Should We Trust This Summary? Bayesian Abstractive Summarization to The Rescue

Alexios Gidiotis
Aristotle University of Thessaloniki
gidiotis@csd.auth.gr

Grigorios Tsoumakas
Aristotle University of Thessaloniki
greg@csd.auth.gr

Abstract

We explore the notion of uncertainty in the context of modern abstractive summarization models, using the tools of Bayesian Deep Learning. Our approach approximates Bayesian inference by first extending state-of-the-art summarization models with Monte Carlo dropout and then using them to perform multiple stochastic forward passes. Based on Bayesian inference we are able to effectively quantify uncertainty at prediction time. Having a reliable uncertainty measure, we can improve the experience of the end user by filtering out generated summaries of high uncertainty. Furthermore, uncertainty estimation could be used as a criterion for selecting samples for annotation, and can be paired nicely with active learning and human-in-the-loop approaches. Finally, Bayesian inference enables us to find a Bayesian summary which performs better than a deterministic one and is more robust to uncertainty. In practice, we show that our Variational Bayesian equivalents of BART and PEGASUS can outperform their deterministic counterparts on multiple benchmark datasets.

1 Introduction

State-of-the-art text summarization methods have achieved remarkable performance in various benchmarks (Song et al., 2019; Dong et al., 2019; Lewis et al., 2019; Zhang et al., 2020). The majority of these methods use very large Transformer models pre-trained on language generation tasks.

Although such methods can generate high quality summaries for texts similar to their training set, they suffer from a couple of issues when the inputs lie far from the training data distribution. They are prone to generating particularly bad outputs (Xu et al., 2020; Kryściński et al., 2020) and are usually fairly confident about them (Gal and Ghahramani, 2016; Xiao et al., 2020). These shortcomings are bound to cause problems once a summarization model is deployed to solve a practical problem.

Since the output of automatic summarization models is usually expected to be consumed by humans, it is very important to know when such an output is of good enough quality to be served to users. In most cases, it is very much preferable to not serve an output at all, instead of serving a bad output. This will in turn increase users' trust to automated summarization systems.

Model uncertainty is one way of detecting when a model's output is likely to be poor on the grounds of predicting far away from it's training distribution. Recent summarization methods have focused heavily on improving the overall performance, but model uncertainty has been explored very little (Xu et al., 2020).

In addition to improving user experience, the development of uncertainty measures for summarization can pave the way for active learning approaches (Gal et al., 2017; Houlsby et al., 2011; Liu et al., 2020; Lyu et al., 2020). The value of active learning stems from the fact that obtaining labeled samples for training is hard, but it is relatively easy to obtain large amounts of unlabeled samples. Summarization is no different in this perspective, since creating good quality target summaries for training can be very costly.

This work explores uncertainty estimation for state-of-the-art text summarization models, from a Bayesian perspective. We extend the BART (Lewis et al., 2019) and PEGASUS (Zhang et al., 2020) summarization models with Monte Carlo dropout (Gal and Ghahramani, 2016), in order to create corresponding Variational Bayesian PEGASUS (VarPEGASUS) and BART (VarBART) models. Sampling multiple summaries from those models allows us to approximate Bayesian inference in a practical way, which in turn enables us to estimate summarization uncertainty. To the best of our knowledge this is the first attempt to apply Bayesian summary generation with large Transformer models.

Findings of the Association for Computational Linguistics: ACL 2022, pages 4119 - 4131

Based on Bayesian approximation, we adapt the Monte Carlo BLEU variance metric (Xiao et al., 2020) to the summarization task, and investigate its efficacy as a measure of summarization uncertainty. Our findings suggest that this uncertainty metric correlates well with the quality of the generated summaries and can be effective at identifying cases of questionable quality.

Finally, we take the summarization uncertainty study one step further, and select the summary with the lowest disagreement out of multiple summaries sampled from our Variational models. Experiments across multiple benchmark datasets show that this method consistently improves summarization performance (see Table 5), and by using it our VarPEGASUS and VarBART models achieve better ROUGE F-scores compared to their original deterministic counterparts.

The rest of this paper is structured as follows. Section 2 discusses related work on Bayesian deep learning and uncertainty estimation methods. Section 3 presents our approach. Section 4 describes our experimental setup, while Section 5 presents and discusses the results. Finally, Section 6 concludes our work and considers its broader impact.

2 Related work

Uncertainty estimation in deep learning is a topic that has been studied extensively. Bayesian deep learning includes a family of methods that attempt to capture the notion of uncertainty in deep neural networks. Such methods have gained increased popularity in the deep learning literature and there exist multiple applications in subfields such as Computer Vision (Kendall and Gal, 2017; Litjens et al., 2017; Gal et al., 2017) and Natural Language Processing (NLP) (Siddhant and Lipton, 2020; Liu et al., 2020; Lyu et al., 2020; Xiao et al., 2020).

Despite their obvious advantage of modeling uncertainty, the main problem with Bayesian deep learning methods is the computational cost of full Bayesian inference. To tackle this problem, Gal and Ghahramani (2016) propose using standard dropout (Srivastava et al., 2014) as a practical approximation of Bayesian inference in deep neural networks and call this method Monte Carlo dropout. Gal et al. (2017) use a convolutional neural network with Monte Carlo dropout in order to obtain an uncertainty estimate for active learning in the task of image classification. Houlsby et al. (2011) sample many networks with Monte Carlo simulation and propose an objective function that takes into account the disagreement and confidence of the predictions coming from these networks.

