C&T Riqchary Revista Journal

Pedestrian identification through Deep Learning with classical and modern architecture of Convolutional Neural Networks

Erech Ordoñez, Ecler Mamani , Yonatan Mamani-Coaquira

Abstract— This article, refers to the research carried out at the National University Micaela Bastidas (UNAMBA), whose specific objectives were: To determine in a first stage of learning the proportion of accuracy of a classical architecture of Convolutionary Neural Network (CNN) in the identification of UNAMBA peoples, to determine in a second stage the proportion of precision in a modern architecture of RNC and finally compare the first stage with the second, to find the highest proportion. The training was given with a quantity of 242 people. Therefore, 27,996 images had to be generated through the technique of Video Scraping and data augmentation, which were divided into 19,700 images for training and 8,296 for the validation. Regarding the results in the first stage, a modified model VGG16-UNAMBA is proposed, with which a ratio of 0.9721 accuracy was achieved; while in the second stage it is proposed to DenseNet121-UNAMBA, with which a proportion of 0.9943 accuracy was achieved. Coming to the conclusion that the use of deep learning allows UNAMBA staff to be identified in a high proportion of accuracy.

Keywords *— recognition of people, convolutional neural network, deep learning*

1 INTRODUCCIÓN

Uno de los problemas más importantes en el reconocimiento automático de personas, es el nivel de precisión en la identificación en tiempo real, para lo cual es necesario generar y entrenar modelos que se adapten a los problemas específicos de reconocimiento en ambientes definidos y para propósitos generales; de ahí la importancia de investigar temas como el Deep Learning, el Machine Learning y la Inteligencia Artificial (IA). Estos conceptos, están íntimamente ligados, todos como sub partes de la Inteligencia Artificial[1].

Se define a la IA como el proceso de simulación de la inteligencia humana mediante máquinas y sistemas informáticos especiales, que incluyen el aprendizaje, razonamiento y auto corrección [2]. En general es un campo de investigación muy amplio, donde se incluyen la visión por computadora y muchos otros [3]. En [4] el campo de la IA se compone de varias áreas de estudio, dentro de las cuales está el Aprendizaje de las Máquinas y una de sus ramas son las Redes Neuronales Artificiales.

Dentro del Machine Learning existen tres enfoques para aprender, el aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por reforzamiento [5]. En esta investigación se trabajó con el enfoque supervisado y dentro de este existen aun dos tipos principales de problemas, denominados clasificación y regresión '[6]. Por los objetivos de la investigación se trabajó con problemas de clasificación; así el Deep Learning, es un conjunto de técnicas y procedimientos algorítmicos basados en Machine Learning para lograr que una máquina aprenda de la misma forma que lo haría un ser humano [7]. Las Redes Neuronales Convolucionales (RNC) son un tipo de modelo popular para el procesamiento imágenes [8]. Las RNC son una variante del perceptrón multicapa, con la diferencia que las RNC's realizan operaciones de "convolución" entre los parámetros de la red y los datos de entrada, en lugar de productos punto; el diseño de la RNC está inspirada en el área del cerebro (corteza visual) que procesa la información visual, y la corteza visual tiene varias especializaciones que le permiten procesar eficazmente los datos visuales; la corteza visual contiene muchas células receptoras que detectan la luz en regiones superpuestas del campo visual; asimismo, estas células receptoras están sujetas a la misma operación de convolución; es decir, todas procesan su entrada de la misma manera[9].

El simple hecho de enseñar a las computadoras a detectar personas en una imagen era un proceso muy difícil de realizar al no contar con buenos algoritmos y tener procesamiento limitado de las computadoras; sin embargo, a partir de las investigaciones de Viola Jones [10], las cosas mejoraron notablemente, en consideración a que propusieron un clasificador efectivo en la detección de rostros frontales en una imagen; sin embargo, aun no se había solucionado el problema de la detección en diferentes ángulos de las personas, hasta que otros investigadores –[11], –[12] mejoraron la detección de Viola Jones [10], ya que sus investigaciones aportaron a la detección de la identidad, sin tener problemas de los diferentes ángulos en las tomas; es así que con la intención de lograr mejoras, se encuentra el artículo "Center and Scale Prediction: A Box-free Approach for Object Detection" [13], en cuya investigación detecta el centro de objetos y a partir de allí se calcula el cuadro característico que encierra al objeto (persona) en una escala determinada, cuando se hace una búsqueda en una imagen.

