Reconhecimento e Tradução de Sinais de Libras para Língua Portuguesa Escrita usando Redes Neurais Profundas

Jhon Lucas S. Silva * **Gabriel S. Vieira** * **Afonso U. Fonseca** *
Fabrizzio Soares *

* *Instituto de Informática, Universidade Federal de Goiás, GO (e-mail: jhonlucas@discente.ufg.br).*
** *Instituto Federal Goiano, Campus Urutaí, Urutaí, GO (e-mail: gabriel.vieira@ifgoiano.edu.br).*
*** *Instituto de Informática, Universidade Federal de Goiás, GO (e-mail: afonsoueslei@ufg.br).*
**** *Instituto de Informática, Universidade Federal de Goiás, GO (e-mail: fabrizzio@ufg.br).*

Abstract: The Brazilian Sign Language (Libras) enables deaf people to understand and interact with others in order to facilitate their access to culture, knowledge and social integration. However, there are only few solutions to reduce the communication barrier between deaf and hearing people. In this work, we propose a solution based on deep neural networks for sign language recognition. This is an exploratory study in which signs in Libras ("Hello", "Good morning", and "Thank you") are used in training, recognition and classification in continuous and real-time mode. We compared two machine learning models that were trained on the LSTM and BiLSTM neural network architectures. The results point to superior assertiveness of the LSTM model, with an accuracy of 84.71% compared to the 77.07% achieved by the BiLSTM model. Therefore, the LSTM architecture is more suitable for classifying the signals investigated in this study. Besides, its use in sign image recognition systems in Libras proves to be viable.

Resumo: A Língua Brasileira de Sinais, ou Libras, possibilita a pessoa surda compreender e interagir com o mundo de modo a viabilizar seu acesso à cultura, ao conhecimento e à integração social. Contudo, há poucas soluções disponíveis para reduzir a barreira de comunicação entre pessoas surdas e ouvintes. Neste trabalho, propomos uma solução baseada em redes neurais profundas para o reconhecimento de língua de sinais. Trata-se de um estudo exploratório em que sinais em Libras ("Oi", "Bom dia", e "Obrigado") são submetidos ao treinamento, reconhecimento e classificação em modo contínuo e em tempo real. Comparamos dois modelos de aprendizado de máquina treinados nas arquiteturas de redes neurais LSTM e BiLSTM. Os resultados apontam para assertividade superior do modelo treinado em LSTM, com acurácia de 84,71% em detrimento de 77,07% alcançado pelo modelo em BiLSTM. Logo, a arquitetura LSTM se mostra mais adequada à classificação dos sinais investigados neste estudo, sendo viável seu uso em sistemas de reconhecimento por imagem de sinais em Libras.

Keywords: Sign language recognition, sign language processing, continuous sign language recognition, machine learning, deep learning, computer vision, sign language.

Palavras-chaves: Reconhecimento de língua de sinais, processamento de língua de sinais, reconhecimento contínuo de língua de sinais, aprendizado de máquina, aprendizado profundo, visão computacional, língua de sinais.

1. INTRODUÇÃO

Segundo o último censo brasileiro realizado pelo IBGE em 2010, estima-se que $5,1\%$ da população, ou cerca de 10 milhões de brasileiros, possuem alguma deficiência auditiva. Além disso, enquanto $89,5\%$ dos ouvintes são alfabetizados, os processos de alfabetização abrangem apenas $75,5\%$ da população surda (IBGE, 2010).

A Língua Brasileira de Sinais, ou Libras, principal instrumento para a educação dos surdos, tem respaldo legal por meio da Lei n.º 10.436/2002 e pelo Decreto n.º 5.626/2005 que a reconhecem como forma de comunicação e expressão. Essa regulamentação garante o direito da pessoa surda em compreender e interagir com o mundo de modo a viabilizar seu acesso à cultura, ao conhecimento e à integração social (BRASIL, 2002).

Em se tratando de tecnologias de comunicação para a língua de sinais, há poucas soluções disponíveis resultando no crescimento da barreira de comunicação entre pessoas surdas e ouvintes (Bragg et al., 2019). Contudo, as pro-

postas existentes abrem caminhos para novos estudos. Por exemplo, as plataformas *Hand Talk* e *VLibras* que realizam a tradução simultânea do português para Libras (Caetano and Passos, 2017) contribuem para reduzir desigualdades ao passo que fomentam a investigação e lançamento de outras propostas que deem conta das limitações presentes em soluções dessa natureza.

Dentre essas limitações, registra-se a ausência de processos consistentes para o reconhecimento e tradução de línguas de sinais de forma contínua, isto é, que associe sinais previamente executados para formar frases gramaticalmente e semanticamente corretas (Bragg et al., 2019). Outro ponto é que parte das soluções disponíveis tratam apenas da tradução do português (escrito ou falado) para a Libras, e não o inverso.

Neste artigo, investigamos essas limitações para propor uma solução baseada em redes neurais profundas para o reconhecimento de língua de sinais. Trata-se de um estudo exploratório em que três sinais em Libras são submetidos ao treinamento, reconhecimento e classificação em duas arquiteturas de redes neurais (LSTM e BiLSTM). Nesse sentido, construímos modelos de aprendizagem de máquina para categorização de imagens em modo contínuo e em tempo real como primeiro passo para a construção de um sistema de tradução de Libras para a língua portuguesa escrita, algo ainda pouco explorado em trabalhos relacionados.

