nesse contexto, a tecnologia desempenha um papel crucial para atingir esse feito, usando técnicas de visão computacional e aprendizado de máquina que permitam a Libras ser usada e traduzida em tempo real (LORETO, 2023), como por exemplo o *VLibras*, que é um conjunto de ferramentas gratuitas e de código aberto que traduz conteúdos digitais (texto, áudio e vídeo) de Português para Libras, tornando computadores, celulares e plataformas Web mais acessíveis para as pessoas surdas (vli).

A plataforma de aprendizagem de Libras oferece benefícios significativos ao promover acessibilidade e conscientização. Ao tornar a língua visual acessível, desempenha um papel fundamental na inclusão de surdos na sociedade e melhora a comunicação entre ouvintes e surdos (LORETO, 2023). A Libras é o meio que permite a comunicação entre os surdos e não surdos, mas acaba sendo uma barreira quando não é apresentada de forma bilíngue (Miranda et al., 2020). Os ouvintes ignorantes no assunto chegam a imaginar que o surdo é incapaz de fazer tudo e que não é como ele, logo não consegue trabalhar, não tem momentos de lazer e que é incapacitado (Magno). Apesar de especialistas e da própria comunidade surda defenderem o ensinamento de libras em escolas tanto bilíngues quanto convencionais para promover o envolvimento e inserção dos mesmos na sociedade, e das tentativas de promover a presença de intérpretes em eventos públicos, como palestras programas de televisão, essa parcela da população ainda se vê negligenciada e de lado pela comunidade não surda por justamente não ter conhecimento para reconhecer os sinais. É com o intuito de romper essa barreira sociocomunicativa o objetivo da pesquisa, em analisar um conjunto de sinais em vídeo e realizar a

tradução dos gestos da Língua Brasileira de Sinais (Libras) para texto em português.

Ressalta-se que a comunicação é um aspecto fundamental para a interação social e o desenvolvimento pessoal. Para a comunidade surda, a língua de sinais, e, neste caso a Língua Brasileira de Sinais (Libras) desempenha um papel fundamental na facilitação da comunicação para com pessoas ouvintes. Grande parte dos hospitais (públicos e particulares), laboratórios, clínicas, postos de saúde, etc., conforme (dos Santos et al., 2021), ainda não oferecem o padrão necessário para atender pessoas com deficiência auditiva. O maior empecilho para um atendimento de qualidade é a impossibilidade de comunicação em Libras (DIF). Dessa forma, principalmente devido à falta de conhecimento e compreensão da Libras pela comunidade ouvinte e a falta de uma abordagem que facilite a comunicação entre ambas pode acabar resultando em uma barreira comunicativa significativa. Exatamente por isso, é essencial adotar estratégias para que os surdos consigam ter mais opções de lazer e cultura usando a Libras como principal recurso de comunicação (DIF).

Desta forma, este trabalho tem como objetivo apresentar uma proposta de uma arquitetura de rede neural para tradução de Libras para texto, utilizando técnicas de visão computacional e aprendizado profundo explorando uma abordagem para abordar essa barreira sociocomunicativa através de uma aplicação prática. A pesquisa tem como objetivo explorar recursos de tradução em Libras para texto, permitindo que pessoas ouvintes se comuniquem de maneira mais eficaz com a comunidade surda, sem precisar recorrer à mímicas, promovendo assim meios para que a comunicação entre ambos e a igualdade de informação seja facilitada. Para isso, será utilizado as bibliotecas *OpenCV* e *TensorFlow*, que são amplamente reconhecidas e utilizadas na área de visão computacional. O OpenCV é uma biblioteca de código aberto, que nos proporciona recursos para a captura de uma imagem em tempo real, ou vídeos e imagens armazenadas no computador facilitando manipulação e processamento de técnicas de pré-processamento de imagem, como, reescala, canais de cores ou algoritmos prontos para remoção de fundo (Bradski, 2000). Já o Tensorflow, é uma interface para expressão de algoritmos de aprendizado de máquina e sua implementação para execução desses algoritmos, permite a criação ampla de algoritmos incluindo treinamento e inferência para modelos de redes neurais profundas essenciais para a leitura e treinamento de detecção de sinais em Libras (Abadi et al., 2015a).

O presente trabalho está organizado da seguinte forma: na Seção 2 apresentamos Língua LIBRAS e na Seção 3, apresentamos os modelos de redes neurais utilizados nos experimentos, principais itens de revisão necessários para o entendimento deste trabalho; na Seção 4, relatamos os trabalhos relacionados; os procedimentos metodológicos e os experimentos realizados são relatados na Seção 5 e Seção 6, respectivamente; os resultados e a sua análise são demonstrados na Seção 7; e, finalmente, na Seção 8 apresentamos nossas conclusões e apontamentos para trabalhos futuros.

2 Língua Brasileira de Sinais

A LIBRAS, como toda Língua de Sinais, é uma língua de modalidade gestual-visual porque utiliza, como canal ou meio de comunicação, movimentos gestuais e expressões faciais que são percebidos pela visão; portanto, diferencia-se da Língua Portuguesa, que é uma língua de modalidade oral-auditiva por utilizar, como canal ou meio de comunicação, sons articulados que são percebidos pelos ouvidos (Ramos, 2004). Na LIBRAS, através do estudo dos seus processos de formação de palavras, pode-se constatar que há várias configurações de mãos que, constituindo seu sistema de flexão verbal para gênero animado/inanimado, sempre estão presas a uma raiz verbal, não ocupando uma posição sintagmática independente.(Felipe, 2006)

A Libras também possui muitos verbos denominais ou substantivos verbais que possuem a mesma forma para os pares verbo/substantivo (Felipe, 2006). Tendo suas características linguísticas como, fonologia, morfologia, sintaxe e semântica, sendo por esse motivo considerada uma língua autêntica. Assim, a fonologia da língua de sinais é constituída pelos seguintes parâmetros que formam os sinais: a Configuração de Mãos (CM), o Ponto de Articulação (PA), o Movimento(M), a Expressão facial e/ou corporal, a Orientação/Direção, conforme demonstrado na Figura 1.