En un ámbito más realista con la necesidad de detectar personas en tiempo real a través de cámaras de video, por parte de un robot en un contexto social, encontramos el artículo "Efficient and robust Pedestrian Detection using Deep Learning for Human-Aware Navigation"[14] en la cual utiliza el aprendizaje profundo para la identificación de humanos, previamente identifica objetos con el método ACF (Aggregate Channel Features) y luego los analiza en una Red Convolucional Profunda para finalmente detectar personas que transitan a su alrededor.Las experiencias de Machine Learning en el ámbito del reconocimiento de objetos y personas ya estaban definidas, sin embargo la tarea de identificar quien es dicha persona es una tarea mayor que[15] en el artículo "Joint Detection and Identification Feature Learning for Person Search", proponen la fusión de estos dos problemas: la detección y luego la identificación en un solo modelo de Red Neuronal Convolucional; sin embargo, para poder poner en práctica este modelo, se requería marcar las coordenadas de cada objeto (persona) en cada imagen y teniendo en cuenta que se tiene 27,996 imágenes para el proceso de aprendizaje de la RNC, sería una tarea muy difícil; por lo que se separo los problemas y se enfocó solo en la identificación de personas (quienes son).

Ahora dentro del ámbito del Aprendizaje Profundo, existen muchos estudios, así en [16] afirman que estudiar las

ISSN 2706-543X - Ing. Informática y Sistemas - UNAMBA

Redes Neuronales Artificiales (RNA) es uno de los más activos en la comunidad científica con múltiples aplicaciones recientes por lo que: [17], –[11], –[12], [14], [15], [18], [19], [20] y muchos otros usan el aprendizaje profundo para resolver diferentes problemas en cada una de las investigaciones a través de las redes neuronales convolucionales de manera exitosa.

La investigación es cuantitativa, por lo que su propósito es hallar el nivel más alto de proporción en la precisión de la identificación del personal docente y administrativo de la Universidad Nacional Micaela Bastidas de Apurímac (UNAMBA) usando el Aprendizaje Profundo; es decir, entrenando Redes Neuronales Convolucionales para dicha identificación en un modelo clásico como es VGG16 y un modelo moderno como es DenseNet121, ambos modelos adaptados y modificados al reconocimiento de personas en un ámbito real, uno con una arquitectura simple y el otro con una arquitectura compleja.

2 MATERIALES Y MÉTODOS

A. Diseño de la investigación

El diseño de la investigación fue cuasi experimental, ya que fueron necesarios varios experimentos para llegar a uno óptimo en cada etapa, considerando que se tuvo dos etapas principales, en la primera se trabajo con el modelo clásico VGG16 y en una segunda etapa se trabajo con DenseNet121 para obtener los mejores resultados en la identificación de 254 personas en cada etapa, como fines principales de la investigación.

B. Lugar de estudio

La investigación se realizó en los ambientes de la Universidad Nacional Micaela Bastidas de Apurímac, Ubicada en la región de Apurímac, provincia de Abancay. En [21], Apurímac es catalogada como una de las regiones más pobres del Perú; sin embargo la UNAMBA y el personal que labora ahí, intenta lograr investigaciones a la par que otros países, por lo que esta investigación es de mucha importancia para el cimiento de futuras investigaciones en el campo de la Inteligencia Artificial, Visión Artificial y específicamente en el Aprendizaje Profundo, dentro de esta área geográfica del Perú.

C. Descripción detallada por objetivos

Uno de los objetivos específicos fue determinar en una primera etapa de aprendizaje la proporción de precisión de una arquitectura clásica de RNC en la identificación del personal administrativo y docente de la UNAMBA. Para el entrenamiento de las RNC es necesario tener abundantes información, que en nuestro caso fue obtenida a través de una cámara IP de video vigilancia, de 2 Mega Pixels de resolución y a través de python y opencv se logro recolectar 27,996 imágenes (fotos) individuales (Figura 1), las cuales fueron organizadas en carpetas correspondientes a cada persona (clase), luego se reorganizo en dos carpetas más uno de entrenamiento y otro de validación con la misma estructura de nombres con la única variación de sus contenidos, las cuales sumados en total ascendían a 19,700 imágenes para el entrenamiento y 8,296 para la validación de los modelos (Figura 2). El modelo VGG16 [22] se modifico a un nuevo modelo de RNC (VGG16-UNAMBA) y con los datos obtenidos, se entrenaron ambos modelos en Google Colaboratory y Keras como librería de Aprendizaje Profundo. Colaboratory es un entorno de notebook de Jupyter gratuito que no requiere conFiguración y se ejecuta completamente en la nube, a su vez contempla procesamiento con GPU (Graphics Processing Unit).

