18-5-2025

Leslie del Carmen Sanchez Reyes

Tec de monterrey campus querétaro

Primer reporte

Generación o selección del set de datos y Preprocesado de los datos

Contenido

[Obtención y Preprocesamiento del Dataset 2](#_Toc198502622)

[División de los datos: Entrenamiento y Prueba 2](#_Toc198502623)

[Referencias 3](#_Toc198502624)

Clasificación de Pokémon mediante Aprendizaje Supervisado con Imágenes

# Obtención y Preprocesamiento del Dataset

Para construir el dataset se utilizó como base el conjunto de datos “**7,000 Labeled Pokémon**” disponible públicamente en Kaggle [1]. Este dataset contiene imágenes etiquetadas y centradas de Pokémon de la primera generación. Contiene 150 folders cada uno con entre 25 a 50 imágenes por Pokémon.

Para este proyecto se seleccionaron solo 8 clases de Pokémon, con un total de 3500 imágenes:

* Abra (351 imágenes)
* Aerodactyl (540 imágenes)
* Arbok (508 imágenes)
* Articuno (595 imágenes)
* Bulbasaur (392 imágenes)
* Cubone (387 imágenes)
* Diglett (361 imágenes)
* Ditto (366 imágenes)

Como el dataset original contaba con una cantidad muy pequeña de imágenes por Pokémon, para aumentar la cantidad de muestras y la variabilidad del dataset, se empleó un script existente “downloadImg” [2] para descargar imágenes adicionales desde Google y Pinterest, incrementando así la diversidad de las imágenes. Lo anterior, dio lugar a que cada carpeta contara con aproximadamente 800 imágenes. Sin embargo, no todas eran realmente útiles.

Por lo tanto, una vez descargadas las imágenes con ayuda del script anteriormente mencionado, se hizo una depuración manual, evaluando que las imágenes obtenidas fueran de acuerdo con lo buscado y en su defecto eliminarlas o recortarlas para enfocarse en el objetivo. Esta depuración se llevó a cabo entrando a cada una de las carpetas, y seleccionando manualmente aquellas donde no apareciera el Pokémon requerido y se recortaron aquellas donde el Pokémon estaba acompañado de otros elementos.

Logotipo

El contenido generado por IA puede ser incorrecto.

Por ejemplo, en esta imagen que corresponde a la clase de Aerodactyl, están seleccionadas algunas imágenes que no son de valor y que fueron eliminadas.

El proceso anterior, se ejecutó múltiples veces hasta que llegamos a un dataset balanceado.

# División de los datos: Entrenamiento y Prueba

Los datos fueron divididos en dos subconjuntos:

* **80% para entrenamiento.** 2796 imágenes.
* **20% para prueba.** 704 imágenes

Esta distribución fue basada en la división típica o sugerida en otros proyectos.

# Referencias

[1] L. Tian, "7,000 Labeled Pokemon," Kaggle, [Online]. Available: <https://www.kaggle.com/datasets/lantian773030/pokemonclassification>  
[2] O. del Valle Mejía, *QuintupletRecognizer: downloadImg.py*, GitHub repository. [Online]. Available: <https://github.com/OsvalDev/QuintupletRecognizer/tree/master/scripts>