Aplicación de la Arquitectura VGG para la solución de un problema de clasificación multiclase: modelado y optimización usando RNA

Freddy Alvarado B.
Gerencia de Proyectos de IA
CMI Consulting | Lima, Perú
falvarado@cmiconsulting.pe

***Abstract*— The present work aims to introduce the reader to the world of Deep Learning as well as to expose the application of the VGG Architecture in solving a Multiclass Classification problem: design and optimization; using as Dataset the corpus of images from Kaggle Intel Image Classification referring to natural scenes around the world.**

***Resumen*—El presente trabajo tiene por objetivo introducir al lector en el mundo del Deep Learning así como exponer la aplicación de la Arquitectura VGG en la solución de un problema de Clasificación Multiclase: diseño y optimización; utilizando como Dataset el corpus de imágenes de Kaggle Intel Image Classification referidas a escenas naturales alrededor del mundo.**

Keywords—clasificación multiclase, Kaggle, CNN, VGG ImageNet, PCA, Ensemble Neural Networks

I. INTRODUCTION

En la actualidad vivimos el auge de la Inteligencia Artificial (IA, por sus siglas iniciales), es así como observamos un gran desarrollo en el logro de las capacidades que exige la Prueba de Turing[i] [1], las cuales definen a la IA como como tal: 1) Procesamiento del Lenguaje Natural 2) Representación del conocimiento 3) Razonamiento automático 4) Aprendizaje automático 5) Visión computacional y 6) Robótica. En gran medida algunas de estas características como el Procesamiento de Lenguaje Natural y Visión computacional se desarrollan vertiginosamente gracias a las Redes Neuronales Artificiales (RNA) que superó su etapa de desarrollo lento o Invierno de la IA entre otros, debido principalmente al desarrollo de la técnica de optimización de la función de coste denominada Back Propagation en 1986, y ya para inicios de 1987 la IA se convierte en Ciencia la cual goza con una gran popularidad entre la comunidad científica de diferentes países. En lo referente a Vision computacional se logra un desarrollo importante a partir del modelo de redes neuronales convulocionales denominado AlexNet en el año 2012. En lo sucesivo se vienen desarrollado una serie de mejoras en la exactitud de los modelos de regresión y clasificación gracias a nuevas técnicas para el tratamiento de datos.

II. REDES NEURONALES ARTIFICIALES (RNA)

A. Redes neuronales

Se cree que la capacidad de procesamiento de información de los seres humanos proviene principalmente de redes de neuronas, por lo que los algoritmos de redes neuronales artificiales pretenden emular este comportamiento biológico. Desde un punto de vista informal en el ámbito de la IA, la neurona se "dispara" cuando una combinación lineal de sus entradas excede un determinado umbral.

Las redes neuronales están compuestas de nodos conectadas a través de conexiones dirigidas con un peso numérico cada una. Se utiliza una función de activación para producir la salida de una neurona [1]. En la siguiente figura, extraída del libro de Peter Norvig, se presenta un modelo matemático sencillo de una neurona artificial:

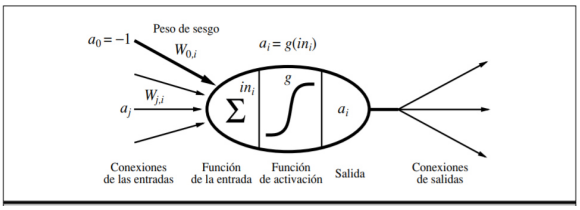


Figura 1: Modelo matemático sencillo para una neurona

B. Redes neuronales de una sola capa

Una red con todas las entradas conectadas directamente a las salidas se denomina **red neuronal de una sola capa**, o red **perceptrón**.

[2] El perceptrón creado en 1957 por Frank Rosenblatt, es un algoritmo de aprendizaje supervisado, bajo su forma más sencilla también se denomina neurona formal (resultado de los trabajos de McCulloch y Pitts en 943) y su objetivo es separar las observaciones en dos clases (o grupos) distintos con la condición de que sus datos sean linealmente separables. Su función es clasificar.

