A DEEP LEARNING TECHNIQUE FOR OPTICAL WORD RECOGNITION FOR INDIC SCRIPT

Sophia Alamanda
Assistant Professor, Department of Computer Science Engineering, SITAM, India.
Email-Id: alamanda000@gmail.com

Dr Siva Rama Krishna T, D D V Sivaram Rolangi and Dr G Jaya Suma
Assistant Professor, Department of Computer Science Engineering, JNTUK-UCEV India.
Email Id: tsrk.cse@jntukucev.ac.in, rddvsr.cse@jntukucev.ac.in, hod.it@jntukucev.ac.in

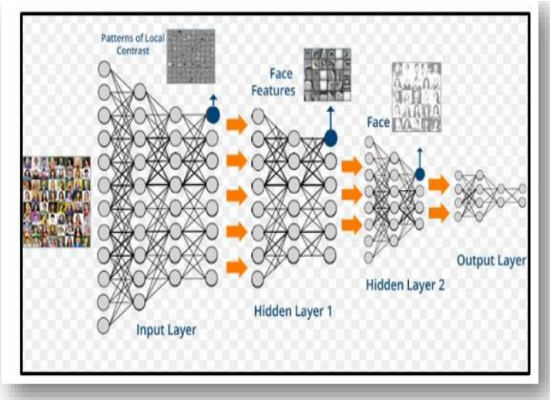
Abstract: This paper presents an Indic Script Optical Character Recognition approach using deep learning techniques called Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) learning techniques for different national language of India. Detecting and recognizing text in natural scene images is a challenging, yet not completely solved task. In recent years several new systems that try to solve at least one of the two sub-tasks (text detection and text recognition) have been proposed. In this project giving a step towards convolutional neural networks for scene text recognition that can be optimized end-to-end. In contrast to most existing works that consist of multiple deep neural networks and several pre- processing steps are applied to use a deep neural network that learns to detect and recognize text from natural images in a supervised way.This network that integrates and jointly learns to detect text regions in an image, and a text recognition network that takes the identified text regions and recognizes their textual content.

Keywords: Deep learning, CNN, VGG16 neural network, RNN, Lstm, BLstm.

I INTRODUCTION

In modern times, data records in the form of printed paper, consisting of passport documents, invoices, bank statements, printouts of static-data, or any appropriate documentation are being stored in the form of digital copies. It is a common practice to digitize printed texts so that it can be edited, searched and stored, and can be used for text mining electronically. Optical character recognition is a method of converting handwritten, typed or printed text in an image to the machine-encoded text that can later be edited, searched and used for further processing. Optical Character Recognition (OCR) can be used to convert printed texts into a digital representation. In the 1900s, an early form of optical character recognition (OCR) was used in the technologies such as telegraphy and reading device for blind people. In 1914, Emanuel Goldberg invented a device that could read characters and translate them into standard telegraphic code. In general, OCR is used to identify and read a natural language from an image and convert it into standard representation. In 1967, the research work of Anderson R.H, there has been a surge in interest for extracting patterns from images for representing them in markup form, which is a correct semantic representation of the images. 1.1 Deep learning: Deep learning is an artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Deep learning is a subset of machine learning in artificial intelligence (AI) that has networks capable of unsupervised learning from data that is unstructured or unlabeled. Also known as deep neural learning or deep neural network. Deep learning is a specific approach used for building and training neural networks, which are considered highly promising decision-making nodes. An algorithm is considered to be deep if the input data is passed through a series of nonlinearities or nonlinear transformations before it becomes output. An Example of Deep Learning: If the machine learning system created a model with parameters built around the number of dollars a user sends or receives, the deep-learning method can start building on the results offered by machine learning. Each layer of its neural network builds on its previous layer with added data like a retailer, sender, user, social media event, credit score, IP address, and a host of other features that may take years to connect together if processed by a human being. Deep learning algorithms are trained to not just create patterns from all transactions, but also know when a pattern is signalling the need for a fraudulent investigation. The final layer relays a signal to an analyst who may freeze the user's

account until all pending investigations are finalized. Deep learning is used across all industries for a number of different tasks. Commercial apps that use image recognition, open source platforms with consumer recommendation apps and medical research tools that explore the possibility of reusing drugs for new aliments are a few of the examples of deep learning incorporation.

