ISSN 1870-4069

ELT: Transformadores para la comprensión de la lengua de señas mexicana a través del preentrenamiento de puntos de referencia en imágenes

Víctor Martínez-Sánchez[1], Iván Villalón-Turrubiates[1], Francisco Cervantes-Álvarez[1], Carlos Hernández-Mejía[2], Delia Torres-Muñoz[3]

[1] Instituto Tecnológico de Estudios Superiores de Occidente, Guadalajara, México

[2] Tecnológico Nacional de México, Instituto Tecnológico Superior de Misantla, México

[3] Instituto Tecnológico Superior de San Martín Texmelucan, México

{ng683728, villalon, fcervantes}@iteso.mx, {cmahernandez, deletsm}@gmail.com

Resumen. Un nuevo modelo de representación de la Lengua de Señas Mexicana llamado Encoded Landmarks from Transformers (ELT) es introducido. ETL está basado en codificadores bidireccionales preentrenados usando puntos de referencia en imágenes no etiquetadas mediante un vector enmascarado en todas las capas. En consecuencia, el modelo ETL preentrenado puede ajustarse finamente con solo una capa de salida adicional con el propósito de generar modelos de clasificación y subtitulación de secuenciación de imágenes. El rendimiento es evaluado mediante el conjunto de datos MX-ITESO-100. ELT evidencia una ganancia de precisión real del 3 % comparado contra un modelo tradicional Long Short-Term Memory (LSTM) bidireccional. Adicionalmente, el modelo ELT reduce el tiempo de entrenamiento hasta en un 28 % y permite realizar un ajuste fino en un tiempo altamente competitivo.

Palabras clave: ETL, transformadores, lengua de señas mexicana.

ELT: Transformers for the Understanding of Mexican Sign Language through Pre-training of Reference Points in Images

Abstract. A new representation model of Mexican Sign Language called Encoded Landmarks from Transformers (ELT) is introduced. ETL is based on pre-trained bidirectional encoders using landmarks on unlabeled images using a masked vector on all layers. Consequently, the pre-trained ETL model can be fine-tuned with only one additional output layer for the purpose of generating image sequencing classification and captioning models. The performance is

evaluated using the MX-ITESO-100 data set. ELT shows a real precision gain of 3% compared to a traditional bidirectional Long Short-Term Memory (LSTM) model. Additionally, the ELT model reduces training time by up to 28% and allows fine tuning to be carried out in a highly competitive time.

Keywords: ETL, transformers, Mexican sign language.

1. Introducción

La Lengua de Señas Mexicana (LSM) es la lengua de señas utilizada por la comunidad sorda en México. Ha sido oficialmente reconocida como un idioma en nuestro país desde 2003, según la Ley General para la Inclusión de Personas con Discapacidad. La LSM sigue reglas particulares gramaticales y léxicas; siendo una herramienta crucial para la comunicación entre la población sorda en México. Por esta razón, es imperativo contar con tecnología que facilite la integración de las personas sordas en la sociedad mexicana.

Los nuevos enfoques basados en redes de transformadores prometen resultados favorables para el procesamiento del lenguaje natural y por lo tanto vincular la tendencia de desarrollo en transformadores con los avances en la LSM. Esta investigación propone vectorizar la representación de un gesto o seña utilizando puntos de referencia de una o ambas manos como un componente fundamental de la propuesta. De manera paralela, el uso del modelo preentrenado ETL permite agilizar el entrenamiento de modelos para otras investigaciones a través del ajuste fino.

2. Trabajos relacionados

El concepto de representación vectorial para unidades lingüísticas, específicamente palabras o lexemas, ha sido un tema destacado en el Natural Language Processing (NLP). La utilización de vectores continuos permite capturar matices semánticas y relaciones dentro del espacio lingüístico. El algoritrmo word2vec [11] ha demostrado eficacia en la extracción de características; logrando que las incrustaciones de palabras capturen relaciones semánticas entre palabras. Matthew E. [12] presenta un novedoso modelo de representación de palabras llamado Embeddings from Language Models (ELMo). ELMo extiende la idea mediante la introducción de incrustaciones contextualizadas que permiten una representación más matizada de las palabras basada en el contexto en donde aparecen.