Similar methods have also been applied to NLP. In machine translation, Xiao et al. (2020) extend the Transformer architecture with MC dropout to get a Variational Transformer, and use it to sample multiple translations from the approximate posterior distribution. They also introduce BLEUVar, an uncertainty metric based on the BLEU score (Papineni et al., 2002) between pairs of the generated translations. Lyu et al. (2020) extend the work of Xiao et al. (2020) to question answering and propose an active learning approach based on a modified BLEUVar version. Similarly, Liu et al. (2020) use a conditional random field to obtain uncertainty estimates for active learning and apply their method to named entity recognition.

Although summarization is a prominent NLP task, summarization uncertainty has not been widely studied. Xu et al. (2020) is the only work that focuses on uncertainty for summarization, but their work does not make use of Bayesian methods. They define a generated summary's uncertainty based on the entropy of each token generated by the model during the decoding phase. Their study includes experiments on CNN/DM and XSum using the PEGASUS and BART summarization models. Their main focus is on understanding different properties of uncertainty during the decoding phase, and their work is not directly comparable to ours.

3 Methods

We first introduce Bayesian inference, in the context of deep neural networks and show how it can be used to measure uncertainty. Subsequently, we show how Bayesian inference can be applied to summarization in order to estimate the uncertainty of a summary generated for a given input. Finally, we show how Bayesian inference can be employed for producing better summaries.

3.1 Monte Carlo dropout

Contrary to standard neural networks, Bayesian probabilistic models capture the uncertainty notion explicitly. The goal of such models is to derive the entire *posterior* distribution of model parameters θ given training data X and Y (Equation 1).

$$P(\theta|X,Y) = \frac{P(Y|X,\theta)P(\theta)}{P(Y|X)} \quad (1)$$

At test time, given some input x, a prediction $\hat{y}$ can be made by integrating over all possible θ values (Equation 2). The predictive distribution's variance can then be used as a measure of the model's uncertainty.

$$P(\hat{y}|x, X, Y) = \int P(\hat{y}|x, \theta)P(\theta|X, Y)d\theta \quad (2)$$

In practice, integrating over all possible parameter values for a deep neural network is intractable, and therefore Variational methods are used to approximate Bayesian inference. A neural network trained with dropout can be interpreted as a Variational Bayesian neural network (Gal and Ghahramani, 2016), and as a result making stochastic forward passes with dropout turned on at test time is equivalent to drawing from the model's predictive distribution. This Monte Carlo (MC) dropout method can be easily applied to any neural network that has been trained with dropout.

3.2 Summary uncertainty

MC dropout is a simple yet effective method that requires no adjustment to the underlying model. It is possible to convert any state-of-the-art summarization model to a *Variational Bayesian model*, with the use of MC dropout. For Transformer based models in particular, the Transformer blocks that make up the encoder and decoder are usually trained with dropout, and therefore the conversion is trivial by simply turning dropout on at test time.

In Variational models, the variance of the predictive distribution can be used to measure the model's uncertainty. For a text summarization model, we can approximate the variance of this distribution, by measuring the dissimilarity of N stochastic summaries $y_1, y_2 \dots y_N$, generated with MC dropout.

The BLEU metric (Papineni et al., 2002) is commonly used for measuring the similarity between a pair of texts. As in Xiao et al. (2020), we approximate the model's predictive variance with the BLEU Variance (BLEUVar) metric over the N summaries generated with MC dropout as shown in Equation 3. BLEUVar is computed by summing the squared complement of BLEU among all pairs of summaries (twice as BLEU is asymmetric) generated for the same input with different dropout masks.

$$\text{BLEUVar} = \sum_{i=1}^{N} \sum_{j \neq i}^{N} (1 - \text{BLEU}(y_i, y_j))^2 \quad (3)$$

Because we sum over all pairs of N samples twice, scores that are computed with different N values are not directly comparable. To alleviate this issue we propose a normalized version of the metric, BLEUVarN, where we divide BLEUVar by $N(N-1)$ (Equation 4). This allows for comparisons between scores computed with different N values.

$$\text{BLEUVarN} = \frac{\sum_{i=1}^{N} \sum_{j \neq i}^{N} (1 - \text{BLEU}(y_i, y_j))^2}{N(N-1)} \quad (4)$$

By running multiple stochastic forward passes for the same input, we essentially create an ensemble of models with different parameters. Making predictions with this ensemble has the following effects. For inputs close to the learned distribution the summaries generated by all models in the ensemble will be similar to one another, and as a result BLEUVarN will be low. On the other hand, for inputs lying away from the learned distribution, the generated summaries will differ wildly and BLEUVarN will be high, indicating high uncertainty.

3.3 Bayesian summary generation

Inspired by the fact that making multiple predictions with MC dropout is equivalent to ensembling multiple stochastic models, we propose a novel Bayesian approach to summary generation. Instead of generating a single deterministic summary without dropout, as is commonly the case with modern summarization approaches, we consider using the *predictive mean* of multiple predictions made with MC dropout. Because the predictions in our case are summaries their predictive mean is not well defined, so instead we opt for selecting one of the N summaries.

