Polishing Network for Decoding of Higher-Quality Diverse Image Captions

Yue Zheng[1, 2]
zhengy17@mails.tsinghua.edu.cn
Ya-li Li[1, 2]
liyali13@tsinghua.edu.cn
Shengjin Wang[1, 2]
wgsgj@tsinghua.edu.cn

[1] Department of Electronic Engineering
Tsinghua University
Beijing, China

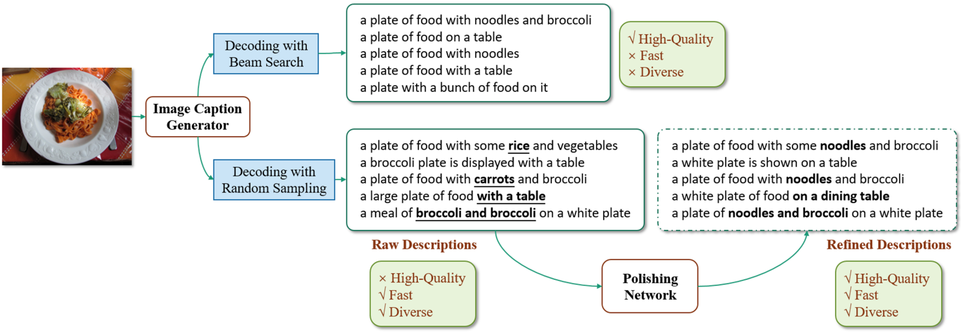
[2] Beijing National Research Center for
Information Science and Technology
Beijing, China

Abstract

Diverse image caption generation has attracted more attention in recent researches. Existing methods usually adopt a single-pass decoding process, that the sampled words at each time step during decoding will not be modified. A mistaken word could affect the whole subsequent sequence. On the other hand, decoders in single-pass approaches only have access to the previously generated words, thus unable to compose the sentences with an understanding of the whole contents. Inspired by the multi-pass process of human generating descriptions, in this paper we propose a novel framework with a Polishing Network (PN) for decoding diverse image captions. PN refines the raw descriptions generated by an original diverse image caption generation model. The refined sentences could modify some of the incorrect words and phrases in the raw descriptions, while still describing similar content. We also propose a novel approach for training PN. The raw-refined caption pairs used as training samples for PN are obtained by sampling both the input and output words of an original model during decoding. The experimental results show that the proposed approach can generate high-quality diverse image captions, achieving a better quality-diversity trade-off. We compare the performance of our method with several existing methods in the diverse image caption generation task. The proposed method achieves the state-of-the-art performance with oracle BLEU-4/CIDEr scores of 0.534/1.709 at sample size 20 on the MS COCO dataset.

1 Introduction

Humans can describe an image with diverse expressions. In recent researches [3, 25, 33, 38, 39], diverse image caption generation has attracted more attention. For each given image in this task, a model is used to generate a set of descriptions which are diverse while related to the image. Methods for diverse descriptions can be roughly divided into two categories: by using different latent variables or control signals as input to the decoder; or by applying sampling methods to the decoding process, where the word at each time step is sampled according to its probability predicted by the model. Most existing diverse image caption generation methods follow a single-pass decoding process. When generating a sentence, the sampled words will appear in the final generation and will not be further modified.



a plate of food with noodles and broccoli
a plate of food on a table
a plate of food with noodles
a plate of food with a table
a plate with a bunch of food on it

a plate of food with some rice and vegetables
a broccoli plate is displayed with a table
a plate of food with carrots and broccoli
a large plate of food with a table
a meal of broccoli and broccoli on a white plate

a plate of food with some **noodles** and broccoli
a white plate is shown on a table
a plate of food with **noodles** and broccoli
a white plate of food **on a dining table**
a plate of **noodles and broccoli** on a white plate

Figure 1: In the task of diverse image caption generation, a set of descriptions obtained with beam search are usually of high quality and low diversity. While with random sampling methods, a set of descriptions with higher diversity can be generated with low calculation consumption. However, the quality of these descriptions are usually lower, with incorrect words and phrases appearing in the descriptions. In this paper, we propose a novel framework with a polishing network to refine the raw descriptions generated by an original model, thus generating a set of refined descriptions with higher-quality. For example, mistaken words **"rice"** and **"carrots"** in the figure can be refined as **"noodles"** by the polishing network.

This would bring two problems. First, when a mistaken word is sampled, it may affect the subsequent description. This problem is particularly prominent when decoding descriptions with random sampling methods, where low-quality words are frequently sampled. Second, in single-pass processes decoders only have access to the previously generated words, thus the decoders are unable to predict words with an understanding of the whole contents in a sentence. The situation is different when humans generate descriptions. Rough drafts are usually generated first by humans, then the drafts are refined to compose the final description. This process of polishing can correct mistaken words in the drafts, and generate each word according to the whole content. Inspired by the way humans generate descriptions, we apply a similar polishing process in generating diverse image captions.