De acuerdo a los avances más reciente en NLP, existe en la literatura dos conceptos fundamentales: modelos preentrenados [14] y la arquitectura de transformadores [16]. Alec Radford [13] propone un enfoque innovador conocido como Generative Pre-training (GPT) en donde es posible destacar que la fase de preentrenamiento capacita al modelo con una comprensión integral del lenguaje y por lo tanto permite capturar información contextual y adquirir conocimiento de estructuras sintácticas y semánticas. De forma semejante a el modelo de lenguaje GPT, Jacob Develin [5] introduce un modelo llamado Bidirectional Encoder Representations from

Transformers (BERT) basado en la arquitectura de transformadores y preentrenamiento en un extenso conjunto de datos. El descubrimiento clave de BERT radica en la comprensión contextual bidireccional a diferencia de los modelos anteriores que procesan texto unidireccionalmente.

BERT considera tanto el contexto izquierdo como el derecho de cada palabra en una oración y por lo tanto recolecta información contextual más significativa. El proceso de preentrenamiento permite el aprendizaje de representaciones de palabras contextualizadas. Esta investigación destaca la eficacia del preentrenamiento en diversas tareas mediante el ajuste fino del modelo BERT preentrenado en tareas específicas posteriores.

Conjuntamente a lo anterior, existen modelos adicionales construidos sobre el modelo BERT y centrados en el preentrenamiento optimizado. Tal es el caso de Robustly Optimized BERT Pretraining Approach (RoBERTa) [7] cuyo objetivo principal es mejorar la metodología de preentrenamiento de BERT y afrontar las limitaciones. Los investigadores también enfatizan la importancia de entrenar con secuencias más largas y utilizar conjuntos de datos más grandes para el preentrenamiento.

Los mecanismos de atención en las redes de transformadores pueden ser extendidas hacia imágenes estáticas, como evidencian las arquitecturas de modelos de visión preentrenados como ViT [6], BEiT [2] y Swin [8]. Las investigaciones afirman que los transformadores pueden capturar dependencias y relaciones globales dentro de una imagen y por lo tanto son convenientes para el reconocimiento de imágenes a gran escala. Además, los transformadores también pueden ser integrados en el campo de la visión por computadora para modelar datos de video. Javier Selva [15] explora el uso de grandes redes neuronales convolucionales (Convolutional Neural Networks, CNNs) como la base de Vision Transformers (VT); aprovechando los sesgos inductivos y capacidades de reducción de dimensionalidad.

Esta investigación también destaca la explotación explícita de la estructura del video a través de la tokenización. Sin embargo, a pesar de que existen modelos para introducir estrategias específicas en el desarrollo de interacciones espacio-temporales detalladas; aún es necesario la exploración de estrategias para la LSM cuya orientación principal sea el movimiento relativo de las manos mas allá de las características de la imagen. Necati C. [4] presenta una metodología innovadora denominada Sign Language Translation Transformer (SLTT).

Esta metodología incorpora una capa de incrustación espacial y utiliza CNNs para el procesamiento de imágenes. Sin embargo, este enfoque depende de un conjunto de datos completo para lograr un rendimiento robusto y que las CNNs usadas en el modelo puedan extraer información espacial de toda la imagen ocasionando dependencias inherentes en el contexto ambiental. Por lo tanto, hemos considerado las investigaciones de Necati como punto de partida fundamental para nuestra propuesta de extender los articuladores de señas a través de puntos de referencia en un modelo preentrenado que pueda ser ajustado de manera fina.

