INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Deep Neural Network Calibration for E2E Speech Recognition System

Mun-Hak Lee and Joon-Hyuk Chang

Department of Electronics Engineering
Hanyang University, Seoul, Republic of Korea
lullaby0804@hanyang.ac.kr, jchang@hanyang.ac.kr

Abstract

Cross-entropy loss, which is commonly used in deep-neural-network-based (DNN) classification model training, induces models to assign a high probability value to one class. Networks trained in this fashion tend to be overconfident, which causes a problem in the decoding process of the speech recognition system, as it uses the combined probability distribution of multiple independently trained networks. Overconfidence in neural networks can be quantified as a calibration error, which is the difference between the output probability of a model and the likelihood of obtaining an actual correct answer. We show that the deep-learning-based components of an end-to-end (E2E) speech recognition system with high classification accuracy contain calibration errors and quantify them using various calibration measures. In addition, it was experimentally shown that the calibration function, which was being trained to minimize calibration errors effectively mitigates those of the speech recognition system, and as a result, can improve the performance of beam-search during decoding.

Index Terms: E2E speech recognition, deep neural network calibration

1. Introduction

Speech recognition is the task of finding the most appropriate word sequence W, corresponding to a given speech X. By modeling such a problem probabilistically, it can be expressed as follows:

$$P(W|X) \propto P(X|W)P(W),$$

where the posterior probability $P(W|X)$ is expressed as the product of the prior $P(W)$ and conditional probabilities $P(X|W)$ using Bayes' rule. The trained automatic speech recognition (ASR) system searches for this probability distribution in the decoding process, and outputs a word sequence with the highest probability value. The practical benefit of dividing the speech recognition system into two units is that it can be used to improve speech recognition performance by modeling the prior probabilities of strings with a language model (LM) that is independently trained using unpaired text data. Therefore, many speech recognition systems train $P(X|W)$ and $P(W)$ separately, and find the word sequence with the highest probability by searching the joint probability distribution. As a representative example, the hidden-Markov-model-based (HMM) speech recognition system combines the output of the acoustic model (AM) and the language model using a weighted-finite-state-transducer. The end-to-end (E2E) speech recognition system also improves recognition performance by combining a language model trained from external texts with a speech recognition system via shallow fusion.

Therefore, to ensure high speech recognition performance, it is important for the output distributions of two independently trained classification models to approximate the actual probability distribution well. Among the algorithms applied to modern speech recognition, deep learning performs best. A deep neural network (DNN) estimates the probability distribution by normalizing the output of a multilayer neural network and is trained using cross-entropy loss. DNNs first replaced the acoustic model for HMM-based speech recognition systems, and studies recently been conducted to apply a deep-learning-based language model to speech recognition systems. In addition, there is active research on E2E speech recognition systems that model an entire ASR system as a single neural network, such as the connectionist temporal classification (CTC) or attention-based sequence-to-sequence (Seq2Seq) model [1, 2].

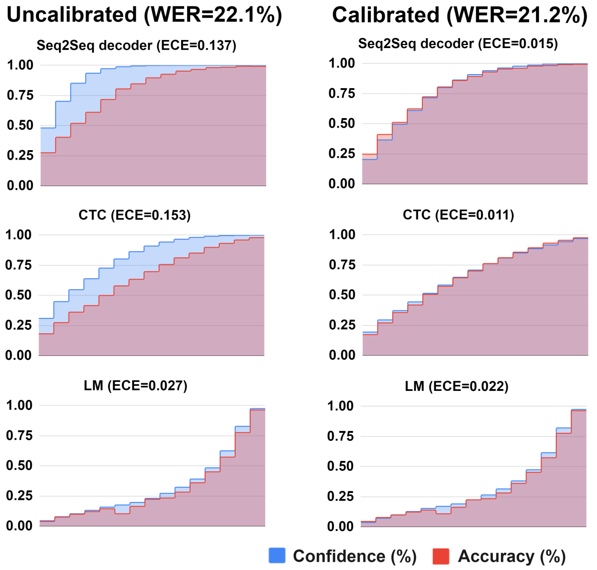

Figure 1: *Calibration error of E2E ASR system components: this graph shows the bin-wise* ***accuracy(red)*** *and* ***confidence(blue)*** *for the top 1 class. While LM is relatively well calibrated, it can be seen that CTC and Seq2seq decoder tend to overconfide in the class with the highest probability. We trained the system with the 100h trainset from the Librispeech dataset and measured WER with the test-other from the Librispeech dataset.*