In this paper, we propose a novel framework for diverse image caption generation with a multi-pass process, in which the diverse image captions generated by an original model are refined by a Polishing Network (PN). First, the original model is used to decode a set of raw diverse descriptions. Then PN encodes the raw descriptions and regenerates refined results. The refined descriptions remain describing similar contents with the raw descriptions, while some of the low-quality words and phrases can be modified. During the decoding process, PN has access to the whole raw description. This enables the model to predict words with an understanding of the whole sentence, thus has the potential to generate descriptions with higher quality. As shown in Fig. 1, mistaken words and phrases in the raw descriptions can be refined by the polishing network.

We also propose a novel approach for training PN in this paper. The descriptions obtained from PN should be similar to the corresponding raw descriptions in content to remain diversity. This requires raw-refined pairs of descriptions with similar contents for training PN. However, it is intractable to annotate a refined description for each lower-quality raw description generated by original models. Therefore, we propose a novel approach with two sampling modules to obtain the raw-refined pairs. Instead of annotating refined descriptions, raw descriptions are generated according to human labeled descriptions in this paper. To reduce the gap between the distributions of raw descriptions during training and inference, both the output and the input words of an original model are obtained with a random sampling process. The proposed approach is not restricted to a specific original model, and can be used as a plug-in for various diverse image caption generation methods.

We combine polishing network with multiple existing sampling methods for decoding of diverse image captions. The experimental results show that diverse descriptions with higher quality can be obtained by using PN. When combined with random sampling methods, quality of the generated diverse descriptions can be improved while maintaining good diversity and speed. We demonstrate both quantitative and qualitative results of our method. The proposed method is compared with existing diverse image caption generation models, achieving the state-of-the-art performance with oracle BLEU-4/CIDEr scores of 0.534/1.709 at sample size 20 on the MS COCO dataset [21].

2 Related Work

Diverse image caption generation. Image caption generation has been extensively studied in recent years [2, 6, 12, 16, 23, 42]. Improvements of the methods include image feature representation [2, 43], the structure of attention modules [2, 5, 12, 16, 23, 42], and the use of discrete machine translation evaluation scores as training objectives [22, 31, 44]. Many recent works have studied generating diverse image captions [3, 24, 25, 33, 38, 39, 40, 45]. Generative models are adopt in [3, 25, 33, 38]. AG-CVAE in [38] uses a VAE model for generating diverse descriptions. Seq-CVAE [3] further uses a sequence of latent variables during decoding. LNFMM [26] uses normalizing flows [9] to learn the distribution of the latent space. COS-CVAE [25] uses a factorized latent variable to leverage contextual diversity in the dataset. Part-of-Speech labels are used as control signals in POS [7]. Different from these methods, diverse image captions are generated and refined through a multi-pass process in our approach.

Sampling methods for diverse sequence generation. Methods [10, 15, 19, 35, 37] for sampling diverse sequences are leveraged in various NLP applications. Beam search (BS) based methods include the top-g BS [19] which limits the number of next words for each reserved sequence, the DBS [37] method which involves the difference between generated sentences as objectives. Random sampling based methods include top-s sampling [10, 29] in which only top s words can be sampled, and top-p sampling [15] in which only words with a probability greater than p can be sampled. We evaluate these sampling methods when combined with PN in our framework for diverse image caption generation.

Multi-pass language generation. In machine translation tasks, multi-pass methods have been proved efficient [41]. In image caption generation, [11] proposes a deliberate network to generate descriptions in a multi-pass process. Work [32] applies iterative editing on generated captions to obtain descriptions with high quality. These works for standard image caption generation cannot be directly applied to diverse image caption generation, because there is a large gap between distributions of diverse descriptions from a single-pass model during training and inference. To solve the problem, we propose a novel approach for obtaining the raw-refined description pairs to supervise the training of polishing networks in diverse image caption generation.

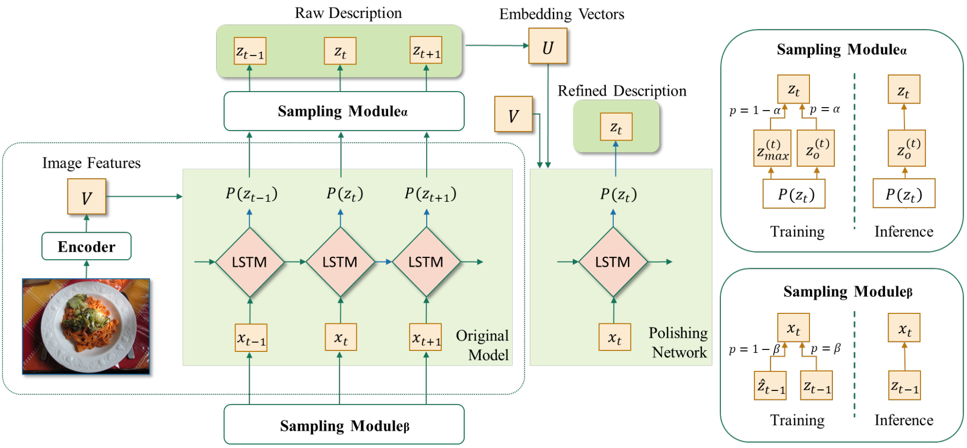


Figure 2: An overview of the proposed approach. During training, the input and output words of the original model are randomly sampled with the sampling modules. The raw-refined description pairs are then generated as training samples for the polishing network. Embedding vectors of a generated raw description are used as input to the polishing network to generate a refined description. During inference, a set of diverse raw descriptions is generated by the original model, then the polishing network refines each of the raw descriptions.

