Machine Learning
Agosto – Diciembre 2023
Dr. Luis Carlos Padierna García

Traductor inglés-español por reconocimiento de texto en imagen

Proyecto Final

Jiménez Pano Daniela, 427292, d.jimenezpano@ugto.mx

Calderón Reyes Miriam, 381390, mcalderonreyes@ugto.mx

Márquez Sulca Norma Angélica, 427278, na.marquezsulca@ugto.mx

Introducción

En la búsqueda continua de soluciones pragmáticas en el ámbito de la inteligencia artificial, presentamos un proyecto concreto que fusiona la visión computarizada y el procesamiento de lenguaje natural. Nuestro enfoque se centra en el desarrollo de un traductor automático que, a través de una Red Neuronal Convolucional (CNN), analiza imágenes para extraer texto, y posteriormente utiliza una Red Transformer para realizar traducciones precisas.

El uso de CNN permite identificar y aislar regiones de texto en imágenes, convirtiendo datos visuales en información legible. Este proceso sirve como punto de partida para la siguiente etapa, donde una Red Transformer se encarga de contextualizar y traducir el texto reconocido.

Esta combinación técnica no solo simplifica la experiencia del usuario al eliminar la necesidad de ingreso manual de texto, sino que también demuestra aplicaciones prácticas en entornos donde la comunicación multilingüe basada en imágenes es esencial.

A lo largo de este proyecto, analizaremos detalladamente la implementación de estos modelos, evaluaremos su rendimiento y exploraremos posibles mejoras para maximizar la eficacia de nuestro traductor.

Objetivo

Desarrollar un programa capaz de reconocer texto en imágenes mediante una red neuronal convolucional y, posteriormente, emplear técnicas de traducción para proporcionar la salida con el texto traducido de forma bidireccional español-inglés.

Justificación

La relevancia de realizar este proyecto es que se obtendrá una herramienta útil para la comprensión de contenido textual en imágenes, además, se ampliará el conocimiento obtenido de la materia, explorando un área que no ha sido tratada en clase.

En el escenario actual de interconexión global, donde la diversidad lingüística es una realidad, surge la necesidad de superar las barreras idiomáticas que pueden obstaculizar la comunicación eficaz. Este proyecto se fundamenta en la premisa de que la información, en sus diversas formas visuales, debe ser accesible y comprensible para todos, independientemente de las diferencias lingüísticas.

Machine Learning
Agosto – Diciembre 2023
Dr. Luis Carlos Padierna García

La propuesta de desarrollar un sistema de reconocimiento de texto en imágenes, integrado con una capacidad de traducción bidireccional entre inglés y español mediante CNN y Redes Transformers, responde directamente a esta necesidad. La justificación de este proyecto se cimienta en diversos beneficios que abarcan desde la inclusión y accesibilidad hasta la mejora de procesos empresariales y educativos.

En términos prácticos, la capacidad de nuestro sistema para analizar visualmente el texto y proporcionar traducciones precisas y contextualmente relevantes en tiempo real tiene un impacto significativo en la accesibilidad multilingüe. Elimina las barreras que podrían dificultar la comprensión de información clave, fomentando así una participación más amplia en la sociedad globalizada.

Para empresas que buscan operar a escala internacional, nuestro proyecto ofrece una herramienta estratégica para superar las limitaciones lingüísticas y facilitar la comunicación efectiva con diversos stakeholders. Además, en entornos educativos, el sistema presenta un potencial disruptivo al facilitar el aprendizaje de idiomas a través de la interacción con materiales visuales.

Marco Teórico

En el marco de nuestro proyecto final, nos enfocamos en la tarea crucial de reconocimiento de letras en imágenes y la posterior conversión de dichas letras a texto, con el objetivo de realizar traducciones. Para abordar esta tarea, hemos seleccionado la arquitectura de red neuronal convolucional (CNN) LeNet, que es considerada la primera red neuronal convolucional de la historia, además fue desarrollada con el objetivo de reconocer dígitos escritos a mano. (recordar poner cita)

En nuestro programa, se utilizó una variante de esta red neuronal, la cual contiene 7 capas, 2 de convolución, 2 de pooling, 2 capas totalmente conectadas y 1 de salida. La fase de captura de características en la arquitectura LeNet se realiza mediante las capas convolucionales, específicamente C1 y C3. Estas capas aplican convoluciones a la imagen de entrada, buscando activamente patrones locales como bordes y texturas. En este proceso, se extraen características clave que son representativas de las letras presentes en la imagen. La aplicación de funciones de activación no lineales en estas capas contribuye a la captura de información distintiva y la construcción de una representación significativa de la entrada.

