Convolutional Neural Network Model for the Translation of Colombian Sign Language Gestures to Text Focused on the Colombian Hotel Sector

N. D. Naranjo Astaiza
Pontificia Universidad Javeriana Cali
nicolle01@javerianacali.edu.co

M. J. Suarez Peña
Pontificia Universidad Javeriana Cali
mariasuarez@javerianacali.edu.co

ABSTRACT

The main objective of this project is to support individuals with auditory sensory disabilities, enabling them to carry out basic actions in environments where personnel are not adequately trained to offer completely suitable assistance, as is often the case in the Colombian hotel sector. Therefore, developing a computational tool capable of recognizing and translating manual gestures from Colombian Sign Language (LSC) to text is increasingly necessary. This tool aims to address a real need in society and facilitate inclusive attention and a positive experience for all involved.

KEYWORDS

Colombian Sign Language (LSC), Sensory disabilities, Machine learning, Convolutional neural network (CNN), Accessibility.

I INTRODUCTION

Auditory sensory diversities present significant challenges for individuals in various aspects of daily life, especially in contexts where communication barriers persist, such as in tourism an the hotel industry. The difficulty in reading and writing in Spanish among people with auditory disabilities in Colombia demonstrates the need for tools that enable effortless expression in the language they know. Additionally, inadequate training among hotel staff affects the independence and communication of people who depend on sign language, highlighting the urgent need to implement measures to support them.

In response, the development of a computational tool capable of detecting and translating manual gestures from LSC into text emerges as a promising proposal. This tool aims to facilitate communication between individuals who depend on LSC and people such as hotel staff who lack sign language skills, improving accessibility and contributing to a more inclusive experience in hotel environments.

The formulation and development of this computational tool involve addressing challenges such as dataset preparation, model training, evaluation, and the development of context-adapted prototypes. By exploring these aspects, this research seeks to provide practical insights into leveraging technology to support individuals with auditory sensory disabilities and create a prototype as a proposed solution.

II KEY CONCEPTS

For this research, it is necessary to define some concepts related to technology and applied theory:

A MACHINE LEARNING

Machine learning is a branch of artificial intelligence (AI) that focuses on the use of data and algorithms to imitate the way that humans learn [1].

B NEURAL NETWORK

A neural network serves as a type of machine learning framework or model that imitates the decision-making mechanisms of the human brain. It works by replicating how biological neurons cooperate to identify patterns, evaluate different options, and make deductions. [2].

C CONVOLUTIONAL NEURAL NETWORK

A Convolutional Neural Network (CNN) is a specialized neural network used for image recognition, relying on extensive labeled datasets and high-performance processors like GPUs or NPUs to yield fast results [3].

D TRANSFER LEARNING

It is a machine learning technique in which knowledge gained through one task or dataset is used to improve model performance on another related task and/or different dataset. Transfer learning uses what has been learned in one setting to improve generalization in another setting[4].

III METHOD

In this section, the origin and composition of the dataset are initially depicted, followed by the presentation of the models utilized for training.

A DATASET

The dataset consists of two folders: 'Training' and 'Validation', each containing 39 subfolders representing the classes of the gestures [5]. Additionally, a new folder named 'Test' was created, which contains 50 randomly selected images of each subfolder from the 'Training' folder.

B CNN MODELS

For the identification and classification of images, CNNs were chosen due to their architecture's ability to easily adapt to image dimensions and characteristics. Their structure is designed to capture relevant visual aspects and patterns efficiently. One research challenge is to propose and find the most competent convolutional network models to achieve optimal results to apply to the prototype. Therefore, three types of CNN architectures were compared: a simple initial model created for this research, a simplified model from other researchers with more layers, and a well known pre-trained model: VGG16. Additionally, hyperparameters and metrics such as: precision, recall, f1-Score and accuracy were implemented in each model as tools and indicators for precise evaluation and better feedback.

IV RESULTS

In this section, the results of all models will be presented and compared to identify the model with the highest accuracy.

A INITIAL MODEL

	Precision	Recall	F1-score	Support
Accuracy			0.97	1950
Macro avg	0.98	0.97	0.97	1950
Weighted avg	0.98	0.97	0.97	1950

Figure 1: Initial model metrics

As shown in Figure 1, this model achieves an accuracy of 97%, indicating a promising start. This suggests that the dataset does not necessitate a complex network to achieve high metrics.

	Precision	Recall	F1-score	Support
Accuracy			0.98	1950
Macro avg	0.98	0.98	0.98	1950
Weighted avg	0.98	0.98	0.98	1950

Figure 2: Initial model metrics with hyperparameters

As seen in Figure 2, the model's accuracy has improved to 98%, indicating a slight enhancement. In this instance, the discovered parameters outperformed those of the basic model (see Figure 1), resulting in improved accuracy.

B SIMPLIFIED MODEL

	Precision	Recall	F1-score	Support
Accuracy			0.96	1950
Macro avg	0.97	0.96	0.96	1950
Weighted avg	0.97	0.96	0.96	1950

Figure 3: Simplified model metrics

In this case, the model achieved an accuracy of 96%, which while satisfactory, falls below that of the two preceding models as illustrated in Figures 1 and 2).

