CALIBRATION OF ENCODER DECODER MODELS FOR NEURAL MACHINE TRANSLATION

Aviral Kumar[*]
University of California Berkeley
aviralk@berkeley.edu

Sunita Sarawagi
Indian Institute of Technology Bombay
sunita@iitb.ac.in

ABSTRACT

Calibration of black-box prediction models like neural networks is required for interpretability of its confidence. For a well calibrated model, the average prediction confidence is equal to average prediction accuracy, thereby providing an interpretable notion to the model confidence. We study the calibration of several state of the art neural machine translation (NMT) systems built on attention-based encoder-decoder models. For structured outputs like in NMT, calibration is important not just for reliable confidence with predictions, but also for proper functioning of beam-search inference. We show that most modern NMT models are surprisingly miscalibrated even when conditioned on the true previous tokens. Our investigation leads to two main reasons severe miscalibration of EOS (end of sequence marker) and suppression of attention uncertainty. We design recalibration methods based on these signals and demonstrate improved accuracy, more intuitive results from beam-search and better sequence-level calibration.

1 INTRODUCTION

Calibration of supervised learning models is a topic of continued interest in machine learning and statistics (Niculescu-Mizil and Caruana, 2005; Candela et al., 2005; Crowson et al., 2016; Guo et al., 2017). Calibration requires that the probability a model assigns to a prediction equals the true chance of correctness of the prediction. For example, if a calibrated model M makes 1000 predictions with probability values around 0.99, we expect 990 of these to be correct. Calibration is important in real-life deployments of a model since it ensures interpretable probabilities. In this paper we show that for structured prediction models calibration is also important for sound working of the inference algorithm that generates structured outputs.

Much recent work have studied calibration of modern neural networks for scalar predictions (Guo et al., 2017; Lakshminarayanan et al., 2017; Hendrycks and Gimpel, 2017; Louizos and Welling, 2017; Pereyra et al., 2017; Kumar et al., 2018; Kuleshov et al., 2018). Modern neural networks have been found to be miscalibrated in the direction of over-confidence, in spite of a statistically sound log-likelihood based training objective.

We investigate calibration of attention-based encoder-decoder models for sequence to sequence (seq2seq) learning as applied to neural machine translation. We measure calibration of token probabilities of three modern neural architectures for translation — NMT (Bahdanau et al. (2015)), GNMT (Wu et al. (2016)), and the Transformer model (Vaswani et al. (2017)) on six different benchmarks. We find the output token probabilities of these models to be poorly calibrated. This is surprising because the output distribution is conditioned on *true* previous tokens (teacher forcing) where there is no train-test mismatch unlike when we condition on *predicted* tokens where there is a risk of exposure bias (Bengio et al. (2015); Ranzato et al. (2016); Norouzi et al. (2016); Wiseman and Rush (2016)). We show that such lack of calibration can explain the counter-intuitive BLEU drop with increasing beam-size (Koehn and Knowles (2017)).

We dig into root causes for the lack of calibration and pin point two primary causes: poor calibration of the EOS token and attention uncertainty. Instead of generic temperature based fixes as in (Guo

[*] Work done when the author was a student at IIT Bombay

et al., 2017), we propose a parametric model to recalibrate as a function of input coverage, attention uncertainty, and token probability. We show that our approach leads to improved token-level calibration. We demonstrate three advantages of a better calibrated model. First, we show that the calibrated model better correlates probability with BLEU and that leads to BLEU increment by up to 0.4 points just by recalibrating a pre-trained model. Second, we show that improved calibration diminishes the drop in BLEU with increasing beam-size. Third, we show that the calibrated model has better calibration on the per-sequence BLEU metric, which we refer to as sequence-level calibration and was achieved just by recalibrating token-level probabilities.

2 Background and Calibration Measures

Attention-based NMT State of the art NMT systems use an attention-based encoder-decoder neural network for modeling $\Pr(\mathbf{y}|\mathbf{x}, \theta)$ over the space of discrete output translations of an input sentence $\mathbf{x}$ where θ denotes the network parameters. Let $y_1, \ldots, y_n$ denote the tokens in a sequence $\mathbf{y}$ and $x_1, \ldots, x_k$ denote tokens in $\mathbf{x}$. Let V denote output vocabulary. A special token EOS $\in V$ marks the end of a sequence in both $\mathbf{x}$ and $\mathbf{y}$. First, an encoder (e.g. a bidirectional LSTM) transforms each $x_1, \ldots, x_k$ into real-vectors $\mathbf{h}_1, \ldots, \mathbf{h}_k$. The Encoder-Decoder (ED) network factorizes $\Pr(\mathbf{y}|\mathbf{x}, \theta)$ as $\Pr(\mathbf{y}|\mathbf{x}, \theta) = \prod_{t=1}^{n} \Pr(y_t|\mathbf{y}_{<t}, \mathbf{x}, \theta)$ where $\mathbf{y}_{<t} = y_1, \ldots, y_{t-1}$. The decoder computes each $\Pr(y_t|\mathbf{y}_{<t}, \mathbf{x}, \theta)$ as $\Pr(y_t|\mathbf{y}_{<t}, \mathbf{x}, \theta) = \text{softmax}(\theta_{y_t} F_\theta(\mathbf{s}_t, \mathbf{H}_t))$ where $\mathbf{s}_t$ is a decoder state summarizing $y_1, \ldots y_{t-1}$; $\mathbf{H}_t$ is attention weighted input: $\mathbf{H}_t = \sum_{j=1}^{k} \mathbf{h}_j \alpha_{jt}$, $\boldsymbol{\alpha}_t = \text{softmax}(A_\theta(\mathbf{h}_j, \mathbf{s}_t))$ and $A_\theta(.,.)$ is the attention unit. During training given a $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}$, we find θ to minimize negative log likelihood (NLL): $\text{NLL}(\theta) = -\sum_{i \in D} \sum_{t=1}^{|\mathbf{y}_i|} \log \Pr(y_{it}|\mathbf{y}_{i,<t}, \mathbf{x}_i, \theta)$. During inference given a $\mathbf{x}$, we need to find the $\mathbf{y}$ that maximizes $\Pr(\mathbf{y}|\mathbf{x})$. This is intractable given the full autoregressive structure. Approximations like beam search with a beam-width parameter B (typically between 4 and 12) maintains B highest probability prefixes which are grown token at a time. At each step beam search finds the top-B highest probability tokens from $\Pr(y|\mathbf{y}_{b,<t}, \mathbf{x}, \theta)$ for each prefix $\mathbf{y}_{b,<t}$ until a EOS is encountered.

