Global-aware Beam Search for Neural Abstractive Summarization

Anonymous Author(s)
Affiliation
Address
email

Abstract

This study develops a calibrated beam-based algorithm with global awareness for neural abstractive summarization, aiming to improve the local optimality problem of the original beam search in a rigorous way. Specifically, a novel global protocol is proposed based on the attention distribution to stipulate how a global optimal hypothesis should attend to the source. A global scoring function is then developed to regulate beam search to generate summaries in a more near-global optimal fashion. This novel design enjoys a distinctive property, i.e. the global attention distribution could be predicted before inference, enabling stepwise improvements on the beam search through the global scoring function. Extensive experiments on 9 datasets show that the global-aware inference significantly improves state-of-the-art summarization models even using empirical hyper-parameters. The algorithm is also proven robust as it remains to generate meaningful texts with corrupted attention distributions. The codes and a comprehensive set of examples are available in supplementary materials.

1 Introduction

As the barriers exist from the auto-regressive design of neural probabilistic text generators to predicting the global optimum directly [26], the heuristic algorithm beam search that factorizes global optimization to multiple local optimizations, has been popularly used for text decoding [20]. In the original beam search setting, the global optimum is a hypothesis $\boldsymbol{g}$ of the highest probability among all possible sentences, and consists of words in vocabulary $\mathcal{V}$. Given the global optimum at step t denoted as $\boldsymbol{g}_{\leq t}$, the local optimums $\boldsymbol{l}_{\leq t}$ refer to $\mathcal{K}$ candidate sequences with the highest probabilities at each step. While it is necessary to compromise on the beam size $\mathcal{K} \ll \mathcal{V}$ to ensure text quality [3, 21, 20] and search efficiency, beam search suffers from a major limitation mainly due to its local property. Namely, it has to operate solely with the local information available at each step, by relying on the ideal assumption that the global optimal hypothesis is within the $\mathcal{K}$ local optimal hypotheses of the highest probabilities, i.e. $p(\boldsymbol{g}_{\leq t}) \geq p(\boldsymbol{l}_{\leq t})$, for all t until the termination T. In practice, the aforementioned assumption may however fail in the case that the probability of the global optimum at step $s < T$ is less than those of the local optimums, i.e. $p(\boldsymbol{g}_{\leq s}) < p(\boldsymbol{l}_{\leq s})$, but is adjusted to a higher level in the later steps, $p(\boldsymbol{g}_{>s}|\boldsymbol{g}_{\leq s}) > p(\boldsymbol{l}_{>s}|\boldsymbol{l}_{\leq s})$ and $p(\boldsymbol{g}_{\leq s})p(\boldsymbol{g}_{>s}|\boldsymbol{g}_{\leq s}) > p(\boldsymbol{l}_{\leq s})p(\boldsymbol{l}_{>s}|\boldsymbol{l}_{\leq s})$. This often leads beam search to get stuck in the local optimum from step s onward in generating texts .

This study proposes a calibrated beam-based algorithm with global awareness at all searching steps. Generally, our novel algorithm is implemented in two phases. Before the beam search (Phase I), the *global attention distribution* is predicted in order to be included as a protocol to calibrate beam search at each step, encouraging the generated hypotheses to attend to source in a more near-global optimal way. Specifically, the global attention distribution is essentially the distribution of the total attention that all reference tokens assign to each source token, which could be predicted from the source by training an attention-prediction model. The training is fairly straightforward and resembles a sequence

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

tagging task [11], except that the predicted attention distribution from the source is a regression result. There are several advantages of using the attention distribution as the global protocol. 1) Attention distributions are sensitive to the decoder input, suggesting that any input to the decoder leads to a unique attention distribution with fixed model parameters; 2) attention distributions are accessible for almost all text generation tasks thanks to the recent advances in attention models [23, 4, 27]; 3) relying on the source only, the global attention distribution can be predicted before beam search, thus offering a rigorous mechanism to calibrate a global-aware beam search.

During beam search (Phase II), we develop a novel step-wise *global scoring function* to guide beam search based on the predicted global attention distribution. As one main theoretical result, we show that the global score can be considered as the probability that generated texts attend to sources in line with the predicted global attention distribution. Specifically, the generated tokens in each step update the local attention distribution to source tokens dynamically, where the attention values grow monotonically as the generation proceeds. Since the desired situation is that the local distribution reaches exactly the global distribution at the terminal step, we regulate the inference by discouraging local attention values from exceeding their corresponding predicted global attention values at all steps.

With regards to the core target to investigate the possible paradigm that improves beam search with global awareness during decoding, contributions of this study are summarized as follows:

- We argue that the limitation of beam search roots from its defect in finding the global optimal hypothesis. We improve the algorithm by proposing a global protocol to regulate beam search step-by-step. This paper is the first to predict and deploy the global attention distribution to calibrate the inference in a rigorous way, thus returning a hypothesis that attends to source tokens in a more near-global optimal manner. Previous works [33, 7, 17, 25, 18] try to use attention distributions to improve beam search, but ignore that the global attention distribution is predictable.
- A novel global scoring mechanism is designed to evaluate the generated sequences at each inference step based on the desired relationship between local attention distributions and the predicted global attention distribution. As theoretically justified, it can be elegantly integrated into beam search in the form of a probability so that there only increases $\mathcal{O}(\mathcal{K})$ time complexity in each step (see Section § 3.1 for more details).
- The proposed algorithm with global awareness manifests a robust and plug-and-play property in enhancing beam search for neural abstractive summarization. Without requiring any model or parameter modification, the global-aware inference shows excellent performance in generating meaningful texts, even if the attention distribution used is corrupted or not of its own. Further, it is found that summaries generated by global-aware inference are rather different from beam search hypotheses and are of higher quality (see *Global-aware* in Table 1). More interestingly, we find that the generation style of a dataset could be transferred by the designated global attention distribution. For instance, summaries of higher abstractness for CNN/DM could be generated by only replacing its global attention distribution with a highly abstractive distribution during inference, as presented in *Global-aware*† of Table 1.
- On the empirical side, we show that the proposed global-aware inference can boost two state-of-the-art summarization models BART [16] and PEGASUS [35] to produce higher quality summaries on 9 datasets. The improvements are observed stable and significant even if only the empirical hyper-parameters are used.

