COS 429 Final Project - Hindi Scene Text Recognition

Vaibhav Mehta

vaibhavm

Abstract

Scene Text Recognition is one of the most important practical problems in Computer Vision. Since Hindi is amongst the most spoken languages in the world, studying it in this context becomes even more important. In this project, I explore the effect of Focal CTC Loss function on the baseline model for the IIIT-ILST dataset. I find that the focal loss function does improve the accuracy, achieving a Word Recognition Accuracy of 49.1%, significanlty higher than the baseline.

1 Background

Scene Text Recognition is a long standing problem in Computer Vision with a wide range of applications across many areas. There are two steps to Scene Text Recognition. The first is text localisation and the second is the actual detection. While many modern text detection systems do both these parts, in this project, I focus only on the second part, i.e text recognition. In particular, I investigate improvements to a method published by Mathew et. al [9]. Although, their results are no longer State of the Art, I try to improve on them. In particular, I try to mitigate the problem of unbalanced data through a focal loss function.

1.1 Related Work

Much work has been done on text recognition over the years. Of late, Deep Learning approaches have performed very well over the years on English Text Recognition. Early Deep Learning methods involved using a Convolutional Neural Network (CNN) to recognise individual characters and then combine these results to achieve word level recognition [14] [2] [7]. While these methods worked fine for English, they were not suitable for Indic Languages for two reasons. The first reason is that these methods require accurate character segmentation. Character Segmentation in Hindi is harder than in English due to the presence of a 'shirorekha', that is a line that runs along the top of full letters, and half characters that fuse with each other. The second reason is that the number of available characters is much larger, and characters can be fused to for visually distinct characters. Despite this, progress was made in applying machine learning methods to Hindi OCR [3] [8]. However, when it came to text scene text detection these methods failed. This changed when a segmentation free, transcription based approach was introduced [12]. Such a system fed handcrafted features to an LSTM network with a CTC loss. These advances when applied to Hindi OCR, greatly improved the accuracy. This was improved upon by [11], who instead of naively selecting handcrafted features, used a CNN + RNN hybrid model with a CTC loss. This was applied to Hindi Scene Text Recognition and achieved State of the Art Results on the IIIT-ILST dataset in[9]. The current State of the Art Results are held by [10], who incorporate a multiheaded attention to the model.

2 CTC and Focal CTC Loss

In 2006, [5] introduced a CTC loss function to model the conditional probability of label sequences given probability distribution of each predicted label. This turned out to be very useful for tasks where alignment of data is very important. The CTC loss essentially calculates a loss between a continuous data with time

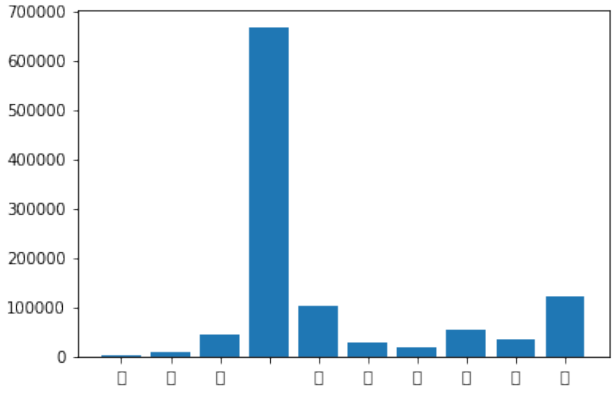

Figure 1: Graph showing frequency of 10 random characters in training data. Labels not displayed due to Unicode formatting error.

steps and a target sequence. It does this by adding the probability of possible alignments of input to target and producing a loss. Mathematically, this can be represented as follows:

Let $\mathbb{R}$ and $\mathbb{L}$ be the a set of real number and labels. Let $X = \mathbb{R}^{m \times t}$ be the feature space of the input, where m is the feature dimension and and t is the number of time steps. Let $Y = \mathbb{L}^s$ be the label space of size s. The CTC function models a joint probability distribution over X and Y, denoted as $D_{X \times Y}$

