Active Learning for Speech Recognition: the Power of Gradients

Jiaji Huang*, **Rewon Child***, **Vinay Rao***
{huangjiaji, rewon, vinay}@baidu.com

Hairong Liu, Sanjeev Satheesh, Adam Coates
{liuhairong, sanjeevsatheesh, adamcoates}@baidu.com
Baidu Silicon Valley AI Lab
1195 Bordeaux Dr
Sunnyvale, CA, 94089

Abstract

In training speech recognition systems, labeling audio clips can be expensive, and not all data is equally valuable. Active learning aims to label only the most informative samples to reduce cost. For speech recognition, confidence scores and other likelihood-based active learning methods have been shown to be effective. Gradient-based active learning methods, however, are still not well-understood. This work investigates the Expected Gradient Length (*EGL*) approach in active learning for end-to-end speech recognition. We justify *EGL* from a variance reduction perspective, and observe that *EGL*'s measure of informativeness picks novel samples uncorrelated with confidence scores. Experimentally, we show that *EGL* can reduce word errors by 11%, or alternatively, reduce the number of samples to label by 50%, when compared to random sampling.

1   Introduction

State-of-the-art automatic speech recognition (ASR) systems [1] have large model capacities and require significant quantities of training data to generalize. Labeling thousands of hours of audio, however, is expensive and time-consuming. A natural question to ask is how to achieve better generalization with fewer training examples. Active learning studies this problem by identifying and labeling only the most informative data, potentially reducing sample complexity. How much active learning can help in large-scale, end-to-end ASR systems, however, is still an open question.

The speech recognition community has generally identified the informativeness of samples by calculating confidence scores. In particular, an utterance is considered informative if the most likely prediction has small probability [3], or if the predictions are distributed very uniformly over the labels [7]. Though confidence-based measures work well in practice, less attention has been focused on gradient-based methods like Expected Gradient Length (*EGL*) [4], where the informativeness is measured by the norm of the gradient incurred by the instance. *EGL* has previously been justified as intuitively measuring the expected change in a model's parameters [4].We formalize this intuition from the perspective of asymptotic variance reduction, and experimentally, we show *EGL* to be superior to confidence-based methods on speech recognition tasks. Additionally, we observe that the ranking of samples scored by *EGL* is not correlated with that of confidence scoring, suggesting *EGL* identifies aspects of an instance that confidence scores cannot capture.

---

*Equal contribution.

In [4], *EGL* was applied to active learning on sequence labeling tasks, but our work is the first we know of to apply *EGL* to speech recognition in particular. Gradient-based methods have also found applications outside active learning. For example, [9] suggests that in stochastic gradient descent, sampling training instances with probabilities proportional to their gradient lengths can speed up convergence. From the perspective of variance reduction, this importance sampling problem shares many similarities to problems found in active learning.

2 Problem Formulation

Denote $\mathbf{x}$ as an utterance and y the corresponding label (transcription). A speech recognition system models the conditional distribution $p(y|\mathbf{x}, \theta)$, where θ are the parameters in the model, and $p(y|\mathbf{x}, \theta)$ is typically implemented by a Recurrent Neural Network (RNN). A training set is a collection of $(\mathbf{x}, y)$ pairs, denoted as $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$. The parameters of the model are estimated by minimizing the negative log-likelihood on the training set:

$$\hat{\theta}_n = \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(\mathbf{x}_i, y_i, \theta) \triangleq -\log p(y_i|\mathbf{x}_i, \theta) \right]. \tag{1}$$

Active learning seeks to augment the training set with a new set of utterances and labels $\{(\mathbf{x}_i^*, y^*)\}_{i=1}^{m}$ in order to achieve good generalization on a held-out test dataset. In many applications, there is an unlabeled pool U which is costly to label in its entirety. U is *queried* for the "most informative" instance(s) $\mathbf{x}_i^*$, for which the label(s) y_i^* are then obtained. We discuss several such *query strategies* below.

2.1 Confidence Scores

Confidence scoring has been used extensively as a proxy for the informativeness of training samples. Specifically, an $\mathbf{x}_i^*$ is considered informative if the predictions are uniformly distributed over all the labels [7], or if the best prediction of its label is with low probability [3]. By taking the instances which "confuse" the model, these methods may effectively explore under-sampled regions of the input space.

