UNIVERSIDAD SAN FRANCISCO DE QUITO
USFQ

Colegio de Ciencias e Ingenierías

Resultados Teórico - Prácticos para Modelo de Red Neuronal Convolucional de Reconocimiento de Gestos Faciales

Christopher Alexander León Viracucha

Cristian Olimpo Vizcaino Viñan

Ingeniería Electrónica y Automatización

Trabajo de fin de carrera presentado como
requisitopara la obtención del título de
Ingeniero Electrónico

Quito, 21 de Diciembre de 2021

UNIVERSIDAD SAN FRANCISCO DE QUITO
USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Resultados Teórico - Prácticos para Modelo de Red Neuronal Convolucional de Reconocimiento de Gestos Faciales.

Christopher Alexander León Viracucha

Cristian Olimpo Vizcaino Viñan

Nombre del profesor, Título académico **Diego Benítez, Ph.D.**

Quito, 21 de Diciembre de 2021

DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manualesde la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánicade Educación Superior.

Nombres y apellidos: Christopher Alexander León Viracucha

Código: 00200712

Cédula de identidad: 1725496911

Lugar y fecha: Quito, 21 de Diciembre de 2021

DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manualesde la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánicade Educación Superior.

Nombres y apellidos: Cristian Olimpo Vizcaino Viñan

Código: 00140220

Cédula de identidad: 1719144113

Lugar y fecha: Quito, 21 de Diciembre de 2021

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017)Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de QuitoUSFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

En la actualidad existen muchas propuestas de técnicas de reconocimiento de expresiones faciales basadas en distintas ramas de investigación. Las Redes Neuronales Convoluciones entrenadas por bases de datos de imágenes destacan entre las mejores herramientas debido a su gran porcentaje de precisión. En este trabajo se presenta un estudio y comparación de los resultados teóricos y prácticos de un modelo CNN para el reconocimiento de seis gestos faciales. El modelo propuesto fue entrenado y evaluado para dos bases de datos logrando una puntuación de precisión media de 72,37 % y 71,4 %. Estos resultados fueron comparados con una prueba en vivo realizada a veinte personas donde se logró una precisión media de 73,58 % y 70,53 %.

Palabras claves: CNN, aprendizaje profundo, reconocimiento de expresiones faciales, bases de datos, error porcentual, optimización y entrenamiento.

ABSTRACT

Nowadays there are many models for facial expression recognition based on different branches of research. Trained Convolutional Neural Networks are one of the most used models due their high percentage of precision. The following paper presents a study and comparison of the theoretical and practical results of a CNN model for the recognition of six facial gestures. The proposed model was trained and evaluated for two data sets, achieving a precision score of 72.37% and 71.4%. These results were compared with a live test performed on twenty people where a precision of 73.58% and 70.53% was achieved.

Key Words: CNN, Deep Learning, facial expression recognition, data sets, percent error, optimization and training.

Tabla de Contenidos

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

En el desarrollo tecnológico que hemos evidenciado en las últimas tres décadas, se ha planteado la duda de si las computadoras podrán volverse autónomas y más inteligentes. El concepto de unir lo que entendemos por inteligencia y tecnología fue tratado por varias ramas importantes dentro de la Inteligencia Artificial (Wolfgang, 2009). De las ramas existentes se destaca la conexionista que buscó modelos que repliquen el comportamiento y estructura del cerebro. Por otro lado, en el Instituto Tecnológico de Massachusetts se investigaba una IA enfocada en que los resultados del procesamiento tuviesen el carácter de proporcionar resultados inteligentes sin priorizar las características del cuerpo humano. Pese a ambos enfoques para resolver un mismo problema todas las ramas de investigación de la inteligencia artificial buscan comprender la inteligencia y aprendizaje humano para desarrollar una tecnología que pueda adquirir conocimiento y resolver problemas en diversas áreas como por ejemplo la estadística, la comunicación y el procesamiento de imágenes (Dafonte, 2007).

Desde el punto de vista electrónico los avances de los microprocesadores han revolucionado el desarrollo tecnológico durante las tres décadas pasadas, dando como resultado una mayor velocidad de procesamiento y mejor capacidad de almacenamiento (Harris, 2013). Dentro de la inteligencia artificial los avances tecnológicos permitieron desarrollar y resaltar la importancia del procesamiento de datos para introducir la rama del aprendizaje automático. El aprendizaje automático o *machine learning* se convierte en un subcampo de las ciencias de la computación con la finalidad de crear algoritmos y técnicas que permitan a las computadoras analizar datos y a partir de esta información poder generar predicciones (Hinestroza, 2018).