While DNNs exhibit remarkable classification accuracy, some recent studies revealed that very complex and sophisticated deep neural networks are frequently overfitted on the training dataset, and they generate overconfidence probability distribution than the actual distribution [3, 4]. This is called a calibration problem, and in [3, 5], the difference between the output distribution of the DNNs and the distribution of the ac-

http://dx.doi.org/10.21437/Interspeech.2021-176

tual correct answer label is quantitatively measured using various calibration measures. Overconfidence can adversely affect beam search during decoding in speech recognition systems that use joint probability distributions derived from independently trained models [4]. For example, if the entropy values of the output probability distribution of the independently trained acoustic and the language models are significantly different, the final combined probability distribution follows the probability distribution with low entropy, regardless of the actual performance. As a result, this raises a problem in that simply improving the accuracy of individual components does not guarantee a reduction in the word error rate (WER) of the combined ASR system.

In this study, we experimentally confirm the calibration error of neural-network-based components (CTC, attention-based Seq2Seq and LM) that were used in the E2E speech recognition system and quantified it through various calibration measures. In addition, we conducted experiments using the latest calibration methods that can reduce calibration errors and demonstrated that these methods improved the performance of beam-search during decoding in speech recognition systems.

2. Miscalibration of E2E ASR system

The goal of model calibration is to approximate the true probability distribution of target Y using the output probability distribution $\hat{\mathbf{p}}$ of the k-class classification model.

$$P(Y = i|\hat{\mathbf{p}}(X) = \mathbf{p}) = p_i \quad \text{for } i = 1, ..., k$$

where $\mathbf{p} = (p_1, ..., p_k)$. For this, we calculate the classification accuracy of the network by comparing the predicted value of the network $\hat{Y}$ with the actual correct answer Y. Next, we measured how much the normalized output distribution $\hat{\mathbf{p}}(X)$ represents the classification accuracy of the network. This is generally evaluated using the validation set, and the difference between the bin-wise accuracy and bin-wise confidence of the validation set was considered a calibration error.

The attention-based E2E model has a Seq2Seq network structure and estimates the probability distribution of the token existence of the current time step using the output of the previous time step and feature sequence. To measure the calibration error of the sequence classification model, we assume conditional independence of time and use the ground truth label of the previous time step as the input of the decoder instead of the output of the previous time step. Based on this assumption, the E2E speech recognition model can be considered as a time-independent token classifier, and the calibration error of the model can be measured using multi-class calibration measures. The most representative calibration measure is the expected calibration error (ECE) in Eq. (1).

$$ECE = \sum_{i=1}^{b} \frac{|B_i|}{n} \left| acc(B_i) - conf(B_i) \right|, \tag{1}$$

$$acc(B_i) = \frac{1}{|B_i|} \sum_{m \in B_i} \mathbf{1}(\hat{y}_m = y_m),$$

$$conf(B_i) = \frac{1}{|B_i|} \sum_{m \in B_i} \hat{p}_m,$$

where b is the number of bins and n is the total number of data points. The entire batch is divided into bins (B) and the distance between the bin-specific accuracy and the average confidence is defined as the calibration error. In addition, we extend this to multi-class problems by measuring the ECE classwise.