Este artigo está organizado da seguinte forma: na Seção 2, trazemos uma revisão acerca dos trabalhos relacionados à problemática deste estudo. Na Seção 3, apresentamos os materiais e métodos utilizados no planejamento e desenvolvimento da pesquisa. Na Seção 4, lançamos os resultados obtidos com a experimentação e discutimos os resultados. Por fim, na Seção 5, apresentamos as principais conclusões e trabalhos futuros.

2. TRABALHOS RELACIONADOS

Em Bull et al. (2020), investigou-se a segmentação automática da língua francesa de sinais através do uso de unidades de legendas para o reconhecimento de frases completas. Os autores utilizaram uma base de dados chamada *MEDIAPI-SKEL* a qual possui dados rotulados na forma de legendas de vídeo e em que o interprete é representado por um esqueleto 2D, permitindo a rastreabilidade de vários pontos de um sinal de linguagem que está sendo sinalizado. Utilizaram redes neurais recorrentes do tipo *BiLSTM* (bidirectional LSTM) e obtiveram 92% de acurácia nos experimentos realizados.

Raval and Gajjar (2021) realizaram o reconhecimento de língua inglesa de sinais em tempo real utilizando modelos de redes neurais convolucionais. Os autores fizeram uso de uma base de dados privada com 10 imagens para cada sinal do alfabeto e nos testes de validação, o modelo proposto chegou a uma acurácia de 83%.

No trabalho de Park et al. (2021), foi apresentado um sistema para a tradução de língua coreana de sinais utilizando câmeras frontais de *smartphones*. Denominado *SUGO*, esse sistema foi projetado a partir da rede neural convolucional *3DCNN* que classifica sequências de *frames* recebidas como entrada. Foi construída uma base de imagens própria com cerca de 5 mil amostras de vídeo em diferentes cenários e nos testes realizados, obteve-se 91% de acurácia com o modelo proposto.

Yin et al. (2021) prepararam uma revisão sistemática sobre publicações que envolvem a temática de Processamento de Língua de Sinais. Discutiram a aplicação de técnicas de processamento de linguagem natural em língua de sinais como forma de representar e identificar estruturas morfológicas, léxicas, sintaxes, e outras estruturas linguísticas, assim como se faz na comunicação convencional, escrita ou falada. Concluíram que há um crescimento no número de pesquisas que buscam resolver este tipo de problema e que trabalhos dessa natureza deveriam considerar a linguística presente na língua de sinais.

Passos et al. (2021) trabalharam com a identificação de gestos por meio de uma técnica chamada Gait Energy Image (GEI), usada no trabalho para codificar informações de movimento de mãos, braços e cabeça. Lidaram com o reconhecimento de sinais isolados utilizando algoritmos clássicos de aprendizagem de máquina, com resultados similares as arquiteturas de redes neurais mais sofisticadas.

Mittal et al. (2019) apresentam um modelo de *Long Short Term Memory (LSTM)* modificado que tem como objetivo reconhecer sequência continua de sinais sem conexão, ou uma sequência de sinais contínuos que possuem conexão. Os sinais capturados foram divididos em sub-unidades para serem usados com modelos com redes neurais, ou seja para cada sequência de sinais, há uma quebra em diversas unidades para que cada sinais seja passado ao modelo como um sinal isolado.

Os trabalhos relacionados apresentados nesta seção possuem objetivos semelhantes, mas com a aplicação de diferentes técnicas de aprendizagem de máquina e visão computacional para os problemas propostos. Alguns lidaram com o reconhecimento de língua de sinais em situações dinâmicas que envolvem movimentações contínuas de gestos. Outros lidaram com sinais estáticos, com limite de escopo de sinais em cenários individualizados de comunicação não-verbal. Contudo, o reconhecimento de sinais para a reconstrução de frases em processos de tradução ainda é limitado devido a fatores que envolvem a preparação de bases de dados mais representativas, a elaboração de modelos que contemplem a formação de palavras e frases, e etapas de correção gramatical dos sinais observados em processos de tradução da língua de sinais para a língua escrita. Além disso, como cada língua possui o seu próprio sistema de sinais, faz-se necessária a preparação particularizada de modelos de referências com bases rotuladas para cada língua, como é o caso da Libras.

Neste trabalho, construímos uma base de dados própria para o reconhecimento de três sinais da Língua Brasileira de Sinais ("Oi", "Bom dia", e "Obrigado"). Usamos uma arquitetura de rede neural para o treinamento de um modelo de aprendizagem cuja tarefa é a detecção e classificação de sinais de maneira contínua por meio de *frames* de vídeo. Ademais, a execução é realizada em tempo real com o processamento de pelo menos 30 *frames* por segundo.

