Supplementary Material: Text-to-Image Generation Grounded by Fine-Grained User Attention

Jing Yu Koh*, Jason Baldridge, Honglak Lee and Yinfei Yang
Google Research
{jykoh,jridge,honglak,yinfeiy}@google.com

Appendix

We provide additional implementation details, further results for randomly sampled LN-COCO and LN-OpenImages examples, and greater detail on our human evaluation procedure.

A. Implementation Details

A.1. BERT Sequence Tagger

The sequence tagger is implemented in TensorFlow [1]. We fine-tuned the public pretrained uncased BERT-Large model [1] to perform sequence tagging on our HMM tags. Weights were optimized using Adam [3] with a learning rate of 1e-5. The model was trained with a batch size of 128 over 10 epochs, on a 2x2 TPU. During training, we set the class probabilities of the COCO-Stuff `other` and `background` tags to 0. This assists in downstream image generation, as we found that these classes were not meaningful when presented to the mask-to-image translation models, i.e. SPADE [5] and CC-FPSE [4].

A.2. Mask Retrieval

For the text modality, we use the pretrained public uncased BERT-Base model[2]. The outputs of the BERT model are passed into a transformer and an FC-layer which maps it into $\mathbb{R}^{2048}$. We use in-batch negative sampling for training the dual encoder model. The use of the BERT-Base model instead of BERT-Large allows us to experiment with larger batch size.

For the image modality, we use an Inception v3 model [7] pretrained on ImageNet[3]. The outputs have dimensionality $\mathbb{R}^{2048}$.

The dual-encoder model is implemented in TensorFlow. Weights are optimized with Adam [3] with a learning rate of 1e-3, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e-8$. We impose a learning rate schedule of stepwise exponential decay, with decay parameter 0.999 and 1000 steps for each decay step. Gradients are clipped to have l_2 norm between 0.1 and 1.0. The model was trained with a batch size of 2048 on a 4x4 TPU. We train the dual-encoder until loss convergence on the LN-COCO validation set.

A.3. Mask-to-Image Translation

For SPADE[4] and CC-FPSE[5], we used the official models released by the authors. All models are pretrained on COCO-Stuff [2].

A.4. Evaluation Metrics

We use a standard codebase[6] to compute IS. Similarly, we compute FID scores using a popular PyTorch repository.[7]

A.5. Dataset Details

We conduct all experiments on Localized Narratives [6].[8] The LN-COCO training set contains 130,810 examples, and the validation set contains 8573 examples. In addition, we run evaluations on a held out test set of the Open Images subset of Localized Narratives (LN-OpenImages). For this, we sampled 10,000 random images for computing IS and FID.

B. Human Evaluation

We collected two human evaluation metrics for comparing image generation models: (1) realism of generated images, and (2) language alignment of generated images. We

*Work done as a member of the AI Residency program.

[1] https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-24_H-1024_A-16.zip

[2] https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip

[3] https://tfhub.dev/google/imagenet/inception_v3/feature_vector/4

[4] https://github.com/NVlabs/SPADE

[5] https://github.com/xh-liu/CC-FPSE

[6] https://github.com/openai/improved-gan/tree/master/inception_score

[7] https://github.com/mseitzer/pytorch-fid

[8] https://google.github.io/localized-narratives/

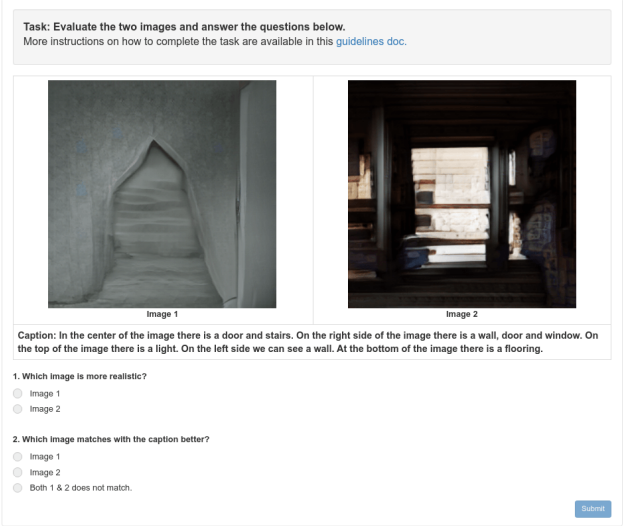

Figure 1. User interface presented to human evaluators.

mostly compare our results with AttnGAN [8], and hence utilize a side-by-side comparison format to minimize bias and variance in collected ratings. Figure 1 displays the user interface that is shown to the human evaluators. Each annotator is asked to answer the two questions displayed, selecting the image that is more realistic, and a better match to the caption. The presentation order of images are random to avoid bias (i.e. TReCS images appear as *Image 1* 50% of the time).

In addition, we observed in our experiments that a small proportion of images generated by both models are not well aligned with the provided captions. For example, a blank image might be generated by both TReCS and AttnGAN in certain cases. For these scenarios, evaluators may select a 3rd option that indicates that neither image is matched.

We collected five independent ratings for a randomly selected subset of 1000 LN-COCO validation examples and 1000 LN-OpenImages examples (sampled from the 10,000 LN-OpenImages test examples). We use majority voting over the five ratings as the final quality rating, but also display the full range of votes for each model in our results figures. The 1000 synthesized images (for both LN-COCO and LN-OpenImages) will be publicly released to facilitate reproducibility and comparison with other models.

Annotators were employed as contractors and were paid hourly wages that are competitive for their locale. They have standard rights as contractors. They are fluent non-native English speakers.

