Clasificación de Imágenes de Paisajes utilizando Redes Neuronales Convolucionales

Instituto Tecnológico y de Estudios Superiores de Monterrey

Autor: Daniel Felipe Hurtado

Abstract

Este proyecto busca construir un modelo de red neuronal convencional (CNN) capaz de clasificar correctamente imágenes de paisajes. Se utilizó el dataset de *Intel Image Clasificación* de Kaggle, compuesto por aproximadamente 25000 imágenes de 150x150 píxeles distribuidos por seis categorías. La CNN utilizada, fue inspirada en el estado del arte *Analyzing Deep Learning Techniques in Natural Scene Image Classification.* Se obtuvo un resultado de un 88% sobre el conjunto de pruebas. Como referencia del desempeño, se utilizó el modelo pre entrenado VGG 16 con capas congeladas, obteniendo un resultado comparable del 87%.

Introducción

La clasificación de imágenes es una de las tareas fundamentales en el campo del *Deep Learning* y la visión por computadora. Esta técnica permite asignar etiquetas a imágenes con base en sus características visuales y ha demostrado ser especialmente eficaz cuando se emplean redes neuronales convolucionales (CNN), gracias a su capacidad de extraer patrones espaciales y jerárquicos automáticamente.

En este proyecto se aborda el problema de clasificar imágenes de paisajes en seis categorías: edificios, selva, glaciar, montaña, mar y calle. Para esto se utiliza el dataset de *Intel Image Classification,* disponible en Kaggle [1], que contiene alrededor de 25000 imágenes de tamaño 150x150 pixeles. El objetivo es diseñar y entrenar un modelo CNN desde cero que logre un rendimiento robusto en esta tarea de clasificación multiclase.

Además, con el fin de obtener una referencia (benchmark) con un modelo que se considera *Estado del Arte*, se emplea el modelo VGG 16 como referencia, siguiendo el enfoque presentado por Mayanja et al. [2].

Descripción del Dataset

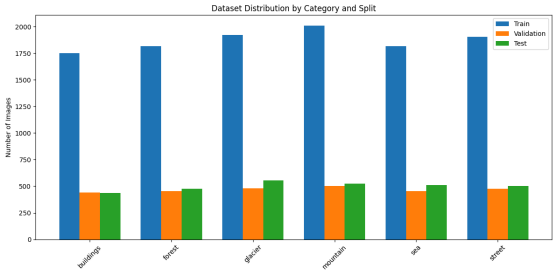
Para este proyecto se utilizó el *Intel Image Classification Dataset*, disponible públicamente en Kaggle. Este conjunto contiene aproximadamente 25000 imágenes a color con resolución de 150×150 píxeles, distribuidas equitativamente en seis clases de paisajes naturales:

- buildings
- forest
- glacier
- mountain
- sea
- street

La base de datos fue dividida manualmente en tres subconjuntos:

- **Entrenamiento (train - 60%)**: contiene la mayor parte de los datos para el aprendizaje del modelo.
- **Validación (valid - 20%)**: se usa para ajustar hiper parámetros y prevenir sobreajuste.
- **Prueba (test - 20%)**: se emplea exclusivamente para evaluar el rendimiento final del modelo.

A continuación, se muestra una visualización de algunas imágenes del dataset, junto con sus etiquetas, utilizando *matplotlib*. Esta inspección visual permitió verificar la calidad de las imágenes, así como la representación balanceada entre clases.



Metodología

El proceso de implementación del modelo de clasificación se desarrolló bajo un enfoque estructurado de tipo ETL (*Extract, Transform, Load*). En la fase de extracción, se seleccionó el conjunto de datos *Intel Image Classification Dataset* desde la plataforma Kaggle, validando su integridad y almacenándolo localmente. La estructura del proyecto siguió una organización jerárquica basada en subdirectorios por clase dentro de las carpetas */train*, */valid* y */test*, permitiendo cargar automáticamente las imágenes mediante herramientas de *TensorFlow*.

El entorno de desarrollo fue configurado utilizando *Python* y *JupyterNotebook* con las siguientes librerías principales: *TensorFlow* y *Keras* para modelado, *NumPy* para operaciones numéricas y análisis de etiquetas, *MatplotLib* para visualización y *Scikit-Learn* para la evaluación cuantitativa del rendimiento del modelo.

