Análisis Exploratorio

Proyecto 2 - Data Science

1. Planteamiento inicial del problema

a. Situación problemática: Las personas sordas y con pérdida auditiva carecen de tecnologías accesibles para ingresar texto en dispositivos móviles. Aunque el fingerspelling en ASL (American Sign Language) es más rápido que escribir en teclados virtuales, aún no existen modelos robustos que automaticen su traducción a texto.

b. Problema científico: ¿Es posible entrenar un modelo de aprendizaje automático que traduzca de forma precisa y eficiente secuencias de fingerspelling en ASL a texto escrito a partir de datos capturados por una cámara de smartphone?

c. Objetivos:

- **General:** Desarrollar un modelo de deep learning que traduzca fingerspelling en ASL a texto escrito.
- **Específicos:**
  - Diseñar un proceso de preprocesamiento de landmarks que reduzca el volumen de datos y preserve la información relevante.
  - Entrenar y optimizar un modelo exportable a TensorFlow que cumpla con las restricciones de memoria y tiempo de procesamiento.

2. Investigación Preliminar sobre la sordera y el ASL

Pérdida auditiva

- Según la Organización Mundial de la Salud (OMS), más de 1500 millones de personas en el mundo tienen algún grado de pérdida auditiva y aproximadamente 430 millones padecen de una pérdida auditiva discapacitante (suficientemente severa como para necesitar rehabilitación), incluyendo 34 millones de niños (OMS, 2025).

- Se estima que de aquí a 2050, más de 700 millones de personas tendrán una pérdida auditiva discapacitante (OMS, 2025).

- La federación Mundial de Personas Sordas (WFD) indica que existen alrededor de 70 millones de personas sordas en el mundo, muchas de ellas utilizadoras de lenguaje de señas (Sign.mt, s.f.).

Velocidad de fingerspelling vs escritura en smartphone

- La elevada velocidad del fingerspelling, junto con la falta de alternativas tecnológicas accesibles, evidencia una oportunidad crucial para desarrollar sistemas de reconocimiento que permitan a personas sordas o con discapacidad auditiva ingresar texto de manera más rápida y eficiente.

- El impacto educativo, social y económico de la pérdida auditiva es enorme. Sin embargo, el uso de tecnologías de señalización visual - como el fingerspelling - puede ayudar a cerrar esta brecha si se impulsa correctamente.

Componentes lingüísticos del ASL

a. **Forma de la mano (Handshape):** la configuración de los dedos y la palma al realizar una seña (Jones, 2023).

b. **Ubicación (Location):** el lugar en el espacio donde se realiza la seña, que puede incluir áreas como la cabeza, el torso o la mano dominante (Jones, 2023).

c. **Movimiento (Movement):** la dirección, velocidad y trayectoria del desplazamiento de la mano o las manos (Jones, 2023).

d. **Orientación de la palma (Palm Orientation):** la dirección en la que la palma de la mano está orientada durante la seña (Jones, 2023).

e. **Señas no manuales (Non-manual signals):** expresiones faciales y movimientos corporales que complementan el significado de la seña (Jones, 2023).

Uso del cuerpo en ASL

El cuerpo entero es esencial en la comunicación en ASL. Además de manos, se utilizan:

a. **Expresiones faciales:** indican emociones, preguntas, negaciones y otras intenciones comunicativas (Strickland, 2021).

b. **Movimientos de la cabeza y torso:** complementan la información y pueden modificar el significado de una seña (Strickland, 2021).

c. **Espacio tridimensional:** el uso del espacio alrededor del hablante es fundamental para indicar relaciones espaciales y temporalidad (Strickland, 2021).

Clasificadores en ASL

Los clasificadores son formas especiales de seña que representan categorías de objetos o personas, y su forma y movimiento varían según el tipo de objeto o la acción que se describe.

Por ejemplo, un clasificador puede representar una persona caminando en un vehículo en movimiento (Strickland, 2021).

3. Investigación sobre tecnologías a utilizar

Métodos actuales de reconocimiento automático

Así como en medicina el diagnóstico por imágenes depende de detectar patrones complejos en radiografías o resonancias, en el caso del fingerspelling en ASL se requiere reconocer movimientos rápidos y cambios sutiles en la posición de la mano. Con el avance de la tecnología se han desarrollado varias maneras distintas de hacer esto:

- **Modelos tradicionales:** usaban técnicas basadas en extracción de características (contornos, HOG, SIFT) y clasificadores como SVM.

- **Modelos modernos:** actualmente se usan redes neuronales profundas, en particular CNNs (Convolutional Neural Networks) y transformers para visión (Vision Transformers, ViTs), que permiten identificar secuencias de imágenes con mayor precisión.

- **Velocidad de comunicación:** investigaciones muestran que el fingerspelling puede alcanzar 57 palabras por minuto, mientras que escribir en un teclado en pantalla promedio llega a 36 palabras por minuto (TensorFlow, 2023). Esto resalta el potencial de un sistema automático para superar barreras tecnológicas.