3 Approach

3.1 Diverse Image Caption Generation

In the task of diverse image caption generation, a model is supposed to generate a set of diverse sentences $\{\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_N}\}$ for each image I, where N is the size of the set. The sentence $\mathbf{z_n} = (z_1^{(n)}, z_2^{(n)}, ..., z_{T_n}^{(n)})$ of length T_n consists of a sequence of words $z_t^{(n)} \in \mathbf{C}$, where $\mathbf{C}$ is a word vocabulary. The encoder-decoder structure is widely used in image caption generation and methods in this paper are also based on this structure. In a common image caption generation method, the sentence $\mathbf{z_n}$ is usually generated one word at a time step by a decoder and the predicted probability $P(\mathbf{z_n})$ can be decomposed as

$$P(\mathbf{z_n}) = \prod_{t=1}^{T_n} P(z_t^{(n)} | \mathbf{z}^{(\mathbf{n})}_{[\mathbf{1:t-1}]}, \mathbf{V}, \Theta_o) \qquad (1)$$

where $\mathbf{z}^{(\mathbf{n})}_{[\mathbf{1:t-1}]} = (z_1^{(n)}, ..., z_{t-1}^{(n)})$ denotes the generated words at this time step t. The feature vectors $\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_M}\}$ are extracted by an encoder from the image I, where M is the number of the vectors. Θ_o denotes the parameters in the model, which will be omitted for brevity. A CNN [13, 34] or an object detector [30] is usually used as the encoder. Attention mechanism is widely used in image caption generation models, where the decoders generate a vector $\mathbf{a_t} = (a_1^{(t)}, a_2^{(t)}, .., a_M^{(t)})$ to re-weight different features at each time step.

$$a_m^{(t)} = Attention(\mathbf{v_m}, [\mathbf{s_{t-1}}; \mathbf{x_t}]); \sum_{m=1}^{M} a_m^{(t)} = 1 \ . \qquad (2)$$

The vector $\mathbf{s_{t-1}}$ is the previous state of the decoder, $\mathbf{x_t}$ is the input vector to the decoder at time step t, the operator $[;]$ denotes concatenated vectors. Index n is omitted for brevity.

Then an attended image feature $\mathbf{\tilde{v}_t} = \sum_{m=1}^{M} a_m^{(t)} \mathbf{v_m}$ is used for predicting the word at this time step with probability $P(z_t|\mathbf{s_{t-1}}, \mathbf{x_t}, \mathbf{\tilde{v}_t})$.

3.2 Polishing Network and Multi-pass Decoding

Different from single-pass generation process, we apply a polishing network to refine each description $\mathbf{z_n}$ in the generated set $\{\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_N}\}$ from an original model. We denote the refined descriptions from polishing network as $\{\mathbf{z_1^*}, \mathbf{z_2^*}, ..., \mathbf{z_N^*}\}$. The proposed polishing network adopts an encoder-decoder structure, encoding the raw description $\mathbf{z_n}$ from the original model as input, as shown in Fig. 2. An LSTM [14] is used as the decoder. The predicted probability of the refined description $\mathbf{z_n^*}$ is

$$P(\mathbf{z_n^*}) = \prod_{t=1}^{T_n^*} P(z_t^{*(n)}|\mathbf{z}^{*(\mathbf{n})}_{[\mathbf{1:t-1}]}, \mathbf{V}, \mathbf{z_n}, \Theta_p) \ . \tag{3}$$

T_n^* is the length of refined description $\mathbf{z_n^*}$. Θ_p denotes parameters in the polishing network and will be omitted for brevity. Each raw description $\mathbf{z_n}$ is embedded as a sequence of vectors $\mathbf{U} = (\mathbf{u_1}, \mathbf{u_2}, ..., \mathbf{u_{R_n}})$, where R_n equals the length T_n of the raw description $\mathbf{z_n}$. Position embedding is applied to vectors $\mathbf{u_r}$ to encode the position of the words [8]. Then an attention vector $\mathbf{a_t^*}$ is predicted by the decoder for the embedding vectors $\mathbf{u_r}$ as

$$a_r^{*(t)} = Attention^*(\mathbf{u_r}, [\mathbf{s_{t-1}^*}; \mathbf{x_t^*}]); \sum_{r=1}^{R} a_r^{*(t)} = 1 \ . \tag{4}$$

The decoder in the polishing network then uses the re-weighted image features $\mathbf{\tilde{v}_t}$ and embedding vectors of the input raw description $\mathbf{\tilde{u}_t} = \sum_{r=1}^{R} a_r^{*(t)} \mathbf{u_r}$ to predict the word in this time step with probability $P(z_t^*|\mathbf{s_{t-1}^*}, \mathbf{x_t^*}, \mathbf{\tilde{v}_t}, \mathbf{\tilde{u}_t})$. The decoder of the polishing network is able to access the whole raw description $\mathbf{z_n}$, thus it can predict the next word conditioned on both the predicted words $\mathbf{z}^{*(\mathbf{n})}_{[\mathbf{1:t-1}]}$ and future information contained in the input raw description $\mathbf{z_n}$. We use cross entropy of $P(\mathbf{z_n^*})$ and the ground truth descriptions as the objective function.