En la fase de transformación, se aplicaron operaciones de preprocesamiento sobre las imágenes. Las imágenes fueron redimensionadas a una resolución uniforme de 150 × 150 píxeles y normalizadas dividiendo los valores de píxeles entre 255, con el fin de escalar la imagen en un rango entre 0 y 1.

Para el conjunto de entrenamiento se implementó una estrategia de aumento de datos (*data augmentation*) con el objetivo de mejorar la generalización del modelo. Esta incluyó rotación aleatoria de hasta 10 grados, desplazamientos horizontales y verticales del 10%, zoom aleatorio, transformación de cizalladura y volteo horizontal. Estas técnicas se aplicaron utilizando la clase *ImageDataGenerator* de *Keras*.

El conjunto de validación y prueba solo fue sujeto a normalización, asegurando que las imágenes se mantuvieran sin alteraciones para una evaluación objetiva del rendimiento del modelo.

El modelo propuesto se basa en una red neuronal convolucional (CNN) diseñada específicamente para la clasificación de escenas naturales. La arquitectura fue desarrollada tomando como referencia el estudio de Kanavos et al. [3], que destaca la importancia de estructuras bien balanceadas que combinen eficiencia computacional con alta capacidad de generalización.

La red consta de cinco bloques convolucionales, cada uno compuesto por una capa *Conv2D* con activación ReLU seguida de una operación de *max pooling*. La cantidad de filtros se incrementa progresivamente (16, 32, 64, 64, 128) para capturar características visuales de baja y alta complejidad (bordes, texturas, formas).

La salida del modelo está conformada por una capa *Dense* con activación *softmax*, la cual contiene seis neuronas, correspondientes a las clases del conjunto de datos: edificios, selva, glaciar, montaña, mar y calle. Esta configuración permite al modelo generar una distribución de probabilidad sobre las clases y tomar una decisión de clasificación basada en la probabilidad más alta.

A diferencia de otras arquitecturas analizadas en el estudio citado, que incorporan

componentes como *batch normalization* la arquitectura presentada busca un equilibrio entre simplicidad, eficiencia y desempeño. Esta elección permite evaluar el rendimiento de una CNN desde cero sin depender de modelos pre entrenados ni de recursos computacionales especializados.

El modelo fue compilado utilizando la función de pérdida *categorical_crossentropy*, adecuada para clasificación multiclase con codificación one-hot. Como optimizador se seleccionó *Adam* con una tasa de aprendizaje inicial de 0.001, dada su capacidad de adaptación del gradiente durante el entrenamiento.

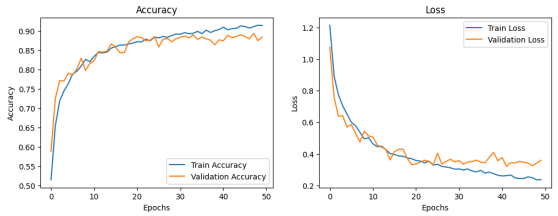
Se entrenó el modelo durante 50 épocas con un tamaño de lote (*batch size*) de 64 imágenes. El rendimiento se validó de forma continua contra el conjunto de validación para detectar sobreajuste y evaluar la progresión del aprendizaje. No se observaron señales significativas de sobreentrenamiento, y el modelo mostró una curva de aprendizaje estable a lo largo de las épocas.

Resultados

Tras el entrenamiento del modelo CNN propuesto durante 50 épocas, se evaluó su desempeño sobre el conjunto de prueba, compuesto por 3,000 imágenes distribuidas equitativamente en seis clases. Los resultados obtenidos reflejan un rendimiento robusto y generalizable en la tarea de clasificación de paisajes naturales.

Se registraron las curvas de precisión (*accuracy*) y pérdida (*loss*) tanto en el conjunto de entrenamiento como en el de validación. Estas gráficas muestran una tendencia de mejora progresiva en ambas métricas, sin indicios severos de sobreajuste. La estabilidad observada sugiere que el modelo fue capaz de generalizar adecuadamente sobre datos no vistos.

La ligera diferencia entre la curva de entrenamiento y la de validación es esperada y refleja un buen equilibrio entre complejidad del modelo y tamaño del dataset.