3. MATERIAIS E MÉTODOS

Nesta seção serão apresentados os materiais e métodos utilizados no desenvolvimento deste trabalho. Serão apresentadas subseções sobre aquisição dos dados, rastreabilidade de pontos em imagens extraídas a partir de vídeos capturados, arquitetura dos modelos utilizados, trazendo detalhes sobre as etapas que permitiram chegar aos resultados que serão apresentados na próxima seção. Em suma, o método utilizado consiste na aquisição de dados e posteriormente a rastreabilidade de pontos para extrair variáveis que são utilizadas para serem submetidas aos modelos que farão classificação dos sinais, sendo estes modelos treinados e validados em uma base de dados separada para esta finalidade, o processo de separação é descrito com mais detalhes nas subseções.

3.1 Rastreabilidade de pontos

Para a detecção e rastreamento de pontos em imagens, usamos a solução *open source* do Google, chamada Media Pipe. Usamos essa solução para fazer a rastreabilidade de pontos das mãos, face e corpo. A Tabela 1 mostra a quantidade de pontos rastreáveis para cada objeto de interesse através do Media Pipe.

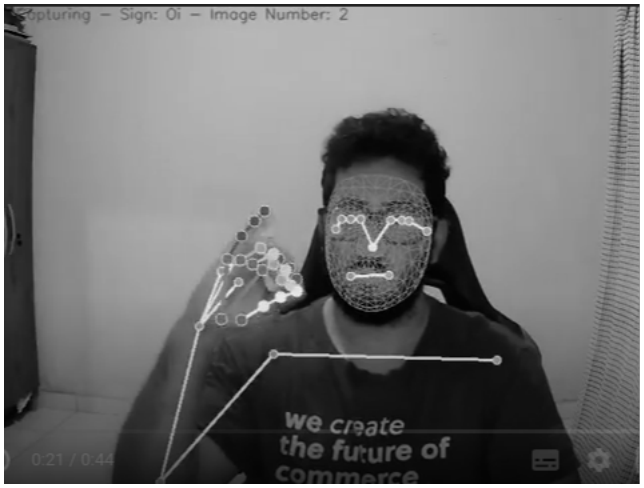
Tabela 1. Pontos rastreáveis por cada objeto no Media Pipe.

Objeto de interesse	**Quantidade de Pontos**
Face	468
Corpo	33
Mãos	21

3.2 Aquisição de dados

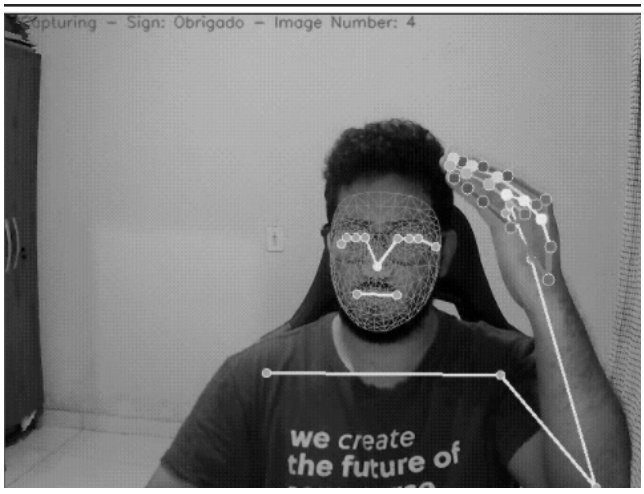
Neste estudo, consideramos três sinais de Libras: (1) "Oi", (2) "Bom dia", e (3) "Obrigado". Um sinal em Libras é composto por gestos e expressões, realizados em sequência e geram uma série temporal contínua de poses, por este motivo a aquisição de dados se dá por meio de vídeos. Para cada um desses sinais foram capturados 350 vídeos com taxa de quadros por segundo igual a 30, sendo 30 um número de quadros necessário para que seja possível capturar todos os movimentos e expressões que são realizados durante a interpretação de um sinal de Libras. A Figura 1 exemplifica a captura dos dados na qual pontos de interesse são detectados e rastreados quadro a quadro com uso de *webcam*. Para cada sinal temos uma sequência de 350 vídeos, sendo cada vídeo com duração aproximada de 3 segundos, totalizando mais de 94 mil *frames*. É importante ressaltar que cada sinal tem uma sequência de *frames* que permite classificar o seu significado. Além disso, um sinal é composto não apenas por gestos, mas também por expressões faciais e pequenos movimentos que podem, em alguns casos, alterar o significado do sinal.

O processo de aquisição de dados foi implementado utilizando a linguagem Python com o apoio do *framework* OpenCV e Media Pipe. Cada frame passa por um processamento de transformação, extraindo do frame vetores que possuem as coordenadas de cada movimento realizado durante a sinalização de um sinal, o resultado desta transformação, no final é uma matriz com tamanho igual a



(a) Sinal "Oi"

(b) Sinal "Bom dia"



(c) Sinal "Obrigado"

Figura 1. Exemplo de aquisição de dados em imagem.

1050x30x1662, sendo 1050 a quantidade de vídeos capturados (350 vídeos multiplicados por 3 sinais), 30 o número de *frames*, e 1662 é a quantidade de pontos capturados considerando a face, mãos e corpo em um total 522 pontos para cada sinal. Esta matriz é a base com todos os sinais que foram sinalizados e que possui valores que se referem as

Tabela 2. Camadas utilizadas no modelo LSTM.