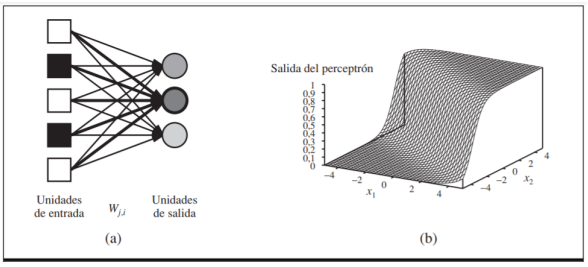


Figura 2: (a) Red perceptrón (b) Gráfico función de activación sigmoide.

C. Redes neuronales multicapa

Las redes neuronales multicapa, presentan grupos de neuronas intermedias en lo que se denomina capa oculta. La ventaja de añadir capas ocultas es que se amplía el espacio de hipótesis que puede representar la red (ver Figura 3).

D. Deep Learning

Se considera que el término Deep Learning fue introducido por primera vez en 1986, pero no ha sido hasta el siglo XXI cuando su uso se ha puesto realmente de moda. El principal objetivo del Deep Learning es emular los enfoques

de aprendizaje de los seres humanos, utilizando para ello una serie de características específicas: transformaciones no lineales y diferentes niveles de abstracción. Podemos enfocar el Deep Learning dentro de otros paradigmas de la Ciencia de Datos como una evolución natural de los mismos.

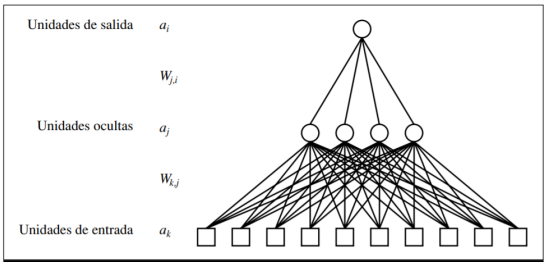


Figura 3: Ejemplo de red neuronal multicapa

El tamaño de un modelo de Deep Learning o red neuronal lo determina el número de capas intermedias, el número de neuronas de estas capas, y grado de interconexión entre neuronas. Según las últimas estimaciones, el tamaño de las redes se dobla cada cerca de 2 años y medio, y no será hasta 2050 cuando se genere una red con la complejidad del cerebro humano.

En la actualidad el trabajo con Deep Learning se ha visto favorecido por el acceso gratuito a un conjunto de librerías y especificaciones técnicas para su uso como son los frameworks: **TensorFlow** de Google, **PyTorch** de Facebook, **MXNet** de Amazon, entre otros.

E. Datasets

El uso de Datasets estándar nos permite comparar los resultados obtenidos frente a otras técnicas. Evita la tarea de tener que etiquetar las muestras. Muchos de los datasets se encuentran ya integrados dentro de los frameworks de Deep Learning, de forma que su uso se facilita ampliamente. Algunos Datasets de referencia:

- **Iris**: 150 muestras (5 características numéricas) y 3 categorías.
- **MNIST**: 60,000 muestras (28x28) y 10 categorías.
- **CIFAR-100**: 60,000 muestras (32x32) y 10 categorías.
- **Kaggle**: una subsidiaria de Google LLC, es una comunidad en línea de científicos de datos y profesionales del aprendizaje automático. Kaggle permite a los usuarios encontrar y publicar conjuntos de datos, explorar y crear modelos en un entorno de ciencia de datos basado en la web; trabajar con otros científicos de datos e ingenieros de aprendizaje automático y participar en concursos para resolver desafíos de ciencia de datos.

F. Hacia la Estandarización

Durante muchos años, el Deep Learning únicamente era accesible para desarrolladores con grandes conocimientos de aprendizaje automático y capacidad de computación. Desde 2013 han aparecido diferentes plataformas donde se ofrecían modelos ya entrenados que proporcionan etiquetas a partir de imágenes. Esto ha marcado la línea hacia la estandarización, facilitando la reutilización de recursos.

Utilizando modelos ya generados, podemos aprovechar el conocimiento codificado en redes neuronales sin la necesidad de poseer los recursos necesarios para su entrenamiento. También evita comenzar desarrollos desde 0, de forma que se pueda partir de una red que quizá haya encontrado codificación óptima para la mayoría de las imágenes, pero que no distinga bien entre categorías específicas.