Una de las técnicas más investigada del aprendizaje automático es el aprendizaje profundo o *Deep Learning,* donde el término aprendizaje destaca por el concepto conexionista de las redes neuronales. Al principio de su investigación, los modelos de redes neuronales artificiales de propagación de retroalimentación destacaron, ya que conceptualmente las

neuronas artificiales se organizan desde un nivel de entrada pasando por distintos niveles ocultos y terminando en un nivel de salida. Esto permitió que durante el proceso de retro - propagación se calcule errores de las neuronas de salida comparándolas con una salida deseada para así poder propagar el error desde la salida hasta niveles anteriores de la red. Esta teoría demostró el equivalente a un entrenamiento basado en los valores de error de salida, donde la repetición de este proceso hace que la red produzca salidas que correspondan a las entradas con un valor de error cada vez menor logrando un aprendizaje automático para redes neuronales artificiales. La funcionalidad de los modelos de redes neuronales artificiales de propagación de retroalimentación depende del número de niveles ocultos existentes, donde dos niveles ocultos son suficientes para representar funciones discontinuas (Jackson, 2019).

El estudio y propuestas actuales de diferentes arquitecturas de redes neuronales artificiales mantienen un concepto similar al de propagación de retroalimentación o *feedforward.* Siendo sistemas de procesamiento computacional con nodos o neuronas artificiales interconectados de manera distribuida para aprender conjuntamente de las entradas generando una auto - optimización mediante el aprendizaje (O'Shea, 2015).

Las arquitecturas más relevantes para resolver aplicaciones complicadas de reconocimiento de patrones en base a imágenes son las redes neuronales convolucionales o CNN por sus siglas en inglés. Las CNN fueron diseñadas para el reconocimiento de patrones de imagen para una entrada en formato de dato de imagen. A partir de la entrada, el algoritmo genera una extracción de valores propios de imágenes, reconocimiento y la convolución para la identificación y análisis de imágenes (Lou, 2020). Dentro de la visión por computadora, el reconocimiento de gestos faciales se ha vuelto un tema relevante, principalmente porque nos permite interpretar el estado de ánimo de una persona frente a una situación y desde un punto de vista neurológico se puede interpretar información valiosa sobre el comportamiento humano. Por ejemplo, Espinel propone cinco modelos de CNN para el reconocimiento de siete

gestos faciales con un entrenamiento basado en tres bases de datos. Como resultados en las precisiones obtenidas (ACC) se obtuvo en el peor de sus modelos un ACC de 15 %, mientras que para el mejor de sus modelos 96 % (Espinel, 2021).

Pese a los distintos estudios de arquitecturas de redes artificiales para el reconocimiento de gestos faciales, es importante unir todas las características de una red convolucional para obtener resultados que tengan congruencia con la base de datos y la aplicación que se vaya a dar a la CNN con un específico aparato biométrico. En ese sentido, en el siguiente documento se presenta un estudio teórico y práctico sobre los ACC obtenidos con la aplicación de un modelo CNN. Los resultados presentados están basados en distintos cambios de características de la CNN para la obtención de los mejores resultados de ACC teóricos y resultados de precisión basados en una prueba de la red neuronal convolucional en vivo para un grupo de veinte personas.

METODOS Y MATERIALES

Bases de Datos de Gestos Faciales

Se trabajó con dos bases de datos de la plataforma pública *kaggle.com,* la cual es una comunidad apoyada por Google para compartir bases de datos y códigos para aplicaciones de *Deep Learning.* A continuación, comentamos características de cada base de datos utilizadas.

La primera base de datos fue compartida en *Kaggle* por Manas Sambare bajo el nombre de *Facial Expressions Recognition* – 2013. La base de datos cuenta con un *set* de entrenamiento de 24,612 imágenes divididas a partir de los seis gestos faciales: 3995 imágenes para ira, 436 disgusto, 7215 felicidad, 4965 neutral, 4830 tristeza y 3171 para sorpresa. Mientras que el *set* de prueba cuenta con 6154 imágenes. Las imágenes están en escala de grises a 48 x 48 píxeles. Como características relevantes resalta la variedad de edades y etnias dentro de las imágenes,

esto se puede observar en la Figura 6, donde se muestra una imagen de ejemplo para cada gesto facial.