We assume that the *predictive mean* of the N summaries generated with MC dropout should be the one having the lowest *disagreement* with the rest of the $N-1$ summaries. Since the pairwise complement of BLEU between all pairs of the sampled summaries has already been computed when estimating BLEUVarN uncertainty, it can be further used to help us find the lowest disagreement summary. In practice, we select the summary $\hat{\mu}$ that maximizes the sum of symmetric BLEU similarity with the rest of the summaries (Equation 5) (Xiao et al., 2020). This summary could be seen as the *median* of all the summaries generated with MC dropout, although this is not a mathematically

correct expression.

$$\hat{\mu} = \underset{y_i}{\operatorname{argmax}} \sum_{j \neq i}^{N} [\text{BLEU}(y_i, y_j) + \text{BLEU}(y_j, y_i)] \quad (5)$$

The intuition behind this approach is based on the following assumption. We expect the *median* summary to integrate the key concepts that all individual summaries agree on. Consequently, for inputs close to the model's learned distribution, the individual summaries will be similar to one another and as a result the *median summary* will be the best choice. On the other hand, for out-of-distribution inputs, the *median* out of a number of very different summaries will result in a more robust and overall better final summary. In practice, even for well trained models, we expect to have a fairly large number of inputs that are not close to the models' learned distribution, and therefore we expect to benefit from the positive effects of ensembling multiple outputs.

4 Experimental Setup

We first present the three datasets that are involved in our experiments, their main statistics and the reasons for including them in our empirical study. Then we present the two summarization models that we employed, along with their parameters and details on stochastic summary generation.

4.1 Datasets

In order to verify the effectiveness of our Bayesian abstractive summarization approach, we conducted a series of experiments on three well-known summarization benchmarks:

- **XSum** (Narayan et al., 2018) is a dataset of 227k BBC articles on a wide variety of topics. Each article is accompanied by a human written, single-sentence summary.

- **CNN/DailyMail** (Hermann et al., 2015) is a dataset containing a total of 93k articles from the CNN, and 220k articles from the Daily Mail newspapers. All articles are paired with bullet point summaries. The version used is the non-anonymized variant similar to (See et al., 2017).

- **AESLC** (Zhang and Tetreault, 2020) is a dataset of 18k emails from the Enron corpus (Klimt and Yang, 2004). The body of each email is used as source text and the subject as summary.

Table 1: Basic statistics for each one the datasets used in our experiments. The document and summary length are measured in words.

	Size		Length	
Dataset	**Val.**	**Test**	**Doc.**	**Sum.**
XSum	11,332	11,334	431	23
CNN/DM	13,368	11,490	760	46
AESLC	1,960	1,906	75	4

The main criteria for selecting these datasets are the availability of recent, open source models trained on them and their relatively short texts that would allow us to run a number of different experiments quickly. Since our methods do not involve training, we will only focus on the validation and test set of each dataset. All datasets are obtained from the Hugging Face datasets repository[1]. Table 1 presents some basic statistics for these datasets.

4.2 Models

BART (Lewis et al., 2019) and PEGASUS (Zhang et al., 2020) are Transformer based sequence-to-sequence models, pre-trained on massive corpora of unsupervised data (Web and news articles). Since our experiments do not involve training, we utilize open-source models fine-tuned on the training set of each dataset. These models can be found in the Hugging Face models repository[2].

Our BART models follow the $\text{BART}_{\text{LARGE}}$ architecture with 12 Transformer blocks for the encoder and the decoder. BART is pre-trained as a denoising autoencoder, where the text is corrupted and the model learns to reconstruct the original text. Open-source fine-tuned BART models are only available for XSum and CNN/DM. Our PEGASUS models follow the $\text{PEGASUS}_{\text{LARGE}}$ architecture and have 16 Transformer blocks for the encoder and the decoder. PEGASUS is pre-trained on the C4 and HugeNews datasets, on a sentence infilling task. Open-source fine-tuned PEGASUS models exist for all three datasets considered in our experiments.

In order to convert BART and PEGASUS to Variational models, we enable dropout for all Transformer blocks of the encoder and decoder. For each sample, we generate N summaries using beam

[1] https://huggingface.co/datasets
[2] https://huggingface.co/models

search decoding with 8 beams. We experimented with N equal to 10 (MC-10) and 20 (MC-20). The rest of the hyper-parameters used were identical to the original papers.

5 Results

Our main experiment evaluates BLEUVarN's effectiveness in quantifying uncertainty for summarization models. A second experiment investigates the potential of the Bayesian summarization method proposed in Section 3.3 as a way of improving summarization performance at test time.

5.1 Evaluating Bayesian uncertainty

We here evaluate the effectiveness of BLEUVarN in measuring the model's uncertainty. The performance versus data retention curve (Filos et al., 2019) measures how well a model would perform if we completely removed the $k\%$ most uncertain outputs from the test set. In the x-axis we have the fraction of data from the test set that are removed, while in the y-axis we have the performance metrics. An effective uncertainty measure should show a consistent improvement in performance as we discard more samples based on high uncertainty. In this experiment, we arrange samples by decreasing BLEUVarN score and gradually remove the samples with the highest score.

Figure 1 shows, for each dataset, the performance of our Variational models in terms of ROUGE-1 F-score versus the fraction of data discarded based on BLEUVarN. ROUGE-2 and ROUGE-L F-scores follow similar patterns and can be found in the Appendix A. For reference, we are also plotting the performance of the deterministic models using all data as straight lines. Also, in Table 2 we quantify the percentage increase in ROUGE F-scores as we discard different fractions of the full test datasets based on BLEUVarN.

All ROUGE F-scores improve as we gradually discard samples with high BLEUVarN, an observation that is consistent across all test datasets and models. More specifically, we notice that the increase is linear for the first 80% of the data, but then becomes almost exponential. From these observations we can draw two conclusions. First, models indeed perform significantly worse on samples with high uncertainty. Second, BLEUVarN is effective at quantifying uncertainty and can be used to identify high uncertainty samples.

Furthermore, we notice that the performance increase is significantly higher in the XSum and AESLC datasets compared to CNN/DM. In particular, VarPEGASUS-20 shows a staggering 58 point increase in ROUGE-1 score. We think that this difference might be related to the more extractive nature of CNN/DM summaries as opposed to the other two datasets. Such a finding would mean that Bayesian uncertainty filtering is more beneficial in abstractive rather than extractive setups.

