Análisis de Grad-CAM++ y Score-CAM en modelos de clasificación basados en redes neuronales convolucionales

Eduardo Díaz Gaxiola
Posgrado en Ciencia de la Información
Universidad Autónoma de Sinaloa
Culiacán, México
eduardogaxiola@uas.edu.mx

Arturo Yee Rendón
Facultad de Informática Culiacán
Universidad Autónoma de Sinaloa
Culiacán, México
arturo.yee@uas.edu.mx

Inés Fernando Vega López
Parque de Innovación Tecnológica
Universidad Autónoma de Sinaloa
Culiacán, México
ifvega@uas.edu.mx

Gerardo Beltrán Gutiérrez
Facultad de Informática Culiacán
Universidad Autónoma de Sinaloa
Culiacán, México
gerardo@uas.edu.mx

***Abstract*— Las tareas de identificación y clasificación de objetos en imágenes digitales son cada vez más importantes en diversas áreas del conocimiento. Recientemente, se han aplicado diferentes técnicas computacionales en búsqueda de solucionar este tipo de problemas. Un ejemplo de ello son las Redes Neuronales Convolucionales (RNCs) que son una técnica de aprendizaje profundo para dar solución a problemas de clasificación de imágenes. Las RNCs extraen las características distintivas de los objetos a clasificar mediante convoluciones. En una primera instancia, no se conocen visualmente cuáles fueron las características distintivas extraídas por las RNCs, siendo éste un proceso de caja negra. El conocer las características distintivas extraídas por las RNCs nos ayudaría a entender las decisiones que toman los modelos para la clasificación. En este trabajo, presentamos un estudio comparativo del desempeño de diferentes métodos de visualización aplicados a modelos de clasificación basados en RNCs. Al usar los métodos de visualización, buscamos analizar el proceso de extracción de características para determinar si existe una relación entre el valor de cobertura del objeto de interés y la tasa de aciertos de los modelos de clasificación. Una vez entrenados los modelos, utilizamos como métricas de evaluación el valor de cobertura de los objetos de interés y la tasa de aciertos de los modelos. Como caso de estudio evaluamos las arquitecturas de RNCs VGG19 [1] y Xception [2] utilizando un conjunto de datos conformado por más de 10,000 imágenes digitales de flores de 100 especies de plantas de la flora de México. Por último, los resultados experimentales muestran que los modelos de clasificación obtienen un 0.85 de valor de cobertura y una tasa de aciertos de hasta 88.75% Top-5.**

Keywords— Clasificación de objetos, Flores, Métodos de visualización, Plantas, RNC.

I. INTRODUCCIÓN

La clasificación de objetos en imágenes digitales es una tarea rutinaria para los seres humanos, pero ha demostrado ser compleja para las máquinas. En el área de aprendizaje de máquina, la clasificación de objetos en imágenes digitales se ha convertido en uno de los principales retos. Las Redes Neuronales Convolucionales (RNCs) son técnicas de aprendizaje profundo que realizan el procesamiento de la información utilizando diferentes tipos de capas: convolucionales, reducción y completamente conectadas [5]. Las RNCs realizan dos procesos: el proceso de extracción de características y el proceso de clasificación. En el proceso de extracción de características se procesan un conjunto de imágenes de entrenamiento para aprender e identificar sus características distintivas. En el proceso de clasificación, se utilizan las características encontradas en el proceso de caracterización como datos de entrada a un clasificador basado en redes neuronales tradicionales. Una vez entrenado el modelo, se desconocen las áreas de la imagen utilizadas por el modelo durante el proceso de extracción de características. Esto genera un problema de falta de información al validar si el objeto de interés fue tomado en cuenta y sus características fueron extraídas para ser utilizadas en el proceso de clasificación. Una manera de esclarecer el proceso de caracterización es utilizar métodos de visualización. Estos métodos permiten conocer las áreas de la imagen utilizada durante la extracción de características. Los métodos de visualización se dividen en dependencia de su enfoque, es decir, la información que usan para realizar la visualización. Por ejemplo: enfocados en el valor de gradiente, enfocado en activación o enfocado en perturbación. El funcionamiento de los métodos de visualización se basa en que, dada una imagen y un enfoque, generan una imagen que representa las áreas de la imagen utilizadas por el modelo en el proceso de clasificación, generalmente en mapa de calor. En este trabajo se propone usar como métrica de evaluación del proceso de extracción de características el valor de cobertura de los objetos de interés.

