Supplementary Materials of “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

A Appendix

A.1 Model Details

In the experiments, we use the top-down model in [1] as the original model. Image features are extracted with a Faster R-CNN [9] trained with attribute labels from the Visual Genome dataset [5]. The polishing network consists of a top-down model and an encoder for the input raw descriptions. The embedding vector of raw descriptions has a dimension of 512.

During training. The hyperparameters α and β in the sampling modules of the original model is set to 0.7 and 0.1 respectively, chosen according to the studies in Tab. 4 in the paper. Models are trained with Adam algorithm [4] with a learning rate starting from $5e-4$, batch size 32. The original model is trained with 100 epochs and polishing networks are trained with 50 epochs. Both the original model and the polishing networks are trained on a GeForce RTX 2080 GPU.

During inference. The polishing network is decoded using beam search with a small beam size of 2.

A.2 Statistics of the Error Endings

Discussion. We count the number of descriptions ending with words (‘of’, ‘on’ , ‘in’, ‘with’, ‘a’) from different encoding methods on the m-RNN test split [7] on MS COCO dataset [6] in Tab. 5 in the paper. Here we will discuss the statistics in more detail. In a normal image caption generation process, we can manually filter these bad ending words through an additional post-processing process. However, this post-processing requires a lot of manual work. First, bad ending words are not limited to those we count. Moreover, similar errors in descriptions do not only appear at the end of sentences. Repeated descriptions such as “there is a cat and a cat” are common errors. It is intractable to deal with all these errors with

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

manually defined rules. Therefore, in this paper, we forbid the post-processing and leave this work to PN. We use the statistics of bad ending words as a probe to show the effect of PN. To a certain extent this result can reflect the role of PN in correcting errors of generated descriptions.

A.3 More Examples of Generated Descriptions

More examples on the m-RNN test split of MS COCO dataset generated by random sampling methods (RS with temperature $t_o = 0.7$, top-s [2, 8] and top-p [3]) and corresponding refined descriptions by applying PN are shown in Fig. A1 and A2. Sample size is 5.