Figura 1. Proceso de captura de imágenes

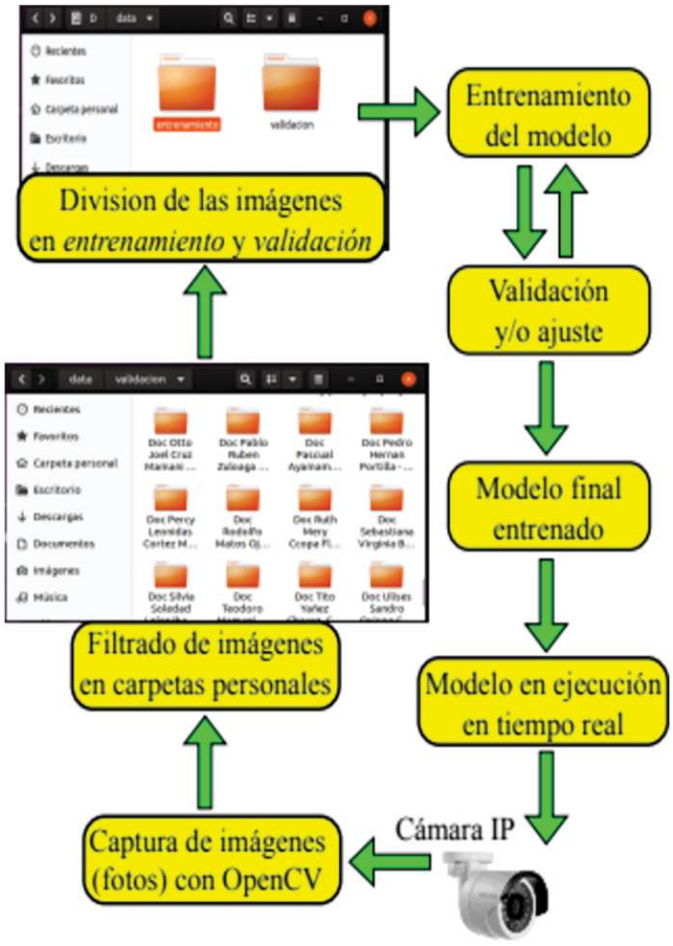

Figura 2. Proceso de modelamiento

El segundo objetivo de esta investigación fue determinar en una segunda etapa de aprendizaje la proporción de precisión con una arquitectura moderna de RNC en la identificación del personal administrativo y docente de la UNAMBA; para lo cual se utilizo el modelo DenseNet121 [23] y se planteo una arquitectura moderna de RNC (DenseNet121-UNAMBA), las mismas que fueron entrenadas con la misma cantidad de datos de la primera etapa, en el mismo entorno y con las mismas librerías.

Finalmente el tercer objetivo de esta investigación fue comparar el desempeño de cada modelo, desde los entrenamientos finales (Figura 3), hasta las proporciones halladas en la Tabla III, de los diferentes modelos entrenados con arquitecturas diferentes (clásica y moderna), para luego discutir y comentar los resultados finales.