2 Preliminary

The proposed decoding strategy is applied in BART [16] and PEGASUS [35] to perform summarization. BART is a pre-trained seq-to-seq model whose structure essentially follows a vanilla Transformer encoder-decoder [27]. PEGASUS has a similar structure but is pre-trained on a larger dataset differently. **Note that both models used in this paper are fine-tuned and parameter-fixed.**

2.1 Cross Attention

Cross attention refers to the attention value that a token in the target sequence gives to a source token. For example, the cross attention in Transformer [27] is calculated by a multi-head attention

Table 1: Use BART [16] fine-tuned in CNN/DM to generate summaries. *Global-aware* uses the attention distribution learned from CNN/DM, while *Global-aware*† takes the attention distribution learned from XSUM. See the supplementary document (App. F) for more examples.

Beam search	President Obama says climate change is a public health issue that affects all of us . Obama: "No challenge poses more of a public threat than climate change" Obama: "Millions of people would lose their health insurance" if Affordable Care Act is not upheld . Obama: "I am not anticipating the Supreme Court would make such a bad decision"
Global-aware	President Obama says climate change is a public health issue that affects all of us . He says the average American can do their part to reduce their own carbon footprint . Obama did not appear particularly concerned about the Supreme Court challenge to the Affordable Care Act .
Global-aware†	President Barack Obama says climate change is a public health issue . He says the average American can do their part to reduce their carbon footprint .

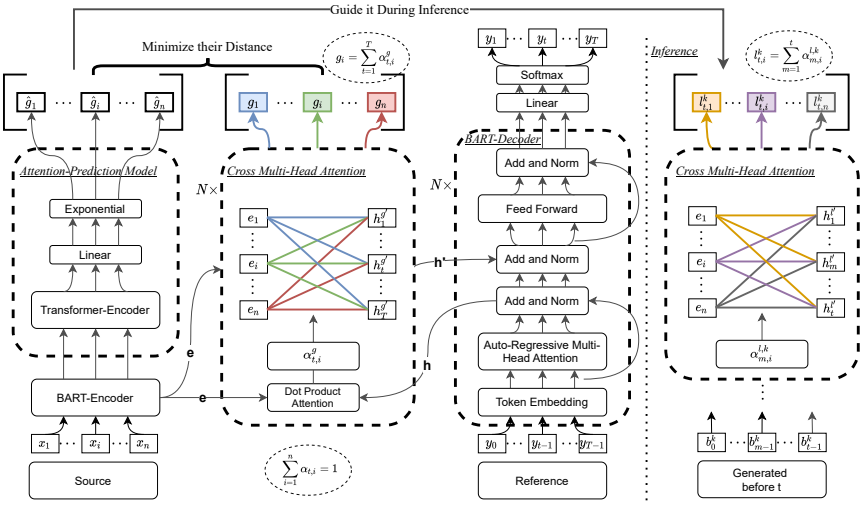

Figure 1: Process of calculating global/local attention distributions and training the attention-prediction model. The attention distribution is composed of the summation of $\alpha_{t,i}$ on the same-colored lines, distinguished from the summation of different-colored lines which always equals 1 due to softmax. In the decoding stage, the reference tokens are replaced by the generated sequence before t to obtain the local attention distribution at t. Note that parameters in BART are fixed.

which connects the encoder and decoder[1], as diagrammed in Figure 1. Given a source sequence $\boldsymbol{x} = (x_1, \cdots, x_n)$, a target sequence $\boldsymbol{y} = (y_0, \cdots, y_T)$ and a fine-tuned BART, the encoder output is $\boldsymbol{e} = [e_1; \cdots; e_i; \cdots; e_n] \in \mathbb{R}^{n \times d}$, and the auto-regressive hidden representation for $\boldsymbol{y}_{<t} = (y_0, \cdots, y_{t-1})$ is $\boldsymbol{h} = [h_1; \cdots; h_t] \in \mathbb{R}^{t \times d}$, where d is the dimension of hidden states.

$$\alpha_{t,i} = \frac{\exp\left(h_t \boldsymbol{w}_q (e_i \boldsymbol{w}_k)^T / \sqrt{d}\right)}{\sum_{i=1}^{n} \exp\left(h_t \boldsymbol{w}_q (e_i \boldsymbol{w}_k)^T / \sqrt{d}\right)}, \ t \geq 1 \tag{1}$$

where $\alpha_{t,i}$ indicates the cross attention that the t_{th} target token y_t assigns to the i_{th} source token x_i, and $\boldsymbol{w} \in \mathbb{R}^{d \times d}$ are two projection matrices. Although we regard $\alpha_{t,i}$ as the attention from y_t, it is only dependent on $\boldsymbol{y}_{<t}$ because the decoder is an auto-regressive model. With both sources and reference targets available in the training set, the attention given by reference tokens is computable, called global cross attention $\alpha^g_{t,i}$. During decoding, the local cross attention at step t is calculated by the sequence generated before t. Since there are often several different candidate sequences at each inference step, we denote the local cross attentions as $\alpha^{l,k}_{t,i}$ where k indexes the k_{th} candidate sequence.

