Wat heb je gezegd? Detecting Out-of-Distribution Translations with Variational Transformers

Tim Z. Xiao
OATML
University of Oxford
tim.z.xiao@outlook.com

Aidan N. Gomez
OATML
University of Oxford
aidan.gomez@cs.ox.ac.uk

Yarin Gal
OATML
University of Oxford
yarin@cs.ox.ac.uk

1 Introduction

We use epistemic uncertainty to detect out-of-training-distribution sentences in Neural Machine Translation. For this, we develop a measure of uncertainty designed specifically for long sequences of discrete random variables, corresponding to the words in the output sentence. This measure is able to convey epistemic uncertainty akin to the Mutual Information (MI), which is used in the case of single discrete random variables such as in classification.

Our new measure of uncertainty solves a major intractability in the naive application of existing approaches on long sentences. We train a Transformer model with dropout on the task of German-English translation using WMT 13 and Europarl, and show that using dropout uncertainty our measure is able to identify when Dutch source sentences, sentences which use the same word types as German, are given to the model instead of German.

2 Related Work

A significant amount of attention has been given to the calibration and uncertainty of neural networks in the context of vision tasks, whereas natural language tasks have been left comparatively neglected. In this work, we explore the calibration properties of Transformers (Vaswani et al., 2017) – a popular architecture for natural language tasks – and identify a major failure model of Transformer calibration.

There have been some attempts at investigating calibration and uncertainty in neural machine translation (NMT) models. In particular, Kumar & Sarawagi (2019) have investigated the calibration in various NMT models at the token level. Kumar & Sarawagi (2019) found that most models are indeed ill-calibrated at the token level, leading to the resulting probability distribution over the vocabulary used during decoding is not a good reference for model uncertainty. To correct for this, Kumar & Sarawagi designed a parametric model based on input coverage, attention uncertainty and token probability for recalibration. Another study of uncertainty in NMT models comes from Ott et al. (2018); they found that the model has a highly uncertain output distribution in the way that probability mass at the sequence level spread widely over the hypothesis space.

The existing work in this area seems to miss one of the key values a well-calibrated model provides: *out-of-distribution detection*. Our work explores this direction and offers a new uncertainty estimation technique that empirically out-performs existing methods by a significant margin.

3 Methods

Principally, we care about measuring the variance of a model's outputs around some input point. In the context of a simple classifier model, the solution is often found by measuring the mutual information between the predicted discrete distribution and model parameters, evaluating the output's entropy, or simply computing the variance of model outputs (Gal, 2016). In the domain of language, however, there are many semantically equivalent alternatives to the same prediction, and it is a difficult matter to measure the disagreement between the predicted discrete sequences. Much worse, when attempting to naively use MI or entropy with long sequences or large sets of discrete random

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

variables, we quickly discover that even approximate integration over the product space becomes prohibitive (Kirsch et al., 2019).

In order to capture epistemic uncertainty in the task of NMT, we investigate several measures of uncertainty appropriate for long sequences of discrete variables:

1. **Beam Score:** we assign a confidence to output y generated (using beam search) from input x using the score assigned to y's beam (Wu et al., 2016).

$$BS = \log(p(y|x)) \ / \ \text{length\_penalty}(y; 0.6) \tag{1}$$

$$\text{length\_penalty}(y; \alpha) = \left(\frac{5 + |y|}{5 + 1}\right)^{\alpha}$$

2. **Sequence Probability:** we assign a confidence to output y generated from input x by taking the log predictive probability under the weight distribution.

$$SP = \log(\mathbb{E}_{\theta \sim q(\theta)} \, p_\theta(y|x)) \ / \ \text{length\_penalty}(y; 0.6) \tag{2}$$

3. **BLEU Variance:** we assign uncertainty at an input x by producing pairs of outputs from the model and measuring the squared complement of the BLEU (Papineni et al., 2002) to judge disagreement between model outputs on input x.

$$\text{BLEUVar} = \mathbb{E}_{\theta \sim q(\theta)} \mathbb{E}_{y, y' \sim p_\theta(y|x)} \left(1 - \text{BLEU}(y, y')\right)^2 \tag{3}$$

For the *beam score* we use the deterministic model found by gradient descent and simply take the probabilities from under its predictive distribution. For *sequence probability* and *BLEU variance* we use MC Dropout (Gal, 2016) and take a number of samples (N) to estimate the expectations[1]:

$$SP \approx \log\left(\sum_{i=1}^{N} p_{\theta_i}(y|x)\right) \ / \ \text{length\_penalty}(y; 0.6)$$

For the BLEUVar approximation, we opt for decoding outputs using beam search applied to different model samples and measuring the complement BLEU between pairs of these examples.

$$\text{BLEUVar} \approx \sum_{i=1}^{N} \sum_{j \neq i}^{N} \left(1 - \text{BLEU}(\text{decode}_{\theta_i}(x), \text{decode}_{\theta_j}(x))\right)^2 .$$

3.1 Evaluating uncertainty in sequence models

A number of uncertainty evaluation metrics have been proposed for standard classifier networks such as the popular ECE and MSE metrics (Guo et al., 2017); however, for sequence modelling these classification-specific metrics are not applicable. Instead we opt for the *performance versus retention* curve method for evaluating our uncertainty measures following Filos et al. (2019).

The performance-retention curve indicates how well an uncertainty measure would perform if the $k\%$ least certain outputs were deleted from the test dataset. The x-axis ranges along the fraction of data retained, while the y-axis measures some performance metric of the model on the retained data.