Calibration: Definition, Measures We study, analyze, and fix the calibration of the next token distribution $Pr(y_{it}|\mathbf{y}_{i,<t}, \mathbf{x}_i, \theta)$ (shortened as $P_{it}(y)$ here). A model $P_{it}(y)$ is well-calibrated if for any value $\beta \in [0, 1]$, of all predictions $y \in V$ with probability β, the fraction correct is β. A graphical measure of calibration error is via reliability plots that bins β values into small ranges within which it shows the average confidence of the highest probability token against the average accuracy of that prediction. In a well-calibrated model the plot lies on the diagonal. Figure 1 shows several examples of calibration plots of two models with 20 bins of β each of size 0.05. The absolute difference between the diagonal and the observed plot scaled by bin frequency is called **expected calibration error (ECE)**. We formally define ECE and a weighted version of it that measures calibration error of the entire distribution in Appendix A.

Importance of Calibration For sequence models, we show that calibration of P_{it} is important not just for interpretability but also for the sound working of beam-search inference. Consider an example: say we have an input for which the correct two-token sequence output is "*That's awesome*". Let's say for the first token position the outputs from a miscalibrated model P_1, and a true distribution (P^*) be

$$P_1(y = \textit{It's}) = 0.4, P_1(y = \textit{That's}) = 0.6 \qquad P_1^*(y = \textit{It's}) = 0.3, P_1^*(y = \textit{That's}) = 0.7.$$

Assume at $t = 2$, the model is calibrated and

$$P_2(\textit{ok}|\textit{It's}) = 0.91, P_2(\textit{awesome}|\textit{That's}) = 0.6.$$

The highest probability prediction from the uncalibrated model is *It's ok* with probability 0.4×0.91, whereas from the calibrated model is *That's awesome*. Thus, accuracy of model P is 0 and the calibrated P^* is 1 even though the relative ordering of token probabilities is the same in P and P^*. If we used beam size of 1, we would get the correct output although with the lower score 0.6×0.6, whereas the higher scoring (0.4×0.91) output with beam size of 2 is wrong. More generally, increasing the beam size almost always outputs a higher scoring prefix but if the score is not calibrated that does not guarantee more correct outputs. The more prefixes we have with over-confident (miscalibrated) scores, the higher is our chance of over-shadowing the true best prefix in the next step, causing accuracy to drop with increasing beam-size. We observe this on real data too.

3 Calibration of existing models

We study the calibration of six state-of-the-art publicly available pre-trained NMT models on various WMT+IWSLT benchmarks. These include: En-De GNMT (4/8 layers), De-En GNMT, De-En NMT, En-Vi NMT and Transformer En-De or En-De T2T. Details are provided in the appendix.

Figure 1 shows calibration as a reliability plot where x-axis is average weighted confidence and y-axis is average weighted accuracy. The blue lines are for the original models and the red lines are after our fixes (to be ignored in this section). The figure also shows calibration error (ECE). We observe that all six models are miscalibrated to various degrees with ECE ranging from 2.9 to 9.8. Five of the six models are overly confident. The transformer model attempts to fix the over-confidence by using a soft cross-entropy loss that assigns a probability 0.9 to the correct label and probability $\frac{0.1}{V-1}$ to others. With this loss, over-confidence changes to slight under-confidence.

This observed miscalibration was surprising given the tokens are conditioned on the *true* previous tokens (teacher forcing). In teacher-forcing, the NLL loss is statistically sound and should be calibrated.

3.1 Reasons for Miscalibration

In this section we seek out reasons for and symptoms induced by the observed miscalibration of modern NMT models. For scalar classification, Guo et al. (2017) discusses reasons for poor calibration of neural nets. A primary reason is that the high capacity of NN causes the negative log likelihood (NLL) to overfit without overfitting 0/1 error (Zhang et al. (2017)). We show that for seq2seq models based on attention and with large vocabulary, a different set of reasons also come into play. We identify three of these. While these are not exclusive reasons, we show that correcting them improves calibration and partly fixes other symptoms of miscalibrated models. These include: extremely **poor calibration of the EOS token** (Appendix C.1), high **uncertainty of attention** leading to higher miscalibration (Appendix C.2) and the differences in the calibration of **head and tail tokens** (Appendix C.3). We include these in the appendix due to lack of space.

