Intérprete automático de Lengua de Señas

Aybar Lourdes Maria Magdalena[1], Benitez Josefina Victoria[1] y Juarez Victor Manuel[1].

[1] Laboratorio de Investigación y Desarrollo de Software e Inteligencia Artificial , Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Argentina.
{lourdes.aybar, josefina.victoria.benitez, manuel.juarez}@mi.unc.edu.ar

Resumen.

El artículo "Intérprete automático de Lengua de Señas" contribuye al desarrollo de un sistema basado en inteligencia artificial que interprete la lengua de señas argentina, priorizando la captura de matices emocionales y expresivos. Se espera impactar positivamente para mejorar la comunicación entre personas sordas y oyentes, así como en la accesibilidad para personas con discapacidad auditiva.

La metodología incluye la recopilación y preprocesamiento de bases de datos compuestas por videos, seguido de la implementación de diversos modelos de inteligencia artificial y su evaluación comparativa enfocada en la precisión y capacidad para captar la expresión emocional. Se basa en antecedentes recopilados de investigaciones previas sobre la traducción automática de lengua de señas utilizando lenguaje natural y aprendizaje automático.

Este trabajo forma parte de una línea de investigación cuyo objetivo consiste en desarrollar aplicaciones para la traducción de lengua de señas, dentro del proyecto de investigación "Enfoque Integral de la Inteligencia Artificial Aplicada, orientado a Problemas Emergentes". Está radicado en el Laboratorio de Investigación y Desarrollo de Software e Inteligencia Artificial (LIDeSIA) de la FCEFyN en la Universidad Nacional de Córdoba, que sostiene como práctica habitual incorporar estudiantes de pregrado, dirigidos por investigadores expertos, para contribuir con la formación de talento humano.

Palabras claves: Intérprete, Lengua de señas, Aprendizaje automático, Lenguaje natural, Inteligencia artificial.

1 Introducción

La comunicación es un derecho fundamental que todos los individuos deberían poder ejercer plenamente, independientemente de sus capacidades físicas o sensoriales. En este contexto, la lengua de señas desempeña un papel crucial en la comunicación de las personas con discapacidad auditiva. Sin embargo, la traducción y comprensión de esta lengua por parte de individuos no familiarizados con ella puede representar un desafío significativo. En respuesta a esta necesidad, surge "Intérprete automático de Lengua de Señas", parte del proyecto de investigación acreditado y financiado por la Secretaría de Ciencia y Tecnología (Se.C. y T.) U.N.C.[1], titulado "Enfoque Integral de la Inteligencia Artificial Aplicada, orientado a Problemas Emergentes".

Los autores de este trabajo, Aybar Lourdes María M., Benítez Josefina Victoria y Juárez Víctor Manuel, somos investigadores en formación de pregrado, dirigidos por la Dra. Ing. Laura C. Diaz Dávila, y acompañados por docentes y egresados que conforman el equipo de investigadores del laboratorio. Es práctica habitual de los proyectos de investigación que se radican en LIDeSIA[2] y en la UNC incorporar a estudiantes de los últimos años de ingeniería como investigadores en formación, como una forma de contribuir a la formación del talento humano.

El mencionado proyecto busca desarrollar un sistema capaz de interpretar y traducir la lengua de señas. Si bien existen soluciones aisladas para algunos idiomas o arquitecturas planteadas, buscamos profundizar la exploración para lograr un intérprete basado en inteligencia artificial que ofrezca una traducción menos robotizada, priorizando la captura de matices emocionales y expresivos.

1.1 Antecedentes

LIDeSIA se dedica a abordar desafíos desde cuatro ejes fundamentales: 1- la obtención y procesamiento de datos y los requerimientos de infraestructura para los modelos de inteligencia artificial (IA), 2- el desarrollo de modelos adecuados para la solución de problemas, 3- la

matemática asociada a los algoritmos de IA y 4- el estudio del impacto ético y regulatorio de estas soluciones.

En este marco para dar tratamiento a la problemática, se llevó a cabo una recopilación de antecedentes relacionados con investigaciones previas. Estos trabajos han establecido una base sólida de conocimientos y experiencias que han influido de manera significativa en nuestro interés por esta área. A continuación, se presenta una síntesis de los artículos más relevantes, entre los cuales se destacan las investigaciones que han abordado la captura y análisis de gestos manuales y expresiones faciales para la traducción de lengua de señas.