III. COMPUTER VISION

La visión por computadora es una disciplina científica que incluye métodos para adquirir, procesar, analizar y comprender las imágenes del mundo real con el fin de producir información numérica o simbólica para que puedan ser tratados por un ordenador. Tal y como los humanos usamos nuestros ojos y cerebros para comprender el mundo que nos rodea, la visión artificial trata de producir el mismo efecto para que los ordenadores puedan percibir y comprender una imagen o secuencia de imágenes y actuar según convenga en una determinada situación.

Una definición pragmática de visión por computadora lo podemos encontrar en el portal de Azure de Microsoft: "El servicio Computer Vision de Azure proporciona acceso a algoritmos avanzados que procesan imágenes y devuelven información basada en las características visuales de interés."

Microsoft ofrece un conjunto de servicios agrupados en 1) Reconocimiento óptico de caracteres (OCR), extrae texto de las imágenes. 2) Análisis de imágenes, extrae muchas características visuales de las imágenes, como objetos, caras, contenido para adultos y descripciones de texto generadas automáticamente. 3) Análisis espacial, analiza la presencia y el movimiento de personas en una fuente de video y genera eventos a los que pueden responder otros sistemas.

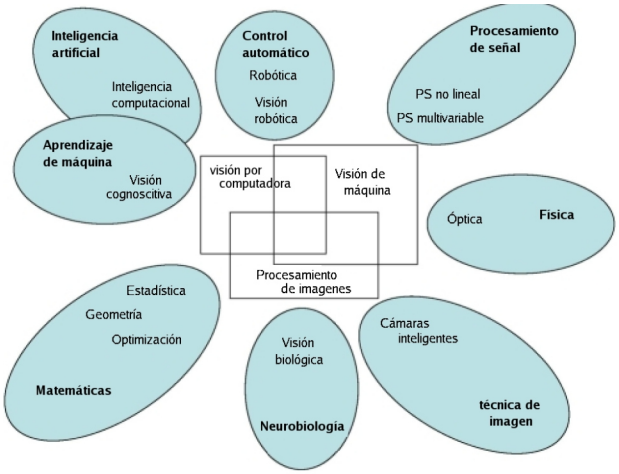


Figura 4: Relación entre la visión por ordenador y otras áreas afines.

IV. REDES NEURONALES CONVOLUCIONALES

Las redes neuronales convolucionales, comúnmente conocidas como CNNs (Convolutional Neural Networks) son un tipo específico de Red Neuronal (Feed Forward) basada en la organización de las neuronas de los cerebros biológicos. Su comienzo se enmarca en un artículo llamado “Gradient-based learning applied to document recognition” publicado en 1988 por Yann LeCun, el cual se basa asimismo en el artículo Neocognitron de Kunihiko Fukushima (1980).

Las redes convolucionales no dejan de ser redes neuronales, y por tanto incluyen una capa de entrada, una capa de salida, y diferentes capas ocultas. Su especialización radica en las capas ocultas, las cuales se relacionan con operaciones de **convolución**, **max-pooling** o **subsampling**, y capas totalmente conectadas (**fully-connected**).

En la Figura 5 se muestra la arquitectura de una red convolucional básica.

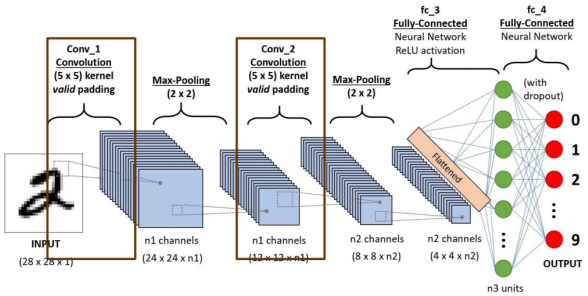


Figura 5: Arquitectura de una CNN Básica

Entre las arquitecturas más exitosas de CNNs tenemos AlexNet 2012. Esta red fue la encargada de poner el foco sobre el Deep Learning al ganar la competición ILSVRC. Introdujo algunos principios que se mantienen hoy en día (uso de ReLU) mientras otros han perdido interés (normalizaciones). Para su entrenamiento se hizo un uso efectivo de paralelismo, distribuyendo las operaciones entre dos CPUs.