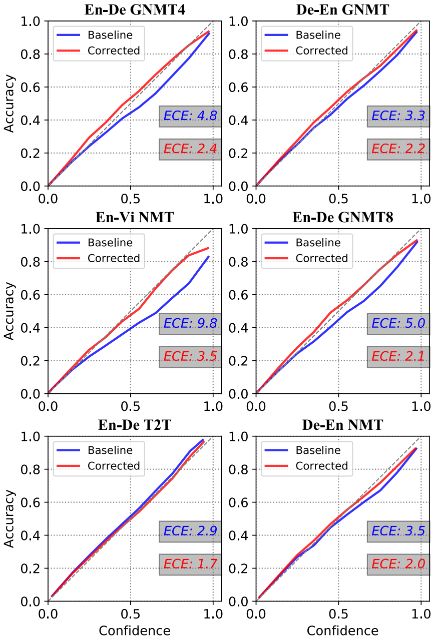


Figure 1: Reliability Plots for various baseline models on the test sets along with their ECE values(Blue) and calibrated models(Red). The x-axis is expected confidence after binning into 0.05 sized bins and y-axis is accuracy in that confidence bin. ECE values in corresponding colors.

4 Reducing Calibration Errors

For modern neural classifiers Guo et al. (2017) compares several post-training fixes and finds temperature scaling to provide the best calibration without dropping accuracy. This method chooses a positive temperature value T and transforms the $P_{it}(y)$ distribution as $\propto P_{it}(y)^{\frac{1}{T}}$. The optimal T is obtained by maximizing NLL on a held-out validation dataset.

Our investigation in Section 3.1/Appendix C showed that calibration of different tokens in different input contexts varies significantly. We propose an alternative method, where the temperature value is not constant but varies based on the entropy of the attention, the log probability of the token, the token's identity (EOS or not) and the input coverage. At the t-th decoding step, let $a_t = \mathcal{H}(\boldsymbol{\alpha}_t)$ denote the entropy of the attention vector $\boldsymbol{\alpha}_t$ and the logit for a token y at step t be $l_{ty} = \log \Pr(y|\mathbf{y}_{<t}, \mathbf{x}, \theta)$. We measure coverage c_t as the fraction of input tokens with cumulative attention until t greater than a threshold δ. We used $\delta = 0.35$. Using (a_t, c_t, l_{ty}) we compute the (inverse of) temperature for scaling token y at step t in two steps. We first correct the extreme miscalibration of EOS by learning a correction as a function of the input coverage c_t as follows: $l'_{ty} = l_{ty} + [[y = \text{eos}]] \log \left(\sigma(w_1(c_t - w_2))\right)$ This term helps to dampen EOS probability when input coverage c_t is low and w_1, w_2 are learned parameters. Next, we correct for overall miscalibration by using a neural network to learn variable temperature values as follows: $T_{ty}^{-1}(a_t, l'_{ty}, c_t|\mathbf{w}) = g_{\mathbf{w}}(a_t) \cdot h_{\mathbf{w}}(l'_{ty})$ where $g_{\mathbf{w}}(.)$ and $h_{\mathbf{w}}(.)$ are functions with parameters $\mathbf{w}$. For each of $g_{\mathbf{w}}(.)$ and $h_{\mathbf{w}}(.)$, we use a 2-layered feed-forward network with hidden ReLu activation,

Model Name	ECE			BLEU		
	Base	Our	T	Base	Our	T
En-Vi NMT	9.8	**3.5**	3.8	26.2	**26.6**	26.0
En-De GNMT4	4.8	**2.4**	2.7	**26.8**	**26.8**	26.7
En-De GNMT8	5.0	2.2	**2.1**	**27.6**	27.5	27.4
De-En GNMT	3.3	**2.2**	2.3	29.6	**29.9**	29.6
De-En NMT	3.5	**2.0**	2.2	28.8	**29.0**	28.7
T2T En-De	2.9	**1.7**	5.4	27.9	**28.1**	28.1
T2T En-De(B=4)				**28.3**	**28.3**	28.2

Table 1: Weighted Expected Calibration Errors on test data of models and baseline calibrated by two different methods. BLEU is without length normalization. Beam Size(B) is 10 in each case if not mentioned.

Model	B=10	B=20	B=40	B=80
En-Vi NMT	23.8	-0.2	-0.4	-0.7
+ calibrated	24.1	-0.2	-0.2	-0.4
En-De GNMT4	23.9	-0.1	-0.2	-0.4
+ calibrated	23.9	-0.0	-0.0	-0.1
En-De GNMT8	24.6	-0.1	-0.3	-0.5
+ calibrated	24.7	-0.1	-0.4	-0.6
De-En GNMT	28.8	-0.2	-0.3	-0.5
+ calibrated	28.9	-0.1	-0.2	-0.3
De-En NMT	28.0	-0.1	-0.4	-0.6
+ calibrated	28.2	-0.0	-0.2	-0.2

Table 2: BLEU with increasing beam on the devset. Note the steep decrease in the case of the baseline model, which is countered by calibrating the model.

three units per hidden layer, and a sigmoid activation function to output in range $(0, 1)$. We learn parameters $\mathbf{w}$ (including w_1 and w_2) by minimizing NLL on temperature adjusted logits using a validation set D_V. The objective is: $-\sum_{i \in D_V} \sum_{t=1}^{|\mathbf{y}_i|} \log(\text{softmax}_y(l'_{ity} T_{ty}^{-1}(a_{it}, l'_{ity}, c_{it}|\mathbf{w}))[y_{it}])$, where $l_{ity} = \log \Pr(y|\mathbf{y}_{i,<t}, \mathbf{x}_i, \theta)$ and l'_{ty} is as defined earlier. The held-out validation set D_V was created using a 1:1 mixture of 2000 examples sampled from the train and dev set.