3 Redes Neurais Artificiais

Uma rede neural artificial (RNA) tem duas facetas elementares: a arquitetura e o algoritmo de aprendizagem. Essa divisão surge naturalmente pelo paradigma como a rede é treinada.(Rauber, 2005)

Tendo sua origem inspirado no cérebro humano,

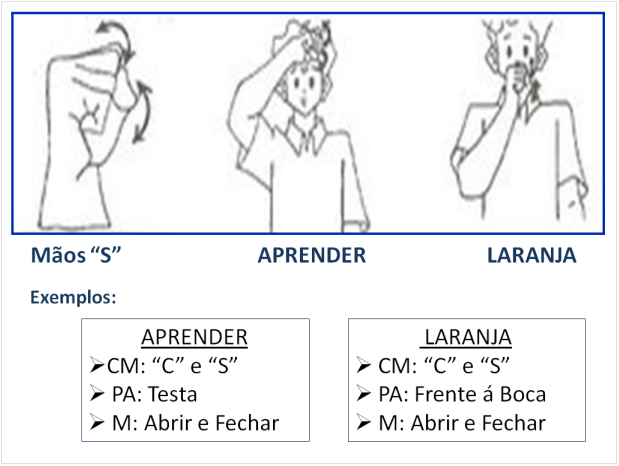


Figure 1: Parâmetros da Libras.

utiliza neurônios que se conectam e comunicam entre si, através de pesos e funções de ativação que ativam os neurônios. A saída é calculada a partir da soma dos produtos de entrada e peso que passam através da função de ativação. Esses modelos processam dados complexos para facilitar decisões inteligentes com pouca intervenção humana.

3.1 CNN

As Redes Neurais Convolucionais são o modelo de rede de aprendizado profundo mais conhecidos e utilizados atualmente. O conceito de CNNs foi originado da grande capacidade de redes neurais com várias camadas realizarem abstrações de alto nível. Conseguindo assim, modelar um grande conjunto de dados através de transformações lineares e não lineares. Consequentemente, amplificando a sua capacidade de aprendizado e proporcionando uma grande performance na classificação de dados (Bengio et al., 2013).

Com seu foco em classificação de imagens, detecção de objetos ou segmentação de imagens, adotam componentes, chamado de filtros, onde são aplicados a entrada para extrair recursos como, bordas, texturas ou formas. O uso de arquiteturas CNN faz a rede ser rápida de treinar, através da arquitetura ConvNet que foi inspirada no cérebro humano, os neurônios focam em uma região no campo visual, conhecida como Campo Receptivo, uma região sobrepõe a outra até cobrir toda a área visual, conforme Figura 2. Uma Rede Neural Convolucional (CNN) processa dados de grade, como imagens, usando convolução e pooling para destacar características. Camadas de convolução aplicam filtros à imagem para gerar mapas de características. Camadas de pooling reduzem a dimensionalidade preservando características importantes. Funções de ativação introduzem não linearidades. Camadas totalmente conectadas processam recursos para tarefas específicas. Dropout e normalização melhoram o desempenho. A saída da CNN é usada para classificar objetos ou segmentar imagens.

3.2 RNN

Redes neurais recorrentes são tradicionalmente utilizadas para processar sequências de informações textuais, onde se busca compreender, por exemplo, as correlações entre a ordem de ocorrência de palavras de uma frase. Em uma interpretação mais ampla, as palavras podem ser compreendidas como variáveis categóricas, que no contexto deste trabalho correspondem aos elementos que compõem cada evento (Capanema et al., 2020).

RNNs são adequadas para dados que possuem uma sequência que pode ser cuja ordem dos dados interpretada como um fator temporal, sendo essa sequência por exemplo, reconhecimento de fala, geração de música, análise de sentimento e reconhecimento de atividade em vídeo. As Redes Neurais Recorrentes (RNNs) possuem uma arquitetura com loops, conforme ilustrado na figura 3, que mantém informações anteriores, influenciando o processamento. Os dados são fornecidos sequencialmente, com cada passo de tempo tendo uma entrada específica. A RNN calcula saídas com base na entrada atual e estado oculto anterior, capturando padrões temporais. O estado oculto é atualizado em cada passo de tempo com uma função de ativação não linear.

As RNNs são eficazes para processar sequências de dados por sua capacidade de manter informações através do tempo mas enfrentam dificuldades com problema de explosão e desaparecimento de gradiente, tornando ineficientes para capturas de dependência de longo prazo. Para resolver esse problema, foi projetada uma nova arquitetura de RNN, a Long Shot-Term Memory (LSTM)(Hochreiter and Schmidhuber, 1997). O LSTM contém células de memórias capazes de manter informações por longos períodos e são controladas por três tipos de portas: entrada, esquecimento e saída como ilustrado na imagem 4). A célula de memória é essencial em uma unidade LSTM, mantendo e atualizando informações por longos períodos para evitar decaimento do gradiente. A porta de esquecimento decide o que manter ou descartar na memória, enquanto a porta de entrada decide as

