Environment Descriptor for Visually Impaired People Implemented on Raspberry Pi Based on Convolutional and Recurrent Artificial Neural Networks

Rafael Chourio *,[a], Wilmer Sanz [b]

[a]*Maestría en Ingeniería Eléctrica, Facultad de Ingeniería. Universidad de Carabobo. Valencia, Venezuela.*
[b]*Laboratorio de Robótica y Visión Industrial. Escuela de Ingeniería Eléctrica. Facultad de Ingeniería, Universidad de Carabobo. Valencia, Venezuela*

La selección de este artículo fue realizada en el marco de la Jornada de Investigación de la Escuela de Ingeniería Eléctrica "Prof. César Rodolfo Ruiz" Octubre 2020, siendo la evaluación, arbitraje, aceptación y edición a cargo de Revista Ingeniería UC.

https://doi.org/10.54139/revinguc.v28i1.15

Abstract.- Vision problems and blindness are disorders of the human body that according to figures from the World Health Organization affect 217 million people in low-income countries. The quality of life of at least 75 million of them can be improved with the development of systems that allow guiding them safely in their daily tasks; this is where it is important to look for technological alternatives oriented to solve this problem and it is precisely where the idea of this research work was born. The idea presented here is based on the development of an image description system trained with deep learning algorithms based on convolutional and recurrent neural networks, implemented in a single-board computer. This implementation uses a low-cost camera for taken images of the environment and obtains a description of it that can be converted to an audible voice signal through a hearing aid system so that visually impaired people can improve their standard of living by obtaining real-time information from the environment surrounding.

Keywords: Artificial Neural Network; Natural Languaje Processing; Raspberry Pi; Deep Learning; Features Extraction.

Descriptor del Entorno para Personas con Discapacidad Visual Implementado en Raspberry Pi Basado en Redes Neuronales Artificiales Convolucionales y Recurrentes

Resumen.- Los problemas de visión y la ceguera son desórdenes del cuerpo humano que de acuerdo con cifras de la Organización Mundial de la Salud afectan a 217 millones de personas en países de bajos recursos. La calidad de vida de al menos 75 millones de estas personas puede ser mejorada con el desarrollo de sistemas que permitan guiarlas de manera segura en sus tareas de desplazamiento diario; allí es donde resulta importante buscar alternativas tecnológicas orientadas a solventar esta problemática y es donde precisamente nace la idea de este trabajo de investigación. La idea aquí presentada se basa en el desarrollo de un sistema de descripción de imágenes entrenado con tecnología de aprendizaje profundo, con redes neuronales convolucionales y recurrentes implementado en un computador de placa única, de modo tal que con una cámara de bajo costo se puedan tomar imágenes del entorno y obtener una descripción del mismo que pueda ser convertida a una señal de voz audible a través de un sistema de audífonos, para que personas con deficiencia visual puedan mejorar su nivel de vida a través de la obtención de información en tiempo real del entorno que los rodea.

Palabras clave: Redes Neuronales; Procesamiento Natural de Lenguaje; Raspberry Pi; Aprendizaje Profundo; Extracción de Características.

Recibido: 22 de octubre, 2020.
Aceptado: 29 de noviembre, 2020.

* Autor para correspondencia:
*Correo-e:*rafaelchourio@gmail.com (R. Chourio)

1. Introducción

Una de las discapacidades que mayores limitaciones puede producir al normal desenvolvimiento de un ser humano se refiere a las correspondientes

al sentido de la vista, dada nuestra alta dependencia de este sentido para la interacción con el medio ambiente (en relación por ejemplo con otros sentidos o mecanismos de interacción como el oído, el habla, el gusto o el olfato). Los problemas de visión y la ceguera son desórdenes del cuerpo humano que generan la pérdida o falta de observación visual, y que en general se relacionan con problemas fisiológicos o neurológicos.

A través de la historia de la humanidad se han ideado un sinfín de soluciones basadas en la técnica y la práctica a fin de dar una mayor calidad de vida a las personas con diversos grados de discapacidad visual; todos estos esfuerzos se han visto reflejados en la creación y desarrollo de instrumentos como espejuelos, lentes de contacto o anteojos, pasando por el desarrollo de métodos de lectura para invidentes como el sistema Braille hasta los actuales esfuerzos basados en el uso de sistemas de visión artificial que ayuden a describir el entorno que rodea a las personas con deficiencias visuales.

De acuerdo con la Clasificación Internacional de Enfermedades (CIE-10) actualizado al año 2008 [1], el desempeño de la función visual se puede clasificar en cuatros categorías principales: Visión Normal, Discapacidad Visual Moderada, Discapacidad Visual Grave y Ceguera. Según datos de la Organización Mundial de la salud, para Octubre del año 2017 la cifra estimada a nivel mundial de personas con discapacidad visual era de 253 millones de individuos, de los cuales 36 millones estaban diagnosticados con ceguera permanente y 217 millones con discapacidad visual que iba de moderada a grave.

De estas 217 millones de personas con baja visión a nivel mundial el 90 % de ellas (195 millones aproximadamente) viven en países de ingresos bajos. De este universo se estima que unos 120 millones de ellos padecen dicha afección por errores de refracción no corregidos (problema que en gran medida puede ser resuelto con el uso de lentes correctivos u operaciones quirúrgicas). Por lo tanto la población restante (al menos 75 millones de personas) va a requerir algún tipo de apoyo visual para dar mayor normalidad a sus vidas, ya que para sus afecciones hoy día no existe solución tecnológica ni médica posible. A todas estas cifras también debemos sumar que el 28 % de las personas con baja visión se encuentran en edad laboral productiva.

La calidad de vida de las personas con impedimento visual puede ser elevada enormemente en cuanto se desarrollen sistemas que permitan guiarlas de manera segura en sus tareas de desplazamiento pedestre diario, siendo este uno de los grandes problemas que confrontan tanto las personas que sufren este tipo de desórdenes como sus familiares y allegados [2]. Para poder ayudar a una persona con discapacidad visual a describir el entorno que le rodea, es necesario que la información generada pueda ser de alguna manera transmitida a dicha persona de forma tal que pueda ser interpretada, para lo cual la forma más simple es a través del uso de alguno de los otros sentidos; para esta labor el sentido que se muestra con mayor capacidad de cumplir con esta función es el oído.