Definimos nuestro problema formalmente de la siguiente forma. Sea X un conjunto de imágenes de entrenamiento. Cada imagen X_i perteneciente a X cuenta con un conjunto de objetos de interés O_j donde O representa el conjunto de objetos de interés de todos los elementos de X. Sea M un conjunto de matrices de información obtenidas del modelo de clasificación f denotado como $f : X \rightarrow M$. Sea V un conjunto de métodos de visualización, tal que $V : M \rightarrow H$ donde H denota el conjunto de métodos de mapas de calor de cada elemento del conjunto X. Nuestro problema se define formalmente como sigue:

$$\underset{v \in V}{argmax\ cover} : H, O \rightarrow \mathbb{R}$$

Encontrar el método de visualización $v \in V$ que maximice la función de cobertura y de como resultado un valor de cobertura.

El resto del documento se organiza de la siguiente forma. En la Sección II realizamos un estudio en el estado del arte sobre el uso de RNCs para la clasificación de objetos y del uso de los métodos de visualización en el entrenamiento de modelos basados en RNCs. En la Sección III describimos nuestras propuestas para evaluar el proceso de extracción de características de los modelos entrenados, las arquitecturas de RNCs que utilizamos y los métodos de visualización que aplicamos. En la Sección IV presentamos los resultados de este trabajo, describiendo el conjunto de datos utilizado y los experimentos realizados. Por último, en la Sección V presentamos nuestras conclusiones.

II. Estado del arte

Recientemente el aprendizaje profundo ha logrado un gran éxito para reconocer y clasificar objetos en imágenes digitales. Por ejemplo, las RNCs se han aplicado exitosamente a problemas de clasificación de especies de plantas, identificación de enfermedades de plantas, detección de vasos sanguíneos de la retina, entre otros. Particularmente, las RNCs han tenido resultados sobresalientes en las diferentes ediciones del desafío de reconocimiento visual a gran escala (o ILSVRC por sus siglas en inglés). Un ejemplo es la arquitectura AlexNet que en el 2012 obtuvo un 84.7% de tasa de aciertos en Top-5 [3] siendo la primera arquitectura de RNC en ganar el reto. En el 2013, Zeiler *et al.* [4] propusieron una variante de AlexNet llamada ZFNet, ganando el reto con un 88.8% de tasa de aciertos. En el 2014, el equipo de desarrollo de Google propone la arquitectura Inception [5], que obtuvo una tasa de aciertos de 93.33%. He *et al.* [6] en el 2015 proponen la arquitectura ResNet, ganando el reto con una tasa de aciertos de 96.43%. En el 2016, Xie *et al.* [7] realizan una adecuación a ResNet llamada ResNeXt, ganando el reto con un 95.9% de tasa de aciertos. En la última edición del reto en el 2017, la arquitectura SENet ganó el reto con 97.74% de tasa de aciertos Top-5 [8]. En el área de la botánica, se han utilizado las RNCs para la clasificación de especies de plantas y de órganos distintivos. En el 2016, Lee *et al.* [9] propusieron una RNC con una arquitectura secuencial de 16 capas, denominada VGG16 para el reto LifeCLEF 2016. Este reto consiste en identificar un conjunto de datos conformado por más de 113,000 imágenes correspondiente a 1,000 especies vegetales. Para este reto, los autores lograron una media de tasa de aciertos por encima del 70%. En el 2017 Barre *et al.* [10] diseñaron una arquitectura de RNC orientada a la clasificación de hojas, denominada LeafNet. Esta arquitectura demostró tener un desempeño superior a métodos tradicionales para la clasificación de imágenes en los conjuntos de datos de Foliage, LeafSnap y Swedish Leaf. En el 2018, Carpentier *et al.* [11] entrenaron una red con arquitectura ResNet para identificar especies de árboles nativos de Canadá a partir de imágenes de cortezas, logrando una tasa de aciertos de 97.81%.

