Sequential Classification Criteria for NNs in Automatic Speech Recognition

Guangsen WANG, Khe Chai SIM

School of Computing, National University of Singapore
COM1, 13 Computing Drive, Singapore 117417
wangguangsen@comp.nus.edu.sg, simkc@comp.nus.edu.sg

Abstract

Neural networks (NNs) are discriminative classifiers which have been successfully integrated with hidden Markov models (HMMs), either in the hybrid NN/HMM or tandem connectionist systems. Typically, the NNs are trained with the frame-based cross-entropy criterion to classify phonemes or phoneme states. However, for word recognition, the word error rate is more closely related to the sequence classification criteria, such as maximum mutual information and minimum phone error. In this paper, the lattice-based sequence classification criteria are used to train the NNs in the hybrid NN/HMM system and the tandem system. A product-of-expert-based factorization and smoothing scheme is proposed for the hybrid system to scale the lattice-based NN training up to 6000 triphone states. Experimental results on the WSJCAM0 reveal that the NNs trained with the sequential classification criterion yield a 24.2% relative improvement compared to the cross-entropy trained NNs for the hybrid system.

Index Terms: lattices, neural networks, discriminative training

1. Introduction

Discriminative training of acoustic models is widely used in state-of-the-art large vocabulary continuous speech recognition systems. Sequential classification criteria such as Maximum Mutual Information (MMI) [1] and Minimum Phone Error (MPE) [1] are commonly used to estimate the GMM/HMM parameters. On the other hand, discriminative classifiers, such as Neural Networks (NNs), have also been successfully integrated with HMMs to improve speech recognition performance. For example, the NN/HMM hybrid system [2] uses NNs instead of GMMs to model the HMM state emission probabilities. NNs have also been used to extract discriminative features, with which better HMM systems can be estimated. PCA projected log posteriori features have been introduced in a tandem connectionist model [3] and bottleneck features have also been extracted from NN hidden layer [4].

NN/HMM hybrid systems have been found to be high-quality phone recognizers [2]. However, the cross-entropy criterion commonly used to estimate the NN parameters suffers from two major limitations for word recognition: 1) It is a frame-based criterion which may not be optimum for sequential classification tasks such as speech recognition; 2) It does not take into consideration the language model during parameter optimization. Therefore, sequential classification criteria such as MMI and MPE may be more suitable. Lattice-based sequence classification optimization of the hybrid system was proposed in [5]. However, their study is relatively rudimental: the NN used has a very simple 3-layer structure and models only 384 quadphone context dependent states, clearly underparameterized for the English broadcast news task [5] which usually requires tens of thousands of contexts. Other works on global optimization and sequence classification of NNs [6, 7] are also mainly dealing with small tasks, e.g., phone recognition, for the sake of an efficient gradient computation and a proper representation of hypotheses.

In this paper, we investigate how the lattice-based sequence classification criteria can be incorporated in both hybrid and tandem systems with a much more complex NN model for the large vocabulary word recognition task. Maximum Mutual Information (MMI) [1] is used. For large tasks, predicting the triphone states using a single NN as in [5] requires the NN to have thousands of output units, both system complexity and robust estimation would become issues. In this paper, a framework based on product-of-expert (PoE) [8] is proposed to factorize the hybrid system into two sets of NNs, namely context independent NN (CI NN) and context dependent NNs (CD NNs). Meanwhile, the sequence classification training can be applied to both the CI NN and the CD NNs.

The remaining of the paper is organized as follows. A brief review of the cross-entropy criterion is given in section 2. Section 3 shows the relation between the lattice-based sequence classification criterion and the cross-entropy criterion and gives the methodology of training NNs using such sequence-based criteria. The PoE-based factorization and smoothing paradigm is given in section 4. This is followed by some implementation issues discussed in section 5. Experimental results are presented in section 6. Section 7 summarizes the findings of the study and concludes the paper.

