Paper Number: 3877

Confidence Estimation for Attention-Based Sequence-to-Sequence Models for Speech Recognition

Qiujia Li[1], David Qiu[2], Yu Zhang[2], Bo Li[2], Yanzhang He[2], Philip C. Woodland[1], Liangliang Cao[2], Trevor Strohman[2]

[1]University of Cambridge, UK, [2]Google LLC, USA

Abstract

- Studied impact of regularisation methods on confidence scores for attention-based sequence-to-sequence models
- Proposed *confidence estimation module* (CEM)
- Showed the effectiveness of CEM for token and word confidence
- Showed CEM generalises well and may help downstream tasks

Introduction

- Confidence scores are very useful for many ASR-related applications, *e.g.* semi-supervised / active learning
- For conventional HMM-based systems
  - confidence based on word posteriors from lattices are reliable
  - model-based approaches can further improve confidence
- For end-to-end trainable systems
  - only a small part of the hypothesis space is observed
  - model-based auto-regressive decoder can easily overfit
  - confidence scores based on softmax probabilities are poor
- To compare conventional and end-to-end systems
  - two systems have similar WERs
  - filter utterances with confidence $>$ threshold (x-axis)
  - plot the WER of the filtered subset (y-axis)

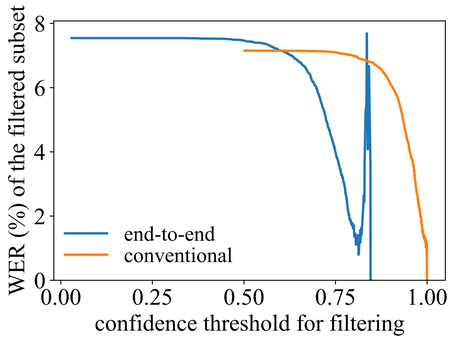


Figure 1: Filtering behaviour of a conventional HMM-based system and an attention-based sequence-to-sequence system.

- Observations
  - poor calibration of confidence scores for end-to-end system
  - the spike indicates that end-to-end model is overconfident

Attention-Based Sequence-to-Sequence Models

- Model components

$$\begin{aligned}
\mathbf{e}_{1:L} &= \textsc{Encoder}(\mathbf{x}_{1:L}) \\
\mathbf{a}_t &= \textsc{Attention}(\mathbf{a}_{t-1}, \mathbf{d}_{t-1}, \mathbf{e}_{1:L}) \\
\mathbf{d}_t &= \textsc{Decoder}(\mathbf{a}_t, \mathbf{d}_{t-1}, \textsc{Emb}(y_{t-1})) \\
p(y_t|y_{1:t-1}, \mathbf{x}_{1:L}) &= \textsc{Softmax}(\mathbf{d}_t)
\end{aligned}$$

Confidence Estimation Module (CEM)

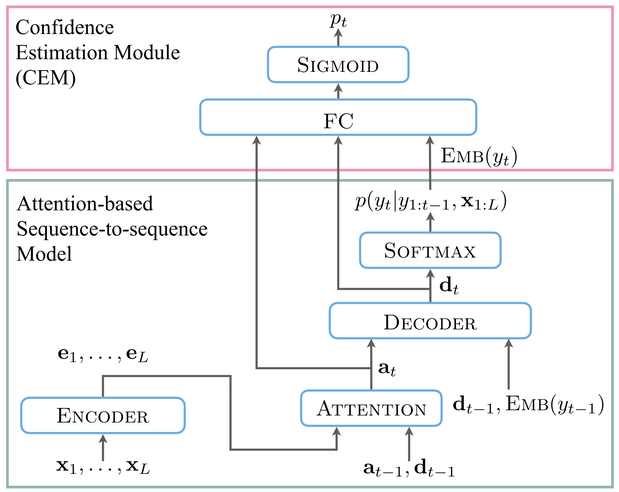


Figure 2: CEM for an attention-based sequence-to-sequence model.