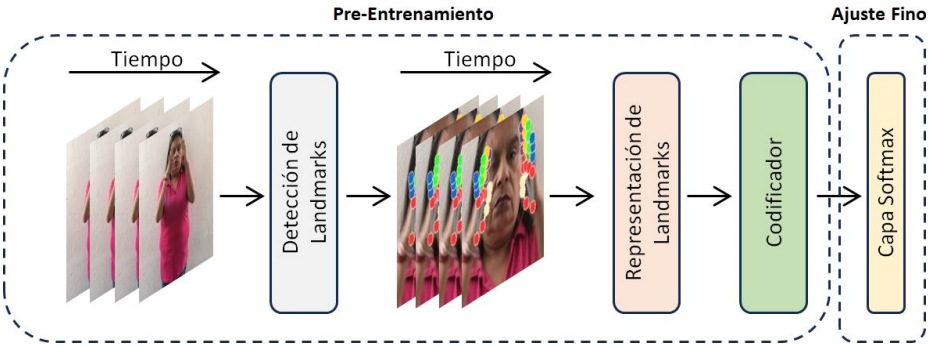

Fig. 1. Estructura general de la estrategia ELT.

3. ELT

El modelo ELT propone una estructura general en donde la representación de señas dinámicas es llevada a cabo dentro de una secuencia de imágenes descritas como una secuencia de puntos de referencia de mano en un espacio vectorial que alimenta a un codificador de las redes de transformadores, como es mostrada en la Figura 1. En conjunto con mecanismos de atención y codificación posicional, posee la capacidad de adquirir los matices cinemáticos que constituyen gestos del léxico de la LSM.

3.1. Puntos de referencia embebidos

Una representación exhaustiva de puntos de referencia en las manos es expuesta de manera profunda en [9]. Este modelo extrae 21 puntos clave de ambas manos. Cada punto clave consta de tres coordenadas espaciales $\{x,\ y,\ z\}$. Las coordenadas $\{x,\ y\}$ se normalizan al rango $\{0{,}0,\ 1{,}0\}$ en relación con la imagen. La coordenada $\{z\}$ es normalizada con la misma magnitud que la coordenada $\{x\}$ para generalizar y detectar correctamente los puntos en una variedad de situaciones y condiciones. Los 21 puntos extraídos de la mano forman un lexema.

Un lexema puede expresarse como un vector continuo en un espacio dimensional en donde cada elemento posee características distintas. Los algoritmos como word2vec emplean un enfoque estructurado para la extracción de características. Nuestro enfoque conlleva el aprovechamiento de los detalles intrincados de gestos de las manos en lugar de la dependencia de características lingüísticas tradicionales. De manera particular, son considerados 21 puntos clave asociados con los movimientos de las manos y, adicionalmente, es construido un nuevo vector, denominado E_{lado} en (1), a través de la concatenación de estos puntos clave. El vector resultante es caracterizado por una dimensionalidad de 63:

$$E_{\text{lado}} = \bigcup_{i=0}^{20} [x_i, y_i, z_i]. \tag{1}$$

La extracción de puntos clave de la segunda mano mejora la integralidad de la representación de características.