A performance-retention curve of a well-calibrated uncertainty measure will see a clear and sustained improvement in performance as low-confidence predictions are excluded from the test set; while a poorly-calibrated model will yield a curve that either lacks a trend or oscillates unstably.

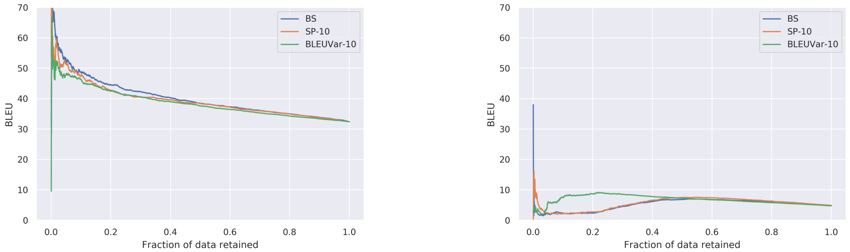
4 Detecting out-of-training-distribution sentences in NMT

We trained a Transformer model on WMT 13 and Europarl DE (German) to EN (English) sentence pairs (obtaining BLEU 33 on the WMT 14 test set). We then evaluated the model on out-of-training-distribution input sentences in NL (Dutch), which shares a large overlapping vocabulary with German

[1] The approximations below should have constant scaling factors, but these don't impact our evaluation metric (performance-retention curves) so we leave them out for simplicity.

(hence input sentences look plausible to non-native speakers). One would hope that such data falling outside of the training distribution would produce model predictions with high uncertainty.

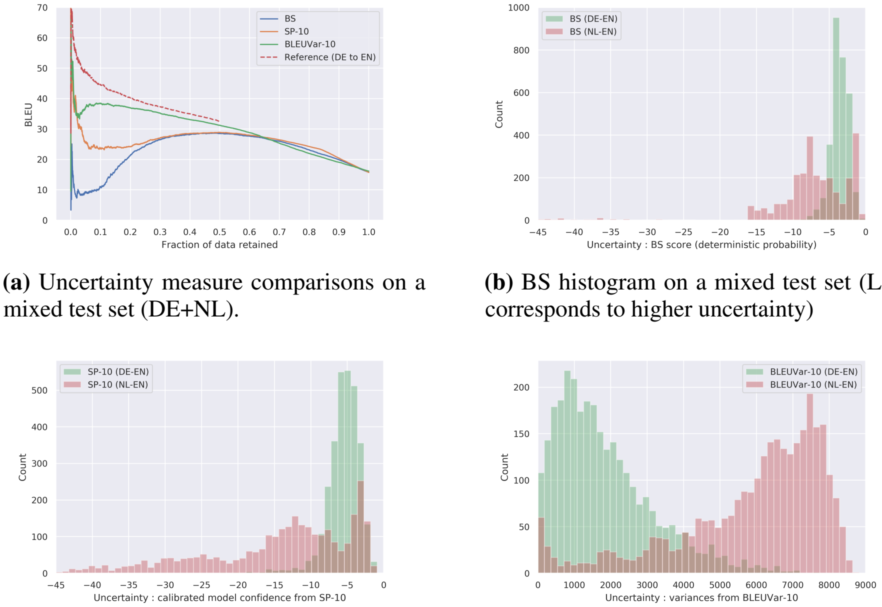
Evaluating the different measures of uncertainty, we find empirically (see Figures 1 and 2) that the deterministic Transformer baseline tends to place remarkably high confidence on its outputs for a previously unseen language (while performing very poorly in terms of BLEU). We find that taking multiple samples of the model and evaluating the averaged probabilities – that is, using equation 2 – results in markedly improved calibration. And even more drastic improvement is found by using our proposed $\mathrm{BLEUVar}$ uncertainty estimation technique (equation 3); $\mathrm{BLEUVar}$ produces a substantial improvement in the separation between in-distribution and out-of-distribution sentences (Fig. 2d).



(a) DE to EN (size 3k)

(b) NL to EN (size 3k)

Figure 1: Uncertainty measure comparisons using the in-distribution DE-EN test set (left) and out-of-distribution NL-EN test set (right). The Transformer model was trained for DE to EN tasks with the full EN-DE training set (size 4.6m) using 250k steps. We denote the results from methods *beam score*, *sequence probability* and *BLEU Variance* as *BS*, *SP* and *BLEUVar* correspondingly. If a suffix is added such as *BLEUVar-10*, then the suffix *10* represents the number of results sampled during MC dropout. Note the poor performance of the model on the out-of-distribution sentences.



(a) Uncertainty measure comparisons on a mixed test set (DE+NL).

(b) BS histogram on a mixed test set (Left corresponds to higher uncertainty)

(c) SP-10 histogram on a mixed test set (Left corresponds to higher uncertainty)

(d) BLEUVar-10 histogram on a mixed test set (Right corresponds to higher uncertainty)

Figure 2: Comparison and histograms for different uncertainty measures on a combined DE+NL test set (6k size). Histograms show uncertainty value for DE-EN (blue) and NL-EN (red). Note how BLEUVar-10 is able to clearly separate the in-distribution (blue) from out-of-distribution (red).

References

Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. Benchmarking bayesian deep learning with diabetic retinopathy diagnosis. `https://github.com/OATML/bdl-benchmarks`, 2019.

Yarin Gal. *Uncertainty in deep learning*. PhD thesis, PhD thesis, University of Cambridge, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *NIPS*, 2019.

Aviral Kumar and Sunita Sarawagi. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*, 2019.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.