5 Experimental Results

We first show that our method manages to significantly reduce calibration error on the test set. Then we present two outcomes of a better calibrated model: (1) higher accuracy, and (2) reduced BLEU drop with increasing beam-size.

Figure 1 shows that our method (shown in red) reduces miscalibration of all six models – in all cases our model is closer to the diagonal than the original. We manage to both reduce the under-estimation of the T2T model and the over-confidence of the NMT and GNMT models. We compare ECE of our method of recalibration to the single temperature method in Table 1 (Column ECE). Note the single temperature is selected using the same validation dataset as ours.

More Accurate Predictions For structured outputs with beam-search like inference, temperature scaling can lead to different MAP solutions unlike scalar classification where it doesn't affect the identity of the prediction. In Table 1 we show the BLEU score with different methods. We report BLEU without length norm here. In almost all cases, our informed recalibration improves inference accuracy. The gain with calibration is more than 0.3 units in BLEU on three models: En-Vi, De-En GNMT and En-De T2T. The increase in accuracy is modest but significant because they came out of only tweaking the token calibration of an existing trained model using a small validation dataset. Also, note fixed temperature actually *hurts accuracy (*BLEU*).* In five of the six models, the BLEU after recalibrating with temperature drops, even while the ECE reduction is comparable to ours. This highlights the importance of accounting for factors like coverage and attention entropy for achieving sound recalibration.

BLEU drop with increasing Beam-Size One idiosyncrasy of modern NMT models is the drop in BLEU score as the inference is made more accurate with increasing beam-size. In Table 2 we show the BLEU scores of original models and our calibrated versions with beam size increasing from 10 to 80. BLEU drops much more with the original model than with the calibrated one. For example for En-De GNMT4, BLEU drops from 23.9 to 23.8 to 23.7 to 23.5 as beam width B is increased from 10 to 20 to 40 to 80, whereas after calibration it is more stable going from 23.9 to 23.9 to 23.9 to 23.8. The BLEU drop is reduced but not totally eliminated since we have not achieved perfect calibration. One fix to reduce this BLEU drop is length normalization which is a heuristic. Recalibration is more principled that also provides interpretable scores as a by-product.

We further define a measure of sequence-level calibration and show how our fixes for token-wise miscalibration can also improve sequence-level calibration as well in Appendix D.1.

6 Conclusion

Calibration is an important for interpretability, bias reduction, and for structured outputs also for sound functioning of beam-search or any approximate inference method. We measured the calibration of six state-of-the-art attention-based NMT models and found token probabilities of all to be miscalibrated even when conditioned on true previous tokens. Going deeper, we identified two main reasons for miscalibration – attention uncertainty and differential miscalibration of crucial tokens like EOS. We designed a parametric model to recalibrate based on these factors. We achieved significant reduction in ECE and an increase in accuracy. Improved calibration leads to greater correlation between probability and error and this manifests as reduced BLEU drop with increasing beam-size. In the appendix, we show how improving token calibration also leads to better sequence-level calibration.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. https://arxiv.org/abs/1409.0473 Neural machine translation by jointly learning to align and translate. *ICLR*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*.

Joaquin Quiñonero Candela, Carl Edward Rasmussen, Fabian H. Sinz, Olivier Bousquet, and Bernhard Schölkopf. 2005. Evaluating predictive uncertainty challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 1–27.

Cynthia S Crowson, Elizabeth J Atkinson, and Terry M Therneau. 2016. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25(4):1692–1706.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1321–1330.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with smt features. In *AAAI*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. 2018. Fast and scalable bayesian deep learning by weight-perturbation in adam. In *Proceedings of the 35th International Conference on Machine Learning, ICML*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2801–2809.

Volodymyr Kuleshov and Percy S Liang. 2015. Calibrated structured prediction. In *NIPS*, pages 3474–3482.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. 2018. Trainable calibration measures from kernel mean embeddings. In *ICML*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, pages 6405–6416.

Christos Louizos and Max Welling. 2017. Multiplicative normalizing flows for variational Bayesian neural networks. In *ICML*, volume 70, pages 2218–2227.

Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. 2017. Neural machine translation (seq2seq) tutorial. *https://github.com/tensorflow/nmt*.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *EMNLP*, pages 1587–1598.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *ICML*.

Mohammad Norouzi, Samy Bengio, zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. 2016. http://papers.nips.cc/paper/6547-reward-augmented-maximum-likelihood-for-neural-structured-prediction.pdf Reward augmented maximum likelihood for neural structured prediction. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1723–1731. Curran Associates, Inc.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *ICLR workshop*.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press.

M Ranzato, S Chopra, M Auli, and W Zaremba. 2016. Sequence level training with recurrent neural networks. *ICLR*.

Pavel Sountsov and Sunita Sarawagi. 2016. Length bias in encoder decoder models and a case for global conditioning. In *EMNLP*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *EMNLP*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. http://arxiv.org/abs/1609.08144 Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018a. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *EMNLP*.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. 2018b. https://openreview.net/forum?id=HkwZSG-CZ Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *ACM SIGKDD*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. http://arxiv.org/abs/1611.03530 Understanding deep learning requires rethinking generalization. *ICLR*.

