2019 IEEE Pune Section International Conference (PuneCon)
MIT World Peace University, Pune, India. Dec 18-20, 2019

SangCTC-Improving Traditional CTC Loss Function for Optical Music Recognition (August 2019)

Jimit K. Shah, Anuya S. Padte, Purnima N. Ahirao

***Abstract*—Optical Music Recognition (OMR) is a branch of AI analogous to Optical Character Recognition (OCR) in which we train the machine to interpret sheet music to produce a playableor editable form of Music. To solve this problem in an End-to-End manner, Convolutional Recurrent Neural Network (CRNN) architecture is used. It considers both spatial and sequential nature of this problem. CTC loss function is proved to be a favorable choice in these types of sequence problems as it trains the models directly from input images to their corresponding musical transcripts without the need for a frame-by-frame alignment between the image and the ground-truth thereby solving the purpose of End-to-End training. Though traditional CTC seems to solve a major chunk of the problem, it suffers from some limitations due to overfitting/underfitting. It tends overfit/underfit because of uneven frequency distribution of symbols in Datasets and also makes overconfident predictions leading to bad generalization of model. No attempt has been made to overcome the aforementioned limitations collectively. Hence in this paper we propose a method and analyze the solution in the form of SangCTC. SangCTC is an enhanced variation of traditional CTC which attempts to overcome these limitations of overfitting/underfitting simultaneously using the concepts of focal theory and entropy.**

Index Terms*—Connectionist Temporal Classification, Deep Learning, End-to-End Training, Optical Music Recognition, Overfitting

I. INTRODUCTION

Optical Music Recognition(OMR), parallel to Optical Character Recognition(OCR), is a system in which we use various computer vision algorithms to interpret the corresponding music symbols into digital playable format [1]. Optical Music Recognition is not a trivial task. The graphical properties of musical objects are significantly different to those of optical characters. The characters use the height of the line space in a traditional manner whereas musical notes broaden the use of the y-axis. Even if the symbols have the same shape, the pitch can be determined using its position on the y-axis. Another important difference between text and music is that the graphical appearance of most character/symbol in text is quite different, while in music many shapes are pictorially similar, and the minor differences carry decisive details [2].

Manuscript received August 16, 2019.

Jimit K. Shah, studies at K. J. Somaiya College of Engineering, Vidyavihar, Mumbai 400077, India (e-mail: jimit.ks@somaiya.edu).

Anuya S. Padte, studies at K. J. Somaiya College of Engineering, Vidyavihar, Mumbai 400077, India (e-mail: anuya.padte@somaiya.edu).

Purnima N. Ahirao, Asst. Prof. at K. J. Somaiya College of Engineering, Vidyavihar, Mumbai 400077, India (e-mail: purnimaahirao@somaiya.edu).

Conventionally, we solve the problem of OMR by breaking it down into small sub-tasks [3] [4]. The General Framework consists of sub-tasks tasks such as:

1) Image Input preprocessing

2) Music Symbol Detection and Recognition

3) Semantic Reconstruction [5] [6] [7], with each step consisting of their own obstacles and complexities. Therefore, we strive for a solution that solves the task of OMR in an end-to-end manner, without the need of dividing the problem into smaller sub-tasks and falling prey to their individual limitations. To solve this problem holistically or in a single step, we use Connectionist Temporal Classification (CTC) loss function, with which the neural network can be trained in an end-to-end fashion [8]. CTC is used to train deep neural networks in many sequence learning applications such as speech recognition [9], handwriting recognition [10] and scene text recognition [11]. It is actually a loss function that concentrates on the output sequence we require, without considering which frames corresponds to which symbol in the input. It only targets that the model converges to assemble the expected sequences, and turns a blind eye to the region in which symbols are produced [12]. It allows us to train the model without knowing the exact location of symbol within the input image corresponding to the ground-truth. CTC can play an important role in solving this problem in an end-to-end manner. Traditional CTC has the following limitations leading to overfitting and underfitting:

1) It does not train well on datasets that are extremely unbalanced or contain large amount of low-frequent samples.

2) It also tends to focus on overconfident paths which leadsto bad generalization of model.

