DEEP LEARNING BASED IMAGE DEDUPLICATION FOR EFFICIENT STORAGE

ABSTRACT

The rapid increase in digital images has caused a significant rise in the need for data storage. Repetitive images worsen this problem, especially in advanced applications with large datasets. This research aims to create effective image deduplication strategies to enhance storage efficiency. The research also investigates how expanding neural network designs could improve deduplication efficiency.

Three separate models have been suggested for detecting and removing duplicate images. The models are created to utilize developments in neural network technology to enhance the effectiveness of image deduplication operations. The research intends to enhance the accurateness and speed of duplicate image identification and removal by using neural networks to optimize storage. Also it demonstrates the effectiveness of increasing the complexity of neural networks in improving image deduplication efficiency through testing and analysis. It determine the most effective methods for managing duplicate images in storage systems by rigorously assessing several neural network designs. Also this research enhance image deduplication algorithms through thorough experimentation and analysis to optimize storage resources.

The first model applies a novel technique called Deep Learning Image Deduplication via Feature Extraction, which finds duplicate images inside a

dataset by using cosine similarity to evaluate image similarity and a pretrained deep Convolutional Neural Network (CNN) for feature extraction. The procedure includes transforming images into arrays for easier processing and then extracting features using a pre-trained deep Convolutional Neural Network (CNN). The extracted characteristics are compared using cosine similarity to measure the image's similarity. Images are classified as duplicate or non-duplicate based on the cosine similarity score using a predetermined threshold. Evaluating the approach involves implementing it with several pretrained CNN models, including VGG16, VGG19, and ResNet50. The experimental results confirm the efficacy of the Deep dup approach in precisely detecting duplicate images in datasets.

The second Model utilizes Deep Convolutional Generative Adversarial Networks (DCGANs) to identify and remove identical images in storage environments efficiently. The methodology involves training a DCGAN to create artificial images, which will be analyzed using similarity measurement methods to detect duplicates. Redundant images are removed through a rigorous deduplication procedure to optimize storage space. This technology provides a new solution to image duplication in storage, showing potential for improving storage economy and data management in substantial image collections.

The third Model presents a novel method to identify and delete duplicate images in storage systems. The methodology includes a series of steps, starting with data

pre-processing methods and then creating synthetic images using a Deep Convolutional Generative Adversarial Network (DCGAN). An enhanced dataset is created by combining natural and synthetic images, which is then used to fine-tune a ResNet50 model. The refined Model is used to carry out the deduplication process, efficiently detecting duplicate images in the dataset. This holistic method greatly enhances image deduplication precision and storage effectiveness. This research has promising implications for improving data management procedures in situations where there are widespread, large-scale image repositories.

These three research models offer a comprehensive framework for eliminating duplicate images in storage systems by utilizing a blend of deep learning models. The integration of these models is intended to improve accuracy of 96.5, maximize storage efficiency, and potentially save operational costs by decreasing the computing time with 24 seconds. The results highlight the flexibility and potential to grow the suggested methods, providing helpful knowledge for improving data management procedures in image repositories.