2. The cross-entropy criterion

Cross-entropy is a frame-based criterion widely adopted to optimize the NN parameters. Let \( \mathcal{O} \) denote the \( r \)-th training utterance and \( Y_r \) the corresponding phone state sequence. The label sequence \( Y_r \) has the target class labels \( y_{rt} \) for each frame \( t \) of \( \mathcal{O} \). These target labels are usually obtained through a forced alignment using the word-level transcription \( S \). Let \( T_r \) denote the frame number of \( \mathcal{O} \). The cross-entropy criterion can be formulated as [5]:

\[
F_{\text{XENT}}(\theta) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{i=1}^{N} \hat{y}_{rt}(i) \log \frac{\hat{y}_{rt}(i)}{y_{rt}(i)},
\]

where \( \theta \) is the NN parameter set, \( N \) is the target label size, i.e., the total number of the phone states, \( y_{rt}(i) \) is the output of the NN for phone states \( i \) at frame \( t \) of \( \mathcal{O} \). During NN training, the error back-propagation (EBP) adjusts \( \theta \) so that \( F_{\text{XENT}} \) is minimized. Softmax is usually used as the NN output nonlinearity for classification purpose:

\[
y_{rt}(i) = \frac{e^{\theta y_{rt}(i)}}{\sum_{j=1}^{N} e^{\theta y_{rt}(j)}},
\]

Copyright © 2011 ISCA
where \( a_{rt}(i) \) is the input to the output unit \( i \) before the softmax activation function. If the cross-entropy is used with the softmax nonlinearity, the derivative of the \( F_{\text{XENT}} \) loss function with respect to \( a_{rt}(i) \) can be expressed as [5]:

\[
\frac{\partial F_{\text{XENT}}}{\partial a_{rt}(i)} = y_{rt}(i) - \hat{y}_{rt}(i).
\]  

(3)

3. Sequential classification criterion

Sequential classification criteria, such as MMI and MPE, can be optimized using extended Baum-Welch (EBW) [1] where lattices are used as a compact representation for the hypotheses and the references space. Note that, for the ease of reference, the notations and formulae given in this section are borrowed from [5]. There are two essential terms in the standard EBW update formulae [1], namely \( \gamma_{rt}^{DEN}(i) \) and \( \gamma_{rt}^{NUM}(i) \). They are the expect occupancies that the frame \( t \) of utterance \( r \) resides in phone state \( i \). The occupancies can be obtained through a standard forward-backward procedure over the numerator and denominator lattice. For the sequence classification criterion, the gradient of loss with respect to the HMM state log-likelihood used in the EBW updates is calculated as [1]:

\[
\frac{\partial F_{\text{SEQ}}}{\partial l_{rt}(i)} = k(\gamma_{rt}^{DEN}(i) - \gamma_{rt}^{NUM}(i)),
\]  

(4)

where \( F_{\text{SEQ}} \) can be any sequence classification criterion, \( l_{rt}(i) \) is the log-likelihood of phone state \( i \) at frame \( t \) in utterance \( r \), \( k \) is a scaling factor to enhance generalization. In hybrid NN/HMM, the scaled likelihood \( y_{rt}(i)p(i) \) is used to represent the emission probability of the HMM state, where \( p(i) \) is the prior of physical state \( i \) which is estimated using relative frequency. Therefore, the log-likelihood term can be expressed as \( l_{rt}(i) = \log y_{rt}(i) - \log p(i) \). The derivative of \( F_{\text{SEQ}} \) with respect to the NN output \( y_{rt}(i) \) can be derived by applying the chain rule [5]:

\[
\frac{\partial F_{\text{SEQ}}}{\partial y_{rt}(i)} = k\frac{\gamma_{rt}^{DEN}(i) - \gamma_{rt}^{NUM}(i)}{y_{rt}(i)}.
\]  

(5)

Moreover, if softmax is used as the output nonlinearity, the gradient with respect to the softmax activations \( a_{rt}(i) \) under the sequence classification criterion becomes [5, 7]:

\[
\frac{\partial F_{\text{SEQ}}}{\partial a_{rt}(i)} = k(\gamma_{rt}^{DEN}(i) - \gamma_{rt}^{NUM}(i)).
\]  

(6)

By comparing Eqn. 3 and Eqn. 6, the frame-based gradient \( y_{rt}(i) - \hat{y}_{rt}(i) \) is replaced with the lattice-based sequence classification gradient \( k(\gamma_{rt}^{DEN}(i) - \gamma_{rt}^{NUM}(i)) \) with an additional scaling factor \( k \). Therefore, the NN can be trained using a lattice-based sequence classification criterion under the EBW framework: the gradient is calculated through a forward-backward procedure on the lattices (Eqn. 6) instead of using the frame-based gradient in the cross-entropy criterion (Eqn. 3).

