Starting Small Learning Strategies for Speech Recognition

by

Hari Vydana, Brij Mohan lal, Anil Kumar Vuppala, Manish Shrivastava

in

13th International IEEE India conference INDOCON 016

IISc Bengaluru, India

Report No: IIT/TR/2016/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
December 2016
Starting Small Learning Strategies for Speech Recognition

Conference Paper · December 2016
DOI: 10.1109/INDICON.2016.7838940

4 authors:

- Hari Krishna
  International Institute of Information Technology
  12 PUBLICATIONS 4 CITATIONS
  SEE PROFILE

- Brij Mohan Lal Srivastava
  International Institute of Information Technology
  7 PUBLICATIONS 2 CITATIONS
  SEE PROFILE

- Manish Shrivastava
  International Institute of Information Technology
  23 PUBLICATIONS 83 CITATIONS
  SEE PROFILE

- Anil Kumar Vuppala
  International Institute of Information Technology
  38 PUBLICATIONS 154 CITATIONS
  SEE PROFILE

Some of the authors of this publication are also working on these related projects:

- IIIT-H System Submission for FIRE2014 Shared Task on Transliterated Search View project

- Medical Domain, Multilingual Virtual Assistant View project

All content following this page was uploaded by Brij Mohan Lal Srivastava on 03 November 2016.

The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.
Abstract—Designing various learning strategies has been gaining a lot of scientific interest during the recent progress of deep learning methodologies. Curriculum learning is a learning strategy aimed at training the neural network model by presenting the samples in a specific meaningful order rather than randomly sampling the training examples from the data distribution. In this work, we have explored starting small paradigm of curriculum learning technique for speech recognition. The starting small paradigm of curriculum learning is performed by a two step learning strategy. Training dataset is re-organized as a set of easily classifiable examples followed by the actual training dataset and the model is trained on the re-organized dataset. We hypothesize that by following the starting small learning paradigm the learning gets initialized in a better way and progresses to attain a better convergence. We propose to rank the toughness of the training example based on the posterior probabilities obtained using a pre-trained model. Apart from re-arranging the training corpus starting small paradigm of curriculum learning is applied at model level. We consider the broad manner class classification objective function as the smoother version of the phone class classification objective function. The model initially trained for broadclass classification is later adapted for phone classification.

In this work, we have used TIMIT and a subset of Wall Street Journal (WSJ) corpus to validate the experiments, both the learning strategies have shown consistently better performances across the two datasets compared to the baseline system trained by randomly sampling the dataset.

I. INTRODUCTION

Owing to the recent progress in deep neural networks the performance of Automatic Speech Recognition (ASR) has significantly improved [1]. Deep learning methods aim at learning the feature hierarchies with features from higher level of hierarchies formed by composition of lower level features [2]. The major improvements in deep learning methods are due to their initializations and new training mechanisms [3]. Curriculum learning is one such learning mechanism that aims at learning from the non-randomly presented examples i.e., examples arranged in some specific meaning full order. Though there have been studies stressing the improvements shown by curriculum learning strategies there have also been studies where curriculum learning strategies have underperformed for certain tasks [4], emphasizing the need for designing the task specific knowledge for designing curriculum learning strategies. This motivated us to study the effectiveness, applicability of the curriculum learning strategies for speech recognition.

One of the most noted works towards curriculum learning is in [5], a staged learning strategy is designed to train a recurrent neural network (RNN) for the task of learning the grammar of a language. The data is subsetted in to four different parts based on their complexity. The network trained with the datasets gradually in the increasing order of complexity performed better than the network trained by randomly sampling the training examples from the whole data. In [2], a two stage learning is proposed for the task of image classification. The entire task is carried out using two synthetically generated datasets i.e., easy set (Basic shapes) and a tough set (geometric shapes). The easy set comprises of the basic geometrical shapes and the complex set has the geometric shapes with different sizes and orientations. The network trained with the two level strategy worked better than the network trained by randomly sampling the dataset. To replicate similar success of this learning strategy in case of natural datasets or in different domains, it is very crucial to define or rank the toughness to classify the training examples.

Recently, accordion annealing based curriculum learning strategy has been employed for curriculum learning strategy for noise robust speech recognition. In [6], multiple versions of the dataset is created synthetically and these are arranged such that the signal-to-noise ratios (SNR) of the dataset is gradually increased as training progresses and an improvement in the performance is observed.

We propose a new strategy to train a neural network in a starting small paradigm, we use a pre-trained model to rank the toughness of the training example. The posterior probabilities of the pre-trained model are considered as the indicators of the toughness-to-classify the training examples.