[1]Mean pooling is used for multi-layers and multi-heads.

2.2 Beam Search

Since the decoder of the seq-to-seq model is an auto-regressive model, the target sequence probability can be factorized to the conditional probability of tokens at each decoding step t, i.e.,

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{T} p(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}),\ T \geq 1 \tag{2}$$

when $T = 0$, $\boldsymbol{y} = (y_0)$ and $p(y_0|\boldsymbol{x}) = 1$ because y_0 is a fixed start token. Beam search [33] is a decoding strategy to predict a target sequence by maximizing this factorization. Given a vocabulary set $\mathcal{V}$, at each inference step t, beam search selects a beam set $\mathcal{B}_t^{\mathcal{K}} = \{\boldsymbol{b}_t^1, \cdots, \boldsymbol{b}_t^{\mathcal{K}}\}$ (where each beam $\boldsymbol{b}_t^k = (b_0^k, \cdots, b_t^k)$ is a candidate sequence) from a huge beam set $\mathcal{B}_t$ of size $\mathcal{K} \times |\mathcal{V}|$, namely,

$$\mathcal{B}_t = \{\boldsymbol{b}_{t-1}^k \circ v \mid \boldsymbol{b}_{t-1}^k \in \mathcal{B}_{t-1}^{\mathcal{K}} \textbf{ and } v \in \mathcal{V}\} \tag{3}$$

$$\mathcal{B}_t^{\mathcal{K}} = \{\boldsymbol{b}_t^k | \boldsymbol{b}_t^k = \text{argtopk}\left(\log p(\boldsymbol{b}_t|\boldsymbol{x})\right),\ \boldsymbol{b}_t \in \mathcal{B}_t,\ t \geq 1\} \tag{4}$$

where argtopk$(\cdot)$ outputs $\mathcal{K}$ beams with the highest conditional probability, and $\circ$ is the concatenation operation. Besides, $\mathcal{B}_0^{\mathcal{K}} = \{b_0^1, \cdots, b_0^{\mathcal{K}}\}$ where b_0^k is the start token. By Eq. 2, $\log p(\boldsymbol{b}_t|\boldsymbol{x})$ is an accumulated value, its calculation can be simplified as:

$$\log p(\boldsymbol{b}_t|\boldsymbol{x}) = \begin{cases} \log p(\boldsymbol{b}_{t-1}^k|\boldsymbol{x}) + \log p(v|\boldsymbol{x}, \boldsymbol{b}_{t-1}^k), & t \geq 2 \\ \log p(v|\boldsymbol{x}, b_0^k), & t = 1 \end{cases} \tag{5}$$

where the value of $\log p(\boldsymbol{b}_{t-1}^k|\boldsymbol{x})$ is computed from the previous step. Therefore, at each step, we only need to calculate the condition probability of each token in the vocabulary set.

A beam is terminated after it generates an ending token, and the beam set of $\mathcal{K}$ terminated beams is defined as $\mathcal{Y}$. The final hypothesis $\boldsymbol{y}^*$ is chosen from $\mathcal{Y}$ based on the beam probability normalized by $length^a$ where α is a hyper-parameter of length [33]:

$$\boldsymbol{y}^* = \underset{\boldsymbol{y}^k \in \mathcal{Y}}{\text{argmax}}\ \frac{\log p(\boldsymbol{y}^k \mid \boldsymbol{x})}{(|\boldsymbol{y}^k| - 1)^a} \tag{6}$$

where $\boldsymbol{y}^k = (y_0^k, \cdots, y_T^k)$. $|\boldsymbol{y}^k| - 1$ is used since the start token is not considered in calculating the length.

3 Proposed Global-aware Inference

3.1 Global Score

The global attention distribution G can be attained by the global cross attention $\alpha_{t,i}^g$, i.e., $G = [g_1, g_2, \cdots, g_n]$, where $g_i = \sum_{t=1}^{T} \alpha_{t,i}^g$ (see Figure 1). Thereinto, $G \in \mathbb{R}^n$ and n indicates the source length, g_i refers to the total attention that the reference attends to the i_{th} source token. Moreover, its sum $\sum_{i=1}^{n} g_i = \sum_{i=1}^{n} \sum_{t=1}^{T} \alpha_{t,i}^g = T$ due to $\sum_{i=1}^{n} \alpha_{t,i}^g = 1$, where T is the reference length, or equivalently the optimal length Z.

Correspondingly, we define the attention distribution calculated by the local cross attention $\alpha_{t,i}^{l,k}$ as the dynamic local attention distribution L_t^k, i.e., $L_t^k = [l_{t,1}^k, l_{t,2}^k, \cdots, l_{t,n}^k]$, where $l_{t,i}^k = \sum_{m=1}^{t} \alpha_{m,i}^{l,k}$ (similar to G, but computed from different decoder input) and $l_{t,i}^k$ denotes the total attention paid to the i_{th} source token by the beam sequence $(b_1^k, \cdots, b_t^k)$. As mentioned in Section § 2.1, $l_{t,i}^k$ is only dependent on $\boldsymbol{b}_{t-1}^k = (b_0^k, \cdots, b_{t-1}^k)$. Thus the global scoring function $\mathcal{G}(\cdot)$ at the step t depends on $\boldsymbol{b}_{t-1}^k$,

$$\mathcal{G}(\boldsymbol{b}_{t-1}^k) = \frac{\sum_{i=1}^{n} \min(l_{t,i}^k, g_i)}{\zeta_t^k},\ \zeta_t^k = \sum_{i=1}^{n} l_{t,i}^k,\ t \geq 1 \tag{7}$$