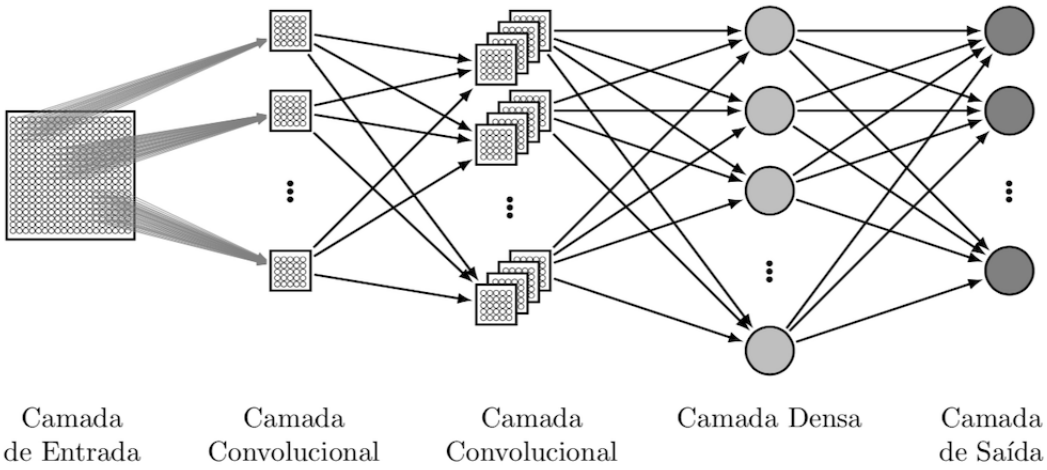


Figure 2: Exemplo de uma rede neural convolucional.

novas informações a serem adicionadas. A atualização da célula combina esses resultados, mantendo partes importantes do estado anterior e adicionando novas informações. Por fim, a porta de saída determina a saída da unidade LSTM, usando o estado atual da célula de memória modificado.

A célula de memória é responsável por preservar as informações importantes ao longo da execução, a porta de entrada controla quais informações devem entrar na célula de memória, a porta do esquecimento decide quais informações devem ser mantidas na célula de memória e quais devem ser esquecidas, e a porta de saída define qual parte da célula de memória deve ser calculada para a saída atual da rede. Essa arquitetura permite que a rede mantenha informações relevantes por longos períodos evitando assim a explosão e desaparecimento de gradiente.

3.3 CNN x RNN

As RNNs são comumente usadas como modelos de linguagem, enquanto as CNNs são usadas principalmente para processamento de imagens e vídeos. (Mogren, 2016). As CNNs são objetivas quando tem que trabalhar com imagens e vídeos, embora consigam um bom trabalho com áudio e texto. São principalmente usadas nas áreas de visão computacional e processamento de imagem, como, classificação de objetos e detecção de objetos, como exemplos mais famosos, detecção facial e detecção de objetos para veículos autônomos.

As RNNs tem seu foco em dados sequenciais, através de sua capacidade de desenvolver a compreensão contextual de sequências. Portanto são usados em reconhecimento de fala e processamento de linguagem natural, como, geração de resumos e traduções. Então é possível estarmos juntando as duas para estarem trabalhando, por exemplo, na criação de legendas de um vídeo, onde a CNN pode extrair dados sobre os quadros de vídeo e a RNN usando esses dados extraído pode escrever as legendas.

Ao combinar o uso de uma CNNs e RNNs podemos criar uma nova arquitetura, sendo ela a CNN-RNN em cascata, essa combinação pode usar a extração de características dos dados de entrada e os resultados dessa extração são usados pela RNN para modular a dependência sequencial dos dados. Trazendo vantagens como, ao combinar CNNs e RNNs, a arquitetura consegue explorar simultaneamente padrões espaciais e temporais, pode ser aplicada a problemas onde uma arquitetura não consegue aprofundar, por exemplo, ao classificar um conjunto de imagens podemos separar em sub grupos e identificar sem conhecimento especializado.

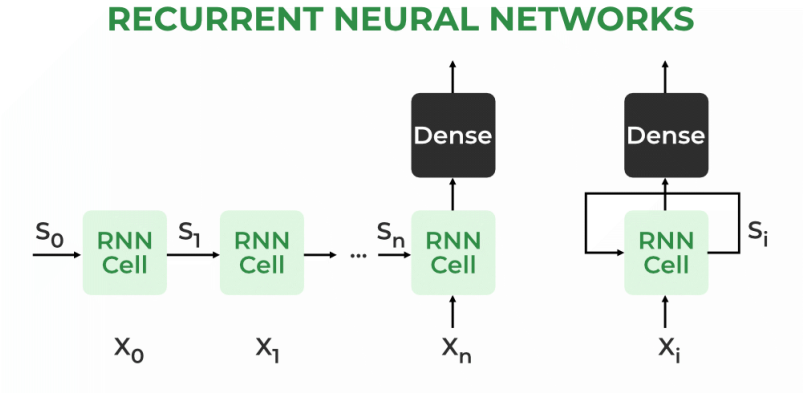


Figure 3: Exemplo de uma rede neural recorrente.