During training, the original model is treated as a black box for generating diverse raw descriptions $\{\mathbf{z_n}\}$. Raw-refined description pairs are needed as supervision to train the polishing network, while it is intractable to assign a refined description for each decoded raw description. Therefore, we propose a novel training approach based on sampling to generate the raw-refined description pairs for training PN, which is described in subsection 3.3.

3.3 Generating Training Samples for Polishing Network

Refined descriptions generated from a polishing network need to be of high quality while describing similarly with the raw descriptions, rather than becoming correct but irrelevant. This requires the input of PN being similar with the ground truth label during training. In order to obtain the raw-refined description pairs as training samples, we apply an opposite process that relevant raw descriptions are generated from the human labeled high-quality annotations. Then the raw descriptions and the corresponding ground truth descriptions can be used as raw-refined description pairs for training PN.

Specifically, two sampling modules are applied to the decoding process of raw descriptions during training. Similar to the training process of a normal image captioning model, a ground truth annotation $\mathbf{\hat{z}_n}$ is used as the input at each time step for the decoder of an original model, so as to ensure that the generated raw description $\mathbf{z_n}$ and the ground truth $\mathbf{\hat{z}_n}$

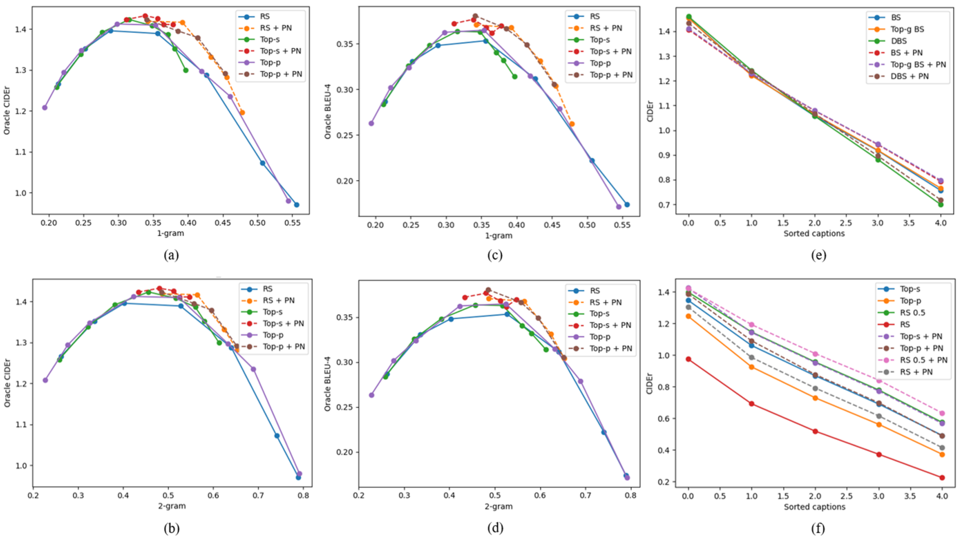


Figure 3: (a-d) Quality-diversity trade-off. Random sampling methods achieve better oracle CIDEr/BLEU-4 scores under same diversity scores when combined with PN . Each sampling method is evaluated with a varying temperature t_o from 0.1 to 1.2. (e-f) Overall quality across the whole set of descriptions. The generated descriptions are sorted with CIDEr score. The quality of descriptions from the random sampling based methods are improved considerably. It is worth noting that although the best-quality descriptions obtained by the beam search based methods are not improved by PN, the descriptions with lower quality for each image are improved. Thus the overall quality of descriptions can still be improved by PN, which is also reflected by the higher average quality scores shown in Tab. 1.

are related. Thus, the probability of the next word predicted by the original model becomes $P_o(z_t^{(n)}|\hat{\mathbf{z}}^{(\mathbf{n})}_{[\mathbf{1:t-1}]}, \mathbf{V})$ at each time step.

Firstly, we randomly sample an output $z_o^{(n)}$ according to the predicted $P_o(z^{(n)})$ at each time step. Then the word $z^{(n)}$ is sampled as

$$z^{(n)} = I_{\{y\leq\alpha\}} z_o^{(n)} + I_{\{y>\alpha\}} z_{max}^{(n)}; y \sim U(0,1) \; . \tag{5}$$

Subscript t is omitted for brevity. The word $z_o^{(n)} \sim P_o(z^{(n)})$ and $z_{max}^{(n)} = \arg\max P_o(z^{(n)})$, where $z_{max}^{(n)}$ is the word with the maximum probability. $I_{\{y\leq\alpha\}}$ is a characteristic function, which equals 1 when $y \leq \alpha$ and equals 0 otherwise. The hyperparameter α is used to control the probability of using the random sampling results. A larger α can augment the raw descriptions with more possible words while introducing larger noise. The generated descriptions tend to have higher quality and lower diversity when α is large.