Los métodos de visualización se han aplicado en distintos problemas de clasificación utilizando RNCs. En el 2019, Liu *et al.* [12] entrenaron modelos basados en las arquitecturas VGG16 y ResNet50 para clasificar entre distintos tipos de la especie de flor *Chrysanthemum morifolium Ramat.* Los autores utilizaron el método de visualización Grad-CAM para interpretar el comportamiento de los modelos. Al aplicar Grad-CAM, los autores se percataron que los modelos se enfocaban en los bordes y en el centro de la flor, prestando poca atención a las hojas o al fondo negro. En el 2020, Shukla *et al.* [13] detectaron distintas enfermedades en hojas de manzana con modelos entrenados utilizando Efficientnet y ResNet50. Para la visualización, utilizaron los métodos CAM y Grad-CAM. Los autores reportaron una tasa de aciertos de hasta 89%, con un enfoque de la visualización en el centro de las hojas de manzana. En el trabajo de Yebasse *et al.* [14] en el 2021, utilizaron los métodos de visualización Grad-CAM, Grad-CAM++ y Score-CAM en la clasificación de hojas de café. Los autores entrenaron modelos basados en ResNet y en Unet, obteniendo hasta un 98% de tasa de aciertos. Para este estudio, los métodos de visualización mostraron las diferencias entre los procesos de caracterización de ambas arquitecturas, en donde la arquitectura Unet se enfocó en las características distintivas de las hojas, por lo que obtuvo una mejor tasa de aciertos.

III. Materiales y métodos

En esta sección, describiremos la metodología de nuestra propuesta y los pasos a seguir para obtener el valor de cobertura del objeto de interés. Conocer el comportamiento durante el proceso de caracterización nos permitirá evaluar que tan confiable es la tasa de aciertos obtenida por nuestros modelos entrenados, además de validar si el entrenamiento de los modelos se realizó tomando en cuenta características relevantes en los objetos de interés. Partimos desde la idea de, mientras más se enfoque el modelo en las características de los objetos de interés durante la caracterización, más relevantes serán estas características para el modelo de clasificación y su porcentaje de tasa de aciertos será mayor. Para realizar el cálculo de la cobertura del objeto de interés, comparamos el mapa resultante de aplicar un método de visualización con respecto al área del objeto de interés.

Nuestra propuesta pretende analizar el problema del desconocimiento de las características distintivas extraídas por las RNCs. Esto con el propósito de entender las decisiones de los modelos de clasificación. Los pasos a seguir son los siguientes:

1. Pre-procesamiento de las imágenes: Como primer paso en nuestra metodología preprocesamos las imágenes del conjunto de datos obteniendo manualmente una máscara que indica el área del objeto de interés. Esta mascara es necesaria para realizar el cálculo de la cobertura del objeto de interés.
2. Aplicar método de visualización: En el siguiente paso se aplica un método de visualización a la imagen de entrada, generando un mapa de calor. Este mapa de calor se utiliza como una máscara, tomando en cuenta los valores que superaran un umbral del 45% del valor máximo posible.
3. Calcular el valor de cobertura del objeto: Por último, se realiza el cálculo del valor de cobertura del objeto de interés aplicando la máscara resultante del método de visualización sobre la máscara del área del objeto de interés, obteniendo un valor de la cobertura.

La metodología de este trabajo se muestra en la Fig 1.

A. Descripción de arquitecturas

En esta sección se describen las arquitecturas de RNC utilizadas para entrenar los modelos de clasificación. Estas arquitecturas cuentan con diferencias en cuanto a la cantidad

de capas convolucionales como en el número de parámetros que entrenan.