La principal función de la ciencia aplicada es la resolución de problemas que mejoren las condiciones de vida de los seres humanos, bien sea para potenciar las limitaciones inherentes a nuestra forma biológica, para mejorar la dinámica y/o funcionamiento social o para superar las limitaciones de nuestros sentidos producidas por discapacidades que puedan mermar nuestras condiciones de vida. Es por ello que se hace importante buscar alternativas tecnológicas que permitan mejorar la calidad de vida de las personas con desórdenes de baja visión, de modo tal de poder brindarles las herramientas necesarias para que puedan llevar una vida más independiente y productiva. Dada la alta incidencia de este tipo de desórdenes en países de bajos ingresos (lo cual implica elevados niveles de pobreza) se hace importante tomar en cuenta que dichas alternativas, para poder ser de ayuda efectiva a estas personas, deben ser de bajo costo y ser capaces de funcionar sin conexión a internet dedicada.

Es en base a todo ello que el presente trabajo propone la implementación de un sistema de bajo costo orientado a la ayuda de personas con deficiencia visual en países de bajos ingresos económicos. La idea presentada se basa en el desarrollo de un sistema de descripción de

imágenes entrenado con tecnología de aprendizaje profundo e implementado en un computador de placa reducida, de modo tal que con una cámara de bajo costo se puedan tomar imágenes del entorno y obtener una descripción del mismo que sea convertida a una señal de voz audible a través de un sistema de audífonos, para que así las personas con deficiencia visual puedan mejorar su nivel de vida a través de la obtención de información en tiempo real del entorno que los rodea.

1.1. Aprendizaje Automático: Redes Neuronales Artificiales

El Aprendizaje Automático se refiere a la rama de las ciencias de la computación que estudia las técnicas que le permiten a los computadores aprender en lugar de seguir instrucciones programadas; más específicamente trata acerca del desarrollo de programas que le permitan a un computador generalizar comportamientos tomando como base un conjunto de ejemplos dados a partir de los cuales debe aprender [3]. Si se parte de la premisa de que el método de investigación lógico inductivo se basa en la elaboración de conclusiones de corte general partiendo de observaciones particulares, no es difícil concluir que los algoritmos de máquinas de aprendizaje terminan siendo un proceso de inducción del conocimiento.

De acuerdo a la forma que tienen los datos a partir de los cuales se pretende aprender con una máquina de aprendizaje, se pueden reconocer varios tipos de algoritmos. Entre los más conocidos se encuentran el Aprendizaje Supervisado, el Aprendizaje Semisupervisado y el aprendizaje No Supervisado; con respecto al tipo específico de técnicas utilizadas para implementarlos para labores de clasificación o regresión sobre un conjunto de datos a partir del cual se quiere realizar un proceso de aprendizaje, estos enfoques se encuentran agrupados en 7 principales tipos a saber: Árboles de Decisiones, Reglas de Asociación, Algoritmos Genéticos, Redes Neuronales Artificiales, Máquinas de Vectores de Soporte, Algoritmos de Agrupamiento y Redes Bayesianas.

El desarrollo de las redes neuronales artificiales se inspiró en el funcionamiento del cerebro humano para procesar información y aprender a partir de su entorno. El descubrimiento de este mecanismo de funcionamiento del cerebro fue realizado en 1888 por Ramón y Cajal [4], quien descubrió la estructura celular del sistema nervioso. Inspirado en este mecanismo, en 1986 Rumelhart, Hinton y McClelland [5] definieron una neurona artificial como un dispositivo capaz de generar una única salida a partir de un set de entradas. En la Figura 1 se puede observar una red neuronal artificial con una única salida.

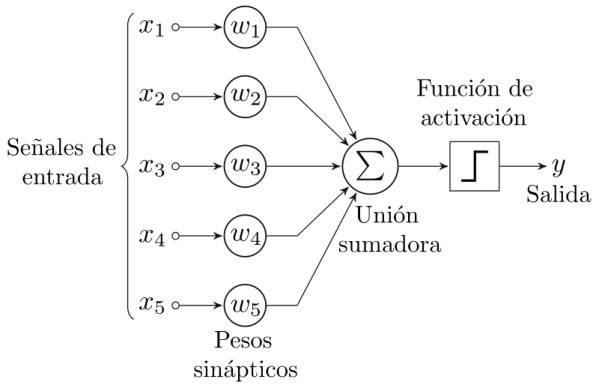

Figura 1: Red Neuronal Artificial

Las señales de entrada son denotadas como x, cada una de ellas es modificada por un factor denominado peso sináptico, el cual es diferente para cada entrada. La suma de todas las entradas ponderadas por su respectivo peso sináptico es aplicada a una función de activación, que es la función encargada de romper la linealidad del modelo para generar una salida y. Esta red neuronal puede ser utilizada para clasificación binaria o regresión. En la Figura 2 se puede observar una red neuronal artificial con múltiples salidas.

A fin de que el algoritmo de aprendizaje pueda ser efectivo en la tarea de clasificación o regresión la función de salida obtenida a partir de los datos de entrenamiento (en el caso del aprendizaje supervisado) se debe comparar con la etiqueta que cada ejemplo posee. A esta función de comparación se le llama función de costo, que depende de las variables w y b (pesos y bias de la red entrenada).El ajuste óptimo de los pesos se basa en la minimización de la función de coste, lo cual se traduce en un alto nivel de predicción. Para ello una de las técnicas más usadas es la del

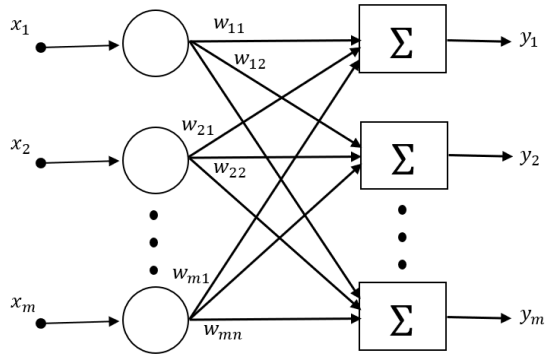

Figura 2: Red Neuronal Artificial Multi-salida

gradiente descendiente,la cual consiste en calcular la dirección del gradiente de la función de coste en el punto establecido por cada evaluación del conjunto de entrenamiento, para así ajustar los pesos y bias en la dirección donde dicho gradiente descienda.