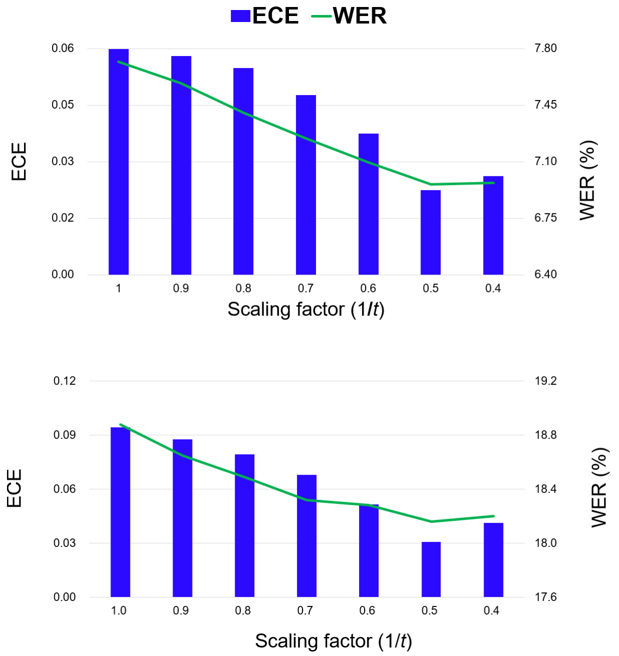

Figure 2: *We separately evaluated $ECE_{adaptive}$ and WER in LibriSpeech test-clean (top) and test-other (bottom). This graph shows a strong correlation between calibration error and WER.*

$$ECE_{classwise} = \frac{1}{k} \sum_{j=1}^{k} \sum_{i=1}^{b} \frac{|B_{ij}|}{n} \left| acc(B_{ij}) - conf(B_{ij}) \right|,$$

where k denotes the number of classes. Class-wise ECE measures multi-class calibration errors, but if the number of classes gets large, the accuracy of most bins is reduced to approximately 0; this is not suitable for classification systems with many classes. The adaptive calibration error configures all bins ($\dot{B}$) to contain the same number of correct answer labels, thus allowing a higher weight to be provided to calibration errors in regions with high prediction values.

$$ECE_{adaptive} = \frac{1}{\dot{b}k} \sum_{j=1}^{k} \sum_{i=1}^{\dot{b}} \left| acc(\dot{B}_{ij}) - conf(\dot{B}_{ij}) \right|,$$

where $\dot{b}$ is the number of adaptive bins. It is also important to determine a suitable binning method for the calibration measurements. For this, we used a method of generating bins by sorting the classification results in mini-batch according to confidence scores. Alignment was performed once for each class and once for each mini-batch. If binning is performed in this way, the variance in the bin is minimized; this helps to identify calibration errors for each confidence value. Figure 1 shows the confidence and accuracy of the trained E2E ASR system.

3. Calibration methods for E2E ASR system

Network calibration is affected by various factors, such as regularization (weight decay and batch normalization) and model size (depth and hidden unit size). Therefore, it is very difficult to clearly identify the cause of calibration errors [3]. However, a few studies show that overconfidence is common in high-capacity models where overfitting occurs [3, 4, 6]. In addition to overconfidence, calibration errors are common in most DNNs, and various methods have been suggested to mitigate this problem. In this study, these calibration methods are applied to individual components of the speech recognition system to minimize the calibration errors of each component.

3.1. Label smoothing

The method that is often used to solve overconfidence in speech recognition systems is label smoothing. Label smoothing uses the target vector smoothed through the following equation for network training, and prevents the network from generating excessively large output values for one class. Many previous studies have shown that label smoothing is helpful for calibrating neural networks [3, 4, 6]. We conducted experiments to show that label smoothing prevents overfitting of the ASR system and lowers calibration errors in the test set (Table 1).

$$y_{smooth} = y_{1hot} - \epsilon(y_{1hot} - \frac{1}{k}y_{ones}),$$

where $y_{ones} = (1, \ldots, 1)$, y_{1hot} are the one-hot target vectors, k is the number of classes, and $\epsilon \in [0, 1]$, respectively.

3.2. Weighted sum

The most common method when combining the outputs of LM and AM during decoding in speech recognition systems is to weight sum the network output before going through softmax (logit) or log probability using an empirically determined weight value α_m, where m is the number of ASR system components. Most decoding processes in speech recognition systems adopt the weighted sum method. The limitation of this method is that the hyper parameters α_m must be manually determined through a grid search. Finding the optimal value is time-consuming, and the process does not guarantee that the probability distribution calculated using the obtained weight is well approximated to the actual probability distribution.

$$\hat{W} = \underset{W}{argmax} \; \{\alpha_1 \log(P(X|W)) + \alpha_2 \log(P(W))\}$$

3.3. Logit scaling

Logit scaling calibrates the output probability distribution by scaling the logit or log probability of a network, (similar to the aforementioned weighted sum method). We introduce three types of logit scaling methods. The first one is temperature scaling, a method that uses the same scaling factor for all classes in Eq. (2). The second one is a vector scaling, a method that performs scaling with independent scaling factors for each class in Eq. (3). The last one is matrix scaling, which performs scaling by considering the correlation between classes in Eq. (4) [3, 7]. Each method trains the parameters of the calibration function to minimize the cross entropy loss in the validation set while keeping the parameters of the classification network fixed. In this case, if the number of parameters of the calibration function is large, the function may be overfitted for the validation set. To prevent this problem, we regulate the elements of the function (V_m, M_m) that cause interclass polarization through l2-regularization [7].