Table 2: Percentage increase in ROUGE F-scores when discarding 25%, 50% and 75% of the data based on the highest BLEUVarN.

	25%	50%	75%
Model	**R-1/R-2/R-L**	**R-1/R-2/R-L**	**R-1/R-2/R-L**
		XSum	
VarBart-10	6.4/13.8/8.3	12.2/25.2/15.4	22.1/41.9/26.1
VarBart-20	6.5/14.1/8.3	13.2/26.7/16.5	22.5/42.6/27.2
VarPegasus-10	7.5/15.8/9.6	14.9/29.4/18.6	25.2/4.9/29.9
VarPegasus-20	8.0/16.8/10.3	15.8/31.2/19.6	26.3/48.1/31.2
		CNN/DM	
VarBart-10	2.9/7.2/4.8	5.4/13.1/8.5	8.8/20.4/13.3
VarBart-20	3.2/7.8/5.1	5.3/12.8/8.5	8.3/19.4/12.6
VarPegasus-10	4.1/9.9/6.1	7.8/17.4/10.9	12.6/26.1/16.8
VarPegasus-20	4.6/10.7/6.8	8.5/19.0/11.9	14.7/29.6/18.7
		AESLC	
VarPegasus-10	17.5/33.5/17.7	30.6/51.9/31.1	54.4/75.0/54.7
VarPegasus-20	18.7/36.3/18.9	36.0/59.7/36.6	58.4/78.0/58.8

To further illustrate how BLEUVarN behaves across different parts of the data, Figure 2 shows the decrease in the average BLEUVarN of all Variational models as we gradually discard samples with low ROUGE-1 scores from each dataset. We observe that for the samples with the highest ROUGE performance BLEUVarN becomes almost zero. This observation further supports our argument that model uncertainty has a significant impact on model performance.

5.1.1 MC-10 vs MC-20

From Figure 1 we can see that MC dropout with 20 samples performs better than 10 samples, resulting in higher performance. In more detail, for highly uncertain data, both MC-10 and MC-20 converge to similar BLEUVarN values (Figure 2) as well as ROUGE scores (Figure 1). On the other side of the spectrum, for low uncertainty data, using 20 samples leads to bigger performance increase along with a little higher BLEUVarN scores.

Based on these observations, we conclude that MC dropout with 20 samples is generally better in terms of performance. This comes at the cost of increased computational overhead because it requires running twice as many stochastic passes with

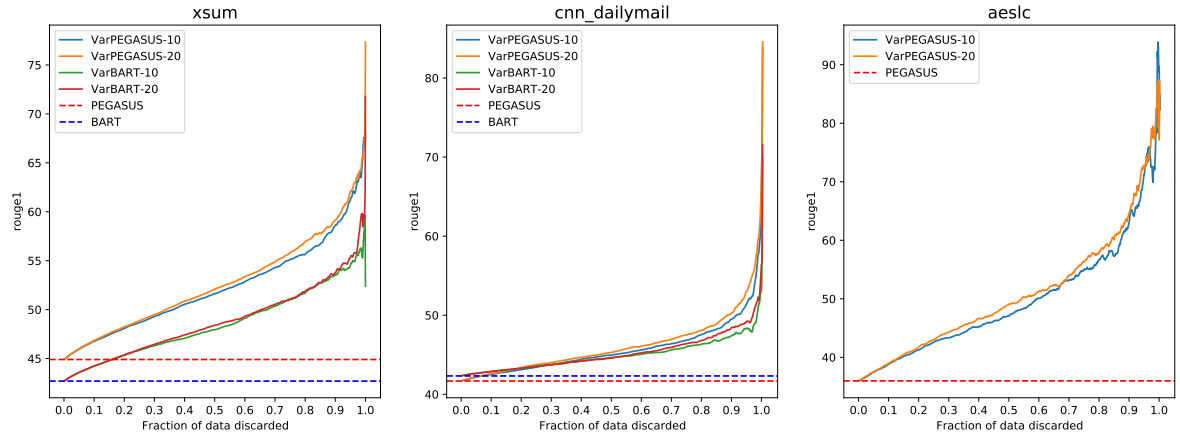

Figure 1: ROUGE-1 scores vs fraction of data discarded due to high BLEUVarN. The straight dashed lines indicate the performance level of the deterministic PEGASUS and BART models.

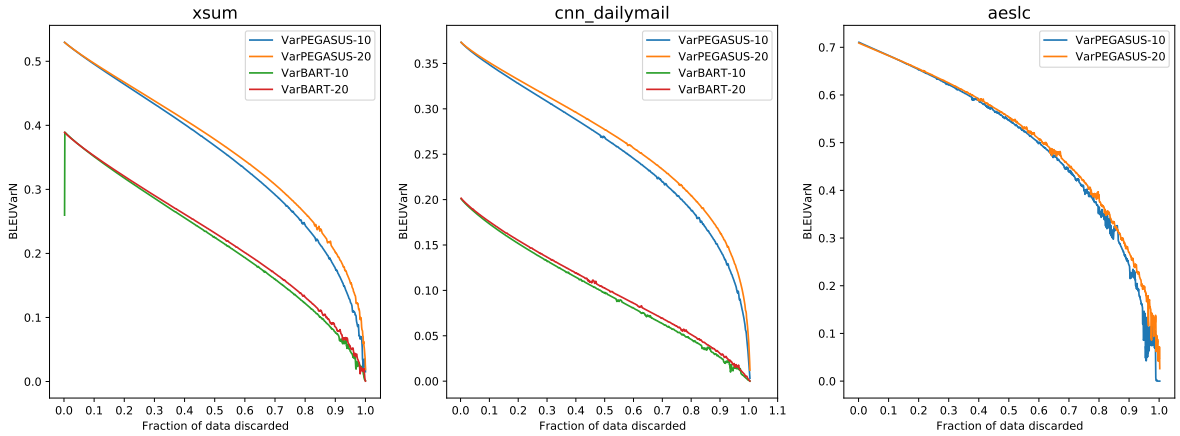

Figure 2: BLEUVarN curves as a function of data discarded due to low ROUGE-1 scores.

