ABSTRACT

We propose a novel approach to summarization based on Bayesian deep learning. We approximate Bayesian summary generation by first extending state-of-the-art summarization models with Monte Carlo dropout and then using them to perform multiple stochastic forward passes. This method allows us to improve summarization performance by simply using the median of multiple stochastic summaries. We show that our variational equivalents of BART and PEGASUS can outperform their deterministic counterparts on multiple benchmark datasets. In addition, we rely on Bayesian inference to measure the uncertainty of the model when generating summaries. Having a reliable uncertainty measure, we can improve the experience of the end user by filtering out generated summaries of high uncertainty. Furthermore, our proposed metric 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.

Keywords: abstractive summarization · Bayesian inference · uncertainty · Monte Carlo dropout

1 Introduction

State-of-the-art text summarization methods have achieved remarkable performance in a variety of tasks [Song et al. [2019], Dong et al. [2019], Lewis et al. [2019], Zhang et al. [2020]. The majority of these methods make use of very large Transformer models that are pre-trained on various language generation tasks. Although these methods are able to generate high quality summaries for texts that are similar to their training set, they are still prone to generating particularly bad outputs when the input text lies far from the distribution of their training data [Xu et al. [2020], Krysciński et al. [2020]. Furthermore, deep neural networks in general, are usually fairly confident about the outputs they produce, regardless of how familiar they are with the input [Gal and Ghahramani [2016], Xiao et al. [2020].

Recent summarization methods have focused heavily on improving the overall performance, but the topic of model uncertainty has been explored very little [Xu et al. [2020]. Since the output of automatic summarization models is in most applications 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 the trust of users to automated summarization systems.

In addition, the development of uncertainty measures for summarization can open 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 such approaches stems from the fact that obtaining labeled samples for training is hard, while at the same time it is relatively easy to obtain large amounts of unlabeled samples. The task of summarization is no different from that perspective, since creating good quality target summaries for training can be very costly. Nevertheless, very little work has been done on applying active learning for summarization [Zhang and Fung [2009], and no work exists that focuses on the state-of-the-art, abstractive summarization models.

This work explores the topic of 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.
Table 1: High uncertainty example from the XSum dataset. Sample summary (1), in bold typeface, is the median sample 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.**
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".

**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:** 86.79

with Monte Carlo dropout [Gal and Ghahramani 2016], in order to create corresponding variational PEGASUS (VarPEGASUS) and BART (VarBART) models. Sampling multiple summaries from those models allows us to approximate Bayesian inference in a practical way. To the best of our knowledge this is the first attempt to apply Bayesian summary generation with large Transformer models.

We investigate model uncertainty based on Bayesian approximation, by presenting a study of the Monte Carlo BLEU variance uncertainty metric [Xiao et al. 2020] in the context of summarization. Our findings suggest that this uncertainty metric correlates with the quality of the generated summaries and can be used to effectively identify cases of questionable quality (see Table 1).

In addition, we propose selecting the summary with the lowest disagreement, out of the summaries sampled from our variational models, as a way of consistently improving summarization performance (see Table 4). Experiments across multiple benchmark datasets show that 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. In Section 2 we present related work on Bayesian deep learning and uncertainty methods. In Section 3 we describe the details of our approach, while in Section 4 we establish our experimental setup. The results of our experiments are presented and discussed in Section 5. Finally, we conclude our work in Section 6.

## 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].

One of the main problems with Bayesian deep learning methods is the computational cost of full Bayesian inference in deep neural networks. [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 (MC) dropout. [Gal...
et al. [2017] use a convolutional neural network with MC dropout in order to obtain an uncertainty estimate for active learning in the task of image classification. Bayesian Deep Active Learning by Disagreement (BALD) Houlsby et al. [2011] samples many networks with Monte Carlo simulation and proposes an objective function that takes into account the disagreement and confidence of the predictions coming from the different 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 the task of question answering and propose an active learning approach based on a modified version of BLEUVar. Similarly, Liu et al. [2020] use a conditional random field to obtain uncertainty estimates for active learning and apply their method to the task of named entity recognition.

