Generating Images From Captions With Attention

Anonymous Author(s)
Affiliation
Address
email

Abstract

At the end.

1 Introduction

Statistical natural image modelling remains a fundamental problem in computer vision and image understanding. This has motivated recent approaches in generative modelling applied to natural images by employing deep neural networks for their inference and generative components. Image generative models studied previously, e.g. variants of Boltzmann Machines [1], [2] and Deep Belief Networks [3], are often restricted to learning unconditional models of images distribution or conditioned on simple structured annotations, for example classification labels. Despite the advances in the variational generative models [4][5], learning the highly structured natural image distribution alone in the high dimensional pixel space alone proves to be a difficult task. In real world, however, images rarely appear in isolation. They often accompanied by their unstructured textual descriptions on web pages and in books. One domain presents substantial amount of relating information of the other domain. The additional information from image and unstructured text description could be used to simplify the image modelling task.

There are two directions in learning a generative model of image and text. One approach is to learn a text generative model conditioned on the images. Significant amount of recent works has been focused on generating captions from images [6], [7], [8] and etc. The models take an image descriptor and generate unstructured texts through a recurrent decoder. By contrast, learning a generative model for image and text may also be studied by generating images correctly interpreting the text description. Generating high dimensional realistic images from their descriptions is a more difficult approach that combines two challenging components of language modelling and image generation. Namely, the model has to capture the semantic meaning expressed in the description and then use that knowledge to generate pixel intensities of the image. Although, the interesting high dimensional natural images lay on a small manifold that is difficult to capture, the additional text description cues of a target image may simplify the learning problem by focusing on the conditional distribution.

In this paper, we address the problem of image generation from unstructured natural language captions. By extending the Deep Recurrent Attention Writer (DRAW)[5], our model iteratively draws the patches on canvas, while attending to the relevant words in the description. Overall, the main contributions of this work are the following: we introduce a conditional alignDRAW model, a generative model of images using soft attention mechanism attending to different input words while generating images. The images generated by our alignDRAW model are post-process by. We then illustrate how our method generalizes over new color combinations and environments.

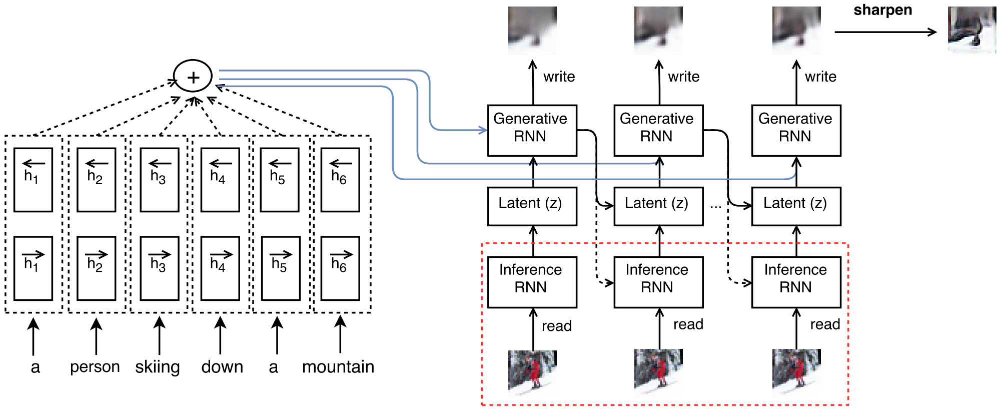

sharpen
write
write
write
Generative RNN
Generative RNN
Generative RNN
Latent (z)
Latent (z)
...
Latent (z)
Inference RNN
Inference RNN
Inference RNN
read
read
read
+
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_4$ $\overleftarrow{h}_5$ $\overleftarrow{h}_6$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_4$ $\overrightarrow{h}_5$ $\overrightarrow{h}_6$
a
person
skiing
down
a
mountain

Figure 1: The image of the proposed model

2 Model

Our proposed model can be viewed as a part of sequence-to-sequence framework [9], [10], [11] where captions are represented as a sequence of consecutive words and images are represented as a sequence of patches drawn on canvas over time $t = 1, ..., T$. Let $\mathbf{y}$ be the input caption, consisting of N words $y_1, y_2, ..., y_n$ and $\mathbf{x}$ be the image corresponding to that caption.