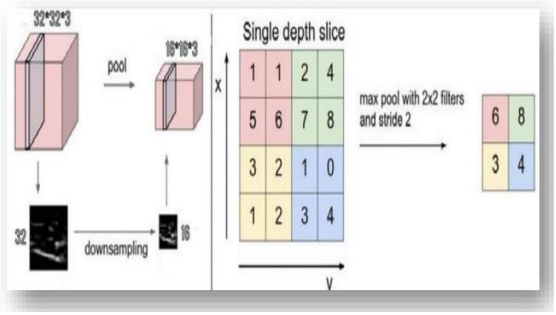


Convolutional Neural Network (CNN) Convolutional Neural Networks (CNNs) fall under the purview of deep learning. They are specifically used for high dimensional data processing such as colored images, videos etc. CNNs are multilayer feed forward networks. Each neuron in the convolution layers performs a dot product of image pixels with a filter. Each convolution layer is followed by Rectified Linear Unit (ReLu) layer and pooling layer. ReLu is a nonlinear activation function which is used to perform transformation on the images. The dimensionality of the image is being reduced as the computation moves forward in successive layers and it is achieved by pooling layer. The output of the pooling layer becomes the input for the next convolution layer. An illustration of a CNN For an instance, in Figure 2.1 an image of 32*32 size with 20 filters each of size 5*5 are used to extract features, which will produce 20 activation feature maps and it is forwarded to pooling layer. Then with a filtersize of 5*5 in pooling layer and a stride of 1 pixel, the image reduces to 28*28. This reduced image is then forwarded tothe convolution layer and pooling which will reduce the image to 14*14 and so on, till the image is reduced to dimension 1*1. CNN architectures are completely relied on four hyper-parameters such as Filters, Pooling, Stride, and Padding to give the optimal results.

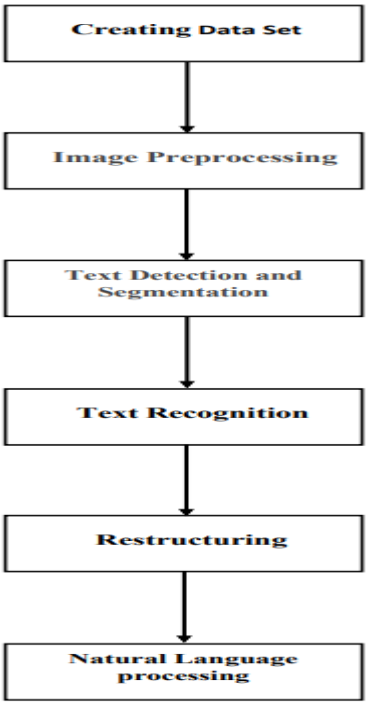
• Convolutional layers have filters which are used to extract features from the input images. Filters (for example filters of size 5*5 as shown in Figure 2.1) are moved across the whole image with strides (for example stride of size 1 pixel, 2 pixels) and produce a feature map. If you set a stride to 1 pixel, the filter will move 1 pixel at a time to cover the whole image.

• Pooling is used to reduce the dimensionality of an image. It determines the highest value among the input pixels in the filter window. For example, if an image of size 32*32*3 and pooling with filter size 2*2 and stride is 2 then the resulting image size will be reduced to 16*16*3.Figure 2.2 shows an illustration of the pooling operation.

• Padding is used to deal with the edges of the images. Sometimes it is convenient to pad the input image with zeros around the boundary. It helps to retain the spatial size of an input at output layer.