Posteriormente, el submuestreo desempeña un papel esencial en las capas S2 y S4 de LeNet. Esta etapa de la red tiene como objetivo reducir la dimensionalidad de la representación sin sacrificar la información crítica relacionada con el reconocimiento de letras. A través del submuestreo, se preservan las características esenciales, manteniendo la capacidad de discernir letras mientras se disminuye la complejidad computacional del modelo.

La Capa Totalmente Conectada (C5) juega un papel crucial al agregar las características extraídas por las capas anteriores. Esta capa proporciona una representación más abstracta de la imagen, permitiendo la identificación de patrones más complejos. La conexión total entre las neuronas en esta capa facilita la combinación y comprensión global de las características, preparando la información para ser utilizada en la tarea final de reconocimiento de letras.

Por esto, la arquitectura LeNet sigue un flujo coherente desde la captura de características clave hasta la creación de representaciones más abstractas, todo lo cual es fundamental para su eficacia en el reconocimiento de letras en imágenes.

La elección de LeNet es fundamental para nuestro proyecto, ya que su capacidad para el reconocimiento de letras en imágenes se integra de manera directa en nuestra tarea de conversión de texto y traducción. La arquitectura modular y estratificada de LeNet facilita la adaptación a las particularidades de nuestro conjunto de datos, mejorando la precisión del reconocimiento de letras y, por ende, la calidad de las traducciones resultantes.

La traducción automática ha sido un desafío constante en el procesamiento del lenguaje natural, y los modelos Transformer han emergido como una solución destacada en este ámbito. En este contexto, el procesamiento eficiente de secuencias de texto y la capacidad para capturar relaciones complejas entre palabras los han posicionado como la herramienta fundamental para resolver el problema de la traducción automática en el campo de la inteligencia artificial.

El Transformer inicia su proceso con una capa de embedding, donde cada palabra se convierte en un vector en un espacio vectorial específico. Además de las palabras, se incorporan tokens especiales como `<START>`, `<END>`, y `<PAD>` para marcar el inicio y el final de la secuencia de traducción, así como para gestionar el relleno y garantizar la uniformidad en la longitud de las secuencias. Este proceso de embedding permite al modelo capturar las relaciones semánticas y sintácticas entre las palabras en la lengua fuente, proporcionando una representación numérica para su procesamiento posterior. La capa de input embedding actúa como la primera etapa del Transformer, preparando la entrada para la aplicación de mecanismos de atención y la generación subsiguiente de la secuencia de salida en el idioma de destino.

Enseguida se lleva a cabo la codificación posicional que asigna a cada palabra en la secuencia un vector posicional único, que se suma al vector de embedding de la palabra. Este enfoque permite al modelo distinguir entre palabras con significados similares pero ubicadas en diferentes posiciones dentro de la secuencia. La función sinusoidal utilizada para la codificación posicional garantiza que la relación entre los vectores posicionales conserve un orden relativo adecuado.

En el proceso de encoder, cada palabra se representa como un vector y pasa a través de múltiples capas de atención y feedforward. Esto permite al modelo capturar relaciones significativas entre los tokens de la secuencia de origen. El decoder, por otro lado, se encarga de generar la secuencia de salida en el idioma destino. Comienza con una capa de output embedding, donde cada palabra de la secuencia objetivo se transforma en un vector. Utilizando la información contextual proporcionada por el encoder, el decoder aplica mecanismos de atención para enfocarse en partes específicas de la entrada durante la generación de la salida. Este proceso es autorregresivo, lo que significa que el modelo predice cada palabra de la secuencia de salida una a la vez, utilizando las predicciones anteriores como contexto.

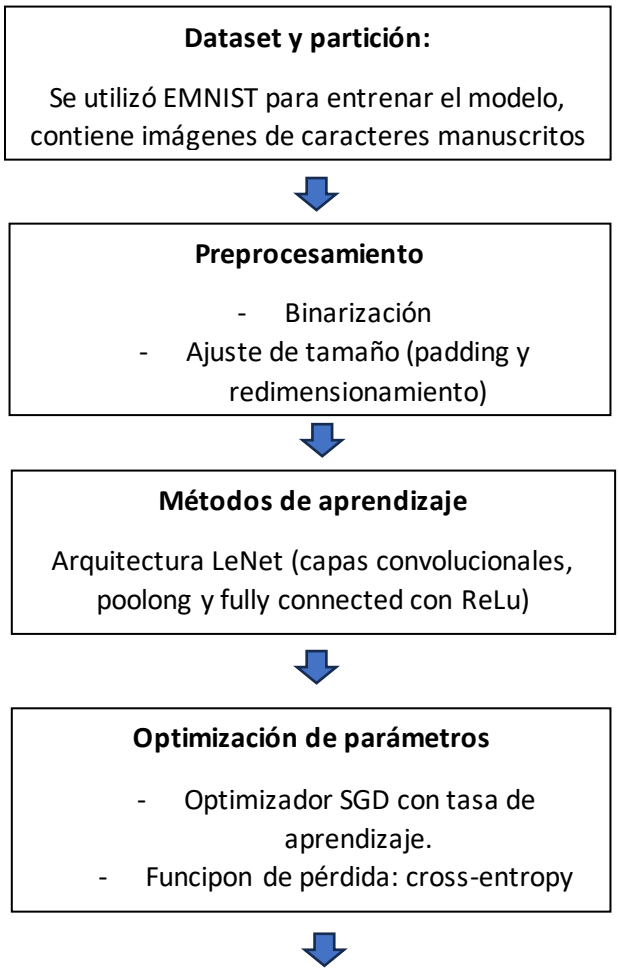
La capa de atención, tanto en el encoder como en el decoder, permite al modelo enfocarse en partes específicas de la secuencia de entrada durante el proceso de codificación y decodificación. Este mecanismo revolucionario mejora significativamente la capacidad del Transformer para capturar relaciones a largo plazo entre palabras. En el encoder, la capa de atención permite ponderar la importancia de cada palabra en relación con otras, generando representaciones contextuales ricas

