International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Harnessing CRNN Technology for Image-Based Sequence Recognition

Ujwal Lonarkar, Saurabh Bhite, Suraj Ingale, Prof. Sunita Bangal

Ujwal Lonarkar, Department of Technology, SPPU
Saurabh Bhite, Department of Technology, SPPU
Suraj Ingale, Department of Technology, SPPU
Prof. Sunita Bangal Department of Technology, SPPU

Abstract - We are studying the challenge of finding text in everyday images using advanced computer techniques, specifically deep learning. Our system can accurately spot and read text in images. It uses Convolutional Recurrent Neural Networks (CRNNs), a type of deep learning model, to locate and understand the text. Instead of relying on manual rules, our model learns from the data to identify text patterns. Our projects unique feature is that our model is trained on single-word images, making it versatile in recognizing different text styles and backgrounds. We trained our model using a large dataset, including the "**MJ Synthetic Word Dataset**", which consists of 50,000 out of which we have taken 35000 as training images, 15000 as validation images. The results have been promising, indicating potential real-world applications such as searching for words in large sets of images.

***Key Words*:** Advanced computer techniques, Deep learning, System, Spot and read text, Convolutional Neural Networks (CRNNs), Deep learning model

1.INTRODUCTION

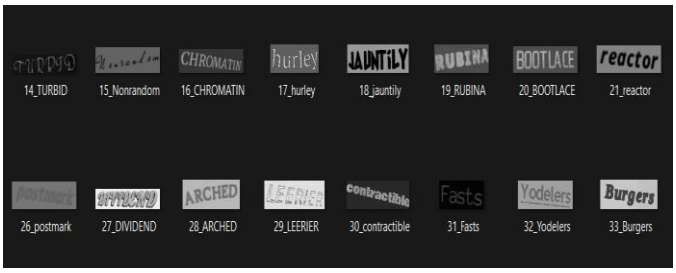
In the midst of today's image-rich landscape, a profound challenge emerges – deciphering the hidden language of text woven into visual narratives. This pursuit transcends intellectual curiosity; it is a key that unlocks the textual treasures concealed within the vast sea of digital images. Traditional methods, reliant on manual interventions, falter under the weight of images brimming with textual significance. In response, we set forth on a journey into the world of cutting-edge computer techniques, guided by the tenets of deep learning, to unveil the intricate tapestry of text that resides within the complex imagery of our daily lives.

At the core of our endeavor lies a groundbreaking amalgamation of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) – a symphony of technology aptly named Convolutional Recurrent Neural Networks (CRNN). These fusion harnesses the strength of CNNs in skillful feature extraction and the prowess of RNNs, specifically Long Short-Term Memory (LSTM) units, in sequencing. While CNNs adeptly capture the nuances of images, LSTMs orchestrate the rhythmic sequences that reveal the tale of text. Our project's distinctiveness lies in this innovative blend of CNNs and LSTMs – a potent recipe for unveiling the intricate threads of textual narratives within images. This method sidesteps traditional rule-based approaches, opting instead for data-driven learning that uncovers the latent textual patterns interwoven with the visual content.

A pivotal facet of our endeavor involves a meticulous training regime, built upon the foundation of single-word images. This approach imparts adaptability to our model, enabling it to navigate diverse text styles and backgrounds with ease. Anchored by a colossal dataset containing 35000 training, 15000 validation images, this corpus serves as our guiding star. As our story unfolds, tangible applications materialize – envision revealing hidden words within the image tapestry or untangling the intricate dialogue between text and visuals. With every stride we take, we unearth the stories concealed within images, harnessing the power of CRNN technology to decipher the unspoken language of text.

2.1 Dataset and Preprocessing:

2.1.1 Dataset: We utilize the MJ Synthetic Word Dataset, which is a comprehensive collection containing a total of 9 million images covering 90k English words, and includes the training, validation and test. But in this project, I get only 50000 These images have been meticulously divided into 35000 training images, 15000 validation images.