To solve all the above mentioned limitations we came up with SangCTC. SangCTC is an enhanced form of traditional CTC which combines the concepts used in Focal CTC and EnCTC [12] [13]. These upgraded versions of CTCs were used to solve the problem in allied applications similar to OMR i.e. OCR. Our SangCTC adopts these concepts in OMR to overcome all the aforementioned limitations simultaneously.

II. BACKGROUND

A. Connectionist Temporal Classification(CTC)

Connectionist Temporal Classification (CTC) is an objective function for End-to-End sequence learning, which adopts dynamic programming algorithms to directly learn the mapping between input and output sequences [12] [14].

Let's consider mapping input images X to corresponding output semantic representations $Y= [y_1,y_2,...,y_u]$. We want to find a precise mapping from X to Y's, i.e. from the input

978-1-7281-1924-3/19/$31.00 ©2019 IEEE

music sheet to its transcripts. The parameters of network are denoted by Θ [15].

1) Y can vary in length.

2) We don't have an accurate alignment of (corresponding to the input and output) X and Y.

For a given X, CTC gives us an output distribution over all possible Y's. We can use this distribution either to infer a likely output or to assess the probability of a given output. We then compute,

1. CTC Loss function: $L_{CTC}(X,Y;\Theta)$.

2. Derivative of Loss Function w.r.t. Θ (parameters of network): $\Delta\Theta L_{CTC}(X,Y;\Theta)$.

This derivative is then used to update the network parameters through back-propagation [16] [17].

B. Generalization, Overfitting and Underfitting

We always want our Deep Learning Model to generalize well. Generalization is the ability of the model to respond appropriately or predict good outputs even if it encounters some unseen or unfamiliar data. Better generalization reduces overfitting underfitting. Generalization behavior depends implicitly on the algorithm used to minimize the training error [18].

Overfitting is the case where the overall cost is really small, but the generalization of the model is unreliable. Learning patterns which are biased to datasets during training or when the model is a stubborn or an inexorable learner, it is not good for the generalizability of the model and often generates unreliable predictions.

Underfitting occurs when the model or the algorithm does get the opportunity to train well enough, again resulting in low generalization and unreliable predictions.

Both overfitting and underfitting affect the model's performance adversely which is evident in the diagram as well.

We can visualize the effect of overfitting and underfitting (in the Figure 1) which demonstrates how it in turn affects the generalization of a model. In the diagram we have graphical representation of two classes. We than infer how the underfitted model fails to learn the trends and overfitted model learns too much from the dataset. We strive a model between both the cases to find the best fit for the model so as to achieve good generalization.

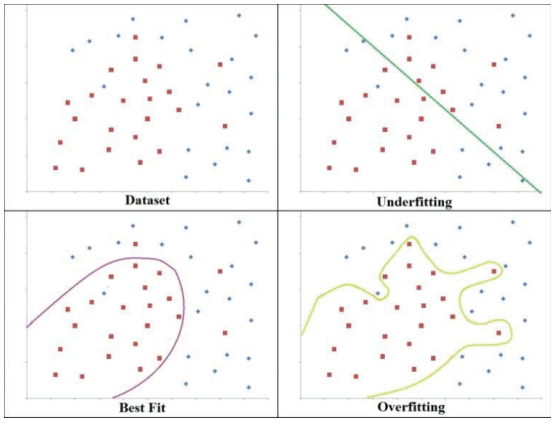

Fig. 1. A binary classification example is taken (for simplicity) demonstrate underfitting and overfitting

C. CTC and Overfitting in OMR

CTC loss function is the preferred choice when it comes to End-to-End OMR (figure 2). The limitation of traditional networks is that a strongly-aligned training set is needed, i.e., the network has to be presented with the required output of the recurring block for each input frame of the image. To focus on the desired output sequence, without considering which frames outputs which symbol the CTC loss function is adopted and is used to update the parameters of Convolutional Recurrent Neural Network (CRNN) model. CTC loss function is primarily used when we just want to directly train the model. It was discovered that CTC presents a way to optimize the CRNN parameters in such a way that it is more probable to obtain the correct sequence Y for an input X. This is because of the way in which the model is trained. We do not expect the model to give details about symbol position while producing the output. From a musical point of view, it is not mandatory to retrieve the exact position of each music symbol in the image but their context in order to precisely interpret it [8].

