GRADIENT-BASED ACTIVE LEARNING QUERY STRATEGY FOR END-TO-END SPEECH RECOGNITION

Yang Yuan, Soo-Whan Chung and Hong-Goo Kang

Department of Electrical & Electronic Engineering, Yonsei University, Seoul, South Korea

ABSTRACT

In this paper, we propose an effective active learning query strategy for an automatic speech recognition system with the aim of reducing the training cost. Generally, training a deep neural network with supervised learning requires a massive amount of labeled data to obtain excellent performance. However, labeling data is tedious and costly manual work. Active learning can solve this problem by choosing and only annotating informative instances, which presents better results even with less transcribed data. In this approach it is vitally important to accurately select informative samples. Based on the preliminary experiment results that true gradient length has the best performance in terms of measuring sample informativeness in ideal conditions, we propose utilizing both uncertainty and the expected gradient length criterion to approximate the true gradient length using a neural network. The experiment results show that our proposed method is superior to the conventional individual criterion when applied to a phoneme-based speech recognition system, and it has both a faster convergence speed and the greatest loss reduction in both clean and noisy conditions.

Index Terms— Active learning, deep learning, combined query strategy, automatic speech recognition

1. INTRODUCTION

As a large amount of data can be collected through the Internet deep learning-based approaches have rapidly grown for many research and industrial fields such as image processing, natural language processing and speech signal processing. However, data cannot be directly used for training because most are unlabeled. Particularly for automatic speech recognition (ASR), manually transcribing hundreds hours of raw speech data individually is tedious and very costly.

Active learning that actively chooses informative and representative training data has been proposed as a solution to relieve the big training dataset issue. Generally, a deep learning network is favorable for training with unseen or less-trained data that causes larger gradients for backpropagation in the training stage. If labels exist for such data, this measurement is called the true gradient length (TGL) [1].

Although TGL is the ideal method for active learning, it cannot be used in practice because it still requires labeled data. Instead, there are two alternative strategies that utilize information uncertainty [1, 2] and the amount of model change [3, 4]. Uncertainty represents the degree to which the current model is not certain when attempting to recognize an instance. The methods that belong to this strategy are the least confidence method [5], the margin sampling method [6], and the entropy-based method [7]. There has also been interest in the query strategies that evaluate how much the model changes given new input data. For example, the expected gradient length (EGL) [8] was proposed to measure the variation in gradients in the backpropagation process, which does not require true labels. Beyond that, query-by-committee [9, 10] and density-weighted methods [11, 12] have been proposed. However, such conventional active learning remains insufficient to represent the information obtained by the TGL-based strategy.

In this paper, we propose a novel active learning method, in which we estimate the true gradient length with a deep learning framework. Based on the analysis that EGL and entropy-based methods provide different types of active learning knowledge, we utilize both features together in the framework. In other words, the proposed method estimates the true gradient length from two different active learning criteria. The performance of the proposed active learning query strategy is evaluated by implementing a phoneme recognition system based on connectionist temporal classification (CTC) [13] to compare its performance with conventional methods. In our clean and noisy speech experiments, the proposed estimated TGL strategy shows better performance than conventional single active learning methods.

This paper is organized as follows: Section 2 describes the active learning strategies that are relevant to the proposed method in detail, Section 3 introduces the proposed active learning method that utilizes a deep learning framework, Section 4 shows the experimental settings and evaluations to prove the performance, and Section 5 concludes the paper.

Part of this work was done when the first author was an intern at Naver corporation.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

2. ACTIVE LEARNING

In this section, we describe the training process for two active learning strategies: uncertainty-based and model change-based methods.

2.1. Uncertainty-based Active Learning

The uncertainty-based approach is the simplest query strategy that directly calculates the learning model's posterior probability. The most prevalent uncertainty method for our work is the following entropy-based active learning:

$$x^*_{en} = \arg\max_x \left[-\sum_{j=1}^{J} p(y_i = j|x_i, \theta)\log p(y_i = j|x_i, \theta) \right]. \quad (1)$$

The entropy-based method considers entropy between output y_i and all possible labels J with regard to input x_i and parameters θ. Well-trained data has small entropy when stable, but the others can be treated as unseen or less-trained data. Since other uncertainty-based methods such as the least confidence and margin sampling methods only consider the influence of one or two labels, they are inappropriate in multi-class problems such as speech recognition or image classification, where the label set is considerably large [1]. The entropy-based approach alleviates the deficiency by considering all possible class labels.

