Diseño de un Experimento: Comparación de Arquitecturas CNN en la Clasificación Industrial de Fresa Andaluza

Eduardo Reyero Ibáñez

I. Motivación del experimento

España es el segundo productor de fresa a nivel mundial, por detrás de Estados Unidos. Según refleja la ficha de producto elaborada por el Observatorio de Precios y Mercados de la Junta de Andalucía en 2018, en Andalucía se produce el 97% de la fresa española y el 26% de la europea [1].

De cara al mantenimiento de un gran volumen de producción en la industria alimenticia, las técnicas de visión artificial han cobrado una gran importancia ya que permiten reducir costes de operación, proveer mayor velocidad y precisión y mejorar el control de calidad del producto. Entre las ventajas de usar esta metodología no solo se libera a los trabajadores de un trabajo tedioso y repetitivo, sino que además amplia las posibilidades de operación como puede ser por ejemplo analizar imágenes fuera del espectro visible humano.

En cuanto a las frutas, y particularmente en las fresas, el aspecto visual cobra un papel determinante a la hora de la venta, y es por ello por lo que se hace necesario potentes sistemas de visión, garantizando de esta forma un producto más atractivo. Durante la inspección, es necesario identificar tres parámetros fundamentales: forma, tamaño y color. Una posible solución es la aplicación de Inteligencia Artificial para solventar el problema.

En el marco de este planteamiento, lo que se propone en este experimento es la comparación de diferentes arquitecturas de *Convolutional Neural Networks* (CNNs), las cuales son un tipo de aprendizaje profundo ampliamente utilizado para el tratamiento de imágenes. Concretamente el objetivo del experimento consistirá en la clasificación de las fresas en dos categorías: aptas y no aptas. Las arquitecturas que se evaluarán serán AlexNet, GoogLeNet y VGGNet y las métricas de comparación serán la *accuracy* y el tiempo de entrenamiento.

II. Descripción del equipo

Para la realización del experimento, el banco de ensayos, inspirado en [2] y representado en la Figura 1, se compone de:

- Lampara fluorescente con un difusor de celulosa, permitiendo obtener una iluminación de tipo difusa en la escena.
- Cámara web HD de marca Logitech y modelo pro C920, que permite la construcción de una base de datos de imágenes de fresas y además la adquisición de imágenes en tiempo real durante el proceso productivo.
- Ordenador personal para el almacenamiento de las imágenes, ejecución del algoritmo de clasificación de fresas y posterior visualización de los resultados. Concretamente, el modelo utilizado es un ordenador de marca Acer y modelo Aspire F15, que cuenta con un procesador Intel-Core i7, 16GB de memoria RAM y una tarjeta gráfica NVDIA GeForce 940MX con 2GB de RAM dedicada.

Figura 1: Banco de ensayos [2]

III. Adquisición de imágenes para la elaboración de la base de datos

Para la adquisición de imágenes se utilizará la webcam marca Logitech HD pro C920 situada a 25 cm sobre una base blanca conformada por una hoja de papel A4 de 75gr. La orientación del objetivo de la cámara será paralelo a la superficie del papel.

Debido a la utilización de algoritmos de aprendizaje para el experimento, será necesario conformar una base de datos con imágenes de las fresas utilizando el banco de ensayos descrito. La base de datos contará con 1000 imágenes, de las cuales 800 corresponderán a fresas etiquetadas como aptas y las 200 restantes estarán compuestas tanto de fresas fuera del estándar buscado como de otros objetos y serán etiquetadas como no aptas.

Las imágenes tendrán una resolución de 250x150 píxeles en formato JPG y el tamaño que ocupará la fresa dentro de la imagen será menor de 100x100 píxeles.

IV. Arquitecturas evaluadas

AlexNet

AlexNet [3] fue una arquitectura propuesta para la clasificación de imágenes en el LSVRC-2010. La arquitectura, representada en la Figura 2, cuenta con cinco capas convolucionales, algunas seguidas de capas max-pooling, y tres capas totalmente conectadas. La salida de la ultima capa totalmente conectada sirve de entrada a una función softmax. Después de cada capa convolucional se utiliza una función de activación ReLU. Además, AlexNet utiliza *data augmentation* y *dropout* con probabilidad de 0.5 en las dos primeras capas totalmente conectadas para reducir el sobreajuste.