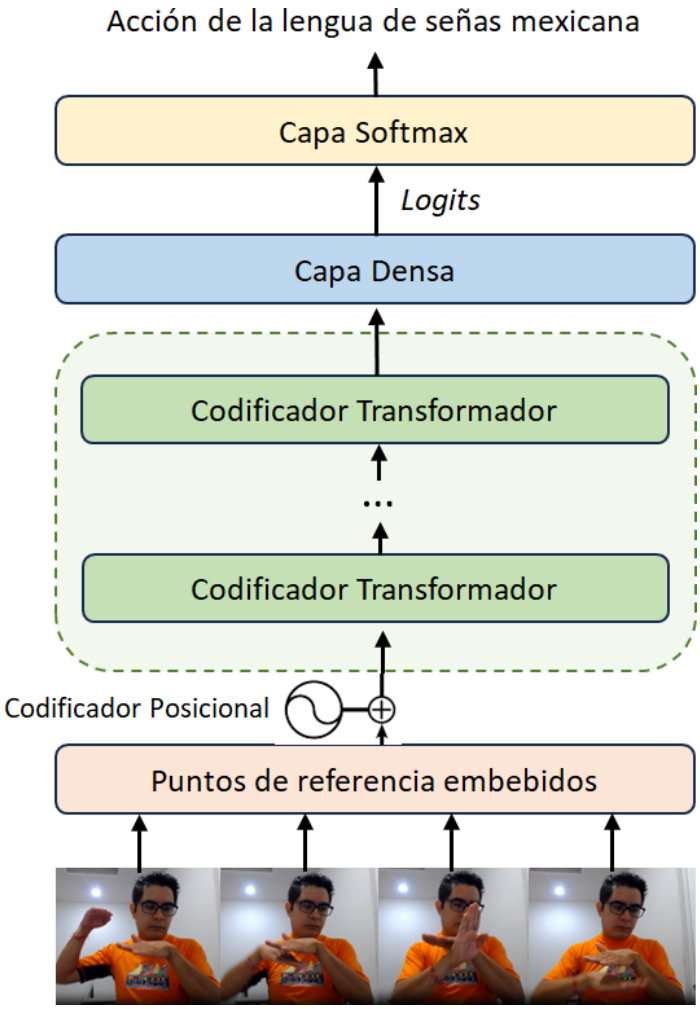

Fig. 2. Arquitectura general del modelo ELT.

Este proceso produce un vector de características dimensionales de 126 y por lo tanto engloba un conjunto rico de atributos espaciales y temporales. La ampliación de información a través de este enfoque de mano izquierda E_{I} y mano derecha E_{D} contribuye en la robustez y la precisión del sistema general de reconocimiento de LSM, como es mostado en (2):

$$E = [E_{\mathrm{I}},\ E_{\mathrm{D}}]. \tag{2}$$

Es importante mencionar que, en el caso de señas llevadas a cabo con una sola mano, el algoritmo rellena el vector E con 63 valores de 10^{-9} cada uno.

3.2. Arquitectura del modelo

El modelo ETL, inspirado en el trabajo fundamental de Vaswani [17] sobre las redes de transformadores, sigue una estructura semejante al modelo original del codificador. De manera notable, la capa Input Embedding en [17] es reemplazada con la capa de Puntos de referencia embebidos, como es mostrado en la Figura 2, ocasionando un cambio hacia las entradas de puntos de referencia en imágenes. En la configuración propuesta, especificamos parámetros para mejorar las capacidades del modelo.

Esto incluye un codificador de 12 capas ($L = 12$), 14 cabezas de autoatención ($A = 14$) para mecanismos de atención robustos, y una dimensión de capa oculta de 126 ($H = 126$). La elección de la dimensión L está alineada deliberadamente con el modelo ampliamente utilizado en [5] y por lo tanto promueve la consistencia y la comparabilidad en las métricas de rendimiento. Debido a que cada vector E contiene la representación de las manos en un momento específico de una secuencia S de imágenes, la secuencia completa para un gesto de lenguaje de señas debe estar compuesta por n vectores E_n, como es mostrado en (3), en donde el número máximo de vectores está establecido por el valor de longitud de secuencia:

$$S = [E_0,\ E_1,\ E_2,\ \ \ldots,\ E_n]. \tag{3}$$

3.3. Etapa de preentrenamiento

Durante la etapa de preentrenamiento, el modelo utiliza 60 vectores para representar el movimiento de una secuencia de señas. Habitualmente, una seña tiene una duración mínima y máxima entre 1 y 2 segundos. En el caso de suponer una duración promedio de 30 cuadros por segundo para un video, el proceso completo consume un total de 60 cuadros y por lo tanto corresponde con la secuencia propuesta. En una situación donde la duración de la seña es inferior a dos segundos, la secuencia es completada mediante un vector de relleno especial [PAD] para asegurar una dimensionalidad consistente en la secuencia S. El valor para cada elemento en el vector [PAD] es 10^{-9} [17].

El entrenamiento del modelo ELT es llevado a cabo con la estrategia propuesta por el modelo BERT, la cual está relacionada con tareas no supervisadas. De manera similar, un vector especial [MASK] es utilizado para enmascarar de forma aleatoria un vector E para una secuencia de entrada dada S. Por lo tanto, el proceso de entrenamiento implica predecir el vector E enmascarado usando el contexto de la secuencia. El valor propuesto para cada elemento en el vector [MASK] es 0. Es importante mencionar que, el modelo ELT mantiene la misma dimensión de entrada a diferencia de BERT que conecta la salida a una capa lineal de tamaño de vocabulario para generar logits.