2.1 Language Model: the Bidirectional Attention RNN

The input caption sentences are fed into a deterministic Bidirectional LSTM[] that encodes the variable size sentences into the vector representation s. Bidirectional LSTM consists of one Forward LSTM and Backward LSTM which combine information from past and future respectively. The Forward LSTM computes the sequence of forward hidden states $[\overrightarrow{h}_1^{lang}, \overrightarrow{h}_2^{lang}, ..., \overrightarrow{h}_N^{lang}]$, whereas the Backward LSTM computes the sequence of backward hidden states $[\overleftarrow{h}_1^{lang}, \overleftarrow{h}_2^{lang}, ..., \overleftarrow{h}_N^{lang}]$. Then these hidden states are concatenated together into the sequence $[h_1^{lang}, h_2^{lang}, ..., h_N^{lang}]$, where $h_n^{lang} = [\overrightarrow{h}_n^{lang}, \overleftarrow{h}_n^{lang}], 1 \leq n \leq N$.

2.2 Image Model: the Conditional DRAW Network

The DRAW network [5] is a sequential probabilistic model generating images by accumulating the output at each iterative step. While the original DRAW network assumes the latent variables are independent, it has shown in [12] the model performance is improved by including the dependencies of latent variables.

We extended the architecture the DRAW network generative process to include additional input caption from the language model described in Sec. (2.1). Similarly to the original DRAW network, the conditional DRAW network is a stochastic recurrent neural network that consists of Inference LSTM that infers the distribution of latent variables of image x given y and then the Generative LSTM that uses the inferred latent variables in order to reconstruct the image x given y. The $align$ function is used to compute the alignment between the input caption and intermediate image generative steps as in [13]:

Formally, the image is generated by iteratively computing the following equations for $t = 1, ..., T$

$$\hat{x}_t = x - \boldsymbol{\sigma}(c_{t-1}) \tag{1}$$

$$r_t = read(x_t, \hat{x}_t, h_{t-1}^{dec}) \tag{2}$$

$$h_t^{enc} = LSTM^{enc}(h_{t-1}^{enc}, [r_t, h_{t-1}^{dec}]) \tag{3}$$

$$z_t \sim Q(Z_t | h_t^{enc}) \tag{4}$$

$$h_t^{dec} = LSTM^{dec}(h_{t-1}^{dec}, z_t, s_{t-1}) \tag{5}$$

$$s_t = align(h_{t-1}^{dec}, \boldsymbol{h^{lang}}) \tag{6}$$

$$c_t = c_{t-1} + write(h_t^{dec}) \tag{7}$$

where $read$ and $write$ are the same attention operators as in [5]. Given the caption representation from the language model, $\boldsymbol{h^{lang}} = [h_1^{lang}, h_2^{lang}, ..., h_N^{lang}]$, the $align$ operator computes the final sentence representation s_t through a weighted sum using alignment probabilities $\alpha_{1...N}$:

$$s_t = align(h_{t-1}^{dec}, \boldsymbol{h^{lang}}) = \alpha_1 h_1^{lang} + \alpha_2 h_2^{lang} + ... + \alpha_N h_N^{lang}. \tag{8}$$

The corresponding alignment probabilities $\alpha_{1...n}$ at each step are obtained by:

$$e_{tj} = v^T tanh(U h_j^{lang} + W h_t^{dec} + b) \tag{9}$$

$$\alpha_j = \frac{exp(e_{tj})}{\sum_{j=1}^{N} exp(e_{tj})}. \tag{10}$$

Here h_0^{lang} is initialized to the learned bias. Setting $\alpha_{1...N}$ to $\frac{1}{N}$ turns the encoder into the vanilla model introduced in [10] without the attention.

2.3 Learning

The model is learned by the modified version of Stochastic Gradient Variation Bayes (SGVB) algorithm introduced by [4]. The model is trained to maximize the lower bound of marginal likelihood $\mathcal{L}$ of the correct image x given the input caption y. The $\mathcal{L}$ is decomposed into the latent loss $\mathcal{L}^z$ and the reconstruction loss $\mathcal{L}^x$.

The reconstruction loss $\mathcal{L}^x$ equals to $\frac{1}{L}\sum_{l=1}^{L}(log\, p(x_t|y, z)$ where L is the number of samples used during training, which was set to 1 in our experiments.

The latent loss $\mathcal{L}^z$ is a negative sum of Kullback–Leibler divergence terms between distribution $Q(Z_t|h_t^{enc})$ and some prior distribution $P(Z_t)$ over time $t = 1, ..., T$, which can be seen as a regularization term. Since the patches drawn on canvas over time are not independent of each other, naturally the sufficient statistics of the prior distribution at time t should be dependent on the sufficient statistics of the prior distribution at time $t - 1$. Therefore, instead of setting $P(Z_1), ..., P(Z_T)$ to be independent unit gaussian distributions, the mean and variance of $P(Z_t)$ depends on the h_{t-1}^{dec}, which forms a Markov chain $P(Z_1), P(Z_2 \,|\, Z_1), ..., P(Z_T \,|\, Z_{T-1})$ as in [12], where

