ANALYZING UNCERTAINTIES IN SPEECH RECOGNITION USING DROPOUT

Apoorv Vyas , Pranay Dighe, Sibo Tong, Hervé Bourlard

Idiap Research Institute, Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

The performance of Automatic Speech Recognition (ASR) systems is often measured using Word Error Rates (WER) which requires time-consuming and expensive manually transcribed data. In this paper, we use state-of-the-art ASR systems based on Deep Neural Networks (DNN) and propose a novel framework which uses "Dropout" at the test time to model uncertainty in prediction hypotheses. We systematically exploit this uncertainty to estimate WER without the need for explicit transcriptions. In addition, we show that the predictive uncertainty can also be used to accurately localize the errors made by the ASR system. We study the performance of our approach on Switchboard database where it predicts WER accurately within a range of 2.6% and 5.0% for HMM-DNN and Connectionist Temporal Classification (CTC) ASR systems, respectively.

Index Terms— dropout uncertainty, WER estimation, word confidence, error localization

1. INTRODUCTION

Dropout-based training [1] of Deep Neural Network (DNN) acoustic models is a standard regularization technique often used to improve generalization properties (hence robustness) of state-of-the-art ASR systems. While dropout is typically used during training to prevent overfitting of DNNs, it was recently shown in [2] that dropout during inference can also provide a way to compute the model's uncertainty on its predictions. Computing the *prediction uncertainty* of a DNN model using Monte Carlo sampling with dropout has been successfully used not only to characterize model errors but also to improve the system performance in various applications [2, 3, 4, 5]. The present work is a novel attempt to study the usage and utility of dropout uncertainty in the context of automatic speech recognition (ASR) systems and explore its implications, including its potential use in semi-supervised training and/or adaptation of DNN-based ASR systems.

ASR systems have made rapid progress in recent years, leading to various applications across multiple domains. Thus, it is of utmost importance to quickly and reliably estimate the accuracy as well as the errors made by an ASR system. Typically, the Word Error Rate (WER) metric is used as a straightforward way to evaluate and compare the performance of ASR systems. Word errors are typically caused by the inputs which are noisy or exhibit a mismatch with the training conditions. Such inputs also lead to a higher predictive uncertainty of the ASR model which is an indication of potential speech recognition errors. Therefore, this work proposes to exploit dropout-based uncertainty to predict the errors made by ASR. Specifically, we focus on analyzing how confident or uncertain the acoustic modeling component is in making a prediction about the input acoustic evidence.

Given an already trained DNN-based acoustic model, we can employ the dropout mechanism during inference for computing the frame-level state posteriors (or data likelihoods) for the test speech utterances. These posteriors can then be used to generate a decoding lattice which can output a hypothesis word sequence for the test utterance. If this process is repeated for the same test utterance multiple times, the acoustic model will predict different posterior probability outputs for the same acoustic input due to a different random selection of the active neurons. As shown in [2], this process leads to a Bayesian inference over the acoustic model weights. The *acoustic model uncertainty* about a test utterance is therefore reflected in the variations observed in the predicted hypotheses for each Monte Carlo sample. As discussed in Sections 2 and 3, the variations in different decoded hypotheses for any utterance are often highly localized at certain word positions and depict locations where the ASR decoding might be inaccurate. We capitalize on these localized uncertainties in the predicted ASR hypotheses to estimate the WER of our speech recognition systems without the need of comparing them against the oracle transcriptions.

In prior research, a prominent method for quantifying ASR uncertainty has been in terms of computing lattice-based confidence measures. The posterior probability of a recognized word can be estimated from a word lattice [6, 7, 8] or a word confusion network [9, 8] without any additional training. However, these confidence measures do not specifically exploit the uncertainty of the acoustic model as the predictions made by acoustic model are static (point estimates). Many classifier-based approaches have also been proposed to detect errors in ASR output, for example, conditional random fields (CRFs) [10, 11, 12], feedforward neural networks [13, 14, 15, 16] and recurrent neural networks

(RNNs) [17]. The classifier-based approaches show good performance in detecting error as well as in WER estimation. However, training the classifier and preparing the training data are time-consuming. In comparison, the proposed dropout uncertainty approach does not require any additional model training step.