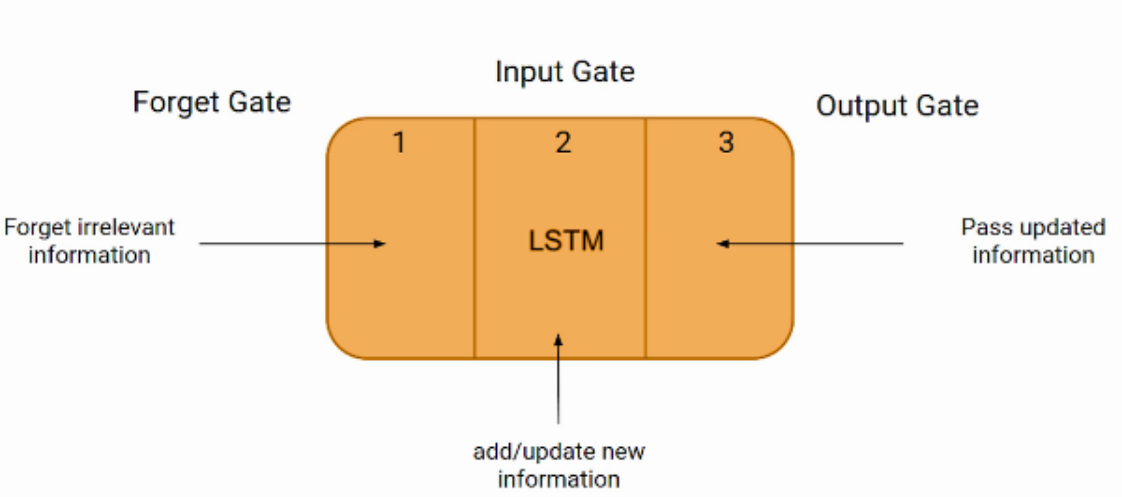


Figure 4: Exemplo de uma LSTM.

Como o gerador CNN-RNN pode gerar simultaneamente os rótulos grosseiros e finos, nesta parte comparamos ainda mais seu desempenho com redes 'específicas grosseiras' e 'específicas finas'. (Guo et al., 2018) Com essa arquitetura ao passar um vídeo, a arquitetura CNN consegue retirar as características em cada quadro e a arquitetura RNN a cada quadro usa as características passadas pela CNN e analisar a sequência do contexto do vídeo, assemelhando muito ao aprender a LIBRAS, onde olhamos cada movimento nosso cérebro remove características principais a cada movimento e analisa a sequência do movimento para aprender o sinal.

4 Trabalhos Relacionados

Em Porfirio et al. (2013), os autores apresentam uma abordagem para o reconhecimento de configurações de mão da Língua Brasileira de Sinais (LIBRAS) a partir de malhas 3D. Foi utilizada uma base de dados com vídeos de diferentes usuários realizando cada configuração de mão várias vezes. As imagens da mão foram segmentadas, pré-processadas e usadas para gerar malhas 3D utilizando a técnica de *Shape from Silhouette*. O reconhecimento das configurações de mão foi realizado utilizando um classificador SVM e características extraídas através do método de Harmônicos Esféricos. Os resultados demonstraram uma taxa média de acerto de 96,83 com Rank 3, evidenciando a eficiência do método proposto.

Para desenvolvimento da malha 3D foi utilizado *Blender*. Executando técnicas para formar o deslocamento de imagem 2D para 3D assim formando a silhueta da mão e usando algoritmos de remapeamento 3D é possível reorganizar a malha de modo que o mesmo objeto possa ser criado com uma estrutura tridimensional mais organizada e suave.

Já em Camgoz et al. (2018), o texto discute a diferença entre o Reconhecimento de Língua de Sinais (RLS) e a Tradução de Língua de Sinais (TLS). Enquanto o RLS se concentra em recon-

hecer sequências de sinais contínuos, negligenciando a estrutura gramatical e linguística da língua de sinais, a TLS busca gerar traduções em linguagem falada levando em consideração as diferenças gramaticais. Os pesquisadores propõem o uso da Tradução Neural de Máquina (TNM) para abordar o problema, utilizando conjuntos de dados específicos para avaliar o desempenho dos modelos de tradução. Os resultados mostram um limite superior para o desempenho da tradução e os resultados alcançados pelos modelos desenvolvidos pelos pesquisadores. O texto destaca a importância da pesquisa futura nesse campo em desenvolvimento.

Através de uso de redes neurais *CNN* e *RNN* para um modelo *attention-based encoder-decoders*.

Finalmente, em Silva et al. (2022), a solução proposta faz o rastreamento da estrutura linguística da libras, separando e processando cada etapa, sendo elas, detecção das configurações de mãos, detecção dos movimentos das mãos, detecção das expressões faciais, detecção do ponto de articulação, detecção da orientação da palma da mão, para após detectar todos esses pontos que são essenciais na comunicação em Libras, faz a tradução do sinal para o português, a estratégia usada foi armazenas todos esses pontos em uma estrutura de dados e ao usar o modelo é feito a busca, que é feito a um período de tempo pré-determinado pelo algoritmo, identifica os parâmetros que são feitos no vídeo e salva em uma nova estrutura para após o algoritmo buscar o sinal que foi feito entre os sinais que já foram mapeados e salvos na mesma estrutura.

Após uma revisão detalhada da literatura existente, a pesquisa aborda várias lacunas identificadas nos estudos anteriores e oferece novas contribuições significativas. Utilizando uma rede que combina a extração de características de uma arquitetura CNN e a modelagem da dependência sequencial dos dados com uma arquitetura RNN, é possível alcançar uma contribuição significativa sem a necessidade de um conjunto de dados extenso ou de uma separação específica para rastrear a estrutura linguística. Esta abordagem híbrida supera métodos tradicionais e oferece tempos de treinamento mais rápidos se comparada a arquiteturas mais complexas como os encoder-decoders.

5 Metodologia

O estudo adotou a metodologia proposta por *CRISP-DM*, uma abordagem ágil na qual cada fase é iniciada somente após a validação da anterior. Essa abordagem flexível permite ajustes contínuos ao longo do projeto, com o intuito de aprimorar a precisão da pesquisa no contexto das técnicas, abordagens e estudos fundamentais relacionados à tradução de Libras para texto por meio de visão computacional e aprendizado de máquina (Figura 6).