While sampling on $P_o(z_t^{(n)})$ provides diversity of each independent word in the raw descriptions, there is still a gap between distribution of the generated raw descriptions in training and inference process. During inference, previously generated words are used as input to the original model. Therefore, we apply a sampling module to the generating process of inputs in the original model, in order to reduce the difference in the training and inference processes. With the sampling module, the input word x_t is sampled either from the predicted

#	Method	Oracle		Average		Top-one	
		B4	C	B4	C	B4	C
1	Arg-max	0.337	1.100	0.337	1.100	0.337	1.100
	Arg-max+PN	**0.338** (0.001)	**1.104** (0.004)	**0.338** (0.001)	**1.104** (0.004)	**0.338** (0.001)	**1.104** (0.004)
5	Top-g BS [19]	**0.378**	**1.454**	0.224	1.085	**0.356**	**1.125**
	Top-g BS+PN	0.366 (-0.012)	1.412 (-0.043)	**0.227** (0.003)	**1.092** (0.008)	0.353 (-0.003)	1.116 (-0.009)
	DBS [37]	**0.380**	**1.460**	0.207	1.068	**0.358**	**1.134**
	DBS+PN	0.377 (-0.003)	1.433 (-0.027)	**0.211** (0.004)	**1.071** (0.002)	0.355 (-0.003)	1.126 (-0.008)
	RS	0.177	0.975	0.050	0.556	0.189	0.720
	RS+PN	**0.310** (0.133)	**1.305** (0.330)	**0.110** (0.059)	**0.822** (0.266)	**0.246** (0.057)	**0.911** (0.191)
	Top-p [15]	0.285	1.247	0.099	0.767	0.253	0.889
	Top-p+PN	**0.354** (0.070)	**1.387** (0.141)	**0.142** (0.044)	**0.908** (0.141)	**0.283** (0.030)	**0.972** (0.083)
	Top-s [10]	0.330	1.348	0.134	0.892	0.294	0.995
	Top-s+PN	**0.374 (0.043)**	**1.425** (0.077)	**0.167** (0.034)	**0.972** (0.080)	**0.312** (0.018)	**1.035** (0.040)
10	Top-g BS [19]	**0.461**	**1.613**	0.215	1.059	**0.352**	**1.114**
	Top-g BS+PN	0.435 (-0.026)	1.553 (-0.060)	**0.220** (0.006)	**1.074** (0.0143)	0.349 (-0.003)	1.104 (-0.010)
	DBS [37]	**0.454**	**1.582**	0.184	1.009	**0.348**	**1.110**
	DBS+PN	0.450 (-0.004)	1.564 (-0.018)	**0.191** (0.007)	**1.023** (0.014)	0.347 (-0.000)	1.101 (-0.009)
	RS	0.253	1.139	0.049	0.554	0.193	0.751
	RS+PN	**0.405** (0.151)	**1.463** (0.325)	**0.108** (0.059)	**0.812** (0.258)	**0.246** (0.053)	**0.901** (0.150)
	Top-p [15]	0.372	1.407	0.099	0.774	0.268	0.957
	Top-p+PN	**0.452** (0.080)	**1.552** (0.145)	**0.145** (0.046)	**0.914** (0.139)	**0.290** (0.022)	**1.020** (0.063)
	Top-s [10]	0.428	1.515	0.131	0.889	0.302	1.032
	Top-s+PN	**0.453** (0.025)	**1.576** (0.061)	**0.164** (0.033)	**0.969** (0.079)	**0.308** (0.005)	**1.045** (0.013)

Table 1: Scores of quality metrics using the m-RNN test split on MS COCO dataset. When combined with PN, random sampling based methods (**RS**, **top-p** and **top-s**) achieve better results on oracle/average/top-one scores. Note that for beam search methods (**top-g BS** and **DBS**), the oracle/top-one scores drop slightly, which is reasonable since the BS methods search for the best-quality descriptions for an image and the oracle/top-one scores only measure the best descriptions in a generated set. It is worth noting that, descriptions with lower quality in beam search can be improved by PN, that the average scores are slightly improved. This indicating an overall improvement of quality of the generated descriptions. **Arg-max** denotes the results with greedy sampling. Improvement of scores by using PN are in parentheses.

word $z_{t-1}^{(n)}$ or from the ground truth word $\hat{z}_{t-1}^{(n)}$ at last time step.

$$x_t^{(n)} = I_{\{y \leq \beta\}} z_{t-1}^{(n)} + I_{\{y > \beta\}} \hat{z}_{t-1}^{(n)}; y \sim U(0,1)\ . \tag{6}$$

The hyperparameter β is used to control the proportion of the generated words used as the input. A larger β can better diminish the gap between raw descriptions in training and inference. On the other hand, a large β would make the generated results deviate from the ground truth, which will cause the learned polishing network no longer focusing on the input raw descriptions, generating high-quality but irrelevant results. This could harm the diversity of the refined descriptions. Experimental results show that by selecting an appropriate setting of the hyperparameters α and β, the learned polishing network can improve the quality of the raw descriptions while maintaining diversity.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate the proposed method on the MS COCO dataset [21]. Each image in the dataset is labeled with five human generated descriptions. Following previous works