APPENDIX: CALIBRATION OF ENCODER DECODER MODELS FOR NEURAL MACHINE TRANSLATION

A FORMAL DESCRIPTION OF ECE AND WEIGHTED ECE

In this section, we describe in detail the formulae of ECE and weighted ECE, and also provide a concrete example which indicates why weighted ECE is desired over vanilla ECE used in scalar calibration. We first revisit ECE more formally and then describe weighted ECE.

Expected Calibration Error (ECE) ECE is defined when a model makes a single prediction $\hat{y}$ with a confidence p. In the case of scalar prediction or considering just the topmost token in structured prediction tasks, the prediction is $\hat{y}_{it} = \arg\max_y P_{it}(y)$ with $P_{it}(\hat{y}_{it})$ as confidence. Let $C_{it}(y) = \delta(y_{it} = y)$ denote if y matches the correct label y_{it} at (i, t).

First partition the confidence interval [0..1] into M equal bins $I_1, \ldots, I_M$. Then in each bin measure the absolute difference between the accuracy and confidence of predictions in that bin. This gives the expected calibration error (ECE) as:

$$\frac{1}{L} \sum_{b=1}^{M} \left| \sum_{i,t: P_{it}(\hat{y}_{it}) \in I_b} C_{it}(\hat{y}_{it}) - P_{it}(\hat{y}_{it}) \right| \tag{1}$$

where $L = \sum_i^N |\mathbf{y}_i|$ is total output token lengths (or total number of scalar predictions made). Since beam-search reasons over probability of multiple high scoring tokens, we wish to calibrate the entire distribution. If V is the vocabulary size, we care to calibrate all LV predicted probabilities. A straightforward use of ECE that treats these as LV independent scalar predictions is incorrect, and is not informative.

Weighted ECE Weighted ECE is given by the following formula: (various symbols have usual meanings as used in the rest of this paper)

$$\frac{1}{L} \sum_{b=1}^{M} \left| \sum_{i,t} \sum_{y: P_{it}(y) \in I_b} P_{it}(y)(\delta(y_{it} = y) - P_{it}(y)) \right|.$$

We motivate our definition as applying ECE on a classifier that predicts label y with probability proportional to its confidence $P_{it}(y)$ instead of the highest scoring label deterministically.

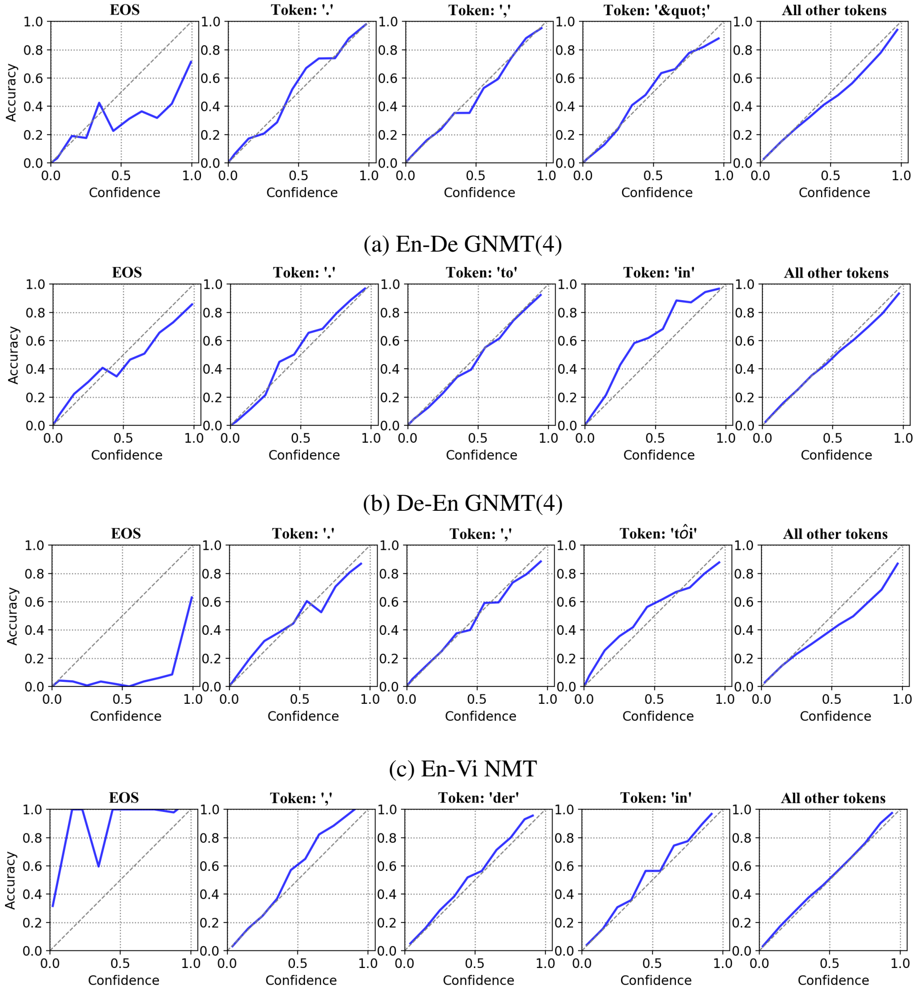
We go over an example to highlight how weighted ECE calibrates the full distribution. Consider two distributions on a V of size 3: $P_1(.) = [0.4,\ 0.1,\ 0.5]$ and $P_2(.) = [0.0,\ 0.5,\ 0.5]$. For both let the first label be correct. Clearly, P_1 with correct label probability of 0.4 is better calibrated than P_2. But ECE of both is the same at $|0 - 0.5| = 0.5$ since both of theirs highest scoring prediction (label 3) is incorrect. In contrast, with bins of size 0.1, weighted ECE will be $0.4|1 - 0.4| + 0.1|0 - 0.1| + 0.5|0 - 0.5| = 0.42$ for P_1 which is less than 0.5 for P_2. Such fine-grained distinction is important for beam-search and any other structured search algorithms with large search spaces. In the paper we used ECE to denote weighted ECE.