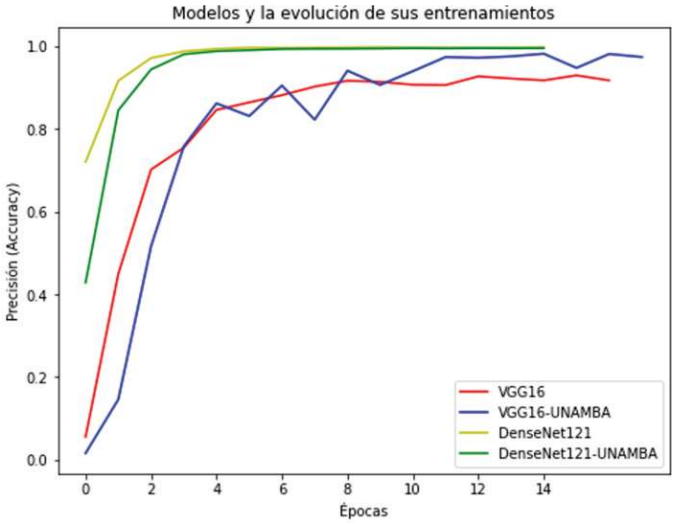

Figura 3. Resultado del proceso de entrenamiento de los modelos

La Figura 3 muestra de manera conjunta los resultados del proceso de entrenamiento de cada modelo y sus dificultades en la precisión de identificación de personas en cada época, hasta acercarse a 1.0, los modelos DenseNet son los más óptimos; sin embargo hay que considerar que acercarse demasiado a uno, no es lo óptimo, puesto que estaríamos hablando de un problema de sobre entrenamiento (overfitting), en donde el modelo puesto en ejecución, no podría tener el mismo performance, debido a que no generalizo correctamente y le sería difícil identificar a personas con diferentes atuendos o diferentes circunstancias en la que fue entrenado.

3 RESULTADOS Y DISCUSIÓN

A. Primera etapa

El primer objetivo específico, fue llegar a la proporción más alta, con el modelo de Red Neuronal Convolucional (RNC) con una arquitectura clásica VGG16, el cual se logró a través múltiples afinamientos y entrenamientos que se hicieron con los datos obtenidos por la cámara de video. El modelo resultante fue VGG16-UNAMBA (Figura 4).

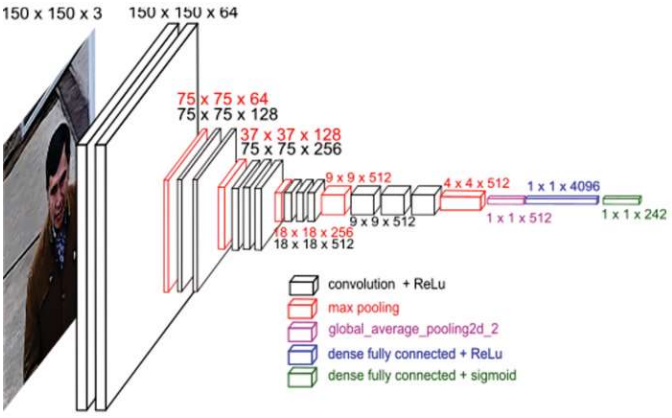

Figura 4. Modelo con arquitectura clásica VGG16-UNAMBA

La arquitectura clásica con la cual se trabajo VGG16 fue planteada por Simonyan y Zisserman [22], dicho modelo fue preparado para 242 clases y con tamaño de imágenes de 224 x 224 (longitud x altura) de entrada como lo especifican en sus artículos, este modelo al ser implementado tal cual, a excepción de la cantidad de clases (personas), generaron 135,252,018 parámetros entrenables (neuronas artificiales) y a través de los ajustes y experimentos, se logró llegar al modelo óptimo modificado VGG16-UNAMBA, el cual consta de los siguientes cambios: Tamaño de entrada de las imágenes a 150 x 150, se cambio la función Flatten() con la función GlobalAveragePooling2D(); así como también, se redujo los dos bloques del original, la capa fully connected (4,096) a uno solo y finalmente en la capa para la clasificación se cambio la función "softmax" por "sigmoid" y con estos cambios se generaron 17,807,410 parámetros entrenables, mucho menor a VGG16.

Evaluación de los modelos con arquitectura clásica

Los modelos VGG16 y VGG16-UNAMBA, fueron evaluados en 20 épocas y 1000 pasos por cada época; con 8,296 imágenes positivas y 240 imágenes negativas, se consideró una tasa mayor a 5% de probabilidad, para la identificación de una persona y menor o igual a esta, se considera no identificada. En el modelo VGG16 se introdujeron en total 8,536 imágenes como muestras, de los cuales clasificó correctamente 7,608; por lo cual se alcanzó una proporción de 0,8912. En el modelo VGG16-UNAMBA se ingreso igual número de imágenes (8,536), de las cuales clasificó correctamente 8,298 muestras; por lo cual alcanzó una proporción de 0,9721 de precisión en la identificación de las 242 personas entre administrativos y docentes de la UNAMBA (Tabla 1).

