Data Collection and Preprocessing Phase

Date	11 November 2024
Team ID	739849
Project Title	Pixelprose - crafting visual stories with intelligent image captioning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	Preprocessing involves preparing image datasets for deep learning models by ensuring consistency and quality. Key steps include resizing images to a fixed dimension, normalizing pixel values for efficient training, and applying data augmentation techniques like rotation and scaling to increase dataset diversity. Additional steps such as denoising, edge detection, and color space conversion help enhance feature clarity and relevance. Cropping focuses on regions of interest, and batch normalization stabilizes training by standardizing layer activations. These techniques ensure a robust dataset for effective model training.
Data Collection Plan	The data collection process focuses on acquiring diverse and representative datasets to support preprocessing and model training. Open datasets like MS COCO and ImageNet, custom data from sensors or cameras, and user-provided images serve as primary sources. Ethical considerations, proper permissions, and secure storage protocols ensure data integrity and compliance. This approach ensures the dataset is reliable and well-suited for subsequent machine learning tasks.

Raw Data Sources Identified	• **Open Datasets**: Publicly available resources like MS COCO, ImageNet, and Flickr30k provide annotated image collections suitable for training. • **Custom Data**: Domain-specific images collected using cameras, sensors, or web scraping tools to meet project requirements. • **User-Provided Data**: Data contributed by users or clients, ensuring compliance with permissions and privacy regulations.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset-1	A large-scale dataset containing images	https://www.kaggle.com/datasets/adityajn105/flickr8k	JPEG	1.03 GB	Public
Dataset 2:	A dataset containing captions related to images	https://www.kaggle.com/datasets/adityajn105/flickr8k	TEXT DOCUMENT	1.05 GB	Public