La segunda base de datos fue compartida por Jonathan Oheix en 2019 bajo el nombre de *Face Expression Recognition Dataset*. La base de datos cuenta con 24,718 imágenes de entrenamiento y 6048 imágenes para la validación. Las imágenes están en escala de grises y contienen 2304 píxeles. Entre las imágenes hay los mismos seis gestos faciales que la primera base de datos, FER *2013* y de igual manera hay una gran diversidad de etnias y edades. Como se observa en la Figura 7, las imágenes del *dataset* son de bebés, jóvenes y adultos mayores.

Bibliotecas y GPU

El entrenamiento y el procesamiento de datos de una red neuronal requiere de unidades de procesamiento modernas y eficientes. Esta limitación relacionada a la potencia de cómputo ha generado que empresas como Google y Facebook apuesten al desarrollo bibliotecas de código abierto para el *Deep Learning;* entre las bibliotecas más utilizadas destacan los enfocados al procesamiento de imágenes, reconocimiento de voz y pronósticos (Nguyen, 2019). Para el entrenamiento de la red neuronal convolucional propuesta en este proyecto se usaron las bibliotecas Scikit-Learn (Pedragosa, 2011), Keras (Chollet, 2015) con su ImageDataGenerator y Tensorflow (Abadi, 2015) como backend. Las tres bibliotecas se basan en el lenguaje de programación python 3.7.0. Como solución a las limitaciones de una unidad de procesamiento se utilizó la herramienta en línea, Google Colaboratory (Bisong, 2019), ideal para proyectos de aprendizaje automáticos, ya que nos brinda acceso gratuito a unidades de procesamiento para procesar de manera correcta la cantidad de imágenes requeridas para el entrenamiento de la red neuronal convolucional del proyecto.

Modelo Propuesto

El tamaño de imagen requerido a la entrada de la red neuronal implementada es de (48x48x1) esto implica que cada imagen ingresada tiene un alto y ancho de 48 pixeles y está en escala de grises. En el modelo propuesto la CNN consta de tres etapas, la primera constituida por cinco capas de convolución y una capa de *maxpooling*, que tienen como objetivo identificar patrones gráficos; la segunda etapa constituida por dos capas ocultas que se encargan de clasificar los datos recibidos y por último, la tercera etapa constituida netamente por la salida de la red neuronal. Tal como se puede apreciar en la Figura 1 donde se muestra el diagrama de bloques para el modelo de la red neuronal convolucional propuesta.

Con respecto a la Figura 1, se observa como en la primera etapa hay una capa de convolución que posee 32 filtros y un tamaño de kernel de 3x3, que tienen la función de extraer las principales características de la imagen, como por ejemplo detalles que sean relevantes para la identificación de un gesto facial. La siguiente capa es la de *batch-normalization* donde se fija las medias y las variaciones de las entradas de cada capa mediante la transformada de normalización de lotes con el fin de tener distancia entre datos de 0 o 1, con este método se logra mejorar el rendimiento de la red neuronal dado el dato antes mencionado de la distancia de datos. La capa de *batch-normalization* está seguida por una capa de *max-pooling* de tamaño 2x2, que tiene como objetivo reducir la varianza espacial para la salida de datos obtenidos en la primera capa de convolución anteriormente descrita. El concepto varianza espacial proviene de la estadística espacial. Este proceso se repite cuatro veces más, con la diferencia del tamaño de los *kernels* para cada repetición. En la primera repetición el tamaño de los *kernels* es de 5x5, mientras que en las otras tres repeticiones el tamaño de los *kernels* es de 3x3. De igual manera, en las repeticiones el tamaño de los filtros se va multiplicando por dos hasta llegar a la última capa con un total de filtros de 512 respectivamente. Con respecto al tamaño de la capa max-*pooling*, esta se mantiene igual en todas las repeticiones. Continuando con la descripción de la

Figura 1, las dos etapas siguientes solo usan dos capas ocultas que se conectan a la salida con una capa, *fully connected,* que provee la salida de la CNN o en otras palabras la última clasificación obtenida por la CNN.