Xu et al. [2020] is the only work that focuses on uncertainty for the task of summarization, but their work does not make use of Bayesian methods. They define the uncertainty of generated summaries based on the entropy of each token generated by the model during the decoding phase. Their study involves experiments on two summarization datasets, namely CNN/DM and XSum, using the PEGASUS and BART summarization models. Their work mainly focuses on understanding different properties of uncertainty during the decoding phase, and 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 analyze Monte Carlo BLEU variance, an uncertainty metric that is based on Bayesian inference. Finally, we show how Bayesian inference can be applied to the task of summarization.

3.1 Monte Carlo dropout

As opposed to standard neural networks, Bayesian probabilistic models can explicitly model the notion of uncertainty. The goal of Bayesian models is to derive the entire posterior distribution of the model parameters \( \theta \) given the training data \( X \) and \( Y \) (Equation 1).

\[
P(\theta|X,Y) = \frac{P(Y|X,\theta)P(\theta)}{P(Y|X)}
\]

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

\[
P(\hat{y}|x,X,Y) = \int P(\hat{y}|x,\theta)P(\theta|X,Y)d\theta
\]

In practice, integrating over all possible parameter values for a deep neural network is intractable. Therefore, approximate methods are used for inference. Making stochastic forward passes with dropout Srivastava et al. [2014] turned on at test time is equivalent to drawing from the predictive distribution. This method is called Monte Carlo (MC) dropout Gal and Ghahramani [2016] and can be easily applied to any neural network that has been trained with dropout, by simply using a different dropout mask for each forward pass.

3.2 Summary uncertainty

It is possible to extend any state-of-the-art summarization model with MC dropout, this way converting it to a Bayesian model, and sample \( N \) stochastic summaries for a given input text. In Bayesian models, the variance of the predictive distribution can be used to measure the prediction uncertainty of the model. For a text summarization model, we will measure the variance of the samples \( y_i, i = 1 \ldots N \), generated with MC dropout, as the uncertainty at input \( x \). This follows the paradigm of Bayesian Active Learning by Disagreement (BALD) Houlsby et al. [2011] that has been successfully applied to various deep learning problems.

The BLEU metric Papineni et al. [2002] is a common way of measuring the similarity between a pair of texts. As in Xiao et al. [2020], we measure the variance of the generated summaries with the BLEU Variance (BLEUVar) metric over the \( N \) summaries generated during the Bayesian inference as shown in Equation 3. In essence, BLEUVar is
computed by measuring the sum of the squared complement of BLEU among all pairs of summaries generated with different dropout masks for the same input.

\[
BLEUVar = \sum_{i=1}^{N} \sum_{j \neq i}^{N} (1 - BLEU(y_i, y_j))^2
\]  

(3)

By running multiple stochastic forward passes for the same input, we effectively create an ensemble of models with different parameters. Such an approach has the following effects. For inputs that are close to the model’s learned distribution, the model will be confident about its predictions and the generated summaries will be similar to one another. On the other hand, for inputs that lie away from that distribution, the model does not have a clear focus and the generated summaries will differ wildly.

3.3 Bayesian summary generation

We propose a novel Bayesian approach to summary generation. Instead of taking the deterministic summary produced with dropout turned off as the final summary of a summarization model for a given input, we instead consider selecting one of the \(N\) stochastic summaries produced with MC dropout. Conceptually, what we need is an approximation for the median of all samples, but this is hard to define when the samples are text sequences. We assume that such a median sequence should have the lowest disagreement with the rest of the \(N - 1\) samples.