In spite of CTC solving our problem to an extent, it falls trap to major learning obstacles like overfitting-underfitting due to biased datasets and over confident predictions which leads to poor performance of model. Datasets like PrIMuS: Printed Images of Music Staves (discussed more in the later section) which are used in OMR are highly unbalanced. For example a music symbol like tie occurs less than 1% of the time and 2-3 bar lines are present in every image in the entire dataset [19]. As a result the model fails to learn it and ends up ignoring that symbol. The other problem is that the model is not able to adapt to other datasets which are different in nature. Also while training, entropy of feasible paths decrease too fast and become the victim of being dominated by a single path which leads to the peaky distribution problem. Attempts have been made to solve these issues independently in similar fields [20].

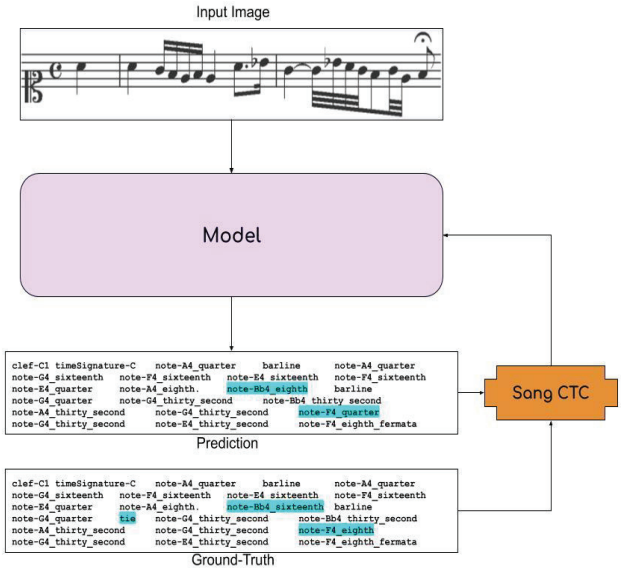

Fig. 2. End-to-End training using CTC

D. Problem of Unbalanced Dataset

Datasets used for OMR like Primus Dataset generally contain high variance in from mean frequency of individual symbols. For example, as seen in the diagram, which represents approximate distribution of music symbol

frequency, symbols like bar line have very high frequency of occurrence and symbols like ties have very low frequency of occurrence with respect to other symbols.

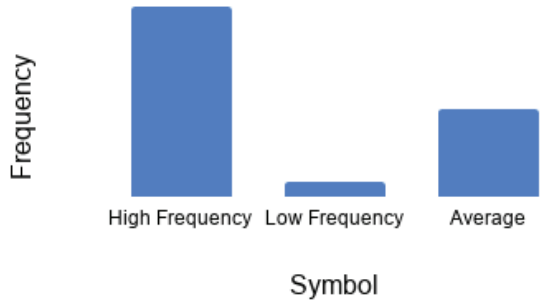

Fig. 3. Unbalanced Music Symbol Frequency Distribution in Dataset.

The symbols with low frequency of occurrence have less impact on the model during training and symbols with high frequency of occurrence have more impact on the model during training. This unbalance of datasets results in severe overfitting for symbols with high frequency of occurrence and underfitting for symbols with low frequency of occurrence. To overcome this issue of high variance in frequency of individual symbols in a dataset, we use focal loss based CTC function. This method was previously used in Chinese OCR which modifies the traditional CTC by combining focal loss with it. [13]

E. Problem of Over Confident Predictions

Confident predictions correspond to output distributions that have low entropy. A network is overconfident when it places all probability on a single class in the training set, which is often a symptom of overfitting. There can be many paths by which we can reach the desired output. Generally CTC chooses one such path and then focuses on that path only to reach the output. This happens because it produces peaky predictions at each timestamp as segmentation boundary is ambiguous. The peaky distribution is not desirable for sequence segmentation tasks when the model needs to densely predict labels for each time-step, as in sheet music where we have many symbols in a single staff line. Even if only the temporal order of labels is required, learning the correct segmentation will improve model generalization and interpretation abilities. We can solve this problem if somehow we are able to preserve all the possible paths, so when we encounter other situations, the model will be able to generalize well there too. One such solution is using EnCTC which was used in various sequence learning tasks. It prevents the entropy of feasible paths from decreasing too fast and encouraging exploration during training and hence providing more options to adapt during training leading to better generalization [12].