La precisión global del modelo alcanzó un 88%, lo cual indica que el 88% de las imágenes en el conjunto de prueba fueron clasificadas correctamente. Además, se calcularon las métricas promedio de precisión (precision), exhaustividad (recall) y puntaje F1 (f1-score), todas con un valor promedio también cercano al 88%. Estas métricas se calcularon utilizando las siguientes fórmulas estándar en clasificación multiclase:

- **Precisión (Precision)**: mide la proporción de verdaderos positivos sobre el total de elementos predichos como positivos:

$$Precision = \frac{TP}{TP+FP}$$

- **Exhaustividad (Recall)**: mide la proporción de verdaderos positivos sobre el total de elementos que realmente pertenecen a la clase

$$Recall = \frac{TP}{TP+FN}$$

- **F1-score**: es la media armónica entre precisión y recall, útil cuando existe un desequilibrio en las clases:

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Donde:

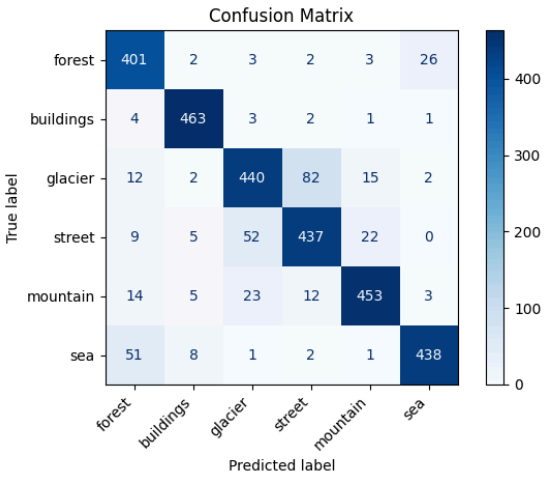
- **TP**: verdaderos positivos,
- **FP**: falsos positivos,
- **FN**: falsos negativos.

A continuación, se presenta un resumen del *classification report* para el modelo CNN:

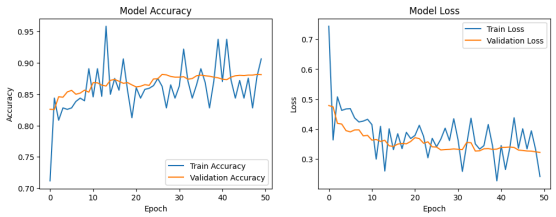
Clase	Precisión	Recall	F1-score	Soporte
Forest	0.82	0.92	0.86	437
Buildings	0.95	0.98	0.97	474
Glacier	0.84	0.80	0.82	553
Street	0.81	0.83	0.82	525
Mountain	0.92	0.89	0.90	510
Sea	0.93	0.87	0.90	501

La matriz de confusión correspondiente al modelo CNN muestra un desempeño equilibrado entre las clases, con una alta proporción de aciertos en categorías como *buildings* y *sea*. Sin embargo, se observa cierta confusión entre clases visualmente similares como *glacier* y *mountain*, lo que sugiere que estos casos podrían beneficiarse de resoluciones más altas o preprocesamiento más específico.

Esta visualización permite identificar los tipos de errores más frecuentes, ofreciendo una herramienta diagnóstica complementaria al *classification report*.



Para validar la solidez del enfoque implementado, se utilizó el modelo pre entrenado VGG 16 como referencia (*benchmark*). El modelo fue evaluado bajo las mismas condiciones (dataset, partición y tamaño de entrada), conservando todas sus capas convolucionales congeladas. El rendimiento de VGG 16 fue similar al de la CNN personalizada, con una precisión global del 87%, aunque con ligeras variaciones por clase.



Clase	Precisión	Recall	F1-score	Soporte
Forest	0.89	0.89	0.89	437
Buildings	0.98	0.97	0.98	474
Glacier	0.78	0.87	0.82	553
Street	0.87	0.75	0.81	525
Mountain	0.88	0.86	0.87	510

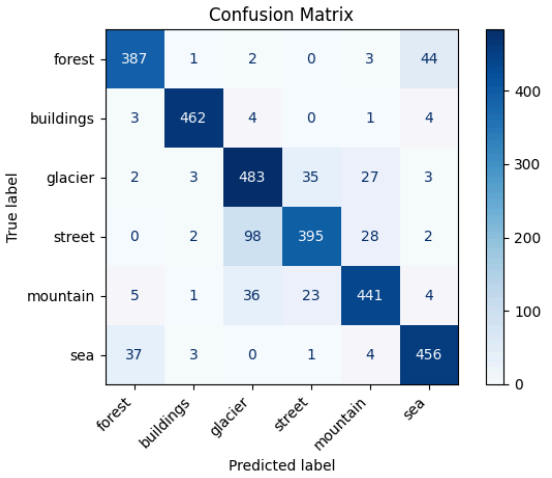
Sea	0.89	0.91	0.90	501

La matriz de confusión correspondiente al modelo VGG 16 revela un rendimiento notable en clases como *buildings* (462 aciertos), *glacier* (483) y *sea* (456), donde el modelo logra una alta tasa de predicciones correctas. Estos resultados reflejan la capacidad del modelo pre entrenado para reconocer patrones visuales distintivos y bien definidos en escenas urbanas o naturales consistentes.

