Information Problem Essay

Shreyas Kalvankar *Master of Information Management and Systems applicant*

In my senior year of engineering, I undertook the task of developing a system for the automatic colorization of black-and-white astronomical images as part of my bachelor's thesis. The motivation behind this endeavor stemmed from the realization that space archives contain an extensive collection of grayscale images that may never be processed due to the overwhelming volume of data present. Streamlining image processing with automated colorization and super-resolution can significantly aid astronomers in analyzing vast datasets, which are continuously expanding with the introduction of new telescopes. Upon commencing work on this project, I encountered a significant data challenge. While automated colorization has been explored in-depth, no dataset existed specifically for astronomical images. Given the nature of the problem, our focus was on generating colorized images using Generative Adversarial Networks, with data sourced from the Hubble Legacy Archive (HLA). The Hubble Legacy archive is slow and produces grainy images with lots of noise and majority are unprocessed and unfit to be used for training. This challenge presented an interesting problem to solve: filtering the data to extract useful information. Upon investigation, it was discovered that the archive contained processed images of the M101 galaxy, making them ideal for training purposes. We decided to scrape these using a web scraper and ended up with 80,000 images. Subsequently, we encountered another set of challenges: a significant portion of these images depicted empty space with minimal color information and, while not identical, bore similarities to many other images in the subset. Manual sorting through approximately 80,000 images was impractical. The solution was twofold: first, eliminate images predominantly consisting of black pixels, which was done using statistical methods, and second, employ a similarity metric to filter out similar images from the subset. This involved comparing the embeddings of images obtained from a pre-trained neural network with other embeddings.

Undoubtedly, this presented one of the most practical data-related challenges I have ever encountered. In contrast to my previous projects, which utilized curated datasets, this project necessitated the creation of a dataset from the ground up. It involved a meticulous search for databases, the extraction of valuable information, and the curation of a dataset specifically tailored to address the unique requirements of the problem. This process underscores the inherent necessity of dataset creation and extracting information when working on novel research problems.