Compositional Generation of Images

Amit Raj*, **Cusuh Ham***, **Huda Alamri***, **Vincent Cartillier***, **Stefan Lee, James Hays**
Georgia Institute of Technology
{amit.raj, cusuh, halamri, vcartillier3, steflee, hays}@gatech.edu

Abstract

We propose a novel approach to synthesize images from text descriptions by exploiting the compositionality of language. We decompose the input text to basic visual primitives and dynamically train a network to generate visual attributes for the desired image. By incorporating compositionality into the network design, we aim to make the generative process more interpretable and controllable. We define a set of compositional blocks and introduce them into generative networks in two different settings–one at the conditioning level and one at the generation level. We evaluate our approaches on the SHAPES dataset and present the qualitative results.

1 Introduction

With growing interest towards vision and language tasks such as captioning [24, 23], VQA [3, 8], visual dialog [5, 6], and text-based image synthesis[20], compositionality and interpretability of architechtures have become increasingly important aspects of network design. Many approaches to these tasks involve combining state-of-the-art networks from visual understanding and language modeling. However, the exploration of these modalities for image generation has been limited to treating the language component as a simple conditioning signal in the generation process. There are several drawbacks to these conventional approaches. First, they are not scalable to incremental generation or editing. Second, such systems have been shown to work only on restricted domains (e.g. faces, birds) [15] or in generation of low variance images [19, 7] (e.g. LSUN). However, authoring image content through textual interactions has a huge impact on image generation and editing. One can imagine a scenario where an image editing tutorial is parsed by a system which can then generate or edit images according to the parsed instructions. Finding effective methods to combine information between the text and image modalities is an important step to improve the scalability and reliability of text-based image editing tools.

To address these concerns, we build upon ideas from Andreas et al. [1] and propose the first compositional approach to text-based image generation. We explicitly design a set of fundamental blocks to better aid in interpretability and to encourage the image generation process to be more compositional. We demonstrate that disentangling the generation process into pre-defined components and injecting conditional information from the input text to the appropriate blocks not only improves the generation quality but also make the editing process easier. We present results that show promise towards exploration in this direction.

The contributions of this paper are two-fold:

1. We design a set of fundamental blocks which can be used to generate images in an interpretable manner.
2. We present an approach to dynamically inject conditions at different stages of the network to aid in quick editing of the generated image.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Related Work

2.1 Generative Networks

Generative adversarial networks [9, 4, 10, 19, 15, 16] have received considerable attention in recent literature for image generation tasks. Of these the conditional variants are particularly popular in generating images conditioned on various kinds of signal such as class[18], attributes[14], pose[17] or text [22, 21, 2]. The authors of [20, 25] generate images of birds conditioned on text. Zhu et al. [26] present a framework to edit garments based on text descriptions. In all these works, the sentence descriptions are encoded into a fixed length embedding. In contrast, we posit that by exploiting the richness and composition of language, we can inject better conditioning information into the network instead of using a fixed embedding. To this end, we present a framework in which different parts of the sentence condition different stages of the generation process, making the generation more compositional.

2.2 Neural Module Networks

Andreas et al. [1] propose the neural module networks to address the VQA task and present a set of building blocks to dynamically construct networks based on the question. In the earlier works, a rule-based parser is employed to extract the structure of the dynamic network from the query text. Later works have focused on determining these programs in an end-to-end fashion [11, 13]. In our exploration, we extend this initial work to the generative task.

3 Approach

In this section, we formulate our problem and describe two variants of our approach: Compositional C-GAN and Modular GAN. The two architectures differ in which aspect of the generative network is made compositional. The compositional C-GAN is modular at the embedding level to generate a conditioning signal, while the modular-GAN has discrete blocks dynamically instantiated during run-time. We use the SHAPES dataset for our experiments [1]

3.1 Problem Formulation

We define our problem as a compositional generative task, given a fixed set of generative functions. For the SHAPES dataset we define our set $\mathcal{F}$ of compositional functions as follows:

$$\mathcal{F} = \{\mathcal{S} : c_{shape} \rightarrow I, \mathcal{C} : (I \times c_{color}) \rightarrow I, \mathcal{L} : (I \times c_{location}), \cup_k \mathcal{J}_k : (I \times M)^k \rightarrow I\}$$

where $\mathcal{S}$ is a shape block that generates the basic shape, $\mathcal{C}$ is a color block, $\mathcal{L}$ is a location block that places the generated shape in one of 9 locations, and J_k is a combine block that takes in multiple images and a mask and returns a single image. The conditions c_{shape}, c_{color}, $c_{location}$ and mask M (determined by the location block) are fed into these blocks, respectively. We assume that any image can be generated as:

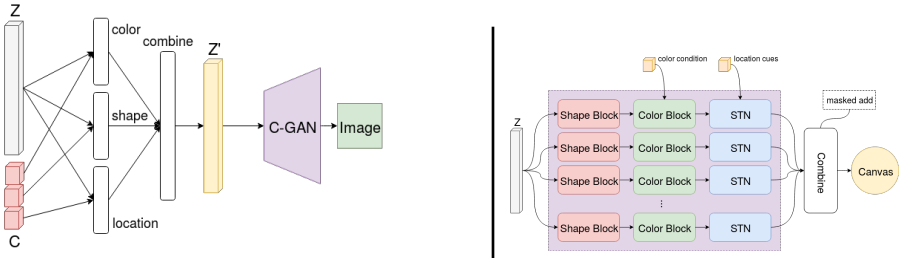
$$\Phi(\bar{\mathcal{F}}), \text{ for some } \bar{\mathcal{F}} \subseteq F$$

where Φ is some composition of functions. For instance, "A red square" can be represented as $\mathcal{C}(\mathcal{S}(square), red)$. Similarly, "A red square next to the green circle" is represented as:

$$f_1 = \mathcal{L}(\mathcal{C}(\mathcal{S}(square), red), c_{location}) \quad (1)$$
$$f_2 = \mathcal{L}(\mathcal{C}(\mathcal{S}(circle, green), c_{location}) \quad (2)$$
$$\Phi = \mathcal{J}_2(f_1, f_2, M_1, M_2) \quad (3)$$

For some masks M_1, M_2 corresponding to the query locations.

We use a rule-based parser to parse and conform the sentence to the formulation above. Based on the generated tuple of functions, we dynamically instantiate the blocks associated with these functions.



(a) Compositional C-GAN (b) Modular GAN

Figure 1: Our two proposed architectures.

3.2 Compositional C-GAN

The Compositional C-GAN architecture is shown in Figure 1a. Given an image I and query Q, we parse Q to extract the shape, color, and location attributes $\{(c_{shape}, c_{color}, c_{location})_i\}_{i=1}^{n}$ for every object in the query. The input to our model is a noise vector z and condition vectors $c_1, c_2, ..., c_n$ where c_i is a one-hot encoding for the particular condition type. Each c_i along with z is fed into the respective conditional blocks, which are 4-layer multilayer perceptrons (MLPs). The outputs of these blocks are fed into the $combine$ block, another 4-layer MLP, which produces the final embedding z' for the input to a standard conditional GAN (C-GAN) [18].

3.3 Modular GAN

The Modular GAN architecture is shown in Figure 1b. We describe the 4 fundamental blocks that the modular block entails.

3.3.1 Shape Block

The shape block is a decoder that generates a single channel image. There is one instance of a shape block for every shape. That is, the shape block instantiates different set of weights for different shape queries. The shape block has lesser number of parameters than a standard generative network since every shape block has to keep track of only one kind of shape.

3.3.2 Color Block

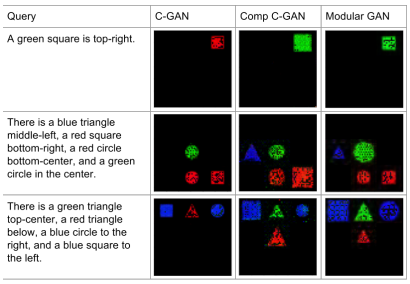
The color block is a shallow 4-channel encoder-decoder architecture that operates on the single channel output of the shape block to generate an RGB image. Two variants are possible: the first is when there is a different instance of the block for each color and the second involves employing a single instance of the color block that is then conditioned on the target color. To demonstrate that the image generation process can be made fully compositional, we present the results from experiments with the first variant.