1.2. Aprendizaje Profundo (Deep Learning)

Las redes neuronales artificiales pueden ser utilizadas en arquitecturas de múltiples capas, donde cada capa sea capaz de procesar y extraer determinadas características de la información a analizar para lograr un mayor nivel de abstracción, del mismo modo que lo hacen las capas de neuronas en el cerebro humano. A pesar de haberse diseñado modelos de redes neuronales desde los años 80 del siglo XX, la capacidad computacional de los procesadores existentes en esa época era insuficiente para aplicaciones con muchas capas de profundidad. No es sino hasta bien entrada la segunda década del presente siglo cuando la capacidad de cálculo de los modernos procesadores y el desarrollo de potentes tarjetas de video (GPU) han potenciado el desarrollo de aplicaciones de redes neuronales de múltiples capas. Esto es lo que se conoce como aprendizaje profundo (referido como Deep Learning en idioma inglés).

El uso de múltiples capas para la extracción de características complejas en los datos de entrenamiento ha permitido el análisis automático de imágenes, el procesamiento natural de lenguaje y la mejora en la predicción de sistemas complejos (clima, análisis de variables en bolsa de valores entre otras).Junto con la evolución en capacidad de cálculo se ha requerido el desarrollo de investigaciones matemáticas, cálculo numérico y optimización de algoritmos para hacer frente a problemas que se presentan en dichas redes neuronales, entre los cuales se cuentan el desvanecimiento del gradiente, la explosión del gradiente o la sensibilidad a la traslación o rotación en el reconocimiento de imágenes.

1.2.1. Redes Neuronales Convolucionales

Las redes neuronales convolucionales están basadas en los trabajos realizados por Hubel y Wisel en 1959 [6], que dieron luz sobre la comprensión del mecanismo de funcionamiento de la corteza visual del cerebro. Su origen está en la arquitectura de red neuronal presentada por Kunihiko Fukushima en 1980 [7], a la cual le dio el nombre de Neocognitron; no obstante no es sino hasta el año 2011 [8] en que se logra refinar esta técnica con uso de backpropagation e implementarla en un GPU para hacer clasificación de imágenes.

En las redes neuronales convolucionales se programan neuronas con campos receptivos de creciente complejidad, aplicando técnicas de filtrado al campo visual de cada neurona a fin de generar, entre otras cosas, insensibilidad con respecto a la posición de un objeto dentro del campo visual completo de la imagen, proveer un mecanismo de reducción de complejidad computacional y extracción de características de las imagen. La aplicación de la técnica de convolución a un conjunto de pixeles de una imagen o a la salida de una capa de neuronas se realiza a través de la ecuación (1).

$$\left[Y_j = \left[b_j + \Sigma_i K_{ij} Y_i\right]\right]. \qquad (1)$$

En donde Y_j es la salida de la capa de convolución agregada a la salida de una capa de neuronas o los pixeles de una imagen Y_i, K_{ij} es la matriz o núcleo de convolución correspondiente y que depende de la característica que se desea extraer y b_j es un factor de desplazamiento o bias. Esta operación que es similar en cierto sentido a la operación de convolución matemática en el tiempo que es ampliamente usada en los

sistemas de comunicaciones, tiene el efecto de lograr un filtrado de características de la capa de neuronas anterior a través de los parámetros del núcleo aplicado. Aplicando capas de convolución a imágenes, combinadas con funciones de activación y otras técnicas de regularización y reducción de dimensionalidad, es posible el análisis de imágenes complejas para extraer de ellas características correspondientes a objetos presentes en ellas e incluso la relación entre dichos objetos (encima, debajo, delante, detrás, etc.).

1.2.2. Redes Neuronales Recurrentes

El uso de Redes Neuronales secuenciales o convolucionales en aplicaciones para datos de tipo secuencial tales como vídeo, música o análisis de lenguaje son limitados debido a que no poseen capacidad de tomar en cuenta las decisiones anteriores como base para realizar una predicción; por ejemplo, para poder hacer la traducción de un texto de un lenguaje a otro, es importante hacerlo en el contexto de las palabras ya traducidas a fin de que la respuesta obtenida tenga sentido. Por ello se crearon las redes neuronales recurrentes, las cuales poseen lazos de transmisión de información que permiten que la información de decisiones anteriores persista.

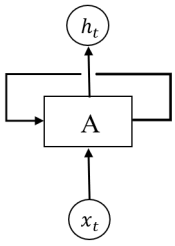

Figura 3: Red Neuronal Recurrente Enrollada

En la Figura 3 se puede observar un diagrama simplificado donde además de los estados de entrada y salida existe un estado interno oculto que se transmite de forma cíclica a través del lazo dibujado, el cual representa un estado de memoria que permite recordar las decisiones anteriores y es utilizado en la celda A para decidir la salida h_t. Este diagrama se puede desenrollar, de manera de ver con mayor claridad que estos estados internos básicamente trasmiten información del pasado a través del tiempo (Figura 4).

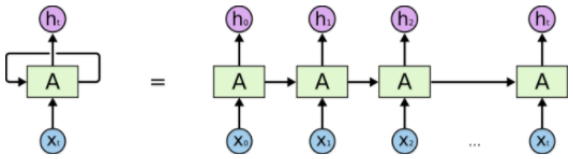

Figura 4: Red Neuronal Recurrente Desenrollada

Las primeras redes neuronales recurrentes a pesar de poder transmitir información de estados pasados para poder tomar mejores decisiones presentes, tenían limitaciones en cuanto a la cantidad de información que pueden efectivamente almacenar. Si los estados temporales transmitidos incluyen memorización a largo plazo, los algoritmos de optimización utilizados tienden a ser poco efectivos para esta arquitectura de aprendizaje [9]. Para contrarrestar esto se desarrollaron las Redes Neuronal Recurrente LSTM, que son una arquitectura de aprendizaje desarrollada para aumentar la efectividad del entrenamiento con dependencias temporales largas [10].