A CTC loss function has an input of a softmax layer.A blank label is added to obtain a new label $L' = L \cup B$. An input sequence $x \in \mathbb{R}^{mxt}$ is transformed to $y \in \mathbb{L}^{m \times t}$ sequence through the softmax layer. Denote activation of output unit k at time t as y_k^t. Then this is interpreted as the probability of observing label k at time t, which defines a distribution over the set $\mathbb{L}^{+T}$ of length T sequences of the lexicon $L \cup B$. The authors of [] refer to the elements of $\mathbb{L}^{+T}$ as paths and denotes them as π . We assume that the distribution of the outputs of the network is conditionally independent. Then the probability of path π can be expressed as follows:

$$P(\pi|x) = \prod_{t=1}^{T} y_{\pi_i}^t \tag{1}$$

We then define a mapping function β on $\pi \in \mathbb{L}^{+T}$ that maps π onto l by removing repeated blanks. Summing the probabilities of all π from β to l gives

$$P(l|y) = \sum_{\pi:\beta(\pi)=l} p(\pi|y) \tag{2}$$

This explanation of CTC loss is drawn from [5] [15] [4] [1]

2.1 Focal CTC Loss

The regular CTC loss defined above works well, and was put to use in Hindi OCR by [9]. However, one feature of Hindi is the different characters have very different frequencies, as can be seen in Figure 1, where 10 randomly selected characters and their frequencies in the training data have been plotted. Since this bias is naturally occurring, any dataset based on real Hindi words will tend to face this problem. The authors of [15] propose a focal CTC loss function that mitigates the effect of unbalanced data. They apply this to Chinese OCR and see improvements in the accuracy of low frequency data. The focal CTC loss is given by

$$F_{CTC}(l|y) = -\alpha_t(1 - P(l|y))^\gamma log(P(l|y)) \tag{3}$$

where α and γ are hyper-parameters.

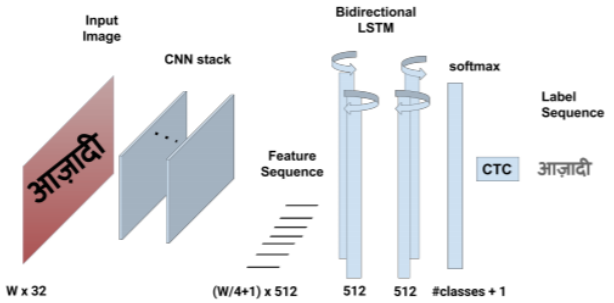

Figure 2: A visualization of the model architecture, taken from [9]

	HF Accuracy	LF Accuracy	CRR	WRR
CTC (no α, γ)	78.4 %	35.7 %	74.3 %	43.8%
α (= 0.75), γ (= 0.5)	80.6 %	48.4%	79.1 %	41.3%
α (= 0.75), γ (= 1)	**79.3 %**	**57.5%**	**81.8 %**	**49.1%**
α (= 0.5), γ (=0.5)	79.8%	36.1%	75.7%	42.9%

Figure 3: Accuracies with different values of α and γ

3 Dataset, Model Architecture and Training

In order, to investigate whether or not a Focal CTC loss makes a difference to the model used by [9], I first implemented the model used by them. A hybrid CNN-RNN network with it's convolutional stack inspired from the VGG-style architecture and minor modifications made to the layers to better fit a script recognition setting was used. In the 3rd and 4th max-pooling layers, the pooling windows used are rectangular instead of the usual square windows used in the VGG architecture. All input images are converted to grey scale and re-scaled to a fixed height. The convolutional stack is followed by two BLSTM layers each of size 512. The second BLSTM layer is connected to a fully connected layer of size equivalent to the number of labels + 1 (extra label for blank). Finally Softmax activation is applied to the outputs at the last year and the CTC loss is computed between the output probabilities and the expected, target label sequence. To accelerate the training process, batch normalization is performed after the 3rd and 4th convolutional layers. This can be visualised in Fig 2. Although many implementations of this system existed online, I chose to write my own from scratch, so I had complete control of the loss function when I had to modify it.