$$\mu_t^{prior} = tanh(W_\mu h_{t-1}^{dec}) \tag{11}$$

$$\sigma_t^{prior} = exp(tanh(W_\sigma h_{t-1}^{dec})) \tag{12}$$

Overall, the loss function $\mathcal{L}$ is calculated as follows:

$$\begin{aligned}\mathcal{L} =& \mathbb{E}_{Q(Z_{1:t} \,|\, \mathbf{y}, \mathbf{x})}\left[-\log p(\mathbf{x} \,|\, \mathbf{y}, Z_{1:T}) + \sum_{t=2}^{T} \mathrm{D_{KL}}\left(Q(Z_t \,|\, Z_{t-1}, \mathbf{y}, \mathbf{x}) \,\|\, P(Z_t \,|\, Z_{t-1}, \mathbf{y})\right)\right] \\ &+ \mathrm{D_{KL}}\left(Q(Z_1 \,|\, \mathbf{y}, \mathbf{x}) \,\|\, P(Z_1 \,|\, \mathbf{y})\right).\end{aligned} \tag{13}$$

The expectation can be approximate by L Monte Carlo samples $\tilde{Z}_{1:T}$ from $Q(Z_{1:T} \,|\, \mathbf{y}, \mathbf{x})$:

$$\begin{aligned}\mathcal{L} \approx& \frac{1}{L}\sum_{l=1}^{L}\left[-\log p(\mathbf{x} \,|\, \mathbf{y}, \tilde{Z}_{1:T}^l) + \sum_{t=2}^{T} \mathrm{D_{KL}}\left(Q(Z_t \,|\, \tilde{Z}_{t-1}^l, \mathbf{y}, \mathbf{x}) \,\|\, P(Z_t \,|\, \tilde{Z}_{t-1}^l, \mathbf{y})\right)\right] \\ &+ \mathrm{D_{KL}}\left(Q(Z_1 \,|\, \mathbf{y}, \mathbf{x}) \,\|\, P(Z_1 \,|\, \mathbf{y})\right).\end{aligned} \tag{14}$$

2.4 Generating Images from Captions

During the image generation step, we throw away the inference network from the decoder and instead sample from the prior distribution. Due to the blurriness of samples generated by DRAW model, we do an additional post processing step, where we use the generator of the adversarial network trained on residuals of laplacian pyramid to sharpen the generated images, similar to [14]. By fixing the prior of adversarial generator to the mean of uniform distribution, it gets treated as a deterministic neural network which allows us to calculate the lower bound of likelihood. The reconstuction loss becomes the loss between sharpened image and correct image, whereas the latent loss stays the same. We also noticed that sampling from the mean of uniform distribution allowed us to generate much less noisy samples than by sampling from the uniform distribution itself.

3 Experiments

3.1 COCO

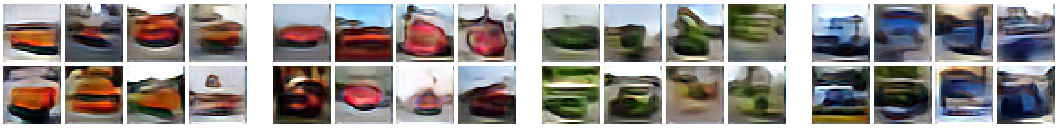
Microsoft COCO [15] is a very large dataset of roughly 83k images, each annotated with 5 captions. The rich collection of images with a wide variety of styles, backgrounds and objects makes the task of learning a good generative model conditioned on a caption very challenging. For consistency with related work on caption generation, we disregard four of the five captions when training our model.

In the following subsections, we analyze both the qualitative and quantitative aspects of our model as well as compare its performance with that of other, related generative models.

3.1.1 Analysis of Generated Images

The main goal of this work is to learn a model that can understand the semantic meaning expressed in the textual descriptions of images, such as the properties of objects, the relationships between them, etc. and then use that knowledge to generate relevant images. To verify that, we wrote a set of captions inspired by the COCO dataset and changed some words in the captions to see whether the model made the relevant changes in the generated samples.

First, we wanted to see whether the model understood one of the most basic properties of any object, the color. In Figure 2, we generated images of school buses with four different colors: yellow, red, green and blue. Although, there are images of buses with different colors in the training set, all mentioned school buses are specifically colored yellow. Despite that, the model managed to generate images of an object that is visually reminiscent of a school bus that is painted with the specified color.



A yellow school bus parked in a parking lot.