4. PoE-based factorization and smoothing

In previous work [5], sequential classification criterion has been applied to small NN/HMM systems. Although improvements were observed compared to the cross-entropy criterion, given the simple structure of their NN, it was not clear whether the NN systems would still outperform the GMM systems if more complex model is used. In this paper, we have successfully scaled the lattice-based sequence classification framework to larger context dependent (CD) networks under the product-of-expert (PoE) framework [8]. According to this framework, CI state posteriors are regarded as \textit{experts} and a set of 2-layer CD NNs are built to map the CI log posterior to CD state posteriors. The overall system architecture is depicted in Fig. 1. The scaled likelihood expression for each CD state is given by:

\[
\frac{p(o_{i}|m_{i}, s_{j}, c_{k})}{p(o_{i})} = \frac{P(m_{i}, s_{j}, c_{k}|o_{i})}{P(m_{i}, s_{j}, c_{k})} = \frac{P(m_{i}, s_{j}|o_{i})P(c_{k}|m_{i}, s_{j}, f_{i})}{P(m_{i}, s_{j}, c_{k})}
\]

where \( m_{i}, s_{j} \) and \( c_{k} \) denote the \( i \)th monophone, \( j \)th state and \( k \)th context cluster respectively. \( o_{i} \) is the observation at time \( t \). The class prior probability \( P(m_{i}, s_{j}, c_{k}) \) can be estimated using relative frequency. The posterior terms \( P(m_{i}, s_{j}|o_{i}) \) and \( P(c_{k}|m_{i}, s_{j}, f_{i}) \) can be predicted using a cascade of NNs through the PoE factorization scheme as shown in Fig. 1. \( P(m_{i}, s_{j}|o_{i}) \) is produced by a CI NN and \( P(c_{k}|m_{i}, s_{j}, f_{i}) \) is generated by a set of CD NNs whose inputs are the log posterior probabilities of CI states, \( f_{i} = \{ \log P(m_{i}, s_{j}|o_{i}) : \forall i \} \). To ensure the robust estimation of the conditional probabilities, Weighted Geometric Smoothing (WGS) [8] is applied such that the smoothed CD state posteriors are given by:

\[
P(m_{i}, s_{j}, c_{k}|o_{i}) = P(c_{k}|m_{i}, s_{j}, f_{i})^{\alpha} P(m_{i}, s_{j}|o_{i})^{1-\alpha}
\]

where \( \alpha (0 \leq \alpha \leq 1) \) is the smoothing factor. It is interesting to note that the proposed PoE CD NN/HMM system is essentially an instance of canonical state model [9]. CI NN is the canonical state and the CD NNs are the discriminatively learned non-linear transforms for context dependent state modeling.

5. Implementation issues

The sequence classification based NN training framework has several implementation issues that are worth further explanations. The numerator and denominator lattices are generated by a cross-entropy trained hybrid NN/HMM system. These lattices are generated once and used for all subsequent NN training iterations. During each training iteration, lattice forward-backward is performed to compute the gradient of the sequential classification criterion in Eqn. 6 for each frame. Error backpropagation (EBP) is then performed to update the NN parameters. The updated NN/HMM model is then used to perform the next iteration of lattice forward-backward. During each iteration, if the accuracy on the held-out set fails to improve for one epoch, the NN weights are restored with the values of the previous epoch but with the learning rate halved. The training ceases if the accuracy again fails to increase. The NNs are initialized using frame-based cross-entropy learning.
6. Experiments

6.1. Experimental setup

This section presents the experimental results comparing the frame-based and sequential classification criterion for NNs using the WSJCAM0 [10] corpus. There are 18.3 hours of training data, comprising 9889 utterances. The 5k WSJ0 tasks are used for performance evaluation. The “si_dt5a” set is used as the development set to tune the smoothing factors; “si_dt5b” is used as the test set for evaluation purpose. The phone set has 41 monophones including one silence model “sil” and one short pause model “sp”. The features for CI NN and the bottle-neck NN are the standard 39-dimensional MFCC which consists of 13 static coefficients (12 MFCC plus one C0 energy term) and its first and second derivatives. HTK1 was adopted for decoding and NN models were trained with a modified version of QuickNet2. Word recognition was performed using a bigram full decoding followed by a trigram rescoring.

