

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

Jing Yu Koh*, Jason Baldridge, Honglak Lee and Yinfei Yang
Google Research

{jycoh, 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 model¹ 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 model². 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 ImageNet³. 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 SPADE⁴ and CC-FPSE⁵, 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 codebase⁶ 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.

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

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

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

⁴<https://github.com/NVlabs/SPADE>

⁵<https://github.com/xh-liu/CC-FPSE>

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

⁷<https://github.com/mseitzer/pytorch-fid>

⁸<https://google.github.io/localized-narratives/>

Task: Evaluate the two images and answer the questions below.
More instructions on how to complete the task are available in this [guidelines doc](#).






Image 1 Image 2

Caption: In the center of the image there is a door and stairs. On the right side of the image there is a wall, door and window. On the top of the image there is a light. On the left side we can see a wall. At the bottom of the image there is a flooring.

1. Which image is more realistic?

☐ Image 1

☐ Image 2

2. Which image matches with the caption better?

☐ Image 1

☐ Image 2

☐ Both 1 & 2 does not match.

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 emphasizes 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.

Caption	Original	AttnGAN	TRECS	Caption	Original	AttnGAN	TRECS
In this picture we can see a train in blue colour on a railway track. We can see two persons inside a train. This is a signal. These are boards ...				In this image i can see a few persons standing. On the table there is a system, bottle, speaker, mouse, cup and a tissue. At the back side the woman ...			
He is standing. He is wearing a cap, he is holding a bat. There is a toy on the left side. We can see in the background green color banner.				This is picture of a place where we have some people sitting on the chairs in front of the table on which there are some things like laptop, bottles and ...			
In this picture there are trees in the right side. Towards the left there is a wooden tray, on the tray there are banana peels and some bird are on the tree. In the ...				In this image, there are two persons standing and they are smiling, in the back-ground there is a white color wall.			
In this image in the center there is one television and under the television there is a table on the table there is one remote, books and some objects and some wires ...				This is a picture of a red color car, a person standing beside another car, and at the back ground there are group of cars in the street, trees, sky.			
In this image in the center there is one person who is holding a racket and he is playing tennis, and on the background there is a board and on the left side there ...				This is a picture of inside of the house, in this picture on the top there are some cupboards. In the bottom also there are some cupboards and on the ...			
In this picture we can see a glass with drink in it, and the glass is on the table.				This is the picture outside of the building. At the right side of the image there is a clock on the pole. At the back there are ...			
In this image, in the middle there is a table, on that table there are some plates which are in white color, there are some glasses, there are ...				A picture inside of a hall. This is window with curtain. On floor there is a couch. Pictures on green wall. This is lamp ...			
Here in the front we can see boats travelling in the river and beside that we can see a train travelling on a track and we can see plants and trees present ...				In this image there is a bench on stone slab. Bottom of image is a grassy land. Background of image there are few plants.			
This image is clicked in a open lad. To the left there is a recliner chair. On it there is a teddy bear, clothes and beside it there are many toys. To the teddy ...				In this image there is a person standing with the snow skating sticks in a snow mountain ,and at the back ground there are plants ...			
This picture shows a stop sign board on the road and we say few paper stucked on it and we see a brick wall on the side				This is the picture might be taken on sea shore. In the image in middle there is a person and we can also see another person ...			
In the image there is a polar bear sat, beside that there is a stone wall.				This is a picture taken in the outdoors. There are two persons skating in snow. In front of the people there are trees and ...			

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 Com-*

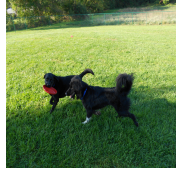
puter Vision and Pattern Recognition, 2019.