Optical Character Recognition Pipeline



II LITERATURE REVIEW

In the early 2000s, Andrew Kae and Erick Miller addressed the OCR problem in an efficient way with the computational power that existed during that time [3]. However, with the advancement in computational power both in hardware and software, a great deal of research interest has emerged in OCR. The availability of graphical processing units (GPUs) in hardware and the

development of pattern recognition algorithms based on deep learning have given a thrust to the new algorithms of OCR based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) etc. [4]. Deep Learning is a part of a big family of machine learning methods suitable for high dimensional data such as images, text,speech etc. Deep learning uses artificial neural networks that contain several hidden layers. One main principle of deep learning is that it uses raw features as inputs. The deep networks use a cascade of layers of neurons with non-linear activation function. The non-linear activation functions are used to provide a non-linearity in the network, which will help networks to perform feature extraction and transformation from the input. Each successive layer in the deep network uses the output from the previous layer as input and feeds forward it to next layer. For example, A simple explanation for a deep network is that it takes a high-dimensional input such as images, videos, speech etc. and perform non-linear activation function to extract features from an image and send it to next layer for further processing. Various deep learning algorithms have been used to solve complex problems, such as face recognition, facial expression, edge detection and many more. The deep learning techniques that have been used in this thesis are as follow:

• *Gomez and Karatzas*ref[8]propose a text-specific selective search algorithm that, together with a DNN, can be used to detect (distorted) text regions in natural scene images.

• *Gupta et al.* ref[9]propose a text detection model based on the YOLO-Architecture that uses a fully convolutional deep neural network to identify text regions.

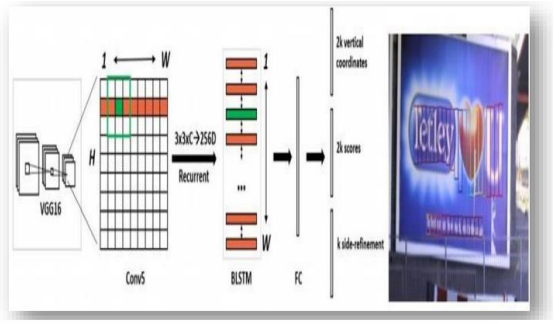
• *Bissacco et al.*ref[10]propose a complete end-to-end architecture that performs text detection using hand crafted features.

• *Jaderberg et al.*ref[11]propose several systems that use deep neural networks for text detection and text recognition

• *Good fellow et al.* propose a text recognition system for house numbers, that has been refined by Jaderberg et al.

PROPOSED MODEL

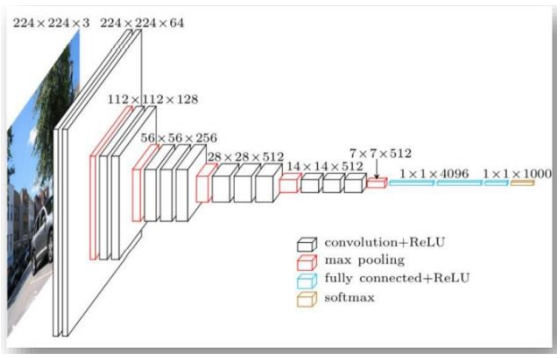
This work uses CNN followed by RNN to get the best result. CNN extracts features from an image which is then fed to RNN for getting final output. In this model we use very deep 16-layer vggNet (VGG16), a small spatial window, 3×3, to slide the feature maps of last convolutional layer. Using a sliding window in the convolutional layer allows it to share convolutional computation, which is the key to reduce computation of the costly sliding-window based methods. A vertical anchor mechanism that simultaneously predicts a text/non-text score and y-axis location of each fine-scale proposal. Text have strong sequential characteristics where the sequential context information is crucial to make a reliable decision. Including the RNN layer upon the conv5, which takes the convolutional feature of each window as sequential inputs, and updates its internal state recurrently in the hidden layer helps detector should be able to explore this important context information to make a more reliable decision, when it works on each individual proposal.