Nº	Camada	Neurônios	Saída	Função de Ativação
1	*LSTM*	64	64	RELU
2	*LSTM*	128	128	RELU
3	*LSTM*	64	64	RELU
4	*Dense*	64	64	RELU
5	*Dense*	32	32	RELU
6	*Dense*	3	3	SOFTMAX

Tabela 3. Camadas utilizadas no modelo BiLSTM.

Nº	Camada	Neurônios	Saída	Função de Ativação
1	*Bidirectional*	64	128	RELU
2	*Bidirectional*	128	256	RELU
3	*Bidirectional*	64	128	RELU
4	*Dense*	64	64	RELU
5	*Dense*	32	32	RELU
6	*Dense*	3	3	SOFTMAX

coordenadas x, y, e z de cada movimento realizado durante a captura do frame. Os vídeos que foram capturados foram de apenas uma pessoa sinalizando, porém considerando movimentos e posicionamento diferentes durante a captura de cada frame para termos uma base generalizada em relação as coordenadas capturadas.

3.3 Modelos de aprendizagem

Neste trabalho, dois modelos de Redes Neurais Recorrentes (RNN) foram considerados, LSTM (*Long-Short Term Memory*) e BiLSTM (*bidirectional LSTM*). Ambos os modelos classificam os dados de entrada conforme a sequência de *frames* é inserida no fluxo de execução (*pipeline*) e destinam-se ao tratamento de dados em séries temporais (Bull et al., 2020; Mittal et al., 2019). As redes LSTM possuem a capacidade de guardar informações e, por esse motivo, também são denominadas memórias de longo e curto prazo (Misra and Li, 2020). Redes BiLSTM são extensões de redes LSTM nas quais os dados de entrada são submetidos considerando a série temporal mais antiga para a mais nova, e depois submetidos novamente em ordem inversa, por isso, são ditas bidirecionais (Siami-Namini et al., 2019).

Os modelos foram implementados utilizando a biblioteca Keras com *interface* à biblioteca de código aberto para aprendizado de máquina, Tensorflow. As entradas são submetidas aos modelos com um tamanho fixo de 30x1662, sendo 30 o número de *frames* capturados e 1662 o número de pontos rastreados pelo Media Pipe.

Os modelos foram treinados em um computador com 64 GB de memória RAM, placa gráfica NVIDIA RTX 2060 e CPU Ryzen 5 5600X. As Tabelas 2 e 3 mostram as camadas utilizadas nos modelos LSTM e BiLSTM para a classificação dos sinais.

Para realizar o treinamento dos modelos a base foi dividida em três partes distintas, sendo 70% da base para treinamento, 15% para validação e 15% para testes. E em todas as bases os dados foram escolhidos de forma aleatória para não gerar nenhum viés durante do treinamento, validação e teste do modelo.

3.4 Arquitetura dos modelos

Nas Tabelas 2 e 3 podemos observar quais são as camadas que são usadas na arquitetura de cada modelo, LSTM e BiLSTM, respectivamente. As camadas foram inspiradas no trabalho de (Liu et al., 2016), que faz o uso de LSTM para reconhecimento de sinais da língua Chinesa. O trabalho publicado por (Liu et al., 2016) tem uma abordagem diferente no que se trata de aquisição do dado, porém o método de reconhecimento de sinais é semelhante ao que foi proposto neste trabalho. A arquitetura da rede LSTM possui um total de 6 camadas, sendo as 3 primeiras camadas do tipo LSTM, sendo a primeira camada com um total de 64 neurônios, a segunda camada com 128 neurônios e a terceira camada com 64 neurônios. No trabalho de (Liu et al., 2016) foram usados 512 neurônios, porém para este trabalho reduzir o número de neurônios não foi um problema, tendo em vista a quantidade de dados utilizados para treinamento do modelo. Para o modelo BiLSTM a abordagem é a mesma, o que muda neste caso é que para o modelo LSTM a rede é totalmente conectada, o que significa que a quantidade de neurônios usados para entrada, é a mesma quantidade para saída, já o modelo BiLSTM tem a quantidade de neurônios de saída duplicada, exatamente porque é um modelo que possui camadas LSTM bidirecionais.

3.5 Fluxo de processos

O processo de reconhecimento e classificação de sinais em Libras inicia-se com a captura de *frames* por *webcam* os quais são submetidos aos modelos de rede neural recorrente investigados nesse estudo, LSTM e BiLSTM. Os processos são executados em tempo real e consideram os movimentos corporais que formam os sinais de interesse.

Fluxo de treinamento do modelo Na etapa de treinamento, as imagens em vídeos são submetidas aos modelos de aprendizagem de máquina do Média Pipe para rastreamento das mãos, face e corpo. Em seguida, os pontos rastreados são transformados em uma matriz tridimensional e são usados pelos modelos de classificação para treinamento. A Figura 2 mostra o fluxo de processo de treinamento do modelo para realizar o reconhecimento de sinais.

Fluxo de reconhecimento de novas entradas O fluxo é semelhante ao de treinamento, recebendo as imagens em vídeo como entrada, e que também passam por uma etapa de processamento que transforma os pontos rastreados em uma matriz tridimensional, que é submetida ao modelo treinado que faz a inferência sobre os dados e retorna a classificação dos sinais. A Figura 3 mostra o fluxo de classificação de novas entradas.