Curriculum learning can also be considered as a continuation method, the basic idea of continuation method is to initially consider a smooth objective function for optimization and gradually decrease the smoothness during the training process. It is hypothesized that use of a smoother version of the actual objective function enriches the model with global nature of the classification problem. In [5], a model level constraint is imposed on a recurrent neural network by gradually increasing the memory as the training progresses in a grammar learning task. Gradually increasing the memory of an RNN starting from a limited memory is hypothesized as initially presenting a smoother version of the original objective function.

We initially train the model with a broad manner class
objective function intended to classify the input features to vowel, nasal, stop and fricative. We hypothesize that the manner class classification objective function is a smoother version of the phone classification objective function. The initially trained network with broad manner class objective function is then adapted to the actual phone classification task.

From the studies of [2] and [5], it has been hypothesized that the paradigm of starting small (curriculum learning) has its influence on the speed of model convergence during the training and attaining better local minima. Both the above mentioned hypothesis are successfully validated in [2], [5] for image classification and language learning tasks. A contrary argument to the hypothesis has been proven for language acquisition task in [4]. In this study, we are interested in exploring the implications of curriculum learning strategies for speech signals. The effectiveness of the curriculum learning by training a deep neural network and the performance metric per-frame phone error rate is used. We opted this metric opposed to conventional phone error rate as per-frame phone error rate is a direct measure of classification accuracy.

The major contribution of the present work can be summarized as follows

- Exploring the curriculum strategies for speech recognition.
- Use of a pre-trained or a parallel model posteriors to rank the toughness to classify of the training example.
- Exploring the continuation methods and use of broad manner class objective function as a smoother version of the phone recognition.
- Attempts to combine both the approaches.

The datasets, model architecture and feature extraction are described in section II. Various curriculum learning strategies comprising the corpus re-arranging methods and broadclass continuation methods are presented in section III.

II. EXPERIMENTAL SETUP

A. Database

During the course of the study we have used two datasets TIMIT [7] and Wall Street Journal corpus (WSJ) [8]

a) TIMIT: We have used TIMIT database with all the 632 speakers. Data from 50, 30 speakers is used as validation and test set and the data from remaining speakers are used for training the phone classifiers [7]. The phone alignments generated using kaldi DNN are used during to obtain the phone labels.

b) WSJ: We have used a subset of WSJ [8] corpus, seven hours of speech data from si284 is used for training the models and dev93, eval93 have been used as validation and test sets.

B. Feature Extraction

The type-4 features proposed in [9] are used in the present study. The conventional MFCC features spliced in time over 9 frames (±4) resulting feature vector of 117 is reduced to 40 dimensions by linear discriminant analysis (LDA). A speaker normalization using feature-space maximum likelihood linear regression (fMLLR) is performed. The resulting speaker normalized 40 dimensional features are again spliced in time over 6 frames and their dimensionality is reduced to 200 by an LDA. This entire feature extraction is performed using kaldi toolkit. To authenticate the feature extraction we have implemented an MLP presented in [10] and obtained comparable per-frame phone error rate of 56 %.

C. Architecture

A deep neural network with three hidden layers comprising of rectified linear units (R) and softmax (S) units with architecture 200R-800R-800R-800R-39S is used during the course of the present study. The categorical entropy of the outputs is used as the loss function. ADADELTA [11] is used as an optimizer. The dropout of 0.8 for the input layer and 0.5 is used for the hidden layers [12].

The broad manner class identification network is also trained using the same architecture i.e., (200R-800R-800R-800R-4S). The broadclass network trained using the TIMIT labels is used for both the datasets.

III. STARTING SMALL CURRICULUM LEARNING STRATEGIES

In this work, we have explored two different types of starting small curriculum learning strategies i.e.,

- Corpus-re-arranging strategies
- Broadclass Continuation methods

Both the above learning strategies are aimed at exploring the paradigm of starting small while training the neural network. In the former approach, the paradigm of starting small is achieved by arranging the dataset such that the initial examples are easily classifiable by the classifier. In the later approach, the paradigm of starting small is achieved by initially training the classifier using a smoother version of the actual phone classification objective function. The details of the experiments in both these approaches is presented in the below subsections.

A. Corpus re-arranging strategy

The basic idea of curriculum is to start small, learn easier examples and gradually increase the difficulty level [2]. In this work we have explored two step training process of the curriculum learning i.e.,

1) Initially train with the set of easy examples
2) Train with the full training set

To obtain the set of easy examples, we train a DNN classifier with architecture mentioned in section II-C and get the predictions of the classifier on the training data. The examples that are correctly predicted from the training data are considered as the easily classifiable examples (easy set). The Epoch at which we switch the dataset from the set of easy examples to the actual examples is termed as switch epoch.