Configuración del modelo

Como una técnica para el entrenamiento de la CNN, se optimizó un hiper-parámetro llamado en inglés *epochs*, el cual permite las iteraciones durante el entrenamiento. Las variaciones realizadas para los *epochs* fueron de 5, 50, 100, 150 y 200. En las Tablas 1 y 2 se puede observar los resultados obtenidos de precisión para cada variación de los *epochs.* El valor 200 el elegido para las pruebas en vivo, debido a que un mayor aumento en los *epochs* dejó de generar cambios significativos para los porcentajes de precisión del entrenamiento de la red. El valor elegido además nos proporcionó un tiempo de compilación menor a tres horas para el entrenamiento de la CNN. También, se consideró otra técnica de entrenamiento añadiendo un parámetro *adam optimizer*, el cual es un algoritmo de optimización que sirve de alternativa al descenso de gradiente estocástico clásico con el fin de actualizar los pesos de la red neuronal de manera iterativa en función de los datos de entrenamiento (Brownlee,2017). La tasa de entrenamiento fue probada a $1x10^{-4}$ y el valor de los *droputs 0.2* o hablando en porcentajes este valor correspondería al 20%.

Para el entrenamiento y validación se usaron las bases de datos *FER 2013* y *Face Expression Recognition Dataset.* Con respecto a *FER 2013* se utilizaron 24,612 imágenes para el entrenamiento de la CNN y 6154 imágenes para la validación. Por otro lado, con el *dataset, Face Expression Recognition,* se utilizaron 24,718 imágenes para entrenamiento y 6048 para la validación. En ambos *datasets* de entrenamiento, el gesto facial con más imágenes es la felicidad que cuenta con 7215 imágenes para *FER 2013* y 7164 para *Face Expression Recognition.* De manera contraria, el gesto facial con menos imágenes dentro de ambos

datasets para entrenamiento es el disgusto con 436 imágenes para *FER 2013* y 426 para *Face Expression Recognition.* Con respecto a las pruebas en vivo, se probó la red neuronal convolucional para veinte personas hay tres niños, catorce adultos y tres adultos mayores. Las veinte personas realizaron los seis gestos propuestos para el proyecto, dando un total de 120 resultados de precisión obtenidos con las pruebas en vivo. Como ejemplo de las pruebas en vivo se adjunta las Figuras 3, 4, 5 y las Tablas 3 y 4. Dentro de las figuras se puede observar un cuadrado azul, que tiene el fin de identificar las caras en las pruebas en vivo y además mostrar el resultado de precisión para la efectividad de reconocimiento de cada uno de los seis gestos faciales con la prueba en vivo de la red neuronal convolucional. Por otro lado, en las Tablas 3 y 4 se observan los 120 resultados de precisión obtenidos en las pruebas en vivo.

RESULTADOS Y DISCUSIÓN

Las pruebas realizadas para el entrenamiento de la red neuronal convolucional se realizaron con la herramienta Google Colab (Bisong, 2019), a partir del hiper-parámetro anteriormente mencionado, *epochs*, se realizaron cinco pruebas para los dos *datasets* y los resultados se observan en las Tablas 1 y 2 donde para 5 *epochs* con el *dataset* FER *2013* se obtuvo un ACC de 45,23 %, mientras que para el *dataset Face Expression Recognition* con 5 *epochs* se obtuvo un ACC de 43,67 %. En el valor final de variación de *epochs* se obtuvo una mejora del 27,14 % para el *dataset* FER *2013* y una mejora del ACC del 27,73 % para el *dataset Face Expression Recognition.* Para el proyecto no se aumentó más el valor de los *epochs* debido a que la diferencia de mejora de ACC para cada *dataset* comenzó a ser del 1 % para cada aumento de 50 *epochs*, tomando como conclusión dejar el valor 200 *epochs* debido al tiempo de compilación de la red y la poca variación de su ACC. De igual manera, en la Figura 2 se presenta una matriz de confusión para el *dataset FER 2013* a 200 *epochs*, la cual permite observar los porcentajes de ACC obtenidos para cada gesto facial y además permite ver la

cantidad de veces que la CNN confundió un gesto con otro gesto, la matriz de confusión es un resultado del entrenamiento de la red neuronal.

Los cambios realizados fueron interpretados en primer lugar a base de prueba y error para ver la evolución de la red neuronal dependiendo de los valores que se probaron; se puede seguir aumentando los *epochs* pero por cuestiones de optimización no es recomendable, ya que el aprendizaje de una red neuronal tiene sus límites. Al sobrepasar dicho límite la tasa de aprendizaje tiende a ser la misma o inclusive puede disminuir, como se observa en la tabla 2, entre 150 y 200 *epochs* sí existe un mayor porcentaje, sin embargo, la diferencia no es tan grande. Además, los recursos computacionales necesarios para llegar a 200 *epochs* son altos y esto se demuestra en la compilación de 2 a 3 horas de la CNN. El entrenamiento de la CNN en una computadora con características computacionales de un estudiante universitario promedio resultó en el doble de tiempo, demostrando la importancia del uso del Google Colab (Bisong, 2019).