3.6 Métricas de avaliação

As seguintes métricas são consideradas nesse trabalho:

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

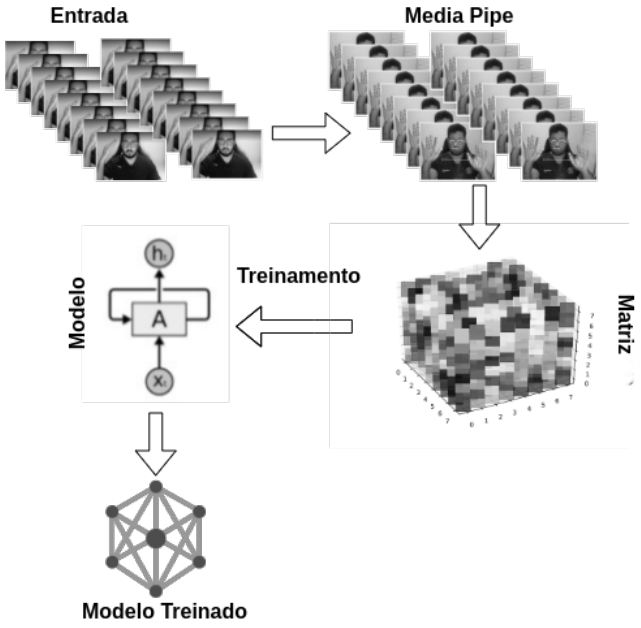


Figura 2. Pipeline de treinamento do modelo.

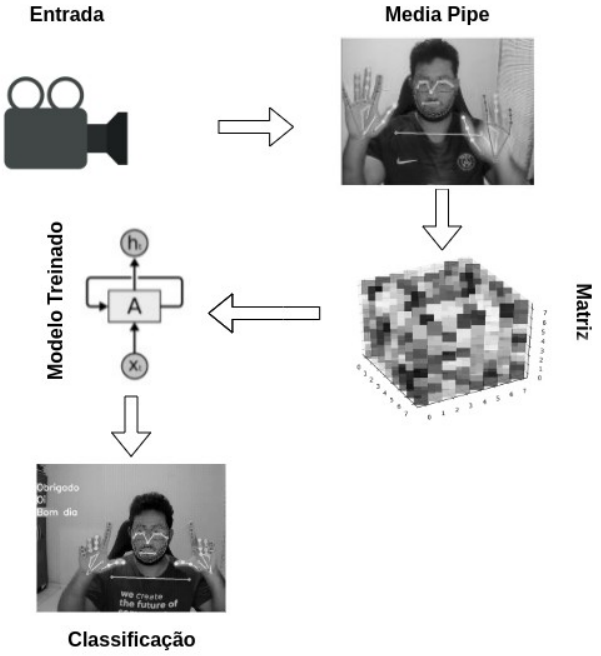


Figura 3. Pipeline de classificação de sinais em Libras.

$$\text{F1-Score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (4)$$

onde TP (*True-Positive*) indica o número de sinais corretamente classificados. TN (*True-Negativo*) indica o número de sinais que não foram relacionados a nenhuma das classes, ou seja, ocorre quando o sinal não pertence nem ao rótulo esperado, nem ao rótulo que foi previsto. FP (*False-Positive*) indica o número de sinais classificados como pertencentes a uma classe, mas que na verdade não pertence, uma classificiação incorreta. FN (*False-Negative*) indica a quantidade de sinais que não foram classificados em acordo à classe a que pertencem.

4. RESULTADOS E DISCUSSÃO

Para selecionar os hiper-parâmetros dos modelos LSTM e BiLSTM, foram realizados um total de 26 experimentos, sendo 13 experimentos para cada um dos modelos em estudo. Ao final dessa etapa de experimentação, identificou-se os parâmetros mais adequados para cada um dos modelos, conforme apresentado na Tabela 4. *Batch_size* e *Épocas* referem-se a hiper-parâmetros de modelos de aprendizagem em redes neurais. O primeiro refere-se a quantidade de vezes que os dados serão repassados ao modelo para treinamento. O segundo refere-se a quantidade de dados submetidos ao modelo em cada época de treinamento. Os resultados são medidos pela assertividade das respostas dos modelos nas etapas de treinamento e teste, medido pela *Accuracy* (Eq. 4).

Tabela 4. Parametrização dos modelos LSTM e BiLSTM.