where ζ_t^k indicates the total attention that the generated sequence $(b_1^k, \cdots, b_t^k)$ gives to the source, and $\zeta_t^k = |\boldsymbol{b}_{t-1}^k| = t$ because the assignable attention for each generated token is 1. Notably, Eq. 7

gets the maximum score provided that each $l_{t,i} \leq g_i$. As mentioned in Section § 1, the reason for this design is that we desire the local attention value $l_{T,i}$ at the termination is exactly g_i, since the final hypothesis is expected to attend to source tokens in the global-optimal manner. Meanwhile, we have $l_{T,i} > l_{t,i}$ for $t < T$ because $l_{t,i}$ monotonically increases on t with $\alpha^l_{m,i} > 0$. Therefore, at any step, the local attention $l_{t,i}$ should not surpass g_i. Otherwise, the global score will decline, and the penalty depends on the total amount by which these $l_{t,i}$ exceed g_i (see Theorem 3.1). Further, the global score could be considered as the proportion of correctly assigned attention to the total attention given by the generated sequence, where correct assignment indicates that all parts of $l_{t,i}$ do not exceed g_i. Also, it could be interpreted as the probability of correct allocation of attention (see Corollary 3.1.1). In this case, the total global score can be expressed as the same multiplicative form as Eq. 2 to be elegantly integrated into beam search.

Theorem 3.1. *Given* $\mathcal{S}^k_t = \{s | s = l^k_{t,i} - g_i,\ l^k_{t,i} > g_i,\ i = 1, \cdots, n\}$ *and* $S = \sum_{s \in \mathcal{S}^k_t} s$ *where* $S \geq 0$*, then* $\mathcal{G}(\boldsymbol{b}^k_{t-1})$ *decreases as* S *increases.*

Corollary 3.1.1. *The bound of* $\mathcal{G}(\boldsymbol{b}^k_{t-1})$ *is between* 0 *and* 1.

Proof. See the supplementary document (App. B). □

In addition to the constraint for $l_{t,i}$, we still desire $l_{T,i} = g_i$ for each token. A strictly satisfied hypothesis should have two characteristics simultaneously. Namely, its global score at the termination is the maximum 1, and its length equals the optimal length. Therefore, we introduce a length reward mechanism to cooperate with the global score to penalize the situation $l_{T,i} \neq g_i$, which will be discussed at the end of this subsection.

As mentioned before, the total global score at the decoding step t is defined as:

$$\overline{\mathcal{G}}(\boldsymbol{b}^k_{t-1}) = \prod_{m=1}^{t} \mathcal{G}(\boldsymbol{b}^k_{m-1}) \tag{8}$$

Thus, the jointly scoring function $J(\cdot)$ is modified from Eq. 5:

$$\begin{aligned} J(\boldsymbol{b}_t, \boldsymbol{x}) &= \log p(\boldsymbol{b}^k_{t-1}|\boldsymbol{x}) + \beta \log \overline{\mathcal{G}}(\boldsymbol{b}^k_{t-1}) + \log p(v|\boldsymbol{x}, \boldsymbol{b}^k_{t-1}) \\ &= \sum_{m=1}^{t-1} \left(\log p(b_m|\boldsymbol{x}, (\boldsymbol{b}^k_{t-1})_{<m}) + \beta \log \mathcal{G}(\boldsymbol{b}^k_{m-1}) \right) + \log p(v|\boldsymbol{x}, \boldsymbol{b}^k_{t-1}) + \beta \log \mathcal{G}(\boldsymbol{b}^k_{t-1}) \\ &= J(\boldsymbol{b}^k_{t-1}, \boldsymbol{x}) + \log p(v|\boldsymbol{x}, \boldsymbol{b}^k_{t-1}) + \beta \log \mathcal{G}(\boldsymbol{b}^k_{t-1}),\ t \geq 2 \end{aligned} \tag{9}$$

and $J(\boldsymbol{b}_1, \boldsymbol{x}) = \log p(v|\boldsymbol{x}, b^k_0) + \beta \log \mathcal{G}(b^k_0)$, where β is a hyper-parameter to regulate the trade-off between the probability and global score. Similar to $\log p(\boldsymbol{b}_t|\boldsymbol{x})$ in Eq. 5, $J(\boldsymbol{b}_t, \boldsymbol{x})$ is also an accumulative score. Consequently, at each step t, we only need to compute $p(v|\boldsymbol{x}, \boldsymbol{b}^k_{t-1})$ and $\mathcal{G}(\boldsymbol{b}^k_{t-1})$. Compared with Eq. 5, the time complexity of each step only increases $\mathcal{O}(\mathcal{K})$ as there are $\mathcal{K}$ global scores. Replacing $\log p(\boldsymbol{b}_t|\boldsymbol{x})$ in Eq. 4 by $J(\boldsymbol{b}_t, \boldsymbol{x})$, we can select the top $\mathcal{K}$ beams of each decoding step according to both probability and attention distribution.

Considering the length reward mechanism, the final hypothesis is thus defined as:

$$\boldsymbol{y}^* = \underset{\boldsymbol{y}^k \in \mathcal{Y}}{\operatorname{argmax}} \left(\frac{J(\boldsymbol{y}^k, \boldsymbol{x})}{|\boldsymbol{y}^k| - 1} + \beta\gamma R(\zeta^k_T, Z) \right) \tag{10}$$

where $R(\cdot)$ is the length reward function dependent on the optimal length Z and the candidate hypothesis length ζ^k_T. Exactly, ζ^k_T is the total attention that a candidate hypothesis $(y^k_1, \cdots, y^k_T)$ pays to the source and equals $|\boldsymbol{y}^k| - 1$. Besides, the global score and length reward are weighted by a hyper-parameter γ, and the role of β is to ensure that the two are at the same level relative to the probability. We remove α in Eq. 6 as it only adjusts the length preference without really controlling the length, as one should expect from a length reward mechanism.