2.1.2 Preprocessing: - Before these images undergo any model training, they are subjected to several preprocessing steps:

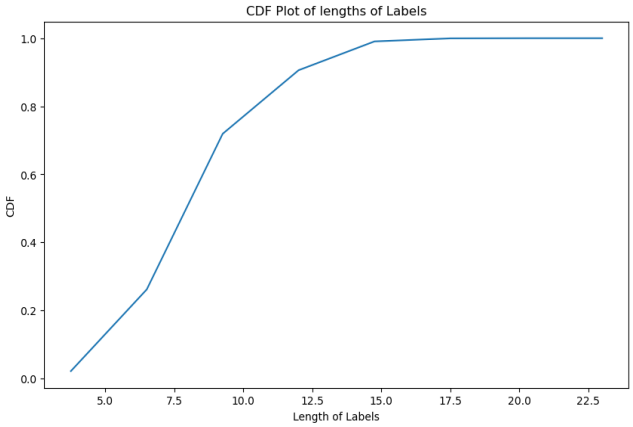
2.1.2.1 Grayscale Conversion: - Each image is converted to grayscale. This transformation simplifies the dataset, making it less computationally expensive.

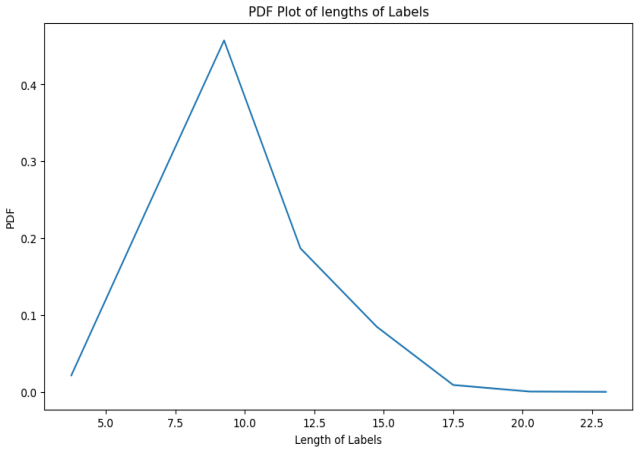
2.1.2.2 Normalization: - Image pixel values are normalized, ensuring they lie within the 0-1 range. This step ensures faster and more stable convergence during model training.

2.1.2.3 Data Augmentation: - A pivotal technique that amplifies the diversity of our training set. Here, images are subjected to transformations such as rotations, zooms, and shifts, broadening the models understanding and increasing its generalization capabilities.

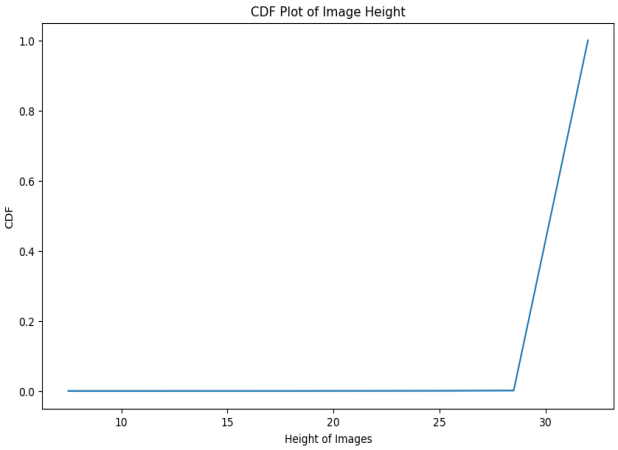
2.2Data Analysis:

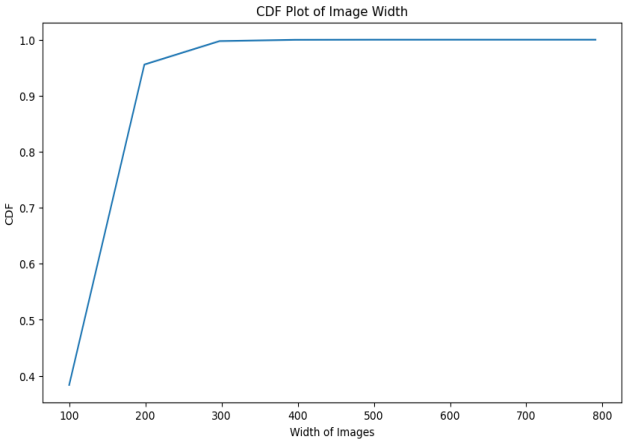
2.2.1 Label Analysis: Every image in our dataset is associated with a label representing a word. We have ensured a balanced distribution among these labels to prevent any undue bias during model training. (Labels with Lengths 7 and 8: The number of labels with lengths 7 and 8 are almost equal, comprising approximately 16.75% each of the total labels in the Data. Labels with Length 9: Labels of length 9 comprise approximately 14.8% of the total labels in the Train Data Labels with Lengths 6 and 10: The counts of labels with lengths 6 and 10 are almost equal, contributing approximately 12.1% each to the total number of labels





2.2.2 Height and Width Analysis: Based on the dataset's properties, we've discerned that the images predominantly maintain dimensions conducive for model optimization. These dimensions help strike a balance between computational efficiency and feature retention (Most of the Images have a Height of 31 Almost 95 % of the Images have a width of 200 or less Based on these observations we can resize the images to 32 (Height) x 170 (Width)





2.2.3 Digit Analysis: Almost all of the Labels in the Data do not have digits present in them comprising 99.809% of total data Very Few labels in Data have digits present with a very less percentage of 0.190% of total data.