Since the pairwise complement of BLEU between all pairs of the sampled summaries has already been computed when estimating the BLEUVar uncertainty, it can be used further to help us find the median summary. In practice, we select the summary \(\hat{\mu}\) that minimizes Equation 4 Xiao et al. [2020].

\[
\hat{\mu} = \arg\min_{y_i} \sum_{j \neq i}^{N} [(1 - BLEU(y_i, y_j)) + (1 - BLEU(y_j, y_i))]
\]  

(4)

We expect the median of the generated summaries to integrate the most important aspects that all individual summaries agree on. For inputs that are very close to the model’s learned distribution, the individual summaries will be similar to each other and as a result their median will be an equally good choice. At the same time, when it comes to out-of-distribution inputs, taking the median of a number of very different outputs will result in a more robust and overall better final summary.

In real world situations, even for a well trained model, we expect to have a fairly large number of inputs that will not be as close its’ learned distribution, and therefore we expect to benefit from the positive effects of ensembling multiple opinions.

4 Experimental Setup

We first present the three datasets that are involved in our experiments, their main statistics and 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.

The main criteria for the selection of these datasets are the availability of recent, open source models trained on those datasets and the relatively short texts that would allow us to run a number of different experiments fairly quickly. Since
Table 2: Basic statistics for each of the datasets used in our experiments. The document and summary length are measured in words.

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

our methods do not involve training, we will only focus on the validation and test set of each one of the datasets. All datasets are obtained from the Hugging Face datasets repository. Table 2 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 make use of models trained and open-sourced for each one of our test datasets. These models can be found in the Hugging Face models repository.

The BART models we are using follow the BART\textsubscript{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 available for the XSum and CNN/DM datasets only.

Our PEGASUS models follow the architecture of PEGASUS\textsubscript{LARGE} 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 the Transformer blocks of the encoder and decoder. For each sample, we generate $N$ summaries using beam search decoding with 8 beams. We experimented with $N$ equal to 10 and 20 and measured the differences. The rest of the hyper-parameters used were identical to the original papers.

5 Results

Our empirical study focuses on evaluating two aspects of Bayesian abstractive summarization. First, we explore the potential of the BLEUVar metric, which relies on Bayesian inference, as an uncertainty measure for the summaries generated by summarization models. Second, we investigate the effectiveness of Bayesian inference as a way to improve summarization performance on test data.

5.1 Evaluating uncertainty

We want to evaluate the effectiveness of BLEUVar in measuring the uncertainty in the generations of the model. The performance versus data retention curve [Filos et al. 2019] measures how well an uncertainty metric would perform if we completely removed the $k$% most certain outputs from the test set. In the $x$-axis we have the fraction of data from the test set that are retained, while in the $y$-axis we can have some performance metric. An effective uncertainty metric should show a consistent improvement in performance as we exclude from the evaluation high uncertainty samples.

In Figures 1-3, we show, for each dataset, the performance of our Variational models in terms of ROUGE-1, ROUGE-2 and ROUGE-L F-scores versus the fraction of data retained using the BLEUVar metric. For reference, we are also plotting the performance of the deterministic models as straight lines.

As shown in Figures 1-3, all ROUGE F-scores are steadily improving as we exclude samples with the highest BLEUVar. This observation is also consistent across all test datasets, which indicates that there is correlation between BLEUVar and summarization performance. It is interesting to see that we are getting a very significant increase in performance by simply removing 25% of the samples with the highest BLEUVar uncertainty.
Figure 1: Performance vs fraction of data retained using BLEUVar for the XSum dataset. The straight dashed lines indicate the performance level of the deterministic PEGASUS and BART models.

Figure 2: Performance vs fraction of data retained using BLEUVar for the CNN/Dailymail dataset. The straight dashed line indicates the performance level of the deterministic PEGASUS and BART models.

Figure 3: Performance vs fraction of data retained using BLEUVar for the AESLC dataset. The straight dashed line indicates the performance level of the standard PEGASUS model.
Finally, we can see that in most cases the 20 samples curve is a bit higher than the 10 samples curve, which shows that sampling more summaries is also beneficial in this setup. At the same time, all Variational models are well above the deterministic lines.

It should be noted here, that the decline in performance of VarBART is much steeper than VarPEGASUS on both the XSum and the CNN/DM dataset. This observation, which is particularly noticeable on the XSum dataset, hints us that the BART model is in general more confident about the outputs it generates.

