Image-Based Localization of User-Interfaces

A Project Report

Presented To

Dr. Chris Pollett
Department of Computer Science
San José State University

In Partial Fulfillment
Of the Requirements for the
Class CS 297

By
Riti Gupta
May, 2019
Abstract

There is an increasing to make the data available in all the languages so that people all over the world can understand it. Most of the data is still available in English. Data can be available in various formats, it can be text, images, books and sound. In the past, lot of work has been done to translate the data from one language to another to make it globally available. The aim of the research proposal is to study translation in different images. The focus would be to translate the image from English to Hindi. This can be extended to other languages in the future. The project would use various Artificial Intelligence and Machine Learning technologies and evaluate the results achieved using various metrics. Image translation mainly involves two basic tasks, extraction of text and translating from one language to other. Both these aspects would be discussed in this proposal.

*Index terms* – Artificial Intelligence (AI), machine learning (ML), user interface (UI), Convolutional Neural Networks (CNN)
# TABLE OF CONTENTS

I. Introduction...........................................................................................................1

II. Deliverable 1: Understanding ML techniques for image processing..............3

III. Deliverable 2: Detect text in images using openCV and Machine Learning.....5

IV. Deliverable 3: Extract text in images using machine learning.....................6

V. Deliverable 4: Secure the dataset.....................................................................7

VI. Conclusion.........................................................................................................8

References.............................................................................................................9
I. INTRODUCTION

There are many different languages that are used throughout the world. In order to avoid the digital divide, we need to translate the information into every possible language so that every human being will be able to understand it. Even today, most of the information is still available in English which is understood by just 3 per cent of the total world population. The data can be in various formats like text-based data, audio-based, image-based and many more. The aim of this project is to translate the images from one natural language to other using automated computing. As a part of this project, the focus would be on just one pair of language i.e. from Hindi to English. Due the advancement in technology, lot of vital information is captured in the form of images. The other reason being due to easy access to such electronic devices, people take lot of photos to store the information.

The image translation primarily involves two steps, first step is efficiently extracting the text from images and second step is to translate the extracted text from Hindi to English. In order to determine the text region, textual features need to be understood which can be done using edge detection and connected component analysis on the images. Character extraction from images can be done using Machine Learning models that use Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). The translation from one language to other language needs lot of information for high accuracy. Knowledge of language system, knowledge of usage and situation can be of vital importance in such conversions. As discussed in paper by Sandeep et al. [1], translation between languages that have similar structures is easy than languages that are completely different. The translation can be done using various techniques like using rules of the language system, word-by-word translation, statistical-based translation.
Each technique has its own pros and cons. Developing a complete set of rules for rule-based system might be challenging, word-based systems do not take into account the context of the complete sentence ignoring the relationship between words which might not generate meaningful translation. Statistical-based information needs large corpus which might be costly in certain cases.

In this semester, as described in above paragraph, along with reading and understanding the work done in translation area, I tried to understand how images should be processed using Machine Learning and various algorithms related to it. I also explored text recognition and extraction as part of the deliverables in the sections discussed below. Various techniques to secure the dataset have been discussed as well.

![Deliverables Diagram](Fig 1. Diagram showing the deliverables overview)
II. Deliverable 1: Understanding ML Techniques for Image Processing

The aim of deliverable one was to classify images into their suit and value in a playing card of fifty-two cards. Image processing is an important mechanism to extract useful information and learn from them. This can be done by extracting various features from the image (edge detection, pixel density, etc.). There are various techniques to extract the relevant features from images and get the required statistics from them. As a part of this deliverable, I explored one of the ML technique known as Convolutional Neural Networks (CNN) which is widely used in extracting information from images. Various filters of different sizes and values to determine different kinds of images are passed. These filters transform the images to extract the relevant features from an image which helps in determining the image category. To make the images of uniform size and give similar weightage to all the pixels, the images are padded with a certain value (most of the times with 0, -1, 1).

The dataset for this deliverable was secured by capturing images of various cards in a playing deck. These were transformed to create a diverse dataset. Along with these, the images were secured from internet as well. The dataset of images was split into training and test data. Fig.2 shows subset of images from the dataset.
First step was to transform the images to make them suitable for training the model. The images were transformed into 28x28x3 size where height and width of pixel array was twenty-eight and the number of channels was three (each channel was for Red, Green and Blue (RGB)). The values in pixel array ranged from 0 to 255.