2.2. Model change-based Active Learning

The maximum model change claims to query the samples that will lead to the biggest change in the existing model if the true label is known, such as the true gradient length (TGL). A measure of the change can be inferred by the gradient length $|| \bigtriangledown_\theta L(\mathbf{x}_i, y_i; \theta)||$. However, the learning algorithm does not know the ground truth annotation of y in advance in practical applications, so it needs to calculate the expected value of the gradient over all possible labels then pick the instances that have the largest expected gradient length (EGL):

$$x^*_{egl} = \arg\max_x \sum_i P_\theta(y_i|x) || \bigtriangledown_\theta L(D^L \cup (x^*, y^*_i))||. \quad (2)$$

where $\bigtriangledown_\theta L(D^L)$ is the gradient of the loss function L with respect to the parameters θ and labeled data D^L. $\bigtriangledown_\theta L(D^L \cup (x^*, y^*))$ represents the new gradient derived by adding a new training instance (x^*, y^*). $|| \cdot ||$ is the Euclidean norm of each gradient vector. Note that the $|| \bigtriangledown_\theta L(D^L)||$ term is actually nearly zero, since L has converged at the previous training iteration. Thus, this term can be ignored and the $\bigtriangledown_\theta L(D^L \cup (x^*, y^*_i))$ is simplified to $\bigtriangledown_\theta L((x^*, y^*_i))$ [1].

EGL could be expensive if the computational complexity of the feature and label space are very large; for example, speech recognition usually contains many phoneme labels.

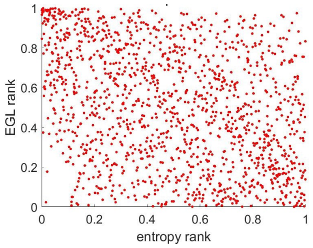

Fig. 1: EGL rankings vs. entropy rankings

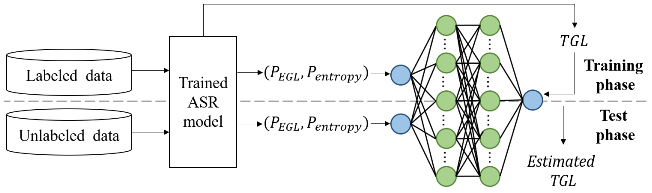

Fig. 2: Combining EGL and entropy to estimate TGL through the neural network

3. PROPOSED METHOD

3.1. Combined query strategy

Huang et al. [14] claimed that the two strategies described in the previous section tend to choose different data. Figure 1 depicts that EGL and entropy are not correlated. Since they operate at disparate scales, the plot axes represent the normalized rankings that correspond to their values. A plot close to the diagonal implies that the two methods evaluate informativeness in very similar ways. EGL and entropy criteria are uncorrelated, so we can conclude that EGL can identify the unique aspects of informativeness that uncertainty-based measurement cannot capture. Therefore, it is more powerful to use EGL and entropy in combination. To combine two different criteria, the strength of each strategy should be analyzed or needs to be processed with reinforcement learning or meta-learning [15, 16]. Since these approaches require very complicated training processes with huge computational complexity, its practical use is unfavorable. This work uses a neural network to jointly combine the characteristics of the two criteria to estimate TGL criterion which is better than the other criteria [17, 18]. Before estimation, as each criterion has a different dynamic range, all the criteria (*i.e.* EGL, entropy and TGL) are converted into a percentile scale and fed into the neural network to approximate TGL as demonstrated in Figure 2.