In Figure 4 we show the increase in the average BLEUVar of all Variational models as we remove data from the test sets of XSum and CNN/DM. Once again, we observe that the BLEUVar curves of the VarBART models are much more steep compared to the VarPEGASUS ones. Also, by looking at the difference between the minimum and the maximum BLEUVar values for each model, it is clear that VarPEGASUS has a much larger range of uncertainty about it’s predictions. Anecdotally, we can say here that PEGASUS is more aware of the things it does not know.

Table 3: A comparison of our varBART and VarPEGASUS models against the deterministic BART and PEGASUS. VarBART-10 and VarPEGASUS-10 sample 10 summaries with MC dropout, while VarBART-20 and VarPEGASUS-20 sample 20.

| Model     | XSum   | CNN/DM | AESLC |
|-----------|--------|--------|-------|
|           | 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 |

5.2 Bayesian vs deterministic summarization

Table 4 reports the ROUGE-1, ROUGE-2 and ROUGE-L F-scores of our VarBART and VarPEGASUS models on all benchmark datasets. For comparison, we also report the performance of the deterministic BART and PEGASUS models in the same datasets.

The trained BART and PEGASUS models as well as the ROUGE implementation we are using are different from the ones in the original papers, so we decided to re-evaluate these models on the test set of each dataset in order for the results to be comparable.

From the reported results we can see that both VarBART and VarPEGASUS improve over their deterministic counterparts on all test datasets, which shows that Bayesian summarization is indeed effective. This comes at the cost of increased computational complexity since it involves running the models multiple times for each input. However, this operation
can be optimized in modern deep learning systems by batching MC dropout generations and using different dropout masks for each sample in the batch.

It should be noted here that our goal in these experiments was not to compete with other state-of-the-art models. This works aims to show that relying on the agreement between multiple Bayesian summaries for the same input, is an effective way to boost summarization performance compared to deterministic models. Furthermore, increasing the number $N$ of samples generated during the Bayesian inference further improves performance for all datasets except for AESLC, although this increases computational complexity even further.

We also want to study in more detail the difference in performance of the variational models compared to the deterministic ones. In Figures 5 and 6 we plot the difference in ROUGE of each variational model with its deterministic counterpart as a fraction of the data retained for the XSum and CNN/DM datasets. Those plots give us a better view of how the variational models fare against the deterministic ones for different levels of uncertainty. Positive values indicate that the variational model achieves a higher score than the deterministic.

By looking at the curves we can clearly see that the scores fluctuate between positive and negative values for the top 10% – 20% of the data with the lowest uncertainty. Intuitively, when the model uncertainty is very low, we can expect both variational and deterministic models to converge to equally good outputs. Consequently, it is not clear if the variational or deterministic summary would be better. This behavior is consistent for all models and both datasets.

On the other hand, as the model uncertainty becomes higher, we observe that all curves become positive, indicating that the variational models have a clear advantage there. This pattern is also consistent across all models and datasets, and suggests that for samples with high uncertainty ensembling multiple summaries leads to overall better results.
Table 4: Low uncertainty example from XSum. Sample summary (7), in bold typeface, is the median sample according to our approach. We also 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.  | Torquay United have signed defender Myles Anderson and striker Ruairi Keating. (R1: 62.5) | Torquay United have signed defender Myles Anderson and striker Ruairi Keating. (R1: 52.63) |
| 7. Torquay United have signed defender Myles Anderson and striker Ruairi Keating. |                                                                 |                                                                 |
| 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. |

Moreover, we can see that for the variational models using 20 MC dropout samples the performance difference is increasing as uncertainty grows, in most but not all cases. This observation indicates that increasing the number of MC dropout samples would be particularly beneficial in cases of very high uncertainty.