The contribution of this work is twofold. First, we systematically show that the localized uncertainties in the hypotheses generated by dropout sampled acoustic models are highly correlated with the actual word errors. Second, we use this observation to predict the WER, as well as to localize the errors made by various ASR systems. Specifically, we experiment with HMM-DNN and Connectionist Temporal Classification (CTC) acoustic models and provide experimental analysis on Switchboard database.

The remainder of this paper is organized as follows: in Section 2, we briefly discuss the proposed method for WER estimation. In Section 3, experimental setup is described. Results and analysis are provided in Section 4. Finally, Section 5 concludes the paper.

2. OUR METHOD

In this section, we explain our dropout uncertainty based approach to (1) estimate WERs without using oracle transcriptions and (2) localize errors in the decoded ASR hypotheses. Unlike previous approaches, the proposed method does not require a lattice N-best list or a dedicated DNN to predict word-level confidences. We only exploit the uncertainty in the output of the acoustic models through the Monte Carlo sampling of the neural networks using dropout at the test time.

For each utterance, we forward-pass it N times through a dropout enabled neural network acoustic model. Each of the N acoustic model outputs is then processed though the decoding pipeline to generate N dropout-hypotheses. Separately, we also obtain a hypothesis by keeping the dropout off during test time, as is done traditionally. The resulting $N+1$ hypotheses are then used to get an estimate of both the WER and the word-level confidences for the given utterance.

Figure 1 below shows an example where an utterance is decoded $N = 4$ times with dropout turned on. Here D_{off} refers to the decoded output with dropout turned off and D^1_{on}-D^4_{on} refers to the decoded output with dropout on. GT refers to the ground truth transcript and $\mathcal{C}_w$ is the word level estimated confidence. The blue boxes refer to the errors between the D_{off} case and GT which are used in the traditional WER computation. The red and green boxes show the word positions where D^1_{on}-D^4_{on} hypotheses disagree with each other. The differences in decoded hypotheses is due to the acoustic model's uncertainty in predicting the posterior probabilities for acoustic features in certain regions of the utterance. For the utterance shown below, we observe that two uncertain word positions actually overlap with the mismatches between D_{off} and GT (true positive detections), one uncertainty is a false positive and one mismatch is not detected.

Fig.1 Predicting ASR errors using dropout uncertainty.

D_{off} : i i agree with the a hundred percent there or
GT : i say agree with you a hundred percent there --
****
D^1_{on} : i i agree with you a hundred percent there --
D^2_{on} : yes i agree with the a hundred percent there --
D^3_{on} : i i agree with you a hundred percent there or
D^4_{on} : yes i agree with the a hundred percent there --
$\mathcal{C}_w$: 0.5 1 1 1 0.5 1 1 1 1 0.25
True Positive False Positive Missed Detection

2.1. Word Error Rate Estimation

For WER estimation of an utterance i, $\binom{N}{2}$ pairwise edit distances between all the N hypotheses are calculated and sorted in descending order. The mean edit distance, ${}^iE_\mu$, is obtained as the mean of the top-K edit distances, where K is the hyper-parameter tuned on the development data. Similarly, we then obtain the mean length, ${}^iL_\mu$, of the decoding corresponding to the top-K edit distances. The WER for the utterance is then estimated as the ratio of ${}^iE_\mu$ and ${}^iL_\mu$. Let D denote the whole dataset with $|D|$ utterances, then the WER of this dataset is calculated as the ratio of total mean edit distances and total mean lengths summed over all the utterances.

Let ${}^i_kE$ and ${}^i_kL$ denote the top-K edit distances and corresponding lengths for the utterance i. The WER for the utterance i is given by:

$$ {}^iE_\mu = \frac{\sum_{j=1}^{K} {}^i_kE_j}{K} \tag{1} $$

$$ {}^iL_\mu = \frac{\sum_{j=1}^{K} {}^i_kL_j}{K} \tag{2} $$

$$ {}^iWER = \frac{{}^iE_\mu}{{}^iL_\mu} \tag{3} $$

$$ WER_{data} = \frac{\sum_{i=1}^{|D|} {}^iE_\mu}{\sum_{i=1}^{|D|} {}^iL_\mu} \tag{4} $$

2.2. Error Localization Using Word-Level Confidence

In addition to estimating WER, we can also exploit dropout uncertainty to localize the errors in ASR. To achieve this, we propose a method that relies on computing the word-level confidences of the ASR hypotheses i.e. the D_{off} case. Word confidences represent the word-by-word reliability of the D_{off} decoding. If the confidence $\mathcal{C}_w$ of a word w in the decoding D_{off} is less than a pre-defined threshold τ, our system predicts a *potential location of error*. If the predicted error locations are the same as the mismatch positions between D_{off} and ground truth transcription GT, we can claim that the error localization is accurate. Our error localization method then works as follows.

