International Institute of Professional Studies

I I P S
DAVV
International Institute
of Professional Studies

Project Synopsis

Investigation of efficient approach for alphabet/text mining from an image

Submitted in partial fulfillment of the Requirement for the Award of the Degree of

Master of Computer Application
Semester VI
Session Jan - May, 2023

Project Guide:
Dr. Shaligram Prajapat

Submitted By:
Shruti Khandelwal
IC - 2K20 - 77

Table of Contents:

1. Introduction to the Proposed System
2. Existing System
3. Problem Statement
4. Scope of Project
5. Technologies to be Used
6. Bibliography and References

Introduction to the Proposed System

In today's world of business, data is of utmost importance. Data extracted can be helped in decision making, data understanding and data analysis. Extraction of meaningful data is important to understand, and reduces time,effort as well as human error during data entry.

Extraction of data from an image can be done using methods of deep learning. Deep learning is a subset of Machine learning, which in itself is a subset of Machine Learning, which in itself is a subset of artificial intelligence.

Artificial intelligence is simply defined as a field of computer science in which instructions don't need to be explicitly given to complete a particular task. It is the ability of a computer to perform certain tasks which are commonly associated with human intelligence. This includes speech recognition, computer vision and language translation.

With advancements in deep learning, it has become possible to extract a variety of data from images. This has been made possible because of neural networks, which are complex data structures that form the building blocks of deep learning.

Neural networks consist of nodes/neurons, which form layers. There are mainly three types of layers in neural networks. They are: Input layer, Hidden Layer and Output Layer. A neural network has two methods of learning information: forward propagation and backward propagation.

Convolutional Neural Networks are particularly used for image classification and pixel classification. This data structure is particularly used for image/object recognition. It is particularly used for computer vision.

Existing Systems

The OpenCV, stands for Open Computer Vision Library of Python, enables users to extract data using real time computer vision. It is a large open source library for computer vision, machine learning, and image processing.

Computer vision is a process by which we can understand the images and videos, how they are stored and manipulated. Computer vision is mostly used in Artificial Intelligence. Computer Vision has played a major role in self-driving cars, robotics, as well as photo correction apps.

One of the major technologies that have come into existence with regards to deep learning is Intelligent Document Processing. It is able to combine various AI technologies for automatically classifying images, as well as extracting elements and writing short sentences with proper grammar.

IDP utilizes a deep learning network, called Convolutional Neural Network(CNN) to learn patterns that naturally occur in the pixelated data. It is then able to adapt, learn and modify as new data is processed, which is instrumental in advancing its accuracy regarding predictions.

IDP uses Imagenet, which is one of the biggest databases of labeled data. Subsets of big datasets are also available on Kaggle.

IDP automates data extraction from complex, unstructured documents that help in business. Data extracted by IDP can be used for business systems to drive automation, digitization and other efficiencies. IDP offers fast, low cost and scalable operations to automate data extraction from unstructured, complex documents.

IDP performs data extraction, document classification and categorisation, data validation and offers Intelligent insights from the data that it has extracted.

Data Extraction: IDP offers solutions to automatically extract data from complex and unstructured documents. For a lot of businesses, this task is complex enough to require a trained professional for the extraction process. IDP is a powerful tool that eliminates human efforts as well as error for extraction and manual data entry.

Document Classification: IDP also automates the task of classifying documents into different categories based on their structure and content. More advanced softwares offered by IDP, such as Infrrd, offer multiple documents analysis in a single image, then automatically split and classify them so they can be categorized to resolve different queries.

Data Validation: IDP classifies and validates data with features regarding business rules, document constraints, and other sources. It is important to verify extracted data to ensure data accuracy. Data that passes validation is sent for further processing, and data that fails validation can be corrected.

Intelligence and Insights: IDP is used by enterprises to analyze the extracted data to gain insights, predict the next steps and drive better business decisions.

Steps in the IDP Process:

1. **Clean and Organize the documents**
   Documents are classified and categorized in this step to make them ready for conversion. IDP aims to integrate, validate, repair/impute problems, split images and enhance images.
2. **Convert**
   The documents and images are then converted to textual data. Various AI and OCR technologies are used for optimum performance. IDP is also able to find and maintain context within the extracted data.
3. **Validate, Enrich and Understand**
   The extracted data is raw data and is not yet ready for consumption. The data can be enriched, extended, enhanced, validated and classified during this step.

Instead of simple extraction, IDP understands the context , and thus allows more accuracy than other extraction technologies.

4. **Analytics and Insights**

   IDP can use its various AI capabilities to transform data to gain insights, and produces recommendations and predictions. AI used by IDP technologies can help in predicting the next step by identifying the missing document. IDP also uses Natural Language Processing(NLP) and Natural Language Generation(NLG) to generate summary reports based on the extracted data.

   This has been particularly useful in Insurance Providers to resolve small claims. They can identify the missing documents and notify the borrower that they need to send it.

IDP can process different types of documents:

- **Complex Documents**

  Complex documents are documents with endless text and image complexity. Text complexity includes mixed fonts, text mixed with images, long documents and multiple document types in a single PDF.

- **Unstructured Documents**

  Unstructured documents are documents in which the format and relevance of data changes over time. OCR can't manage these types of documents because it won't know where to look if the structure varies.

Although IDP is extremely proficient in tackling the toughest extraction tasks, true IDP solutions have the adaptability to effectively handle both complex as well as simple documents, allowing enterprises to cater to a wide range of documents using a single platform.