No obstante, se observan errores relevantes en ciertas clases, como por ejemplo:

- **forest**: 44 imágenes fueron clasificadas erróneamente como *sea*, lo que puede atribuirse a similitudes en tonalidades.
- **street**: 98 imágenes fueron clasificadas como *glacier*, lo cual sugiere que el modelo presenta dificultades reconociendo patrones complejos.
- **mountain**: presenta dispersión de errores hacia *glacier* y *street*, lo que refuerza la idea de que algunas características visuales entre estas clases no son suficientemente diferenciadas.

Estos errores ponen en evidencia una limitación del enfoque de VGG 16 cuando se utiliza sin fine-tuning, es decir, sin permitir que sus capas convolucionales se ajusten al dominio específico del dataset. Aunque el modelo logra una alta precisión general, no alcanza una adaptación completa a las características particulares de las escenas naturales en este conjunto.



Ambos modelos mostraron un comportamiento competitivo y estable, lo cual demuestra que una arquitectura CNN construida desde cero, bien diseñada y entrenada adecuadamente, puede alcanzar niveles de rendimiento cercanos a los de modelos pre entrenados. No obstante, se observó que la red VGG 16, aunque robusta, mostró menor rendimiento en la clase *street*, mientras que la CNN personalizada mantuvo un desempeño más equilibrado entre clases.

El análisis también incluyó la matriz de confusión, la cual mostró un bajo nivel de error entre clases visualmente similares, como *glacier* y *sea*. Este tipo de errores es común en tareas de clasificación de escenas naturales, donde los paisajes pueden compartir patrones de color y textura similares.

Discusión

Los resultados obtenidos reflejan que el modelo CNN diseñado desde cero logró un rendimiento competitivo en la tarea de clasificación de imágenes de paisajes, alcanzando un 88% de precisión sobre el conjunto de prueba. Este desempeño es comparable al de VGG 16, un modelo pre entrenado ampliamente utilizado como referencia, que obtuvo un 87% en las mismas condiciones.

Uno de los aspectos más destacados del modelo propuesto fue su comportamiento equilibrado entre clases, con valores consistentes de *precision* y *f1-score*, particularmente en categorías como *buildings*, *mountain* y *sea*. Sin embargo, también se identificaron desafíos específicos en clases como *glacier* y *street*, donde el modelo cometió errores atribuibles a similitudes visuales con otras categorías. Estas confusiones son comunes en escenas naturales, especialmente cuando el tamaño de imagen (150x150 píxeles) limita el nivel de detalle disponible para la red.

Desde el punto de vista arquitectónico, la profundidad progresiva de filtros en los cinco bloques convolucionales permitió al modelo extraer representaciones jerárquicas efectivas sin incurrir en un sobreajuste evidente. El uso de *dropout* contribuyó a la regularización, evitando dependencia excesiva de rutas específicas en la red. Asimismo, el diseño compacto del modelo resultó eficaz para entrenarse con recursos computacionales moderados.

En contraste, VGG 16 ofreció un rendimiento competitivo pero sin ajuste fino (fine-tuning). Su ventaja radica en la transferencia de características visuales generales aprendidas a partir de un conjunto amplio y variado de imágenes (ImageNet), lo cual le permitió generalizar adecuadamente al problema, aunque sin adaptarse de forma específica al dominio de paisajes naturales. Esto refuerza la idea, discutida en estudios recientes [3], de que los modelos personalizados bien estructurados pueden igualar o superar a arquitecturas preentrenadas cuando se dispone de un dataset balanceado y suficientemente amplio.

Finalmente, la revisión de métricas por clase y la matriz de confusión sugieren que existe margen de mejora en la discriminación entre categorías similares, lo que podría abordarse en trabajos futuros mediante técnicas como *batch normalization*, mayor resolución de entrada o el uso de mecanismos de atención.