MC dropout. However, this computation is embarrassingly parallelizable in modern hardware, and can be easily optimized by batching MC dropout generations with different dropout masks for each sample within the batch.

Although we have shown that MC dropout with 20 samples performs better than 10 samples, we observed that further increasing this number, for example to 30 or 50 samples, was beginning to bring diminishing returns. Furthermore, the performance increase we got by using 10 and 20 samples was substantial enough while the runtimes for MC dropout with more samples were becoming a lot longer. For these reasons we refrained from increasing it even further in order to keep computational capacity manageable.

5.1.2 VarBART vs VarPEGASUS

Out of the two models, VarPEGASUS is consistently showing the biggest increase in performance as more uncertain samples are dropped from the dataset. It should be noted here, that the decline in performance as data uncertainty increases, is much steeper for VarBART than it is for VarPEGASUS on both the XSum and the CNN/DM dataset. This coincides with the fact that VarPEGASUS also has much higher BLEUVarN uncertainty as shown in Figure 2, which hints us that the PEGASUS model is in general less confident about the outputs it generates. Anecdotally, we can say here that PEGASUS is more aware of the things it does not know, and as a result it seems to benefit more from the uncertainty estimates.

5.2 Bayesian vs deterministic summarization

The next experiment focuses on the Bayesian summarization method proposed in Section 3.3. We compare the performance of Bayesian summarization using the VarBART and VarPEGASUS models against the standard summarization paradigm using the deterministic BART and PEGASUS models. Our goal is to verify the efficacy of Bayesian summarization as a post-hoc way of improving summarization performance.

Table 3 reports the ROUGE-1, ROUGE-2 and ROUGE-L F-scores of our VarBART and VarPEGASUS models along with the deterministic BART and PEGASUS models on all benchmark datasets, re-evaluated for consistency. The results show that Bayesian summarization is effective, with both VarBART and VarPEGASUS outperforming their deterministic counterparts on all datasets. Furthermore, increasing the number, N, of samples generated during the Bayesian inference, improves performance for all datasets except for AESLC, at the cost of increased computational complexity as discussed in Section 5.1.

Note that our goal in this work was not to compete with other state-of-the-art models. What we want to show is that relying on the agreement between multiple Bayesian summaries for the same input, is an effective way to boost the summarization performance of deterministic models. Also, this is a post-hoc method and does not involve training new models, which makes it easily applicable to many different scenarios.

Figure 3 plots the difference in ROUGE-1 of each Variational model with its deterministic counterpart versus the fraction of the data discarded due to high uncertainty. Similar plots for ROUGE-2 and ROUGE-L can be found in Appendix A. Positive values indicate that the Variational model achieves a higher score than the deterministic one. These plots give us a better view of how the Variational models fare against the deterministic ones for different levels of uncertainty. As far as we know, this is the first study to directly compare Variational and deterministic models on data with varying levels of uncertainty.

Looking at the curves, we clearly see that the differences are positive for most uncertainty levels but start decreasing as more data with high uncertainty are discarded. For the top $10\% - 20\%$ most certain samples we start seeing a fluctuation between positive and negative values. This pattern is in line with the observations made in Figure 1, and leads us to believe that there is a significant gap between the deterministic model's performance on the 20% most certain samples and the rest of the data.

These observations lead us to the following conclusions. For samples of very low uncertainty, we can expect both Variational and deterministic models to converge to equally good outputs. In contrast, as uncertainty becomes higher, we observe a clear advantage of the Variational summaries over the deterministic ones. This pattern is consistent across all models and datasets, and underpins our case that Bayesian summarization is beneficial for the majority of inputs.

5.3 Qualitative analysis

In order to better illustrate our findings in this work, we present a couple of real examples from VarPEGASUS-10 on XSum. For each example, we show the 10 sample summaries generated with MC dropout for the same input, as well as the corresponding BLEUVarN score. We have highlighted the median summary in bold typeface and for the sake of comparison we also show the summary generated by the deterministic PEGASUS model.

The first example (Table 4) is a case of high uncertainty from the XSum dataset. We can see that all 10 samples are considerably different from one another, which leads to a high BLEUVarN score. In contrast, the second example (Table 5) has much lower uncertainty. In this case all 10 samples seem to mostly agree on the main points and as a result BLEUVarN is fairly low. Here, the median summary is the one that represents better this agreement.

Looking at the ROUGE-1 score for both examples we can see there's a rather drastic difference as well. For the sample in Table 4 we can see that neither the deterministic nor the Bayesian summary show a strong performance, yet even in that case the median Bayesian summary scores a bit higher. On the other hand, the sample in Table 5 showcases a solid performance from both the deterministic and the Bayesian summary. Here it is evident that the median Bayesian summary is close but slightly better than the deterministic summary in terms of ROUGE.

6 Conclusions and future work

This work explored Bayesian methods in the context of text summarization. We extended state-of-

Table 3: A comparison of our VarBART and VarPEGASUS models against the deterministic BART and PEGASUS.