3.1 Dataset and Training

The authors of [9] use a Synthetic dataset for the training their model. They used 4 million images. In contrast to this, I gathered data from 2 different sources. The first source was a synthetic dataset of 100k words, taken from [13] generated using the method described in [6]. The second source of data was the Hindi images in the ICDAR MLT 2019 Challenge. For both datasets, using the bounded boxes, I cropped the images to words as 32x128 rectangles. This was still a small dataset. I added random noise, and transformations to augment the dataset. This gave a total dataset of 700k images. I trained both the baseline model and the focal model for 50 epochs each using the Adam optimiser. In the case of the Focal CTC, I experimented with different values of α and γ. The results are summarised in Figure 3.

4 Results and Evaluation

First I trained the baseline model with a regular CTC till it had comparable accuracy as [9]. The results of the baseline can be seen in the first row of Figure 3. All the models were evaluated on the IIIT-ILST [9] dataset to be consistent with [9]. To measure the effect of the CTC loss in mitigating the effect of data set imbalance,I used the following evaluation metrics :

- Character Recognition Rate (CRR), as defined in [9] = $\frac{nCharacters - \sum(Levenshtein\ Distance(RT,GT))}{nCharacters}$

	CRR	WRR
Hybrid CNN-RNN with CTC (Mathew et. al)	75.6 %	42.9%
Hybrid CNN-RNN with focal CTC	81.1 %	49.1%
E2E model w/t 8 head attention (Saluja et. al)	N .A (not reported)	51.09%

Figure 4: Comparison of various models based overall CRR and WRR

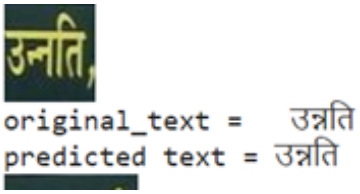

Figure 5: This word has low frequency characters that are identified correctly by the Focal Model, but the Baseline model fails

- Word Recognition Rate (WRR): $\frac{nCorrectWords}{noOfWords}$
- Accuracy of High Frequency Characters(HF Accuracy) = $\frac{Correctly\ identified\ HFchars}{Number\ of\ HF\ chars}$
- Accuracy of Low Frequency Characters(LF Accuracy) = $\frac{Correctly\ identified\ LFchars}{Number\ of\ LF\ chars}$

The results of training the Focal Model with different values of α and γ, are summarised in Figure 3. $\alpha = 0.75$ and $\gamma = 1$, has the highest accuracy across metrics. From the table it is clear that the Focal CTC loss increases the LF Accuracy significantly without decreasing HF accuracy. This leads to an overall increase in accuracy across metrics. The Overall results of this model on the benchmark IIIT - ILST dataset compared with others is summarised in Figure 4. We see improvement from [9], however the accuracy is still marginally less than the State of the Art Results in [10] In Figure 5, we see the the Focal Model has identified a word with many Low Frequency characters correctly. The Baseline model, on the other hand failed at this. Figure 6, demonstrates how the model generalizes over a variety of fonts and backgrounds.

5 Summary and Conclusions

This work has shown that using a Focal CTC loss as defined in [15] can improve the accuracy of Scene Text Recognition greatly. Although the results are not State of the Art, they come close and this significant since this model is simpler than the Attention based model used in [10], although that model performs text localisation as well. Further improvements can perhaps be made by experimenting more with the hyper parameters of the Focal CTC loss function and changing the architecture of the CNN used a feature extractor.. The main bottleneck in Hindi OCR, however, is the lack of a large public data-set to train on. Most authors rely on different synthetic datasets that do not co-respond with real scenes. A large dataset that can be used for training will help drive research in this area forward.

References

[1] *An Intutive Understanding of CTC Loss.*

[2] Alessandro Bissacco et al. "PhotoOCR: Reading Text in Uncontrolled Conditions". In: *Proceedings of the 2013 IEEE International Conference on Computer Vision*. ICCV '13. USA: IEEE Computer Society, 2013, pp. 785–792. ISBN: 9781479928408. DOI: 10.1109/ICCV.2013.102. URL: https://doi.org/10.1109/ICCV.2013.102.