The design of $R(\cdot)$ could be straightforward – one only needs to ensure that it is increasing as ζ^k_T approaches Z, and reaches the maximum only at $\zeta^k_T = Z$. In this paper, we design a step-wise length reward mechanism $\mathcal{R}(\zeta^k_t, Z)$ to better balance the relationship between the global score and the length reward and make the whole searching process as succinct as beam search. We put the design details of the step-wise length reward in the supplementary document (App. C), and we regard Eq. 10 as the general scoring formulation of global-aware inference.

3.2 Predicting the Global Attention Distribution

Since the reference is unknown in reality, the global attention distribution could only be predicted from the source. We construct an attention-prediction model to learn the relationship between the source tokens and the global attention distribution, as presented in Figure 1.

The input of the attention-prediction model is the fixed encoder output $\boldsymbol{e}$ plus learnable positional encodings $\boldsymbol{p} \in \mathbb{R}^{n \times d}$ which are randomly initialized. The input is fed to a vanilla Transformer-encoder to obtain $\boldsymbol{e}' \in \mathbb{R}^{n \times d}$ that is encoded with additional context information, followed by a linear transformation with an exponential function:

$$\hat{G} = \exp(\boldsymbol{e}'\boldsymbol{w}_g + \boldsymbol{b}_g) \tag{11}$$

where $\hat{G} \in \mathbb{R}^{n \times 1}$ refers to the prediction of G, $\boldsymbol{w}_g \in \mathbb{R}^{d \times 1}$ and $\boldsymbol{b}_g \in \mathbb{R}^{n \times 1}$ are the learnable weights and biases. The exponential function is imposed to ensure $\hat{G} > 0$. We choose the exponential function for this operation because it is shown stable in the training and testing stage. Given the objective of minimizing the distance between $\hat{G}$ and G, the loss is defined as their Euclidean distance:

$$\mathcal{L} = \|\hat{G} - G\|_2 \tag{12}$$

The predicted optimal length $\hat{Z}$ is the sum of elements in $\hat{G}$. Note that the length reward mechanism is not affected by whether $\hat{Z}$ is an integer or not. The pseudo-code is available in the supplementary document (App. A).

4 Experiment

4.1 Setup

Datasets. We evaluate the performance on totally 9 summarization datasets, where 2 datasets (CNN/DM [9], XSUM [5]) with BART [16] and 8 datasets (XSUM [5], BillSum [14], Multi-News [6], NewsRoom [8], WikiHow [15], Reddit TIFU [30], arXiv and PubMed [2]) with PEGASUS [35]. Thereinto, XSUM [5] is a highly abstractive dataset whose summaries are all expressed in a short sentence.

Implementation Details. We adopt a randomly initialized 2-layer transformer-encoder in the attention-prediction model, and the structure of each layer is the same as the BART-encoder layer. The optimizer is the Adabelief-optimizer [37] with eps $1e-16$, betas $(0.9, 0.999)$, weight decay $1e-4$ and learning rate $2e-5$. The attention-prediction model is trained on the training set for about $50,000$ steps, and checkpoints are saved per $10,000$ steps to select the best checkpoints on the development set. Since the attention prediction is slightly different from common sequence tagging tasks, we have summarized two notable points after several attempts – the dropout rate should be 0, and a low learning rate is preferred. All experiments are conducted on 3 *RTX 6000*. We include the global-aware inference in the generation code of *HuggingFace transformers* [32]. At the time of evaluation, ROUGE-1, ROUGE-2 & ROUGE-L (R-1, R-2 & R-L) scores [19] are computed from the ROUGE code[2] used by BART [16].

Hyper-parameter Selection. Although the global-aware inference requires two new hyper-parameters γ and β, some original hyper-parameters of beam search, namely length penalty, minimum and maximum length, are omitted. The searching scopes of β and γ are in $\{2, 4, 6, 10, 12, 15, 18, 20\}$ and $\{0, 0.5, 1, 1.5, 2\}$, respectively. According to the numerical tests on the development set, we finally choose $\beta = 12,\ \gamma = 1.5$ for CNN/DM and $\beta = 4,\ \gamma = 0$ for XSUM. Since limited improvement could be observed from larger γ's, we recommend $\gamma = 1$ for normal or longer targets. When testing the global-aware inference with PEGASUS [35], we directly use **empirical hyper-parameters** for each dataset, namely $\beta = 4,\ \gamma = 0$ for XSUM and $\beta = 12,\ \gamma = 1$ for others. The beam size $\mathcal{K}$ follows the setups in BART [16] and PEGASUS [35].

4.2 Results

Comparison with Beam Search. Beam search is a hard-to-beat baseline which has stood the test of time and has proven its superiority in practice for countless times [20]. In Table 2, we compare

[2] https://github.com/pltrdy/files2rouge

Table 2: ROUGE F_1 scores of summaries generated by global-aware, in comparison to beam search with length regularizations. Notably, global-aware uses empirical hyper-parameters.