3.2. Application on CTC-based ASR

The proposed method is applicable for any type of deep learning-based system, but we apply it to the active learning for ASR trained using CTC-based phoneme recognition system. CTC [13] is favorable because it does not require

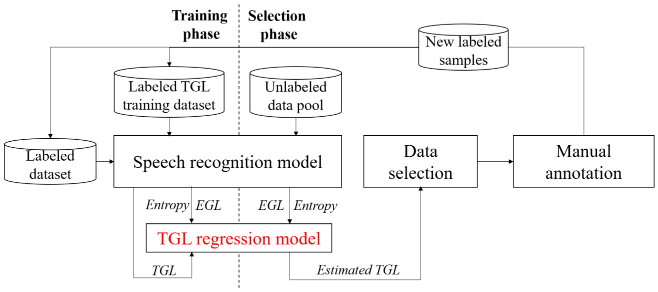

Fig. 3: The overall proposed active learning process

Table 1: Details of dataset configuration.

Dataset types	Clean utt.	Noise	Noisy utt.
Pretraining	1,200	three types	3,000
Active learning pool	2,000	four types	10,000
TGL training	200	four types	600
Validation	200	four types	1,000
Evaluation	192	four types	1,000

an accurate forced alignment process to determine phonetic labels. Since CTC uses a decoding process that computes possible graph and training loss, it is essential to modify the calculation of EGL and entropy. Note that CTC has an extra ”blank” label to distinguish temporal label changes. Since its probability is much higher than other labels, the ”blank” label should be removed before calculating the total entropy to avoid bringing any bias caused by the label. To reduce the computational cost of EGL, it uses the most probable top K labels obtained by the beam search decoding process. Besides, it is favorable to consider the decoded path along with their probabilities since the CTC-based model directly generates the phoneme sequence.

$$x_{egl}^{*'} = \arg\max_x \sum_i || \bigtriangledown_\theta \{P(y_i|x,\theta)L_{ctc}(y_i|x,\theta)\}|| \qquad (3)$$

Equation (3) decides that the path probability is first multiplied with the decoding results and the gradient of this weighted CTC loss function is then calculated.

The overall active learning process of the proposed method is demonstrated in Figure 3. The ASR model is pretrained by a labeled dataset. In this paper, we utilize the bidirectional long short-term memory recurrent neural network (BLSTM-RNN)-based structure. Next, the large dataset including only unlabeled raw speech is fed into the pretrained model. By utilizing active learning criteria such as uncertainty, EGL, and our proposed approach, the most valuable utterances are chosen and given to an oracle for annotation. After labeling, the instances are merged into the existing dataset to re-train the existing recognition model. This process is repeated for several iterations until we obtain a desirable ASR performance.

4. EXPERIMENTS

Here, the proposed method’s performance is compared to that of conventional methods for ASR.

4.1. Experimental Settings

The speech recognition experiments include two settings operated in clean and noisy environments. For the clean environment setting, experiments were performed with the TIMIT corpus with 1,200, 2,000, 200, 200, and 192 utterances for a small labeled dataset for pretraining, unlabeled data pool for active learning, distinct speech data for TGL estimation, validation set for ASR training, and evaluation set for ASR, respectively. To perform experiments in noisy environment setting, each dataset includes the same amount of utterances as was used in the clean speech experiment, and four types of noise from the CHiME3 dataset are used; bus, cafe, street, and pedestrian. Pretraining utterances were mixed with bus, cafe, and pedestrian noises at -5, 0, 5, and 10 dB signal-to-noise ratio (SNR) levels in a uniformly random manner. The other sets include all noise types with the same SNR levels. Noises recorded in different conditions were used for the evaluation set. Consequently, 3,000, 10,000, 600, 1,000, and 1,000 utterances were generated for each set as described in Table 1.

For DNN-based ASR, a 40-dimensional Mel-filterbank was extracted with a 25ms window and 10ms frame shift. Relying on [19], the folded 39 phonetic labels were used instead of the overall 64 phone labels. The phoneme-error-rate (PER) shows how efficiently the network learns in active learning strategies as an evaluation metric. The ASR neural network consists of three hidden layers with 256 cells for a clean simulation and four hidden layers with 512 cells for noisy environments. Model weights were initialized by Xavier initialization and trained by an Adam optimizer that was trained to the CTC training criterion. We reduced the computation by using only the top 50 probable paths determined by the CTC beam search decoding process to compute EGL.