	XSum			CNN/DM			AESLC		
Model	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
BART	42.69	20.66	35.29	42.32	20.28	36.21	-	-	-
VarBART-10	42.97	20.86	35.56	42.65	20.64	36.56	-	-	-
VarBART-20	**43.07**	**20.97**	**35.68**	**42.76**	**20.76**	**36.69**	-	-	-
PEGASUS	44.90	23.33	37.74	41.68	20.24	36.17	35.97	20.28	35.09
VarPEGASUS-10	44.93	23.54	38.01	42.04	20.75	36.76	36.36	**21.40**	**35.58**
VarPEGASUS-20	**45.32**	**23.87**	**38.29**	**42.25**	**20.98**	**36.94**	**36.41**	21.00	35.53

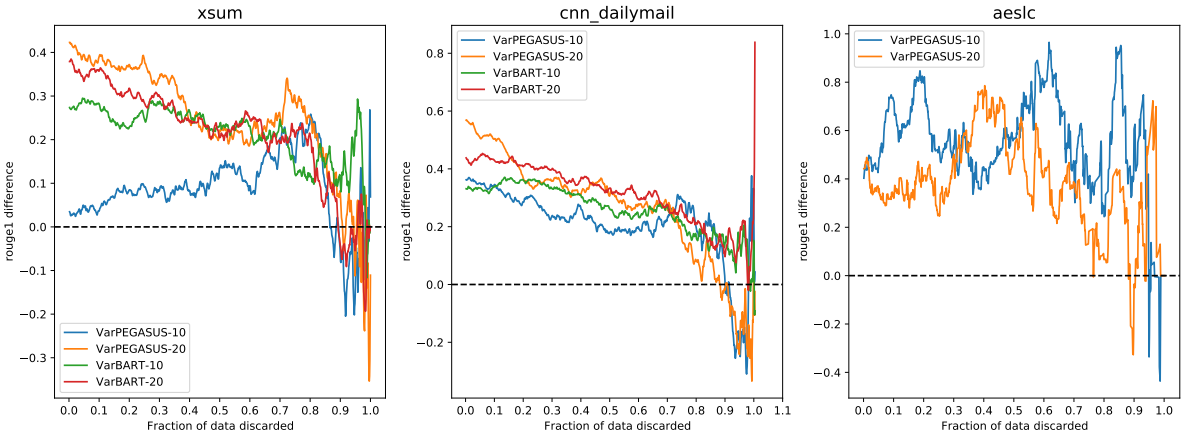

Figure 3: Difference in ROUGE-1 between Variational models and their deterministic counterparts versus the fraction of data discarded. Positive values indicate that deterministic ROUGE-1 is lower than Variational.

the-art summarization models with MC dropout to approximate Bayesian inference, and demonstrated how BLEUVarN can be used to quantify model uncertainty. This allows us to effectively identify high uncertainty summaries at prediction time, which can be a significant advantage.

Furthermore, we show that ensembling multiple stochastic summaries generated by Variational Bayesian models can lead to improved performance compared to similar deterministic models. This finding is verified by experiments for two different models and across 3 benchmark datasets.

It should be noted here that the proposed methods are directly applicable to other abstractive summarization datasets as well. We acknowledge that some of the more interesting summarization problems involve much longer summaries, for example scientific abstracts. In this work we focused on datasets of short summaries in order to be more resource efficient and conduct more experiments. There's a lot of interesting work that focuses on the summarization of longer documents (Gidiotis and Tsoumakas, 2020; Zaheer et al., 2020) that could potentially be applied in combination with the methods we propose here.

Our work can have a broader impact in several ways. To the research community, being the first work to study Bayesian uncertainty for abstractive summarization and paving the way for other similar methods. To the industry, because it improves automatic summarization systems and can be paired nicely with active learning and human-in-the-loop approaches. Finally, to the end users, improving their experience and building up confidence towards automatic summarization systems.

References

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13042–13054. Curran Associates, Inc.

Table 4: High uncertainty example from the XSum dataset. Sample summary (1), in bold typeface, is the *median summary* according to our approach.

Bayesian samples:

1. **When John Choe launched his first hotel in Singapore, he had no idea what he was getting himself into. (R1: 22.22)**
2. When Singapore's Frasers Centre hired him as its first managing director, he was told it would take him five years to get off the ground.
3. In his early 20s, when he was working as a waiter in a luxury hotel in Hong Kong, David Choe always dreamed of running his own business.
4. "When I was a teenager, I used to say to myself 'I want to start my own company'."
5. When John Choe was appointed chief executive of a Singapore-based property firm in the early 1990s, he said he wanted to "make a difference to people's lives".
6. When David Choe was asked if he would ever run a hotel company, he thought it would be a good idea.
7. "I always wanted to be a hotelier," says Fraser Choe.
8. As a young entrepreneur with no experience in hospitality, John Choe had no idea what he was about to achieve.
9.).
10. "When I started the company, I said 'let's see what we can do, let's see what we can achieve, let's see what we can achieve'."

Deterministic summary: When Choe Swee Swee was appointed chief executive of one of Singapore's biggest property firms, he told the BBC he wanted to "make the world a better place". (R1: 21.81)

Target summary: On the first day in his new job, Choe Peng Sum was given a fairly simple brief: "Just go make us a lot of money."

BLEU variance: 0.96

Table 5: Low uncertainty example from XSum. Sample summary (7), in bold typeface, is the *median summary* selected according to our approach. In the parentheses we show the ROUGE-1 score for the median Bayesian summary and the deterministic summary.