- A. Krizhevsky, I. Sutskever, and G.E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. Advances in neural information processing systems. 2012

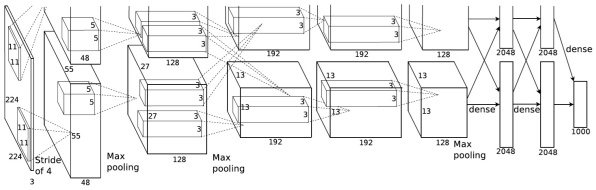


Figura 6: Arquitectura de AlexNet

A. Arquitectura de Red Neuronal Convolucional VGG

VGG es un modelo de Deep Learning presentado por Karen Simonyan y Andrew Zisserman en el artículo **Very deep convolutional networks for large-scale image recognition**. Se trata de un modelo basado en redes convolucionales de 16 o 19 capas. El modelo recibe como input una matriz de 224 x 224 x 3 a la que se le aplican un conjunto de capas ocultas que dan como resultado un vector de mil componentes con la probabilidad de que cada objeto este en la imagen. La arquitectura de la red se puede apreciar en la Figura 7:

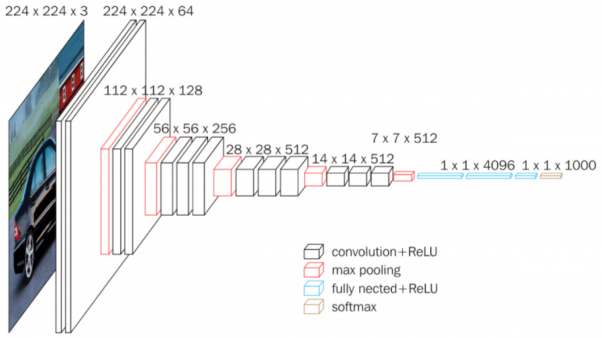


Figura 7: Arquitectura VGG

El modelo presentado por Karen y Andrew fue entrenado durante 2 semanas sobre Image-Net. Una iniciativa para proporcionar a investigadores de todo el mundo una base de datos de imágenes de fácil acceso. Actualmente cuenta con más de 14 millones de imágenes y más de 1 millón de imágenes anotadas.

V. LA ARQUITECTURA VGG 16 Y LA CLASIFICACION MULTICLASE

Para el presente trabajo se trabajó con el DataSet de imágenes de Kaggle (disponibles es el sitio web: https://www.kaggle.com) denominado **Intel Data Classification**, la cual contiene imágenes sobre escenas naturales alrededor del mundo. Las imágenes están clasificadas en 7 categorías:

['mountain', 'street', 'glacier', 'buildings', 'sea', 'forest']

Se utilizó una función para descargar las imágenes y sus respectivas etiquetas del portal de Kaggle. Es importante resaltar que se require registro prelimienar y generación de un API Token para poder descargar DataSets de Kaggle (revisar la sección videos).

Inicialmente se utilizó una **Red Neuronal Básica** a fin de considerarla como línea base:

```
model = tf.keras.Sequential([
    tf.keras.layers.Conv2D(32, (3, 3), activation = 'relu', input_shape = (150, 150, 3)),
    tf.keras.layers.MaxPooling2D(2,2),
    tf.keras.layers.Conv2D(32, (3, 3), activation = 'relu'),
    tf.keras.layers.MaxPooling2D(2,2),
    tf.keras.layers.Flatten(),
    tf.keras.layers.Dense(128, activation=tf.nn.relu),
    tf.keras.layers.Dense(6, activation=tf.nn.softmax)
])
```

Figura 8: Red Neuronal Básica

Luego se procedió a la extracción de características de la data de entrenamiento: train_features usando PCA:

```
from sklearn import decomposition
pca = decomposition.PCA(n_components = 2)
X = train_features.reshape((n_train, x*y*z))
pca.fit(X)
C = pca.transform(X)
C1 = C[:,0]
C2 = C[:,1]
```