Los modelos de RNC VGG16 y VGG16-UNAMBA aprendieron a identificar a las 242 personas (clases) sin mayor problemas en coincidencia con [17] que sugiere que los entrenamientos convergen en menor cantidad de épocas si se cuenta con una gran cantidad de imágenes para su entrenamiento.

A. Segunda etapa

El segundo objetivo específico fue llagar a la proporción más alta en la identificación del personal administrativo y docente de la UNAMBA, con un modelo de Red Neuronal Convolucional (RNC) con arquitectura moderna como es DenseNet121, la cual fue ajustada y modificada en consideración a la inmensa cantidad de falsos negativos que se obtenían, es decir que de 240 imágenes que no correspondían a las 242 clases, sólo rechazaba correctamente 24 imágenes, el resto era identificada como falsos negativos (ver Tabla 2), por lo que se ajusto y modifico en múltiples afinamientos, hasta llegar al modelo DenseNet121-UNAMBA, el cual fue entrenado con el mismo grupo de imágenes que la etapa I.

TABLA 1. EVALUACIÓN DE LOS MODELOS CON ARQUITECTURA CLÁSICA.

Variables (Imágenes ingresadas al modelo)	Cantidad	Modelo: VGG16 Imágenes clasificadas correctamente	Modelo: VGG16-UNAMBA Imágenes clasificadas correctamente
verdaderos positivos	8,296	7,606	8,075
verdaderos negativos	240	2	223
TOTAL	8,536	7,608	8,298

ISSN 2706-543X - Ing. Informática y Sistemas - UNAMBA

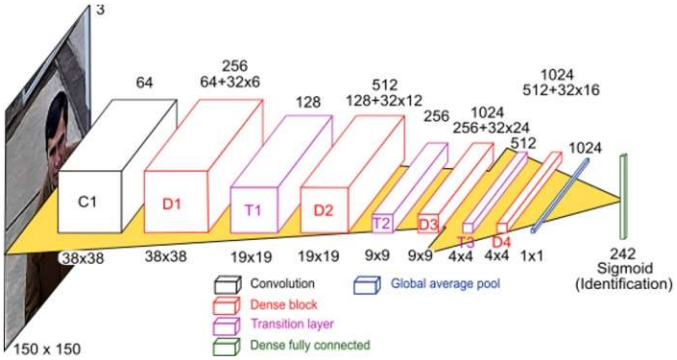

Figura 5. Modelo general con arquitectura moderna DenseNet121-UNAMBA

En la arquitectura moderna DenseNet121 [23], cada bloque se sub dividen en sub bloques que tienen una distribución de capas mucho más compleja que VGG16, por lo que su representación gráfica no es sencilla de mostrar a detalle; por lo cual, para mayor detalle referirse al artículo planteado por sus autores [23]; sin embargo, de manera general se puede apreciar la estructura general del modelo obtenido DenseNet121-UNAMBA (Figura 5). El modelo original DenseNet121 al ser preparado para 242 clases y con un tamaño de imágenes de 224 x 224 (longitud x altura) de entrada, se generaron 7,201,906 parámetros entrenables; sin embargo, a esta arquitectura se realizaron ajustes por la cantidad de falsos negativos descritas anteriormente, y esto ajustes fueron: Tamaño de entrada de las imágenes a 150 x 150 (longitud x altura) y la capa de salida (clasificación) donde se redujo el número de neuronas de 1000 a 242 neuronas y se cambio la función de activación por defecto de "softmax" a "sigmoid", al igual que el modelo I lográndose una disminución drástica de falsos negativos en las pruebas como se aprecia en la Tabla Tabla 2.

Evaluación de los modelos con arquitectura moderna

Los modelo DenseNet121 y DenseNet121-UNBAMBA, fueron evaluados en 15 épocas y 1000 pasos por cada época; con 8,296 imágenes positivas y 240 imágenes negativas, al igual que el modelo I, se consideró una tasa mayor a 5% de probabilidad, para la identificación de una persona y menor o igual a esta, se considera no identificada, con dichos parámetros se alcanzó una proporción de 0,9504 con DenseNet121 y 0,9943 de proporción con DenseNet121-UNAMBA en la precisión de identificación de las 242 personas entre administrativos y docentes de la UNAMBA (Tabla 3).