$\mathcal{K} = 8$	XSUM R-1	XSUM R-2	XSUM R-L	BillSum R-1	BillSum R-2	BillSum R-L	Multi-News R-1	Multi-News R-2	Multi-News R-L	WikiHow R-1	WikiHow R-2	WikiHow R-L
Beam search [35]	47.05	24.53	39.33	57.00	39.65	52.70	47.29	18.91	43.31	41.86	19.04	40.40
Global-aware	**47.33**	**24.66**	**39.50**	**58.66**	**40.12**	**53.96**	**47.95**	**19.08**	**43.93**	**42.82**	**19.68**	**41.43**

	Reddit TIFU R-1	Reddit TIFU R-2	Reddit TIFU R-L	NewsRoom R-1	NewsRoom R-2	NewsRoom R-L	PubMed R-1	PubMed R-2	PubMed R-L	arXiv R-1	arXiv R-2	arXiv R-L
Beam search [35]	27.55	8.67	22.12	42.05	29.88	38.70	44.25	19.19	41.11	43.82	16.75	39.28
Global-aware	**28.31**	**9.13**	**23.30**	**44.68**	**31.71**	**41.28**	**45.78**	**20.16**	**42.62**	**44.92**	**17.41**	**40.31**

our global-aware inference to beam search with length regularizations (i.e., α in Eq. 6, accompanied with two hard constraints, namely minimum length and maximum length). We strictly implement the hyper-parameter setups of PEGASUS [35] in terms of beam search, while we only adopt empirical hyper-parameters for our method. Even so, significant improvements can be observed on all the data sets, especially when the summary is of normal or longer length.

Table 3: ORACLE and ablation results.

$\mathcal{K} = 4$	CNN/DM R-1	CNN/DM R-2	CNN/DM R-L
ORACLE global-aware	**51.85**	**28.13**	**48.68**
-w/o length reward	50.46	27.53	47.43
Global-aware	**45.13**	**21.77**	**42.04**
-w/o length reward	44.39	21.58	41.41
-w/o global score	44.12	21.29	40.91

$\mathcal{K} = 6$	XSUM R-1	XSUM R-2	XSUM R-L
ORACLE global-aware	**49.50**	26.24	41.13
-w/o length reward	48.92	**26.48**	**41.45**
Global-aware	45.44	22.15	37.11
-w/o length reward	**45.57**	**22.60**	**37.61**
-w/o global score	45.23	21.88	36.73

Table 4: Comparison with other methods.

	CNN/DM R-1	CNN/DM R-2	CNN/DM R-L
Beam search [16]	44.12	21.21	40.89
+ Coverage	44.74	21.69	41.48
+ Repetition penalty [12]	44.11	21.14	40.87
+ Attention masking [1]	**45.54**	**22.24**	**42.44**
Global-aware	45.13	21.77	42.04

	XSUM R-1	XSUM R-2	XSUM R-L
Beam search [16]	45.38	22.32	37.15
+ Coverage	44.54	21.82	36.97
+ Repetition penalty [12]	45.40	22.31	37.13
+ Attention masking [1]	45.35	22.31	37.15
Global-aware	**45.57**	**22.60**	**37.61**

ORACLE Results and Ablation Study. ORACLE refers to the global-aware inference combined with (true) global attention distribution instead of predicted global attention distribution. The related results have been presented in Table 3, and the significant boosting certifies that the proposed method could improve beam search with the global attention distribution. On the other hand, we conduct ablation experiments on ORACLE and global-aware, and both results indicate length reward plays an important role in generating normal-length text but causes adverse effect on generating text of very short length. Besides, the performance declines significantly when only length reward is applied, due to the fact that sole length reward can not calibrate beam search stepwise.

Comparison with Other Attention Approaches. In this part, we focus on comparing other approaches which also exploit certain attention distributions to improve beam search. The first is the coverage penalty [33, 18]. To enhance its performance in summarization, we replace its preset attention distribution with our predicted global attention distribution. Note that the coverage function can only evaluate the generated sentence at the terminal step. Instead of comparing the global-aware inference to the methods [25, 17, 7] that aim to reduce repetition using the dynamic attention distribution, we compare our algorithm with the CTRL repetition penalty [12] which has similar motivation but is more systematical and independent of training. Table 4 lists the comparison results against different algorithms. It can be observed that our global-aware approach can improve the performance of beam search stably and fundamentally. We also observe that attention head masking [1] appears to outperform the global-aware approach on CNN/DM, but it fails to gain any improvement on XSUM. To further show the advantage of the proposed approach, we will take a closer examination on the attention head masking [1] and our proposed approach in the next part.

Further Comparison with Attention Head Masking. First, one should bear in mind that attention head masking [1] acts on the model instead of beam search, which is opposite to us. Specifically, it selects contents during decoding by disabling partial attention heads for unimportant tokens to

decrease the attention to these tokens. According to the reported results presented in Table 4, we can see that although attention masking achieves amazing results on CNN/DM, it does fail completely on XSUM. Since hiding unimportant tokens from some heads results in the loss of context information of salient tokens, this would lead to its instability. Thus, it could be ineffective for tasks that require contextual information of the whole source such as XSUM. Taking a further comparison, we deploy the attention head combinations selected for CNN/DM and XSUM to examine its effect on PEGASUS [35]. These comparison results are shown in Table 5, implying that our method enjoys a robust feature that is able to boost summary inference on various datasets and models even with the same set of hyper-parameters. In contrast, attention masking [1] behaves much sensitive to the changes of models and datasets. Besides, attention masking has to construct its training saliency labels based on the longest common subsequences between a reference summary and the source. This may be hardly achieved in some text generation tasks (e.g. translation) where no common subsequence exists at all. Such drawback presents one main limitation for attention masking.

Table 5: Improvements of attention head masking and global-aware on beam search [35] in terms of ROUGE-L F_1 score. Both use empirical setups.

	XSUM	**BillSum**	**Multi-News**	**WikiHow**	**Reddit**	**NewsRoom**	**PubMed**	**arXiv**
Attention head masking [1]	-0.31	0.23	0.34	0.10	-0.16	1.24	0.35	0.21
Global-aware	**0.17**	**1.26**	**0.62**	**1.03**	**1.18**	**2.58**	**1.51**	**1.03**

Table 6: Generate CNN/DM summaries with XSUM's style.