Para criação do dataset para treinar o modelo, utilizou-se um *dump* da Wikipédia, da data de $20/03/2024$, na língua português brasileiro (PT-BR) para extrair as palavras com maior frequência de utilização. Após exclusão das *stopwords*, uma contagem de frequência foi feita de cada palavra e foram verificadas a existência de vídeos dos sinais que representam aquela palavra no repositório aberto "Dicionário de Libras"do governo brasileiro[1]. Desta forma, foram selecionadas as 500 palavras mais frequentes do *dump* e ao verificar a existência dessas palavras no "Dicionário de Libras", foram encontradas 228 palavras para compor o dataset. Por ser um conjunto relativamente pequeno, foram geradas automaticamente 5 variações de cada vídeo utilizando a biblioteca "Video Augmentation Techniques for Deep Learning" [2] para inclusão nos dataset de treino enquanto as imagens originais foram selecionadas para compor o dataset de validação. Dentre as variações aplicadas nos vídeos, estão inclusas rotação do vídeo, ajustes de resolução, redução de escala, entre outros. Além disso, todas as imagens extraídas dos vídeos tiveram sua resolução reduzida para tamanho 128×128 pixels.

A etapa de modelagem envolveu o uso de técnicas avançadas de *deep learning*. Foram empregadas arquiteturas de redes neurais *CNN* e *RNN* no treinamento do modelo. O objetivo é desenvolver modelos capazes de reconhecer e traduzir os gestos da Libras para texto de maneira eficiente. A implementação da pesquisa foi realizada em linguagem *Python*, utilizando as capacidades da visão computacional. O pacote resultante conterá algoritmos de detecção e rastreamento de gestos, tradução de Libras para texto e recursos adicionais para facilitar estudos futuros.

A fase de avaliação incluiu testes para avaliar a precisão da pesquisa em diferentes cenários. Serão considerados fatores como a velocidade de execução do gesto, a distância em relação ao dispos-

[1] https://www.ines.gov.br/dicionario-de-libras/

[2] https://github.com/okankop/vidaug

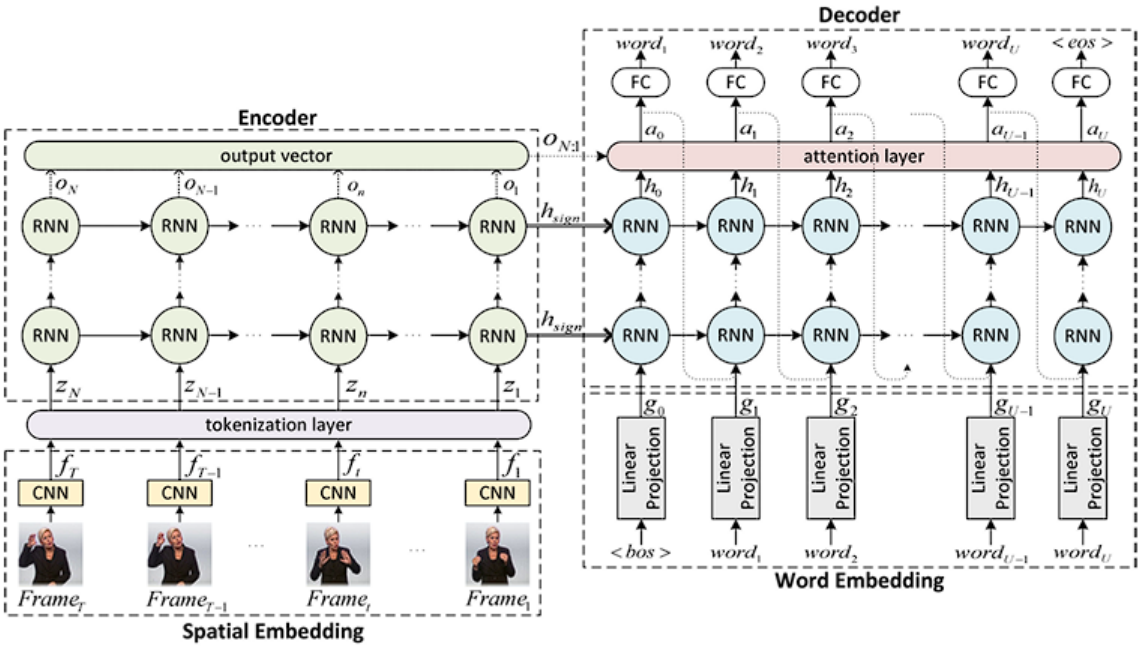


Figure 5: Uma visão geral da abordagem SLT que gera traduções em língua falada de vídeos em língua de sinais.(Camgoz et al., 2018).

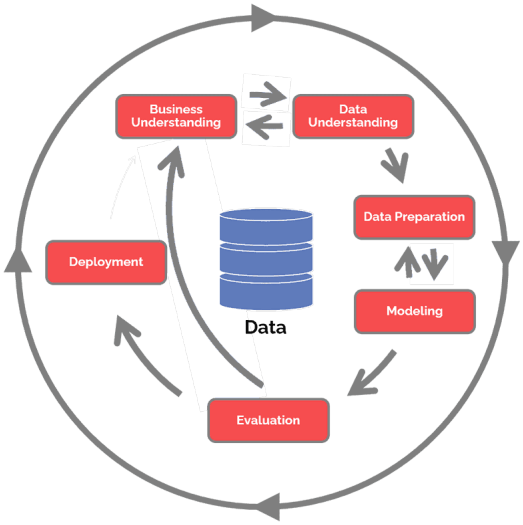


Figure 6: CRISP-DM Diagram.

itivo de captura e a mão utilizada para realizar o sinal. A avaliação incluiu cálculo de métricas, confiabilidade e tempo de resposta do sistema desenvolvido. Análises qualitativas e quantitativas serão conduzidas para avaliar a eficácia do sistema e identificar possíveis melhorias, tanto em desempenho quanto em precisão.

