Data Collection and Preprocessing Phase

|  |  |
| --- | --- |
| Date | 20 October 2024 |
| Team ID | 739734 |
| Project Title | Ai-powered vehicle damage assessment and cost estimation for insurance claims |
| Maximum Marks | 6 Marks |

**Preprocessing Template**

The preprocessing stage involves collecting and processing vehicle damage images and corresponding claims data. This includes data cleaning, image resizing and normalization, and annotation of damage locations and types. Additionally, data augmentation techniques will be applied to increase dataset diversity and reduce overfitting.

The preprocessed data will be split into training, validation, and testing sets to support the development and evaluation of the AI-powered damage assessment and cost estimation model.

|  |  |
| --- | --- |
| **Section** | **Description** |
| Data Overview | The data for AI-powered vehicle damage assessment and cost estimation consists of approximately 100,000+ vehicle damage images and 500,000+ corresponding claims records, sourced from insurance company records and public databases. |
| Resizing | Resizing images to a uniform size (e.g., 512x512 pixels) to facilitate model training and improve efficiency.  - Maintaining aspect ratio to preserve image integrity and prevent distortion.  - Using interpolation techniques (e.g., bilinear, bicubic) to minimize image degradation during resizing. |
| Normalization | * Convert all text to lowercase. * Remove unwanted characters such as URLs, HTML tags, and special characters. |
| Data Augmentation | * Generate synthetic data by:   + Replacing words with synonyms.   + Back-translation. * Example:   + Original: "You are terrible." |

|  |  |
| --- | --- |
|  | * Synonym Replacement: "You are awful." * Back-translation (via Spanish): "You are horrible." |
| Denoising | * Remove stopwords (e.g., "is", "and", "the"). * Example: * Original: "This is an offensive comment." * Denoised: "offensive comment" |
| Edge Detection | * Adaptation: Extract key phrases or n-grams from text. * Example: * Original: "I hate you, you are useless." * Key Phrases: ["hate you", "are useless"] |
| Color Space Conversion | * Convert sentences to embeddings (e.g., Word2Vec, GloVe, or BERT embeddings). |
| Image Cropping | * Adaptation: Truncate text to relevant portions, e.g., first 50 words. * Example: * Original: "This is a very lengthy comment that exceeds the limit." * Cropped: "This is a very lengthy comment." |
| Batch Normalization | * Normalize word frequencies in text data (e.g., TF-IDF). * Example: * Comment: "This is toxic toxic toxic." * After Normalization: ["toxic": 3/6, "this": 1/6, "is": 1/6]. |
| **Data Preprocessing Code Screenshots** | |
| Loading Data |  |

|  |  |
| --- | --- |
| Data Description & Null values |  |
| Coreleation between variables |  |
|  |  |

|  |  |
| --- | --- |
| Data Preprocessing |  |
| clean the comment \_text in both the datasets. & training and testing |  |
| Vectorize the data |  |
| Train \_test\_ split& Transform the data& saving word vectorizer |  |

|  |  |
| --- | --- |
| Loading the pickle file |  |