Method	# Sample	B4	B3	B2	B1	C	R	M	S
DBS [37]	20	0.383	0.538	0.687	0.837	1.405	0.653	0.357	0.269
AG-CVAE [38]		0.471	0.573	0.698	0.834	1.259	0.638	0.309	0.244
POS [7]		0.449	0.593	0.737	0.874	1.468	0.678	0.365	0.277
PN (0.7)		**0.534**	**0.670**	**0.789**	**0.911**	**1.709**	**0.719**	**0.408**	**0.315**
Seq-CVAE [3]	20	0.445	0.591	0.727	0.870	1.448	0.671	0.356	0.279
PN (1.0)		**0.486**	**0.626**	**0.755**	**0.896**	**1.622**	**0.700**	**0.386**	**0.309**
DBS [37]	100	0.402	0.555	0.698	0.846	1.448	0.666	0.372	0.290
AG-CVAE [38]		0.557	0.654	0.767	0.883	1.517	0.690	0.345	0.277
POS [7]		0.550	0.672	0.787	0.909	1.661	0.725	0.409	0.311
PN (0.7)		**0.654**	**0.756**	**0.853**	**0.950**	**1.950**	**0.780**	**0.473**	**0.352**
Seq-CVAE [3]	100	0.575	0.691	0.803	0.922	1.695	0.733	0.410	0.320
LNFMM [26]		0.597	0.695	0.802	0.920	1.705	0.729	0.402	0.316
COS-CVAE [25]		0.633	0.739	0.842	0.942	1.893	0.770	0.450	0.339
PN (1.0)		**0.653**	**0.749**	**0.848**	**0.952**	**1.926**	**0.774**	**0.459**	**0.352**

Table 2: Scores of quality metrics at sample size 20 and 100 using the m-RNN test split on MS COCO dataset. PN (t_o) denotes results of random sampling with temperature t_o combined with PN. Results show that a better quality-diversity trade-off can be achieved with the proposed method.

Method	# Sample	Distinct	# Novel	mBLEU-4	1-gram	2-gram
DBS [37]	20	**100%**	3106	0.81	0.20	0.26
AG-CVAE [38]		69.8%	3189	0.66	0.24	0.34
POS [7]		96.3%	3394	0.64	0.24	0.35
PN (0.7)		90.9%	**3498**	**0.53**	**0.35**	**0.49**
Seq-CVAE [3]	20	94.0%	**4266**	0.52	0.25	0.54
PN (1.0)		**98.2%**	4224	**0.31**	**0.42**	**0.60**
DBS [37]	100	**100%**	3421	0.82	0.20	0.25
AG-CVAE [38]		47.4%	3069	0.70	0.23	0.32
POS [7]		91.5%	3446	0.67	0.23	0.33
PN (0.7)		90.5%	**3522**	**0.53**	**0.34**	**0.48**
Seq-CVAE [3]	100	84.2%	4215	0.64	0.33	0.48
LNFMM [26]		97.0%	**4741**	0.60	0.37	0.51
COS-CVAE [25]		96.3%	4404	0.53	0.39	0.57
PN (1.0)		**98.3%**	4218	**0.31**	**0.42**	**0.61**

Table 3: Diversity scores at sample size 20 and 100 using the m-RNN test split on MS COCO dataset.

[3, 25, 38, 45] in diverse image caption generation, we use the m-RNN split in [27] to train and evaluate our method, with 118287, 4000 and 1000 images for training, validation and testing.

Quality and diversity evaluation. Automatic evaluation metrics in machine translation are usually used to evaluate the quality of image captions, including BLEU (B) [28], METOER (M) [4], ROUGE (R) [20], CIDEr (C) [36], and SPICE (S) [1]. In the task of diverse image caption generation, the quality of a set of sentences are evaluated. **Oracle scores** of the metrics are most commonly used in this task, evaluating the sentences with the highest score of each metrics in each set. We also evaluate the **average scores** of the metrics across each set, in order to evaluate the overall quality of the generated descriptions. Following [25, 45], we also evaluate the **top-one scores**, which evaluate the quality of the most probable or rank first sentence in each set. Following previous works [3, 25, 38, 45], the diversity of results are evaluated with **Distinct** (higher: more diverse), **#Novel** (higher: more diverse), **mBLEU-4** (lower: more diverse), and **n-gram** (higher: more diverse).

Model implementation. In the following experiments, we use the top-down model in [2] as the original model. Image features are extracted with a Faster R-CNN [30] trained with attribute labels from the Visual Genome dataset [18]. The polishing network is built based on a top-down model and an encoder for the input raw descriptions. The polishing network