2.2 Expected Gradient Length

Intuitively, an instance can be considered informative if it results in large changes in model parameters. A natural measure of the change is gradient length, $\|\nabla_\theta \ell(\mathbf{x}_i, y_i; \theta)\|$. Motivated by this intuition, Expected Gradient Length (*EGL*) [4] picks the instances expected to have the largest gradient length. Since labels are unknown on U, *EGL* computes the expectation of the gradient norm over all possible labelings. [4] interprets *EGL* as "expected model change". In the following section, we formalize the intuition for *EGL* and show that it follows naturally from reducing the variance of an estimator.

2.3 Variance in the Asymptote

Assume the joint distribution of $(\mathbf{x}, y)$ has the following form,

$$p(\mathbf{x}, y|\theta_0) = p(y|\mathbf{x}, \theta_0)p(\mathbf{x}),$$

where θ_0 is the true parameter, and $p(\mathbf{x})$ is independent of θ_0. By selecting a subset of the training data, we are essentially choosing another distribution $q(\mathbf{x})$ so that the $(\mathbf{x}, y)$ pairs are drawn from

$$q(\mathbf{x}, y|\theta_0) = p(y|\mathbf{x}, \theta_0)q(\mathbf{x}).$$

Statistical signal processing theory [6] states the following asymptotic distribution of $\hat{\theta}_n$,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to \mathcal{N}(0, \mathbf{I}_q^{-1}(\theta_0)]), \tag{2}$$

where $\mathbf{I}_q(\theta_0) \overset{\text{def}}{=} \mathbb{E}_{q(\mathbf{x},y)}\left[\nabla_\theta \log p(\mathbf{x}, y|\theta_0)\nabla_\theta^\top \log p(\mathbf{x}, y|\theta_0)\right]$ is the Fisher Information Matrix with respect to $q(\mathbf{x}, y)$. Using first order approximation at $\ell(\mathbf{x}, y; \theta_0)$, we have asymptotically,

$$\sqrt{n}(\ell(\mathbf{x}, y; \hat{\theta}_n) - \ell(\mathbf{x}, y; \theta_0)) \to \mathcal{N}(0, \nabla_\theta^\top \ell(\mathbf{x}, y; \theta_0)\mathbf{I}_q^{-1}(\theta_0)\nabla_\theta \ell(\mathbf{x}, y; \theta_0)). \tag{3}$$

Eq. (3) indicates that to reduce $\ell(\mathbf{x}, y; \hat{\theta}_n)$ on test data, we need to minimize the expected variance $\mathbb{E}_{p(\mathbf{x},y)}[\nabla_\theta^\top \ell(\mathbf{x}, y; \theta_0)\mathbf{I}_q^{-1}(\theta_0)\nabla_\theta \ell(\mathbf{x}, y; \theta_0)]$ over the test set. This is called Fisher Information Ratio criteria in [8], which itself is hard to optimize. An easier surrogate is to maximize $\mathrm{tr}(\mathbf{I}_q(\theta_0))$. Substituting Eq. (2.3) into $\mathbf{I}_q(\theta_0)$, we have

$$\mathbf{I}_q(\theta_0) = \mathbb{E}_{q(\mathbf{x},y)}\left[\nabla_\theta \log p(y|\mathbf{x}, \theta_0)\nabla_\theta^\top \log p(y|\mathbf{x}, \theta_0)\right] = \mathbb{E}_{q(\mathbf{x},y)}\left[\nabla_\theta \ell(\mathbf{x}, y; \theta_0)\nabla_\theta^\top \ell(\mathbf{x}, y; \theta_0)\right],$$

which is equivalent to $\max_q \int q(\mathbf{x}) \int p(y|\mathbf{x}, \theta_0)\|\nabla_\theta \ell(\mathbf{x}, y; \theta_0)\|^2 dy d\mathbf{x}$.

A practical issue is that we do not know θ_0 in advance. We could instead substitute an estimate $\hat{\theta}_0$ from a pre-trained model, where it is reasonable to assume the $\hat{\theta}_0$ to be close to the true θ_0. The batch selection then works by taking the samples that have largest gradient norms,

$$i^* = \arg\max_i \sum_y p(y|\mathbf{x}_i, \hat{\theta}_0)\|\nabla_\theta \ell(\mathbf{x}_i, y; \hat{\theta}_0)\|^2. \tag{4}$$

For RNNs, the gradients for each potential label can be obtained by back-propagation. Another practical issue is that *EGL* marginalizes over all possible labelings, but in speech recognition, the number of labelings scales exponentially in the number of timesteps. Therefore, we only marginalize over the K most probable labelings. They are obtained by beam search decoding, as in [5]. The *EGL* method in [4] is almost the same as Eq. (4), except the gradient's norm is not squared in [4].