C. Qualitative Results

C.1. Randomly Sampled Images

Figure 2 and 3 show several randomly sampled images from LN-COCO and LN-OpenImages respectively. The outputs are consistent on these two subsets, indicating TReCS is able to generate plausible looking images for most provided narratives. While there is room for improvement, we observe that in most outputs, TReCS is able to generate relevant objects, despite occasional glitches.

We observe that some images are blank or with very few objects (e.g. the image in the first column and third row from the bottom in Figure 3). We believe it is due to the limitation of the candidates pool used in the mask retrieval module. If there are no relevant masks returned from the retrieval module, the system is unable to draw anything meaningful. This would likely be improved significantly by expanding the corpus of retrievable masks.

C.2. Effect of Captions on Generated Images

It is possible that one image in the Localized Narratives dataset can be annotated by multiple independent annotators, providing different captions for the same image. TReCS is able to read the differences from different captions and generate corresponding images, which emphasis the importance of language grounding and user attention in this text-to-image synthesis problem. For example, the image in the left column of Figure 4 depicts two black dogs playing with a frisbee. However, the first annotator doesn't describe *dog* explicitly, and the resulting image is unable to depict the dogs. For the other images, the dogs are captured and well generated.

Similarly, we observe variations in the second column of Figure 4, with background scenes differing based on the provided description. In the first image, the annotator describes a mountain and a lake, which is depicted in the generated image. Additionally, we observe that the third annotator mistakenly described the giraffe as a zebra, resulting in the generated image depicting a zebra. This emphasizes the integrality of language grounding within the TReCS system.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

[4] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems*, 2019.

Figure 2. Randomly sampled TReCS images from LN-COCO.

[5] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[6] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language

Figure 3. Randomly sampled TRECS images from LN-OpenImages.

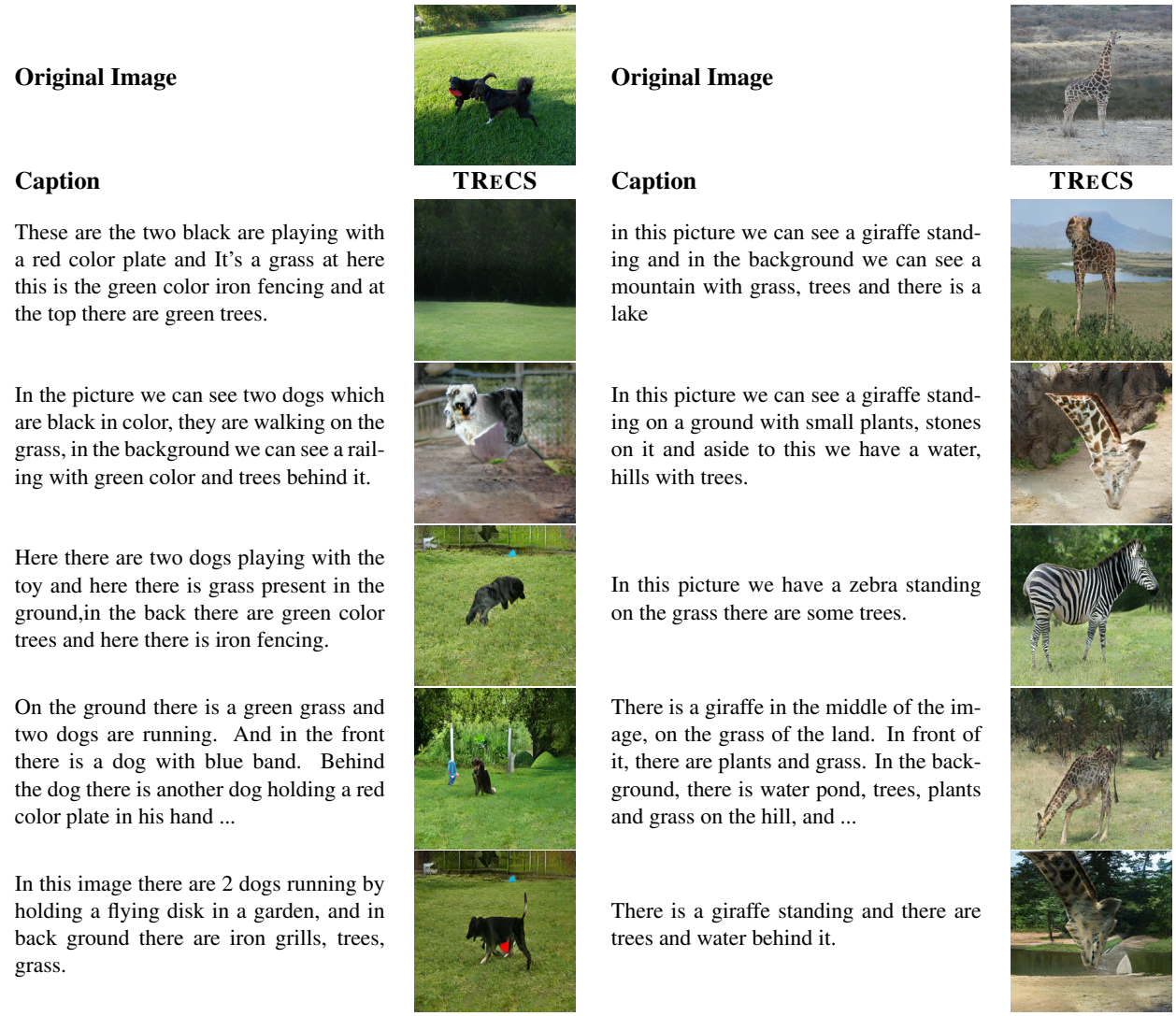

Figure 4. TRECS images based off different descriptions of the same image from LN-COCO. Descriptions are shown verbatim.

with localized narratives. *Proceedings of the European Conference on Computer Vision*, 2020.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.