To estimate the word level confidences, we first align the N dropout turned-on hypotheses $\{D^i_{\text{on}}\}_{i=1}^N$ against the hypotheses with the dropout turned-off D_{off}. Then, for each word, we use the mean agreement between all the hypotheses to estimate its confidence. Formally, for an utterance, the confidence for its w^{th} word in ASR decoding D_{off} is given by

$$C_w = \frac{\sum_{k=1}^{N} I(D^k_{\text{on}}[w] = D_{\text{off}}[w])}{N} \quad (5)$$

where $D^k_{\text{on}}[w]$ and $D_{\text{off}}[w]$ denote w^{th} words in decoding D^k_{on} and D_{off} respectively and function $I()$ is the indicator function. The confidences C_w lie in the range $[0, 1]$ and we predict an error location wherever C_w is lower than a threshold τ.

To evaluate the efficacy of our error localization method, we use an "Intersection over Union" (IoU) metric where "Intersection" refers to the intersection between the true errors and the predicted errors and "Union" refers to the set union of the true errors and predicted errors. Note that true errors refer to mismatches between D_{off} and GT. IoU lies in the range $[0, 1]$ and is highest when the predicted errors exactly match the true errors. It penalizes for both false negatives (when a true error is not detected) and false positives (when a correct word is predicted as an error). In the example presented in Figure 1, if $\tau = 0.6$, there are 3 error locations, their intersection with true error locations is 2 and the union is 4. Hence, the IoU will be 0.5. Similarly, if $\tau = 0.4$, the IoU will be 0.33 (intersection=1, union=3). In this work, we use the mean of the IoU over all the sentences as the indicator of localization performance.

3. EXPERIMENTAL ANALYSIS

3.1. Dataset

We evaluate our dropout uncertainty-based WER estimation approach on Switchboard database [18]. A 110h subset *train_100k* of the actual 300h training data is used for training the acoustic models. We tune the hyperparameters K and τ for better WER estimation on a 11.5h subset *dev* taken from the rest of the training data. Finally, we evaluate our approach on a *test* subset which has 8.4h of speech. We ensure that there are no common speakers in our data partitions. As used in [19], the *train_100k* subset contains the first 100k utterances of the actual training data and is used for a faster turnaround time. Nevertheless, preliminary results have also shown the validity of our approach on the complete Switchboard dataset.

3.2. ASR Models

Two different acoustic models are used in the evaluation, namely DNN-HMM and CTC-based model. We use the Kaldi [20] *nnet1* recipe for training a feedforward DNN-HMM ASR system. Kaldi *tri4* setup based on LDA+MLLT+SAT system is used for generating senone alignments (8564 units) and fMLLR transformed MFCC features (1320 dimensional, after appending delta features and context of 11). DNN acoustic model has 6 hidden layers having 2048 neurons each. We set a dropout rate of 0.2 for all hidden layers during training and the same is used during testing.

The phoneme-based CTC model is trained in Pytorch using Baidu's CTC implementation for Deepspeech 2 [21]. It has 4 layers of Bidirectional Long Short-Term Memory (BLSTM), with 320 cells in each layer and direction. we use 40-dimensional log-mel filterbank coefficients as acoustic features together with their first and second-order derivatives. Dropout was applied for all BLSTM layers. The dropout rate was set to 0.2. A trigram LM is used for decoding for both DNN-HMM and CTC models and no further LM-based rescoring of lattices is done.

To compute the oracle WERs, we use the basic scoring script `compute-wer` provided with Kaldi instead of NIST sclite tool which involves text normalization. Although the oracle WERs computed this way are usually high, they pose as a more suitable ground truth to compare with the estimated WERs using our approach.

3.3. Baseline Systems and Hyper-parameters

As a baseline for WER estimation, we replace the N dropout-on hypotheses by the N-best hypotheses of the decoding lattice and use (1)-(4) to estimate WER.

