# A Comprehensive Analysis of Semantic Compositionality in Text-to-Image Generation

**Chihiro Fujiyama** **Ichiro Kobayashi**
Ochanomizu University
Ohtsuka 2-1-1, Bunkyo, Tokyo, 112-8610, Japan
fujiyama.chihiro@is.ocha.ac.jp
koba@is.ocha.ac.jp

## Abstract

In this study, we analyze the structure of a feature representation space in a deep neural text-to-image generative model in order to explore the possibility that the model implicitly acquires the semantic compositionality while generating images from captions as an explicit task. This is a fundamental approach toward concept acquisition by grounding between linguistic expressions and images. We analyze the semantic compositionality in a text embedding space in a generative model. Our experimental result suggests that the semantic compositionality appears among words indicating positions. This study is the first attempt to explore the semantic compositionality in text-to-image generation.

## 1 Introduction

Recently, many novel methods based on deep neural networks have been proposed in natural language processing tasks and they have achieved highly impressive results, however, it remains a difficult problem how those models interpret natural language. In this study, we attempt to disentangle how a deep neural network model interprets natural language through a text-to-image generation task. A concept behind this is that it is necessary to interpret natural language and to ground language representations and image representations in order to generate reasonable images reflecting the content of the given natural language. In the context of text-to-image generation with deep neural networks, generative adversarial networks (GANs) have made a great stride and they successfully generate high resolution realistic images (Reed et al., 2016; Zhang et al., 2017; Xu et al., 2018; Zhang et al., 2018; Qiao et al., 2019). On the other hand, most of such researches focus on generating high resolution images, and there are no research to examine how those models connect the two different modalities, natural language description and images in detail. Another stream to generate images from natural language description is the method based on Variational Autoencoder (Kingma and Welling, 2014). We focus on a Variational Autoencoder model because it performs image generation in a more straightforward way compared with the models based on GANs. In this study, we will explore the structure of a feature representation space to examine how much semantic compositionality in words is realized at an internal representation space in a text-to-image generative model. The contribution of this paper is that this is the first attempt to explore the semantic compositionality in text-to-image generation.

## 2 AlignDRAW

Mansimov et al. (2016) proposed a model, called alignDRAW, which generates images from natural language descriptions. alignDRAW imitates the procedure of human drawing, that is, it iteratively draws patches on a canvas while attending to the relevant words in the description. This model

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

is an extension of Deep Recurrent Attentive Writer (DRAW) (Gregor et al., 2015) built based on Variational Autoencoder. It employs soft attention mechanism (Bahdanau et al., 2015) to strengthen the relation between linguistic representations and images in text-to-image generation, and then performs stepwise elaboration of drawings. alignDRAW takes a caption, a sequence of words, as input and encodes it using a bidirectional LSTM (Hochreiter and Schmidhuber, 1997), and then it iteratively represents an image as a sequence of patches drawn on a canvas by using attention mechanism.

## 3 Analysis of Semantic Compositionality

We explore the semantic compositionality in a feature representation space of the alignDRAW model. While the original alignDRAW encodes input captions in the form of sequences of one-hot vectors with a bidirectional LSTM, we extend the model by adding an embedding layer in order to pass each input word to a bidirectional LSTM after mapping it to a distributed representation. The outline of our extension of alignDRAW is illustrated in Figure 1.

After training the model on a dataset which consists of captions and images, we analyze the semantic compositionality among the distributed representations acquired in the embedding space. First, we compose estimated representations in the following two methods: a simple element-wise addition of two vectors, and vector operation which takes account of meta meaning included in the word embedding vectors. Then, we evaluate cosine similarities between estimated and actual word representations.

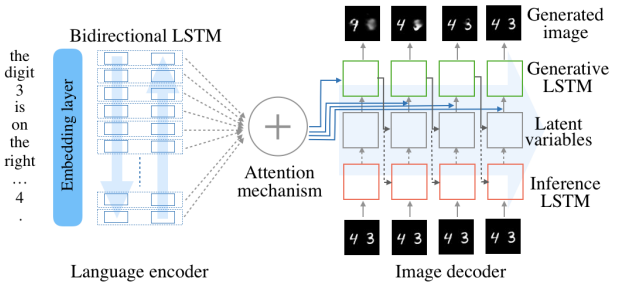


Figure 1: Extended alignDRAW model.