TABLA 2. EVALUACIÓN DE LOS MODELOS CON ARQUITECTURA MODERNA.

Variables (Imágenes ingresadas al modelo)	Cantidad	Modelo: DenseNet121 Imágenes clasificadas correctamente	Modelo: DenseNet121-UNAMBA Imágenes clasificadas correctamente
verdaderos positivos	8,296	8,140	8252
verdaderos negativos	240	24	235
TOTAL	8,536	8,164	8,487

Los modelos de RNC DenseNet121 y DenseNet121-UNAMBA aprendieron a identificar a las 242 personas (clases) un poco mejor que la arquitectura clásica en coincidencia a [23] que manifiestan que DenseNet tiene mejor desempeño que VGG16 y utiliza menos parámetros entrenables.

Tercera etapa

Como tercer objetivo específico se compara la proporción de la primera etapa con la segunda etapa, y vemos que las proporciones más altas corresponden a los dos modelos modificados VGG16-UNAMBA y DenseNet121-UNAMBA. Estadísticamente a un nivel de significancia de 0.01, se comprobó que las proporciones son diferentes; por lo cual, la proporción más alta de 0.9943 de precisión en la identificación del personal administrativo y docentes de la UNAMBA, corresponde al modelo DenseNet121-UNAMBA (Tabla N° 3).

TABLA 3. PROPORCIONES HALLADAS EN LOS DIFERENTES MODELOS DE RNC

Modelo	Tamaño de las imágenes de entrada	Total de imágenes ingresadas positivas y negativas	Total de imágenes clasificadas correctamente	Proporción de precisión en la identificación
VGG16	224 × 224	8,536	7,608	0,8919
VGG16-UNAMBA	150 × 150	8,536	8,298	0.9721
DenseNet121	224 × 224	8,536	8,164	0,9564
DenseNet121-UNAMBA	150 × 150	8,536	8,487	0,9943

La Tabla 3 resume todo el proceso de entrenamientos de los modelos con arquitectura clásica y moderna, los cuales se dieron de manera exitosa y en pocas épocas de 20 y 15 respectivamente en coincidencia con [17] que manifiesta que los entrenamientos convergen en menor cantidad de épocas, si se tiene gran cantidad de imágenes. A su vez que se logró obtener la proporción más alta entre ambas arquitecturas correspondiente a 0.9943, es decir que de 8,487 imágenes positivas y negativas, el modelo DenseNet121-UNAMBA, logra identificar correctamente 8,536; por lo cual la proporción correspondiente es muy alta y se corrobora con lo que [23] manifiestan de su modelo DenseNet, en cuanto a ser muy bueno en la identificación de imágenes; sin embargo la proporción 0.9721,obtenida por VGG16-UNAMBA es también alta.

4 CONCLUSIONES

En una primera etapa se implementó un modelo con arquitectura clásica VGG16-UNAMBA, con lo cual se logró obtener una proporción de 0.9721 de precisión en la identificación del personal administrativo y docente de la UNAMBA.

En una segunda etapa se implementó un modelo con arquitectura moderna DenseNet121-UNAMBA con lo cual se logró una proporción más alta de 0.9943 de precisión en la identificación del personal administrativo y docente de la UNAMBA.

Finalmente manifestar que se logró aplicar el Deep Learning a través de las Redes Neuronales Convolucionales de manera satisfactoria y se planteo dos modelos VGG16-UNAMBA y DenseNet121-UNAMBA, siendo ambas muy efectivas en la identificación de personas y resaltando que en ambos modelos se utilizó la función de activación "Sigmoid" como función final y generadora de la probabilidad de salida, ya que se demostró que no incurría en falsos negativos en gran medida a diferencia de "Softmax" que venía por defecto en los modelos originales.

5 AGRADECIMIENTOS

Agradezco a las autoridades de la UNAMBA por haber permitido usar los ambientes de la universidad y lograr que se concretice el trabajo de investigación.

Finalmente agradezco de manera especial a Jesús Utrera Burgal por sus artículos y publicaciones que fueron fuente de inspiración y admiración.

REFERENCIAS

[1] F. Chollet, Deep learning with Python. Shelter Island, NY: Manning, 2018.

