Channel Islands
CALIFORNIA STATE UNIVERSITY

OCR to Text Summary System

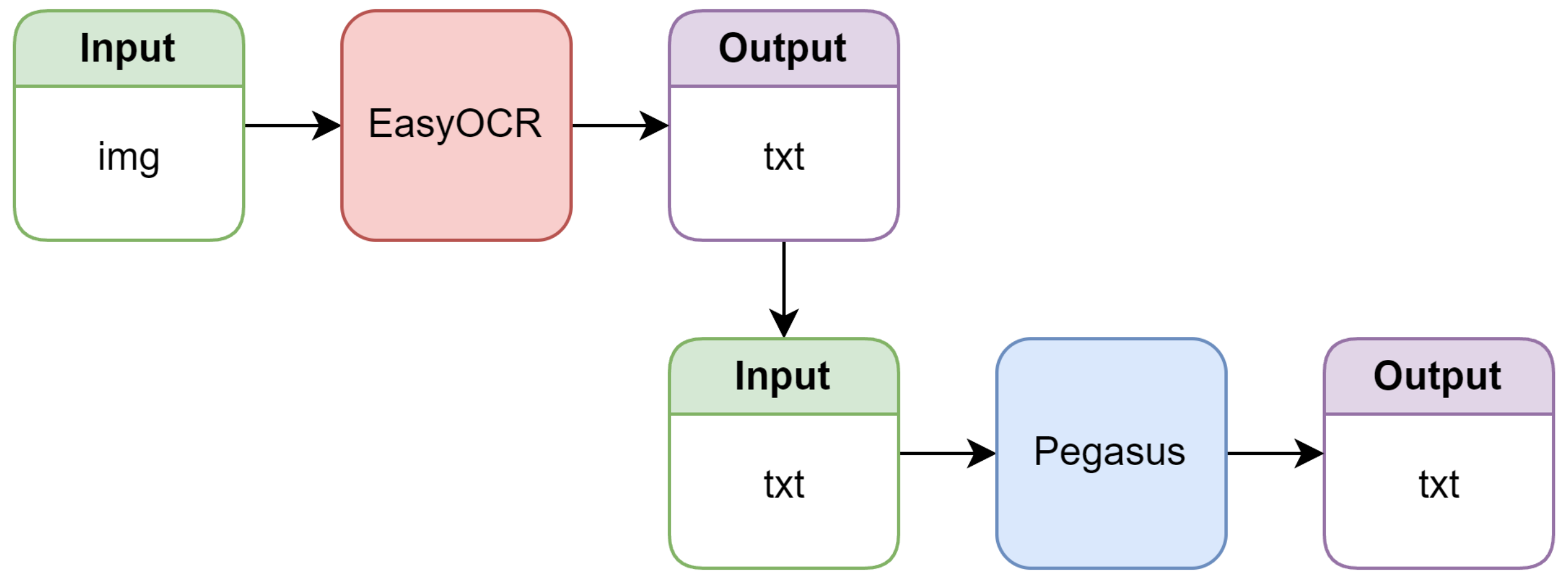
Prince Singh • Professor Bahareh Abbasi • COMP 499-Capstone

Overview

This project is designed to streamline the process of converting visual text information into concise summaries. It leverages two key technologies: EasyOCR[1] for optical character recognition (OCR) and Pegasus[2] for text summarization. The goal is to extract text from an image and then generate a summary that captures the essence of the content. The project compares two approaches: a sequential pipeline and an enclosed model system.