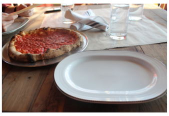
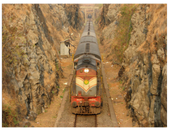
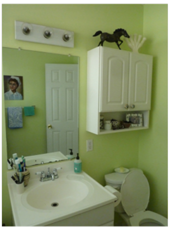
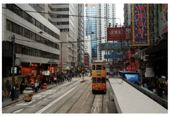
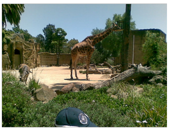
	**RS 0.7** a plate of pizza sitting on a table next to a beverage people standing at a table with a pizza and a glass a plate of a pizza on a table with a fork a pizza sitting on a table with a glass of water a plate of pizza sitting on top of a table **Top-*s*** a plate of pizza on a table with a glass of water a plate of pizza on a table with a person a table topped with pizza and plates on a table two slices of pizza are on on a table table a pizza is sitting on the plate on a table with a fork **Top-*p*** a couple of plates on pizza on a table a white pizza is on a table next to a glass a plate with a large pizzas on it a slice of pizza sitting on a table in front of a table a pizza on a white table with forks and a glass of water	**RS 0.7 + PN** a slice of pizza sitting on a table next to a glass of water someone sitting at a table with a pizza and a drink a plate with a pizza on a table with a fork a pizza sitting on a table with a glass of water a plate of pizza sitting on top of a table **Top-*s* + PN** a plate of pizza on a table with a glass of water a plate of pizza on a table with a person a table topped with pizza and drinks on a table two plates of pizza are sitting on a wooden table a pizza is sitting on a plate on a table **Top-*p* + PN** a couple of plates of pizza on a table a large pizza sitting on a table next to a glass of water a table with two different pizzas on it a slice of pizza sitting on a table in front of a table a pizza on a wooden table with glasses and a glass of water
	**RS 0.7** a train is traveling on the track near a mountain an old train is coming down the tracks a train passing country road with a rock with rocks a train is traveling on a mountain trail a train is coming the tracks at a mountain **Top-*s*** a train on a train track near a mountain a train is on a tracks near a mountain a train traveling down the tracks in a mountainous area a train traveling down train tracks near a mountain a train is going down the track near a mountain **Top-*p*** a train is traveling along a side walk by mountains a train traveling down a train track near a mountain a train going past a rock mountain in the fog a train train going down a mountain in a mountains a train is a track with a mountain in the background	**RS 0.7 + PN** a train is sitting on a track near a cliff an old train is coming down the tracks a train on a track near a cliff a train is traveling down a rocky mountain a train is on the tracks near a cliff **Top-*s* + PN** a train on a train track near a cliff a train traveling down train tracks near a mountain a train going down the tracks in a mountainous area a train traveling down train tracks near a mountain a train is going down the tracks near a cliff **Top-*p* + PN** a train is going down the side surrounded by rocks a train traveling down a train track near a mountain a train traveling along a rocky mountain in the mountains a passenger train traveling down a track near some rocks a train on a track with a mountain in the background
	**RS 0.7** a bathroom with a sink and a toilet a bathroom with a toilet a mirror and a sink a small bathroom with a sink and a dog in it a toilet bathroom with a sink and a toilet a bathroom with a green mirror and white toilet next to a sink **Top-*s*** a bathroom with a sink and toilet and a mirror a small bathroom with a toilet a toilet and a sink the bathroom has a green sink and a mirror a bathroom with a white toilet and a sink in it a bathroom has a sink and a toilet **Top-*p*** a bathroom with a vanity and and a toilet a bathroom with a mirror sink and mirror a white toilet sitting next to a sink a white bathroom with a shower and toilet and mirror a white toilet sitting in to a sink	**RS 0.7 + PN** a bathroom with a sink and a toilet a bathroom with a toilet a sink and a mirror a small bathroom with a sink and a toilet a small bathroom with a sink and a toilet a bathroom with a white sink and a toilet next to a sink **Top-*s* + PN** a bathroom with a sink a toilet and a mirror a small bathroom with a sink a toilet and a mirror a bathroom with a toilet sink and a mirror a bathroom with a white toilet and a sink a bathroom with a sink and a toilet **Top-*p* + PN** a bathroom with a toilet sink and a mirror a bathroom with a toilet sink and mirror a white toilet sitting next to a sink a small bathroom with a sink a toilet and mirror a white toilet sitting next to a sink
	**RS 0.7** a train on a city street with people walking down the street a bus and the street with a buildings busy street a yellow train is coming down the tracks a yellow city decker bus on a street with a train is at a cross walk in the city **Top-*s*** a train is traveling through a busy city street a yellow bus on a city street a yellow double decker bus on a city street the train is traveling down the busy street a yellow bus traveling down the street in the middle of a city **Top-*p*** a train is traveling through a busy city a bus on the street in front city a bus stops for the street to the people a city road with buses parked in it a yellow train is going through a street	**RS 0.7 + PN** a train on a city street with people walking down the street a train on a street near a busy city street a yellow train is coming down the tracks a yellow double decker train on a city street a train stopped at a cross walk in a city **Top-*s* + PN** a train is traveling down a busy city street a yellow train on a city street a yellow double decker train on a city street the train is going down the city street a yellow train traveling down a street in the middle of a city **Top-*p* + PN** a train is traveling down a busy street a train on the tracks in a city a train waiting on the street with many people a city street with people walking on it a yellow train is traveling down the street

Figure A1: Examples of raw descriptions generated by random sampling ($t_o = 0.7$), top-s, top-p and refined descriptions by PN on the m-RNN test split of MS COCO dataset.



RS 0.7
a giraffe standing next to a pile of rocks and a large tree
a giraffe standing in a fenced in area with to a tree
there is a giraffe and a giraffe in a zoo
a giraffe standing on a tree next to a tree
a giraffe eating leaves from a tree and another giraffe is standing in the background

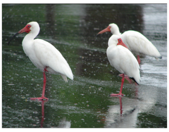
Top-*s*
a giraffe is standing in a zoo exhibit
a giraffe standing in front of a stone wall and a fence
a giraffe standing in a zoo exhibit with a zoo in the background
the giraffe standing next to a large giraffe
two giraffes are standing near a large rock

Top-*p*
a giraffe and a zebra are in a zoo exhibit
a giraffe standing in an enclosed area next to a wooden fence
a giraffe stands in an enclosure with a fenced enclosure in the background
a giraffe is walking near a large giraffe
a giraffe walking standing the ground near a rocks

RS 0.7 + PN
a giraffe standing next to a pile of rocks
a giraffe standing in a fenced in area next to a tree
there is a giraffe and a zebra at the zoo
a giraffe standing by a tree next to a tree
a giraffe eating leaves from a tree while another giraffe is standing in the background

Top-*s* + PN
a giraffe is standing in a zoo enclosure
a giraffe standing in front of a stone wall
a giraffe standing in a zoo enclosure at a zoo
a giraffe standing next to a tall building
two giraffes are standing near a large tree