LSTM es el acrónimo de Long Short Term Memory. Con este esquema, se controla la información de los estados internos permitiendo olvidar o mantener información conforme sea requerido (de la misma forma en que trabaja la mente humana); este diseño está específicamente desarrollado para evitar el problema de larga dependencia temporal, razón por la cual actualmente es ampliamente utilizado en la resolución de una gran variedad de problemas de índole secuencial.

1.3. Antecedentes

Tanto el reconocimiento automático de imágenes como el procesamiento natural de lenguaje han comenzado a converger en un nuevo nicho de investigación relacionado con la extracción de características de imágenes para su correcta descripción. Uno de los trabajos que se ha convertido en una referencia en este campo es el paper titulado “Show and Tell: A Neural Image Caption Generator” [11] desarrollado por integrantes del equipo investigador de Google, Inc. Este trabajo fue una de las primeras aproximaciones

basada en la combinación de un modelo de red neuronal convolucional que alimenta a una red neuronal recurrente para la obtención de una descripción de la imagen procesada.

En "Long-term Recurrent Convolutional Networks for Visual Recognition and Description" [12] se utiliza una variante denominada "Long-term Recurrent Convolutional Network" (LRCN) a fin de procesar secuencias de imágenes de video con la finalidad de generar descripciones coherentes de las acciones que ocurren en el mismo. Para ello hicieron pruebas con varios modelos pre-entrenados, con ajustes de varios hiperparámetros, con diversos tipos y números de capas en redes neuronales recurrentes a fin de lograr descripciones con la mayor robustez posible.

Dado lo novedoso de este campo de aplicaciones, aún se está lejos de alcanzar niveles de descripción comparables al logrado por el ser humano. Es por ello que han surgido propuestas orientadas a la mejora en la precisión y naturalidad de las descripciones obtenidas, bien sea para imágenes o secuencias de imágenes. "Boosting Image Captioning with Attributes" [13] y "Exploring Models and Data for Remote Sensing Image Caption Generation" [14] son dos de los estudios más recientes enfocados a la optimización de modelos y parámetros orientados a la generación de descripciones basadas en imágenes, lo cual es un área de estudio en constante desarrollo para lograr aplicaciones cada mes más precisas en tiempo real.

Un área del aprendizaje supervisado que actualmente se encuentra lo suficientemente madura como para el desarrollo de aplicaciones robustas es el reconocimiento automático de texto; es en este sentido que se han desarrollado multitud de propuestas orientadas a la aplicabilidad de esta tecnología para la asistencia a personas con discapacidad visual. Tanto en la investigación "A Smart Reader for Visually Impaired People Using Raspberry PI" [15] como en "Design of An Electronic Narrator on Assistant Robot for Blind People" [16] se han utilizado computadores de placa reducida Raspberry Pi como base para la construcción de sistemas narrativos que puedan, a través de una cámara instalada en dicho dispositivo, capturar imágenes de un texto y procesar el mismo para convertir dicha información en audio que pueda ser escuchado por una persona con impedimento visual.

Este mismo enfoque puede ser utilizado para describir el ambiente circundante a personas que tengan limitaciones para poder observarlo. Precisamente esta fue la idea presentada en el estudio "Automated Neural Image Caption Generator for Visually Impaired People" [17], que planteó la posibilidad de utilizar las descripciones provenientes de un sistema entrenado con la utilización de una red neuronal convolucional como entrada a una red neuronal recurrente de forma tal de poder explicar de algún modo el ambiente circundante y servir como referencia a personas que posean diversos tipos de discapacidad visual.

2. Metodología

La presente investigación se circunscribe a la aplicación de un sistema de aprendizaje supervisado para la identificación de los principales elementos presentes en una imagen capturada en tiempo real y la interrelación existente entre ellos, a fin de crear una frase coherente que permita describir la escena presentada de una forma concisa. Esta aplicación se implementó en un computador de placa reducida, en aras de sentar este precedente como base a la futura implementación de este tipo de desarrollos en sistemas de bajo costo accesible para países y personas de bajos ingresos. Igualmente es importante el hecho de que el prototipo propuesto puede funcionar sin necesidad de conexión a internet.

El computador escogido para la implementación de este sistema ha sido el Raspberry Pi 3, debido a su versatilidad, bajo costo y software libre para desarrollo. La frase generada para describir la escena capturada se convierte a una señal de voz, a fin de que la persona con impedimento visual tenga la oportunidad de escuchar la descripción de la imagen a través de un sistema de audífonos instalados en el computador de placa reducida. Igualmente se acopla una cámara capaz de hacer captura de la imagen deseada y un sistema de

accionamiento que permite al usuario decidir el momento en que quiere tener una descripción de su entorno.

El software de programación para la implementación del presente trabajo de investigación fue Python, en su versión 2.7, debido a su compatibilidad con las librerías TensorFlow, Keras y OpenCV en el sistema operativo Raspbian utilizado por el Raspberry Pi 3 (las tres librerías mencionadas fueron necesarias para las tareas de entrenamiento del sistema de aprendizaje supervisado, para la construcción del modelo de predicción y para la captura de imágenes desde la cámara instalada en el dispositivo); Raspbian (actualmente Raspberry Pi OS) es una distribución del sistema operativo GNU/Linux basado en Debian, y es de manera oficial el sistema operativo primario de la familia de placas Raspberry Pi.

Es importante recalcar que los objetos identificables en las imágenes para la generación de las descripciones están basados en el set de entrenamiento utilizado para entrenar el sistema de aprendizaje profundo, por lo que objetos fuera de ese conjunto pueden crear también descripciones menos precisas.

El set de datos utilizado para el entrenamiento de la red neuronal recurrente dispone de descripciones en idioma inglés, razón por la cual este fue el idioma de las descripciones generadas tanto en el set de pruebas como en las imágenes capturadas por la cámara instalada en el Raspberry Pi.