The hybrid system is built under the PoE factorization framework: a 3-layer CI NN (585 x 2000 x 120) was firstly trained. Its input is a window of 15 frames of MFCC vectors (585 dimension). The output corresponds to the posterior probabilities of 120 physical states (40 monophones with 3 states each; “sp” comprise one emitting state which is tied to the centre state of “sil”). The size of the hidden units was chosen to optimize development set performance. All CD NNs are 2-layer networks trained with the CI log posteriors as input features. There are 117 CD NNs, each corresponds to one of the 117 monophone states excluding three “sil” states since they are modeled without contexts. The posteriors of the “sil” states are obtained directly from the output of the CI NN. Therefore, they have an input layer of 120 units. Each CI NN is trained to predict the conditional posteriors for all the 50 triphone state contexts given the CI state. The contexts are obtained using the conventional decision tree clustering technique [11]. Therefore, the CD NNs have 50 output units and about 6000 contexts are modelled by these CD NNs. The total number of parameters in the hybrid system is approximately 2.1 million.

For the bottle-neck tandem system, a 4-layer NN (585 x 2000 x 39 x 120) is trained to extract the bottle-neck features. This NN has a similar structure to the CI NN described above, with an additional hidden layer of size 39. This is the bottle-neck layer from which the 39-dimensional bottle-neck features will be extracted. The dimension of the bottle-neck features was chosen to match the dimension of the MFCC features.

6.2. Sequence classification based CI NN/HMM

Firstly, we compare the performance of CI NN/HMM systems trained using both the cross-entropy and lattice-based sequence classification criteria. The Word Error Rate (WER) performance of CI NN/HMM systems trained with MMI and MPE criteria are compared in Fig. 2. The solid lines and dotted lines correspond to the WER performance on “dt5a” and “dt5b” respectively. The performance at iteration 0 corresponds to the NN/HMM systems trained with the frame-based cross-entropy criterion: 12.41% (dt5a) and 13.87% (dt5b). Note the large WER improvements after one iteration. The MMI criterion consistently outperforms MPE on both “dt5a” and “dt5b”. The improvements begin to saturate after the 7th iteration and finally the system achieves a relative word error reduction of

Table 1: WERs of three systems with the same parameter size.

<table>
<thead>
<tr>
<th>System</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM/HMM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML</td>
</tr>
<tr>
<td></td>
<td>11.88</td>
</tr>
<tr>
<td></td>
<td>10.06</td>
</tr>
<tr>
<td>bottle-neck Tandem</td>
<td></td>
</tr>
<tr>
<td>XENT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML</td>
</tr>
<tr>
<td></td>
<td>9.31</td>
</tr>
<tr>
<td></td>
<td>9.15</td>
</tr>
<tr>
<td>NN-MMI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ML</td>
</tr>
<tr>
<td></td>
<td>7.61</td>
</tr>
<tr>
<td>Hybrid NN/HMM</td>
<td></td>
</tr>
<tr>
<td>XENT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CI NN</td>
</tr>
<tr>
<td></td>
<td>12.41</td>
</tr>
<tr>
<td></td>
<td>13.87</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NN-MMI</td>
</tr>
<tr>
<td></td>
<td>CI NN</td>
</tr>
<tr>
<td></td>
<td>8.54</td>
</tr>
<tr>
<td></td>
<td>10.23</td>
</tr>
</tbody>
</table>

26% for MMI and 24.6% for MPE. Therefore, in the following experiments, only MMI performance is reported for all systems.

6.3. Sequence classification based CD NN/HMM

In this section, the WER performance of the hybrid system (NN/HMM) and the tandem system based on the bottle-neck features (bottle-neck Tandem) using both the cross-entropy and MMI criterion will be reported on both “dt5a” and “dt5b”. Table 1 compares the WER performance of various systems with comparable model sizes. The GMM/HMM baseline system has 6 Gaussian components per state. The Maximum Likelihood (ML) trained system achieved 11.88% and 13.07% WER on dt5a and dt5b respectively. The MMI baseline achieved WERs of 10.06% and 10.87%. The bottle-neck tandem system uses a 2-component GMM/HMM system trained with the bottle-neck feature extracted from a 4-layer CI NN trained with either the cross entropy (XENT) or the MMI (NN-MMI) criterion. In both cases, using MMI criterion to train the HMM parameters for the bottle-neck tandem system achieved consistent relative improvements of approximately 17.51% over the ML systems. The quality of the bottle-neck features does not seem to be very much affected by the criteria used to train the CI NN. Both XENT and NN-MMI criteria have comparable WER performance. We do not expect the bottle-neck tandem system to benefit much from NN-MMI training of the CI NN because NN-MMI was designed to specifically to improve the MMI criterion of the NN/HMM system. Furthermore, the CI NN in the bottle-neck tandem system functions as a feature extractor and

1Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk

Figure 2: WERs as a function of MMI and MPE iterations for CI NN/HMM systems.
Table 2: Comparison of the best WER performance among three systems.