"Diseño de Prototipo de Software para Traducción de Lenguaje de Señas en Texto para Personas Sordas en Colombia"[3]: en este trabajo se llevó a cabo el desarrollo de un prototipo de software que convierta la lengua de señas en texto para la comunidad sorda colombiana mediante técnicas de aprendizaje automático. El enfoque se basa en una red neuronal convolucional o Convolutional Neural Network (CNN)[4]. Además de utilizar Transfer Learning[5] y Fine Tuning[6] para entrenar capas adicionales.

"Reconocimiento de lengua de señas colombiana mediante CNN y captura del movimiento"[7]: el artículo describe un modelo computacional basado en redes neuronales convolucionales para reconocer la lengua de señas colombiana (LSC) en el sector hotelero y turístico. Se utiliza inteligencia artificial y aprendizaje profundo para predecir gestos en tiempo real, utilizando datos capturados por cámaras de dispositivos móviles. Se evaluó su rendimiento con medidas categóricas y se compararon configuraciones de red neuronal. Se utilizó Tensorflow[8], OpenCV[9] y MediaPipe[10] para el desarrollo y soporte del proyecto.

"Reconocimiento de Lengua de Señas usando Redes Neuronales Recurrentes (RNN)"[11]: el artículo documenta un sistema de reconocimiento automático de lenguas de señas que convierte señas capturadas en video a texto. Este proceso implica identificar la forma de las manos, su movimiento, posición, expresiones faciales y el fondo en cada fotograma. Posteriormente, las señas se clasifican y traducen a lenguaje escrito, como español o inglés.

"Transformador basado en poses de señas para el reconocimiento de lengua de señas a nivel de palabras"[12]: en este artículo se introduce a un sistema basado en SPOTER un Transformer[13] para el reconocimiento de lengua de señas a nivel de palabras. El sistema se centra en la estimación de la pose del cuerpo humano mediante puntos de referencia en 2D y emplea un esquema robusto de normalización de poses. Este esquema considera el espacio ocupado por las señas y procesa las poses de la mano en un sistema de coordenadas.

"Integración de reconocimiento y traducción de lengua de señas de extremo a extremo"[14]: arquitectura basada en transformer, combina el reconocimiento y la traducción de lengua de señas de manera continua y se entrena de extremo a extremo. Utiliza una pérdida de CTC (Clasificación Temporal Conexionista) para unificar el reconocimiento y la traducción en una sola arquitectura, sin necesidad de información en tiempo real. Planean expandir el enfoque para modelar múltiples articuladores de señas, como rostros, manos y cuerpo, y promover el aprendizaje de las relaciones lingüísticas entre ellos.

Competencia de Google: *"Reconocimiento de Deletreo en Lengua de Señas Americana (ASL)"[15]*. El objetivo de esta competencia fue desarrollar un modelo capaz de detectar y traducir el deletreo de dedos en ASL a texto.

Características del modelo ganador: Realizaron la optimización del modelo "Squeezeformer" para procesar puntos de referencia de MediaPipe en lugar de señales de voz.

Competencia de Google*: "Reconocimiento de lengua de señas aislado para mejorar los juegos educativos de PopSing para aprender Lengua de Señas Americana (ASL)"[16]*.

El objetivo es mejorar los juegos educativos de PopSign[17], una aplicación que facilita el aprendizaje de la ASL mediante juegos interactivos. Google estableció que la solución debía realizar un reconocimiento de lengua de señas aislado utilizando un modelo TensorFlow Lite[18] entrenado con datos etiquetados de MediaPipe Holistic[19].

Características del modelo ganador: Se utilizó una combinación de una CNN unidimensional y un Transformer para abordar el problema de reconocimiento. El modelo se entrenó desde cero con TensorFlow y una Unidad de Procesamiento Tensorial (TPU)[20] en Google Colab[21] para asegurar la compatibilidad con TensorFlow Lite.

"Plataforma de Enseñanza de Lengua de Señas Argentina (LSA) con Inteligencia Artificial 'Eldes'"[22]: Primera plataforma de LSA con IA para enseñanza y preservación de lengua de señas argentina. Es de retroalimentación inmediata basada en detección de movimientos de manos y rostro.