Similarly, for word error localization baseline, we use the N-best list hypotheses in equation 5 to estimate word confidences. Another baseline based on [22] evaluates the word-level confidences from word posterior probabilities computed using forward-backward likelihood computation on a lattice.

ASR Model	Dropout			N-best list			MBR
	N	K	τ	N	K	τ	τ
DNN-HMM	100	5	1.0	60	542	0.8	0.9
CTC	24	119	0.9	60	530	0.8	0.9

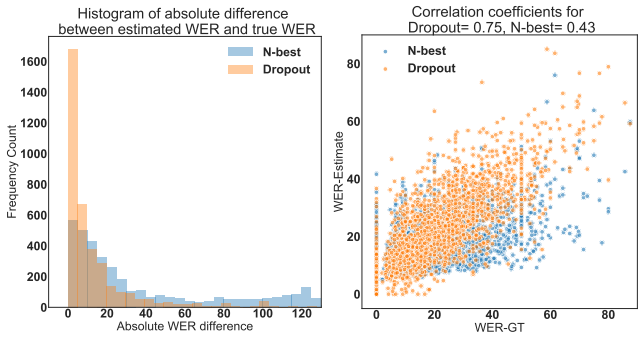
Table 1. Hyperparameter values tuned on *dev* set.

3.4. Word Error Rate Estimation

ASR System	**S.L [1-3]**	**S.L [4-6]**	**S.L [7-10]**	**S.L [11-Max]**	**S.L [1-Max(All)]**
DNN-*test*-Gt	35.5	28.8	25.1	22.3	23.3
DNN-*test*-Dr	33.2	31.0	25.2	21.5	22.7
DNN-*test*-Nb	73.6	42.1	30.2	13.7	19.7
CTC-*test*-Gt	30.6	31.2	25.3	23.3	24.0
CTC-*test*-Dr	34.4	35.5	29.7	23.9	25.2
CTC-*test*-Nb	93.1	49.0	34.3	15.9	22.7

Table 2. Results on estimating WER using dropout uncertainty. *Dr* refers to dropout based estimation, *Nb* refers to N-best list based estimation, and *Gt* refers to the Ground truth WER. *S.L.* refers to ground truth reference sentence length.

Table 2 compares the dropout and N-best list based WER estimation on Switchboard database. Results shown are split according to the length of the ground truth transcription for better analysis. We observe that dropout WER estimate is better across all ground truth sentence lengths and on both DNN-HMM and CTC ASR systems. Although the WER estimate using the N-best list over all the sentences lengths (last column) is reasonable, it severely overestimates on shorter sentence and underestimates the WER on longer sentences. This is because each hypotheses in the N-best list is different and thus, it cannot give a WER of 0. For small length sentences, this results in overestimation of the WER. For example, if we consider sentences with only one word in the ground truth, the n-best list would still contain all different hypotheses resulting in a WER $>= 100$. In contrast to this, dropout outputs change only at word locations where the acoustic model is uncertain. Therefore, if the acoustic model's uncertainty is very low along the whole utterance being decoded, then it is possible that each of N hypotheses is identical. As a result, it does not suffer from the problem of overestimating WER.



(a) Histogram of absolute errors (b) Correlation Plot

Fig. 2. Comparison between dropout and N-best list based WER estimation for CTC system. (a) Histogram of absolute difference between estimated WER and true WER. (b) Correlation between estimated WER and true WER.

Figure 2(a) shows a histogram of the utterance-wise absolute difference between true WER and estimated WER based on dropout and N-best list. For a perfect estimator, the absolute difference for every sentence will be 0. We notice that for the dropout-based estimation, the number of utterances with a very small absolute difference ($\sim$0) is much higher than those for the N-best list based estimation. For a randomly picked utterance, the dropout-based WER estimate is much more likely to be close to the true WER than the N-best list estimation. Figure 2(b) shows the correlation between the ground truth-based true WER and estimated WER. Each point on the scatter plot is an utterance. We observe a high correlation for dropout-based estimation (0.75) as compared to the N-best list based estimation (0.43). This is also evident from the dropout scatter plot being more dense in the diagonal region of the plot. As expected from the discussion above, the estimated WER for N-best list is always greater than 0.