Bayesian samples:

1. Torquay United have signed Torquay United have signed Myles Keating.
2. Torquay United have signed defenders Myles Anderson and Ruairi Keating.
3. National League side Torquay United have signed defender Lewis Anderson and striker Ruairi Keating.
4. Torquay United have signed defender Liam Anderson on a deal until the end of the season, while winger Ruairi Keating has joined until the end of the season.
5. Torquay United have signed defender Matt Anderson on a two-and-a-half-year deal and brought in Republic of Ireland striker Myles Keating on a short-term deal.
6. Torquay United have signed defender James Anderson and striker Myles Keating.
7. **Torquay United have signed defender Myles Anderson and striker Ruairi Keating. (R1: 62.5)**
8. Torquay United have signed defender Lewis Anderson and striker Ruairi Keating.
9. Torquay United have loaned defender Myles Anderson.
10. National League strugglers Torquay United have signed defender Lewis Anderson on a two-and-a-half-year deal and Irish striker Ruairi Keating until the end of the season.

Deterministic summary: Torquay United have signed defender Myles Anderson and striker Ruairi Keating until the end of the season. (R1: 52.63)

Target summary: Torquay United have signed Barrow defender Myles Anderson on a permanent deal, and Irish forward Ruairi Keating on non-contract terms.

BLEU variance: 0.38

Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G.J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon, and Yarin Gal. 2019. A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning, ICML 2016*, volume 3, pages 1050–1059. PMLR.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian active learning with image data. In *34th International Conference on Machine Learning, ICML 2017*, volume 3, pages 1183–1192. PMLR.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A Divide-and-Conquer Approach to the Summarization of Long Documents. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 28.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701. Curran Associates, Inc.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Mate Lengyel. 2011. Bayesian active learning for classification and preference learning.

Alex Kendall and Yarin Gal. 2017. What uncertainties

do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 2017-December, pages 5580–5590. Curran Associates, Inc.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, volume 3201, pages 217–226, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 9332–9346. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis.

Mingyi Liu, Zhongjie Wang, Zhiying Tu, and Xiaofei Xu. 2020. LTP: a new active learning strategy for BERT-CRF based named entity recognition.

Zhihao Lyu, Danier Duolikun, Bowei Dai, Yuan Yao, Pasquale Minervini, Tim Z Xiao, and Yarin Gal. 2020. You Need Only Uncertain Answers: Data Efficient Multilingual Question Answering. In *Workshop on Uncertainty and Robustness in Deep Learning*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

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

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 2017 Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083. Association for Computational Linguistics.

Aditya Siddhant and Zachary C. Lipton. 2020. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2904–2909, Brussels, Belgium. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 2019 International Conference on Machine Learning*, pages 5926–5936. PMLR.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Tim Z. Xiao, Aidan N. Gomez, and Yarin Gal. 2020. Wat zei je? Detecting out-of-distribution translations with variational transformers.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding Neural Abstractive Summarization Models via Uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6275–6281. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 2020-December.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *37th International Conference on Machine Learning, ICML 2020*, pages 11328–11339. PMLR.

Rui Zhang and Joel Tetreault. 2020. This email could save your life: Introducing the task of email subject line generation. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 446–456. Association for Computational Linguistics.

A Appendix

Figures 4 and 5 show the performance versus data retention curves of our Variational models in terms of ROUGE-2 and ROUGE-L F-score respectively. The observations here are similar to Figure 1.

Figures 6 and 7 show the differences in ROUGE-2 and ROUGE-L performance of the Variational

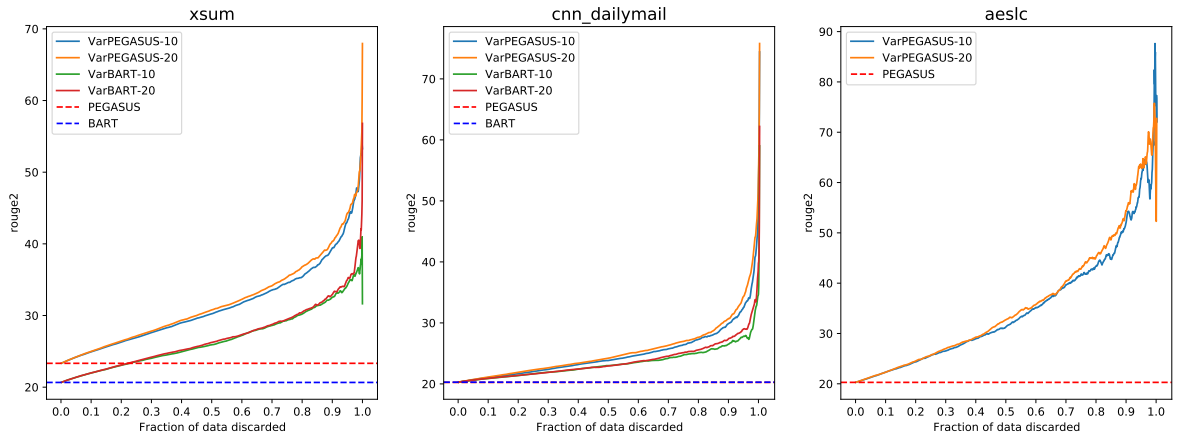

Figure 4: ROUGE-2 scores vs fraction of data discarded due to high BLEUVarN. The straight dashed lines indicate the performance level of the deterministic PEGASUS and BART models.