III. DATASET DETAILS

To train and evaluate our model effectively we require a dataset which has a lot of training samples with well labeled ground truth and simple representation. Therefore to fulfill our requirements we select PrIMuS [18] (The Printed Images of Music Staves dataset). The features which make it suitable for our problem are as follows:

- 87678 real-music incipits
- Monophonic scores
- Semantic representations

This dataset contains sheet music images consisting of a single staff also called as Monophonic scores. Each incipit (an incipit is a sequence of notes used to identify the musical work) is represented in 5 formats out of which we will be using the PNG format for the input and the semantic format for the ground truth mapping. The semantic representation is a simple format containing the sequence of symbols in the score with their musical meaning. An example of input image and its corresponding semantic representation is shown in figure 4.

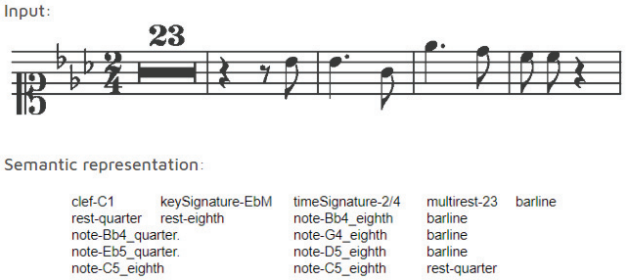

Fig. 4. Example from PrIMuS Dataset containing the monophonic scores input image and it's correspond semantic representation.

IV. PROPOSED METHOD

A. Focal CTC [13]

It is used to overcome the problem of Overfitting and Underfitting which is caused due to unbalanced dataset. Based on focal theory and Cross Entropy, Focal CTC loss function is defined as:

$$L_{Focal\_CTC}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad \ldots\ldots\ldots \quad (1)$$

Where,

p_t = probability of ground truth in the softmax output distribution,

α, γ = parameters used to balance the loss.

B. EnCTC [12]

Motivated by the maximum entropy principle, we propose a maximum conditional entropy based regularization for CTC (EnCTC). It prevents the probability from beingdominated by a single path and solves the peaky distribution problem.The solution of this problem consists of adding a regularization term to the existing traditional CTC.

$$L_{EnCTC} = L_{CTC} - \lambda_{EnCTC} \quad \ldots\ldots\ldots \quad .(2)$$

Where,

λ_{EnCTC}(regularization term) = $\beta H(p_t)$

Where,

β controls the strength of maximum conditional entropy regularization.

C. Deriving SangCTC: Combining Focal CTC & EnCTC

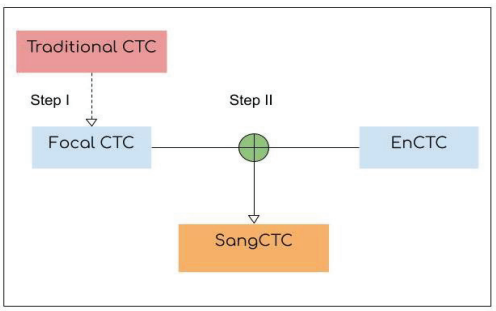

Fig. 5. Formulation of SangCTC

The basic idea to improve traditional CTC is to blend 2 CTC loss functions which are Focal CTC and EnCTC as shown in figure 5. The formulation of SangCTC is as follows:

Step 1: Using the concepts of focal theory and Cross Entropy we improve the Traditional CTC Loss function ($L_{Traditional\_CTC}$) function to Focal CTC Loss Function (L_{Focal_CTC}).

$L_{Traditional\_CTC} \rightarrow L_{Focal\_CTC}$ (Using equation (1)) .(3)

Step 2: We then add an Entropy-Based Regularization Term λ_{EnCTC} to traditional CTC.