3.3.3 Location Block

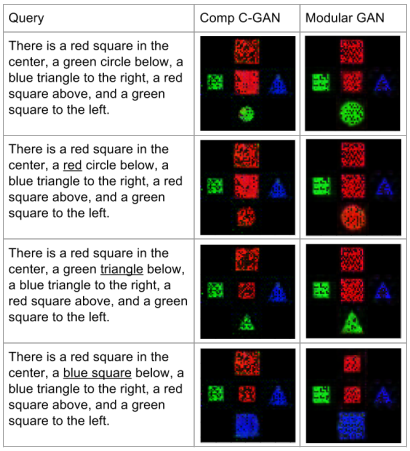
The location block, consists of a spatial transformer network [12] that operates on the output of the color block and locational conditioning information from the text. In our experiments, the affine grid generator of the locational block regresses to fixed affine transforms based on the one-hot encoding of location.

3.3.4 Combine Block

The combine block operates on outputs from multiple location blocks and performs a masked addition of these images to generate the final image. The masks in our experiments are determined based on the parsed location information for the particular shape. Based on an input query, different number of blocks of each kind are dynamically instantiated and combined to form the final image.



(a) We use three queries to generate images: a single object with absolute location, multiple objects with absolute locations, and multiple objects with relative locations. We can notice that the C-GAN is unable to capture all the required objects in the query.



(b) We modify a part of the text (underlined) to edit the generated image.

Figure 2: Qualitative results on the SHAPES dataset.

4 Experiments

In this section we present our results on the synthetic SHAPES dataset [1]. The dataset consists of images of multiple objects of different shapes, colors and locations. Each image is associated with a specific text query. The inherent structure of the dataset provides a controlled environment which allows our results to be more interpretable in the qualitative evaluation process.

We use a rendering script to generate a set of 64×64 images and the corresponding queries. In each image, we randomly create 1-9 objects of varying shapes and colors. There are three possible shapes–$circle, square, triangle$–and three colors–$red, green, blue$. The location of each object can be in any of the 9 cells in a 3×3 grid. For an image with multiple objects, we generate a query that specifies the absolute location of at least one object and the location of the other objects relative to this position. Because we have a fixed number of possible locations for each object, we can parse each $(l_{absolute}, l_{relative})$ pair and map the relative locations to an absolute location in the image grid.

4.1 Qualitative results

We compare our two approaches with the baseline, C-GAN [18], which similarly uses a set of attributes to condition the generated image. However, the C-GAN is being forced to disentangle these attributes whereas by enforcing modularity with the compositional blocks, our models are able to decompose the images more effectively.

We modified C-GAN for our experiments. In the original architecture, the input noise vector z and condition vector c are directly concatenated and fed into the GAN. In our version, we project c to a 32D embedding which is then concatenated with z and fed to the GAN.

In Figure 2a we show that our models are able to correctly generate the given sentences while the vanilla C-GAN fails at certain instances. In Figure 2b we demonstrate the ability to edit parts of the image through slight modifications to the sentence.

5 Conclusion

In this paper, we introduced two compositional methods for generating images from text descriptions by defining a set of compositional generative blocks. We have demonstrated that our approach improves the quality of generating images from text descriptions. In the future, we aim to generalize our approach from this synthetic world to real images and deal with more complex text descriptions.

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016. 1, 2, 4

[2] Anonymous. 3c-gan: An condition-context-composite generative adversarial networks for generating images separately. *International Conference on Learning Representations*, 2018. 2

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 1

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 2

[5] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. *arXiv preprint arXiv:1611.08669*, 2016. 1

[6] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*, 2017. 1

[7] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. 1

[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 1

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[10] R. D. Hjelm, A. P. Jacob, T. Che, K. Cho, and Y. Bengio. Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*, 2017. 2

[11] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. *arXiv preprint arXiv:1704.05526*, 2017. 2

[12] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 3

[13] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. *arXiv preprint arXiv:1705.03633*, 2017. 2

[14] L. Karacan, Z. Akata, A. Erdem, and E. Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 2

[15] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1, 2

[16] N. Kodali, J. Abernethy, J. Hays, and Z. Kira. How to train your dragan. *arXiv preprint arXiv:1705.07215*, 2017. 2

[17] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. *arXiv preprint arXiv:1705.09368*, 2017. 2

[18] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2, 3, 4

[19] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 2

[20] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 1, 2

[21] S. Reed, A. van den Oord, N. Kalchbrenner, V. Bapst, M. Botvinick, and N. de Freitas. Generating interpretable images with controllable structure. 2016. 2

[22] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *Advances in Neural Information Processing Systems*, pages 217–225, 2016. 2

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 1

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 1

[25] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016. 2

[26] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy. Be your own prada: Fashion synthesis with structural coherence. *arXiv preprint arXiv:1710.07346*, 2017. 2