3.4. Ajuste fino para ELT

El proceso de ajuste fino para el modelo ELT es llevado a cabo de manera fluida mediante la incorporación de un clasificador que consta de una capa Softmax, capas lineales adicionales y una función de activación Tanh. Esta estrategia aumenta la adaptabilidad del modelo hacia tareas específicas. La capa lineal, con una dimensionalidad de 2048, está alineada con las recomendaciones presentadas por las redes de avance de posición en el contexto de los transformadores.

Tabla 1. Distribución de los elementos gramaticales en MX-ITESO-100.

Elementos gramaticales	Cantidad
Verbos	30
Adjectivos	29
Sustantivos	25
Adverbios	6
Pronombres	5
Frases	4
Conjunciones	1

Con respecto a la alimentación del clasificador, solo es considerado el primer vector E_0 de la secuencia de salida S. Es importante establecer que, los hiper parámetros para el ajuste fino están estrechamente alineados con los usados en la etapa de preentrenamiento y por lo tanto mantienen la consistencia en la configuración general del modelo. El único ajuste cuantitativo tiene lugar en la tasa de aprendizaje del optimizador Adam [3], de un valor de tasa de 10^{-2} hacia un valor de 10^{-4}, debido a una actualización en el proceso de convergencia. Finalmente, el costo de entrenamiento durante el ajuste fino es significativamente menor en comparación con el de la etapa de preentrenamiento. Esto es debido en parte a que el modelo ha aprendido, a lo largo del proceso de preentrenamiento, la posición relativa de los puntos de referencia en la imagen con respecto a una secuencia.

3.5. Conjunto de datos

El conjunto de datos MX-ITESO-100 [10] consta de 5000 instancias de video para representar los 100 elementos más importantes del léxico mexicano. A diferencia de compilaciones similares, MX-ITESO-100 prioriza un léxico mexicano amplio y heterogéneo, incluyendo elementos gramaticales esenciales. Cada videograbación engloba una noción conceptual transmitida a través de gestos dinámicos; los cuales abarcan dos etapas secuenciales que van desde configuraciones preliminares hasta configuraciones conclusivas. La distribución de los elementos gramaticales son mostrados en la Tabla 1. Con las videograbaciones previamente descritas, los puntos de referencia de las manos de cada fotograma son extraidos usando la biblioteca MediaPipe [9]. El preprocesamiento de los fotogramas conlleva seleccionar puntos clave dentro de una región central; excluyendo todos los puntos que permanecen afuera de la siguiente región normalizada:

$$1 - T \geq \{x, y\} \geq T, \tag{4}$$

donde T representa un parámetro de umbral. De acuerdo a la expresión en (4) $T = 0{,}25$. Además, [1.0] denota la máxima relación ancho/alto del fotograma. Es importante mencionar que, si no se identifican los puntos de referencia en las manos, el proceso continua con el siguiente fotograma.