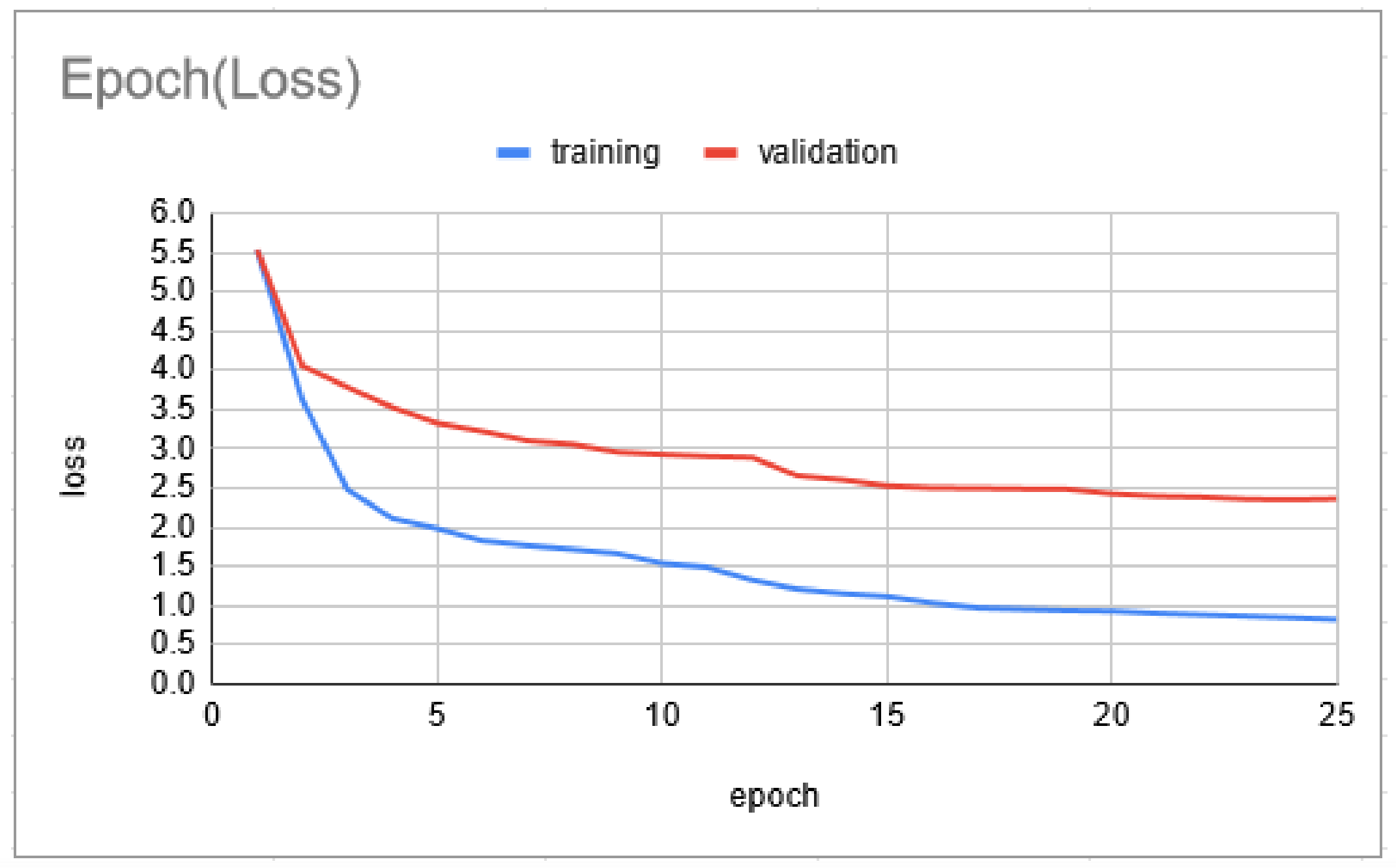
Sequential Model Approach

1. **Text Extraction**: An image is provided as input to EasyOCR, which detects and extracts the text.
2. **Text Summarization**: The extracted text is then fed into the Pegasus model which condenses the information into a summary.



Process

1. **Data Preparation**: I utilized the multi-news[4] dataset for summarization. A script was created to convert text documents into images, pairing each image with its corresponding summary from the dataset. This approach simulates real-world scenarios where summaries are derived from image-based texts.
2. **OCR Selection:** EasyOCR, chosen for its extensive documentation and CRNN architecture, is employed to read images and extract text. This step involves detecting text regions and recognizing characters, which is crucial for accurate text extraction.
3. **Summarization Selection:** The Pegasus model processes the data from the OCR for summarization. Pegasus, with its transformer-based architecture, excels in generating concise summaries.
4. **Training Setup:** The dataset was split into training and validation sets in a 4:3 ratio. With 200 samples for training and 150 samples for validation. Training was only done for 25 epochs with a batch size of 15.
5. **Model Comparison:** Both sequential and enclosed models were evaluated. Criteria included the accuracy of text extraction and the quality of summaries, measured using BERTScore. This comparative analysis helped in identifying the most efficient model configuration.
6. **Training Procedure and Validation:** During training, the model was exposed to batches of image-text pairs. The effectiveness of OCR extraction and subsequent summarization was gauged at each epoch. Validation was performed concurrently to monitor overfitting and generalization capabilities.



