Polishing Network for Decoding of Higher-Quality Diverse Image Captions

Yue Zheng, Ya-li Li, Shengjin Wang

Department of Electronic Engineering, Tsinghua University

Beijing National Research Center for Information Science and Technology

Abstract

Existing methods in diverse image caption generation 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.

Introduction

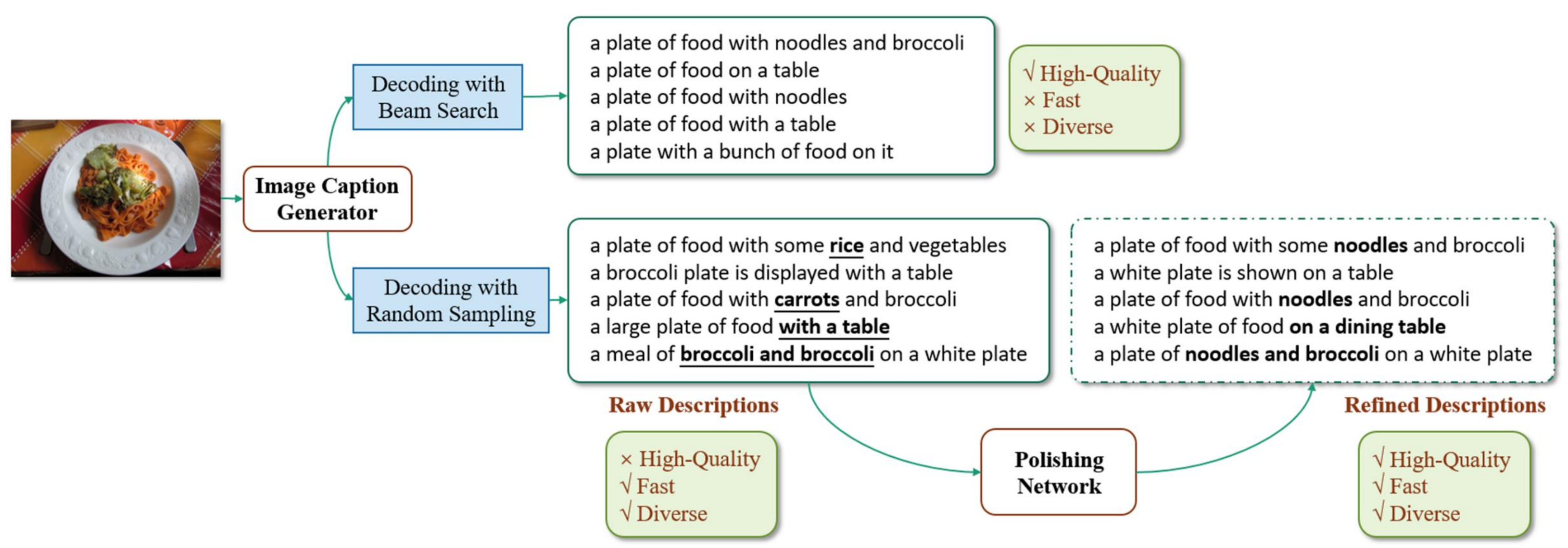

Figure 1: Introduction of our approach.

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.

Approach

Polishing Network and Multi-pass Decoding

Refine each description $\mathbf{z}_n$ in the original generated set $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$.

Refined descriptions $\{\mathbf{z}_1^*, \mathbf{z}_2^*, \ldots, \mathbf{z}_N^*\}$.

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

Predicted probability of the refined description $\mathbf{z}_n^*$ denoted as $\mathbf{P}(\mathbf{z}_n^*)$.