[2] D. TutorialsPoint, «Tensor Flow, Simply Easy Learning», 2018. .

[3] S. J. Russell y P. Norvig, Inteligencia artificial: un enfoque moderno. Pearson Educación, 2008.

[4] P. Ponce, Inteligencia Artificial con aplicaciones a la ingeniería., Primera Edición. Ciudad de México: Alfaomega, 2010.

[5] J. Hurwitz y D. Kirsch, Machine Learning For Dummies. John Wiley & Sons, 2018.

[6] C. Andreas y G. Sarah, Introduction to Machine Learning with Python, Third Realese. United States of America: OReilly Media, 2017.

[7] B. García, «Implementación de Técnicas de Deep Learning», Trabajo de Fin de Grado, Universidad de la Laguna, La Laguna, 2015.

[8] N. Shukla, Machine Learning with TensorFlow. Manning Publications Co, 2017.

[9] J. Hearty, Advanced Machine Learning with Python, Primera edición. Reino Unido: Packt Publishing Ltd, 2016.

[10] P. Viola y M. J. Jones, «Robust Real-Time Face Detection», International Journal of Computer Vision, 2004.

[11] M. Mathias, R. Benenson, M. Pedersoli, y L. V. Gool, «Face Detection without Bells and Whistles», Springer International Publishing Switzerland, pp. 720735, 2014.

[12] D. Chen, S. Ren, Y. Wei, X. Cao, y J. Sun, «Joint Cascade Face Detection and Alignment», Computer Vision ECCV, pp. 109-122, 2014.

[13] W. Liu, S. Liao, y I. Hasan, «Center and Scale Prediction: A Box-free Approach for Object Detection», arXiv.org, abr. 23, 2019.

[14] A. Mateus, D. Ribeiro, P. Miraldo, y J. Nascimento, «Efficient and robust Pedestrian Detection using Deep Learning for Human-Aware Navigation», Robotics and Autonomous Systems, n.o 113, pp. 23-37, 2019.

[15] T. Xiao, S. Li, B. Wang, L. Lin, y X. Wang, «Joint Detection and Identification Feature Learning for Person Search», arXiv.org, n.o arXiv:1604.01850v3 [cs.CV], abr. 06, 2017.

[16] J. Rojas y R. Trujillo, «Algoritmo meta-heurístico Firefly aplicado al pre-entrenamiento de redes neuronales artificiales», Rev. Cuba. Cienc. Informáticas, vol. 12, n.o 1, pp. 14-27, mar. 2018, Accedido: oct. 13, 2018. [En línea].

[17] R. Vizcaya, «Deep Learning para la Detección de Peatones y Vehículos», Maestro en Ciencias de la Computación, Universidad Autónoma del Estado de México, México, 2018.

[18] K. Tsampikos, G. Triantafyllidis, y L. Nalpantidis, «Deep learning-based visual recognition of rumex for robotic precision farming», Computers and Electronics in Agriculture, vol. 165, pp. 85-90, 2019.

[19] Z. Kastrati, A. S. Imran, y A. Kurti, «Integrating word embeddings and document topics with deep learning in a video classification framework», Pattern Recognition Letters, vol. 128, pp. 85-92, 2019.

[20] L. Ye, L. Gao, R. Martinez, D. Mallants, y B. Bryan, «Projecting Australias forest cover dynamics and exploring influential factors using deep learning», Environmental Modelling & Software, vol. 119, pp. 407-417, 2019.

[21] INEI, «Evolución de la pobreza monetaria 2007 -2018», Instituto Nacional de Estadística e Informática, Perú, Informe Técnico, abr. 2019. [En línea]. Disponible en: https://bit.ly/2TKzaiF.

[22] K. Simonyan y A. Zisserman, «Very Deep Convolutional Networks for Large-Scale Image Recognition», ArXiv14091556 Cs, abr. 2015, Accedido: dic. 29, 2019. [En línea]. Disponible en: http://arxiv.org/abs/1409.1556.

[23] G. Huang, Z. Liu, L. Maaten, y K. Q. Weinberger, «Densely Connected Convolutional Networks», ArXiv160806993 Cs, ene. 2018, Accedido: dic. 29, 2019. [En línea].

ISSN 2706-543X - Ing. Informática y Sistemas - UNAMBA