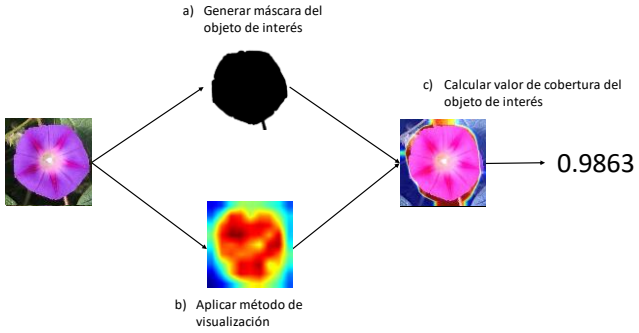

Fig. 1. Metodología seguida para el cálculo de cobertura del objeto de interés.

1) VGG

La arquitectura VGG fue propuesta por Simonyan *et al.* [1] en el 2014. Esta arquitectura se conforma principalmente de dos tipos de bloques: el bloque A y el bloque B. El bloque A se conforma de dos capas convolucionales seguidas de una capa de agrupación por valor máximo. Por su parte, el bloque B se conforma de cuatro capas convolucionales seguidas de una capa de agrupación por valor máximo. La estructura de VGG19 se basa en la siguiente combinación de dos bloques A seguidos de tres bloques B. Al final, se tiene tres capas completamente conectadas (densas), la última de ellas es una capa Softmax que da como salida 1000 valores. El 19 representa la cantidad de capas entrenables contenidas en la arquitectura: 16 capas convolucionales y 3 capas completamente conectadas, con un total de 143,700 parámetros a entrenar.

2) Xception

En el 2017, la arquitectura Xception surge como una adaptación de Inception propuesta por Chollet *et al.* [2]. La arquitectura Xception se basa en capas de convolución separables en profundidad. Las convoluciones separables en profundidad son diferentes a las convoluciones normales de manera en que, para un tensor de entrada (32, 32, 3) podemos usar cualquier cantidad de filtros convolucionales. Cada uno de estos filtros se ejecutará en los tres canales y la salida será la suma de todos los valores. Sin embargo, en las convoluciones separables en profundidad cada canal tiene un solo kernel de convolución. Por lo tanto, al realizar convoluciones separables en profundidad se puede reducir la complejidad computacional, ya que cada kernel es solo de dos dimensiones y es convolucional solo en un canal. Con esta propuesta la convolución separable en profundidad modificada es la convolución de un filtro de 1 x 1 seguida por una convolución en profundidad. Xception cuenta con 81 capas entrenables y un total de 22,900 parámetros

B. *Descripción de métodos de visualización*

Se describen a continuación los métodos de visualización aplicados a las arquitecturas de RNC. Estos métodos se diferencian principalmente en el enfoque de visualización que tienen. Un tipo de enfoque es la visualización basada en el valor de gradiente. Un enfoque distinto es la visualización basada en el valor de perturbación de la imagen.

1) CAM

El mapeo de activación de clases (CAM por sus siglas en inglés) es un método de visualización propuesto por Zhou *et al.* en el 2015 [15]. Este método utiliza la agrupación del promedio global (APG) de la RNC para generar un mapa de activación para una clase en particular. Dicho mapa indica las áreas de la imagen utilizadas por la RNC para identificar dicha clase. Para hacer uso de este método, es necesario utilizar una arquitectura de RNC que cuente con una capa APG antes de la salida de la capa final. Una vez aplicado APG en los mapas de características, estos son usados como entrada para una capa completamente conectada que producirá la salida deseada. Dada esta estructura, se puede identificar la importancia de las áreas de la imagen propagando hacia atrás los pesos de la capa de salida en los mapas de características.