Regarding the neural network for TGL estimation, since the training data was a small set with few dimensions, a shallow structure consisting of two hidden layers with 10 nodes with the ReLU activation function and sigmoid function for outputs was used. It was trained with a mean-squared-error (MSE) criterion to estimate the TGL.

4.2. Experiment Results

To shed light on how the TGL is superior to other query strategies and demonstrate that our proposed method has a similar performance to TGL, we compare TGL and estimated TGL with EGL, entropy, and random selection methods.

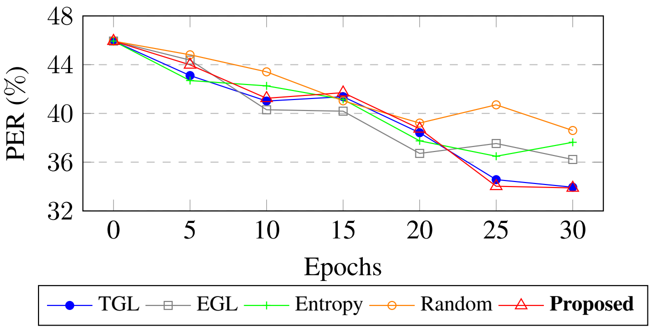

Fig. 4: PER results of clean speech ASR

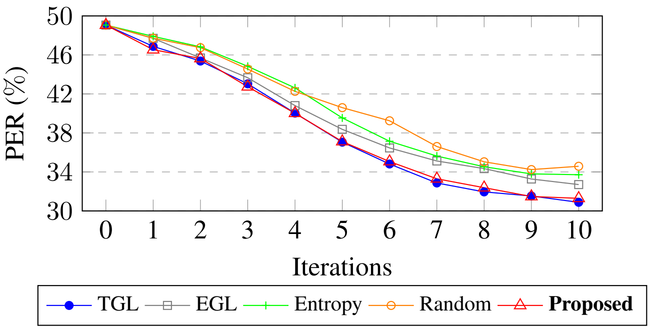

Fig. 5: PER results of noisy speech ASR

4.2.1. *Clean-speech experiments*

In this experiment, we compared five active learning strategies in the clean-speech recognition task. At every selection phase, we chose a fixed number of samples; 400 utterances from the unlabeled data pool.

Figure 4 represents each method's PER. The results show that all four active learning query strategies outperformed the random selection method (*baseline*), and TGL significantly reduced the error and had a much faster convergence speed compared to the other three approaches. We also verified that the training trend of the proposed method was similar to that of TGL, which derives the lowest PER and requires less training time. Consequently, we confirm that integrating EGL and entropy allows us to approximate TGL accurately and can be utilized as an active learning query strategy.

4.2.2. *Noisy-speech experiments*

We verified the generalization performance of the proposed active learning strategy by applying it to the noisy-speech recognition system and evaluating the performance for each noise type separately. The selection dataset for each noise contains 2,500 utterances and each test set includes 250 utterances. Due to the expansion of the noisy dataset, we chose 1,000 utterances at each iteration selection.

Figure 5 displays the PER curve of the whole evaluation set. Similar to the results in the clean-speech condition, active learning strategies outperformed the random selection method. TGL and the proposed estimated TGL method show much faster training times and larger PER reductions. Figure 6 depicts the selection ratio and PER reduction rate for each noise type when the amount of selected data was increased in the active learning process with the proposed method. We can observe that for the first several iterations, utterances from the street set were selected more often than the other noise types because the street noise was not included in the pretraining set. However, upon increasing the street-mixed samples, the model adapted to the street noise characteristics and the selected noise type amount gradually decreased. Meanwhile, the PER reduction which shows error variation compared to each previous step has the similar variation tendency as the number of selections; it is steeper initially and tends to become gentle.