2.1. Modelo Utilizado

El modelo elegido para ser implementado en el Raspberri Pi se basó en el NIC (Neural Image Caption Generator) publicado por el equipo de Google Inc. en el año 2015 [11]. Este modelo permitió generar descripciones nuevas en imágenes nunca vistas, tomando ventaja de la técnica de extracción de características de modelos previamente entrenados y que resultaron exitosos en tareas de clasificación de imágenes. Se escogió el mismo debido a su mecanismo de validación bien documentado (lo cual es un punto fuerte de este trabajo y que no pudo ser apreciado en muchos de los trabajos revisados), además de que la potencia de cálculo requerida sería lo suficiente para ser manejado por una laptop de potencia media (por el uso de la mencionada extracción de características).

El modelo seleccionado consta de dos sistemas de redes neuronales a saber: una red neuronal convolucional que se encarga de procesar la imagen objetivo (etapa de Visión), y una red neuronal recurrente que genera una frase coherente que describa con mayor probabilidad a la imagen (etapa de Generación de Lenguaje).

La etapa de Visión se compone de una red neuronal profunda convolucional, cuya entrada se compone únicamente de la imagen objetivo, y que se debe encargar de extraer las características de dicha imagen. Debido a las exigencias computacionales de entrenar una red de este tipo desde cero, y dado que ya existen modelos pre entrenados y públicos para descarga libre en la red suficientemente probados, se decidió utilizar uno de estos modelos para la extracción de las características necesarias. La etapa de Generación de Lenguaje se implementó con una red neuronal LSTM, cuyo estado inicial se alimenta con las características que se extraen de la imagen a fin de escoger dentro del vocabulario definido la próxima palabra con mayor probabilidad de ocurrencia, repitiendo esta operación hasta completar una frase completa que describa a la imagen. En la Figura 5 se muestra el diagrama de bloques del modelo.

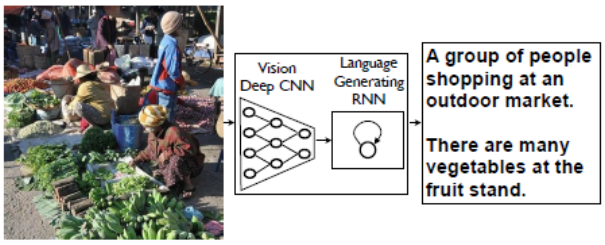

Figura 5: Modelo NIC [11]

Para la programación de la etapa de Visión se compararon varios modelos pre entrenados que están incluidos dentro de la librería Keras. Debido a la limitación en la capacidad disponible de cálculo para el desarrollo del presente trabajo, donde se contó sólo con una laptop con un procesador Intel i5 con dos núcleos, de 8 GB de memoria RAM y

sin GPU instalado, la escogencia de un modelo que minimizase la cantidad de operaciones requeridas para extraer las características de las imágenes en la fase de entrenamiento era primordial; lo mismo aplica a la limitada capacidad en procesador y memoria del Raspberry Pi para que el tiempo de respuesta por cada imagen no sea excesivo. Por ello se decidió utilizar el modelo InceptionV3 debido a su tamaño reducido (92 MB en total) y elevada precisión (78,8 % en el Top-1 Accuracy de Imagenet y 94,4 % en el Top-5 Accuracy del mismo set de imágenes).

Para elegir el set de datos óptimo para entrenamientos, se verificó en Vinyals [11] que el set de datos COCO mostró el mejor desempeño a nivel de evaluación de las frases generadas. Es por ello que se decidió utilizar este mismo set de datos, en específico el correspondiente al año 2014. El mismo está compuesto por imágenes de 91 clases, además de poseer un archivo de datos de referencia donde, por cada imagen, se tienen 5 descripciones que la representan. A partir de las imágenes se obtienen las características que se extraen por medio del modelo inceptionV3 pre entrenado sobre Imagenet, y a partir de las descripciones se obtiene el vocabulario que sirve como parámetro de evaluación para la etapa de generación de lenguaje; la forma de obtener las características de la imagen (y no la probabilidad de pertenencia a cada una de las clases sobre la que fue entrenado) es eliminando la última capa de decisión (SoftMax); con esto se obtiene el conjunto de características extraídas propias de cada imagen que van a servir como parámetros iniciales para la red neuronal recurrente LSTM. El set de datos usado estuvo compuesto por 82.783 imágenes de entrenamiento (con un tamaño aproximado de 13 GB) y por 40.504 imágenes de validación (con un tamaño aproximado de 6,5 GB). De este último conjunto de imágenes, se utilizaron 104 imágenes para pruebas, quedando el conjunto de validación compuesto por 40.400 imágenes en total. En la Figura 6 se observa el diagrama de bloques del modelo entrenado.

A fin de optimizar el entrenamiento de la red neuronal recurrente (la cual es la única que requiere ser entrenada, puesto que para la etapa de visión se

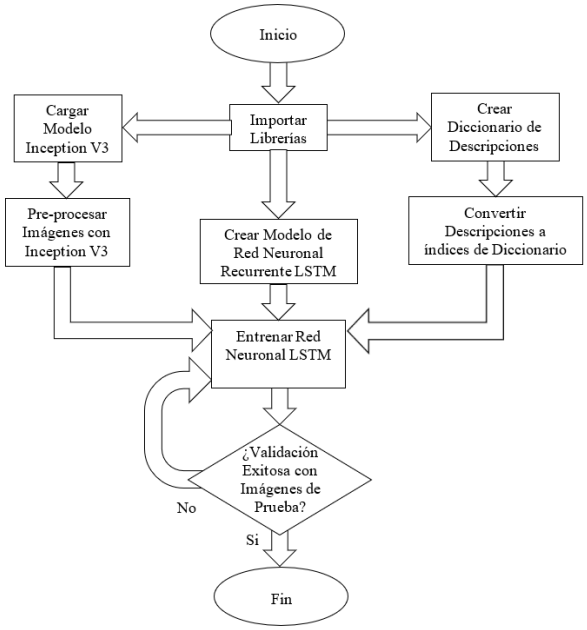

Figura 6: Diagrama de Bloques del Modelo Entrenado

está utilizando una red pre entrenada), se hizo una extracción previa de las características de cada una de las imágenes de entrenamiento y validación. Con el computador utilizado, se requirieron alrededor de 15 horas de cómputo para completar esta tarea.