Here we have provided a more formal characterization of *EGL* to complement its intuitive interpretation as "expected model change" in [4]. For notational convenience, we denote Eq. (4) as *EGL* in subsequent sections.

3 Experiments

We empirically validate *EGL* on speech recognition tasks. In our experiments, the RNN takes in spectrograms of utterances, passing them through two 2D-convolutional layers, followed by seven bi-directional recurrent layers and a fully-connected layer with softmax activation. All recurrent layers are batch normalized. At each timestep, the softmax activations give a probability distribution over the characters. CTC loss [2] is then computed from the timestep-wise probabilities.

A base model, $\hat{\theta}_0$, is trained on 190 hours ($\sim$100K instances) of transcribed speech data. Then, it selects a subset of a 1,700-hour ($\sim$1.1M instances) unlabeled dataset. We query labels for the selected subset and incorporate them into training. Learning rates are tuned on a small validation set of 2048 instances. The trained model is then tested on a 156-hour ($\sim$100K instances) test set and we report CTC loss, Character Error Rate (CER) and Word Error Rate (WER).

The confidence score methods [3, 7] can be easily extended to our setup. Specifically, from the probabilities over the characters, we can compute an entropy per timestep and then average them. This method is denoted as *entropy*. We could also take the most likely prediction and calculate its CTC loss, normalized by number of timesteps. This method is denoted as *pCTC* (predicted CTC) in the following sections.

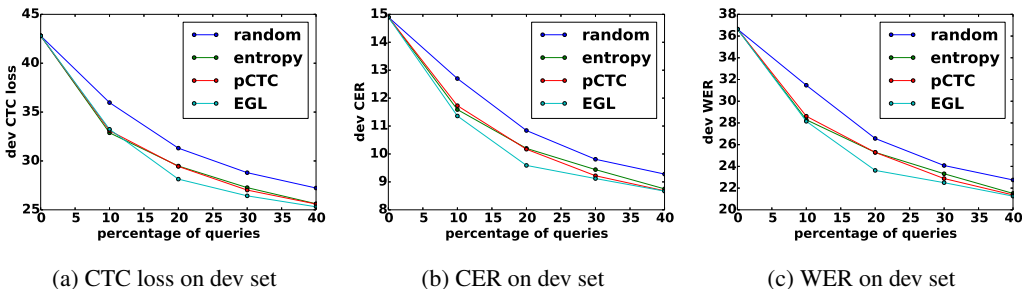| (a) CTC loss on dev set | (b) CER on dev set | (c) WER on dev set |

Figure 1: Performance metrics at various percentages of queries. *EGL* shows a greater reduction in error for smaller amounts of data. By definition, all strategies converge as the query percentage approaches 100%.

3

We implement *EGL* by marginalizing over the most likely 100 labels, and compare it with: 1) a *random* selection baseline, 2) *entropy*, and 3) *pCTC*. Using the same base model, each method queries a variable percentage of the unlabeled dataset. The queries are then included into training set, and the model continues training until convergence. Fig. 1 reports the metrics (Exact values are reported in Table 1 in the Appendix) on the test set as the query percentage varies. All the active learning methods outperform the *random* baseline. Moreover, *EGL* shows a steeper, more rapid reduction in error than all other approaches. Specifically, when querying 20% of the unlabeled dataset, *EGL* has 11.58% lower CER and 11.09% lower WER relative to *random*. The performance of *EGL* at querying 20% is on par with *random* at 40%, suggesting that using *EGL* can lead to an approximate 50% decrease in data labeling.

3.1 Similarity between Query Methods

It is useful to understand how the three active learning methods differ in measuring the informativeness of an instance. To compare any two methods, we take rankings of informativeness given by these two methods, and plot them in a 2-D ranking-vs-ranking coordinate system. A plot close to the diagonal implies that these two methods evaluate informativeness in a very similar way.