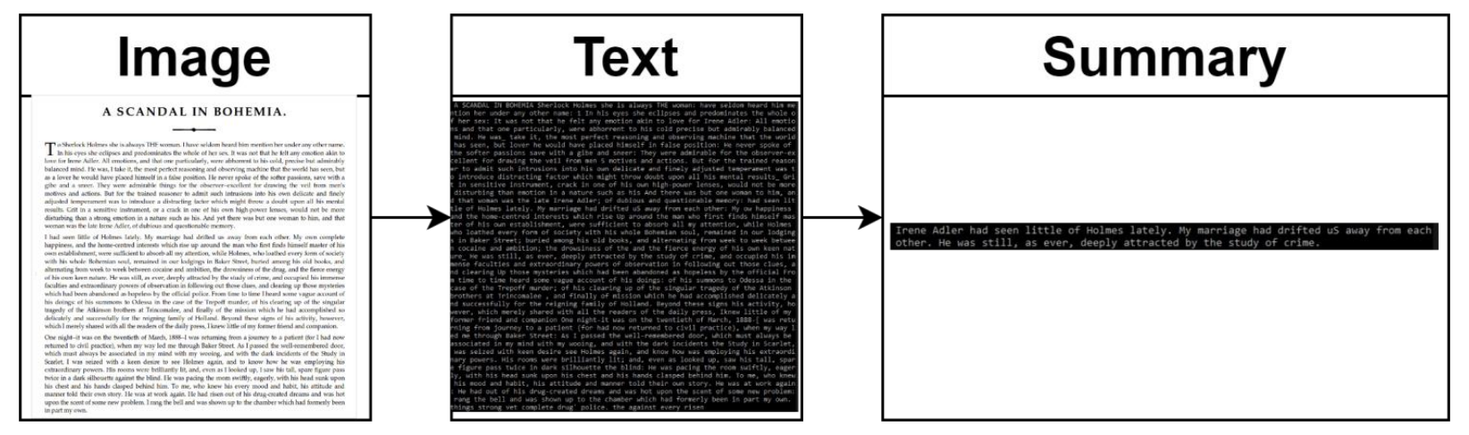
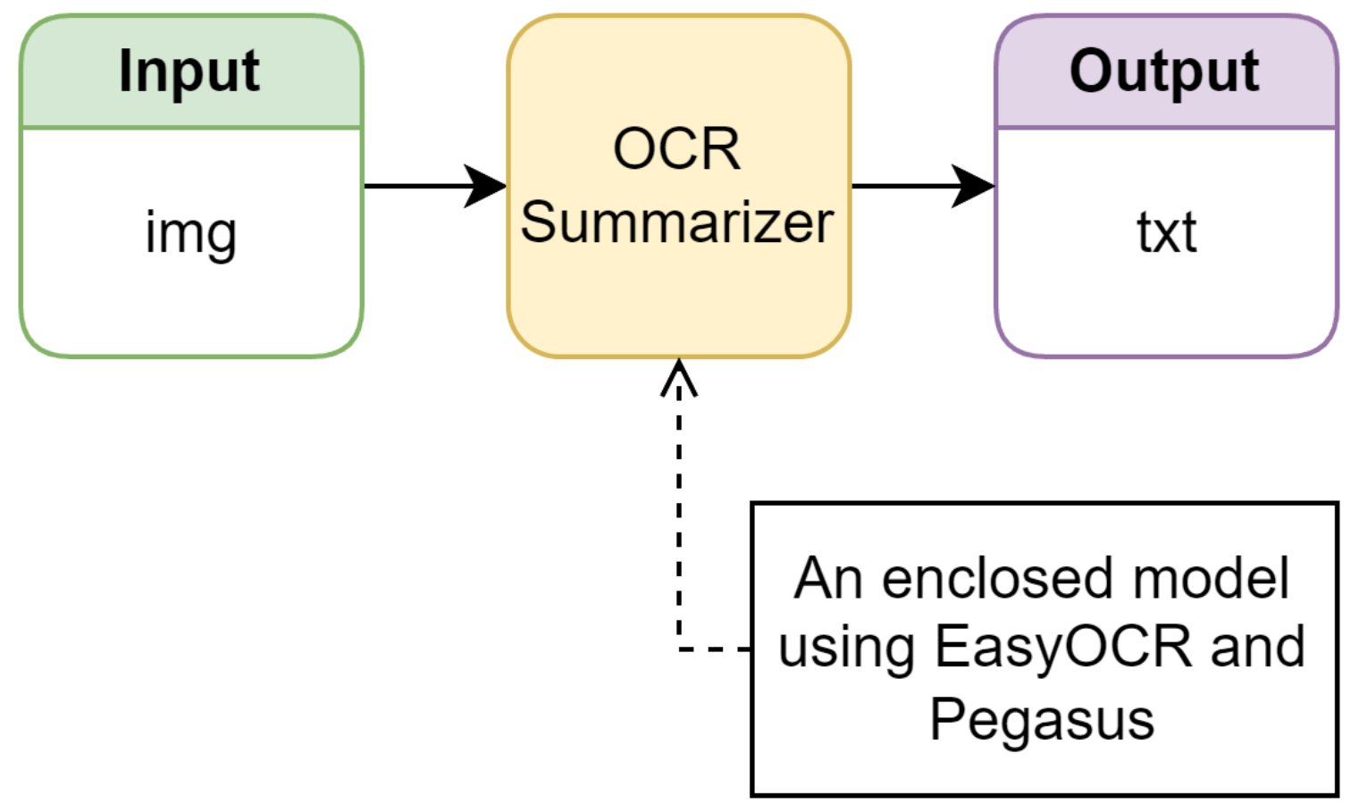


