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 need to make web data available in all the languages so that people all over the world can understand it. Most web data is still available in English only. Web data can be available in various formats, it can be text, images, books and sound. In the past, a lot of work has been done to translate web data from one language to another to make it globally available. The aim of the research project is to study the translation of web interface screenshots. In particular, the focus is to translate the User Interface images from English to Hindi. This can be extended to other languages in the future. As part of CS297, I worked on projects involving various Artificial Intelligence and Machine Learning technologies and evaluated the results achieved using various metrics. I also worked on projects that needed working with image processing and extracting necessary information from them.

***Index terms* – Artificial Intelligence (AI), machine learning (ML), user interface (UI), Convolutional Neural Network (CNN)**

TABLE OF CONTENTS

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 three per cent of the total world population [1]. The data can be in various formats like text-based data, audio-based, image-based and many more. In the past, a lot of work has been done for the translation of these data formats. In [8], the focus has been to translate weather data available in the newspapers from Japanese to English using neural networks. In [9], the translation has been done from Bangla language to English language by identifying the structure of sentence, speech and grammar and using feed-forward neural networks. The aim of this project is to translate the web interface images from one natural language to other using automated computing. For us, the focus will be on just one pair of languages: from Hindi to English for web interface screenshots. Due to the advancement in technology, lot of vital information is captured in the form of images. This is also due to easy access to electronic devices due to which people take lot of photos to store the information.

Image translation involves two steps, the first step is to efficiently extract the text from images and second step is to translate the extracted text from English to Hindi. 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 (CNNs) and Long Short-Term Memory (LSTM). The translation from one language to other language needs lot of information for high accuracy. Knowledge of the language system, knowledge of

usage and situation can be of vital importance in such conversions. As discussed in the paper by Sandeep et al. [1], translation between languages that have similar structures is easier than languages that are completely different. 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.

II. Deliverable 1: Understanding ML Techniques for Image Processing

The aim of Deliverable 1 was to classify images into their suit and value in a playing card of fifty-two cards. Fig. 1 shows subset of images from the dataset.

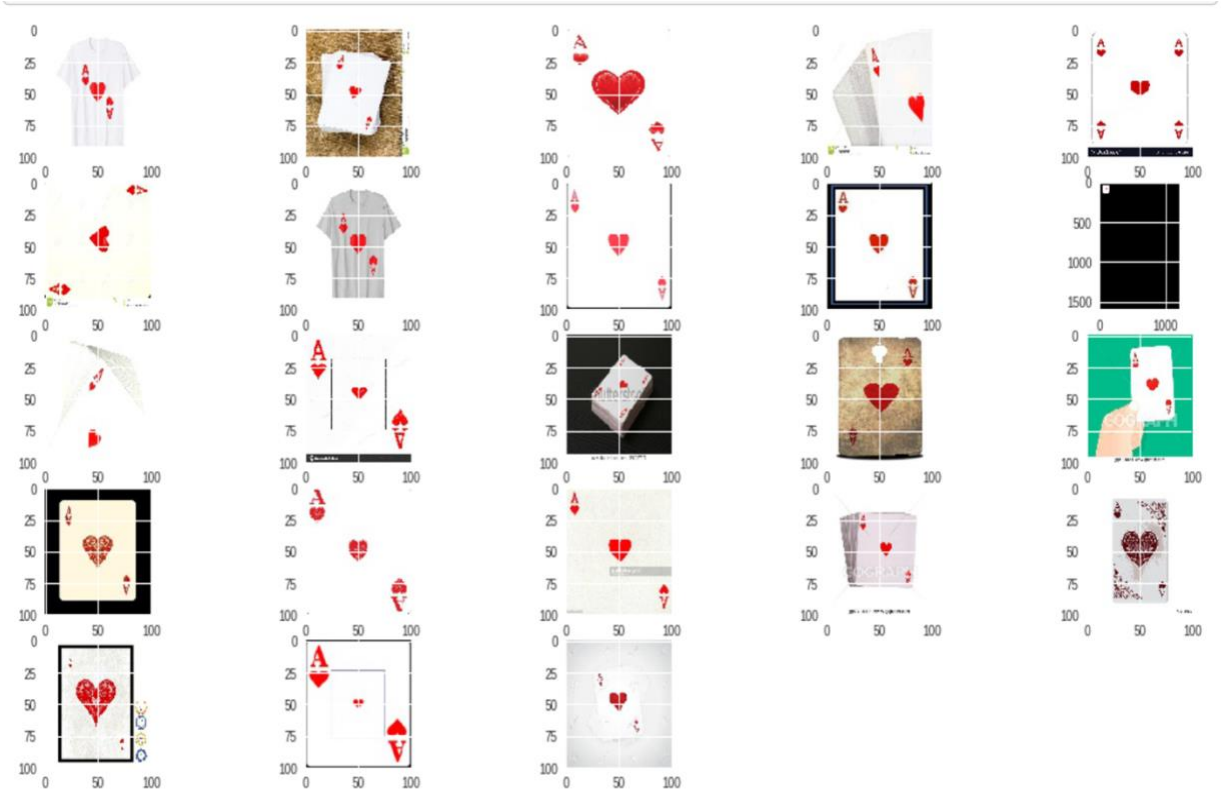

Fig 1. Dataset of images

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 an ML technique known as Convolutional Neural Networks (CNNs) which are 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 (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.