Para implementar la etapa de Generación de Lenguaje es necesario el establecimiento de un vocabulario a partir del cual se van a escoger las palabras para la formación de las oraciones, este se construyó a partir de las descripciones que acompañan a las imágenes del conjunto de entrenamiento y constó de 8765 palabras, agregándose además 4 palabras adicionales: una para inicio de cada frase, una para fin de cada frase, una para identificar palabras desconocidas del vocabulario y una para especificar la ausencia de palabras. En total, el diccionario generado estuvo compuesto por 8769 palabras. El vocabulario generado se almacenó en formato *.pickle para su posterior utilización.

Las características obtenidas de cada imagen por medio del modelo InceptionV3 sirven como estado inicial a la celda LSTM. Para simplificación del modelo, se escogió una red LSTM de 756 unidades, la cual alimenta dos capas de redes neuronales, la última de las cuales tiene el tamaño

del vocabulario utilizado. A esta última se le aplica una máscara para el cálculo de probabilidad solo sobre las palabras que son conocidas, y luego una capa de cálculo SoftMax para selección de la palaba con mayor probabilidad de ser utilizada para construcción de la frase. La palabra escogida, así como las características de la imagen y el vocabulario es nuevamente alimentado a otra celda LSTM de la misma forma anterior para la escogencia de la siguiente palabra; el proceso se repite hasta que la palabra escogida es la palabra de fin de oración o hasta que se alcanza la longitud máxima de oración establecida, la condición que ocurra primero. Finalmente, la función de pérdidas usada para evaluar el aprendizaje fue la de entropía cruzada.

El entrenamiento se efectúo por batches de imágenes. Cada batch se definió de 64 imágenes, y cada época de entrenamiento se estipuló en 1.290. Para validación se tomaron 300 batches. Con estos parámetros se pudo observar un decrecimiento continuo de la función de pérdidas asociadas, y con 18 épocas se pudo obtener resultados de predicción precisos en el set de pruebas. El modelo completo se almacenó por cada época de entrenamiento para referencia en un archivo en el disco duro. En la Figura 7 se observa la curva de pérdidas por cada época de entrenamiento obtenida para ambos sets.

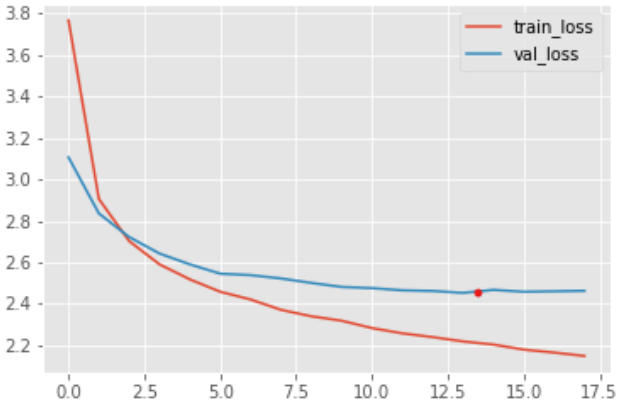

Figura 7: Curva de Precisión del Entrenamiento

2.2. *Implementación en Raspberry Pi*

En la Figura 8 se puede observar el diagrama de bloques del prototipo implementado.

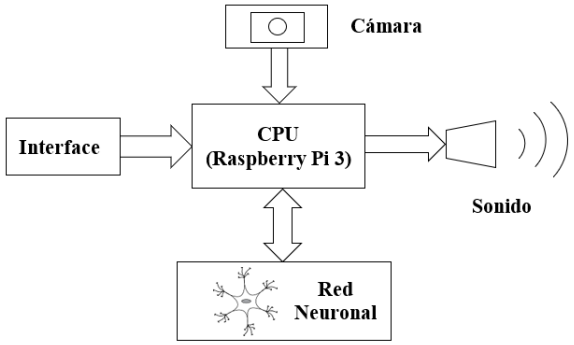

Figura 8: Diagrama de Bloques del Prototipo Implementado

Al computador de placa reducida que se observa en el centro de la Figura 8 se agregó una cámara para poder tomar la imagen del entorno. La interface está compuesta por dos pulsadores, uno para iniciar el proceso de toma de imagen-generación de descripción y otro para repetir la última descripción procesada, un sistema de sonido que está conformada por un set de auriculares, y un modelo de red neuronal previamente entrenado cuya misión es procesar la imagen capturada y generar la descripción de la misma. De la etapa de entrenamiento se obtuvieron dos archivos: el modelo generado con un peso de alrededor de 105 MB y el vocabulario cuyo tamaño fue de unos 167 kB.

La cámara utilizada fue una Kuman modelo SC15, con bus de comunicación para ser instalada en el puerto de cámara del computador de placa reducida. La instalación de la cámara requirió la habilitación de la misma en el entorno de configuración del Raspberry Pi, accesible a través del comando raspi-config en la consola de comando del mismo. Para el módulo de interface se agregó un pequeño protoboard, en el cual se incluyeron dos mini pulsadores; estas señales fueron conectados a las entradas del Raspberry Pi en las GPIO 23 y 24; estos pines debieron ser configuradas en el mismo código del programa implementado para utilizar su resistor de pull-up interno y ser configurados como entradas. Finalmente se añadió un set de audífonos, a través de los cuales se pueden escuchar las descripciones generadas del entorno. A fin de

proveer portabilidad, se le incorporó una batería portátil de equipos celulares a través del puerto micro USB. El prototipo final implementado, con todos sus componentes funcionales operativos instalados se puede observar con detalle en la Figura 9.

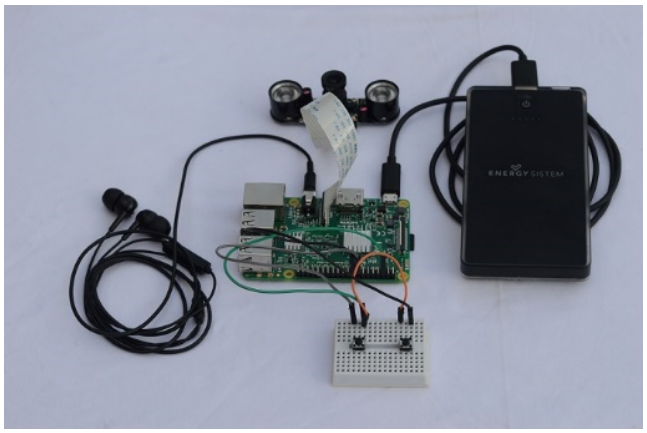

Figura 9: Prototipo Implementado

Por su carácter de licencia pública de uso y facilidad de instalación se utilizó la aplicación de conversión de voz a texto eSpeak, a fin de poder generar las indicaciones audibles que van a servir de guía a las personas con deficiencia visual que requieran usar el prototipo. Finalmente se hizo prueba funcional del código y se configuró el Raspberry Pi para ejecutar este programa cada vez que se encendiese el mismo de forma automática con la intención de poder hacer uso de la aplicación sin necesidad de conectar el equipo a ninguna pantalla ni ordenador. En la Figura 10 se muestra el diagrama de flujo simplificado del código implementado en el computador de placa reducida.