B DETAILS OF THE NMT MODELS USED FOR OUR STUDY

The first five are from Tensorflow's NMT codebase Luong et al. (2017) [1]: En-De GNMT (4 layers), En-De GNMT (8 layers), De-En GNMT, De-En NMT, En-Vi NMT. They all use multi-layered LSTMs arranged either in the GNMT architecture Wu et al. (2016) or standard NMT architecture Bahdanau et al. (2015). The sixth En-De T2T, is the pre-trained Transformer model[2]. (We use T2T and Transformer interchangeably.) The T2T replaces LSTMs with self-attention Vaswani et al. (2017) and uses multiple attention heads, each with its own attention vector.

[1] https://github.com/tensorflow/nmt#benchmarks

[2] https://github.com/tensorflow/tensor2tensor/tree/master/tensor2tensor, pre-trained model at https://goo.gl/wkHexj



(a) En-De GNMT(4)

(b) De-En GNMT(4)

(c) En-Vi NMT

(d) En-De T2T (Transformer)

Figure 2: Tokenwise Calibration plots for some of the models. Note the miscalibration of EOS vs the calibration of other tokens. All other tokens roughly show a similar trend as the overall calibration plot.

C Reasons for Miscalibration (Details and Empirical Evidence)

C.1 Poor calibration of EOS token

To investigate further we drill down to token-wise calibration. Figure 2 shows the plots of EOS, three other frequent tokens, and the rest for four models. Surprisingly, EOS is calibrated very poorly and is much worse than the overall calibration plots in Figure 1 and other frequent tokens. For NMT and GNMT models EOS is over-estimated, and for T2T the EOS is under-estimated. For instance, for the En-De GNMT(4) model (top-row, first column in Fig 2), out of all EOS predictions with confidence in the [0.9, 0.95] bin only 60% are correct. Perhaps these encoder-decoder style models do not harness enough signals to reliably model the end of a sequence. One such important signal is coverage of the input sequence. While coverage has been used heuristically in beam-search inference Wu et al. (2016), we propose a more holistic fix of the entire distribution using coverage as one of the features in Section 4.

C.2 Uncertainty of Attention

We conjectured that a second reason for over-confidence could be the uncertainty of attention. A well-calibrated model must express all sources of prediction uncertainty in its output distribution. Existing attention models average out the attention uncertainty of $\boldsymbol{\alpha}_t$ in the input context $\mathbf{H}_t$. Thereafter, $\boldsymbol{\alpha}_t$ has no influence on the output distribution. We had conjectured that this would manifest as worser calibration for high entropy attentions $\boldsymbol{\alpha}_t$, and this is what we observed empirically. In Table 3 we show ECE partitioned across whether the entropy of $\boldsymbol{\alpha}_t$ is high or low on five[3] models. Observe that ECE is higher for high-entropy attention.

Model Name	**Low $\mathcal{H}$**	**High $\mathcal{H}$**
En-Vi NMT	9.0	13.0
En-De GNMT(4)	4.5	5.3
En-De GNMT(8)	4.8	5.4
De-En GNMT$^{\square}$	3.8	2.3
De-En GNMT$^{\otimes}$	3.9	5.9
De-En NMT	2.3	4.1

Table 3: ECE(%) for the high and low attention entropy zones. High entropy is defined as $\mathcal{H} \geq 1.0$; ($\square$ represents the ECE for the entire set of samples, $\otimes$ represents the ECE for the samples with prediction probability in $0.8 - 1.0$ – this was done to see how attention entropy correlates with calibration in the high confidence prediction range).