## 4 Experiments

### 4.1 Experimental Setup

We created a dataset consisting of captions in Japanese and images based on the MNIST[1] (Lecun et al., 1998) dataset by using the templates shown in Table 1. Using eight types of templates with placeholders, the captions were artificially created by inserting numeric representations corresponding to digits randomly sampled from the MNIST dataset into the placeholders. Images were created by placing the MNIST images to areas corresponding to the captions on 60 × 60 pixel blank images with a latitude of four pixels. We trained the model on 40,000 samples, and 4,000 samples were used as development and test data, respectively. We followed the settings shown in Table 2 for training an alignDRAW model.

On analyzing the semantic compositionality, we focused on eight words indicating positions, i.e. “左" (left), “右"(right), “上"(top), “下"(bottom), “左上"(top left), “左下"(bottom left), “右下"(bottom right), and “右上"(top right). It is assumed that we interpret “top left" semantically as the addition of “top" and “left". In order to see whether this semantic compositionality is found among representations acquired at the embedding layer, we estimated the representations corresponding to complex concepts in the following two methods. In the first method, we estimated the complex representations for “top left", “bottom left", “bottom right", and “top right" by element-wise addition of vector representations corresponding to primitive components; “left", “right", “top", and “bottom". For instance, the estimated representation corresponding to “top left" was composed as the addition of the vector representations corresponding to “top" and “left". This is a naive way which reflects our intuition, however, this does not take account of meta meaning which each representation possesses. Each representation has meta meaning; for example, “right" is the representation that contains the

[1] http://yann.lecun.com/exdb/mnist/

Table 1: Templates for creating captions.

| | Template in Japanese | Translation in English |
|---|---|---|
| 1 | 数字 _ が 画像 の 左 に ある . | the digit _ is at the left of the image . |
| 2 | 数字 _ が 画像 の 右 に ある . | the digit _ is at the right of the image . |
| 3 | 数字 _ が 画像 の 上 に ある . | the digit _ is at the top of the image . |
| 4 | 数字 _ が 画像 の 下 に ある . | the digit _ is at the bottom of the image . |
| 5 | 数字 _ が 画像 の 左 上 に ある . | the digit _ is at the top left of the image . |
| 6 | 数字 _ が 画像 の 左 下 に ある . | the digit _ is at the bottom left of the image . |
| 7 | 数字 _ が 画像 の 右 上 に ある . | the digit _ is at the top right of the image . |
| 8 | 数字 _ が 画像 の 右 下 に ある . | the digit _ is at the bottom right of the image . |

Table 2: Architectural configurations of our alignDRAW model.

| | |
|---|---|
| Vocabulary size | 28 |
| Language encoder | 32-dimensional distributed representation $\rightarrow$ 128units, bidirectionalLSTM |
| Attention mechanism | 256 units, Bahdanau Attention |
| Decoder | 300 units, DRAW LSTM |
| # Iterations for drawing | 32 steps |
| Dimension of latent variables | 150 |
| Optimization algorithm | RMSProp |
| Learning rate | initial lr: 0.001, halving per 15 epochs after 75th epoch |
| Initial parameters | random values $\sim \mathcal{N}(0, 0.1)$ |
| # Epochs to train | 150 |

meaning of either direction or position, “three" is the representation that contains the meaning of numerical value, and so on. In the naive addition, the meta meaning is also straightforwardly added in vector representation operation, although the same meta meaning might redundantly added. Taking account of this problem, as the second method, we consider meta meaning by applying subtraction as well as addition. Table 3 shows how to compose each estimated representation.

Table 3: Composition of estimated representations.

| Complex concepts | Method | Composition |
|---|---|---|
| top left | 1 | top + left |
| | 2 | bottom left - bottom + top, top right - right + left |
| bottom left | 1 | bottom + left |
| | 2 | top left - top + bottom, bottom right - right + left |
| bottom right | 1 | bottom + right |
| | 2 | top right - top + bottom, bottom left - left + right |
| top right | 1 | top + right |
| | 2 | bottom right - bottom + top, top left - left + right |

### 4.2 Results

As we can see in Figure 2, the model generated reasonable images for given captions. Cosine similarities between estimated representations and corresponding actual representations are shown in Table 4. As for representations estimated by naive addition (Method 1), cosine similarities are high for three words other than “右上" (top right). We assume that the reason why cosine similarity for “top right" was small is because the number of samples corre-

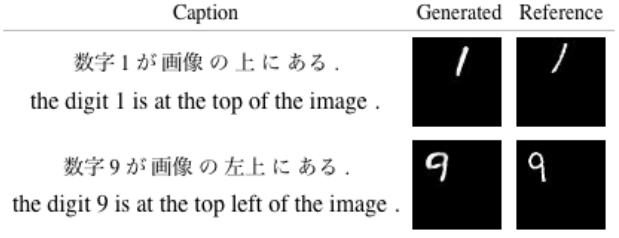


Figure 2: Examples of generated images for the experiments on analyzing the semantic compositionality.