2) Grad-CAM++

EL mapeo de activación de clases por gradiente ponderado (Grad-CAM++) es un método de visualización basado en CAM, propuesto por Chattopadhyay *et al.* en el 2018 [16]. El método CAM cuenta con limitaciones al momento de no poder aplicarse en arquitecturas sin módulos de APG y no ser capaz de visualizar de manera precisa múltiples objetos en una misma imagen. Para dar solución a estas limitaciones, se propuso el método Grad-CAM++. Este método define un valor de ponderación para cada mapa de características de la clase a visualizar, calculado a partir de los coeficientes de correlación para el valor de gradiente correspondiente a la clase a visualizar. Los coeficientes de correlación del gradiente toman en cuenta los valores de múltiples objetos de la clase a visualizar, obteniendo un mapa de calor que abarque los múltiples objetos en la imagen.

3) Score-CAM

El mapeo de activación de clases ponderado por puntaje (Score-CAM) es un método de visualización desarrollado por Wang *et al.* en el 2019 [17]. Los autores identificaron que al visualizar imágenes utilizando valor de gradientes, existe ruido en las imágenes que producen cambios de valores en el gradiente. Este método propone una mejora respecto a otros métodos basados en CAM, tratando de resolver los problemas de ruidos irrelevantes y generar visualizaciones más claras y limpias. Los autores proponen un método basado en perturbaciones que enmascara una parte de las áreas en imagen de entrada de la red y analizan cómo va cambiando el puntaje de la predicción. La máscara de activación obtenida es tratada como un tipo de máscara para la imagen de entrada, haciendo que el modelo prediga sobre la imagen parcialmente enmascarada. El puntaje en la clase a visualizar se utiliza para representar la importancia en el mapa de características.

C. *Métricas de evaluación*

En esta sección, se describen las métricas para evaluar la extracción de características de los modelos de clasificación, la tasa de acierto de los modelos de clasificación y el valor de correlación entre la extracción de características y la clasificación de los modelos. Para evaluar la extracción de características, proponemos un índice de evaluación basado en

el valor de cobertura del modelo con respecto al área del objeto de interés. En la evaluación de la clasificación de los modelos utilizamos como métrica de evaluación el porcentaje de tasa de aciertos. Por último, para evaluar la correlación entre la extracción de características y la clasificación del modelo utilizamos como métrica de evaluación el valor de correlación resultante de aplicar la función de Pearson.

1) Índice de evaluación de la extracción de características.

En esta sección, se describe el índice propuesto para medir la cobertura de los objetos de interés. Esto se hace con el propósito de evaluar el proceso de extracción de características de los modelos de clasificación. Se propuso este índice aplicando un método de visualización y calculando el valor que se cubre del objeto de interés en la imagen en un rango de 0-1. El índice se define con la siguiente ecuación:

$$I_c = \frac{C_o}{A_o}$$

Donde I_c es el índice de cobertura, C_o es área cubierta del objeto de interés y A_o es el área del objeto de interés.

2) Índice de evualuación de la clasificación del modelo.

Para esta métrica de evaluación, utilizamos la tasa de aciertos en la clasificación del modelo. Con esta métrica, valoramos el porcentaje de respuestas correctas con respecto al total de predicciones realizadas con un rango de 0-100%. La tasa de aciertos se define con la siguiente ecuación:

$$TA = \frac{VP + VN}{VP + VN + FP + FN}$$

Donde *TA* es la tasa de aciertos, VP son los verdaderos positivos, VN son los verdaderos negativos, FP son los falsos positivos y FN son los falsos negativos.

3) Índice de evualuación de correlación.

Utilizamos el índice de correlación de Pearson para calcular la correlación entre el valor de cobertura con respecto a la tasa de aciertos. Este índice mide que tan fuerte es la relación entre dos variables a calcular entre un rango de -1 y 1. Un valor cercano a 1 indica una correlación positiva, donde si una variable aumenta, la otra variable también aumenta. Un valor cercano a -1 indica una correlación negativa, donde si una variable aumenta, la otra variable disminuye. La correlación de Pearson se define con la siguiente ecuación:

$$C = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Donde *C* es el valor de correlación, n es el número de elementos en las variables a calcular la correlación, *x y* son las variables a calcular su correlación.