$L_{EnCTC} \rightarrow L_{Traditional\_CTC} - \lambda_{EnCTC}$ (Using equation (2)) … (4)

Combining Equation (3) in Equation (4) we get,

$\mathbf{L_{Sang\_CTC} = L_{Focal\_CTC} - \lambda_{EnCTC}}$ …… ….(5)

Where,

L_{Focal_CTC} – Loss function of Focal CTC

λ_{EnCTC} - Regularization Term from EnCTC

V. Results and Analysis

As discussed earlier, we use End-to-End CRNN model to solve this problem of OMR. We use Symbol Error Rate(%) to find the minimum no. of addition/deletion/modification operations required to match our prediction with the ground truth. We analyzed that Symbol Error Rate(%) using traditional CTC in Primus Dataset(in semantic representation) is very less and query time per sample is also very fast(Table:I)[18]. When we further analyze the errors symbol wise, we find that bar line and ties were amongst the two most erroneously classified symbols(reasons for this is discussed earlier)(Table:II)[18]. To solve the problem of overfitting and underfitting we further analyze the results of FocalCTC and EnCTC. The results of FocalCTC come from a synthesized unbalanced dataset for OCR. They compare the accuracy of High frequency and Low frequency symbols from their FocalCTC models to traditional CTC(Table:III)[13]. We further analyze the EnCTC loss function for model generalizability. They have compared the EnCTC model with Synth5K dataset with traditional CTC(Table:IV) [12].Both the results outperform the traditional CTC as shown in the following tables.

TABLE I. Performance of Traditional CTC[18]

Model (Primus Dataset)	Symbol Error Rate	Average Seconds per Sample
Traditional CTC (CRNN-CTC)	0.8	1

TABLE II. Symbol Wise Error of Traditional CTC on Primus Dataset[18]

Symbol	Error(%)
Barline (HF)	38.6
Ties (LF)	9.4

TABLE III. Comparison of Traditional CTC and Focal CTC[13]

Model (Synthetic Dataset)	Accuracy	HF	LF
Traditional CTC	0.538	0.739	0.337
Focal CTC	**0.628**	**0.755**	**0.501**

TABLE IV. Comparison of EnCTC and Focal CTC[12]

Model	Synth5K
Traditional CTC	38.1
EnCTC	**45.5**

A. Analysis of SangCTC

To solve all the above mentioned limitations in a single model all at once we have proposed a method as well as formulation of SangCTC. *SangCTC* is an upgraded form of traditional CTC which combines the concepts used in Focal CTC and EnCTC to overcome all the aforementioned limitations simultaneously. We can overcome the shortcomings by just updating the parameters of the model. We know that during training, according to the loss, the model parameters are updated. Loss is generated by the loss function used. Hence we enhance only the loss function and not the entire algorithm to solve our purpose.

The equation of SangCTC can also be given as follows:

$\mathbf{L_{Sang\_CTC}} = -\alpha_t (1-p_t)^{\gamma} \log(p_t) - \beta H(p_t)$

Where p_t is the conditional probability of a given target sequence is defined as the sum of probabilities of all feasible paths:

p_t: $p(y|x) = \sum p(\pi|y)$

We can analyze the above equation in 2 parts:

1) **Focal:** $-\alpha_t (1-p_t)^{\gamma} \log(p_t)$ [13]

Given a probability of occurrence of a particular symbol the above equation acts as a weighted loss function.

Case 1: High Frequency

We take p_t close to 1.When we put p_t in the above equation, $p_t$being high, $(1-p_t)$ will be a smaller number $(0<p_t<1)$.

Thus, the overall loss will be less. Hence we will be able to curb the influence of HF symbols. Though this doesn't mean that we are decreasing its effect on training. A HF symbol will occur more number times giving it equal opportunity to train well.

Case 2: Low Frequency

We takep_t close to 0.When we put $p_t$in the above equation, $p_t$being low, $(1-p_t)$ will be a larger number $(0<p_t<1)$.

Thus, the overall loss will be more. Hence we will be able to influence the parameters more during training. Though this doesn't mean that we are making it more

dominant during training. A LF symbol will occur fewer times than other symbols. Therefore, in simple terms it means that it will make the most out of the opportunities it gets to train well.