Method	PN	α	β	BS in PN	Oracle		Average		Diversity	
					BLEU-4	CIDEr	BLEU-4	CIDEr	1-gram	2-gram
RS+PN	✓	0.7	0.1	1	0.357	1.394	0.154	0.936	0.39	0.56
				2	0.371	1.411	0.161	0.947	0.39	0.56
				3	0.369	1.402	0.160	0.945	0.39	0.56
RS+PN	✓	0.7	0.0	2	0.367	1.424	0.156	0.945	0.40	0.57
			0.2		0.368	1.398	0.158	0.951	0.39	0.56
			0.3		0.368	1.426	0.161	0.962	0.38	0.55
			0.5		0.384	1.442	0.177	0.998	0.37	0.52
RS+PN	✓	0.3	0.1	2	0.348	1.379	0.151	0.927	0.40	0.58
		0.5			0.358	1.383	0.150	0.924	0.39	0.57
		1.0			0.367	1.426	0.161	0.962	0.39	0.55
RS+PN	✓	1.0	1.0	2	0.247	1.183	0.221	1.114	0.19	0.21
RS		-	-	-	0.315	1.307	0.121	0.845	0.43	0.64

Table 4: Effect of hyperparameters on the quality and diversity results. **BS in PN** denotes the beam size used during the decoding process of PN.

is trained with Adam algorithm [17] with a learning rate starting from 5e-4, batch size 32. The original model is trained with 100 epochs and the polishing networks are trained with 50 epochs. More details are in subsection A.1 of the appendix.

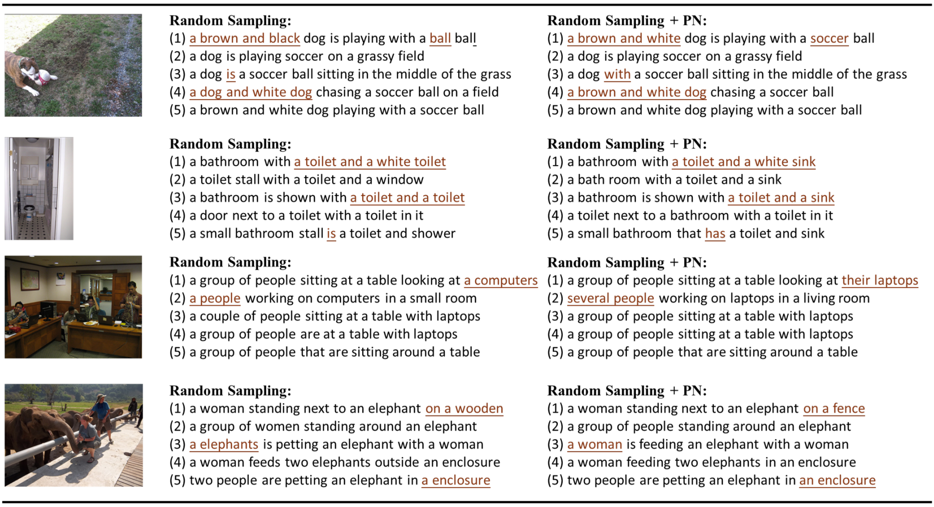
4.2 Results in Diverse Image Caption Generation Task

Combining PN with various sampling methods. We combine PN with multiple sampling methods for decoding diverse sequences, including beam search (BS) based methods (Top-g BS [19] and DBS [37]) and random sampling (RS) based methods (RS, Top-p Sampling [15], and Top-s Sampling [10, 29]). We evaluate the quality improvement from raw descriptions to corresponding refined descriptions. Results are shown in Tab. 1. Applying PN to decoding processes can significantly improve the quality scores of random sampling methods. In Fig. 3, we show the quality-diversity trade-off of generated descriptions. A better trade-off can be achieved when applying PN. It is worth noting that PN shows ability to improve the lower-quality descriptions for each image (in (e) and (f) of Fig. 3).

Comparing with existing diverse image captioning methods. We compare the proposed approach with existing methods for diverse image caption generation, including DBS [37], AG-CVAE [38], POS [7], Seq-CVAE [3], LNFMM [26], and COS-CVAE [25]. Following previous work, we evaluate the oracle scores of quality metrics and the diversity scores. We use random sampling (RS) as the original model and combine it with PN. According to the quality-diversity trade-off analysis, we set the temperature t_o to 0.7 and 1.0 for RS. 20 and 100 samples are generated for each image. A better quality-diversity trade-off can be achieved with the proposed approach comparing to existing methods (in Tab. 2 and 3).

Ablation studies. We analysis the effect of sampling modules in training PN. Results are evaluated with sample size 5. In Tab. 4, with larger hyperparameters α and β, generated descriptions tend to be higher in quality, while the diversity decreases. This is reasonable because when α and β increase, more words and phrases in the raw descriptions are assigned to the same ground truth labels, thus PN tends to generate a set of high-quality but similar descriptions during inference. We evaluate the influence of beam search on the decoding process of PN. The best results are obtained when decoding with a small beam size of 2. *RS* in Tab. 4 shows results without PN, the quality of generated descriptions decreases compared with results with PN.