Una vez optimizada la red neuronal se la descargó y se corrió la CNN en una computadora con las siguientes características: Intel Core i7, 16 GB RAM y sistema operativo de 64 bits. Las pruebas en vivo se realizaron en la computadora descrita con la implementación de un dispositivo biométrico de las siguientes características, cámara *HP Wide Vision HD* de 1280 x 720 píxeles de resolución. Ya realizadas las pruebas en vivo, se obtuvo el rendimiento del modelo propuesto y se compararon los valores teóricos obtenidos para cada *Dataset*. Los resultados de esta comparación se presentan en la Tabla 5, donde se obtuvieron los errores porcentuales para cada gesto facial a reconocer, estos resultados se presentan como el error porcentual de los resultados ACC para cada base de datos con los resultados en vivo de ambos *datasets*. En ambos *datasets* el gesto facial con la menor diferencia de error porcentual entre los resultados ACC y los porcentajes de precisión en vivo fueron el enojo y la felicidad. Mientras que disgusto, neutral, tristeza y sorpresa tuvieron una diferencia porcentual entre

resultados teóricos y prácticos del 30 % al 55 %. Los resultados obtenidos en las Tablas 3 y 4 se obtuvieron mediante la prueba en vivo de la CNN con veinte personas. Se solicitó a cada una de las personas que realicen los seis gestos faciales frente a la cámara y se tomó datos de los porcentajes observados en el programa. Este programa muestra los valores de porcentaje en vivo dentro de un cuadrado azul que tiene la función de identificar la cara de cada persona. El cuadrado azul que se observa en las Figuras 3, 4 y 5 de las pruebas en vivo, muestra su respectivo porcentaje de ACC y fueron adaptadas mediante el software OpenCV (Bradski, 2000). En este software la red neuronal convolucional fue compilada y acoplada a un programa de reconocimiento facial para las pruebas en vivo. Los resultados de las pruebas en vivo y de los valores obtenidos están presentados en las Tablas 3 y 4.

Como parte de la experimentación es necesario mencionar que cada emoción tiene un porcentaje de precisión distinto, demostrando la eficiencia de la red neuronal para el reconocimiento de gestos faciales como el enojo y la felicidad. Mientras que el disgusto, la tristeza, la sorpresa y neutral tienen menores porcentajes de ACC. Un motivo de esta diferencia es la cantidad de imágenes utilizadas en el entrenamiento de los gestos faciales de felicidad y enojo, ya que dentro de ambos *datasets* estas eran los gestos con mayor cantidad de imágenes para el *set* de entrenamiento y validación.

De cierto modo dado que el gesto de sorpresa en la mayoría de las personas es una sonrisa o el de disgusto es muy parecido al enojo, podemos observar estos datos de confusión dentro de la Figura 2, donde en la matriz de confusión se muestra las veces que la CNN confundió un gesto facial con otro. Algo más a agregar es que se encuentran datos atípicos, en los que la expresión neutral de la persona era relacionada con otro gesto, en otras palabras, encontramos casos en los que la persona normalmente tiene una expresión de enojo o tristeza, claro está que no necesariamente esa persona estaba molesta o triste, pero su expresión normal era esa, lo que dio paso nuevamente a confusión en la red.

Como últimas observaciones se puede mencionar que en la parte práctica existen muchos factores que pueden llegar a afectar la precisión de la red neuronal, como es el caso de la cantidad de luz, la calidad de la cámara empleada e inclusive que en los Datasets se utiliza expresiones muy exageradas para cada tipo de emoción.

Con la experimentación realizada de la CNN en tiempo real, es posible notar que en muchos casos las expresiones faciales naturales en un día cotidiano tienden a ser mucho más sutiles y por ende la red neuronal convolucional obtiene esos porcentajes de error. Para mejorar estos resultados es posible realizar un entrenamiento donde se use la misma cantidad de imágenes para cada gesto facial, de tal manera que se obtengan resultados mucho más comparables con respecto a las configuraciones de nuestra CNN propuesta.