Architecture of CNN model with densely slide a 3×3 spatial window through the last convolutional maps (conv5) of the VGG16 model

III METHODOLOGY

Architecture model of VGG 16



The input to cov1 layer is of fixed size 224 x 224 RGB image. The image is passed through a stack of convolutional (conv.) layers, where the filters were used with a very small receptive field: 3×3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations, it also utilizes 1×1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1-pixel for 3×3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers (not all the conv. layers are followed by maxpooling). Max-

pooling is performed over a 2×2-pixel window, with stride 2. Three Fully-Connected (FC) layers follow a stack of convolutional layers (which has a different depth in different architectures): the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the soft-max layer. The configuration of the fully connected layers is the same in all networks. All hidden layers are equipped with the rectification (ReLU) non-linearity. It is also noted that none of the networks (except for one) contain Local Response Normalization (LRN), such normalization does not improve the performance on the ILSVRC dataset, but leads to increased memory consumption and computation time.

Training:This model is trained end-to-end by using the standard back-propagation and stochastic gradient descent (SGD). The standard practice, and explore the very deep VGG16 model pretrained on the ImageNet data. We initialize the new layers (e.g., the RNN and output layers) by using random weights with Gaussian distribution of 0 mean and 0.01 standard deviation. The model was trained end-to-end by fixing the parameters in the first two convolutional layers. We used 0.9 momentum and 0.0005 weight decay. The learning rate was set to 0.001 in the first 16K iterations, followed by another 4K iterations with 0.0001 learning rate. Training labels for text/non-text classification, a binary label is assigned to each positive (text) or negative (non-text). This model was trained on 3,000 natural images, including 229 images from the ICDAR 2013 training set. We collected the other images ourselves and manually labelled them with text line bounding boxes.

IV RESULTS

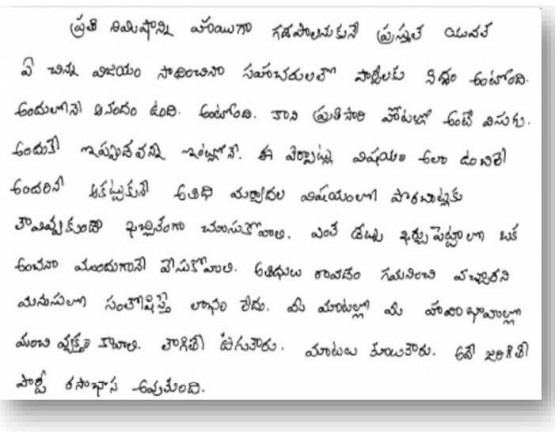


Figure:input of telugu image

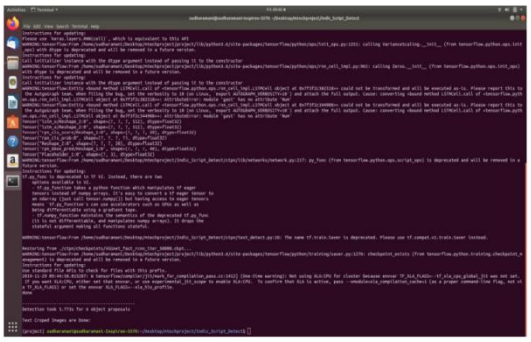


Figure: running the code for word detection(telugu).

Figure: output for telugu word detection.

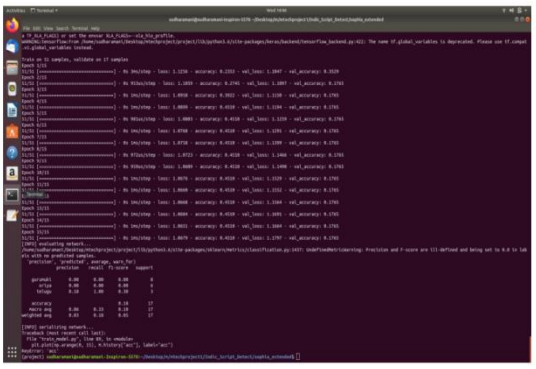


Figure:Accuracy rate for telugu images in a network.