Qualitative results. Examples of generated descriptions are shown in Fig. 4. In order to analyze the effect of PN on the lower-quality words in raw descriptions, we make a statistic



Random Sampling:
(1) a brown and black dog is playing with a ball ball
(2) a dog is playing soccer on a grassy field
(3) a dog is a soccer ball sitting in the middle of the grass
(4) a dog and white dog chasing a soccer ball on a field
(5) a brown and white dog playing with a soccer ball

Random Sampling + PN:
(1) a brown and white dog is playing with a soccer ball
(2) a dog is playing soccer on a grassy field
(3) a dog with a soccer ball sitting in the middle of the grass
(4) a brown and white dog chasing a soccer ball
(5) a brown and white dog playing with a soccer ball

Random Sampling:
(1) a bathroom with a toilet and a white toilet
(2) a toilet stall with a toilet and a window
(3) a bathroom is shown with a toilet and a toilet
(4) a door next to a toilet with a toilet in it
(5) a small bathroom stall is a toilet and shower

Random Sampling + PN:
(1) a bathroom with a toilet and a white sink
(2) a bath room with a toilet and a sink
(3) a bathroom is shown with a toilet and a sink
(4) a toilet next to a bathroom with a toilet in it
(5) a small bathroom that has a toilet and sink

Random Sampling:
(1) a group of people sitting at a table looking at a computers
(2) a people working on computers in a small room
(3) a couple of people sitting at a table with laptops
(4) a group of people are at a table with laptops
(5) a group of people that are sitting around a table

Random Sampling + PN:
(1) a group of people sitting at a table looking at their laptops
(2) several people working on laptops in a living room
(3) a group of people sitting at a table with laptops
(4) a group of people sitting at a table with laptops
(5) a group of people that are sitting around a table

Random Sampling:
(1) a woman standing next to an elephant on a wooden
(2) a group of women standing around an elephant
(3) a elephants is petting an elephant with a woman
(4) a woman feeds two elephants outside an enclosure
(5) two people are petting an elephant in a enclosure

Random Sampling + PN:
(1) a woman standing next to an elephant on a fence
(2) a group of people standing around an elephant
(3) a woman is feeding an elephant with a woman
(4) a woman feeding two elephants in an enclosure
(5) two people are petting an elephant in an enclosure

Figure 4: Examples of raw descriptions generated by random sampling ($t_o = 0.7$) and corresponding refined descriptions with PN. Words and phrases refined by PN are underlined. Mistaken descriptions such as **"a dog is a soccer ball"** can be modified as **"a dog with a soccer ball"**. Grammar errors such as **"a computers"** can be refined as **"their laptops"**.

Method	Top-g BS	DBS	RS	Top-p	Top-s
# **Raw**	38	45	166	114	110
# **Refined**	5	5	41	19	11

Table 5: Number of descriptions with bad endings appearing in the generated raw descriptions (# **Raw**) and the refined descriptions (# **Refined**).

on bad endings in the generated results. We count the number of descriptions ending with ('of', 'on' , 'in', 'with', 'a') in generated descriptions on the m-RNN test split. The results show that most of the counted bad endings can be corrected in the refined descriptions.

5 Conclusion

In this paper, a novel method is proposed for diverse image caption generation with a polishing network, which refines the generated results from an original single-pass method to obtain higher-quality descriptions. A novel training approach is also proposed to generate raw-refined description pairs for training the polishing network. Extensive experiments in diverse image caption generation show that the proposed approach can achieve a better quality-diversity trade-off of descriptions.

Acknowledgement

This work is supported by the state key development program in 14th Five-Year under Grant No. 2021YFF0602103, 2021YFF0602102, 2021QY1702. We also thank for the research fund under Grant No. 2019GQG0001 from the Institute for Guo Qiang, Tsinghua University. This study is also supported by Tsinghua University Toyota Joint Research Center for AI Technology of Automated Vehicle (TTAD 2022-07).

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[3] Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4261–4270, 2019.

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[5] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

[6] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017.

[7] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10695–10704, 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

[9] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[10] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, 2018.

[11] Lianli Gao, Kaixuan Fan, Jingkuan Song, Xianglong Liu, Xing Xu, and Heng Tao Shen. Deliberate attention networks for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8320–8327, 2019.

[12] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10327–10336, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019.

[16] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[19] Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*, 2016.

[20] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[22] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2017.

[23] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.

[24] Ruotian Luo and Gregory Shakhnarovich. Analysis of diversity-accuracy tradeoff in image captioning. *arXiv preprint arXiv:2002.11848*, 2020.

[25] Shweta Mahajan and Stefan Roth. Diverse image captioning with context-object split latent spaces. *Advances in Neural Information Processing Systems*, 33:3613–3624, 2020.

[26] Shweta Mahajan, Iryna Gurevych, and Stefan Roth. Latent normalizing flows for many-to-many cross-domain mappings. In *International Conference on Learning Representations*, 2019.

[27] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*, 2015.

[28] 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 of the Association for Computational Linguistics*, pages 311–318, 2002.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8):9, 2019.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 91–99, 2015.

[31] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[32] Fawaz Sammani and Luke Melas-Kyriazi. Show, edit and tell: a framework for editing image captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4808–4816, 2020.

[33] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27:3104–3112, 2014.

[36] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[37] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[38] Liwei Wang, Alexander G Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5758–5768, 2017.

[39] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4195–4203, 2019.

[40] Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. Diverse image captioning via grouptalk. In *IJCAI*, pages 2957–2964, 2016.

[41] Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Deliberation networks: Sequence generation beyond one-pass decoding. *Advances in neural information processing systems*, 30, 2017.

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[43] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018.

[44] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017.

[45] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *European Conference on Computer Vision*, pages 211–229. Springer, 2020.