CONCLUSIONES

Tras la investigación realizada en el presente escrito se puede concluir que para obtener una mejor precisión en la red neuronal convolucional es necesario ajustar ciertos parámetros e ir observando su evolución con respecto a los cambios planteados para el entrenamiento. A esto hay que agregar que es necesario conocer los parámetros internos de la CNN para modificar correctamente y así lograr mejores resultados con respecto a la precisión y por ende rendimiento.

Se concluye también que a pesar de que el modelo propuesto no tiene un porcentaje de precisión demasiado alto, sí son porcentajes que tienen coherencia y relación con respecto a lo propuesto en la parte práctica**,** ya que se logró obtener resultados similares por emoción y esto indica que lo realizado en el entrenamiento va de la mano con el rendimiento de la red en tiempo real.

Los recursos necesarios tanto para el entrenamiento de la red como para la realización de pruebas son importantes**,** ya que los mismos pueden favorecer a la precisión de la red y por ende una correcta identificación de los gestos, o pueden jugar en contra, en este caso estamos

seguros de que con una mejor cámara le podemos sacar un mejor rendimiento a nuestro modelo de red neuronal. Además, se debe considerar que el uso de un dataset o base de datos también va a tener gran impacto tanto en el entrenamiento como en las pruebas en tiempo real. En nuestro caso optamos por dos *datasets* con imágenes en escala de grises, pero el uso de un dataset a color va a darle una mayor optimización y acercamiento a un entorno real a la red neuronal.

RECOMENDACIONES

Considerando el beneficio obtenido al utilizar Google Colaboratory (Bisong, 2019), se recomienda su uso para el entrenamiento de modelos CNN, ya que no es necesario poseer grandes recursos computacionales para entrenar cualquier tipo de red neuronal artificial. Como un complemento, recomendamos utilizar *datasets* con imágenes en escala de colores, ya que los resultados de ACC que se va a obtener al final van a ser mejores, al considerar un entrenamiento con imágenes a color, también estamos brindando a la red un mayor acercamiento a las características de reconocimiento de gestos faciales.

REFERENCIAS

Abadi, M., & others. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems, Software available from tensorflow.org.

Bisong, E. (2019). Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_7

Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.

Burrueco, D. (s. f.). *Arquitectura de redes neuronales | Interactive Chaos*. Interactive Chaos. Recuperado 10 de octubre de 2021, de https://interactivechaos.com/es/manual/tutorial-de-machine-learning/arquitectura-de-redes-neuronales

Chollet, F., & others. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras.

Ertel, W. (2018). Introduction to artificial intelligence. Springer. Obtenido de: shorturl.at/yQXY8

Espinel.A, Pérez.N, Riofrío.D, Benítez.D and Moyano.R, "Face Gesture Recognition Using Deep-Learning Models," 2021 IEEE Colombian Conference on Applications of Computational Intelligence (ColCACI), 2021, pp. 1-6, doi: 10.1109/ColCACI52978.2021.9469528.

G. Lou and H. Shi. (2020). Face image recognition based on convolutional neural network, in China Communications, vol. 17, no. 2, pp. 117-124, doi: 10.23919/JCC.2020.02.010.

Giger, M. L. (2018). Machine learning in medical imaging. Journal of the American College of Radiology, 15(3), 512-520.

Harris, D., & Harris, S. L. (2010). Digital design and computer architecture. Morgan Kaufmann.

Hinestroza Ramírez, D. (2018). El Machine Learning a través de los tiempos, y los aportes a la humanidad (Doctoral dissertation, Universidad Libre Seccional Pereira).

J. Brownlee, "Gentle introduction to the adam optimization algorithm for deep learning," Machine Learning Mastery, vol. 3, 2017.

Jackson, P. C. (2019). Introduction to artificial intelligence. Courier Dover Publications. Obtenido de: https://link.springer.com/book/10.1007/978-0-85729-299-5

Nguyen, G., Dlugolinsky, S., Bobák, M. (2019) Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artif Intell Rev 52, 77–124. https://doi.org/10.1007/s10462-018-09679-z

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. Obtenido de: https://arxiv.org/pdf/1511.08458.pdf

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Quiza, J. (2021). *Modelos CNN en la clasificación de imágenes clásicas y modernas*. Medium. Recuperado 10 de octubre de 2021, de https://medium.com/datos-y-

ciencia/modelos-cnn-en-la-clasificaci%C3%B3n-de-im%C3%A1genes-cl%C3%A1sicas-y-modernas-d072a6718689

