White Report: Processing Physical Document Images for Text Extraction

## Introduction

This document consolidates four weekly work status reports focused on understanding and improving various training models related to Generative Artificial Intelligence (GenAI). The reports detail the tasks planned, challenges faced, and progress made in enhancing model accuracy and prediction capabilities. The aim is to provide a comprehensive overview of the ongoing efforts and methodologies employed to achieve higher accuracy and efficiency in GenAI projects.

## Overview of Tasks and Approaches

## Common Objectives

* **Understanding Learning Models**: Exploration of different learning models related to GenAI to identify the best approaches for training data and maximizing prediction accuracy.
* **Training Data Strategies**: Investigation of various strategies to improve the accuracy of models and their training stages.
* **Dataset Preparation**: Selection and preparation of datasets consisting of training, testing, and validation sets, primarily in the form of images.

## Techniques and Tools

* **Model Exploration**: Examination of models such as Tesseract, OpenCV (CV2), and EasyOCR for better prediction outputs.
* **Data Preprocessing**: Initial processing of data, including removing unwanted elements and lines from images, to prepare it for training.
* **Training Methodologies**: Evaluation of different training methods to find those that avoid overfitting and underfitting while ensuring high accuracy in predictions.

#### Hardware, Software Requirements

* GPU-enabled system for faster OCR processing
* High-resolution scanner for capturing document images
* Python 3.x
* OpenCV
* EasyOCR
* NumPy
* Matplotlib (for debugging and visualization)
* File handling libraries (os)

# Detailed Weekly Summaries

### Week 1: Initial Exploration and Preprocessing

**Tasks in Bucket**

* Understand various GenAI learning models.
* List approaches for training data to achieve maximum prediction accuracy.
* Improve the accuracy of models through better training stages.

### Planned Tasks

* Understand training strategies and their availability across models.
* Perform initial preprocessing of data before training.

### End of Week Status

* Finalized a training model that offers the highest accuracy while avoiding overfitting and underfitting.
* Continued efforts to improve predictions and accuracy in image text extraction.

## Summary of Work

* Explored various training models and their features, such as preprocessing, text extraction, overfitting, and underfitting.
* Faced challenges in understanding training methods and their applicability across open-source models.

#### Week 2: Enhancing Prediction Models

### Tasks in Bucket

* Understand different approaches to solving accuracy and prediction issues.
* Enlist available training models for better prediction.
* Choose datasets consisting of train, test, and validation sets in the form of images.

### Planned Tasks

* Understand models like Tesseract, OpenCV, and EasyOCR for better prediction outputs.
* Create an appropriate dataset for training the model to improve accuracy.

### End of Week Status

* Finalized a training method that ensures high prediction accuracy and correct data placement.
* Ongoing efforts to improve predictions by removing unwanted elements and lines from images.

### Summary of Work

* Explored training methods and GenAI models, evaluating their pros and cons.
* Faced challenges in achieving high accuracy and prediction from input to output images, ensuring correct word placement.

#### Week 3: Applying Advanced Techniques

### Tasks in Bucket

* Understand different approaches to solving accuracy and prediction issues.
* Enlist available training models for better prediction.
* Choose datasets consisting of train, test, and validation sets in the form of images.

### Planned Tasks

* Use models like Tesseract, OpenCV, and EasyOCR for better prediction outputs.
* Develop code to remove backgrounds and extract data from input images.

### End of Week Status

* Finalized a training model that provides greater accuracy and prediction.
* Ongoing efforts to improve coding for better model prediction and accuracy.

## Summary of Work

* Explored the use of OpenCV in real-life problems related to the project and applications of EasyOCR and Tesseract.
* Encountered challenges in maintaining consistent accuracy across different input images, affecting model performance.

#### Week 4: Refining and Finalizing Models

### Tasks in Bucket

* Further refine models to improve prediction accuracy and consistency.
* Enhance preprocessing techniques to ensure higher quality input data.
* Test models on diverse datasets to validate performance.

### Planned Tasks

* Implement advanced preprocessing techniques to improve data quality.
* Fine-tune model parameters to enhance prediction accuracy.
* Validate models on new datasets to ensure generalizability.

### End of Week Status

* Implemented advanced preprocessing techniques that significantly improved data quality.
* Fine-tuned model parameters, resulting in enhanced prediction accuracy.
* Validated models on new datasets, demonstrating good generalizability and performance.

### Summary of Work

* Achieved significant improvements in preprocessing and model accuracy.
* Successfully validated models on diverse datasets, ensuring robust performance.
* Overcame challenges related to data quality and model generalizability.

### Challenges Faced

* **Understanding Training Methods**: Difficulty in grasping the various training methods and their availability across different models.
* **Maintaining Accuracy:** Challenges in achieving and maintaining high accuracy in predictions, especially when dealing with diverse input images.
* **Preprocessing Issues:** Problems related to preprocessing, such as removing unwanted elements from images and ensuring consistent quality of input data.

#### Progress and Achievements

* **Model Selection:** Successfully identified and finalized training models that provide the highest accuracy while avoiding common pitfalls like overfitting and underfitting.
* **Improved Predictions:** Significant strides made in improving prediction accuracy and ensuring correct data placement within images.
* **Code Development:** Development of code to enhance preprocessing and prediction capabilities, particularly in handling background removal and data extraction from images.

### Conclusion

The consolidated efforts over these four weeks reflect a systematic approach to enhancing GenAI models, focusing on preprocessing, training, and improving prediction accuracy. Despite the challenges faced, substantial progress has been made in understanding and applying different models and techniques. The ongoing efforts aim to achieve robust and high-accuracy models that can be effectively used in various applications, providing valuable insights and solutions in the field of GenAI.