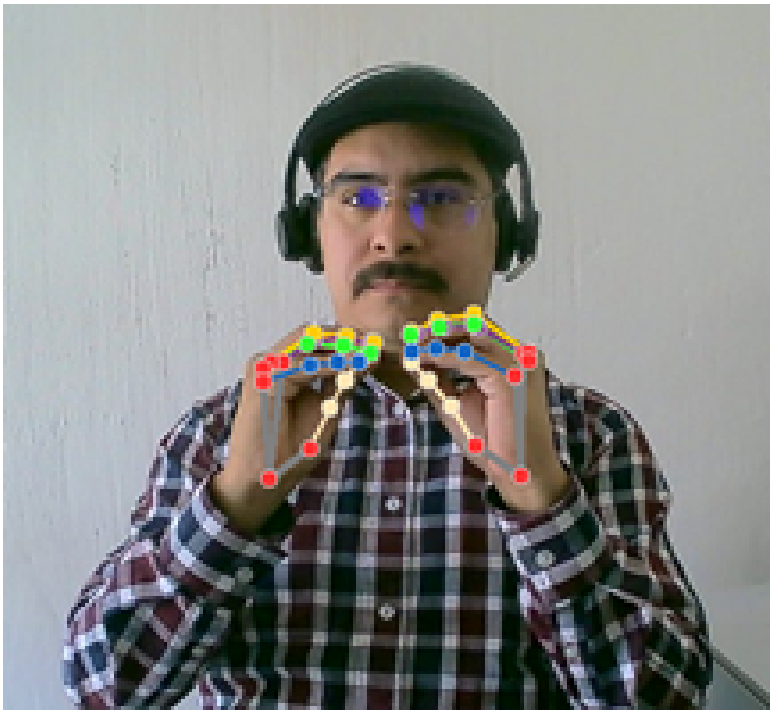

Fig. 3. Reconocimiento de la seña BESAR usando ELT.

4. Experimentación

Durante la etapa experimental, esta investigación usa el conjunto de datos MX-ITESO-100 que consta de 5000 videograbaciones categorizados en 100 elementos gramaticales con 50 videograbaciones para cada elemento gramatical. La arquitectura de modelo consta de 12 capas de codificador, cada una equipada con 14 cabezas de atención. La tasa de abandono (dropout) para cada capa ha sido establecida en 10 % con el propósito de prevenir el sobreajuste. El algoritmo de optimización es Adam con una tasa de aprendizaje de 10^{-2}, $\beta_1 = 0{,}9$ y $\beta_2 = 0{,}98$.

La fase de preentrenamiento incluye 500 épocas mientras que el ajuste fino es llevado a cabo durante 100 épocas. Con respecto a el ajuste fino del clasificador, usamos una función de activación tangente hiperbólica después de una capa lineal con una dimensionalidad de 2048. La función de activación muestra una mejor sensibilidad en valores negativos y por lo tanto el impacto en el sesgo de la red es reducido. Para el optimizador, la tasa de aprendizaje ha sido ajustada en 10^{-4}.

Posteriormente, la secuencia de capas continua con una capa de dropout con valor de 10 %, una capa lineal con 100 neuronas de salida y una capa Softmax para la clasificación final. La Figura 3 muestra la predicción de una seña con el modelo ELT. Los procedimientos de entrenamiento son ejecutados en un sistema Intel XEON habilitado con 128 núcleos, 1TB de memoria RAM y un motor de aceleración Tile Matrix Multiply (TMUL) [1].

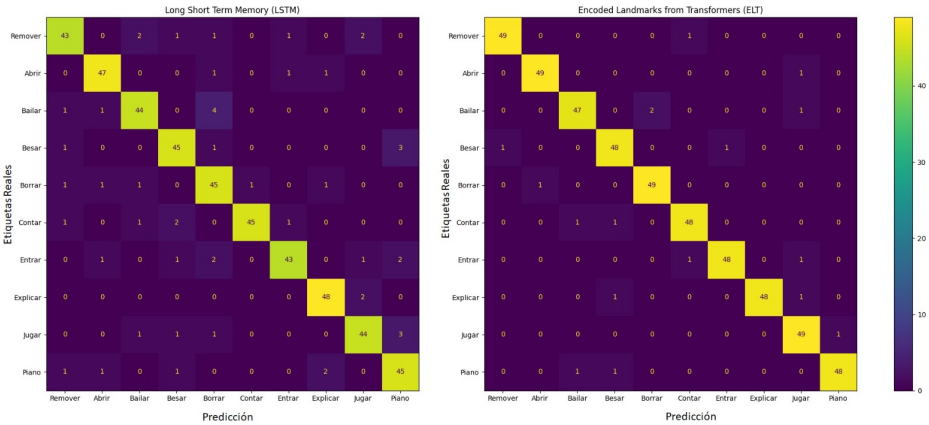

Fig. 4. Matrices de confusión para ETL y LSTM.