The performance of two step curriculum learning i.e., corpus re-arranging strategy is presented in Fig. 1. Initially a DNN with architecture mentioned in subsection II-C is trained on the full training dataset for 200 epochs and the performances of
Fig. 1: Performance of the proposed two step learning strategy (corpus-re-arranging method). The model is trained with the easy set of examples till the switch epoch and with the actual dataset after the switch epoch. Performance of proposed approach at various switch epochs in comparison with the baseline is presented in Fig. 1. All the figures are generated from the subset of WSJ corpus mentioned in subsection II-A.

From Fig. 1, it can be observed that the a rapid increase in the training and a slow increase validation accuracy can be observed when the training is being done on the easily classifiable examples. This rapid increase in the training accuracy can be attributed to the reason that the new model is able learn faster from the easily classifiable set. But the data composition of the validation set resembles the full training dataset but not the easily classifiable set so the validation accuracy increases gradually. During the two step training process when the training is happening on the easily classifiable examples the validation error kept on increasing along with validation accuracy this is slightly different form other sorts of training during this training the network is learning to classify more number of validation examples correctly rater than making the posterior probabilities of correctly predicted class higher than the other classes (i.e., predicting the probability distribution of examples closer to the target distributions), this can be viewed as the model is initially learning the global picture of the full classification problem.

After the switch epoch (when the training is done on the full dataset ) a very rapid decrease in training and validation error can be observed compared to the baseline indicating the faster convergence of the training strategy. From Fig. 1, it can be observed that the model trained with proposed two step learning strategy with any switch epoch has performed better than the baseline though the number of epochs are lesser than
the baseline system. The performance of the proposed strategy of learning has its highest performance in terms of validation and test error around 50 epochs and a further increase in the switch epoch, similar or slightly lower performance is observed. This saturation or decrease in the performance may be due to the reason that the weight reaches to a point in weight space that it cannot re-arrange itself to the actual training dataset. The test accuracy computed form the two step learning strategies is compared with the baseline system in Table I. Column 2 & column 3 in Table I are the test accuracies obtained for the two datasets i.e., TIMIT and WSJ. Row 2 of Table I is the performance of the baseline system, Row 2 to 6 of table. I the performances obtained with different switch epochs. The performance of two step learning strategy on the test set is presented in Table I. Similar to the validation accuracy, the performance of test accuracy is higher than the baseline system in both the datasets. Apart from the proposed learning strategy we have trained the network with with exactly opposite to the proposed criterion i.e., training the network with full dataset and training extra epochs with the dataset comprising examples that are wrongly classified by the initial network, hoping the reinforce the raining for wrongly detected samples but the performance is poorer than the baseline.

<table>
<thead>
<tr>
<th>Corpus Re-arranging Strategy</th>
<th>WSJ</th>
<th>TIMIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>52.66</td>
<td>60.82</td>
</tr>
<tr>
<td>switch epoch 30</td>
<td>53.35</td>
<td>60.92</td>
</tr>
<tr>
<td>switch epoch 50</td>
<td>53.44</td>
<td>60.93</td>
</tr>
<tr>
<td>10 epochs-train set + 30 epochs easyset + 100 epochs-trainset</td>
<td>–</td>
<td>60.83</td>
</tr>
<tr>
<td>100 epochs-train set + 50 epochs toughset</td>
<td>–</td>
<td>60.59</td>
</tr>
</tbody>
</table>
As observed in the Fig. 1(b) to reduce the validation error during the training with easily classifiable set we initially trained 5 epochs for with the full dataset and performed the proposed two step learning process, the increase in the validation error got reduced but the performance is comparable to baseline, thus showing that the improvements are majorly due to starting small learning paradigm.

B. Broadclass based continuation method

Curriculum learning can also be considered as a variant of continuation methods. Continuation methods aim at initially learning an easier version of the objective function and gradually increase the complexity to the original objective function as the training progresses. By starting from the easier version of the complex objective function the network acquires global information about the task.

The problem of phone classification is a major aspect of speech recognition studied under acoustic modelling. The phones of a language (39 in TIMIT English) can be broadly categorized based on place and manner of articulation. In this work, we use broad manner class task as a smoother version of the phone classification task. The rationale for choosing the broad manner class is that in literature the classification accuracy of manner classes is high compared to actual phone classes and place of articulation, moreover in the literature it is a well established fact that in classification hierarchy broad manner classes is a parent class to phones. We hypothesize that manner class classification task is a smoother version of the phone classification task. We initially train the network for the a manner class classification network which is then later trained for phone classification.