R1/R2/RL	**Shorter beam search**	**Global-aware†**
F_1 **score**	**43.63/20.95**/40.35	**43.62**/20.38/**40.58**
Recall	**48.93/23.40/45.22**	40.36/18.81/37.55
Precision	41.40/19.90/38.30	**50.52/23.76/47.01**

Table 7: Generate summaries with corrupted attention distributions.

	R-1	**R-2**	**R-L**
Beam search	27.00	12.21	23.88
Global-aware (10K)	28.78	12.82	25.51
Global-aware (100K)	**29.59**	**13.72**	**26.30**

Robustness. We intend to examine the robustness of our proposed global-aware algorithm in this part. To do so, we substitute the parameters in CNN/DM's attention prediction model to XSUM's to create more abstractive summaries for CNN/DM. For comparison, we set the minimum length of beam search as 0 to allow it to generate shorter summaries. Table 6 shows the F_1 score, recall and precision of the shorter beam search and global-aware†. It is surprising that even by using the attention distribution from a different dataset with distinct properties, the proposed global-aware mechanism still manages to generate meaningful summaries with competitive F_1 scores, proving the robustness of this algorithm. Moreover, the higher Precision and lower Recall of global-aware suggest that although information is partially lost, the algorithm still summarizes core information in a concise format, compared to the standard beam search. On the other hand, we exploit a BART model fine-tuned on CNN/DM to generate summaries of NewsRoom directly, and the ROUGE scores of beam search are shown in Table 7. Next, we randomly select 10K and 100K samples from the training set and use them to fine-tune the attention-prediction model, where the global-aware improves beam search significantly. The experiment once again validates the robustness of the proposed inference algorithm, as it maintains a reasonably good performance even from learning a corrupted attention distribution from a BART of no fine-tuning.

5 Discussion

Degradation of Beam Search. It is widely known that text quality generally degrades with increased beam size [3, 21, 20]. The degradation originates from the gap between training and inference, where teacher forcing training only needs to ensure the local optimality, but beam search pursues global optimality [31, 24, 36]. The proposed algorithm reduces the gap by informing beam search how the global optimal hypothesis attends to the source so that beam search can be calibrated to pursue a more reasonable global optimum instead of that obtained by teacher forcing training. We enlarge the beam size and find that the global attention distribution can well guide beam search to overcome degradation.

According to Figure 2, the average ROUGE F_1 score of beam search summaries declines as beam size increases. Even if there is a prediction bias, the global-aware inference manages to lift ROUGE

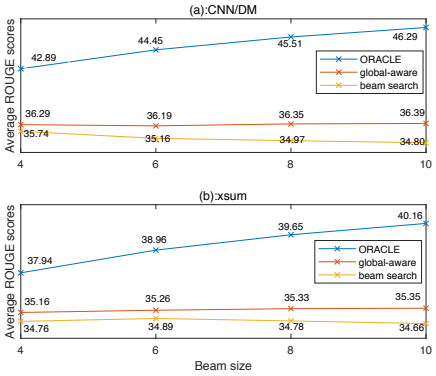

Figure 2: Degradation of beam search.

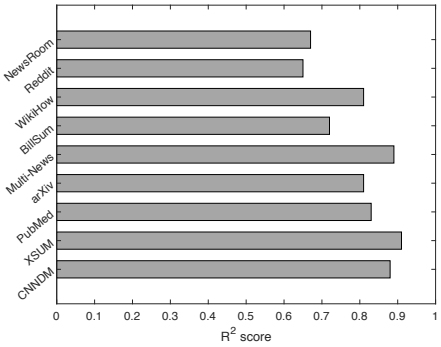

Figure 3: Fitting of att-prediction model.

scores with increased beam sizes. Moreover, a significant boosting in ROUGE scores is observed as beam size increases in the ORACLE global-aware inference, where the true potential of the algorithm is reflected.

Attention-prediction Model Fitting. We randomly select $1,000$ test samples from each dataset and plot their average R^2 scores of the predicted optimal attention distributions in Figure 3. For each sample, its R^2 score is equal to $1 - \frac{\sum_{i=1}^{n}(o_i - \hat{o}_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2}$, where $R^2 \in [-\infty, 1]$. A higher R^2 signifies better model fitting, while $R^2 < 0$ means the model is entirely invalid. More discussions are placed in the supplementary document (App. D).

6 Related Work

In the field of text generation, efforts have been made to boost beam search with regards to the attention distribution. For instance, some studies engage the attention distribution to penalize the situation where sources fail to be fully covered in translation tasks [33, 18], while others [7, 25, 17] incorporate dynamic attention distributions to evade tokens that have been highly regarded to reduce repetition. However, none of the aforementioned studies attempt to apply the global attention distribution to acquire the knowledge that the level of attention should be allocated to each token from the reference. Further, the existing penalty functions used by those studies are rather different from the proposed global score. More details can be seen in the supplementary document (App. E).

In addition to the attention distribution, other techniques are developed to improve beam search in terms of length bias [34], diversity-less [28], vapidity [10] and degradation [21, 3]. These methods are not included for comparison because they are not suitable for summarization [34, 10], or do not aim to enhance beam search as the main purpose [28, 21, 3]. Besides, we argue that these patches to beam search are supposed to be hard for improving the performance stably and fundamentally (as shown in Table 4 given by our global-aware method) because they fail to specify what a final hypothesis should look like and are easy to trap into local optimums.