<table>
<thead>
<tr>
<th>System</th>
<th>No. of Parameters</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM/HMM</td>
<td>—</td>
<td>6.48</td>
</tr>
<tr>
<td>bottle-neck</td>
<td>XENT</td>
<td>6.08</td>
</tr>
<tr>
<td>Tandem</td>
<td>NN-MMI</td>
<td>6.12</td>
</tr>
<tr>
<td>Hybrid</td>
<td>XENT</td>
<td>8.22</td>
</tr>
<tr>
<td>NN/HMM</td>
<td>NN-MMI</td>
<td>6.20</td>
</tr>
</tbody>
</table>

Table 3: Significant test results (p=0.05). The "*" indicates a significant difference at the significant level of 0.05.

| Bottle-neck Tandem | 0.024* | 0.660 |
| Hybrid NN/HMM      | 0.047* | –     |

the subsequent training of the GMM/HMM system already incorporates sequential criteria (ML or MMI). On the other hand, using the NN-MMI criterion to train the CI and CD NNs in the hybrid systems show significant improvements. For the CI NN/HMM system, using the XENT training gave 12.41% and 13.87% WERs while using NN-MMI training achieved 8.54% and 10.23% WERs. These translate to approximately 26.2% relative improvement. The CI NN/HMM system already outperformed the GMM/HMM triphone baseline system. With the smoothing and factorization scheme described in section 4, we extended the NN-MMI training for CD NNs with about 6000 contexts. The performance of the XENT trained CD NN/HMM system is 8.22% and 9.74% on the two data sets. With NN-MMI training, the WERs were lowered to 6.20% and 7.38%. These correspond to about 24.2% relative improvements.

The performance in Table 1 for the GMM/HMM system and the bottle-neck tandem system is not optimal. Therefore, the parameter size of these systems was increased by increasing the number of Gaussian components per HMM state to obtain the best performance. The performance of these systems are given in Table 2. Note the triphone HMMs in both the baseline and the tandem system are trained using MMI and only the CD NN results for the hybrid system are reported. Except for the XENT NN/HMM which gave the worst WER performance 9.74%, all the systems in Table 2 show comparable WER performance. This clearly illustrates the importance of using sequential classification learning for NN/HMM systems. Despite being the smallest model, the hybrid system trained with NN-MMI criterion outperformed the baseline GMM/HMM triphone system and the improvement was found to be statistically significant. Furthermore, it also gave performance comparable to the bottle-neck tandem system which is almost 3 times larger in model size. The performance difference between these two systems was found to be statistically insignificant. All the statistical significance tests were conducted on “dt5b” using SCTK with the best system configurations: the NNs in the CD hybrid NN/HMM and tandem system are trained using NN-MMI, the HMM parameters of the tandem system and the baseline GMM/HMM system are trained with MMI. The results are summarized in Table 3. There is no significant difference found between the CD NN hybrid system and the bottle-neck tandem system. However, both the hybrid system and the tandem system have a significant improvement over the baseline GMM/HMM system.

Table 3: Significant test results (p=0.05). The "*" indicates a significant difference at the significant level of 0.05.

9. References


7. Conclusions

This paper has investigated the use of lattice-based sequence classification criteria for the NN training in the hybrid system and the tandem system. A PoE factorization and smoothing scheme is proposed to scale up the lattice-based CD NN training in the hybrid system. The WER performance is evaluated on the WSJCAM0 corpus. For the hybrid system, NNs trained with the sequence-based criteria (MMI or MPE) significantly outperform the cross-entropy trained system for both CI NN and CD NN. However, for the bottle-neck tandem system, the use of sequence-based criterion in the bottle-neck NN training does not lead to a substantial improvement. The MMI trained NN/HMM system performs significantly better than the baseline GMM/HMM system and the bottle-neck tandem system with similar parameter size. However, with much larger model size, the GMM/HMM and the bottle-neck tandem systems also achieved comparable performance to NN/HMM. Further work includes increasing the NN parameter size, e.g., instead of predicting phone states, the CI NN can be trained at the mixture level. Applying discriminative context clustering techniques [12, 13] to the CD NN training is also an interesting direction.

8. Acknowledgements

This research is done for CSIDM Project No. CSIDM-200806 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

9. References