3. Análisis y discusión de resultados

A fin de hacer la medición cuantitativa de las descripciones generadas por el modelo de predicción entrenado, se escogieron 10 imágenes de forma aleatoria del conjunto de pruebas; 5 de ellas se muestran en la Figura 11 como referencia. Es importante indicar que estas imágenes deben pertenecer a una muestra estadística similar a las imágenes usadas para entrenamiento a fin de lograr un equilibrio Bias-Variance que minimizase el factor reducible del error esperado [18]. Esta es

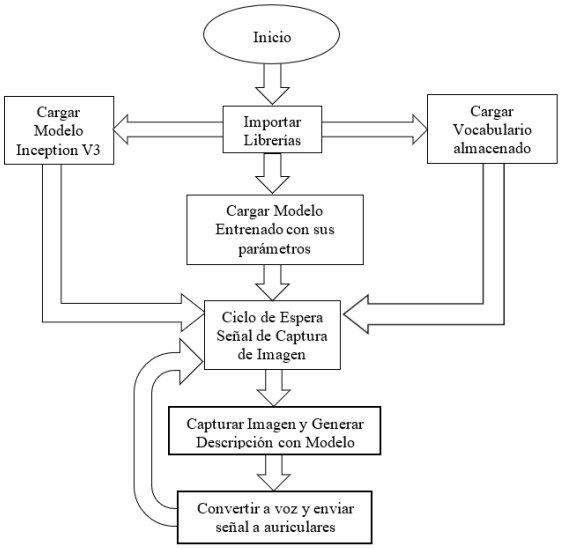

Figura 10: Diagrama de Flujo del Código Implementado en Raspberry Pi

la manera de garantizar la pertinencia del modelo; para ello este set no debe ser utilizado en ninguna de las fases de aprendizaje y debe solo ser reservado para la fase de pruebas.

A nivel de sistemas de traducción automáticos basados en sistemas de inteligencia artificial, e incluso en algunos trabajos de descripción de imágenes se han utilizado diversas técnicas de análisis automático que aumentan la rapidez de las pruebas. Sin embargo, partiendo de las premisas generadas por Vinyals [11], la métrica utilizada para cuantificar la naturalidad y correspondencia de las descripciones generadas para poder medir las sutilezas de lenguaje que un ser humano es capaz de detectar debe ser realizada por una evaluación humana de las mismas. Debido a la indisponibilidad de individuos cuya lengua materna fuese el idioma inglés, la población utilizada para evaluación estuvo compuesta por 5 personas bilingües cuyo idioma materno era el español. La evaluación de las descripciones se hizo a través de una única pregunta, que daba un ranking del 1 al 4 a la calidad de la descripción generada para cada una de las diez imágenes a evaluar.

En la Figura 12 se puede observar la gráfica de distribución normal inversa acumulada de los resultados, tanto para las valoraciones dadas a las descripciones de referencia (las que acompañan

Figura 11: Algunas Imágenes para Validación del Modelo Entrenado

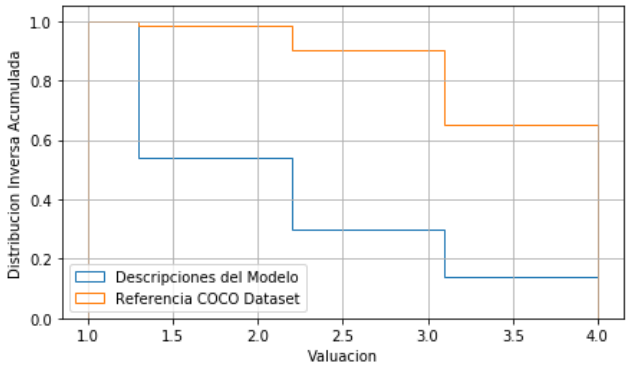

Figura 12: Gráfico de Análisis de Resultados

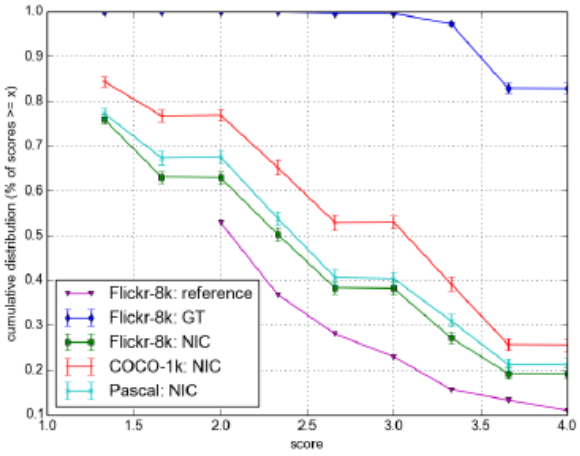

Figura 13: Gráfico de Resultados de Referencia [11]

al set de datos) como para valoraciones dadas a las descripciones generadas por el modelo desarrollado. Allí se puede observar que las puntuaciones dadas en general para las descripciones que acompañaban a cada una de las imágenes fue superior a las generadas por el modelo. Sin embargo, este comportamiento era de esperarse si se compara con los resultados mostrados en la Figura 13, y que se corresponden con la publicación original del modelo que sirvió de base para el presente trabajo de investigación.