La duración del entrenamiento es aproximadamente de 1.2 minutos por época durante la fase de preentrenamiento y en total alrededor de 10 horas. El ajuste fino para 10 gestos exhibe una duración notablemente reducida de 0.13 segundos por época con un tiempo de entrenamiento total aproximado de 2.5 minutos. Adicionalmente, un modelo LSTM bidireccional ha sido entrenado para realizar una comparación de rendimiento. Al igual que el modelo ELT, el modelo LSTM es alimentado con los puntos de referencia extraídos del conjunto de datos MX-ITESO-100. Además, capas TimeDistributed son añadidas a la red LSTM para procesar cada muestra de manera individual. La duración del entrenamiento es aproximadamente de 2.8 minutos por época durante 300 épocas.

Para este caso no hubo un ajuste fino. Por lo tanto, el tiempo de entrenamiento ha sido de 14 horas aproximadamente. La Figura 4 muestra las matrices de confusión producidas por los modelos ETL y LSTM. Para ambos casos, son mostrados únicamente 10 señas relacionadas en su mayoría con verbos. En el caso del modelo clásico LSTM bidireccional, ha sido entrenado únicamente para estas 10 señas ya que no es posible llevar a cabo un ajuste fino. Con relación a la precisión de clasificación, el modelo ELT alcanza un valor de 0.9816 mientras que la precisión para el modelo bidireccional LSTM es de 0.9482. Finalmente, de acuerdo con los resultados experimentales, es posible establecer que el modelo ELT es 28 % mas rápido que el sistema tradicional LSTM durante el proceso de entrenamiento.

5. Conclusiones

En esta investigación, hemos introducido el modelo ELT basado en puntos de referencia de manos usando mecanismos de atención integrados en redes de transformadores. Esta estrategia establece un fundamento sólido para la futura comprensión de la lengua de señas en cualquier contexto lingüístico. El modelo preentrenado puede ser utilizado para la clasificación de señas dinámicas con un costo computacional relativamente modesto y una precisión significativamente elevada por encima del 97 % para un léxico compuesto de 100 elementos gramaticales. Los

resultados experimentales establecen un precedente en la lengua de señas mexicana para contribuir a la mejora en la comunicación de millones de personas sordas y en la integración hacia las actividades sociales rutinarias. Finalmente, el trabajo futuro está relacionado con la representación de puntos de referencia para las expresiones del rostro y la postura del cuerpo. Además, es necesario generar un modelo preentrenado con un léxico más robusto y regionalismos típicos.

Referencias

1. Bal, S., Mummidi, C.S., Da-Cruz-Ferreira, V., Srinivasan, S., Kundu, S.: A novel fault-tolerant architecture for tiled matrix multiplication pp. 1–6 (2023)
2. Bao, H., Dong, L., Piao, S., Wei, F.: BEit: BERT pre-training of image transformers. In: International Conference on Learning Representations. pp. 1–18 (2022)
3. Bernico, M.: Deep learning quick reference: Useful hacks for training and optimizing deep neural networks with TensorFlow and Keras. Packt Publishing (2018)
4. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation pp. 10023–10033 (2020)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North. vol. 1, pp. 4171–4186 (2019)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16×16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. pp. 1–21 (2021)
7. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. In: International Conference on Learning Representations. pp. 1–15 (2019)
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision. pp. 9992–10002 (2021)
9. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: MediaPipe: A framework for perceiving and processing reality. In: 3rd Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition. pp. 1–4 (2019)
10. Martínez-Sánchez, V., Villalón-Turrubiates, I., Cervantes-Álvarez, F., Hernández-Mejía, C.: Exploring a novel mexican sign language lexicon video dataset. Multimodal Technologies and Interaction 7(8), 83 (2023)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations. pp. 1–12 (2013)
12. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. vol. 1, pp. 2227–2237. Association for Computational Linguistics (2018)
13. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018), paperswithcode.com/paper/improving-language-understanding-by
14. Rothman, D., Gulli, A.: Transformers for natural language processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more. Packt Publishing (2022)

15. Selva, J., Johansen, A.S., Escalera, S., Nasrollahi, K., Moeslund, T.B., Clapés, A.: Video transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(11), 12922–12943 (2023)
16. Tunstall, L., von-Werra, L., Wolf, T.: Natural language processing with transformers. O'Reilly Media (2022)
17. Wu, Y.: Attention is all you need for boosting graph convolutional neural network (2024)