El dataset

El reconocimiento automático del deletreo manual en ASL requiere modelos capaces de manejar secuencias espaciales y temporales complejas. Estudios recientes han demostrado que los landmarks 3D de manos, rostro y postura, extraídos mediante bibliotecas como MediaPipe, ofrecen una representación compacta y eficiente para el modelado de señas (Zhang et al, 2023).

El data de la competencia es particularmente valioso porque combina:

- **Gran escala:** más de 3 millones de caracteres, lo que lo convierte en el mayor dataset público de fingerspelling hasta la fecha (Google Research, 2023).

- **Diversidad de participantes:** más de 100 personas sordas, con variaciones en tonos de piel, velocidad y estilo de ejecución de las señas.

- **Variabilidad de condiciones:** grabaciones en diferentes fondos, ángulos y niveles de iluminación, lo que aporta realismo pero incrementa la complejidad del aprendizaje.

Técnicas

Estos datos permiten explorar técnicas de aprendizaje profundo para secuencias, como redes recurrentes (LSTM, GRU), arquitecturas basadas en transformers para series temporales o modelos híbridos (CNN + RNN), que han mostrado buenos resultados en tareas de reconocimiento gestual (Koller et al., 2019).

Además, la estructura de los archivos Parquet y JSON facilita un preprocesamiento eficiente a gran escala, pero plantea un reto de manejo de datos masivos (189 GB), lo que obliga a definir estrategias de carga parcial, muestreo o uso de servicios en la nube.

Librerías

Para el reconocimiento automático del ASL, es fundamental identificar con precisión las posiciones de las manos, los dedos y otras partes relevantes del cuerpo. Esto se logra mediante la detección de landmarks, que consiste en localizar puntos clave (keypoints) en la mano que representan articulaciones, extremos de los dedos y otras referencias anatómicas.

1. **Landmarks de la mano**

   Cada mano se representa con 21 landmarks distribuidos en la palma y dedos:

   - 1. Muñeca
   - 2-5. Dedos pulgar (base, articulaciones, punta)
   - 6-9. Dedo medio
   - 14-17. Dedo anular
   - 18-21. Dedo meñique

   Cada landmark tiene coordenadas x, y, z, donde x e y representan la posición relativa en la imagen y z representa la profundidad o distancia desde la cámara. Estos landmarks permiten capturar con detalle la forma, orientación y movimiento de la mano durante la ejecución de una seña (Google, s.f.).

2. **MediaPipe (Google)**

   MediaPipe es una biblioteca desarrollada por Googleque ofrece pipelines de visión por computadora en tiempo real. Para ASL, MediaPipe proporciona (Google, s.f.):

   - **Hand tracking:** detección de manos en imágenes o videos y estimación de los 21 landmarks de manera rápida y eficiente.

   - **Face Mesh y Pose:** permite complementar la detección de las manos con landmarks faciales y de postura, importantes para captar expresiones no manuales.

   Ventajas para ASL:

- Procesamiento en tiempo real, adecuado para aplicaciones móviles.
- Funciona con videos de baja resolución y condiciones variadas de iluminación
- Los landmarks extraídos son invariantes a la escala y rotación, facilitando el entrenamiento de modelos de aprendizaje profundo.

Aplicación al proyecto: en el dataset de la competencia, los landmarks ya están preprocesados con MediaPipe, lo que permite:

- Reducir el tamaño de los datos frente a trabajar con imágenes completas.
- Enfocar los modelos en secuencia de coordenadas espaciales en lugar de pixeles, optimizando memoria y tiempo de entrenamiento.
- Capturar información estática (forma de la mano) como dinámica (movimiento entre frames), crucial para detectar letras y palabras en fingerspelling.

4. Análisis inicial del problema y los datos disponibles.

➔ Descripción del problema

El objetivo principal de esta competencia es desarrollar un modelo capaz de detectar y traducir el deletreo manual del lenguaje de señas estadounidense (ASL) a texto. Este reto representa un avance significativo en el reconocimiento de señas, con el potencial de hacer la inteligencia artificial más accesible para personas sordas o con dificultades auditivas.

➔ Descripción del conjunto de datos

El dataset contiene más de tres millones de caracteres deletreados por más de 100 personas sordas, capturados mediante la cámara frontal de un teléfono inteligente en condiciones variadas de iluminación y fondo. Los datos incluyen secuencias de coordenadas espaciales extraídas de videos, que detecta landmarks en la cara, manos y postura corporal.

➔ Variables disponibles y descripción del dataset

◆ El conjunto de datos está compuesto por 124 archivos de formatos Parquet, CSV y JSON, con un tamaño total de 189.09 GB. Esto incluye archivos de entrenamiento, metadata suplementaria y los landmarks extraídos. Las variables principales son:

- Path: ruta al archivo de landmarks.
- File_id: identificador único del archivo.
- Participant_id: identificador del contribuyente de datos.
- Sequence_id: identificador único de cada secuencia de landmarks.
- Phrase: etiqueta textual (frase) asociada a la secuencia.
- Frame: número de cuadro dentro de la secuencia.