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

Caption	Original	AttnGAN	TRECS	Caption	Original	AttnGAN	TRECS
In this image we can see a house, a fence, there are plants, trees, there is a vehicle, a board with text on it, also we can see the cloudy sky.				This is an inner view of a building containing a roof, fence and windows. We can also see some painting on the walls and pillars, decors and ...			
In this picture there is a dog in the center of the image and there is a door in the background area of the image.				The picture is taken in a court. In the foreground of the picture there is a girl and a woman playing with ball ...			
As we can see in the image there is a white color wall, rack, lights, baskets, table, sofa, pillows and on table there are photo frames ...				In this image I can see few trees, rock, water, sky and few houses.			
In this image I can see the depiction picture of a man. I can also see he is sitting. I can see he is wearing a blazer, a shirt ...				In this image there is the sky towards the top of the image, there are trees, there is a building, there is text on the building ...			
... at the bottom of the image there is a road. In the background there are a few houses and there are many trees ...				In this image we can see a person wearing white color dress playing baseball and in the background of the image there are ...			
In this image we can see a dog is eating food. This is wooden floor. Background it is blur.				This is green color plant.			
The person wearing black dress is standing on a bicycle and holding a water bottle in his hand and there are few people ...				In this image there is a lake beside that there are so many trees.			
In this image, there are group of people standing, walking and sitting on the bench in front of the table. And on the table ...				In this image I can see the person and the person is wearing black color dress and I can see the gray color background.			
In this image there is an art of two broken coins and there is a text.				In this image there are two womens are doing ramp walk on the stage. The right side women is holding a handbag and ...			
In this image I can see a person standing with open hands at a hill station. I can see buildings and at the top of the image ...				This is a inside view of a building, where there are lights, CC camera , iron grills, door, staircase, name board hanging ...			
In the center of the image there is a woman standing and holding object. In the background we can see decors, cupboards, table.				This seems like panorama shot of a building, where we can see grill on the both the sides and a green wall with text on it.			
In this image I can see the machine in yellow color, background I can see few boards attached to the wall and the wall is in brown ...				In this image we can see a few houses, there are some trees, plants, windows and a pole, also we can see some vehicles on the ...			

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

Original Image



Caption

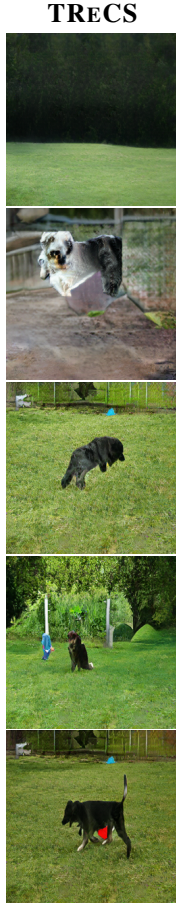
These are the two black are playing with a red color plate and It's a grass at here this is the green color iron fencing and at the top there are green trees.

In the picture we can see two dogs which are black in color, they are walking on the grass, in the background we can see a railing with green color and trees behind it.

Here there are two dogs playing with the toy and here there is grass present in the ground, in the back there are green color trees and here there is iron fencing.

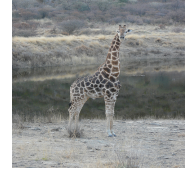
On the ground there is a green grass and two dogs are running. And in the front there is a dog with blue band. Behind the dog there is another dog holding a red color plate in his hand ...

In this image there are 2 dogs running by holding a flying disk in a garden, and in back ground there are iron grills, trees, grass.



TRECS

Original Image



Caption

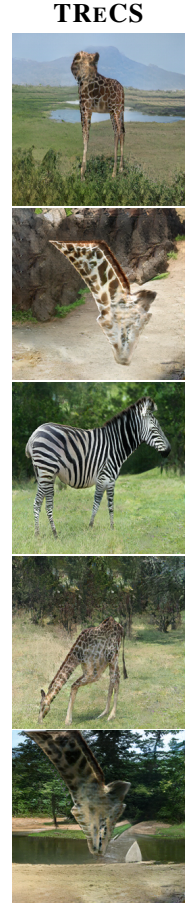
in this picture we can see a giraffe standing and in the background we can see a mountain with grass, trees and there is a lake

In this picture we can see a giraffe standing on a ground with small plants, stones on it and aside to this we have a water, hills with trees.

In this picture we have a zebra standing on the grass there are some trees.

There is a giraffe in the middle of the image, on the grass of the land. In front of it, there are plants and grass. In the background, there is water pond, trees, plants and grass on the hill, and ...

There is a giraffe standing and there are trees and water behind it.



TRECS

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.