3.5. Error Localization Using Word Confidences

Table 3 shows the performance of word confidence-based ASR error localization in terms of the IoU metric. We compare the proposed dropout approach against N-best list and lattice-based word posterior probability based approaches. We observe that the IoU metric is higher using our dropout method as compared to both of the lattice-based approaches. For DNN-HMM as well as CTC-based ASR systems, dropout achieves an IoU of $\sim$0.6 which depicts that a significant number of error locations are accurately identified. Note that the results in Table 3 are averaged over all the utterance lengths. In our experiments, we noticed that IoU metric can be as high as $\sim$0.7-0.8 for very short utterances ($\leq$ 6 words long).

Approach	Dropout	N-best	Lat.Conf.
DNN-*dev*	0.55	0.45	0.54
DNN-*test*	0.59	0.50	0.58
CTC-*dev*	0.56	0.41	0.51
CTC-*test*	0.61	0.41	0.58

Table 3. Error Localization comparison using IoU metric

4. CONCLUSIONS AND FUTURE WORK

We have proposed a novel way to exploit dropout uncertainty in context of measuring the performance of DNN-based ASR systems. We show that the variations in different decoded hypotheses with dropout are often highly localized at certain word positions and depict locations where the ASR decoding might be inaccurate. Experiments on the Switchboard dataset with 2 different acoustic models show that our approach accurately estimates word error rates and word level confidences and is more robust to the length of the sentences, compared to lattice-based approaches. In future, we intend to use word level predictive uncertainty in the output for model combination and for semi-supervised and active learning where the training data from untranscribed data can be picked depending on the model confidence. While this paper primarily focused on the DNN-based acoustic model, it is clear that similar dropout strategies (at test time) could also be used to evaluate the confidence we have in DNN-based Language Models.

5. ACKNOWLEDGMENTS

The research leading to these results has received funding from Swiss National Science Foundation project SHISSM (Sparse and hierarchical Structures for Speech Modeling), grant agreement 200021-175589, and the European Community H2020 SUMMA (Scalable Understanding of Multilingual Media) project No. 688139.

6. REFERENCES

[1] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[2] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016.

[3] Yarin Gal, *Uncertainty in Deep Learning*, Ph.D. thesis, University of Cambridge, 2016.

[4] Alex Kendall, Vijay Badrinarayanan, , and Roberto Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[5] Alex Kendall and Roberto Cipolla, "Modelling uncertainty in deep learning for camera relocalization," *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4762–4769, 2016.

[6] Thomas Kemp and Thomas Schaaf, "Estimating confidence using word lattices," in *EUROSPEECH*, 1997.

[7] Frank Wessel, Klaus Macherey, and Ralf Schluter, "Using word probabilities as confidence measures," in *Proceedings ICASSP*, 1998.

[8] Gunnar Evermann and Philip C Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proceedings ICASSP*, 2000.

[9] Lidia Mangu, Eric Brill, and Andreas Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *EUROSPEECH*, 1999.

[10] Julien Fayolle, Fabienne Moreau, Christian Raymond, Guillaume Gravier, and Patrick Gros, "CRF-based combination of contextual features to improve a posteriori word-level confidence measures," in *Proceedings of Interspeech*, 2010.

[11] Matthew Stephen Seigel and Philip C Woodland, "Combining information sources for confidence estimation with CRF models," in *Proceedings of Interspeech*, 2011.

[12] Atsunori Ogawa, Takaaki Hori, and Atsushi Nakamura, "Error type classification and word accuracy estimation using alignment features from word confusion network," in *Proceedings ICASSP*, 2012.

[13] Thomas Schaaf and Thomas Kemp, "Confidence measures for spontaneous speech recognition," in *Proceedings ICASSP*, 1997.

[14] Dong Yu, Jinyu Li, and Li Deng, "Calibration of confidence measures in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[15] Yik-Cheung Tam, Yun Lei, Jing Zheng, and Wen Wang, "ASR error detection using recurrent neural network language model and complementary ASR," in *Proceedings ICASSP*, 2014.

[16] Ahmed Ali and Steve Renals, "Word error rate estimation for speech recognition: e-WER," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[17] Atsunori Ogawa and Takaaki Hori, "Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks," *Speech Communication*, 2017.

[18] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. IEEE, 1992, vol. 1, pp. 517–520.

[19] Karel Veselý, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of Interspeech 2013*. 8 2013, ISCA.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[21] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016.

[22] Haihua Xu, Daniel Povey, Lidia Mangu, and Jie Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802 – 828, 2011.