5.1 OpenCV

OpenCV (Open Source Computer Vision Library) é uma biblioteca de software de visão computacional e aprendizado de máquina de código aberto. O OpenCV foi desenvolvido para fornecer uma infraestrutura comum para aplicativos de visão computacional e para acelerar o uso da percepção da máquina nos produtos comerciais. Sendo um produto licenciado Apache 2, o OpenCV torna mais fácil para as empresas utilizar e modificar o código (Bradski, 2000).

A biblioteca possui mais de 2.500 algoritmos otimizados, que inclui um conjunto abrangente de visão computacional clássica e de última geração e algoritmos de aprendizado de máquina. Esses algoritmos podem ser usados para detectar e reconhecer rostos, identificar objetos, classificar ações humanas em vídeos, rastrear movimentos de câmeras, rastrear objetos em movimento, extrair modelos 3D de objetos, produzir nuvens de pontos 3D a partir de câmeras estéreo, unir imagens para produzir alta resolução imagem de uma cena inteira, encontrar imagens semelhantes em um banco de dados de imagens, remover olhos vermelhos de imagens tiradas com flash, acompanhar movimentos oculares, reconhecer cenários e estabelecer marcadores para sobrepô-los com realidade aumentada, etc. OpenCV tem mais de 47 mil usuários comu-

nidade e número estimado de downloads superiores a 18 milhões. A biblioteca é amplamente utilizada em empresas, grupos de pesquisa e por órgãos governamentais.

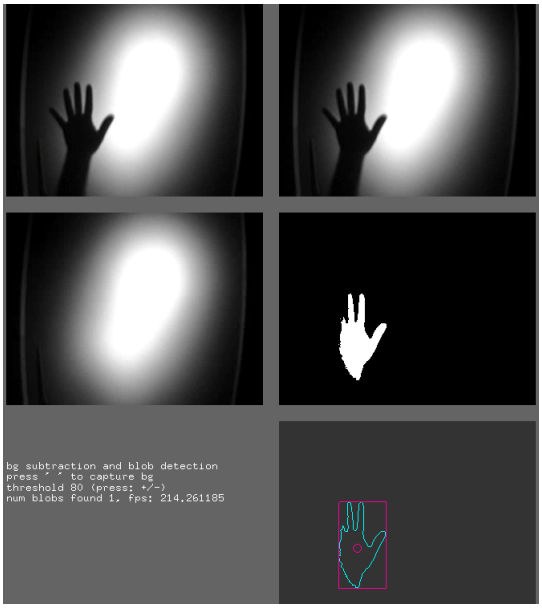


Figure 7: Executando o *OpenCV* em um exemplo.

5.2 Tensorflow

O TensorFlow (Abadi et al., 2015b) é uma biblioteca de código aberto desenvolvida pelo Google para aprendizado de máquina e inteligência artificial. É amplamente utilizada para construir, treinar e implementar modelos de aprendizado profundo. A biblioteca fornece uma ampla gama de ferramentas e recursos que permitem aos desenvolvedores criar modelos complexos de redes neurais, otimizados para diferentes aplicações, incluindo visão computacional, processamento de linguagem natural e reconhecimento de padrões. O TensorFlow facilita a construção de modelos complexos como CNNs e RNNs, utilizadas no projeto. Com o TensorFlow, é possível definir a arquitetura da rede, especificando o número de camadas, tipos de neurônios, funções de ativação, e outras configurações essenciais. Além disso, o TensorFlow oferece otimizadores eficientes, como Adam e SGD que ajustam os pesos do modelo com base no erro de previsão. Além disso, o TensorFlow permite a utilização de técnicas como *backpropagation* (Rumelhart et al., 1986) para atualizar os pesos das redes neurais durante o treinamento.

6 Experimentos e Resultados

Para testar a qualidade dos vídeos e entender a utilização de técnicas de *Deep Learning* para processar e analisar estes vídeos, foram treinados modelos de duas arquiteturas diferentes de redes neurais: uma LSTM que serviu como a primeira *baseline*; e uma CNN-RNN como modelo principal.

Para o modelo baseline, uma RNN com 1 camada contendo 128 células LSTM, ativação $ReLu$ para a entrada da célula e para a célula de memória. O modelo foi treinado utilizando gradiente descendente e *backpropagation through time* (Hochreiter and Schmidhuber, 1997), utilizando batches de 32 exemplos e otimizando a função *categorical cross entropy*:

$$H(X) = -\sum_{x} p(x) \log(p(x)) \tag{1}$$

Para otimização, utilizou-se o otimizador Adam (Kingma and Ba, 2014) com seus hiperparâmetros setados para valores padrão para este algoritmo conforme sugerido por Kingma and Ba (2014). O modelo foi inicialmente configurado para treinar por 20 épocas e um critério de *early stop* com "paciência" de 3 épocas, iniciando-se o contador a partir da época 10 e $\delta = 0.01$.

Para o modelo CNN-RNN (denominado "Ariel"), foram adicionadas 7 camadas convolucionais com filtros de tamanho 3×3 e número de filtros 8, 16, 32, 64, 32, 16 e 8, respectivamente. Após as camadas convolucionais, foi adicionada uma camada LSTM bidirecional com 100 unidades, seguida de uma camada "densa" de 128 neurônios.

