Comparação de Modelos de Aprendizado de Máquina no Reconhecimento de Línguas de Sinais

Guilherme Johann Reckziegel Nunes[1], **Janete Inês Müller**[1], **Fernando Luís Herrmann**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense (IFSul)
Campus de Venâncio Aires
Venâncio Aires – RS – Brasil

guilhermenunes.va015@academico.ifsul.edu.br

janetemuller@ifsul.edu.br

fernandoherrmann@ifsul.edu.br

Abstract. *With the recent developments in the area of Computer Vision, the potential assistance that can be provided for communication between the deaf and the hearing becomes apparent. In light of that, this work compared Machine Learning models to determine whether this hypothesis is viable with the resources available in the current day. To accomplish this, three models trained to classify static signs of three sign languages were tested. It was found that the models were incapable of classifying the tested signs with the expected accuracy. It was supposed that the reason for this is due to the small quantity and variety of the data present in the datasets utilized to train the models.*

Resumo. *Com os desenvolvimentos recentes na área de Visão Computacional, torna-se evidente o potencial auxílio que pode ser oferecido na comunicação entre surdos e ouvintes. Para tanto, este trabalho buscou comparar modelos de Aprendizado de Máquina a fim de verificar se essa hipótese é viável com os recursos disponíveis na atualidade. Para realizar esta tarefa, foram comparados três modelos treinados para classificar sinais estáticos de três línguas de sinais. Foi encontrado que os modelos não são capazes de reconhecer com a acurácia esperada os sinais testados. Supõe-se que a razão disso se deve à pequena quantidade e variedade dos dados pertencentes aos conjuntos de dados utilizados para o treinamento dos modelos.*

1. Introdução

Devido à falta de difusão da Língua Brasileira de Sinais (Libras), há a necessidade de auxiliar a comunidade surda no contexto de inclusão, facilitando as interações diárias. Neste contexto, os novos desenvolvimentos nas tecnologias de Aprendizado de Máquina e Visão Computacional permitem o desenvolvimento de ferramentas que possam auxiliar a comunicação entre surdos e ouvintes, realizando a tradução de Libras para o Português e vice-versa ou até mesmo auxiliando no aprendizado de Libras.

Um dos grandes empecilhos para a utilização de Aprendizado de Máquina no contexto de reconhecimento de sinais em Libras é a escassez de bases de dados com a qualidade, escopo e quantidade necessários para realizar o treinamento [Rodrigues 2021]. A falta de conjuntos de dados em especial para Libras pode ser vista na pequena quantidade

de publicações com este tema. Conforme um Mapeamento Sistemático da Literatura[1] foram encontradas apenas três (3) publicações com o tema de Libras, enquanto outras línguas de sinais apresentam consideravelmente mais publicações, por exemplo, a Língua Americana de Sinais (ASL), com cinquenta e seis (56) publicações e a Língua Indiana de Sinais (ISL), com trinta (30) publicações [Rodrigues 2021].

Com o objetivo de verificar as possibilidades levantadas, este trabalho buscou comparar modelos de Aprendizado de Máquina que utilizam Visão Computacional para reconhecer sinais estáticos de línguas de sinais em tempo real. Os modelos testados utilizam estratégias diferentes para a realização do reconhecimento, sendo elas Rede Neural Convolucional e Vision Transformer. Além disso, os modelos foram treinados para reconhecer línguas diferentes, sendo elas Libras, a Língua Americana de Sinais e a Língua Indiana de Sinais. Os modelos foram comparados usando métricas diversas com base nos resultados encontrados sobre os *datasets* de treinamento, teste e *datasets* desconhecidos pelos modelos.

2. Metodologia

A metodologia envolveu a utilização do ambiente de testes Miniconda através da linguagem de programação Python e com o auxílio de notebooks da biblioteca Jupyter para a realização de análise dos dados [Müller and Guido 2016] [Pedregosa et al. 2011]. Foram realizados testes diversos, verificando a acurácia, a função de perda e a matriz de confusão [Gabriel Filho 2023], assim como, a capacidade dos modelos de reconhecer sinais em tempo real através da utilização das bibliotecas OpenCV e MediaPipe. Para estes testes, foram utilizados sinais estáticos através do sistema datilológico do alfabeto das línguas de sinais mencionadas.

2.1. Modelos

Foram testados três modelos encontrados em plataformas diferentes. O primeiro modelo foi treinado através da plataforma Teachable Machine, um experimento da empresa Google que permite que o usuário treine um modelo para reconhecer imagens definindo as suas próprias classes. Para este treinamento, foram utilizadas imagens da Língua Brasileira de Sinais, providas pelo autor. O modelo é treinado sobre o MobileNet [Howard et al. 2017], um tipo de Rede Neural Convolucional, que, por sua vez, é treinado sobre o conjunto de imagens ImageNet [Deng et al. 2009]. Além disso, o MobileNet [Howard et al. 2017] também é voltado para ambientes com pouca capacidade de processamento, como celulares. Esse pré-treinamento sobre o ImageNet [Deng et al. 2009] permite que o modelo generalize alguns padrões das imagens que já conhece, bem como diminua o custo do processamento no aprendizado de novas imagens.

O segundo modelo foi encontrado na plataforma Hugging Face e foi treinado em um conjunto de dados da Língua Indiana de Sinais sobre o modelo Vision Transformer, definido como ViT, que, por sua vez, foi treinado sobre o conjunto de dados ImageNet-21k, com 14 milhões de imagens e 21.843 classes de diversos contextos.