C.3 Head versus Tail tokens

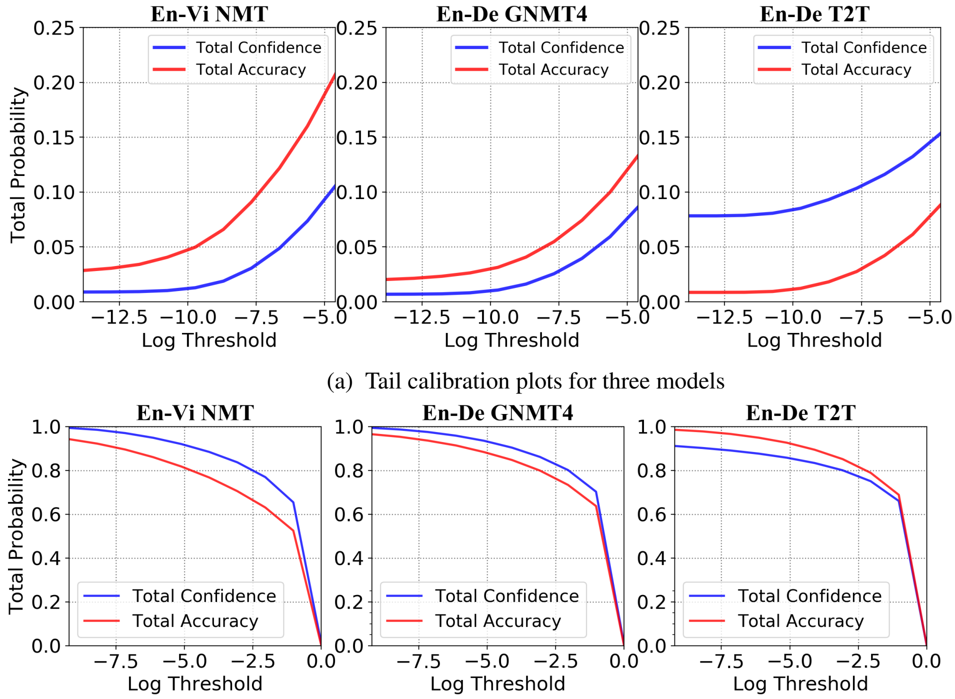
The large vocabulary and the softmax bottleneck Yang et al. (2018b) was another reason we investigated. We studied the calibration for tail predictions (the ones made with low probability) in contrast to the head in a given softmax distribution. In Figure 3a for different thresholds T of log probability (X-axis), we show total true accuracy (red) and total predicted confidence (blue) for all predictions with confidence less than T. In Figure 3b we show the same for head predictions with confidence $> T$. The first two from GNMT/NMT under-estimate tail (low) probabilities while over-estimating the head. The T2T model shows the opposite trend. This shows that the phenomenon of miscalibration manifests in the entire softmax output and motivates a method of recalibration that is sensitive to the output token probability.

D Benefits of Calibrating NMT Models

D.1 An interpretable measure of whole sequence calibration

For structured outputs like in translation, the whole sequence probability is often quite small and an uninterpretable function of output length and source sentence difficulty. In general, designing a good calibration measure for structured outputs is challenging. Nguyen and O'Connor (2015) propose to circumvent the problem by reducing structured calibration to the calibration of marginal probabilities over single variables. This works for tractable joint distributions like chain CRFs and HMMs. For modern NMT systems that assume full dependency, such marginalization is neither tractable nor useful. We propose an alternative measure of calibration in terms of BLEU score rather than structured probabilities. We define this measure using BLEU but any other scoring function including gBLEU, and Jaccard are easily substitutable.

[3] We drop the T2T model since measuring attention uncertainty is unclear in the context of multiple attention heads.



(a) Tail calibration plots for three models

(b) Head calibration plots for three models.

Figure 3: Tail and Head Calibration Plots for 3 models. Note that the head is overestimated in GNMT/NMT, underestimated in T2T and the tail shows the opposite trend. Here the x-axis corresponds to the log of the fraction of vocabulary that is classified as tail prediction.

We define model expected BLEU_θ of a prediction $\hat{\mathbf{y}}$ as value of BLEU if true label sequences were sampled from the predicted distribution $\Pr(\mathbf{y}|\mathbf{x}_i, \theta)$

$$\begin{aligned}\text{BLEU}_\theta(\hat{\mathbf{y}}) &= \text{E}_{\mathbf{y}\sim P(\mathbf{y}|\mathbf{x}_i,\theta)}[\text{BLEU}(\hat{\mathbf{y}}, \mathbf{y})] \\ &\approx \frac{1}{T}\sum_{m=1}^{T}[\text{BLEU}(\hat{\mathbf{y}}, \mathbf{y}_{im})]\end{aligned} \tag{2}$$

where $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT}$ denote T samples from $P(\mathbf{y}|\mathbf{x}_i, \theta)$.[4]

It is easy to see that if $P(\mathbf{y}|\mathbf{x}_i, \theta)$ is perfectly calibrated the model predicted BLEU_θ will match the actual BLEU on the true label sequence $\mathbf{y}_i$ in expectation. That is, if we considered all predictions with predicted $\text{BLEU}_\theta(\hat{\mathbf{y}}) = \alpha$, then the actual BLEU over them will be α when θ is well-calibrated. This is much like ECE for scalar classification except that instead of matching 0/1 accuracy with confidence, we match actual BLEU with expected BLEU. We refer to this as *Structured ECE* in our results (Table 4).

Figure 4 shows the binned values of $\text{BLEU}_\theta(\hat{\mathbf{y}})$ (X-axis) and average actual BLEU (Y-axis) for WMT + IWSLT tasks on the baseline model and after recalibrating (solid lines). In the same plot we show the density (fraction of all points) in each bin by each method. We use $T = 100$ samples for estimating $\text{BLEU}_\theta(\hat{\mathbf{y}})$. Table 1 shows aggregated difference over these bins. We can make a number of observations from these results.

The calibrated model's BLEU plot is closer to the diagonal than baseline's. Thus, for a calibrated model the $\text{BLEU}_\theta(\hat{\mathbf{y}})$ values provide a interpretable notion of the quality of prediction. The only exception is the T2T model. The model has very low entropy on token probabilities and the top

[4] We could also treat various sequences obtained from beam search with large beam width as samples (unless these are very similar) and adjust the estimator by the importance weights. We observed that both explicit sampling and re-weighted estimates with beam-searched sequences give similar results.