2) **Entropy term:**βH(p(y|x))[12][21]

Entropy is the lack of order or predictability. Entropy is denoted by H(Y|X) where it signifies how much uncertainty is removed from Y if X is known. H(Y|X)=0 if and only if the value of Y is completely determined by the value of X. Lower the entropy the more confident it is over classes y given an input x through a softmax function. The entropy of this conditional distribution is given by :

$$H(p(y|x)) = -\sum_i p(y_i|x) \log(p(y_i|x))$$

To penalize confident output distributions, we add the negative entropy to Loss during training,

$$L_{Sang\_CTC} = L_{Focal\_CTC} - \beta H(p(x|y))$$

where β controls the strength of the confidence penalty.

We can also calculate the derivative of the entropy term w.r.t. i^{th} logit by z_i , then

$$\partial H(y|x) / \partial z_i = p(y_i|x) (- \log p(y_i|x) - H(p)),$$

which is the weighted deviation from the mean.

Finally,

$$L_{Sang\_CTC}(p(y|x)) = -\boldsymbol{\alpha} t \ (1-p_t)^{\gamma} \log(p_t) - \boldsymbol{\beta} H(p_t))$$

Can be effectively used, by controlling the parameters of focal and entropy terms: **α, β and γ** - to overcome overfitting/underfitting. This will in turn, in theory, lead to better generalization of model.

VI. Conclusion

In this paper, we have proposed a method to formulate a novel and enhanced CTC loss function which combines the concepts of Focal Theory and Entropy as SangCTC. We also identify that the problem of unbalanced dataset and over confident predictions which causes overfitting and in turn leads to bad generalization of model can be solved simultaneously by combining FocalCTC and EnCTC. We adopted a combination of two CTC loss functions which was earlier used in other domains to solve our OMR problem. Moreover, we also show a concise analysis of how our SangCTC works. We believe that our approach will overcome the limitations of traditional CTC by solving the problem of overfitting and underfitting.

References

[1] A. Zeng, "Optical Music Recognition CS 194-26 Final Project Report".

[2] D. B. a. T. Bell, "The Challenge of Optical Music Recognition".

[3] P. V. Sevy Harris, "Sheet Music Reader".

[4] T. Stramer, "Digitizing sheet music".

[5] A. J. Z. J. C. Cuihong Wena, "A new optical music recognition system based on combined neural network".

[6] C. &. R. A. &. Z. J. &. C. J. Wen, "A new Optical Music Recognition system based on Combined Neural Network".

[7] A. S. &. L. Abramov, "Musical Notes Reader".

[8] J. J. V.-M. Jorge Calvo-Zaragoza, "End-to-End Optical Music Recognition using Neural Networks".

[9] Z. W. J. T. H. N. B. L. Jiangyan Yi, "CTC Regularized Model Adaptation for Improving LSTM RNN Based Multi-Accent Mandarin Speech Recognition".

[10] P. K. M. M. a. C. J. Kartik Dutta, "Improving CNN-RNN Hybrid Networks forHandwriting Recognition".

[11] F. X. Y. L. X. B. C. Y. Zhaoyi Wan, "2D-CTC for Scene Text Recognition".

[12] H. L. S. J. C. Zhang, "Connectionist Temporal Classification with Maximum Entropy Regularization".

[13] H. Y. a. S. Z. Xinjie Feng, "Focal CTC Loss for Chinese Optical Character Recognition on Unbalanced Datasets".

[14] M. G. X. N. T. K. F. M. a. A. W. Yajie Miao, "An Empirical Exploration of CTC acoustic models"

[15] L. K. C. D. N. A. S. Liang Lu, "Multitask Learning with CTC and Segmental CRF for Speech Recognition".

[16] Z. Z. X. L. S. S. Hairong Liu, "Gram-CTC: Automatic Unit Selection and Target Decomposition for Sequence Labelling".

[17] S. F. F. G. J. S. Alex Graves, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks".

[18] Neyshabur, Srinadh Bhojanapalli, David McAllester, Nathan Srebro,Exploring Generalization in Deep Learning Behnam

[19] D. R. Jorge Calvo-Zaragoza, "End-to-End Neural Optical Music Recognition of Monophonic Scores".

[20] K. U. Eelco van der Wel, Optical Music Recognition with convolutional Sequence-to-Sequence Models

[21] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser Geoffrey, Hinton, Regularizing Neural Networks by Penalizing Confident Output Distributions