Centre Modeling and Simulation,SPPU, Pune

Project Anuvad Ideation

Submitted by

Name: Suraj Istari Meshram

Qualification: MTech (SCMS) , B.E(Mechanical)

Email: surajmeshram994@gmail.com

Mob.: 8888675122

Supervised by

Tekdi Technologies Pvt. Ltd.

Aim:

Ideology on Project Sunbird Anuvaad.

Abstract:

Anuvaad is an Open Source text translation tool that translates documents from and to various Indic languages based on Al and ML technologies.

Introduction:

In Anuvaad, text extraction, processing, and translation are based on NLP and Deep Learning models. At Anuvaad, I've been studying various open sources like COCO -text for a tool to split a multi-class annotation dataset with preserving class distribution among train and test sets, Using NLP (BERT) to improve OCR accuracy, End-to-end detector approaches both detect and reorganization text and LASER for parallel corpus generation that could be used to train our models for various language pairs and also to Open-Source a sophisticated dataset of parallel sentences in various Indic languages for the NLP/DL community.

Observation On AI ML Technology:

1) PRIMA: Layout detection mode using Layout parse and Label Studio

A) Since raw data documents have a lot of complicated layouts and existing OCR, we use a layout parser and label studio to tackle this problem.

B) Its pip install library uses in deep learning so, it helps to solve the following problem.

i) Extract structured data from complex documents, such as scientific papers, newspapers, and business analytic papers.

ii) Starting from raw text data, the output is never great and it is difficult to train our models or learn from it.

2) CRAFT: Character Region Awareness for Text Detection

A) it is a scene text detection (STR) method based on a neural network that shows a promising result. A PyTorch implementation that effectively detects text area by exploring each character region and affinity between characters. Region and affinity with character level bounding boxes.

B) The region score is used to localize individual characters in the image, and the affinity score is used to group each character into a single instance.[3]

C) CRAFT detects individual characters even when character level annotation is not given.

Formal Ideation About The Different Tool:

1) LASER: used for sentenced alignment

A) LASER stands for Language-Agnostic SEntence Representations which was developed by Facebook AI to support around 93 languages.

B) Using parallel datasets from various sources like Wikipedia, the LASER tool maps a sentence from one language directly to another in order to maintain consistency between embeddings.

C) LASER being built on PyTorch runs into issues while handling parallel calls, upon making API calls in parallel, we kept getting warnings on the console and eventually, the system broke upon 10 parallel calls simultaneously.

eg. Suppose we have two documents doc1 has an English sentence and doc2 has a Hindi sentence where we match a sentence S1 from D1 to a sentence S2 from D2 if S1 and S2 have the same meaning. In order to achieve this, we make use of pre-trained encoder-decoder models that transform these sentences into vector representations, these vector representations (embeddings) are then used to calculate the similarity between the sentences.

2) COCO-Text:

In terms of recognition accuracy, its reading document is accurate from low graphics images, and its capabilities are based on three tasks: text localization, cropped word recognition, and end-to-end recognition.[3]

3) BERT: Bidirectional Encoder Representations from Transformers

A) OCR is a popular technique used to extract data from scanned documents. As you would expect, the accuracy of an OCR solution is contingent on the quality of images being used as input. One challenge facing practical applications of OCR solutions is the significant drop in word-level accuracy as a function of character-level accuracy.[4]

B) NLP technique developed by Google. The BERT model has been trained using Wikipedia (2.5B words) + BookCorpus (800M words). BERT models can be used for a variety of NLP tasks, including sentence prediction, sentence classification, and missing word prediction. In this blog, we will use a PyTorch pre-trained BERT model to correct words incorrectly read by OCR.

4) End-to-end Text Detectors:

Using recognition results to improve detection accuracy, an end-to-end approach trains both detection and recognition modules simultaneously. Because a relatively small receptive field is sufficient to cover a single character in a large image, which makes CRAFT is robust in detecting scale variant texts.[2]

Conclusion:

As a result of studying AI ML technology, I believe that the above technology can be used to improve recognition, accuracy, and speed during the application phase of a project.

References:

[1] https://towardsdatascience.com/how-to-work-with-object-detection-datasets-in-coco-format-9bf4fb5848a4

[2] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun. An end-to-end text spotter with explicit alignment and attention. In CVPR, pages 5020–5029, 2018 [end to end text detector.

[3] https://rrc.cvc.uab.es/?ch=5com=tasks

[4]https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html