Hindi and Punjabi Continuous Speech Recognition Using CNSVM

Vishal Passricha, National Institute of Technology, Kurukshetra, India
Shubhanshi Singhal, Technical Education and Research Integrated Institute, Kurukshetra, India

ABSTRACT

CNNs are playing a vital role in the field of automatic speech recognition. Most CNNs employ a softmax activation layer to minimize cross-entropy loss. This layer generates the posterior probability in object classification tasks. SVMs are also offering promising results in the field of ASR. In this article, two different approaches: CNNs and SVMs, are combined together to propose a new hybrid architecture. This model replaces the softmax layer, i.e. the last layer of CNN by SVMs to effectively deal with high dimensional features. This model should be interpreted as a special form of structured SVM and named the Convolutional Neural SVM. (CNSVM). CNSVMs incorporate the characteristics of both models which CNNs learn features from the speech signal and SVMs classify these features into corresponding text. The parameters of CNNs and SVMs are trained jointly using a sequence level max-margin and sMBR criterion. The performance achieved by CNSVM on Hindi and Punjabi speech corpus for word error rate is 13.43% and 15.86%, respectively, which is a significant improvement on CNNs.

KEYWORDS

ASR, CNSVM, Maximum Margin, sMBR, sMM

1. INTRODUCTION

The process of mapping spoken utterances into the corresponding text is known as speech recognition. The emergence of CNN-based acoustic models improves the performance of speech recognition task significantly (Abdel-Hamid et al., 2012, 2013; Passricha & Aggarwal, 2018; Sainath et al., 2013; Singhal et al., 2018). CNNs efficiently model the temporal correlation with adaptive content information. In most of the CNNs, the last layer of the architecture is the softmax layer that is used for classification purpose. SVMs also offer good recognition rate in speech recognition task (Shi-Xiong & Gales, 2013). The basic principle of SVMs is maximizing the margin between two classes. The classification in which whole data points are classified into two-classes is known as linear classification and such SVM as linear SVM (Vapnik, 1995). Its extended version that is used for sequence recognition is known as structured SVM. SVM has many unique features: first, it maximizes the margin between the classes and it minimizes an upper bound of generalization errors. Second, SVM guarantees to have a global optimal solution because SVM’s nature is convex for optimization problems. Third, the size of SVM can easily be determined by calculating the number of support vectors that are trained during training. However, SVMs have some shortcomings like there is a need to predefine a feature space, and it does not deal well with arbitrary temporal dependencies and variable length data sequences.

In continuous speech recognition task, some training points are more important than others. Thus, high priority should be assigned to useful training points and low priority should be given
to other training points. SVMs have less effect of outliers and noise so they effectively classify the acoustic features. Inspired by the advantages of both and to overcome the limitations of SVMs, a new model named as convolutional neural support vector machine (CNSVM) is proposed by combining CNNs and SVMs. In this model, SVMs simply replace the last layer of CNNs i.e. the softmax layer. Earlier, Suykens and Vandewalle (1999) proposed MLPs and SVMs as a combined model where SVM optimizes the weights of MLP. Nagi et al. (2012) proposed a similar model convolutional neural support vector machines for multi-robot systems. Instead of training each component separately or jointly, they trained their model using a new training approach. They allowed the model for computationally light incremental learning based on small chunks of training samples. It is tested for finger-count 0 to 5 from finger-count gestures. However, CNSVMs differ from old ones because, in this model, parameters are trained jointly, not incrementally. Agarap (2017) also proposed an architecture by combining CNN and SVM for image classification. This architecture was trained using SGD training method and achieved a test accuracy of 90.72% while the CNN-softmax achieved a test accuracy of 91.86% on popular image recognition dataset MNIST.

Vesely et al. (2013) investigate the sequence discriminative training of neural networks and find that minimum Bayes risk (MBR) criterion performs better than maximum mutual information (MMI). sequence level max-margin (sMM) and state-level minimum Bayes risk (sMBR) are different algorithms that train the parameters of CNNs and SVMs jointly. In this work, the softmax layer is replaced by SVMs and its advantages are explored. The proposed model uses CNNs to learn the feature space from the speech signal. CNNs also handles the temporal dependencies. SVMs perform the classification of the acoustic frame. In CNSVMs, the scores from the SVMs replace the frame-level posterior probability. CNSVM is evaluated on Hindi and Punjabi speech datasets for word error rate (WER) and it achieves 13.43% and 15.86% WER for Hindi and Punjabi speech respectively. Experimental results show that the CNSVMs offer a relative reduction of 4.84% and 4.28% in WER when compared to CNN on Hindi and Punjabi speech datasets respectively.

The rest of the paper is structured as follows: section 2 represents related work on Hindi and Punjabi ASRs and the brief introduction about CNNs & SVMs. In section 3, the compatibility between the softmax activation layer and SVM is shown. Sequence-level max-margin training and state-level minimum Bayes risk training is covered in section 4 and 5 respectively. Section 6 represents the idea of convolutional neural support vector machines. Section 7 provides the details about Hindi and Punjabi language datasets. The detail about the experiment setup and results are given in section 8. Finally, the conclusion of the paper is given in section 9.

2. RELATED WORK

2.1. Hindi and Punjabi ASRs

A large amount of available training and testing data is the backbone of any successful speech recognition system. However, Indian languages like Hindi, Punjabi, Bengali, etc. are lacking in state-of-the-art datasets. Currently, limited work has been done for Hindi and Punjabi language ASRs.

In 2011, Aggarwal and Dave (2011) firstly applied discriminative training methods i.e. MMI, minimum classification error (MCE), and minimum phoneme error (MPE) for the training of the Hindi ASR. They used only 500 words limited vocabulary and showed MPE is superior over MMI and MCE for clean speech. In 2012, Aggarwal and Dave (2012a) optimized the central and side frequencies of MFCC filterbank using Genetic Algorithm and Particle Swarm Optimization methods and claimed optimized features performs better than conventional MFCC features. In the same year, Aggarwal and Dave (2012b) combined conventional, hybrid, and segmental HMM using the ROVER system combination technique. This combined model is evaluated with bigram and trigram language modeling on Hindi speech dataset and claimed 4% reduction in WER as compared to the traditional method. In 2013, Aggarwal and Dave (2013) integrated the conventional feature extraction methods
Challenges Facing Municipal Wireless: Case Studies from San Francisco and Silicon Valley
Heather E. Hudson (2013). Social and Economic Effects of Community Wireless Networks and Infrastructures (pp. 12-26).
www.igi-global.com/chapter/challenges-facing-municipal-wireless/74445?camid=4v1a

TB-WPRO: Title-Block Based Web Page Reorganization
www.igi-global.com/article/wpro-title-block-based-web/59712?camid=4v1a