	Precision	Recall	F1-score	Support
Accuracy			0.93	1950
Macro avg	0.93	0.93	0.92	1950
Weighted avg	0.93	0.93	0.92	1950

Figure 4: Simplified model metrics with hyperparameters

In this instance, the outcome was different. As illustrated in Figure 4, the metrics of this model performed less favorably compared to the model without hyperparameters (refer to Figure 3). The accuracy decreased from 96% to 93%.

C PRE-TRAINED MODEL

Accuracy			1.00	1950
Macro avg	1.00	1.00	1.00	1950
Weighted avg	1.00	1.00	1.00	1950

Figure 5: Pre-trained model metrics

In this model, as depicted in Figure 5, it achieved 100% accuracy. However, it's important to note that these metrics are approximated; the actual accuracy is 99.79%, making it the highest accuracy obtained thus far.

	Precision	Recall	F1-score	Support
Accuracy			0.98	1950
Macro avg	0.98	0.98	0.98	1950
Weighted avg	0.98	0.98	0.98	1950

Figure 6: Pre-trained model metrics with hyperparameters

The final model achieved an accuracy of 98%, which, while slightly lower than the original (refer to Figure 5), still demonstrates promise for effective performance.

V PROTOTYPE

Utilizing the highest accuracy model (refer to Figure 5), the development approach prioritized the creation of a user-friendly interface. The primary objective was to ensure ease of understanding and quick adaptation for users, particularly those new to the system.

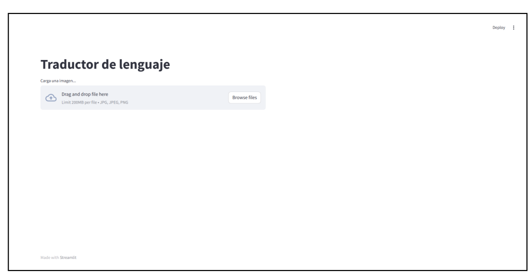

Figure 7: Main interface

The application is developed to upload images in JPG, JPEG, and PNG formats.

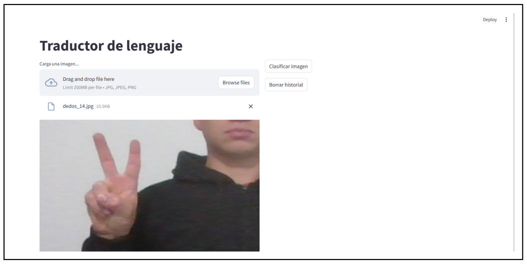

Figure 8: interface

As illustrated in Figure 8, the image is pre-visualized along with its corresponding name. When the 'Clasificar imagen' button is clicked, the app will display the interpretation of the image gesture.

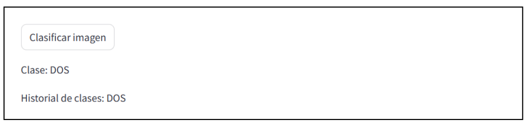

Figure 9: 'Clasificar imagen' button

Moreover, the meaning of every image uploaded is saved to form a complete sentence. For instance, in Figure 10, there is an example of a complete sentence comprised of the words: 'DOS', 'CAMA', and 'POR FAVOR'."

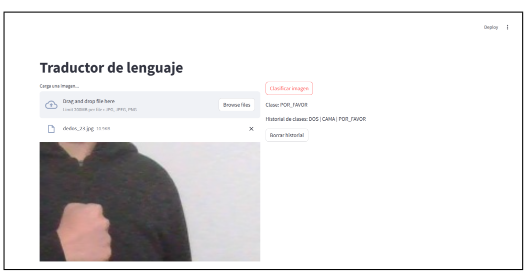

Figure 10: Example

VI CONCLUSION

This research successfully prepared a dataset to train three different convolutional neural network models, all of which yielded strong results and proved to be viable options for the proposed solution, with each achieving an accuracy of over 90% in data prediction. Although hyperparameters were used to optimize results, not all models improved their performance, likely due to limited search ranges and iterations to avoid computational costs and waiting times. Ultimately, the pretrained VGG16 model (without hyperparameters) performed the best, as expected, with 99.79% accuracy, thanks to its deeper architecture and application of transfer learning. This model facilitated the development of a prototype capable of translating gestures from the dataset into text to form sentences successfully, providing a potentially sufficiently effective tool for real-world application.

References

[1] IBM, "What is machine learning?" [Online]. Available: https://www.ibm.com/topics/machine-learning

[2] ——, "What is a neural network?" [Online]. Available: https://www.ibm.com/topics/neural-networks

[3] Arm Ltd. What is a convolutional neural network (cnn). Arm — The Architecture For The Digital World. [Online]. Available: https://www.arm.com/glossary/convolutional-neural-network#:~:text=A%20convolutional%20neural%20network%20(CNN)%20is%20a%20type%20of%20artificial,labelled%20data%20points%20for%20training

[4] IBM, "What is transfer learning?" [Online]. Available: https://www.ibm.com/topics/transfer-learning

[5] LSC - Conjunto de datos sector hotelero, "LSC - Conjunto de datos sector hotelero," https://www.kaggle.com/datasets/jimmyalejandro/lsc-conjunto-de-datos-sector-hotelero?resource=download, 2022, accedido el 14 de marzo de 2022.