The network comprised of CNN, MaxPooling, Flatten, Dropout and Dense layers. The activation functions used were relu and softmax. The relu activation function was used for hidden layers whereas softmax was used for output layer. The softmax activation gives the probability of the image belonging to each category. We chose the one with maximum probability as the candidate suit and value of the card. To avoid overfitting, dropout layers were used which dropped 0.5 fraction of neurons in the hidden layer. The last layers of the network
were Dense layers which considered the complete structure of the image instead of special local structures. Fig 3. shows edge detection of a spade card using CNN at output of one of the layers.

![Fig 3. Edge Detection using CNN](image)

The evaluation of the model was done using cross entropy loss function for every epoch while training the model. The loss function gives large values for incorrect predictions and small values for correct predictions. The objective is to minimize the value given by loss function. The weights were updated using optimizer “adam” which is an adaptive learning algorithm.

### III. Deliverable 2: Detect Text in Images Using OpenCV and Machine Learning

Text extraction from images is an important step to analyze the text and as a part of this project, convert it into from one language to other language. The aim of this deliverable was to understand OpenCV and Machine Learning which are widely used for image processing. Text extraction from images can be quite challenging due to various reasons like noise in images, lighting conditions, blurred images, pixel interpolation, etc. Text extraction from image involves various steps. As a first step, the image is preprocessed and the unnecessary noise which would
decrease the accuracy of the Machine Learning models is removed. This involves rescaling the image, Mean subtraction, resizing the image.

After preprocessing the images and making them ready for training the Machine Learning Models, the boundaries associated with the text is extracted along with the probability distribution. The probability is given by sigmoid activation giving the probability if the region contains text or not. If the probability is above a certain threshold, the image is considered to be a text, else, the it is considered to be non-textual region. We also get the boundary of text in the image given by a feature map. These can be used to draw boundary around the text. Techniques like non-maxima compression were used to avoid overlapping of rectangles. This ignores the boundaries that are overlapping beyond a certain threshold and selects the boundary that would be most appropriate containing most amount of information.

IV. Deliverable 3: Extract the text from images

The purpose of this deliverable was to extract text from images using Machine Learning techniques using openCV and python. The extracted text thus can be used for various other tasks and in the case of this project it can be used for translating the text from one language to other language. Building on deliverable 2, in which the text was recognized using ML techniques, this was extended to extracting the text from images using pretrained opensource models like pytesseract and making it suitable for our project. This involved cleaning of the image and then using it for passing to machine learning models. Various techniques like converting the image to grayscale, applying techniques of dilation and erosion to remove the unnecessary noise which would hinder the text extraction. We also applied adaptive threshold to get the image with only
black and white. The pytesearct does not extract text on a dark background. The image was preprocessed using by changing it from black on white. We passed various images to this model and achieved decent results. Figure 4. shows a sample image and its output.

![Sample Image](image.png)

**Fig 4. a) Input to the model**

**Fig 4. b) Text extracted from the image**

V. Deliverable 4: Securing the Dataset of Images in English and Hindi

We need images in Hindi and corresponding translated image in English to build the model and evaluate them. There can be many sources of image like United Nations (UN) website, India’s Parliament - Raja Sabha’s website, various novels in Hindi that have been translated to English, Wikipedia pages having both English and Hindi versions, etc. The main
challenge securing the dataset is to figure out the corresponding translated image in English programmatically. It is even more challenging for websites like Wikipedia where all pages have not been translated. We have written scripts to capture images on the internet by crawling the web using headless browsers. As mentioned above, the main challenge is while crawling various pages in Hindi might not have English version and vice versa. Various other options like using Wikipedia dumps and other such methodologies would also need to be explored. The number of images that would be needed is around 30,000-40,000 images to train the model and evaluate the models using various metrics.

VI. CONCLUSION

During this semester, as part of CS297 course, I explored various Machine Learning and Image processing techniques. In the first deliverable, I got to learn ML techniques like CNN which aid in image processing and extracting meaningful information from them. I also got to learn text recognition using python and OpenCV. We also used various opensource resources like pytesseract to extract the text from images and can be used in CS298 project. The dataset for the project can be gathered from internet through the methodology discussed in Deliverable 4.

These would help in CS298 project which aims to translate images from Hindi to English. The project would be divided into two parts. The first part would be to efficiently and accurately extract all the text data from the images. The second part would be to build the model using the text extracted from the images. Along with translating the text, we would also like to store the location of each text region so that the translated text can be placed at the correct location in the images.
REFERENCES