- x/y/z_[type]_[landmark_index]: coordenadas espaciales para cada uno de los 543 landmarks, clasificados por tipo (face, left_hand, pose, right_hand).
- Character_to_predictions_index.jason: mapeo entre caracteres y sus índices de predicción.

➔ Observaciones iniciales
Este dataset ofrece una oportunidad única para explorar el reconocimiento de señas a un nivel detallado. La diversidad de participantes y condiciones de grabación añade robustez, pero también plantea desafíos como la variabilidad en la visibilidad de las manos y la calidad de los frames. Además, dado que los landmarks no deben usarse para identificar personas, el enfoque debe centrarse exclusivamente en el reconocimiento gestual.

5. Preprocesamiento de datos

Para el preprocesamiento se implementó un script de Python que permite cargar y organizar las secuencias de landmarks. Se definieron funciones auxiliares:

- Load_data: carga los archivos de datos en formato parquet y garantiza que cada registro contenga un identificador de secuencia (sequence_id).
- Pick_sequence: selecciona una secuencia específica a partir de su identificador facilitando la exploración y verificación de frames.

Con este procedimiento se trabajó directamente con los archivos de landmarks, verificando que las secuencias se encontrarán completas y correctamente indexadas. De esta forma se estableció la base para poder realizar el análisis exploratorio.

6. Análisis exploratorio

Para el análisis exploratorio se realizó un conjunto de gráficos que se puede ver en el archivo .ipynb adjunto.

7. Conclusiones

1. **Viabilidad del reconocimiento:** Las visualizaciones 2D/3D con conexiones de MediaPipe muestran que los landmarks de manos, rostro y postura representan bien las trayectorias del fingerspelling, lo que respalda la factibilidad de un modelo de traducción ASL a texto.
2. **Diagnóstico de calidad de datos:** El conteo de NaN y la inspección por secuencia evidencian la presencia de frames incompletos y variabilidad entre participantes.
3. **Selección de mano:** El criterio automático para elegir la mano más visible mejora la consistencia de las secuencias usadas en las animaciones y futuros experimentos.
4. **Variabilidad y reto:** Se observan diferencias de rango, velocidad y visibilidad entre secuencias; esto anticipa la necesidad de estrategias de regularización/aumentación en el entrenamiento.

5. **Escalabilidad:** El tamaño del dataset obliga a planificar lectura incremental y preprocesamiento eficiente para no exceder memoria en etapas anteriores.

8. Sugerencias para fases

1. **Implementación de batch processing:** Implementar lectura por lotes de .parquet y construcción de tensores por sequence_id con padding/truncamiento, evitando cargar todo en memoria.
2. **Limpieza y normalización:** Formalizar reglas (mínimo de frames, descarte de filas con Nan críticos) y aplicar normalización/estandarización consistente (por participante o global).
3. Reducción de dimensionalidad: Mantener principalmente landmarks de manos y un pequeño set de referencia (cara/pose) para reducir el número de features.
4. Métricas y evaluación: Definir métricas a nivel de carácter o frase (accuracy) y análisis por participante/condición para detectar sesgos.
5. Modelado y despliegue: Probar LSTM/GRU/Transformers con regularización (dropout, early stopping).

Referencias

Redacción. (2023, 24 septiembre). *Sordera: 4 datos que quizás no conocías*. BBC News Mundo. https://www.bbc.com/mundo/articles/cglelv87zgmo#:~:text=%C2%BFCu%C3%A1ntos%20sordos%20hay%20en%20el,cifra%20superar%C3%A1%20los%20700%20millones

Google. (s. f.). *American Sign Language Fingerspelling Recognition*. Kaggle. https://www.kaggle.com/competitions/asl-fingerspelling

World Health Organization (WHO). (2025, 26 febrero). *Deafness and hearing loss*. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

Population | Documentation. (s. f.). https://sign.mt/docs/docs/facts/population.html

Google Research. (2023). *ASL Fingerspelling Recognition Competition*. Kaggle. https://www.kaggle.com/competitions/asl-fingerspelling

Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2019). Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. *Computer Vision and Image Understanding, 141*, 108–125. https://doi.org/10.1016/j.cviu.2015.09.013

Hand landmarks detection guide. (s. f.). Google AI For Developers. https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker

MediaPipe Solutions guide. (s. f.). Google AI For Developers. https://ai.google.dev/edge/mediapipe/solutions/guide

Zhang, Y., Kim, J., & Shi, B. (2023). Landmark-based sign language recognition with deep temporal models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(6), 7254–7268. https://doi.org/10.1109/TPAMI.2023.3245678

Jones, T. (2023, 8 mayo). 5 Parameters of ASL: Sign Language 101. Learn Bright. https://learnbright.org/5-parameters-of-asl-sign-language-101/

Strickland, J. (2021, 15 abril). How sign language works. HowStuffWorks. https://people.howstuffworks.com/sign-language.htm