El promedio de puntuación dado para las imágenes que acompañaban a cada imagen fue de 3.54, en comparación con el valor de 3.89 en el estudio de referencia [11]. En el mismo orden de ideas, el promedio de puntuación dado para las imágenes generadas por el modelo que se desarrolló en el presente trabajo fue de 1,98; en comparación con valores que van desde 2,08 en el estudio referencial. Igualmente se puede observar que los comportamientos de ambas Figura 12 y Figura 13 es similar en cuanto a la forma de la distribución inversa acumulada de probabilidad.

Para validar el funcionamiento del sistema construido con el Raspberry Pi se procedió a la recolección de un pequeño conjunto de imágenes con sus respectivas descripciones en tiempo real, en diversos entornos de tipo interior y exterior. En las Figuras 14 y 15 se expone una muestra de 6 imágenes con sus respectivas descripciones, las cuales fueron obtenidas directamente del prototipo implementado. Queda para trabajos futuros la evaluación de estas descripciones.

Se pudo observar que las descripciones generadas por las imágenes recolectadas por el prototipo funcional de ajustaron de forma bastante aproximada a los elementos capturados y la relación entre ellos. Cabe destacar la capacidad de generalización para describir imágenes con un vocabulario y cantidad de ejemplos relativamente limitados.

Figura 14: Imágenes obtenidas del Prototipo 1-3

Figura 15: Imágenes obtenidas del Prototipo 4-6

4. Conclusión

El presente trabajo demostró la factibilidad de utilizar modelos basados en redes neuronales para la obtención de descripciones del entorno en tiempo real, que sean capaces de ayudar a personas con impedimento visual a mejorar su calidad de vida. En particular, se pudo evidenciar que es posible la obtención de modelos cuyo tamaño y complejidad computacional una vez entrenado sean adecuados para su implementación en computadores de placa reducida, los cuales son una alternativa actual de bajo costo para el desarrollo de prototipos orientados al mercado de personas con bajos recursos económicos.

El nivel de precisión obtenido en las descripciones predichas por un modelo de parámetros reducidos (a fin de generar tiempos de entrenamiento razonables en un computador con un procesador de gama media y sin GPU incorporado) demostró ser similar al obtenido por el estudio referencial [11]. Los tiempos de respuesta entre la captura de la imagen y la obtención de la respuesta en el prototipo fue inferior a los 10 segundos, lo cual resultan ser tiempos poco manejables para aplicaciones que generan descripciones en tiempo real, pero que con implementaciones más orientadas al uso en usuarios finales pueden ser mejorados.

El aspecto más relevante del presente trabajo es el precedente que se genera en cuanto a la posibilidad de desarrollar sistemas de bajo costo, sin necesidad de conexión dedicada a internet y con software de licenciamiento público para la solución de un problema que puede mejorar la vida de más de 75 millones de personas a nivel mundial de forma directa, y de muchos más en forma indirecta (cuando se incluyen los familiares y amigos que sirven de apoyo continuo para soporte vital de las personas con diversos grados de impedimento visual).

El desarrollo de nuevos modelos que mejoren la precisión en la generación de descripciones a partir de una imagen dada, el incremento del tamaño de los datos disponibles tanto para entrenamiento como para validación del entrenamiento del modelo implementado, el aumento de la capacidad

de cálculo en los sistemas computacionales actuales, y la disminución de costos asociado con el desarrollo de nuevos sistemas de implementación de placa reducida, son todos factores que apuntan a la intensificación de los esfuerzos en aras de conseguir soluciones prácticas y factibles para ayudar a todas las personas de bajos recursos en países de bajos ingresos a ser incorporados de manera paulatina al mercado laboral global.

Es la intención que los resultados obtenidos en el presente trabajo, en cuanto a la factibilidad de implementación y la utilización de modelos de máquinas de aprendizaje para la obtención de descripciones del entorno, sirvan como aporte para la mejora futura de los millones de personas que pueden servirse de la técnica y el desarrollo tecnológico para mejorar su calidad de vida y tener la oportunidad de valerse por sí mismos por encima de las limitaciones que alguna discapacidad visual pueda intentar poner en su horizonte.

5. Referencias

[1] Organización Mundial de la Salud, *Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud*. Décima revisión. Organización Panamericana de la Salud, 2008.

[2] N. Chenthamil, N. Rekha, and P. Poovizhi, "Portable Camera Based Identification System for Visually Impaired People," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 19, no. 3, pp. 19 141–19 146, 2016.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, USA: The MIT Press, 2016.

[4] S. Ramón y Cajal, "Estructura de los Centros Nerviosos de las Aves," *Revista trimestral de Histología Normal y Patológica*, vol. 1, no. 1, pp. 314–318, 1888.

[5] D. Rumelhart, G. Hinton, and J. Mcclelland, "A General Framework for Parallel Distributed Processing," *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, 1986.

[6] D. Hubel and T. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[7] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural Networks*, vol. 1, no. 2, pp. 119–130, 1988.

[8] D. Ciresan, M. Ueli, J. Masci, L. Gambardella, and J. Schmidhuber, "Flexible, High Performance Convolutional Neural Networks for Image Classification," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, vol. 2, 2011, pp. 1237–1242.

[9] Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-Term Dependencies with Gradient Descent is Difficult," *IEEE Transactions On Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164.

[12] J. Donahue, L. Hendrick, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

[13] T. Yao, P. Yingwei, Y. Li, Z. Qiu, and T. Mei, "Boosting Image Captioning With Attributes," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4904–4912.

[14] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring Models and Data for Remote Sensing Image Caption Generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2018.

[15] D. Velmurugan, M. Sonam, S. Umamaheswari, S. Parthasarathy, S. Guadarrama, K. Saenko, and T. Darrell, "A Smart Reader for Visually Impaired People Using Raspberry PI," *International Journal of Engineering Science and Computing*, vol. 6, no. 3, pp. 2997–3001, 2016.

[16] R. Ardiansyah, "Design of An Electronic Narrator on Assistant Robot for Blind People," in *MATEC Web of Conferences*, vol. 42, no. 03013, 2016, pp. 03 013p.1–03 013p.5.

[17] C. Elamri and T. de Planque, "Automated Neural Image Caption Generator for Visually Impaired People," *IOSR Journal of Engineering (IOSRJEN)*, vol. 10, pp. 28–33, 2018.

[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, USA: Springer, 2015.