IV. Experimentación y discusión

Los experimentos fueron realizados en una computadora especializada con las siguientes características de hardware/software: sistema operativo Ubuntu 18.04, un procesador Intel Xeon W-2133 de 12 núcleos a 3.6GHz, 32 GB de RAM, 2 TB de almacenamiento tipo HDD interfaz SATA y una unidad de procesamiento de gráficos (GPU) NVIDIA GTX 1080 con 8GB de RAM.

A. Conjunto de datos

Se utilizó un conjunto de imágenes de plantas correspondiente a 100 especies de plantas nativas de México [18]. Cada clase cuenta con 100 imágenes recortadas de flores, contando el conjunto de datos con un total de 10,000 imágenes recortadas. Dentro de este conjunto de datos existen clases que cuentan con alta variabilidad intraclase, es decir, son ejemplares de la misma especie que lucen diferentes entre ellas. Podemos ver en la Fig. 2 un ejemplo de alta variabilidad intraclase entre la especie *Dahlia coccinea Cav. (1791)*, con dos diferentes imágenes de la misma especie que cuentan con diferencias como el color de la flor y forma de los pétalos.

Fig. 2. Comparación entre dos imágenes diferentes de la especie *Dahlia coccinea Cav. (1791)*

B. Conjunto de visualización

Adicional al conjunto de entrenamiento y de prueba utilizado para entrenar los modelos de clasificación, se generó un conjunto de datos para realizar los experimentos de visualización. Este conjunto de visualización fue conformado por una muestra de 10 especies de plantas elegidas aleatoriamente con 10 imágenes por especie, teniendo un total de 100 imágenes para visualizar. Para realizar el cálculo del valor de cobertura de los objetos, se realizó un preprocesamiento manual para una máscara del objeto de interés. Podemos observar en la Fig. 3 el resultado del preprocesamiento de las máscaras. En la parte superior se encuentran las imágenes de flores sin procesar y en la parte inferior se encuentran la máscara generada manualmente del objeto de interés.

C. Entrenamiento de modelos basados en RNC

Se entrenaron modelos de clasificación basados en las arquitecturas de RNC VGG19 y Xception con los siguientes parámetros de entrenamiento:

- 100 épocas, tamaño de lote de 8.
- Descenso estocástico de gradiente como función de optimización.
- Entropía cruzada binaria como función de pérdida.
- Tasa de aprendizaje de 0.01.

A continuación, se describen los resultados del entrenamiento de los modelos de clasificación, indicando el valor de cobertura del objeto de interés, la tasa de acierto Top-1 y Top-5 de los modelos de clasificación, así como el valor de correlación entre el método de visualización con respecto al valor de predicción del modelo. Se tomaron como índices de evaluación de los modelos la tasa de acierto Top-1 y Top-5, es decir, si la clase correcta fue el valor de predicción más alto o si se encontró dentro de los cinco valores de predicción más altos. Los modelos entrenados con VGG19 y Xception

obtuvieron un porcentaje de tasa de aciertos de 39.75% y 71.15% en Top-1 y 65.85% y 88.75% en Top-5. La Tabla I muestra los resultados del método Grad-CAM++, que obtuvo un .3227% de cobertura para VGG19 con un valor de correlación de 0.73 y .6087 de cobertura para Xception con un valor de correlación de 0.84.