1.2 Pruebas de concepto

La recopilación y análisis de investigaciones previas demuestran la diversidad y complejidad de los enfoques adoptados para abordar este desafío. Desde la detección de gestos y expresiones faciales hasta el reconocimiento de patrones y poses del cuerpo humano, cada estudio contribuyó para orientar la línea a seguir en la construcción del sistema buscado. Es por esto que se realizaron pruebas de concepto en paralelo al estudio de antecedentes y arquitecturas.

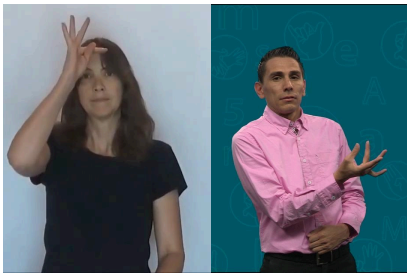
Además, en esta fase se han integrado otros subequipos con distinta expertise disciplinaria y distintivos niveles de investigadores, entre ellos dos estudiantes de la carrera de Ingeniería en Computación en proceso de proyecto integrador para ajustar uno de los antecedentes*[16]* antes mencionado a un prototipo con respuesta en el entorno académico.

Por otro lado, se ha establecido contacto con la Asociación Argentina de Sordos para dar a conocer nuestro proyecto. Así como también establecer comunicación con otros grupos de investigación para fortalecer nuestros vínculos y seguir avanzando en este campo tan importante y necesario.

Implementación de transfer learning[1]*:* se utilizó YOLOv8[23] el cual se centra en la detección de objetos en imágenes y no tiene la capacidad de comprender y traducir lengua de señas como un intérprete humano. Esta última tarea es mucho más compleja, ya que implica la interpretación de gestos, movimientos y expresiones faciales.

Se seleccionaron para la implementación dos palabras extraídas de datos públicos proporcionados por INSOR[24]: "*Hospital"* y *"Dolor"*.

En el preprocesamiento de los datos, se utilizaron 5 videos por palabra y de ellos 4 frames por segundo, lo que da un total de 105 imágenes para el entrenamiento.



Señas INSOR: "Hospital" - "Dolor"

Como resultado se obtuvieron las siguientes métricas generales: *Precisión*[2] (P): 0,951; *Recall*[3] (R): 0,973; mAP50[4]: 0,987; mAP50-95[5]: 0,769.

En cuanto al aporte de datos, es importante destacar el contacto con nuestros aliados colombianos, quienes han avanzado en el campo de la traducción de la lengua de señas. Realizaron pruebas con señas estáticas del alfabeto en Random Forest[25] para su clasificación. Además del uso de redes neuronales recurrentes Long short-term memory (LSTM)[26] para identificar una selección de señas.

[1]IA-LIDeSIA-Dpto-Computacion/Prueba_de_Concepto_YOLOv8_LIDeSIA: https://github.com/lauraceciliadiazdavila/IA-LIDeSIA-Dpto-Computacion/tree/main/Prueba_de_Concepto_YOLOv8_LIDeSIA [Accessed: Jul. 11, 2024]

[2] Precisión (P): "La precisión de los objetos detectados, indicando cuántas detecciones fueron correctas". Ultralytics. (2023). *YOLO Performance Metrics*. Ultralytics.com.

[3] Recall (R): "La capacidad del modelo para identificar todas las instancias de objetos en las imágenes". Ultralytics. (2023). *YOLO Performance Metrics*. Ultralytics.com.

[4] (mAP50): "Precisión media calculada con un umbral de intersección sobre unión (IoU) de 0,50. Es una medida de la precisión del modelo considerando sólo las detecciones "fáciles" ".

[5] (mAP50-95): "métrica que representa la media de la precisión media calculada con distintos umbrales de IoU, que van de 0,50 a 0,95. Ofrece una visión global del rendimiento del modelo en distintos niveles de dificultad de detección". Ultralytics. (2023). *YOLO Performance Metrics*. Ultralytics.com

2 Conclusiones

Tras el estudio de antecedentes y pruebas, se determinó que arquitecturas como YOLOv8, que aunque demostró ser capaz de distinguir señas en movimiento en una prueba simple de dos palabras, se vuelve insuficiente para lograr la traducción completa de la lengua de señas. Esto se debe a que la lengua de señas es un sistema de comunicación visual con su propia gramática, estructura lingüística, variaciones regionales, así como expresiones faciales y corporales.