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 data. For each example, we show the 10 sample summaries generated with MC dropout for the same input as well as the corresponding BLEUVar 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 1) is a case of high uncertainty from the XSum dataset. We can see that all 10 samples are different from one another and that leads to a high BLEUVar score. In contrast the second example (Table 4) has much lower uncertainty. In this case all 10 samples seem to mostly agree on the main points and as a result BLEUVar is fairly low. Here, the median summary is the one that more closely resembles that agreement.

From the example in Table 4 we can also see that the median Bayesian summary is close but slightly better than the deterministic summary in terms of ROUGE. Looking more closely at the deterministic summary we can also see that it makes a factual error as opposed to the median Bayesian summary.

6 Conclusion

This work explored Bayesian methods in the context of text summarization. Based on experimental results, we showcase that a Bayesian summarization model with MC dropout is able to outperform a similar deterministic model on a number of different benchmarks. Furthermore, we demonstrate that Bayesian summarization models are able to effectively model prediction uncertainty. We show that BLEUVar, which is a metric based on Bayesian inference, can be used to identify generated summaries with high uncertainty.

Our work can have a broader impact to both the industry and the end users of summarization systems. Uncertainty measures are commonly used as criteria for selecting samples for annotation and can be paired nicely with active
learning and human-in-the-loop approaches. Such methods have the potential to increase the efficiency of the data labeling process, particularly when the labeling budget is limited, which is the case in most real world situations.

At the same time, when it comes to serving prediction, having the ability to effectively measure the model’s uncertainty is a significant advantage. By detecting generations with high uncertainty and filtering them out, we can greatly improve the experience of the end users as well as increase their confidence towards the automatic summarization system.

References

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie Yan Liu. MASS: Masked sequence to sequence pre-training for language generation. In Proceedings of the 2019 International Conference on Machine Learning, pages 5926–5936. PMLR, 2019. ISBN 9781510886988.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 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, 2019. doi:10.18653/v1/2020.acl-main.703.

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

Jiacheng Xu, Shrey Desai, and Greg Durrett. 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, 2020. doi:10.18653/v1/2020.emnlp-main.508.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 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, 2020. doi:10.18653/v1/2020.emnlp-main.750.

Yarin Gal and Zoubin Ghahramani. 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, 2016.

Tim Z. Xiao, Aidan N. Gomez, and Yarin Gal. Wat zei je? Detecting out-of-distribution translations with variational transformers, 2020. ISSN 23318422.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In 34th International Conference on Machine Learning, ICML 2017, volume 3, pages 1183–1192. PMLR, 2017. doi:10.17863/CAM.11070.

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

Mingyi Liu, Zhongjie Wang, Zhiying Tu, and Xiaofei Xu. LTP: a new active learning strategy for BERT-CRF based named entity recognition, 2020. ISSN 23318422.

Zhihao Lyu, Danier Duolikun, Bowei Dai, Yuan Yao, Pasquale Minervini, Tim Z Xiao, and Yarin Gal. You Need Only Uncertain Answers: Data Efficient Multilingual Question Answering. 2020.

Justin Jian Zhang and Pascale Fung. Active learning of extractive reference summaries for lecture speech summarization. In Proceedings of the 2nd Workshop on Building and Using Comparable Corpora from Parallel to Non-parallel Corpora, pages 23–undefined. Association for Computational Linguistics, 2009. doi:10.3115/1690339.1690346.

Aditya Siddhant and Zachary C. Lipton. 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, 2020. Association for Computational Linguistics. doi:10.18653/v1/d18-1318.
Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. ISSN 15337928.

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*, pages 311–318. Association for Computational Linguistics, 2002.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 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, 2018. Association for Computational Linguistics. doi:10.18653/v1/d18-1206

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

Abigail See, Peter J Liu, and Christopher D Manning. 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, 2017. doi:10.18653/v1/P17-1099

Rui Zhang and Joel Tetreault. 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, 2020. doi:10.18653/v1/p19-1043

Bryan Klimt and Yiming Yang. 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, 2004. Springer Berlin Heidelberg. doi:10.1007/978-3-540-30115-8_22

Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G.J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon, and Yarin Gal. A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks, 2019. ISSN 23318422.