Figura 9: Extracción de características usando PCA

Luego se procedió al entrenamiento de una red neuronal simple de una capa, en las características extraídas de VGG.

```
model2 = tf.keras.Sequential([
    tf.keras.layers.Flatten(input_shape = (x, y, z)),
    tf.keras.layers.Dense(50, activation=tf.nn.relu),
    tf.keras.layers.Dense(6, activation=tf.nn.softmax)
])
model2.compile(optimizer = 'adam', loss = 'sparse_categorical_crossentropy', metrics=['accuracy'])
history2 = model2.fit(train_features, train_labels, batch_size=128, epochs=500, validation_split = 0.2)
```

Figura 10: Entrenamiento

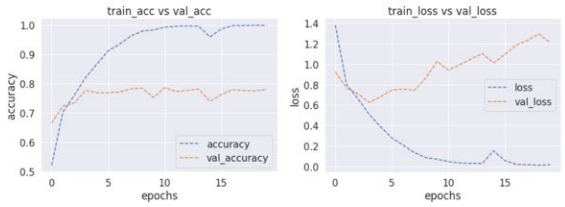
VI. RESULTADOS

Los resultados obtenidos se muestran en la Figura 11, donde se puede apreciar como mejoró el indicador de exactitud a partir del modelo básico de referencia.

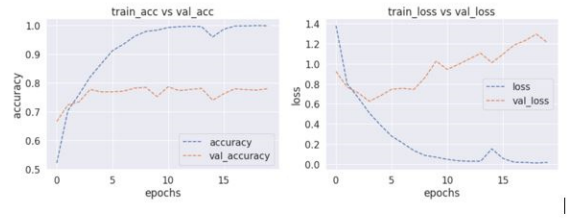
El accuracy del modelo mejoró desde 0.7693 hasta 0.8936, también se puede observer una mejora en el indicador de tiempo promedio por paso, de 6 ms/step a 4 ms/step.

El indicador loss, también mejoró pasando de 1.2475 a 1.2103.

CNN BASICO



VGG ImageNet



Fine Tunning VGG

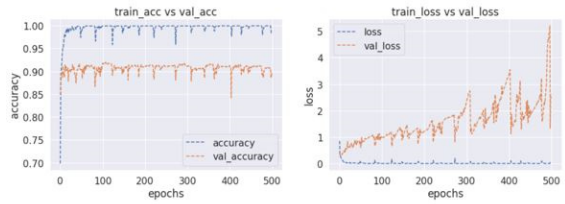


Figura 11: Resultados obtenidos

VII. CONCLUSIONES

Se observa que se logra una mejora en el modelo de clasificación luego de aplicar la arquitectura de red neuronal del tipo VGG asi como técnicas de afinamiento, pasando de un valor de accuracy de **0.7693** hasta **0.8936** tal como se puede apreciar en la Tabla 1.

MODELO/TECNICA	ACCURACY	LOSS	T
CNN BASICO	0.7693	1.2475	6 ms/step
VGG	0.8640	1.2629	2 ms/step
FINE TUNNING VGG	0.8936	1.2103	3 ms/step

Tabla 1: Tabla de resultados obtenidos

VIII. RECOMENDACIONES

El resultado obtenido es susceptible de mejora, se recomienda probar otros modelos de Redes Neuronales Convolucionales como: MobileNet, ResNet, SqueezeNet entre otras.

REFERENCES

[1] Rusell, Stuart, Norvig Peter, Artificial Intelligence a Modern Approach. 3ra ed., Pearson Global Edition, England, pp.2-3.

[2] Vannieuwenhuyze, Aurelien, Inteligencia Artificial. 1ra Ed. Ediciones Eni. 435 p. Barcelona España, 2020.

[3] Castillo-Cara, Manuel. Introducción al Deep Learning, UNI, Lima Perú, 2021.

[4] Lindsay, S. (2002). A tutorial on Principal Components Analysis. Disponible en: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf [Consulta: 22/01/2021]

PAGINAS WEB

Kaggle, https://es.wikipedia.org/wiki/Kaggle, [Consulta: 10/06/2021]

¿Qué es Computer Vision?, https://docs.microsoft.com/es-es/azure/cognitive-services/computer-vision/overview, [Consulta: 14/06/2021]

Visión Artificial, https://es.wikipedia.org/wiki/Visi%C3%B3n_artificial, [Consulta: 14/06/2021]

VIDEOS

¿Cómo importar Kaggle datasets en Google Colab?, https://www.youtube.com/watch?v=57N1g8k2Hwc