Nº	Modelo	Épocas	batch_size	Acurácia - treinamento	Acurácia - teste
1	*LSTM*	100	64	79,59%	79,11%
2	*LSTM*	100	128	**87,62%**	**84,71%**
3	*LSTM*	10	32	57.07%	57,59%
4	*LSTM*	20	32	40.13%	37,34%
5	*LSTM*	20	64	62,04%	57,59%
6	*LSTM*	30	64	70,34%	68,35%
7	*LSTM*	30	32	35,37%	28,48%
8	*LSTM*	30	128	60,82%	55,06%
9	*LSTM*	40	64	82,72%	79,11%
10	*LSTM*	40	128	76,60%	79,11%
11	*LSTM*	50	32	33,88%	36,08%
12	*LSTM*	50	64	70,88%	72,15%
13	*LSTM*	50	128	60,00%	58,86%
14	*BiLSTM*	200	128	35,51%	37,34%
15	*BiLSTM*	200	256	33,88%	27,22%
16	*BiLSTM*	1000	256	31,84%	35,44%
17	*BiLSTM*	100	256	33,47%	29,75%
18	*BiLSTM*	10	256	34,42%	41,14%
19	*BiLSTM*	10	32	32,38%	36,08%
20	*BiLSTM*	10	128	62,86%	61,39%
21	*BiLSTM*	20	256	54,97%	52,53%
22	*BiLSTM*	20	128	37,69%	36,08%
23	*BiLSTM*	40	128	**75,24%**	**77,07%**
24	*BiLSTM*	50	256	31,57%	29,75%
25	*BiLSTM*	50	64	35,51%	30,38%
26	*BiLSTM*	50	128	33,88%	36,08%

Com base na Tabela 4, os experimentos 2 e 20 atingiram os melhores resultados, portanto, sendo escolhidos para a realização de comparativos entre os modelos LSTM e BiLSTM.

Na etapas de treinamento e teste, comparamos os modelos de rede neural LSTM e BiLSTM utilizando as métricas *precision, recall, f1-score, accuracy e confusion matrix* (Eqs.: 1 – 4). Os resultados podem ser acompanhados na Tabela 5 que mostra acurácia superior de 84,71% para o modelo LSTM em detrimento do valor 77,07% alcançado pelo modelo BiLSTM, ambos resultados na etapa de teste.

A Figura 4 mostra a *função de perda* e *acurácia* do modelo LSTM durante o treinamento. Da mesma forma, a Figura 5 mostra a *função de perda* e *acurácia* do modelo BiLSTM.

Para ambos os modelos, os treinamentos foram realizados com base nos hiper-parâmetros mais adequados para cada um deles, conforme apresentado na Tabela 4.

Ao analisar as Figuras 4 e 5 nota-se que o modelo LSTM tem uma performance superior ao modelo BiLSTM. Analisando a acurácia do modelo LSTM, Figura 4, nota-se que não há um *overfitting* no modelo, isto porque a acurácia no treinamento não foi maior que a acurácia na validação, isso nos faz chegar ao resultado de um modelo sem *overfitting*. Analisando a função de perda, nota-se que no treinamento e validação também não tivemos *overfitting*, neste caso, como conclusão, temos que o modelo LSTM não está se ajustando totalmente aos dados.

Além disso, nota-se que o modelo BiLSTM tem performance inferior quando comparado ao LSTM. Contudo, percebe-se que o modelo BiLSTM também não está com *overfitting*, como mostrado pela acurácia e função de perda com as previsões feitas pelo modelo usando os dados de treinamento e validação (Figura 5).

A Figura 6 mostra os resultados da etapa de classificação com o modelo LSTM para cada um dos três sinais considerados como objetos de estudo. Da mesma forma, a Figura 7 apresenta os resultados de classificação para o modelo BiLSTM. Ambas as Figuras se referem a matriz de confusão para cada modelo, matriz esta que tem como objetivo mostrar o resultado dos modelos quando aplicados em dados de teste, ou seja, dados que antes não foram vistos pelos modelos, assim é possível avaliar como o modelo irá se comportar ao realizar a inferência em novas entradas.

Nota-se que o modelo LSTM apresentou maior assertividade na classificação de novos sinais recebidos como entrada, analisando as Figuras 6 e 7, percebe-se que para todos os sinais o modelo BiLSTM tem um percentual maior de FP e FN, resultado que é negativo para o modelo. Em contrapartida, o modelo LSTM tem melhores resultados para FP e FN, ou seja, os percentuais são menores quando comparados aos percentuais da matirz de confusão do modelo BiLSTM.

5. CONCLUSÃO

Nesse estudo, aplicamos técnicas de visão computacional e aprendizado profundo com redes neurais artificiais visando realizar o reconhecimento de um conjunto de sinais básicos em Libras, sendo os sinais para "Oi", "Bom dia", e "Obrigado". Para isto, comparamos dois modelos de arquitetura de aprendizado de máquina, treinados para o reconhecimento da Língua Brasileira de Sinais por meio de análise e classificação em imagens digitais. Além disso, construímos uma base de dados própria com 350 vídeos, com duração aproximada de 3 segundos para cada um dos sinais em Libras, totalizando mais de 94 mil *frames*.

Tabela 5. Comparativo entre os modelos LSTM e BiLSTM (Teste).

Modelo	Precision	Recall	F-Score	Accuracy
LSTM	80,50%	80,90%	80,50%	79,11%
BiLSTM	67,80%	66,20%	66,70%	77,07%

Com os resultados, foi possível identificar que o modelo LSTM obteve assertividade superior em relação ao modelo BiLSTM, com acurácia de 84,71% para o modelo LSTM e 77,07% para o modelo BiLSTM. Logo, a arquitetura LSTM se mostra mais adequada à classificação, sendo viável seu uso em sistemas de reconhecimento por imagem de sinais em Libras. Para trabalhos futuros, pretendemos realizar o reconhecimento de outros sinais e levar em consideração o registro de frases completas que representam situações reais do dia a dia de pessoas surdas.