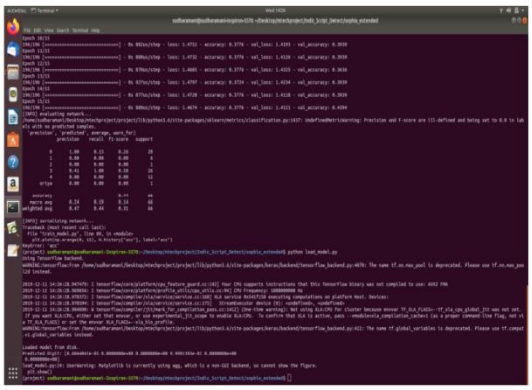


Figure : Probability accuracy for the particular(oriya)image which is an output for word detection.

V CONCLUSION

This paper presented a system that can be seen as a step towards solving end-to-end scene text recognition, using only a single multi-task deep neural network. We trained the text detection component of our model in a semi-supervised way and are able to extract the localization results of the text detection component. The network architecture of our system is simple, but it is not easy to train this system, as a successful training requires extensive pretraining on easier subtasks before the model can converge on the real task. We also showed that the same network architecture can be used to reach competitive or state-of-the art results on a range of different public benchmark datasets for scene text detection. At the current state we note that our models are not fully capable of detecting text in arbitrary locations in the image, as we saw during our experiments with the PHD_INDIC11 dataset. Right now, our model is also constrained to a fixed number of maximum text lines/characters that can be detected at once, in our future work we want to redesign the network in a way that makes it possible for the network to determine the number of text lines in an image by itself.

REFERENCES

1. R. 1. Anderson, Syntax-directed recognition of handprinted mathematics., CA: Symposium, 1967.
2. K. Cho, A. Courville and Y. Bengio, "Describing Multimedia Content Using AttentionBased Encoder-Decoder Networks," in IEEE, CA, 2015.
3. A. K. a. E. Learned-Miller, "Learning on the Fly: Font-Free Approaches to Difficult OCR Problems," MA, 2000.
4. D. Lopresti, "Optical Character Recognition Errors and Their Effects on Natural Language Processing," International Journal on Document Analysis and Recognition, 19 12 2008.
5. WILDML,"WILDML,"[Online].Available: http://www.wildml.com/2015/09/recurrentneural-networks-tutorial-part-1 introduction-to-rnns/.
6. S. a. S. J. Hochreiter, "Long Short-Term Memory. Neural Computation," 1997.
7. S. Yan, "Understanding LSTM and its diagrams," Software engineer &wantrepreneur. Interested in computer graphics, bitcoin and deep learning., 13 03 2016. [Online]. Available: https://medium.com/@shiyan/understanding-lstmandits- diagrams37e2f46f1714.
8.*Gomez and Karatzas*, TextProposals: a Text-specific Selective Search Algorithm for Word Spotting in the Wild.https://www.researchgate.net/publication/301876103_TextProposals_a_Text-specific_Selective_Search_Algorithm_for_Word_Spotting_in_the_Wild

9.Wang, X.; Zheng, S.; Zhang, C.; Li, R.; Gui, L. R-YOLO: A Real-Time Text Detector for Natural Scenes with Arbitrary Rotation. Sensors 2021, 21, 888. https://doi.org/10.3390/s21030888.
10.PhotoOCR: Reading Text in Uncontrolled Conditions.https://www.researchgate.net/publication/271551262_PhotoOCR_Reading_Text_in_Uncontrolled_Conditions

11. STN-OCR: A single Neural Network for Text Detection and Text Recognition Christian Bartz Haojin Yang Christoph Meinel Hasso Plattner Institute, University of Potsdam Prof.-Dr.-Helmert Straße 2-3 14482 Potsdam, Germany {christian.bartz, haojin.yang, meinel}@hpi.de

12. Christian Bartz and Haojin Yang and Christoph Meinel Hasso Plattner Institute, University of Potsdam Prof.-Dr.-Helmert Straße 2-3 14482 Potsdam, Germany {christian.bartz, haojin.yang, meinel}@hpi.de