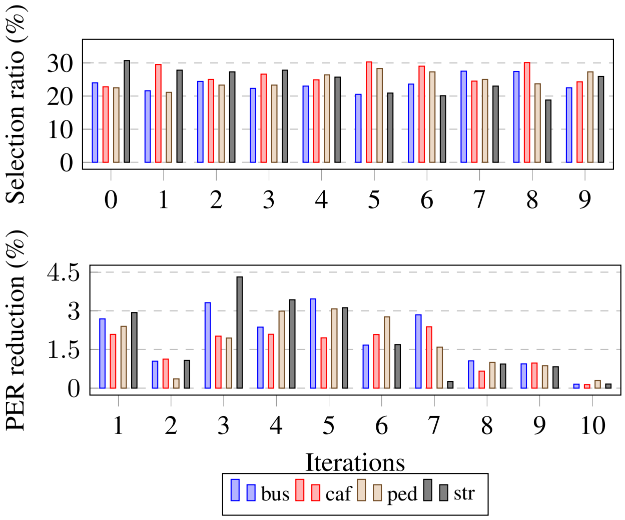

Fig. 6: Results at each retraining of the proposed method (Up: number of selected data of each noise type, Down: PER reduction compared to the last iteration)

5. CONCLUSION

In this paper, we proposed a novel active learning strategy to effectively select informative samples from a large amount of unlabeled data to reduce the manual annotation cost. Since TGL gives more accurate expression of informativeness than conventional active learning methods, our strategy estimates TGL by combining the EGL and entropy-based uncertainty approaches through a neural network. The experimental results via a phoneme recognition task confirmed the proposed estimated TGL method's efficacy, where it improved performance with less transcribed data and reduced training time.

Acknowledgements. We would like to give special thanks to Jisu Choi, Bong-Jin Lee, Min-Seok Choi and Youna Ji at Naver corporation for their helpful advice.

6. REFERENCES

[1] B. Settles, "Active learning literature survey," Tech. Rep. 1648, Univ. of Wisconsin, Madison, WI, USA, 2010.

[2] Giuseppe Riccardi and Dilek Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE transactions on speech and audio processing*, vol. 13, no. 4, pp. 504–511, 2005.

[3] Wenbin Cai, Yexun Zhang, Ya Zhang, Siyuan Zhou, Wenquan Wang, Zhuoxiang Chen, and Chris Ding, "Active learning for classification with maximum model change," *ACM Transactions on Information Systems (TOIS)*, vol. 36, no. 2, pp. 15, 2017.

[4] Alexander Freytag, Erik Rodner, and Joachim Denzler, "Selecting influential examples: Active learning with expected model output changes," in *European Conference on Computer Vision*. Springer, 2014, pp. 562–577.

[5] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM/Springer, 1994, pp. 3–12.

[6] Tobias Scheffer, Christian Decomain, and Stefan Wrobel, "Active hidden markov models for information extraction," in *International Symposium on Intelligent Data Analysis*. Springer, 2001, pp. 309–318.

[7] Claude E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, July 1948.

[8] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Advances in Neural Information Processing Systems (NIPS)*. 2008, vol. 20, pp. 1289–1296, MIT Press.

[9] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the 5th Annual ACM Conference Computational Learning Theory*, 1992, pp. 287–294.

[10] Yuzo Hamanaka, Koichi Shinoda, Sadaoki Furui, Tadashi Emori, and Takafumi Koshinaka, "Speech modeling based on committee-based active learning," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4350–4353.

[11] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[12] L. Yang, Y. Zhang, J. Chen, and Chen D. Z. Zhang, S., "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, 2017.

[13] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[14] J. Huang, R. Child, V. Rao, H. Liu, S. Satheesh, and A. Coates, "Active learning for speech recognition: the power of gradients," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[15] Meng Fang, Yuan Li, and Trevor Cohn, "Learning how to active learn: A deep reinforcement learning approach," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[16] Philip Bachman, Alessandro Sordoni, and Adam Trischler, "Learning algorithms for active learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[17] Peilin Zhao and Tong Zhang, "Stochastic optimization with importance sampling for regularized loss minimization," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[18] Angelos Katharopoulos and Franois Fleuret, "Not all samples are created equal: Deep learning with importance sampling," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[19] K-F Lee and H-W Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.