AGRADECIMENTOS

Agradecimento a empresa Olist, por conceder tempo para dedicação a este projeto, sem este apoio não seria possível desenvolver este trabalho. Os autores agradecem também ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível (CAPES) - Código de Financiamento 001.

REFERÊNCIAS

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Ringel Morris, M. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, 16–31. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3308561.3353774. URL https://doi.org/10.1145/3308561.3353774.

BRASIL (2002). Lei federal nº 10.436, de 24 de abril de 2002. dispõe sobre a língua brasileira de sinais - libras e dá outras providências. URL http://www.planalto.gov.br/ccivil_03/LEIS/2002/L10436.htm. Acessado: 2022-03-20.

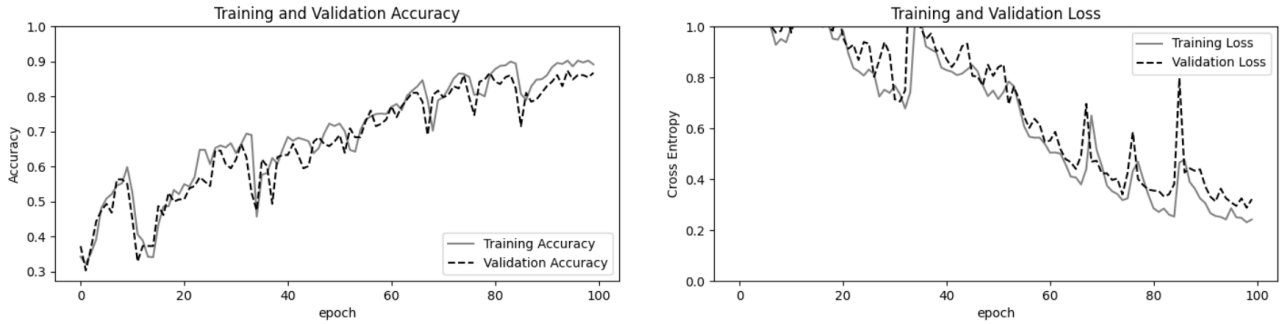
Bull, H., Gouiffès, M., and Braffort, A. (2020). Automatic segmentation of sign language into subtitle-units. In *Computer Vision – ECCV 2020 Workshops*, 186–198. Springer International Publishing. doi:10.1007/978-3-030-66096-3_14. URL https://doi.org/10.1007\2F978-3-030-66096-3_14.

Caetano, M. and Passos, M. (2017). A utilizaÇÃo dos softwares vlibras e hand talk no processo de inclusÃo de alunos com deficiÊncia auditiva em uma escola regular.

IBGE (2010). Censo demográfico 2010: características gerais da população, religião e pessoas com deficiência. URL https://biblioteca.ibge.gov.br/visualizacao/periodicos/94/cd_2010_religiao_deficiencia.pdf. Acesso em: 2022-02-12.

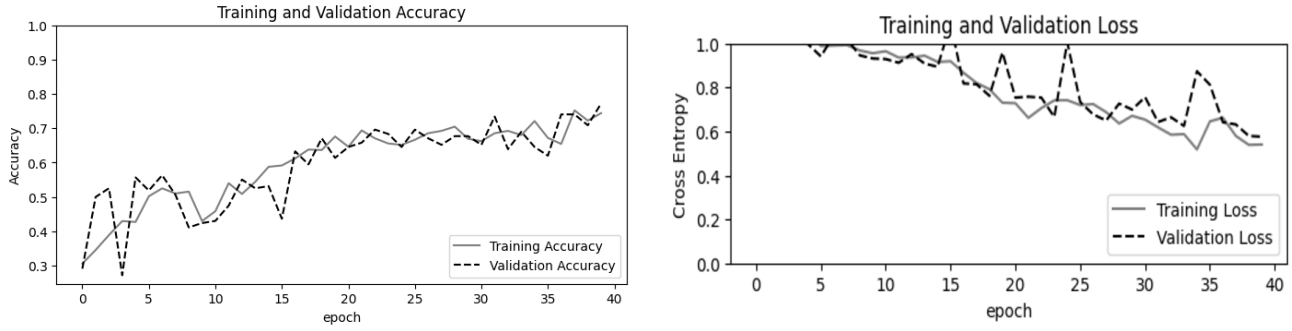
Liu, T., Zhou, W., and Li, H. (2016). Sign language recognition with long short-term memory. In *2016 IEEE International Conference on Image Processing (ICIP)*, 2871–2875. doi:10.1109/ICIP.2016.7532884.

Misra, S. and Li, H. (2020). Chapter 7 - deep neural network architectures to approximate the fluid-filled pore size distributions of subsurface geological formations. In S. Misra, H. Li, and J. He (eds.), *Machine Learning for Subsurface Characterization*, 183–217. Gulf Professional Publishing. doi: https://doi.org/10.1016/B978-0-12-817736-5.00007-7. URL https://www.sciencedirect.com/science/article/pii/B9780128177365000077.