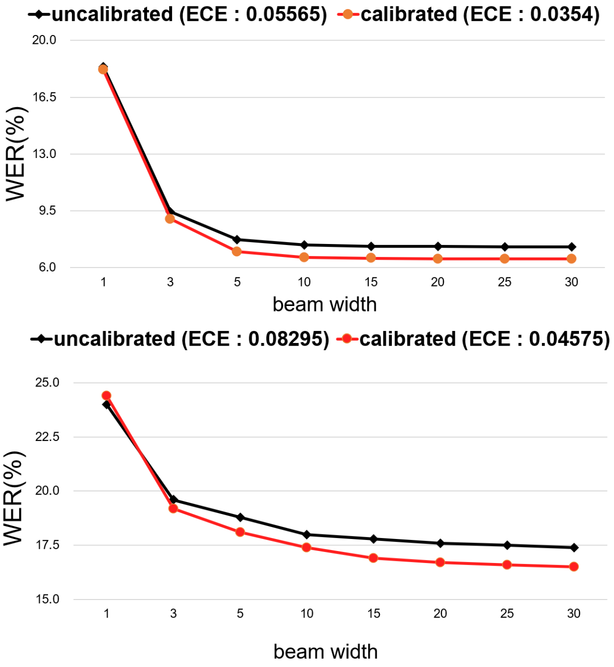

Figure 3: *We present the average $ECE_{adaptation}$ of each speech recognition system component and the WER for the Librispeech test-clean (top) and test-other (bottom) dataset. It is experimentally shown that the calibrated networks improve the performance of the combined speech recognition system via the beam search decoding process.*

$$\hat{W} = \underset{W}{argmax} \left\{ \frac{1}{t_1} \log(P(X|W)) + \frac{1}{t_2} \log(P(W)) \right\}, \tag{2}$$

$$\hat{W} = \underset{W}{argmax} \left\{ \frac{V_1 + I}{t_1} \log(P(X|W)) + \frac{V_2 + I}{t_2} \log(P(W)) \right\}, \tag{3}$$

$$\hat{W} = \underset{W}{argmax} \left\{ \frac{M_1 + I}{t_1} \log(P(X|W)) + \frac{M_2 + I}{t_2} \log(P(W)) \right\}, \tag{4}$$

where t_m is a scalar value, $V_m \in \mathbb{R}^{k \times k}$ is a $k \times k$ diagonal matrix which ${v^m}_{ij} = 0$ for all $i \neq j$, $M_m \in \mathbb{R}^{k \times k}$ is a $k \times k$ matrix and I is the identity matrix.

4. Experiments

4.1. Datasets

We demonstrate our experiments using the LibriSpeech dataset. The LibriSpeech dataset consists of a total of 460h of the clean training set, 500h of the more challenging training set, and separate validation and test sets. In all experiments, we trained the ASR system using only 100h of the training set. The test set of LibriSpeech is divided into “clean” and “other”.

4.2. E2E speech recognition system

For the experiment, we used a speech recognition system consisting of three types of deep-learning-based classifier modules such as CTC, Seq2Seq and LM. CTC and attention-based Seq2Seq play the same role in ASR systems, but they are trained

Table 1: *ECE and the cross-entropy loss of Seq2Seq decoder. We denoted k^{th} best predicted class ECE as ECE_k*

	LibriSpeech train-clean-100				
	Cross Entropy	$ECE_{adaptive}$	ECE_1	ECE_2	ECE_3
Uncalibrated	5.78E-05	5.66E-05	5.65E-05	2.63E-06	1.92E-07
Label smoothing	7.76E-02	7.19E-02	7.20E-02	3.86E-03	1.28E-03
Temperature scaling	1.82E-03	1.68E-03	1.67E-03	6.97E-04	1.70E-04
Matrix scaling	4.63E-03	4.13E-03	4.13E-03	2.03E-03	4.24E-04