Este modelo também foi treinado utilizando gradiente descendente e *backpropagation through time*, com batches de 32 exemplos. Para otimização, utilizou-se o otimizador Adam com uma taxa de aprendizado (*learning rate*) inicial de 0.001. Além disso, aplicou-se uma redução do valor da "norma" $L2$ dos gradientes utilizando técnica de *clipping* para o valor de 1.0. Este modelo também foi treinado utilizando Gradiente descendente e *Backpropagation through time* otimizando a função *categorical cross entropy* (Eq. 1). Este modelo foi inicialmente configurado para treinar por 300 épocas e um critério de *early stop* com "paciência" de 5 épocas, iniciando-se o contador a partir da época 1 e $\delta = 0.0001$.

Após os treinamentos dos dois modelos utilizando o dataset com as variações geradas automaticamente (ver Seção 5), utilizamos os vídeos

originais para calcular o percentual de acertos de cada um dos modelos. Os resultados estão delineados na Tabela 1[3]:

Modelo	Acurácia
LSTM	0.0438
Ariel (CNN-RNN)	0.9641

Table 1: Resultados obtidos pelos modelos LSTM e Ariel (CNN-RNN) no dataset criado.

7 Discussão

O modelo LSTM, utilizado como baseline, apresentou uma acurácia que não foi detalhadamente especificada no artigo. No entanto, é possível discutir de forma geral as características e o desempenho esperado de um modelo LSTM nesta tarefa. A acurácia de um modelo de aprendizado profundo, como a LSTM, é uma métrica fundamental que indica a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões. No contexto do reconhecimento de sinais em Libras, a acurácia reflete a capacidade do modelo de identificar corretamente os gestos realizados e traduzi-los adequadamente para texto em português. Em tarefas de tradução de língua de sinais, as LSTMs são tradicionalmente eficazes para lidar com sequências temporais devido à sua arquitetura, que permite o armazenamento e a atualização de informações ao longo do tempo. No entanto, as LSTMs também enfrentam desafios, especialmente quando se trata de capturar dependências de longo prazo em sequências complexas, como os gestos contínuos de Libras.

Os resultados obtidos modelo CNN-RNN ("Ariel") demonstram um potencial dessas arquiteturas para promover um avanço na pesquisas que possa ajudar inclusão da comunidade surda. A combinação CNN-RNN mostrou-se eficaz na tarefa de reconhecimento de sinais, alcançando uma acurácia alta, acima de 96%. O desempenho deste modelo com uma acurácia de 96.41%, é um indicativo da capacidade dessas redes em capturar e interpretar as nuances dos gestos em Libras. Esse desempenho superior em relação ao modelo baseline (LSTM) reflete a vantagem de utilizar uma abordagem que combina a extração de características espaciais (através das CNNs) com a análise de sequências temporais (através das RNNs). Essa arquitetura permite uma compreensão mais profunda e contextual dos gestos, essencial para a tradução precisa de uma língua gestual-visual.

Apesar dos resultados positivos, alguns desafios e limitações foram identificados durante a pesquisa. Primeiramente, a criação do dataset revelou-se uma tarefa complexa. A quantidade limitada de vídeos disponíveis no Dicionário de Libra restringiu o tamanho do conjunto de dados, exigindo a aplicação de técnicas de data augmentation para aumentar a diversidade dos exemplos de treino. Mesmo com essas técnicas, a variedade e a representatividade dos sinais podem ainda não ser suficientes para capturar todas as variações possíveis dos gestos em Libras. Além disso, a qualidade dos vídeos utilizados para treinamento pode influenciar significativamente o desempenho do modelo. Vídeos com baixa resolução, iluminação inadequada ou ângulos desfavoráveis podem dificultar a detecção precisa dos pontos de referência necessários para a interpretação dos gestos.

A implementação de sistema como este, em tempo real, possui aplicações práticas que podem beneficiar tanto a comunidade surda quanto a ouvinte. Em ambientes educacionais, a ferramenta pode auxiliar na inclusão de alunos surdos, proporcionando-lhes maior acesso ao conteúdo acadêmico. Em serviços de saúde, o sistema pode melhorar a comunicação entre pacientes surdos e profissionais de saúde, garantindo um atendimento mais eficaz e humanizado. No entanto, a usabilidade do sistema em situações do cotidiano ainda necessita de avaliação prática com usuários reais. É essencial realizar testes com diferentes perfis de usuários, em variados contextos, para ajustar e aprimorar a interface e a funcionalidade do sistema. A participação ativa da comunidade surda nesse processo é fundamental para garantir que as necessidades e expectativas dos usuários finais sejam plenamente atendidas.

8 Conclusões

O presente estudo apresentou o desenvolvimento de um sistema para reconhecimento e tradução da Língua Brasileira de Sinais (Libras) para texto, utilizando técnicas de visão computacional e aprendizado profundo. A pesquisa demonstrou que a combinação de Redes Neurais Convolucionais (CNNs) e Redes Neurais Recorrentes (RNNs) é

[3]Outros modelos também foram treinados mas com configurações de hiperparâmetros diferentes. No entanto, nenhum outro modelo performou dentro do esperado e, portanto, não relatamos seus resultados aqui.

eficaz para a interpretação de gestos em Libras, alcançando uma acurácia notável com o modelo CNN-RNN ("Ariel"). A pesquisa apresentada oferece uma contribuição significativa para a inclusão da comunidade surda através da aplicação de tecnologias de visão computacional e aprendizado profundo. O desenvolvimento de um sistema de tradução de Libras para texto mostra-se uma solução viável e eficaz para romper barreiras comunicativas, promovendo maior acessibilidade e interação entre surdos e ouvintes. No entanto, a continuidade do desenvolvimento e a realização de testes práticos com usuários reais são essenciais para validar e aperfeiçoar a aplicabilidade do sistema em situações cotidianas. Os resultados obtidos são promissores e indicam o potencial das tecnologias de aprendizado profundo para promover a inclusão social da comunidade surda, facilitando a comunicação entre surdos e ouvintes. No entanto, foram identificados desafios relacionados à criação e à diversidade do dataset, bem como à qualidade dos vídeos utilizados para treinamento.