O terceiro modelo foi selecionado da plataforma Kaggle, sendo uma Rede Neural Convolucional treinada em um conjunto de dados da Língua Americana de Sinais. Devido

[1] O Mapeamento Sistemático da Literatura realizado por [Rodrigues 2021] utilizou os motores de busca Scopus, ACM Digital Library, IEEExplore e Web of Science.

à falta do modelo no repositório do Kaggle, o código de treinamento foi replicado na plataforma Google Colab e o modelo foi então exportado para testes posteriores.

2.2. Conjuntos de dados

Foram utilizados seis conjuntos de dados, sendo que as 3 línguas de sinais estudadas estão representadas cada uma em dois dos conjuntos de dados — um com os dados de treinamento e teste do modelo e outro inédito, destinado a avaliar sua capacidade de generalização. Os conjuntos A e B correspondem à Libras (criado via Teachable Machine e obtido no Kaggle, respectivamente), C e D à ISL (provenientes do Kaggle e Hugging Face), e E e F à ASL (ambos do Kaggle). Como nem todos os conjuntos tornaram públicas as suas divisões de treino e teste, a metodologia adotada consistiu em realizar as avaliações utilizando o *dataset* completo, assegurando a uniformidade da comparação entre modelos e bases. As características dos *datasets* foram compiladas para referência na tabela 1.

Tabela 1. Características dos Conjuntos de Dados

Língua	***Datasets***	**Nº classes**	**Média de registros por classe**	**Total de registros**
Libras	*Dataset* A	21	300,00	6.300
Libras	*Dataset* B treinamento	21	1.635,05	34.714
Libras	*Dataset* B teste	21	549,90	11.548
ISL	*Dataset* C	35	1.121,29	42.745
ISL	*Dataset* D treinamento	32	268,28	8.585
ISL	*Dataset* D teste	32	47,28	1.513
ASL	*Dataset* E treinamento	24	1.143,96	27.455
ASL	*Dataset* E teste	24	298,83	7.172
ASL	*Dataset* F	29	3.000,00	87.000

O código-fonte desenvolvido, os modelos utilizados, os links para os *datasets* e os resultados compilados encontram-se disponíveis em um repositório[2] público no GitHub, garantindo a reprodutibilidade e a transparência da pesquisa.

Com os modelos e os *datasets* em mãos, foram realizados testes diversos para comparar a eficácia dos modelos no reconhecimento de sinais. A hipótese levantada pelo trabalho foi que os modelos seriam capazes de reconhecer os sinais com alta eficácia e em tempo adequado para o uso diário. Entretanto, conforme os resultados obtidos, a hipótese foi refutada, tendo sido encontrado que os modelos não apresentaram resultados positivos quando testados sobre conjuntos de dados desconhecidos.

3. Resultados

Os resultados encontrados foram compilados em uma tabela agrupando as acurácias dos modelos testados. Conforme a tabela 2, infere-se que os modelos possuem a capacidade de classificar os sinais corretamente, com base nos resultados positivos nos conjuntos de dados de treinamento e teste (*Datasets* A, C e E). Entretanto, quando testados em

[2]https://github.com/GuiJRNunes/sign-language-cv

conjuntos de dados desconhecidos pelos modelos, a acurácia é consideravelmente menor (*Datasets* B, D e F). A análise realizada por este trabalho supõe que isso ocorre devido às características dos conjuntos de treinamento e teste utilizados pelos modelos, sendo elas uma quantidade e qualidade insuficiente de dados. Com bases maiores e de maior variedade, acredita-se que seja possível desenvolver modelos que apresentem resultados muito mais promissores.

Tabela 2. Comparação da acurácia dos modelos

Modelo	**Acurácia nos datasets de treinamento e teste**	**Acurácia em datasets desconhecidos pelos modelos**
Teachable Machine (Libras)	*Dataset* A - 100%	*Dataset* B - 4,76%
VIT-ISLC (Hugging Face - ISL)	*Dataset* C - 100%	*Dataset* D - 7,67%
RNC (Kaggle - ASL)	*Dataset* E - 99,64%	*Dataset* F - 10,99%

4. Conclusão

Conforme descrito, há uma falta de difusão da Língua Brasileira de Sinais na atualidade, o que demonstra uma dificuldade enfrentada pela comunidade surda no âmbito da comunicação. Neste contexto, este trabalho buscou verificar se modelos de Aprendizado de Máquina utilizando Visão Computacional poderiam auxiliar na tradução ou até no ensino de línguas de sinais. Todavia, foi encontrado que os modelos testados não possuem a capacidade de realizar esta tarefa com a acurácia desejada. Acredita-se que isso ocorre devido à insuficiência e pequena variedade dos dados dos conjuntos de dados utilizados para treinar os modelos.

Para trabalhos futuros foi proposta a construção de conjuntos de dados mais extensos, assim como, a análise mais profunda das tecnologias e pesquisas realizadas na área de Visão Computacional e de reconhecimento de sinais de línguas de sinais.

Referências

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Gabriel Filho, O. (2023). Inteligência artificial e aprendizagem de máquina: aspectos teóricos e aplicações. *São Paulo: Blucher*.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861.

Müller, A. C. and Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. "O'Reilly Media, Inc.".

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Rodrigues, A. J. (2021). V-LIBRASIL: uma base de dados com sinais na língua brasileira de sinais (Libras). Master's thesis, Universidade Federal de Pernambuco.