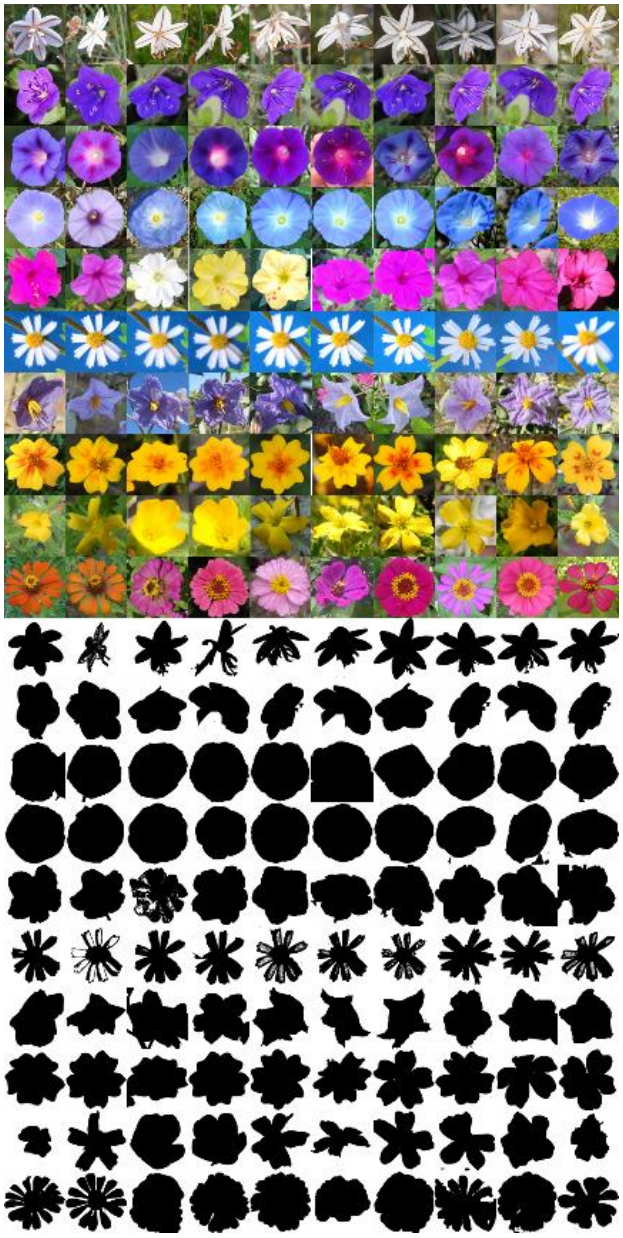

Fig. 3. Resultado del preprocesamiento de las imágenes de flores para la obtención manual de máscaras de los objetos de interés.

TABLE I. RESULTADOS ÍNDICE EN CONDICIONES FAVORABLES GRAD-CAM++.

Arquitectura	*Cobertura*	*% Top-1*	*% Top-5*	*Corr*
VGG 19	0.3227	39.75	65.85	0.73
Xception	0.6087	71.15	88.75	0.84

TABLE II. RESULTADOS ÍNDICE EN CONDICIONES FAVORABLES SCORE-CAM.

Arquitectura	*Cobertura*	*% Top-1*	*% Top-5*	*Corr*
VGG 19	0.3277	39.75	65.85	0.82
Xception	0.8550	71.15	88.75	0.90

La Tabla II muestra los resultados del método Score-CAM, que obtuvo un 0.3277 de cobertura para VGG19 con un valor de correlación de 0.82 y 0.8550 de cobertura para Xception con un valor de correlación de 0.90.

Realizamos un análisis visual del comportamiento de los modelos de clasificación basados en VGG19 y Xception sobre dos imágenes de prueba para demostrar el valor de cobertura de la visualización. Se observa en la Fig. 4 a), donde el valor de cobertura fue de 0.8745 con una predicción del modelo de 98.78% para la clase *Ipomea purpurea (L.) Roth (1787)*. Caso contrario, observamos en la Fig. 4 b), donde el valor de cobertura fue de 0.1656 para la clase *Solanum hindsianum Benth. (1844)* pero con una tasa de aciertos de 1.45%. Al observar los resultados, encontramos que los valores de correlación obtenidos indican una relación entre el valor de cobertura y la tasa de aciertos de la clasificación del modelo, donde a mayor valor de cobertura, mayor será la tasa de aciertos del modelo de clasificación.

a) *Ipomea purpurea (L.) Roth (1787).*

b) *Solanum hindsianum Benth. (1844).*

Fig. 4. Relación entre valor de cobertura y tasa de aciertos del modelo utilizando Score-CAM en la arquitectura Xception a) *Ipomea purpurea (L.) Roth (1787)* con un 0.8745 de cobertura y b) *Solanum hindsianum Benth. (1844)* con un 0.1656 de cobertura utilizando Score-CAM.