Model Name	**ECE**			**BLEU**			**Structured ECE**	
	Base	Our	T	Base	Our	T	Base	Our
En-Vi NMT	9.8	**3.5**	3.8	26.2	**26.6**	26.0	7.3	**0.9**
En-De GNMT4	4.8	**2.4**	2.7	**26.8**	**26.8**	26.7	5.8	**3.4**
En-De GNMT8	5.0	2.2	**2.1**	**27.6**	27.5	27.4	6.4	**3.3**
De-En GNMT	3.3	**2.2**	2.3	29.6	**29.9**	29.6	2.5	**1.3**
De-En GNMT (Lnorm)	3.3	**2.2**	2.3	29.9	**30.1**	30.1	2.5	**1.3**
De-En NMT	3.5	**2.0**	2.2	28.8	**29.0**	28.7	4.0	**2.4**
T2T En-De	2.9	**1.7**	5.4	27.9	**28.1**	28.1	98.8	98.8
T2T En-De (B=4)	2.9	**1.7**	5.4	**28.3**	**28.3**	28.2	98.8	98.8

Table 4: Expected Calibration Errors of baseline and models calibrated by two different methods on test set. Structured ECE refers to the ECE from the reliability plot of expected BLEU. We repeat BLEU and ECE numbers from Table 1 for completeness and for easy comparison. Length Norm in all other cases gave rise to worse BLEUscores except De-En GNMT.

100 sequences are only slight variants of each other, and the samples are roughly identical. An interesting topic for future work is further investigating the reasons behind the T2T model being so sharply peaked compared to other models.

The baseline and calibrated model's densities (shown in dotted) are very different with the calibrated model showing a remarkable shift to the low end. The trend in density is in agreement with the observed BLEU scores, and hence higher density is observed towards the lower end.

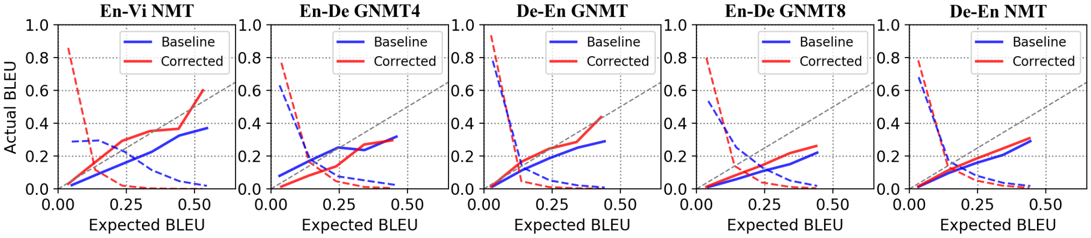


Figure 4: Sequence level calibration plots for various models [Baseline + Corrected(Calibrated)]. The dotted lines shows the densities (fraction of all points) in each bin. Note that the density in all the cases shifts to the low end, showing that overestimation is reduced. This trend in the density is same across all models and the calibrated densities are more in agreement with the observed BLEU on the datasets (test datasets).

E RELATED WORK

Calibration of scalar classification and regression models has been extensively studied. Niculescu-Mizil and Caruana (2005) systematically evaluated many classical models and found models trained on conditional likelihood like logistic regression and neural networks (of 2005) to be well-calibrated, whereas SVMs and naive Bayes were poorly calibrated. Nguyen and O'Connor (2015) corroborated this for NLP tasks. Many methods are proposed for fixing calibration including Platt's scaling Platt (1999), Isotonic regression Zadrozny and Elkan (2002), and Bayesian binning Naeini et al. (2015), and training regularizers like MMCE Kumar et al. (2018). A principled option is to capture parameter uncertainty using Bayesian methods. Recently, these have been applied on DNNs using variational methods Louizos and Welling (2017), ensemble methods Lakshminarayanan et al. (2017), and weight perturbation-based training Khan et al. (2018).

For modern neural networks, a recent systematic study Guo et al. (2017) finds them to be poorly calibrated and finds temperature scaling to provide the best fix. We find that temperature scaling is inadequate for more complicated structured models where different tokens have very different dynamics. We propose a more precise fix derived after a detailed investigation of the root cause for the lack of calibration.

Going from scalar to structured outputs, Nguyen and O'Connor (2015) investigates calibration for NLP tasks like NER and CoRef on log-linear structured models like CRFs and HMMs. They define calibration on token-level and edge-level marginal probabilities of the model. Kuleshov and Liang (2015) generalizes this to structured predictions. But these techniques do not apply to modern NMT networks since each node's probability is conditioned on all previous tokens making node-level marginals both intractable and useless.

Concurrently with our work, Ott et al. (2018) studied the uncertainty of neural translation models where their main conclusion was that existing models "spread too much probability mass across sequences". However, they do not provide any fix to the problem. Another concern is that their observations are only based on the FairSeq's CNN-based model, whereas we experiment on a much larger set of architectures. Our initial measurements on a pre-trained En-Fr FairSeq model[5] found the model to be well-calibrated (also corroborated in their paper) unlike the six architectures we present here (which they did not evaluate). An interesting area of future work is to explore the reasons for this difference.

The problem of drop in accuracy with increasing beam-size and length bias has long puzzled researchers **?**Sountsov and Sarawagi (2016); Koehn and Knowles (2017) and many heuristic fixes have been proposed including the popular length normalization/coverage penalty Wu et al. (2016), word reward He et al. (2016), and bounded penalty Yang et al. (2018a). These heuristics fix the symptoms by delaying the placement of the EOS token, whereas ours is the first paper that attributes this phenomenon to the lack of calibration. Our experiments showed that miscalibration is most severe for the EOS token, but it affects several other tokens too. Also, by fixing calibration we get more useful output probabilities, which is not possible by these fixes to only the BLEU drop problem.

[5] Downloaded from `https://github.com/pytorch/ fairseq`, commit `f6ac1aecb3329d2cbf3f1f17106b74 ac51971e8a`.