Top-*p* + PN
a giraffe and a zebra standing in a zoo enclosure
a giraffe standing in an enclosed area next to a stone wall
a giraffe standing in an enclosure in a zoo
a giraffe is standing near a large giraffe
a giraffe is on the grass near some rocks



RS 0.7
a couple of birds with a long bill beak
two white birds stand in a shallow stream
two colorful birds standing in the grass near water
two ducks standing in front of a marsh of one
two birds standing in the water next to each other

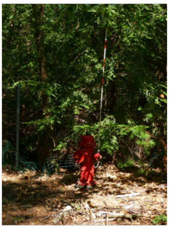
Top-*s*
two birds standing in the water looking for food
two white and red birds standing on a water
a white bird standing next to a bird
two birds standing in a water of water
a couple of birds that on top of a body of water

Top-*p*
a close up of two birds on the water
a couple of long red birds standing next to each other
a pelican and a duck are in the water
a close up of a bird standing on some water
a pair of birds birds standing in the water

RS 0.7 + PN
a couple of birds with a long red beak
two white birds standing in a small pond
two white birds standing in the water near water
two birds standing in front of a pond
two birds standing in the water next to each other

Top-*s* + PN
two birds standing in the water looking for food
a white and red birds standing in the water
a white bird standing next to a bird
two birds standing in the body of water
a couple of birds standing on top of a body of water

Top-*p* + PN
a close up of two birds in the water
a couple of white white birds standing next to each other
a bird and a bird standing in the water
a close up of a bird standing in the water
a couple of white birds standing in the water



RS 0.7
a red fire hydrant in front of a bunch of trees
a red fire hydrant in a wooded area
a red fire hydrant in the middle of the woods
a red fire hydrant by a a lush forest
a red fire hydrant in front of a park and trees

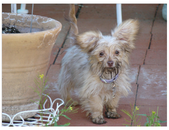
Top-*s*
a fire hydrant is sitting in the middle of a forest
a red fire hydrant in a forest of a forest
a red fire hydrant sitting in a forest next to a tree
a red fire hydrant sitting in the middle of a forest
a fire hydrant sitting in the middle of a forest

Top-*p*
a red fire hydrant in a rural area
a red fire hydrant in a forest in front of trees
a fire hydrant on the side of the woods with the UNK
a red fire hydrant in the middle of a forest
a red fire hydrant sitting in a forest surrounded by trees

RS 0.7 + PN
a red fire hydrant in front of a group of trees
a red fire hydrant in a wooded area
a red fire hydrant in the middle of the woods
a red fire hydrant sitting in a forest
a red fire hydrant in front of a forest

Top-*s* + PN
a fire hydrant is sitting in the middle of a forest
a red fire hydrant in the middle of a forest
a red fire hydrant sitting in a forest next to a forest
a red fire hydrant sitting in the middle of a forest
a fire hydrant sitting in the middle of a forest

Top-*p* + PN
a red fire hydrant in a wooded area
a red fire hydrant in a forest in front of trees
a fire hydrant on the side of the road
a red fire hydrant in the middle of a forest
a red fire hydrant sitting in a forest surrounded by trees



RS 0.7
a small dog is sitting by a pot with a toy
a dog dog is sitting next to a pot
a little dog standing on a stone tile floor
a dog that is standing next to a pot
a small dog sits on a brick floor

Top-*s*
a small dog is standing in a planter
a white dog standing on a brick walkway
a dog dog sitting in front of a planter
a dog dog standing on a sidewalk next to a pot
a small white dog sitting next to an orange pot

Top-*p*
a small dog standing next to a pot
a dog is in a yard near a bowl
a small dog standing on the ground next to a bowl
a dog that is standing by a tree
a white dog standing next to a bowl

RS 0.7 + PN
a small dog is standing in a garden
a small dog is standing next to a plant
a small dog standing on a hard wood floor
a dog that is standing next to a plant
a small dog sitting on a brick floor

Top-*s* + PN
a small dog is standing in a garden
a small dog standing on a brick wall
a small dog standing in front of a wall
a small dog standing on a sidewalk next to a plant
a small white dog standing next to an orange pot

Top-*p*
a small dog standing next to a plant
a dog standing in a garden near a plant
a small dog sitting on the ground next to a plant
a dog that is standing near a plant
a small dog standing next to a plant

Figure A2: Examples of raw descriptions generated by random sampling ($t_o = 0.7$), top-s, top-p and refined descriptions by PN on the m-RNN test split of MS COCO dataset.

References

[1] 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.

[2] 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.

[3] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

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

[5] 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.

[6] 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.

[7] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[8] 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.

[9] 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.