Approach

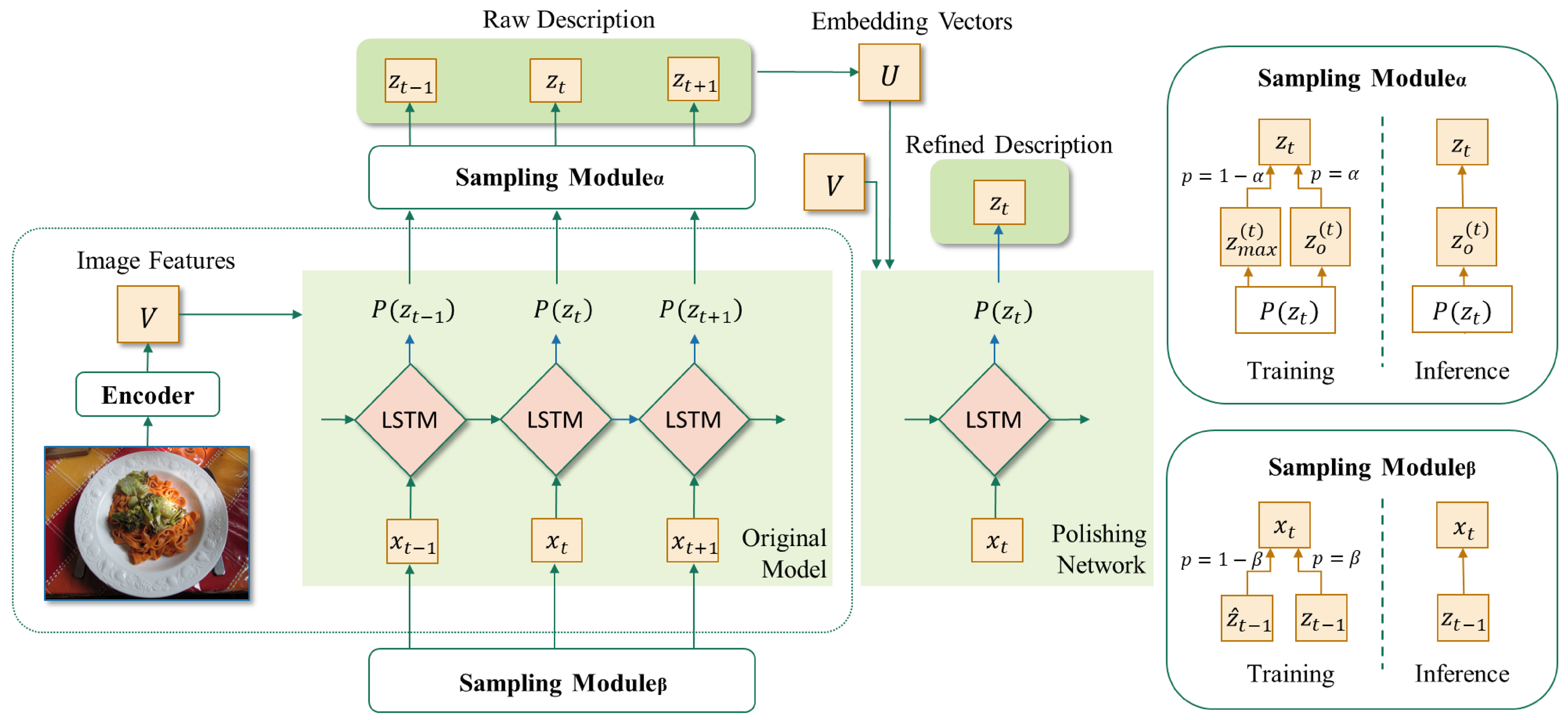

Figure 2: An overview of the proposed approach. During training, the input and output words of the original model are randomly sampled with the sampling modules. The raw-refined description pairs are then generated as training samples for the polishing network. Embedding vectors of a generated raw description 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.

Generating Training Samples for Polishing Network

Refined descriptions: high quality while **describing similarly with the raw descriptions**.

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

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

Two sampling modules α and β.

Results

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 1: Scores of quality metrics using the m-RNN test split on MS COCO dataset.

Quality and diversity scores in Tab. 1 and Tab. 2 show that better quality-diversity trade-off can be achieved with polishing network comparing with existing methods.

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 2: Diversity scores using the m-RNN test split on MS COCO dataset.