Problem Statement

Extraction of data and computer vision are one of the most important aspects of extracting data for business intelligence. Data extraction is key for understanding the context and importance of data. Scanning, analyzing and extracting data, done by computers, hugely reduces the time and human error.

The problem arises when we have to get alphabetical data from an image. An image consists of pixelated data, and can be very difficult to extract computationally.

In the real world, extracting data from an image by humans is easy, but is time consuming and may lead to human error during data entry. When carried out by a computer, this task can lead to results that contain negligible error.

This problem can be solved using deep learning techniques, such as Convolutional Neural Networks and Recurrent Neural Networks. Convolutional Neural Networks are particularly used for image classification.

A neural network consists of inputs, connects/weights and biases. When fed data, it processes the data within its hidden layers, and gives output. If the output is not of desired accuracy, the network in itself adjusts the connects between the neurons of different layers.

Scope of the Project

This project aims to investigate deep mining methods, particularly Convolutional Neural Networks and Recurrent Neural Networks for extraction of data, particularly pixelated data, from an image.

Deep mining uses neural networks, a data structure that forms the building block of deep learning, to analyze data.

Neural networks consist of nodes/neurons, which form layers. There are mainly three types of layers in neural networks. They are: Input layer, Hidden Layer and Output Layer. A neural network has two methods of learning information: forward propagation and backward propagation.

A neural network can have any number of neurons and layers. The more hidden layers in a neural network, the more is its efficiency and accuracy with its predictions. However, the more the number of neurons, the more are the computational complexities.

Convolutional Neural Networks are particularly used for image classification and pixel classification. This data structure is particularly used for image/object recognition. It is particularly used for computer vision.

This project aims to use convolutional neural networks to analyze data from an image using convolutional neural networks, to train the model on image data for accurate predictions regarding alphabetical data.

Technologies to be Used

For a project for investigation of efficient approaches concerning alphabet data extraction, some necessary programming languages, libraries and data structures are required to develop a computer vision program.

The platform used for development of this project is **Google Colab**.

Colaboratory by Google, also known as Google Colab, is a Jupyter notebook based runtime environment which allows the user to run python programs entirely on cloud. It can be used to train large scale Machine Learning and Deep Learning models, even if the user does not have access to powerful machines. It allows both GPU and TPU instances, making it the perfect tool for deep learning and data analytics. Since it is based on cloud, it can be remotely accessed from any machine through a browser.

Services that are offered by Google Colab and are used in this project:

- ➢ Write and execute Python programs
- ➢ Document programs that support mathematical equations
- ➢ Create/Upload/Share notebooks
- ➢ Import/Share notebooks from/to Google Drive
- ➢ Import external datasets, for example, from Kaggle.
- ➢ Integrate TensorFlow, Keras
- ➢ Free cloud service with free GPU.

The programming language involved in this project is **Python**.

Python is a very powerful and highly versatile general purpose language. It is an interpreted, interactive, object-oriented and high-level programming language. It is the primary language used for developments in data science, machine learning, deep learning and artificial intelligence, due to its easy to read and learn syntax and vast libraries.

Python is used in deep learning due to its concise and readable code.

The main Python libraries involved in this Project are:

1. **Numpy**

   It is a library for Python programming, and gives support to large, multi-dimensional arrays and matrices. It stands for Numerical Python. It is a fundamental package for scientific computing with Python. It is an open source software.

2. **Pandas**

   It is a software library for Python programming, and is particularly used for data manipulation and analysis. It particularly offers data structures and operations for facilitating manipulation of numerical tables and time series.

3. **Matplotlib**

   It is a plotting library for the Python programming language. It is a comprehensive library that is used for creating static, animated and interactive data visualization graphics. It is an extension of Numpy. Pyplot, a submodule of matplotlib, is mainly used for this project.

4. **Seaborn**

   It is a python library, built on top of matplotlib, for visually appealing and informative statistical graphs. It is designed to work with pandas dataframes. It offers a variety of powerful tools for visualizing data, including scatter plots, line graphs, bar plots, and many more. It also provides support for advanced statistical analysis, such as regression analysis, distribution plots and categorical plots.

5. **Scikit**

   It is a machine learning library used in the Python language.it features various algorithms, such as classification, regression, random forests, k-means clustering, random forests, etc. It is developed to work alongside Numpy and SciPy.

6. **Tensorflow**

   It is an end to end open source software library for machine learning and artificial intelligence. It can be used for a wide range of tasks, but is particularly focused on training and inference of deep neural networks. It is developed and maintained by Google.

For training the model, datasets from **Kaggle** are used.

Kaggle is a subsidiary of Google LLC. It is an online community of data scientists and machine learning practitioners. It allows users to find and publish datasets. It also allows users to explore and build models in a web based data science environment.

Bibliography and References

1. https://www.infrrd.ai/blog/image-processing-with-deep-learning-a-quick-start-guide
2. https://neptune.ai/blog/how-to-use-google-colab-for-deep-learning-complete-tutorial
3. https://en.wikipedia.org/wiki/NumPy
4. https://www.geeksforgeeks.org/numpy-in-python-set-1-introduction/
5. https://en.wikipedia.org/wiki/Pandas_(software)
6. https://pypi.org/project/matplotlib/
7. https://en.wikipedia.org/wiki/Scikit-learn
8. https://en.wikipedia.org/wiki/TensorFlow
9. https://www.infrrd.ai/products/intelligent-document-processing-idp-101
10. https://www.infrrd.ai/
11. https://www.geeksforgeeks.org/opencv-overview/
12. https://en.wikipedia.org/wiki/Kaggle