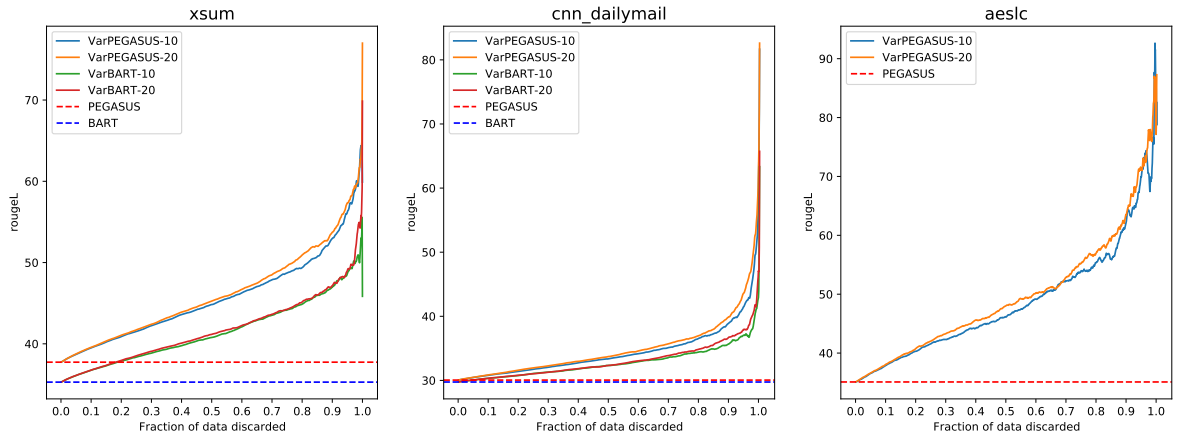

Figure 5: ROUGE-L scores vs fraction of data discarded due to high BLEUVarN. The straight dashed lines indicate the performance level of the deterministic PEGASUS and BART models.

models versus the deterministic ones. What we see here is in aggreement with Figure 3.

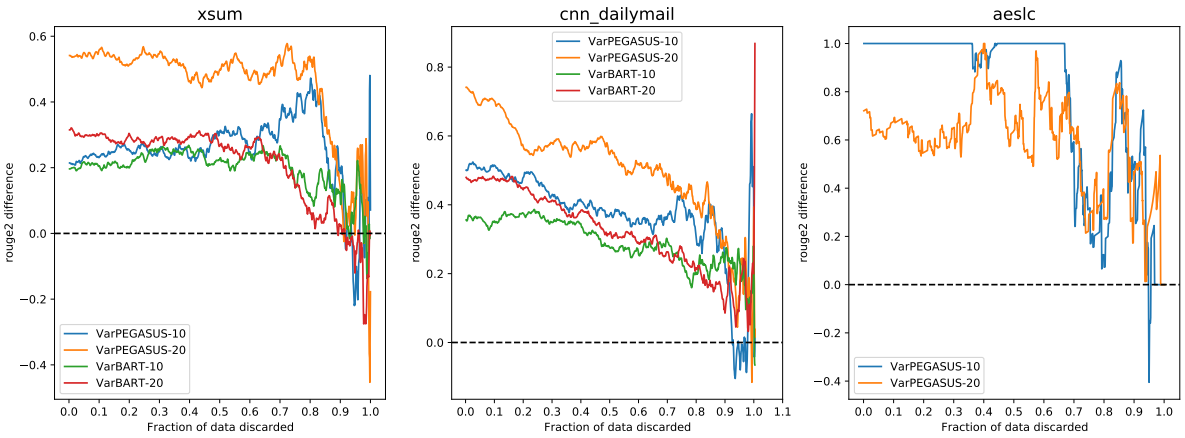

Figure 6: Difference in ROUGE-2 between Variational models and their deterministic counterparts versus the fraction of data discarded. Positive values indicate that deterministic ROUGE-2 is lower than Variational.

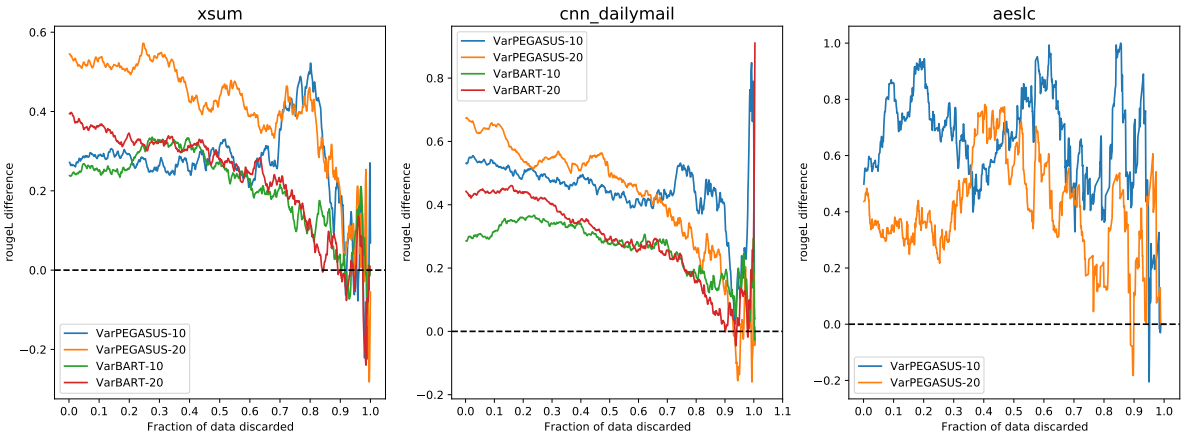

Figure 7: Difference in ROUGE-L between Variational models and their deterministic counterparts versus the fraction of data discarded. Positive values indicate that deterministic ROUGE-L is lower than Variational.