V. CONCLUSIONES

El observar las características distintivas de los objetos encontradas en el proceso de extracción de características nos permite conocer el comportamiento del modelo, y así interpretar los resultados en el proceso de clasificación. Además, concluimos que existe una relación entre el valor de cobertura del método de visualización con respecto a la tasa de aciertos del modelo de clasificación, donde a mayor valor de cobertura del objeto, mayor será el porcentaje de tasa de aciertos. Esto se ve reflejado en los resultados obtenidos por la arquitectura Xception, que fue la arquitectura con mayor

promedio de valor de cobertura del objeto y así mismo, el mayor porcentaje de tasa de aciertos en la clasificación. Esta relación se basa en que, si durante el proceso de extracción de características se utilizaron las características distintivas del objeto de interés contenido en la imagen, el modelo de clasificación se entrenará con las características del objeto de interés. Comparando los métodos de visualización, Score-CAM tuvo un mejor rendimiento al visualizar hasta un 0.8550 en comparación a Grad-CAM++ que alcanzó hasta un 0.6087. Concluimos que el rendimiento de Score-CAM se debe a que no utiliza el valor de gradiente, que puede llegar a causar cambios de valores al momento de realizar la visualización. El enfocarse en el objeto de interés durante la extracción de características de las RNCs ayudará a obtener tasas de aciertos más altas en la clasificación del modelo.

REFERENCES

[1] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classication Models and Saliency Maps. CoRR abs/1312.6034 (2014).

[2] Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, pp. 1800-1807 (2017). https://doi.org/10.1109/CVPR.2017.195.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems,2012, pp. 1097–1105.

[4] Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks CoRR, 2013, abs/1311.2901.

[5] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, pages 1-9, Massachusetts, USA, Oct. 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition ,pages 770_778, Nevada, USA, 2016.

[7] Xie, S.; Girshick, R. B.; Dollár, P.; Tu, Z. & He, K. Aggregated Residual Transformations for Deep Neural Networks CoRR, 2016, abs/1611.05431

[8] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

[9] S. H. Lee, Y. L. Chang, C. S. Chan, and P. Remagnino, "Plant identification system based on a convolutional neural network for the lifeclef 2016 plant classification task," in CLEF (Working Notes), 2016, pp. 502–510.

[10] P. Barré, B. C. St¨over, K. F. M¨uller, and V. Steinhage, "Leafnet: A computer vision system for automatic plant species identification," Ecological Informatics, vol. 40, pp. 50–56, 2017.

[11] M. Carpentier, P. Giguere, and J. Gaudreault, "Tree species identification from bark images using convolutional neural networks," arXiv preprint arXiv:1803.00949, 2018.

[12] Liu, Z., Wang, J., Tian, Y. et al. Deep learning for image-based large-flowered chrysanthemum cultivar recognition. Plant Methods 15, 146 (2019). https://doi.org/10.1186/s13007-019-0532-7

[13] Shukla, N., Palwe, S., Shubham, M. R., & Suri, A. Plant Disease Detection and Localization using GRADCAM. In: International Journal of Recent Technology and Engineering (IJRTE), (2020)

[14] Yebasse, M.; Shimelis, B.; Warku, H.; Ko, J.; Cheoi, K.J. Coffee Disease Visualization and Classification. Plants 2021, 10, 1257. https://doi.org/10.3390/plants10061257

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, pages 29212929, Nevada, USA, June 2016.

[16] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, Nevada, USA, pp. 839-847 (2018). https://doi.org/10.1109/WACV.2018.00097

[17] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P.,Hu, X.: Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, USA, pp. 111-119 (2020). https://doi.org/10.1109/CVPRW50498.2020.00020.

[18] J. A. Campos-Leal, A. Yee-Rendón and I. F. Vega-López, "Simplifying VGG-16 for Plant Species Identification," in IEEE Latin America Transactions, vol. 20, no. 11, pp. 2330-2338, Nov. 2022, doi: 10.1109/TLA.2022.9904757.