2.3 PROPOSED APPROACH:

1) Image Feature Extraction using CNN: Here, we utilize Convolutional Neural Networks for identifying and extracting pivotal text features within images.

2) Sequence Labelling Using RNN: Post feature extraction, we deploy Recurrent Neural Networks, especially LSTM (Bidirectional) units, to handle the sequential character of the extracted text.

3) Transcription Using CTC: To transcribe the predicted sequences into meaningful text, we utilize the Connectionist Temporal Classification (CTC).

2.4 Model Architecture:

2.4.1 Structure: Input → Convolution layer 1 → Relu → Max_pool → Dropout → Convolution layer 2 → Relu → Max_pool → Dropout → Convolution layer 3 → Relu → Max_pool → Convolution layer 4 → Dropout → Convolution layer 5 → Relu → Max_pool → Convolution layer 6 → Dropout → Convolution layer 7 → Dropout → Reshape → Dense → Bidirectional LSTM (2 layers) → Dense → SoftMax → Result

2.4.1.1 Input: Our model is designed to take an image as input, specifically structured in a manner defined by the image_data_format. Depending on the format, the model might accept a 3-dimensional array (width, height, and channel depth).

2.4.1.2 Feature Extraction with CNNs: Convolutional Neural Networks (CNNs) are the cornerstone of this project, responsible for extracting valuable patterns and features from the input image. We use multiple convolutional layers to progressively extract higher and higher-level features. The deeper the layer, the more abstract the extracted features are. Our model utilizes filters of different sizes, ranging from 3×33×3 to 5×55×5. These filters are used to scan the image and produce feature maps.We use the ReLU activation function for introducing non-linearity after every convolution operation.

Max-Pooling layers are incorporated after specific convolutional layers. Their role is to reduce the spatial dimensions of the feature maps while preserving the most important features. Dropout is a regularization technique employed after certain layers, which prevents overfitting by randomly setting a fraction of input units to 0 at each update during training.

2.4.1.3 Transition from CNN to RNN: After the CNN layers, the feature maps are reshaped and passed through a dense (or fully connected) layer. This prepares the data for sequence processing by the RNN layers.

2.4.1.4 Sequence Modelling with RNNs: We use Bidirectional LSTMs, which means our LSTM layers are considering information from both past (backward) and future (forward) steps simultaneously. This is especially useful for text recognition where context plays a vital role. The model consists of two Bidirectional LSTM layers that work in tandem to capture the sequential patterns of the text.

2.4.1.5 Transcription and Output: The final dense layer with a SoftMax activation function provides the probability distribution over our classes (characters in the text). Connectionist Temporal Classification (CTC) is used as a loss function. It helps the model understand the alignment between the predicted sequences and the ground truth, which is especially important in text recognition tasks due to the variable length of text.

2.4.1.6 Training Parameters: Batch Size: Different sizes are used for training and testing to ensure optimal memory usage and faster computations.

2.4.1.7 Dropout: This regularization technique is applied after specific layers. It ensures that the model doesn't overfit the training data.

2.4.1.8 Optimizer: We use the Adam optimizer to adjust the weights of our model. Adam is a popular optimization algorithm in deep learning. It's known for combining the advantages of two other extensions of stochastic gradient descent: AdaGrad and RMSprop. Essentially, Adam computes adaptive learning rates for each parameter, making it particularly effective for problems with large datasets or many parameters.