#	Method	Oracle B4	Oracle C	Average B4	Average C	Top-one B4	Top-one 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)

Table 3: Scores of quality metrics. When combined with PN, random sampling based methods can achieve better results on oracle/average/top-one scores.

Results

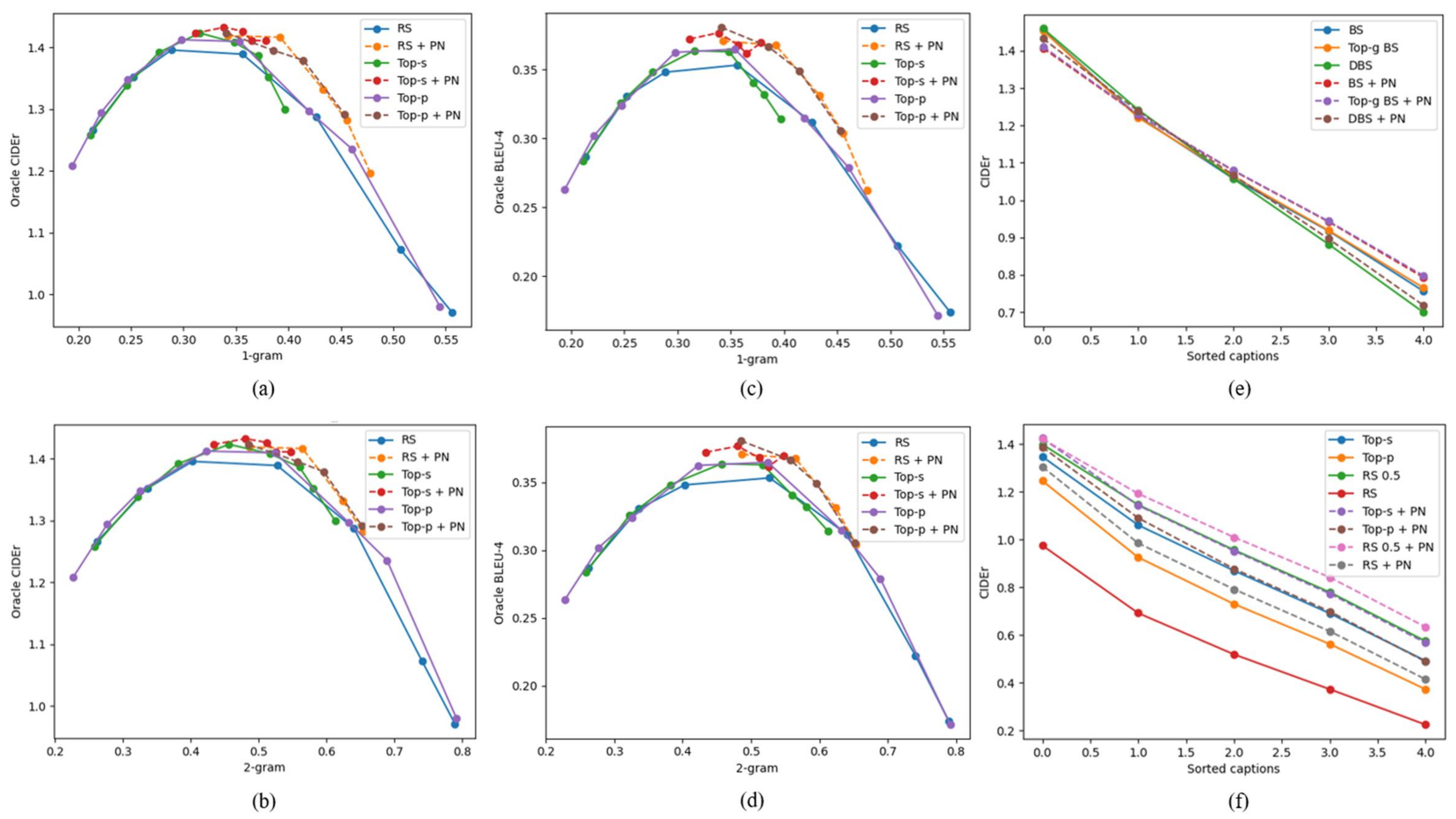

Figure 3: (a-d) Quality-diversity trade-off. Random sampling methods achieve better oracle CIDEr/BLEU-4 scores under same diversity scores when combined with PN. (e-f) Overall quality across the whole set of descriptions. The quality of descriptions from the random sampling based methods can be improved. 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.