First step was to transform the images to make them suitable for training the model. The images were transformed into 28 x 28 x 3 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 as they don't have vanishing gradient problem. Thes softmax was used for output layer for multiclass classification of the card labels. 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. These dropped half of the 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 2. shows edge detection of a spade card using Convolutional Neural Networks (CNNs) at output of one of the layers.

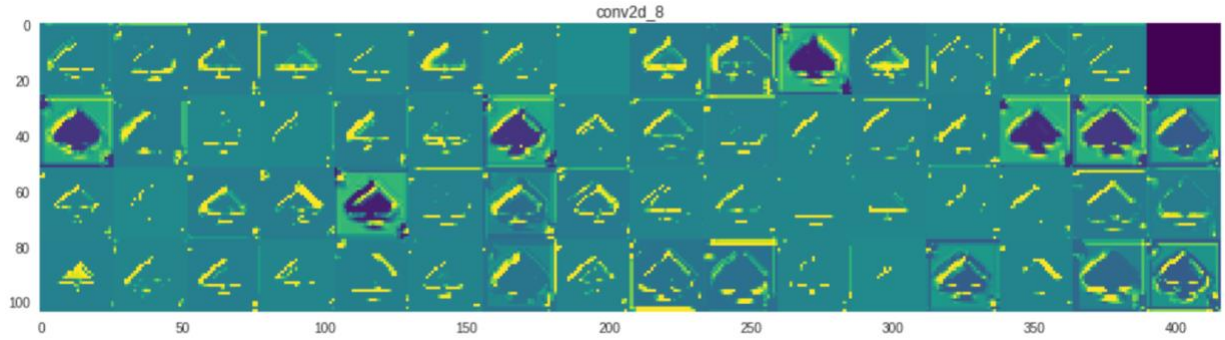

Fig 2. Edge Detection using CNN

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

The aim of this deliverable was to understand OpenCV and Machine Learning which are widely used for image processing. Text extraction from images is an important step to analyze the text and would be needed as part of CS 298 project to convert it into from one language to other language. 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 images involves various steps. As a first step, the image was preprocessed and unnecessary noise, which would decrease the accuracy of the Machine Learning models, was removed. This involved rescaling the image, Mean subtraction and 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 text, otherwise, it is considered to be a 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 pytessearct 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 3. shows a sample image and its output.

Fig 3. a) Input to the model

```
Reconnaissance de caractéres avec TESSERACT-OCR - G33Keries.org
https://geekeries.org » Sécurité v Translate this page

Dec 17, 2015 - En 2005 le code source est finalement libéré sous licence Apache, et maintenu depuis
par des équipe de Google. Tesseract-OCR est LA ...

Artcurial | Linkedin

https://www.linkedin.com/company/artcurial ¥

Learn about working at Artcurial. Join Linkedin today for free. See who you know at Artcurial, leverage
your professional network, and get hired.
```

Fig 3. 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 a 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 investigated text recognition using python and OpenCV. I used various opensource resources like pytesseract to extract the text from images.

The methodologies I studied and investigated in CS297 would be helpful in CS298 project which aims to translate web interface screenshots from English to Hindi. In order to translate the screenshots, we would need to extract the text from them which has been studied as a part of CS297. The ML techniques used as a part of CS 297 can be used in CS 298 project by feeding the text extracted from web interface images to ML algorithms for translation. The dataset for the CS 298 project would comprise of a User Interface image in English and corresponding image in Hindi. These can be gathered from internet through the methodologies explored in Deliverable 4 from websites having both Hindi and English interfaces.

REFERENCES

[1] S. Saini and V. Sahula, "A Survey of Machine Translation Techniques and Systems for Indian Languages," in *IEEE Int. Conf. on Comp. Int. & Comm. Tech.*, 2015.

[2] H.A. Driss, S. ELFKIHI and A. Jilbab, "Features Extraction for Text Detection and Localization," in *5th Int. Symp. On I/IV Comm. And Mobile Network*, 2010.

[3] C.M. Thillou and B. Gosselin, *Natural Scene Understanding*, https://www.tcts.fpms.ac.be/publications/regpapers/2007/VS_cmtbg2007.pdf

[4] X. Zhou, et al., "EAST: An Efficient and Accurate Scene Text Detector," 1704.03155v2 [cs.CV] 10 Jul 2017.

[5] E. Charniak, *Introduction to Deep Learning*, ISBN: 9780262039512192 pp. | 7 in x 9 in75 b&w illus. January 2019.

[6] O. Rippel and L. Bourdev, "Real-Time Adaptive Image Compression," The 34th Int. Conf. on Mach. Learn., 2017. doi: arXiv:1705.05823v1.

[7] G. Toderici et al., "Full Resolution Image Compression with Recurrent Neural Networks," arXiv e-prints.,2016. doi: arXiv:1608.05148.

[8] T. Law, H. Itoh and H. Seki, "A neural-network assisted Japanese-English machine translation system," in Proceedings of 1993 Int. Conf. on Neural Networks.

[9] Md. M. Hossain, K.E.U Ahmed and A.R Uddin, "English to Bangla Translation in Structural Way Using Neural Networks," in 2009 Int. Conf. on Information and Multimedia Tech.