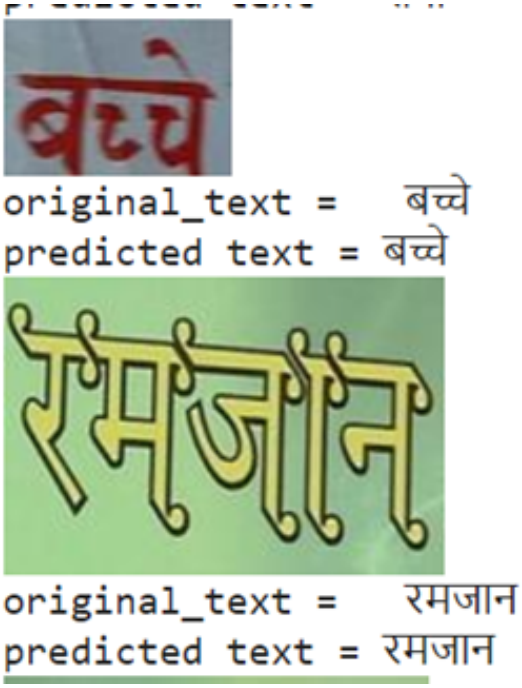

Figure 6: Model succeeding on various fonts and backgrounds

[3] B. B. Chaudhuri and U. Pal. "An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari (Hindi)". In: *Proceedings of the 4th International Conference on Document Analysis and Recognition*. ICDAR '97. USA: IEEE Computer Society, 1997, pp. 1011–1015. ISBN: 0818678984.

[4] *CMU F18 Recitation 8: Connectionist Temporal Classification (CTC)*. URL: https://www.youtube.com/watch?v=GxtMbmv169o.

[5] Alex Graves et al. "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 369–376. ISBN: 1595933832. DOI: 10.1145/1143844.1143891. URL: https://doi.org/10.1145/1143844.1143891.

[6] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. "Synthetic Data for Text Localisation in Natural Images". In: *CoRR* abs/1604.06646 (2016). arXiv: 1604.06646. URL: http://arxiv.org/abs/1604.06646.

[7] Max Jaderberg et al. *Reading Text in the Wild with Convolutional Neural Networks*. 2014. arXiv: 1412.1842 [cs.CV].

[8] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, and S. S. Ravi Kiran. "A Bilingual OCR for Hindi-Telugu Documents and Its Applications". In: *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 1*. ICDAR '03. USA: IEEE Computer Society, 2003, p. 408. ISBN: 0769519601.

[9] Mohit Jain Minesh Mathew and C. V. Jawahar. *Benchmarking Scene Text Recognition in Devanagari, Telugu and Malayalam*.

[10] Rohit Saluja et al. "OCR On-the-Go: Robust End-to-End Systems for Reading License Plates & Street Signs". In: *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*. IEEE, 2019, pp. 154–159. DOI: 10.1109/ICDAR.2019.00033. URL: https://doi.org/10.1109/ICDAR.2019.00033.

[11] Baoguang Shi, Xiang Bai, and Cong Yao. *An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition*. 2015. arXiv: 1507.05717 [cs.CV].

[12] Bolan Su and Shijian Lu. "Accurate Scene Text Recognition Based on Recurrent Neural Network." In: *ACCV (1)*. Ed. by Daniel Cremers et al. Vol. 9003. Lecture Notes in Computer Science. Springer, 2014, pp. 35–48. ISBN: 978-3-319-16864-7. URL: http://dblp.uni-trier.de/db/conf/accv/accv2014-1.html#SuL14.

[13] *Synthetic Hindi Data Generation*. URL: https://github.com/IngleJaya95/SynthTextHindi.

[14] T. Wang et al. "End-to-end text recognition with convolutional neural networks". In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. 2012, pp. 3304–3308.

[15] Shengping Zhang Xinjie Feng Hongxun Yao. "Focal CTC Loss for Chinese Optical Character Recognition on Unbalanced Datasets". In: (2019). DOI: https://doi.org/10.1155/2019/9345861.