Frente a esto se decidió seguir como línea de acción el empleo de arquitecturas Transformer en combinación con frameworks como MediaPipe para alcanzar nuestro objetivo. Para ello, es necesario contar con una base de datos que pueda ser utilizada en su implementación. En Argentina, disponemos de las bases de datos LSA64[27] y LSA-T[28], las cuales están siendo procesadas y adaptadas para ser utilizadas como entrada a esta arquitectura, representando el estado actual del proyecto.

Posteriormente, se planifica llevar a cabo la implementación y prueba de la arquitectura Transformer BERT[29] en su versión DistilBERT[30].

Con esto se busca que el modelo obtenido sea parte de un sistema compuesto por arquitecturas de reconocimiento y procesamiento de imagen en tiempo real, conformando así una prueba de concepto del intérprete automático con las especificaciones definidas.

Referencias

1. "Contactos SECyT," Universidad Nacional de Córdoba, Nov. 21, 2016. https://www.unc.edu.ar/ciencia-y-tecnolog%C3%ADa/contactos-secyt Accessed: Jul. 11, 2024. [Online].
2. FCEFyN, "Laboratorio de Investigación y Desarrollo de Software e Inteligencia Artificial (LIDeSIA)," FCEFyN, UNC, 2023. https://fcefyn.unc.edu.ar/facultad/secretarias/investigacion-y-desarrollo/laboratorios/laboratorio-de-investigacion-y-desarrollo-de-software-e-inteligencia-artificial-lidesia/ Accessed: Jul. 11, 2024. [Online].
3. N. M. Ortiz Farfán, "Diseño y Construcción de Prototipo de Software para Reconocer Lenguaje de Señas de Personas con Discapacidad Auditiva", Trabajo de grado presentado, Univ. Nac. Colomb., Bogotá, Colombia, 2020. https://repositorio.unal.edu.co/bitstream/handle/unal/77440/1030545899.2020.pdf Accessed: Jul. 11, 2024. [Online].
4. C. de, "clase de las redes neuronales profundas, más comúnmente aplicada al análisis de imágenes visuales," Wikipedia.org, Jun. 23, 2014. Available: https://es.wikipedia.org/wiki/Red_neuronal_convolucional. Accessed: Jul. 11, 2024. [Online].
5. C. de, "Aprendizaje por transferencia," Wikipedia.org, Mar. 29, 2024. Available: https://es.wikipedia.org/wiki/Aprendizaje_por_transferencia. [Accessed: Jul. 11, 2024]
6. C. de, "Ajuste fino (aprendizaje profundo)," Wikipedia.org, Aug. 03, 2023. Available: https://es.wikipedia.org/wiki/Ajuste_fino_(aprendizaje_profundo). [Accessed: Jul. 11, 2024]
7. Plazas López., J.A. Gutiérrez Leguizamón., J.J. Suárez Barón., M.J. y González Sanabria., J.S. (2022). "Reconocimiento de lengua de señas colombiana mediante redes neuronales convolucionales y captura de movimiento". Tecnura, 26(74), 70-86. https://doi.org/10.14483/22487638.19213
8. C. de, "biblioteca de software para aprendizaje automático," Wikipedia.org, Mar. 31, 2017. Available: https://es.wikipedia.org/wiki/TensorFlow. Accessed: Jul. 11, 2024. [Online].
9. C. de, "librería de visión para computadoras," Wikipedia.org, Feb. 18, 2008. Available: https://es.wikipedia.org/wiki/OpenCV. [Accessed: Jul. 11, 2024]
10. "Framework de MediaPipe," Google for Developers, 2024. Available: https://ai.google.dev/edge/mediapipe/framework?hl=es-419. [Accessed: Jul. 11, 2024]
11. I. Mindlin, "Reconocimiento de Lengua de Señas con Redes Neuronales Recurrentes", Tesina de Licenciatura, Univ. Nac. Plata, La Plata,Buenos Aires, 2021. https://sedici.unlp.edu.ar/bitstream/handle/10915/129853/Documento_completo.pdf-PDFA.pdf Accessed: Jul. 11, 2024. [Online].
12. M. Boháček and M. Hrúz, "Sign Pose-based Transformer for Word-level Sign Language Recognition." https://openaccess.thecvf.com/content/WACV2022W/HADCV/papers/Bohacek_Sign_Pose-Based_Transformer_for_Word-Level_Sign_Language_Recognition_WACVW_2022_paper.pdf Accessed: Jul. 11, 2024. [Online].
13. C. de, "modelo de aprendizaje automático," Wikipedia.org, Jan. 20, 2023. Available: https://es.wikipedia.org/wiki/Transformador_(modelo_de_aprendizaje_autom%C3%A1tico). [Accessed: Jul. 11, 2024]
14. N. Cihan, O. Koller, S. Hadfield, and R. Bowden, "Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation." https://openaccess.thecvf.com/content_CVPR_2020/papers/Camgoz_Sign_Language_Transformers_Joint_End-to-End_Sign_Language_Recognition_and_Translation_CVPR_2020_paper.pdf Accessed: Jul. 11, 2024. [Online].