Fig. 2 shows the ranking-vs-ranking plots between *pCTC* and *entropy*, *EGL* and *entropy*. We observe that *pCTC* rankings and *entropy* rankings (Fig. 2a) are very correlated. This is likely because they are both related to model uncertainty. In contrast, *EGL* gives very different rankings from *entropy* (Fig.2b). This suggests *EGL* is able to identify aspects of an instance that uncertainty-based measurements cannot capture.

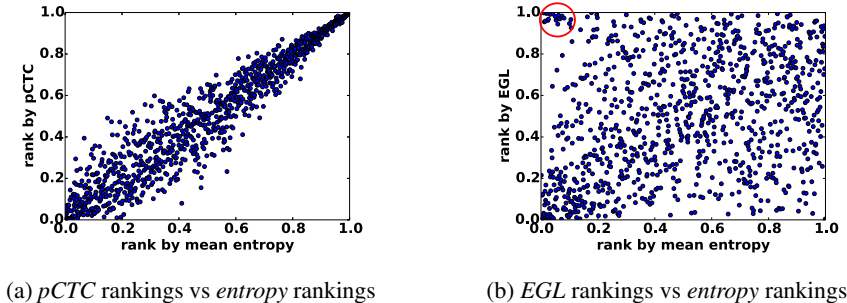(a) *pCTC* rankings vs *entropy* rankings   (b) *EGL* rankings vs *entropy* rankings

Figure 2: The difference in how active learning methods rank informativeness of samples. Rankings are normalized to $[0, 1]$, with 1 being the most informative. In (a), *pCTC* and *entropy* are shown to be very correlated. In (b), *EGL* appears uncorrelated with *entropy* (and *pCTC*). Data samples highlighted in the red circle are considered very informative by *EGL*, but uninformative by *entropy*.

We further investigate the samples for which *EGL* and *entropy* yield vastly different estimates of informativeness, e.g., the elements in the red circle in Fig. 2b. These particular samples consist of short utterances containing silence (with background noise) or filler words. Further investigation is required to understand whether these samples are noisy outliers or whether they are in fact important for training end-to-end speech recognition systems.

4 Conclusion and Future Work

We formally explained *EGL* from a variance reduction perspective and experimentally tested its performance on end-to-end speech recognition systems. Initial experiments show a notable gain over random selection, and that it outperforms confidence score methods used in the ASR community. We also show *EGL* measures sample informativeness in a very different way from confidence scores, giving rise to open research questions. All the experiments reported here query all samples in a single batch. It is also worth considering the effects of querying samples in a sequential manner. In the future, we will further validate the approach with sequential queries and seek to make the informativeness measure robust to outliers.

4

Appendix

Table 1: Performance metrics at various query percentages (smaller is better, best in bold)

query	CTC				CER				WER			
	random	*entropy*	*pCTC*	*EGL*	*random*	*entropy*	*pCTC*	*EGL*	*random*	*entropy*	*pCTC*	*EGL*
10%	35.97	**32.88**	33.10	33.24	12.70	11.59	11.73	**11.36**	31.47	28.29	28.62	**28.15**
20%	31.31	29.48	29.44	**28.14**	10.84	10.20	10.17	**9.59**	26.57	25.29	25.29	**23.63**
30%	28.80	27.27	27.02	**26.43**	9.81	9.44	9.22	**9.12**	24.08	23.32	22.88	**22.50**
40%	27.23	25.62	25.59	**25.31**	9.28	8.75	8.68	**8.67**	22.75	21.51	21.37	**21.26**

References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, et al. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *33rd International Conference on Machine Learning*, 2016.

[2] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 2006.

[3] G. Riccardi and D. Hakkani-Tur. Active learning: Theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(4):504–511, 2005.

[4] B. Settles. Active learning literature survey. Technical report, University of Wisconsin, Madison, 2009.

[5] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics*, 2008.

[6] J. Sourati, M. Akcakaya, T. K. Leen, D. Erdogmus, and J. G. Dy. Asymptotic analysis of objectives based on fisher information in active learning. *arXiv:1605.08798*, 2016.

[7] B. Varadarajan, D. Yu, L. Deng, and A. Acero. Maximizing global entropy reduction for active learning in speech recognition. In *International Conference on Acoustics, Speech and Signal Processing*, 2009.

[8] T. Zhang and F. Oles. The value of unlabeled data for classification problems. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

[9] P. Zhao and T. Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.