	LibriSpeech test-clean				
	Cross Entropy	$ECE_{adaptive}$	ECE_1	ECE_2	ECE_3
Uncalibrated	9.46E-01	5.97E-02	8.31E-02	2.50E-02	1.17E-02
Label smoothing	6.07E-01	6.08E-02	6.53E-02	1.89E-02	6.89E-03
Temperature scaling	5.59E-01	2.61E-02	3.20E-02	1.38E-02	7.26E-03
Matrix scaling	5.45E-01	2.52E-02	3.07E-02	1.27E-02	6.84E-03

	LibriSpeech test-other				
	Cross Entropy	$ECE_{adaptive}$	ECE_1	ECE_2	ECE_3
Uncalibrated	1.56E+00	9.42E-02	1.43E-01	3.98E-02	1.90E-02
Label smoothing	1.09E+00	4.92E-02	5.39E-02	3.38E-02	1.43E-02
Temperature scaling	1.02E+00	3.98E-02	4.98E-02	2.29E-02	1.19E-02
Matrix scaling	9.71E-01	3.75E-02	4.79E-02	2.05E-02	1.16E-02

with different inductive biases, and a model that combines the output probability distributions of the two performs better than individual models [8]. We conducted an experiment using ESPNet [8], a speech recognition toolkit based on Pytorch [9]. We used a 12-layer transformer-based encoder, and let the CTC and Seq2Seq-decoder share the same encoder. We used a 6-layer transformer as the decoder of the Seq2Seq network. For language model training, we used a network structure in which four layers of LSTM were stacked after the embedding layer. Next, CTC, Seq2Seq, and LM were trained by targeting 5,000 subwords using the same technique in [10]. The CTC and Seq2Seq networks were trained using the 80-dimensional filter bank feature, and spec-augmentation was applied to ensure stable performance [11].

5. Discussion

The aforementioned calibration methods were applied to the E2E ASR system, and Table 1 shows that these methods effectively reduce the calibration error in the network. The correlation between the calibration error and WER is shown in Figure 2 and 3. As shown in Figure 2, the scaling factor of the Seq2Seq decoder was changed to show the correlation between $ECE_{adaptive}$ and WER. The WER was calculated using the combined probability distribution of CTC, Seq2Seq decoder, and LM, and in the case of CTC and LM, we used the calibrated models. As shown in Figure 3, the WER for each beam width was measured, and the calibrated networks improved the performance of the combined speech recognition system (using the beam search). As a result, calibrating the output of the network improves speech recognition performance, which demonstrates that the combined probability distribution of calibrated networks approximates the actual probability distribution better than that of uncalibrated networks.

Next, the significance of our experiments is discussed.

- The calibration error of the speech recognition system components were quantified using various calibration measures. The proposed calibration measure also considers the errors of classes that are not correct answers that cross entropy can not consider; therefore, it is possible to measure the distance between the true distribution and the prediction more precisely.

- Calibration methods that can mitigate the calibration errors of components of a speech recognition system were proposed. The methods reduce the overconfidence in neural networks with only a small amount of computation and effectively improve the recognition performance of the speech recognition system.

- The proposed calibration functions are numerically trained, and a manual hyperparameter searching process is unnecessary. In addition, the proposed logit scaling methods can effectively replace the weighted sum method, as they minimize the calibration errors more effectively than the weighted sum method.

6. Conclusions

In this study, we quantified the calibration error of the E2E speech recognition system using the calibration measures of deep learning models and proposed some algorithms to mitigate this error. The proposed algorithms require only a small amount of computation and improve the recognition performance of the speech recognition system. There is room for our research to expand into confidence filtering, in which the confidence score of the speech recognition results is used.

7. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00456, Development of Ultra-high Speech Quality Technology for Remote Multi-speaker Conference System)

8. References

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," 2006, pp. 369–376.

[2] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," 2015.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *International Conference on Machine Learning*, 2017.

[4] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *INTERSPEECH*, 2017.

[5] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning." *CVPR Workshops*, vol. 2, no. 7, 2019.

[6] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" *Advances in Neural Information Processing Systems*, 2019.

[7] M. Kull, M. Perello-Nieto, M. Kängsepp, H. Song, P. Flach *et al.*, "Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration," *Advances in Neural Information Processing Systems*, 2019.

[8] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, 2017.

[9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, 2019.

[10] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018.

[11] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *INTERSPEECH*, 2019.