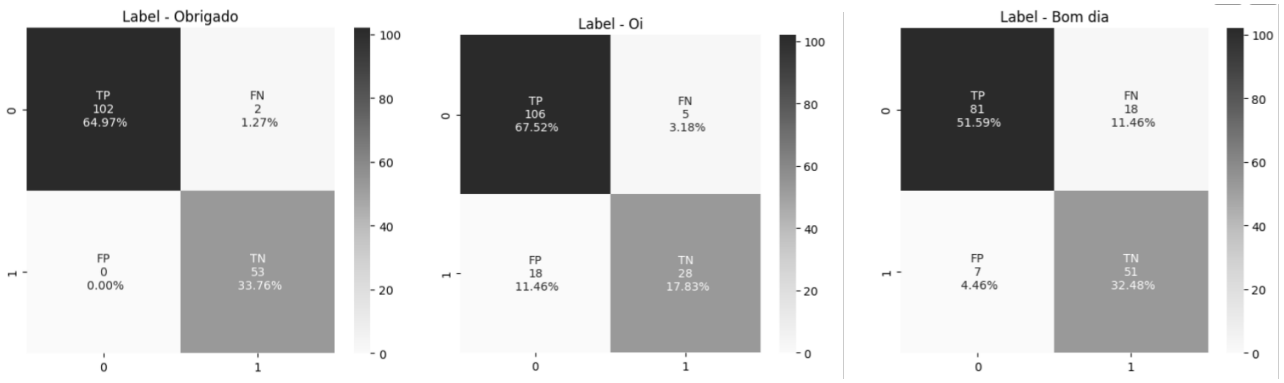
(a) Acurácia (b) Função de Perda

Figura 4. Análise de erro do modelo LSTM durante o treinamento.



(a) Acurácia (b) Função de Perda

Figura 5. Análise de erro do modelo BiLSTM durante o treinamento.



(a) Sinal "Obrigado" (b) Sinal "Oi" (c) Sinal "Bom dia"

Figura 6. Matrizes de confusão para a classificação com o modelo LSTM.

Mittal, A., Kumar, P., Roy, P.P., Balasubramanian, R., and Chaudhuri, B.B. (2019). A modified lstm model for continuous sign language recognition using leap motion. *IEEE Sensors Journal*, 19(16), 7056–7063. doi:10.1109/JSEN.2019.2909837.

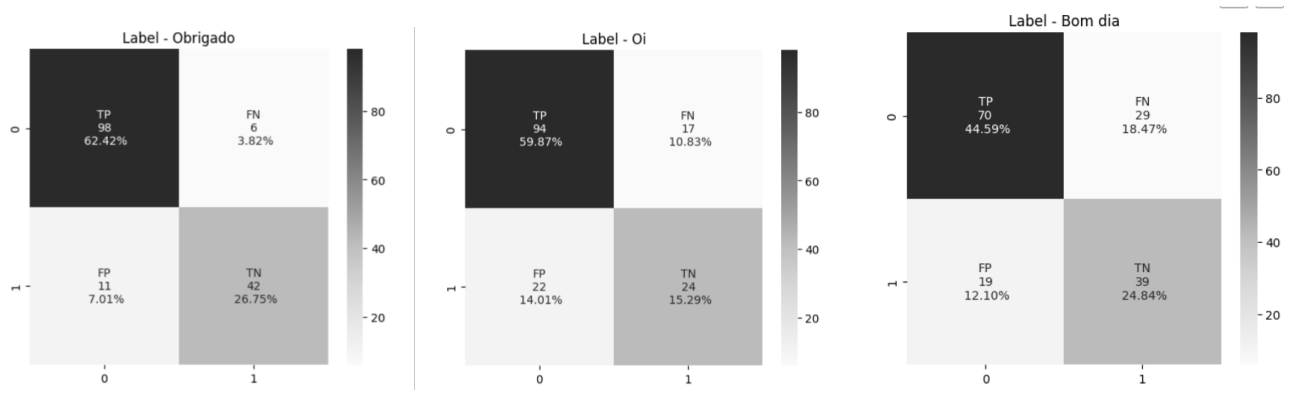
Park, H., Lee, Y., and Ko, J. (2021). Enabling real-time sign language translation on mobile platforms with on-board depth cameras. 5(2). doi:10.1145/3463498. URL `https://doi.org/10.1145/3463498`.

Passos, W.L., Araujo, G.M., Gois, J.N., and de Lima, A.A. (2021). A gait energy image-based system for brazilian sign language recognition. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(11), 4761–4771. doi:10.1109/TCSI.2021.3091001.

Raval, J.J. and Gajjar, R. (2021). Real-time sign language recognition using computer vision. In *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, 542–546. doi:10.1109/ICSPC51351.2021.9451709.

Siami-Namini, S., Tavakoli, N., and Namin, A.S. (2019). The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)*, 3285–3292. doi:10.1109/BigData47090.2019.9005997.

Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., and Alikhani, M. (2021). Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*



(a) Sinal "Obrigado" (b) Sinal "Oi" (c) Sinal "Bom dia"

Figura 7. Matrizes de confusão para a classificação com o modelo BiLSTM.

Joint Conference on Natural Language Processing (Volume 1: Long Papers), 7347–7360. Association for Computational Linguistics, Online. doi:10.18653/v1/2021.acl-long.570. URL `https://aclanthology.org/2021.acl-long.570`.