Trabajo Futuro

Si bien los resultados obtenidos con la arquitectura propuesta fueron satisfactorios, existen diversas líneas de mejora que podrían explorarse en investigaciones futuras para optimizar el rendimiento del modelo y su capacidad de generalización.

Una de las principales extensiones consiste en aplicar técnicas de fine-tuning sobre modelos pre entrenados como VGG 16, descongelando parcialmente las últimas capas convolucionales para permitir un ajuste fino de los pesos en función del dominio específico del dataset. Esta estrategia podría mejorar la discriminación entre clases visualmente similares, aprovechando el conocimiento previo del modelo y adaptándolo al contexto particular de paisajes naturales.

Otra mejora relevante sería el uso de imágenes con mayor resolución (por ejemplo, 224×224 píxeles o más), lo cual permitiría capturar detalles visuales más finos como texturas, bordes complejos o elementos específicos de la escena. Esta mayor riqueza de información visual podría traducirse en representaciones más discriminativas y un aumento en la precisión general del modelo.

Asimismo, podrían incorporarse técnicas de preprocesamiento más avanzadas orientadas a mejorar la calidad de entrada. Por ejemplo, el uso de métodos de realce de contraste, filtrado de bordes, reducción de ruido o transformaciones en espacios de color alternativos (como HSV o YUV) permitiría resaltar características visuales particulares de cada clase, facilitando su aprendizaje por parte de la red.

También es recomendable curar de manera más rigurosa el dataset, revisando posibles inconsistencias en la calidad o el etiquetado de las imágenes, eliminando ejemplos

redundantes o poco representativos, y asegurando una distribución equilibrada entre clases. Una curación más fina del conjunto de datos no solo mejoraría el entrenamiento, sino que también facilitaría una evaluación más precisa del modelo.

Estas posibles mejoras abren el camino para construir modelos aún más robustos y especializados, capaces de abordar con mayor precisión los retos de la clasificación de escenas naturales en entornos reales.

Conclusiones

Este trabajo presentó el diseño e implementación de una red neuronal convolucional (CNN) personalizada para la clasificación de imágenes de paisajes naturales. A través del uso del dataset *Intel Image Classification*, se construyó un modelo desde cero que demostró ser efectivo, alcanzando una precisión del 88% en el conjunto de prueba.

La arquitectura propuesta, basada en cinco bloques convolucionales con aumento progresivo de filtros, acompañada de técnicas de regularización como *dropout*, mostró un desempeño robusto y equilibrado entre clases. Este resultado evidencia que, aun sin recurrir a modelos pre entrenados ni técnicas avanzadas de ajuste, es posible lograr modelos competitivos mediante un diseño arquitectónico adecuado y una estrategia de entrenamiento consistente.

Como referencia, se utilizó el modelo pre entrenado VGG16 con capas congeladas, que alcanzó un rendimiento similar. Esta comparación permitió validar la capacidad de generalización de la CNN personalizada, destacando su eficiencia y simplicidad computacional.

El proyecto también permitió identificar áreas de mejora, como el tratamiento de clases con alta similitud visual, la posible limitación de la resolución de entrada y la necesidad de un preprocesamiento más detallado. Estos aspectos abren oportunidades claras para futuras investigaciones enfocadas en optimización de arquitectura, curación de datos y técnicas de aprendizaje profundo más sofisticadas.

En conjunto, los resultados obtenidos respaldan la viabilidad de desarrollar soluciones efectivas en clasificación de imágenes naturales con arquitecturas CNN propias, sentando una base sólida para proyectos más complejos o especializados en visión por computadora.

References

[1] P. Jindal, "Intel Image Classification," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/puneet6060/intel-image-classification

[2] A. Mayanja, I. A. Özkan, and Ş. Taşdemir, "Utilizing Transfer Learning on Landscape Image Classification Using the VGG16 Model," in *Proc. of the 11th Int. Conf. on Advanced Technologies (ICAT'23)*, Istanbul, Türkiye, 2023, pp. 71–76.

[3] A. Kanavos, O. Papadimitriou, K. Al-Hussaeni, I. Karamitsos and M. Maragoudakis, "Analyzing Deep Learning Techniques in Natural Scene Image Classification," in *Proc. 2024 IEEE Int. Conf. on Big Data (BigData)*, Washington, DC, USA, 2024, pp. 5682–5691. doi: 10.1109/BigData62323.2024.10824948.