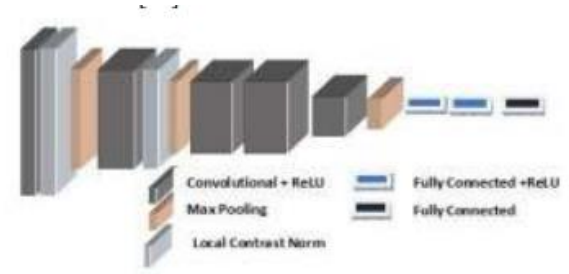

Figura 2: Arquitectura AlexNet [3]

GoogLeNet

GoogLeNet, representada en la Figura 3, es una arquitectura diseñada para la competición ILSCVR-2014 [4]. Para mejorar el funcionamiento de la red neuronal sin ser perjudicado en coste de computación por el aumento del tamaño de la red, la arquitectura utiliza capas escasamente conectadas. GoogLeNet utiliza 9 modulos de incio, consistentes en 22 capas profundas, que son seguidas de 5 capas *pooling*. Además, utiliza una capa *dropout* con una probabilidad de 0.7 y funciones de activación ReLU en todas las capas convolucionales, incluyendo los módulos.

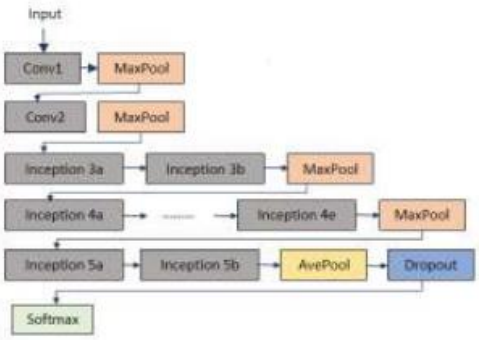

Figura 3: Arquitectura GoogLeNet [4]

VGGNet

VGGNet es una arquitectura de aprendizaje profundo propuesta para la competición ImageNet-2014 [5]. La arquitectura VGGNet, representada en la Figura 4, se trata de una mejora de AlexNet mediante la adición de más capas profundas, y aunque hay diferentes configuraciones en función del número de capas convolucionales, para este experimento se usarán 13 conjuntos de capas convolucionales con ReLU, seguidas de 3 capas totalmente conectadas y una capa con la función softmax al final. Además, algunas de las capas convolucionales son seguidas de capas max-pooling.

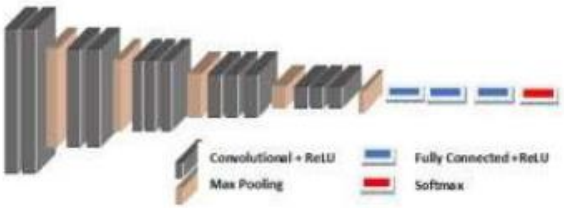

Figura 4: Arquitectura VGGNet [5]

V. Experimento

El flujo de tareas desarrollado en el experimento aparece reflejado en la Figura 5.

En primer lugar, se procederá a la creación de la base de datos de imágenes de fresas. Dicho proceso, descrito en la sección III del documento, producirá un *dataset* de 1000 imágenes para poder entrenar el modelo.