7 Conclusion

Beam search tends to fall into local optimums due to the lack of global information during inference. To calibrate beam search on the premise of global awareness, this study proposes a predictable protocol to stipulate how a global optimal hypothesis should attend to source tokens. By training a simple prediction model of the global attention distribution, a novel global scoring function is then developed to regulate the inference on a step-wise basis. Our experiment shows that the proposed global-aware beam search generates higher-quality summaries, relative to beam search and other counterparts of the similar nature. Besides, the proposed algorithm is proven robust in generating meaningful texts even with corrupted attention distribution, implying its potential to cooperate with user-defined global attention distributions. We plan to focus our future study on generalizing the global-aware inference to a broader range of text generation tasks, including not only text2text but also image caption [29], graph2text [13] and data2text [22].

References

[1] Shuyang Cao and Lu Wang. Attention head masking for inference time content selection in abstractive summarization. ArXiv, 2104.02205, 2021.

[2] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018.

[3] Eldan Cohen and Christopher Beck. Empirical analysis of beam search performance degradation in neural sequence models. In International Conference on Machine Learning, pages 1290–1299, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[5] 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 13063–13075, 2019.

[6] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics.

[7] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. Bottom-up abstractive summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4098–4109, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[8] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[9] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, page 1693–1701, Cambridge, MA, USA, 2015. MIT Press.

[10] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2019.

[11] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015.

[12] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation, 2019.

[13] Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. Text generation from knowledge graphs with graph transformers. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2284–2293, 2019.

[14] Anastassia Kornilova and Vladimir Eidelman. BillSum: A corpus for automatic summarization of US legislation. In Proceedings of the 2nd Workshop on New Frontiers in Summarization, pages 48–56, Hong Kong, China, November 2019. Association for Computational Linguistics.

[15] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. arXiv, 1810.09305, 2018.

[16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin 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, Online, July 2020. Association for Computational Linguistics.

[17] Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. Improving neural abstractive document summarization with structural regularization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4078–4087, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[18] Yanyang Li, Tong Xiao, Yinqiao Li, Qiang Wang, Changming Xu, and Jingbo Zhu. A simple and effective approach to coverage-aware neural machine translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 292–297, 2018.

[19] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[20] Clara Meister, Ryan Cotterell, and Tim Vieira. If beam search is the answer, what was the question? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2173–2185, 2020.

[21] Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation. In International Conference on Machine Learning, pages 3956–3965, 2018.

[22] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with content selection and planning. In Proceedings of the AAAI conference on artificial intelligence, pages 6908–6915, 2019.

[23] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 2019.

[24] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks, 2016.

[25] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017.

[26] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[28] Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[29] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[30] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In Proceedings of the Workshop on New Frontiers in Summarization, pages 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[31] Sam Wiseman and Alexander M. Rush. Sequence-to-sequence learning as beam-search optimization. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1296–1306, Austin, Texas, November 2016. Association for Computational Linguistics.

[32] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[33] Y. Wu, M. Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, M. Krikun, Yuan Cao, Q. Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, M. Johnson, X. Liu, Lukasz Kaiser, S. Gouws, Y. Kato, Taku Kudo, H. Kazawa, K. Stevens, G. Kurian, Nishant Patil, W. Wang, C. Young, J. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, G. Corrado, Macduff Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. ArXiv, abs/1609.08144, 2016.

[34] Yilin Yang, Liang Huang, and Mingbo Ma. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3054–3059, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[35] Jingqing Zhang, Y. Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In ICML, 2020.

[36] Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. Bridging the gap between training and inference for neural machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4334–4343, Florence, Italy, July 2019. Association for Computational Linguistics.

[37] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. Conference on Neural Information Processing Systems, 2020.

Checklist

For all authors

- *Question*: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
- *Answer*: Yes. Theoretically, this paper proposes to use the global attention distribution as a protocol that allows step-wise calibration of Beam search in a more global optimal way, at the same time develops a global scoring function that can be considered as the probability that generated texts correctly attend to sources during the inference (See Section § 3). On the empirical side, we show that the proposed global-aware inference is reliable and robust, boosting two state-of-the-art summarization models to produce higher quality summaries on 9 datasets (See Section § 4).

- *Question*: Have you read the ethics review guidelines and ensured that your paper conforms to them ?
- *Answer*: Yes.

- *Question*: Did you discuss any potential negative societal impacts of your work ?
- *Answer*: Since the proposed method can be potentially adopted in seq-to-seq text generation tasks, it may be used to create Deepfakes or other disinformation.

- *Question*: Did you describe the limitations of your work?
- *Answer*: Yes. See Section § 4.2 – as shown in results, the improvement is relatively insignificant when using global-aware inference to generate very short text, e.g. XSUM.

For authors including theoretical results

- *Question*: Did you state the full set of assumptions of all theoretical results?
- *Answer*: N/A, there is no assumption for the proposed theorem.

- *Question*: Did you include complete proofs of all theoretical results?
- *Answer*: Yes, available in the supplementary material.

For authors running experiments

- *Question*: Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
- *Answer*: Yes, available in the supplementary material.

- *Question*: Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
- *Answer*: Yes. See Section § 4.1.

- *Question*: Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
- *Answer*: No. Following previous methods [16, 35], error bars are not reported.

- *Question*: Did you include the amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
- *Answer*: Yes. See Section § 4.1.

For authors using existing assets

- *Question*: If your work uses existing assets, did you cite the creators?
- *Answer*: Yes. See Section § 4.1.

- *Question*: Did you mention the license of the assets?
- *Answer*: No, they are all public.

- *Question*: Did you include any new assets either in the supplemental material or as a URL?
- *Answer*: Yes, available in the supplementary material.

- *Question*: Did you discuss whether and how consent was obtained from people whose data you're using/curating?
- *Answer*: No.

- *Question*: Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
- *Answer*: No.