A implementação do sistema em diferentes contextos práticos, como educação e saúde, pode proporcionar benefícios significativos, melhorando a acessibilidade e a interação social para pessoas surdas. A avaliação do sistema com usuários reais será essencial para refinar e aperfeiçoar a usabilidade e a funcionalidade da ferramenta. Para dar continuidade a esta pesquisa e melhorar os resultados obtidos, diversas direções podem ser exploradas:

1. Expansão do Dataset: Aumentar o conjunto de dados com mais sinais e variações, possivelmente através de colaborações com instituições de ensino de Libras, pode melhorar a representatividade e a precisão do modelo. A coleta de vídeos em diferentes condições de iluminação, ângulos e resoluções também pode contribuir para a robustez do sistema.

2. Incorporação de Componentes Multimodais: Integrar a detecção de expressões faciais e posturas corporais, que são partes significativas da comunicação em Libras, pode melhorar a precisão da tradução. Isso exigirá o desenvolvimento de modelos mais complexos e o uso de técnicas avançadas de fusão de dados multimodais.

3. Aprendizado por Reforço: Implementar técnicas de aprendizado por reforço para permitir que o modelo aprenda com interações contínuas e feedback dos usuários pode aprimorar a capacidade do sistema de se adaptar a novas variações de sinais e contextos.

4. Testes com Usuários Reais: Realizar testes práticos com diferentes perfis de usuários, em variados contextos, é essencial para validar a usabilidade e a eficácia do sistema. A participação ativa da comunidade surda no processo de avaliação e feedback será crucial para garantir que o sistema atenda às necessidades e expectativas dos usuários finais.

5. Desenvolvimento de Aplicativos e Interfaces: Criar interfaces e aplicativos amigáveis que possam ser utilizados em dispositivos móveis e plataformas web facilitará a adoção do sistema em situações cotidianas. A integração com ferramentas existentes, como o VLibras, pode aumentar a acessibilidade e a utilidade do sistema.

6. Otimização e Escalabilidade: Trabalhar na otimização do modelo para reduzir o tempo de processamento e melhorar a eficiência do sistema permitirá a sua utilização em tempo real, ampliando as possibilidades de aplicação prática.

Essas direções de pesquisa e desenvolvimento podem contribuir para a evolução do sistema de tradução de Libras para texto, promovendo maior inclusão e acessibilidade para a comunidade surda e fortalecendo a interação social e a igualdade de oportunidades para todos.

References

Dificuldades e desafios dos surdos na sociedade (brasil).

Vlibras.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015a. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015b. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

Cláudio Gustavo Santos Capanema, Fabrício Aguiar Silva, and Thais Regina de Moura Braga Silva. 2020. Mfa-rnn: Uma rede neural recorrente para predição de próximo local de visita com base em dados esparsos. In *Anais do XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuidos*, pages 127–140. SBC.

Álvaro Gonçalves de Barros, Carlos Henrique Medeiros de Souza, and Risiberg Teixeira. 2021. Evolução das comunicações até a internet das coisas: a passagem para uma nova era da comunicação humana. *Cadernos de Educação Básica*, 5(3):260–280.

Maria Inês dos Santos, Águida Layse Oliveira Cavalcanti, Valquíria Farias Bezerra Barbosa, Ronny Diógenes de Menezes, Claudia Daniele Barros Leite Salgueiro, and Saulo Santos da Silva. 2021. Dificuldades no acesso da comunidade surda à rede básica de saúde: revisão integrativa. *Enfermagem Brasil*, 20(2):206–221.

Tanya Amara Felipe. 2006. Os processos de formação de palavra na libras. *ETD Educação Temática Digital*, 7(02):200–212.

Yanming Guo, Yu Liu, Erwin M Bakker, Yuanhao Guo, and Michael S Lew. 2018. Cnn-rnn: a large-scale hierarchical image classification framework. *Multimedia tools and applications*, 77(8):10251–10271.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. volume 9, pages 1735–1780, Cambridge, MA, USA. MIT Press.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv*, abs/1412.6980.

Emanoel da Cruz LORETO. 2023. Plataforma de aprendizagem de libras utilizando machine learning e visão computacional para reconhecimento de sinais.

Rodrigo Magno. As dificuldades da pessoa surda na sociedade brasileira.

Antonio Luiz Alencar Miranda, Ana Rosária Soares da Silva, and Shirlane Maria Batista da Silva Miranda. 2020. Educação especial e inclusiva na perspectiva do ensino bilíngue. *The ESPecialist*, 41(1).

Olof Mogren. 2016. C-rnn-gan: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904*.

Andres Jessé Porfirio, Kelly Laís Wiggers, Luiz ES Oliveira, and Daniel Weingaertner. 2013. Libras sign language hand configuration recognition based on 3d meshes. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1588–1593. IEEE.

Clélia Regina Ramos. 2004. Libras: A língua de sinais dos surdos brasileiros. *Disponível para download na página da Ediotra Arara Azul: http://www. editora-arara-azul. com. br/pdf/artigo2. pdf*.

Thomas Walter Rauber. 2005. Redes neurais artificiais. *Universidade Federal do Espírito Santo*, 29.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323:533–536.

Romário Pereira da Silva et al. 2022. Visão computacional: um estudo de caso aplicado à língua brasileira de sinais (libras).