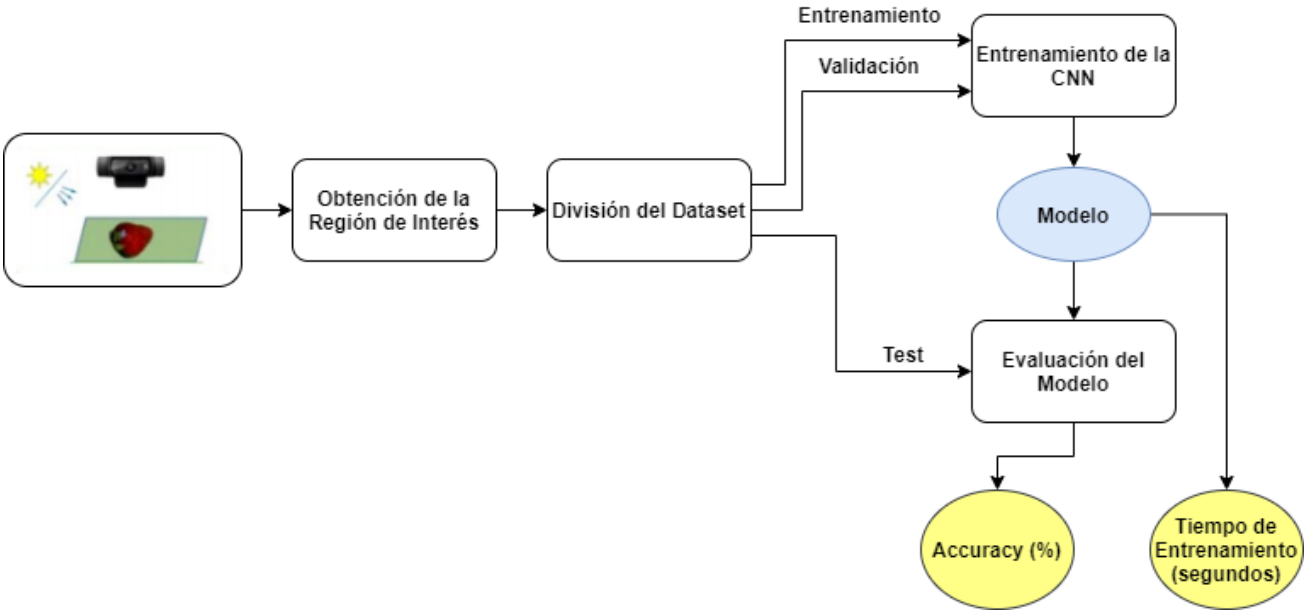

Figura 5: Proceso experimental

El segundo paso del experimento es la obtención de la región de interés de las imágenes. Para ello se llevará a cabo un proceso de segmentación de las imágenes mediante el algoritmo Canny. La región de interés obtenida tendrá unas dimensiones de 100x100 píxeles. Para la segmentación Canny se utiliza un filtrado gaussiano, intensidad de gradiente de la imagen y una relación de umbral superior a umbral inferior de 2.5 [6].

El tercer paso es la división del *dataset* en tres subconjuntos: entrenamiento (80%), validación (10%) y test (10%). Además, para esta división se llevará acabo un proceso de validación cruzada utilizando el método *k-fold* con una k igual a 5, de forma que cada arquitectura será entrenada y probada un total de cinco veces utilizando distintas combinaciones del *dataset* original.

El cuarto paso es el entrenamiento supervisado de cada arquitectura de CNN propuesta. Este paso del experimento se repetirá cinco veces para cada una de las arquitecturas elegidas, de acuerdo con el valor de k elegido en el método de validación cruzada anterior.

El quinto paso consistirá en la evaluación de los modelos entrenados utilizado los subconjuntos de test. La métrica elegida para la comparación de los modelos será la *accuracy*, además del tiempo de entrenamiento. La *accuracy* se calculará como:

$$accuracy\ (\%) = \frac{TP + TN}{TP + TN + FP + FN}$$

TP corresponde a las fresas aptas correctamente clasificadas, TN corresponde a las fresas no aptas correctamente clasificadas, FP corresponde a las fresas no aptas incorrectamente clasificadas como aptas y, por último, FN corresponde a las fresas aptas incorrectamente clasificadas como no aptas. Dichos términos serán extraídos de la matriz de confusión generada para cada prueba de test.

Por último, los resultados obtenidos se organizarán de forma que se puedan comparar las tres arquitecturas. Concretamente, los resultados se organizarán en dos tablas. La primera de ellas incluirá para cada arquitectura la *accuracy* máxima y mínima obtenida, así como la media de los cinco entrenamientos. En la segunda tabla se representarán para cada arquitectura los tiempos máximo y mínimo de entrenamiento, además de la media obtenida en los cinco entrenamientos.

VI. Referencias

[1] Hortoinfo (2018, Octubre 10) Diario Digital de Actualidad Hortofrutícula [Online]. Disponible en: http://www.hortoinfo.es/index.php/99-catotranot/7523-superf-fresa-281116

[2] P. Constante, A. Gordon, O. Chang, E. Pruna, F. Acuna and I. Escobar, "Artificial Vision

Techniques to Optimize Strawberry's Industrial Classification," in *IEEE Latin America Transactions*, vol. 14, no. 6, pp. 2576-2581, June 2016, doi: 10.1109/TLA.2016.7555221.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 1–9.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014

[6] OpenCV Dev Team, Canny Edge Detector, OpenCV 2.4.13.0 documentation. [Online]. Available: http://docs.opencv.org/2.4/doc/tutorials/imgproc/imgtrans/canny_detector/