1. Generate n-best hypotheses with ASR model frozen
2. Align hypotheses with ground-truth using edit distance
   - assign 1 for correct token and 0 for insertion / substitution
3. Estimate confidence score for each token p_t using CEM

$$p_t = \textsc{Sigmoid}\Big(\textsc{FC}\big(\mathbf{a}_t, \mathbf{d}_t, \textsc{Emb}(y_t), p(y_t|y_{1:t-1}, \mathbf{x}_{1:L})\big)\Big)$$

4. Compute loss $\mathcal{L}$ for all tokens and backpropagate

$$\mathcal{L}(\mathbf{c}, \mathbf{p}) = -\frac{1}{T}\sum_{t=1}^{T}\Big(c_t \log(p_t) + (1 - c_t)\log(1 - p_t)\Big)$$

Experimental Setup

- Data: LibriSpeech 100h for training; standard dev/test data
- ASR model: 4-layer Bi-LSTM encoder & 2-layer LSTM decoder

Experimental Results

Effect of Regularisation on Softmax-Based Confidence

	WER ↓	AUC ↑	NCE ↑
baseline	7.5	0.976	-0.195
– dropout	7.8	0.977	-0.204
– EMA	8.2	0.974	-0.189
– label smoothing	10.6	0.985	0.106
– weight noise	12.9	0.978	-0.459
– SpecAugment	10.8	0.952	0.012

Table 1: ASR and token-level confidence performance by removing a regularisation method from the baseline model on LibriSpeech test-clean.

- Regularisation methods can reduce WER
- But they do not necessarily improve confidence scores

Experimental Results (cont.)

CEM Performance

		WER ↓	AUC ↑	NCE ↑
baseline	softmax	7.5/21.6	0.981/0.927	0.269/0.195
	CEM		**0.990/0.962**	**0.350/0.270**
+ LM	softmax	6.8/19.8	0.981/0.928	0.103/0.109
	CEM		**0.991/0.966**	**0.337/0.263**

Table 2: ASR and word-level confidence performance for models with and without RNNLM shallow fusion on LibriSpeech test-clean/test-other.

- CEM can improve confidence estimation performance
- CEM also works well after shallow fusion with a language model

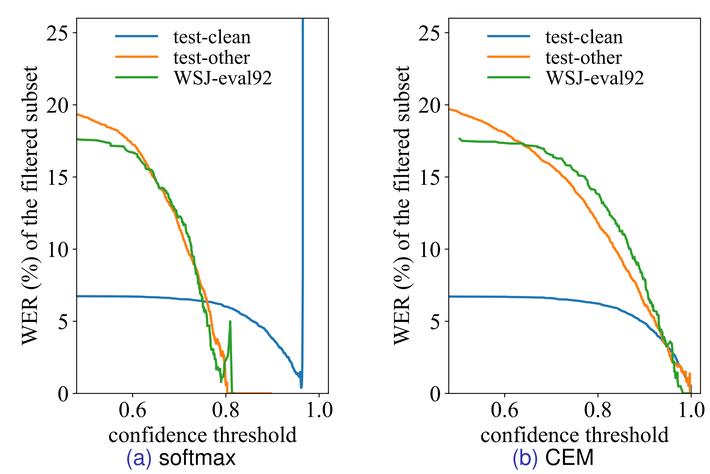
Generalisation to a Mismatched Domain

		WER ↓	AUC ↑	NCE ↑
baseline	softmax	18.7	0.935	0.230
	CEM		**0.970**	**0.280**
+ LM	softmax	17.7	0.933	0.159
	CEM		**0.965**	**0.266**

Table 3: ASR and word-level confidence performance on WSJ eval92.

- CEM also improves the quality of confidence scores for WSJ

Implications for Downstream Tasks



(a) softmax (b) CEM

Figure 3: WERs of filtered utterances w.r.t. confidence thresholds.

- Confidence scores based on CEM reflect WER better
- May benefit data selection for semi-supervised learning

Conclusions

- Using softmax probabilities for confidence is not reliable
- The proposed CEM is very effective as a confidence estimator for attention-based models, and can be extended to RNN-T models.
- CEM can predict deletion errors if trained with deletion targets

E-Mail: ql264@cam.ac.uk