para la secuencia de entrada. En el decoder, la capa de atención se utiliza durante la generación de la secuencia de salida, facilitando la alineación contextual entre la entrada y la salida y mejorando la coherencia y calidad de las traducciones.

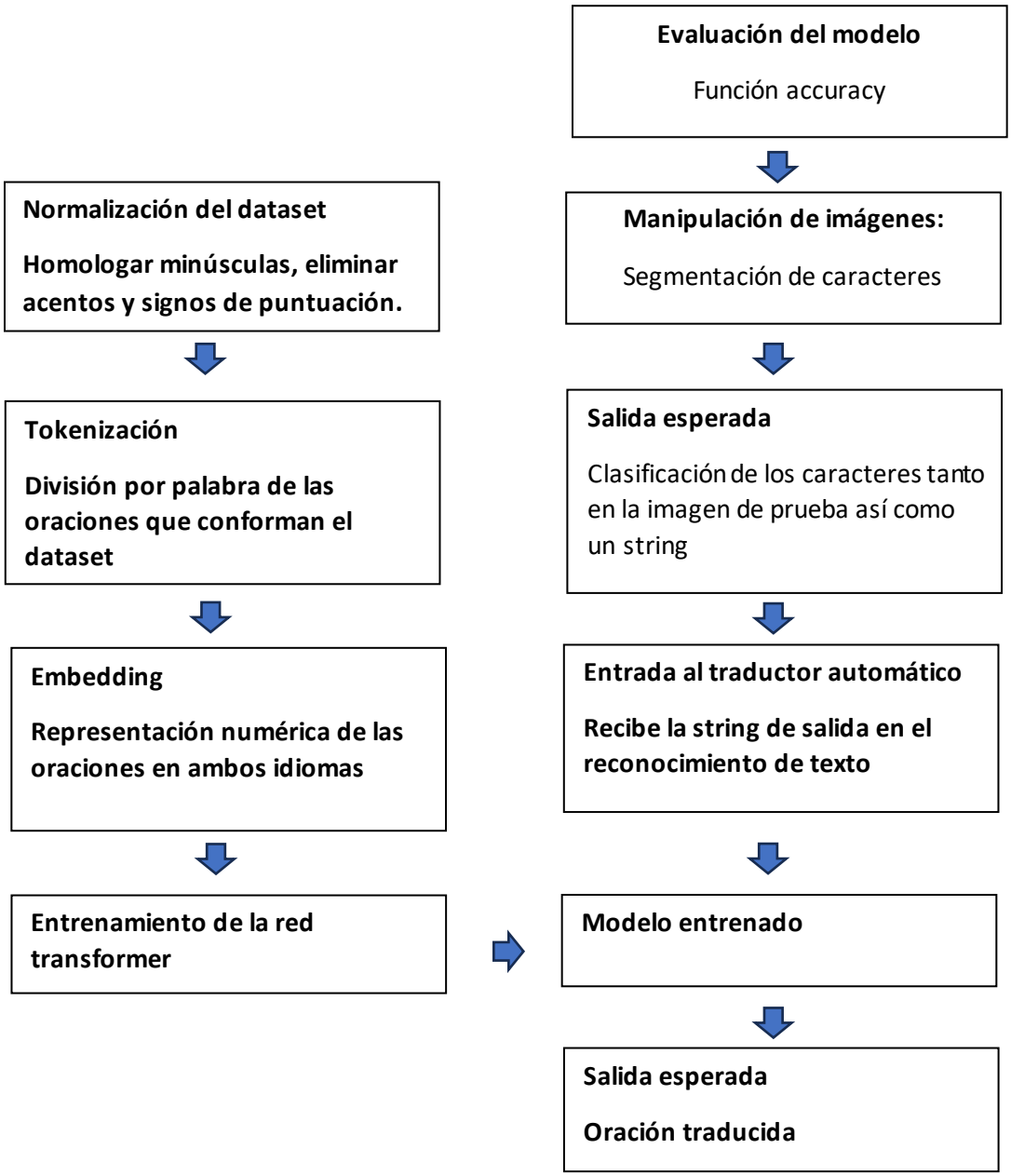
Finalmente, la fase de salida del Transformer involucra la capa lineal y la función softmax. La capa lineal realiza una transformación ponderada de la representación contextualizada de la secuencia, introduciendo no linealidades y ajustes óptimos para capturar la complejidad del lenguaje. Esta transformación prepara el terreno para la función softmax, que normaliza las salidas de la capa lineal en una distribución de probabilidad sobre el vocabulario objetivo. La función softmax asigna probabilidades a cada palabra, permitiendo al modelo seleccionar la palabra más probable en cada posición de la secuencia.