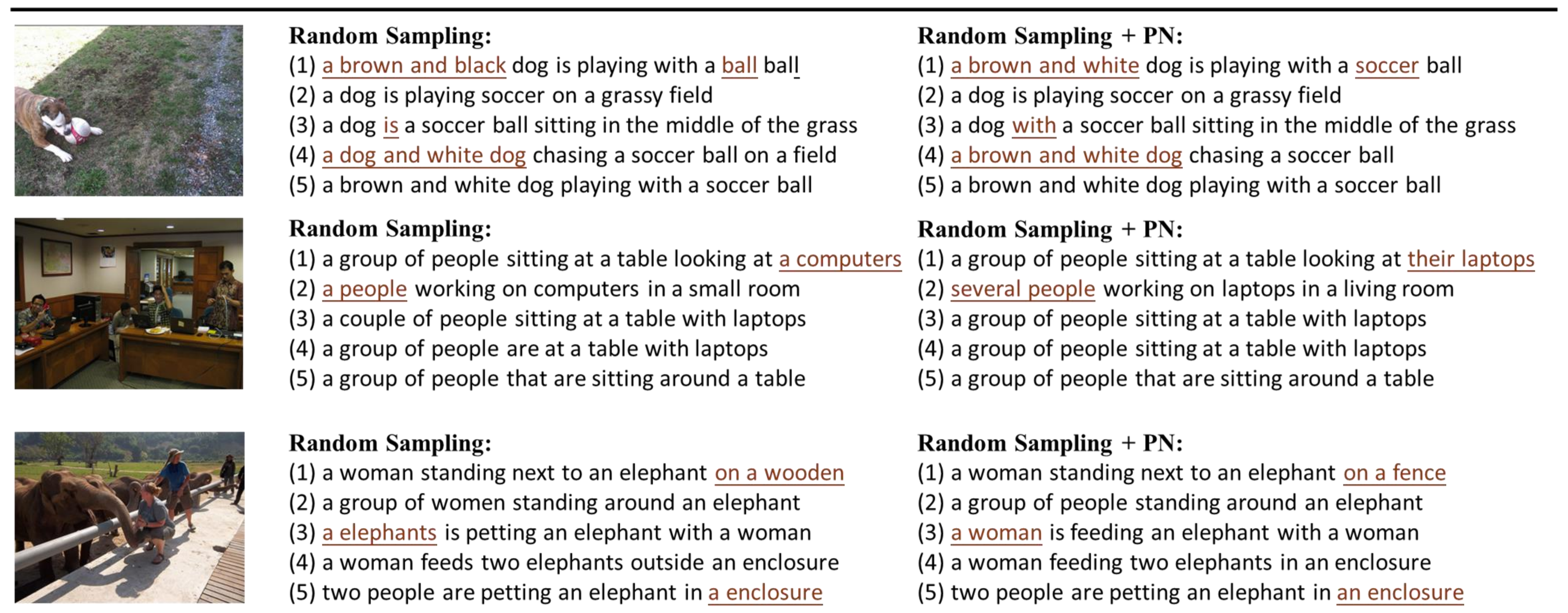

Figure 4: Examples of raw descriptions and corresponding PN refined descriptions. Words and phrases refined by PN are underlined. Mistaken descriptions such as ''a dog is a soccer ball'' can be modified as ''a dog with a soccer ball''. Grammar errors such as ''a computers'' can be refined as ''their laptops''.

Conclusion

We proposed a novel approach 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. Experiments in diverse image caption generation show that the proposed approach can achieve a better quality-diversity trade-off of descriptions.

Contact: zhengy17@mails.tsinghua.edu.cn, {liyali13, wgsgj} @tsinghua.edu.cn