sponding to “top right" was larger than the expected average number of samples and the numbers of samples corresponding to “top" and “right", which were its primitive components, were smaller than the expectation, although we expected the number of samples for each direction was substantially balanced since each template was selected randomly on creating dataset. In the composition, considering meta meaning, there were many cases where cosine similarities were slightly higher or almost the same compared with those in the simple addition. In the case where “top right" were used in the composition, we see that cosine similarities tended to get lower.

Table 5 shows words corresponding to 3-closest representations to the ones estimated by Method 1. For three words, “top left", “bottom left", and “bottom right", we found that the corresponding actual representations successfully ranked in 3-closest representations to their estimated ones respectively. Figure 3 gives a 2-dimensional t-SNE (van der Maaten and Hinton, 2008) visualization of the actual representations and the ones estimated by Method 1. Actual representations corresponding to words indicating positions are colored in red, and representations estimated in Method 1 are colored in blue. Others are displayed in black. For three words other than top right, estimated representations are relatively close to their corresponding actual ones. Thus, we suppose that the result suggests the semantic compositionality is found in the embedding space.

Table 4: Cosine similarities between estimated representations and corresponding actual representations.

| | Method1 | Method2 | |
|---|---|---|---|
| top left | top + left | bottom left - bottom + top | top right - right + left |
| | 0.89 | 0.94 | 0.91 |
| bottom left | bottom + left | top left - top + bottom | bottom right - right + left |
| | 0.80 | 0.35 | 0.92 |
| bottom right | bottom + right | top right - top + bottom | bottom left - left + right |
| | 0.92 | 0.81 | 0.92 |
| top right | top + right | bottom right - bottom + top | top left - left + right |
| | 0.29 | 0.21 | 0.44 |

Table 5: Words corresponding to 3-closest representations to the representations estimated by Method 1.

| rank | top left | bottom left | bottom right | top right |
|---|---|---|---|---|
| 1 | top left | left | right | right |
| 2 | top | top left | 1 | 1 |
| 3 | left | bottom left | bottom right | bottom right |

## 5 Conclusion

We have analyzed the feature representation space in a text-to-image generative model. We focused on alignDRAW model, a text-to-image model based on Variational Autoencoder, and extended it by adding an word embedding layer in order to analyze the semantic compositionality for the obtained word embedding vectors. On analyzing the semantic compositionality, we reported the results suggesting that the semantic compositionality could be found among words representing positions in the embedding space of the model. In the future, we would like to perform further analysis with more diverse datasets.

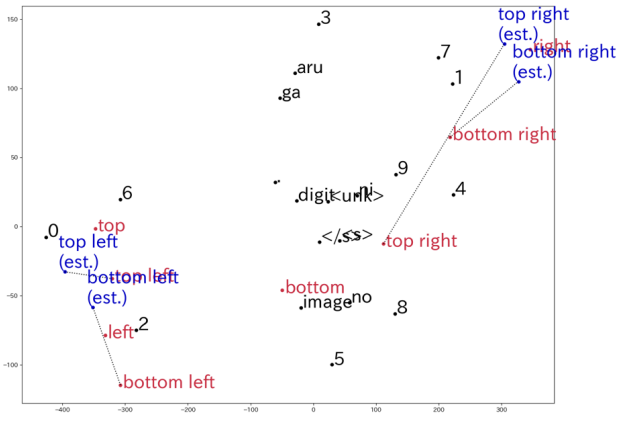


Figure 3: Obtained semantic compositionality.

## References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, Lille, France, volume 37 of *Proceedings of Machine Learning Research*, pages 1462–1471. http://proceedings.mlr.press/v37/gregor15.html.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Diederik P Kingma and Max Welling. 2014. Auto-encding variational bayes. In *Proceedings of the International Conference on Learning Representations*.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324. https://doi.org/10.1109/5.726791.

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2016. Generating images from captions with attention. In *Proceedings of the International Conference on Learning Representations*.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Mirrorgan: Learning text-to-image generation by redescription. *CoRR* abs/1903.05854. http://arxiv.org/abs/1903.05854.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, New York, New York, USA, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069. http://proceedings.mlr.press/v48/reed16.html.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. 2017. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41:1947–1962.

Zizhao Zhang, Yuanpu Xie, and Lin Yang. 2018. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.