15. Google - American Sign Language Fingerspelling Recognition | Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/competitions/asl-fingerspelling Accessed: Jul. 11, 2024. [Online].
16. Google - Isolated Sign Language Recognition | Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/competitions/asl-signs Accessed: Jul. 11, 2024. [Online].
17. Learn about PopSign an educational ASL bubble shooter game – DPAN.TV". DPAN.TV – The Sign Language Channel.https://dpan.tv/learn-about-popsign-an-educational-asl-bubble-shooter-game/ Accessed: Jul. 11, 2024. [Online].
18. "TensorFlow Lite," Google for Developers, 2024. Available: https://ai.google.dev/edge/lite?hl=es-419. [Accessed: Jul. 11, 2024]
19. "Guía de tareas de detección de puntos de referencia holísticos," Google for Developers, 2024. Available: https://ai.google.dev/edge/mediapipe/solutions/vision/holistic_landmarker?hl=es-419. [Accessed: Jul. 11, 2024]
20. C. de, "coprocesador desarrollado por Google para acelerar redes neuronales artificiales," Wikipedia.org, Dec. 25, 2016. Available: https://es.wikipedia.org/wiki/Unidad_de_procesamiento_tensorial. [Accessed: Jul. 11, 2024]
21. "colab.google," colab.google, 2024. Available: https://colab.google/. [Accessed: Jul. 11, 2024]
22. ELdeS | Enseñanza de Lengua de Señas | Aprende a tus tiempos". ELdeS. https://www.somoseldes.com Accessed: Jul. 11, 2024. [Online].
23. "YOLOv8: A New State-of-the-Art Computer Vision Model," Yolov8.com, 2024. Available: https://yolov8.com/. [Accessed: Jul. 11, 2024]
24. admin, "INSOR | Instituto Nacional para Sordos – Trabajando por la Población Sorda Colombiana," Insor.gov.co, Jun. 28, 2024. Available: https://www.insor.gov.co/home/. [Accessed: Jul. 11, 2024]
25. C. de, "Random forest," Wikipedia.org, Jan. 03, 2013. Available: https://es.wikipedia.org/wiki/Random_forest. [Accessed: Jul. 11, 2024]
26. Wikipedia Contributors, "Long short-term memory," Wikipedia, Jul. 05, 2024. Available: https://en.wikipedia.org/wiki/Long_short-term_memory. [Accessed: Jul. 11, 2024]
27. "Papers with Code - LSA64 Dataset," Paperswithcode.com, 2022. Available: https://paperswithcode.com/dataset/lsa64. [Accessed: Jul. 11, 2024]
28. "Papers with Code - LSA-T Dataset," Paperswithcode.com, 2022. Available: https://paperswithcode.com/dataset/lsa-t. [Accessed: Jul. 11, 2024]
29. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv.org, 2018. Available: https://arxiv.org/abs/1810.04805. [Accessed: Jul. 11, 2024]
30. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv.org, 2019. Available: https://arxiv.org/abs/1910.01108. [Accessed: Jul. 11, 2024]