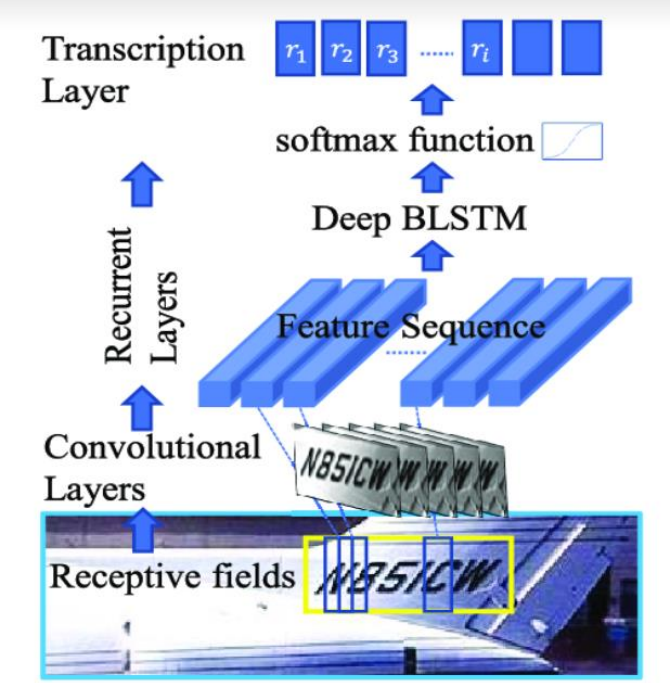
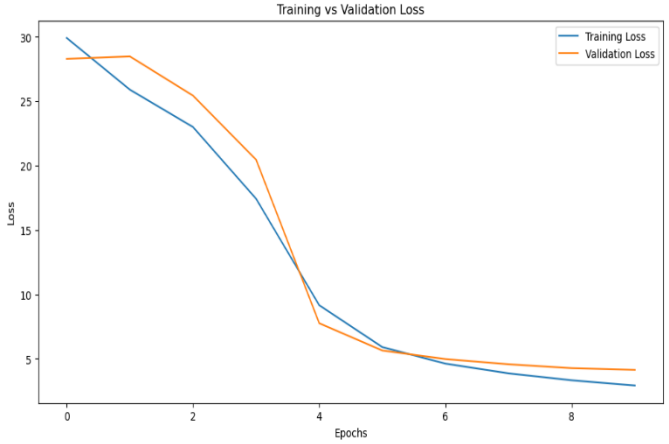


Fig. Model Architecture

3. CONCLUSIONS

During our project, into image text data we utilized methods to uncover the language hidden within images. By leveraging Convolutional Recurrent Neural Networks (CRNNs) we have developed a model of detecting and interpreting text from individual word images. The models consistent training progress resulting in a validation loss of around 4.1384 demonstrates its effectiveness. However, the model performs optimally when working with images that have a width of 128 pixels and a height of 32 pixels. While it can adapt to dimensions doing so may affect the accuracy of full text recognition. This emphasizes the importance of operating within the parameters defined by the model.



Our findings highlight the potential of CRNNs in detecting text within images particularly when trained on datasets like the MJ Synthetic Word Dataset. As digital landscapes become increasingly filled with text integrated images our tool offers a means to extract this embedded information. The future holds possibilities, for enhancing the model to accommodate linguistic and image variations. We are eager to explore these potentials.

REFERENCES

1. G.R. Hemanth, M. Jayasree, S. Keerthi Venii, P. Akshaya, and R. Saranya: Department of Electrical and Electronics Engineering, PSG Institute of Technology and Applied Research, India, ICTACT JOURNAL ON SOFT COMPUTING, OCTOBER 2021, VOLUME: 12, ISSUE: 01.
2. P. Rajeshwari, D. Vinay Sekhar Reddy, G. Pranay, Mohd Arbaz Mazharuddin 1 Assistant Professor, Department of Computer Science and Engineering, Anurag Group Of Institutions, Telangana, India, 2,3,4 Student, Department of Computer Science and Engineering, Anurag Group Of Institutions, Telangana, India: Text Extraction from an Image using CNN.
3. Manolis Delakis and Christophe Garcia Orange Labs, 4, rue du Clos Courtel, 35512 Rennes, France: TEXT DETECTION WITH CONVOLUTIONAL NEURAL NETWORKS
4. Qixiang Ye, Member, IEEE, David Doermann, Fellow, IEEE: Text Detection and Recognition in Imagery: A Survey
5. Rahul Chauhan Kamal Kumar Ghanshala R.C Joshi Graphic Era Hill University: Convolutional Neural Network (CNN) for Image Detection and Recognition 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)
6. Savita Choudhary, Nikhil Kumar Singh, Sanjay Chichadwani Department of Computer Science and Engineering Sir MVIT Bangalore, India: Text Detection and Recognition from Scene Images using MSER and CNN: 2018 Second International Conference on Advances in Electronics, Computer and Communications (ICAECC-2018)