Es por esto por lo que, en el campo del procesamiento del lenguaje natural, los Transformers se han ganado reconocimiento por su eficacia y versatilidad, destacándose como una opción excepcional para abordar los desafíos de la traducción automática.

Diagrama de bloques



Machine Learning
Agosto – Diciembre 2023
Dr. Luis Carlos Padierna García



Actividades de programación

La realización del proyecto se dividió en la programación de dos secciones principales, el reconocimiento de texto en imágenes y la traducción, el primero abarca el uso de una CNN para el reconocimiento de texto en imágenes, mientras que, el segundo, corresponde a la implementación de una red transformer para la traducción de oraciones de inglés a español y viceversa. Principalmente, Daniela Jiménez y Angélica Márquez se encargaron de la primera sección y Miriam

Calderón de la segunda, pero la revisión y resolución de problemas se llevó a cabo entre las tres. Las actividades realizadas se enlistan a continuación:

Para la parte de reconocimiento de texto:

1. Formar el dataset: en este caso, uso de la librería *EMNIST y MNIST* para letras y dígitos respectivamente.
2. Implementación de la arquitectura de Red LeNet (CNN).
3. Entrenamiento de la red: realizado a partir de dos modelos, para letras en mayúsculas y para números.
4. Carga de imágenes para realizar predicciones.
5. Probar imágenes nuevas para ajuste de parámetros en la red neuronal.

Para la parte de traducción:

1. Formar el dataset
2. Preprocesar del dataset: Eliminar signos de puntuación, homologar las letras a solo minúsculas, retirar acentos.
3. Crear los espacios vectoriales de lenguaje (uno de palabras en inglés y otro de palabras en español)
4. Codificar las frases en el dataset para ambos idiomas (asignar vectores con valores numéricos)
5. Definir y entrenar de la red transformer
6. Definir y probar las funciones traducción español-inglés e inglés-español

Referencias

1. Sanz, F. (2021, December 21). *Transformer: la tecnología que domina el mundo*. The Machine Learners. https://www.themachinelearners.com/transformer/

2. Furnieles, G. (2023, April 3). Transformers in depth – Part 1. Introduction to Transformer models in 5 minutes. *Medium*. https://towardsdatascience.com/transformers-in-depth-part-1-introduction-to-transformer-models-in-5-minutes-ad25da6d3cca

3. *OpenCV: Image processing in OpenCV.* (s.f). *https://docs.opencv.org/3.4/d2/d96/tutorial_py_table_of_contents_imgproc.html*

4. Blurredmachine. (2020, 13 de julio). *LeNet Architecture: A complete guide.* Kaggle. https://www.kaggle.com/code/blurredmachine/lenet-architecture-a-complete-guide

5. Murzova, A., & Murzova, A. (2021, 5 mayo). *Otsu's thresholding technique | LearnOpenCV*. LearnOpenCV – Learn OpenCV, PyTorch, Keras, Tensorflow with examples and tutorials. https://learnopencv.com/otsu-thresholding-with-opencv/

6. *What is the use of torch.no_Grad in PyTorch?* (s. f.). Data Science Stack Exchange. https://datascience.stackexchange.com/questions/32651/what-is-the-use-of-torch-no-grad-in-pytorch

7. *EMNIST — TorchVision Main documentation*. (s. f.). https://pytorch.org/vision/main/generated/torchvision.datasets.EMNIST.html

8. Bits, C. [@codificandobits]. (2020, julio 5). *MACHINE TRANSLATION con Redes Transformer en PYTHON (Tutorial)*. Youtube. https://www.youtube.com/watch?v=p2sTJYoIwjo