Romero, J. J., Dafonte, C. A. R. L. O. S., Gómez, Á. N. G. E. L., & Penousal, F. J. (2007). Inteligencia artificial y computación avanzada. Santiago de Compostela: Fundación Alfredo Brañas, 10-15.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach. Obtenido de: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/27702.pdf

TABLAS

Tabla 1: Porcentajes de precisión para distintos epochs del dataset 1 (Face Expression Recognition Dataset)

Epochs	Precisión
5	45,23%
50	69,70%
100	70,89%
150	71,56%
200	72,37%

Tabla 2: Porcentajes de precisión para distintos epochs del dataset 2 (FER 2013)

Epochs	Precisión
5	43,67%
50	68,68%
100	70,59%
150	71,20%
200	71,40%

Tabla 3: Resultados prácticos para la CNN para datase 1 (Face Expression Recognition)

Face Expression Recognition Dataset						
Persona	**Enojo**	**Disgusto**	**Felicidad**	**Neutral**	**Tristeza**	**Sorpresa**
1	70%	62%	100%	60%	99%	50%
2	60%	100%	50%	40%	76%	48%
3	96%	80%	100%	99%	63%	40%
4	75%	73%	100%	50%	94%	33%
5	75%	30%	100%	50%	94%	35%
6	55%	78%	46%	40%	98%	22%
7	100%	48%	100%	98%	63%	60%
8	98%	56%	100%	78%	100%	69%
9	62%	40%	100%	100%	75%	100%
10	78%	43%	100%	100%	91%	60%
11	74%	30%	100%	100%	77%	84%
12	83%	62%	100%	80%	70%	53%
13	70%	10%	100%	100%	50%	70%
14	80%	50%	100%	100%	80%	66%
15	82%	61%	100%	78%	50%	44%
16	82%	55%	100%	95%	77%	60%
17	100%	60%	100%	100%	100%	50%
18	70%	80%	100%	100%	82%	10%
19	82%	10%	100%	100%	96%	60%
20	10%	80%	100%	100%	77%	30%
%Precisión	75%	55%	95%	83%	81%	52%

Tabla 4: Resultados prácticos para la CNN para dataset 2 (FER 2013)

FER 2013 Dataset						
Persona	**Enojo**	**Disgusto**	**Felicidad**	**Neutral**	**Tristeza**	**Sorpresa**
1	50%	40%	99%	99%	100%	20%
2	70%	10%	100%	60%	98%	30%
3	98%	70%	100%	77%	100%	47%
4	77%	73%	100%	88%	96%	78%
5	63%	10%	100%	40%	96%	35%
6	60%	73%	43%	80%	50%	35%
7	73%	40%	100%	98%	100%	44%
8	78%	70%	100%	60%	40%	50%
9	80%	58%	60%	100%	40%	100%
10	78%	73%	100%	80%	100%	10%
11	30%	80%	100%	100%	68%	60%
12	100%	40%	100%	94%	100%	40%
13	80%	78%	100%	100%	40%	40%
14	60%	70%	100%	100%	68%	30%
15	40%	40%	100%	100%	100%	50%
16	40%	78%	100%	100%	80%	10%
17	100%	10%	100%	100%	80%	50%
18	100%	99%	100%	80%	100%	10%
19	70%	70%	100%	78%	100%	10%
20	60%	68%	100%	100%	96%	20%
%Precisión	70%	50%	95%	87%	83%	38%

Tabla 5: Resultados error porcentual de resultados de entrenamiento y resultados en vivo para ambos *datasets*

Error porcentual ACC entrenamiento y pruebas en vivo		
Dataset	**FER 2013**	**Face Expression Recognition Dataset**
Enojo	12,10%	4,49%
Disgusto	30,46%	36,81%
Felicidad	14,59%	14,49%
Neutral	34,52%	40,86%
Tristeza	39,89%	43,23%