Image
A SCANDAL IN BOHEMIA.
Text
Summary
Irene Adler had seen little of Holmes lately. My marriage had drifted uS away from each other. He was still, as ever, deeply attracted by the study of crime.

Enclosed Model Approach

1. **Integrated Processing:** The image input is processed through a custom neural network model. Within this model, EasyOCR is called to extract text, and immediately afterwards, the Pegasus summarization is applied.
2. **End-to-End System:** This approach encapsulates both OCR and summarization in a single step, potentially streamlining the workflow and reducing intermediate handling.



Results

Both the training and validation loss values are decreasing over time. This indicates that the model is learning and improving its predictions as it processes more data over successive epochs. The training loss starts at 5.5284 in the first epoch and decreases consistently to 0.8207 by the 25th epoch. This consistent decrease is a good sign, showing that the model is effectively learning from the training data. The validation loss begins at 5.5352 and also decreases over time, reaching 2.3586 by the 25th epoch. The validation loss is higher than the training loss, which is common as the model is typically better at predicting data it has seen (training data) compared to new data (validation data).

The average F1 BERTscore for the sequential model is 0.87192 and for the enclosed model is 0.67618. The F1 score is a measure used in statistics to evaluate the accuracy of a test, considering both the precision (how many selected items are relevant) and recall (how many relevant items are selected). It ranges from 0 to 1, with 1 being perfect precision and recall.

Conclusion

While the sequential model demonstrated superior performance in summarization tasks, the enclosed model, despite its potential, faces significant challenges that affect its effectiveness. These challenges include dataset limitations, integration complexities, and resultant impacts on summary quality. To enhance the performance of the enclosed model, addressing these challenges is crucial. This might involve expanding and diversifying the dataset, refining the integration process, and implementing additional optimizations to improve its precision and recall. Overall, the sequential model stands out as the more reliable choice for current summarization needs, but with targeted improvements, the enclosed model could also become a viable alternative.

Challenges

Several challenges arose during the project:

1. **Lack of Dataset:**
   A significant challenge was the absence of a dataset containing pairs of images and their corresponding summaries, which is essential for training and evaluating such models.

2. **Post-Processing Integration:**
   Difficulty in customizing the post-processing steps of EasyOCR to seamlessly integrate with the summarization process of the Pegasus model, especially within the enclosed model approach.

3. **Quality of Summaries:**
   The limited dataset, self-compiled for the project, resulted in poor-quality summaries due to insufficient training and diversity in the data.

Resources

1. "Google-Research/Pegasus." GitHub, github.com/google-research/pegasus. Accessed 20 Nov. 2023.
2. JaidedAI. "Jaidedai/EasyOCR: Ready-to-Use OCR." GitHub, github.com/JaidedAI/EasyOCR. Accessed 20 Nov. 2023.
3. Falcão, Fabiano. "Metrics for Evaluating Summarization of Texts Performed by Transformers: How to Evaluate The..." Medium, Medium, 22 Apr. 2023, fabianofalcao.medium.com/metrics-for-evaluating-summarization-of-texts-performed-by-transformers-how-to-evaluate-the-b3ce68a309c3#:~:text=ROUGE%20is%20one%20of%20the,to%20obtain%20an%20overall%20score.
4. The HF Datasets Community. "Multi_news · Datasets at Hugging Face." Multi_news · Datasets at Hugging Face, huggingface.co/datasets/multi_news. Accessed 27 Nov. 2023.

PyTorch

python

EasyOCR