We initially train the network to classify the four broadclass manner class events in speech i.e., vowels, fricatives, nasals and stops. The trained broadclass classification network is later adapted to the final phone classification objective function, this approach is termed as broadclass based continuation approach. The performance of this broadclass continuation approach is presented in Fig. 2. It can be noted that, an improvement in performance of the broadclass continuation method is observed compared to the baseline system. The broadclass based continuation approach is actually trained for 300 epochs (100 epochs on broadclass objective function + 200 epochs on phone classification function), but it can be noted that the performance of phone classification system trained in broadclass continuation method has better performance than the baseline even at 100 epochs.

From Fig. 2, it can be noted that the network trained in broadclass continuation method has reached better local minimum and has faster convergence. The broadclass network trained using TIMIT dataset is later adapted for both the datasets i.e., TIMIT and WSJ. The test accuracy of the broadclass continuation approach is compared with the baseline system in Table. II. An improvement in the test accuracy can be observed in both the datasets compared to the baseline system.

IV. Combined Approaches

Though both the approaches exploit the data and the model perspective of starting small learning paradigm, we attempted to combine both to approaches, the pre-trained broadclass model is initially trained with easily classifiable set till the switch epoch and later the network is trained with full dataset. The combined approach is tested only on the WSJ corpus. The performance of the combined approach is presented in Fig. 3. The performance of the combined approach is higher than the baseline but comparable to both the individual learning strategies. The performance of combined approach in terms of test accuracies in Table.III

Fig. 3: Performance of the system obtained by combining both the curriculum learning approaches (i.e., corpus re-arranging methods and broadclass continuation methods). Figures are generated from the subset of WSJ corpus mentioned in subsection II-A

| TABLE II: Performance of the proposed broadclass continuous method |
|---------------------|--------|--------|
|                     | WSJ    | TIMIT  |
| Baseline            | 52.66  | 60.82  |
| Broadclass based continuation method | 53.22  | 60.9   |

<p>| TABLE III: Performance of combined approach in terms of test accuracies |
|---------------------|--------|--------|</p>
<table>
<thead>
<tr>
<th>corpus re-arranging strategy</th>
<th>WSJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>52.66</td>
</tr>
<tr>
<td>broadclass + 30 epochs easyset + 100 epochs-trainset</td>
<td>53.33</td>
</tr>
<tr>
<td>broadclass + 50 epochs easyset + 100 epochs-trainset</td>
<td>53.35</td>
</tr>
<tr>
<td>broadclass + 60 epochs easyset + 100 epochs-trainset</td>
<td>53.31</td>
</tr>
</tbody>
</table>
The performances of two step learning strategy (re-arranging strategy) and the broadclass continuation methods are compared and the results of comparisons presented in Fig. 4. It is observed that corpus-re-arranging strategy performed slightly better than the broadclass continuation methods in terms of accuracy, but the rate of convergence of broadclass continuation is higher than the corpus re-arranging strategy systems.

V. CONCLUSION AND FUTURE SCOPE

In this work, we have explored curriculum learning strategies for speech recognition. Learning strategies aimed at exploring the starting small paradigm at data and model level have been explored for the task of speech recognition i.e., re-arranging the data in a meaningful order to gradually increase the complexity and constraining the model resources initially and gradually increasing the resources to full capacity of the model.

We propose to use the posterior probabilities from a pre-trained model to rank the easiness of the training example for the present classification task and the ranking is used to reorganize the data in an increasing order of toughness to classify the example. A two step learning strategy (corpus re-arranging strategy) to initially train the model with the easy set of examples and then with the full training dataset is used. The proposed approach has performed the phone classification task better than the baseline which is trained by randomly sampling the dataset. We have observed a consistent improvement in the a performance across the datasets.

Apart from corpus re-arranging strategy we have also explored curriculum learning techniques aimed at constraining the model level resources which are also studied as continuation methods. Broad manner class detection objective function is considered as a smoother version of phone classification objective function. The proposed learning strategy i.e., the model initially trained with broadclass objective function later adapted for the phone classification objective function has performed better than the model directly trained with the phone classification objective function.

During this study, we have attempted to combine both the approaches but the performance of the combined approach is not higher than the individual learning strategies. We hypothesize that as both the approaches are aimed to exploit similar property of starting small in two different perspectives (data and model) so, no significant improvement in the performance of the combined approach has been observed. Curriculum learning techniques are closely related to the multi-task learning and transfer learning in the coming future we are very interested to apply curriculum learning strategies for those tasks.

REFERENCES