FIGURAS

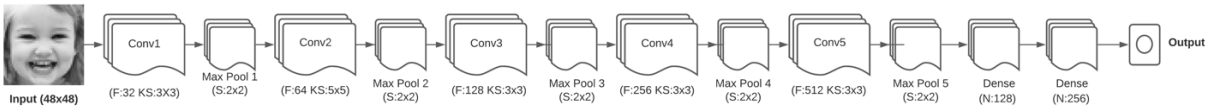

Figura 1: Diagrama de bloques para modelo de CNN

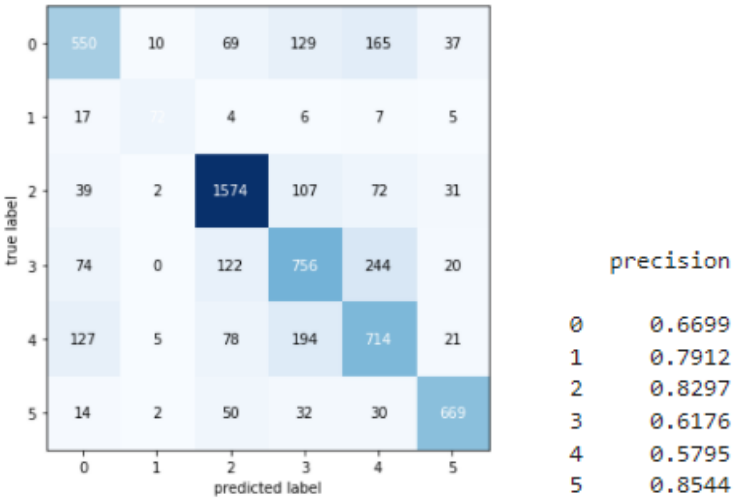

Figura 2: Matriz de confusión para 200 *epochs*, *dataset FER 2013*

Figura 3 Resultado CNN en vivo para gesto: Neutral.

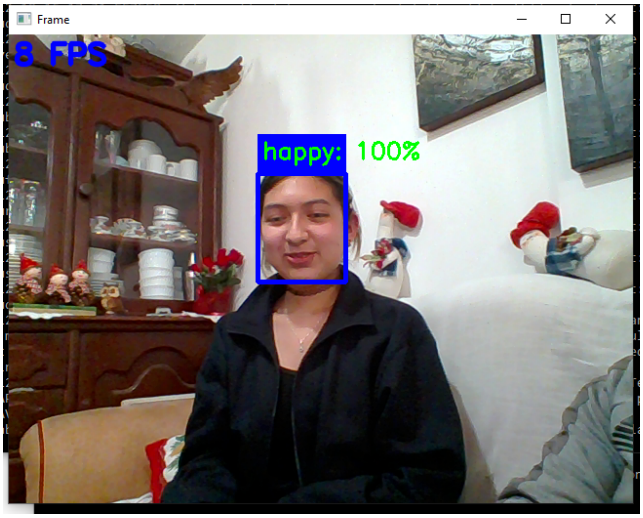

Figura 4 Resultado CNN en vivo para gesto: Felicidad

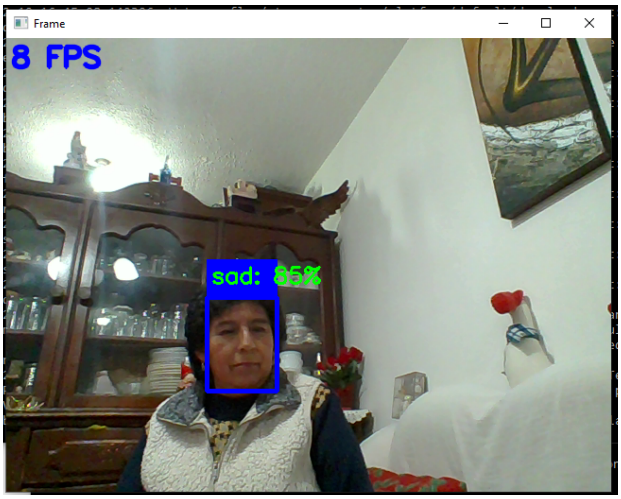

Figura 5 Resultado CNN en vivo para gesto: Tristeza

Figura 6 Ejemplos de imágenes *dataset FER 2013*

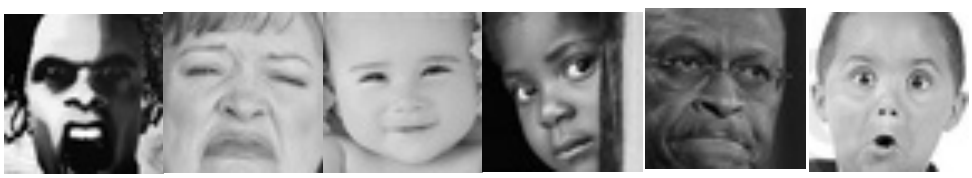

Figura 7 Ejemplos de imágenes *dataset Face Expression Recognition*