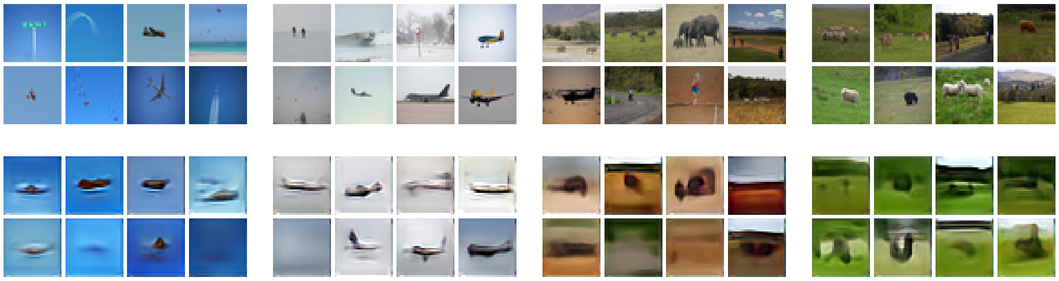
A red school bus parked in a parking lot.

A green school bus parked in a parking lot.

A blue school bus parked in a parking lot.

Figure 2: Examples of changing the color while keeping the caption fixed.

Apart from changing the colors of objects, we were curious whether changing the background of the scene described in a caption would result in the appropriate changes in the generated samples. The task of changing the background of an image is somewhat harder than just changing the color of an object because the model will have to make alterations over a wider visual area. Nevertheless, as shown in Figure 3 changing the skies from blue to rainy in a caption as well as changing the grass type from dry to green in another caption resulted in the appropriate changes in the generated image. The nearest images from the training set also indicate that the model was not simply copying the patterns it observed during the learning phase.



A very large commercial plane flying in blue skies.

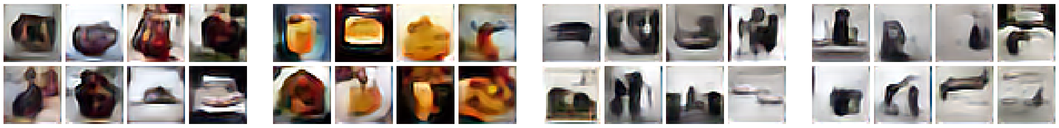
A very large commercial plane flying in rainy skies.

A herd of elephants walking across a dry grass field.

A herd of elephants walking across a green grass field.

Figure 3: Examples of changing the background while keeping the caption fixed. The respective nearest training images based on pixel-wise L2 distance are displayed on top.

Despite an infinite number of ways of changing colors and backgrounds in descriptions, in general we found that the model made appropriate changes as long as some similar pattern was present in the training set. However, the model struggled when the visual difference between objects was very small, such as the objects having the same general shape and colors. In Figure 4, we demonstrate that when we swap two objects that are both visually similar, for example cats and dogs, it is difficult to discriminate solely from the generated samples whether it is an image of a cat or dog, even though we might notice an animal-like shape. This highlights a limitation of the model in that it has difficulty modelling the fine-grained details of objects.



The decadent chocolate desert is on the table.

A bowl of bananas is on the table.

A vintage photo of a cat.

A vintage photo of a dog.

Figure 4: Examples of changing the object while keeping the caption fixed.

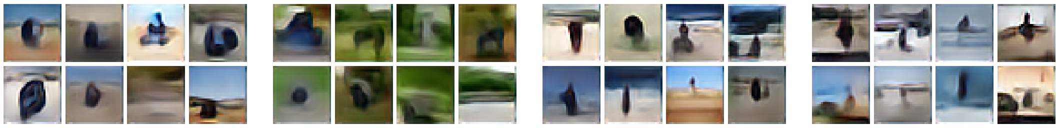
3.1.2 Analysis of Attention

After flipping set of words in the captions, we were curious to see which words did the model attend to when generating images. It turned out that during the generation step, the model mostly focused on the specific words that carried out the main semantic meaning expressed in the sentences. The attention values in sentences helped us interpret the reasons why the model did or didn't make relevant changes in images when flipping words. For example, as shown in Figure 5 despite flipping word "desert" to "forest", the model equally looked at these words and thus made the correct changes. Whereas, flipping between words "beach" and "sun" has resulted in no change, due to the model ignoring the word "sun".

We also tried to analyze the way the model generated images. Unfortunately, we found that there was no connection between the patches drawn on canvas and the most attended words at particular timesteps.

3.1.3 Comparison With Other Models

The quantitative evaluation of generative models has been a subject of ambiguity in a machine learning community. Compared to reporting classification accuracies in discriminative models, the measures defining generative models are intractable most of the times and might not correctly define the real quality of the model. To get a better comparisions between performances of generative models, we report results on two different metrics as well as do some qualitative comparison of different generative models.



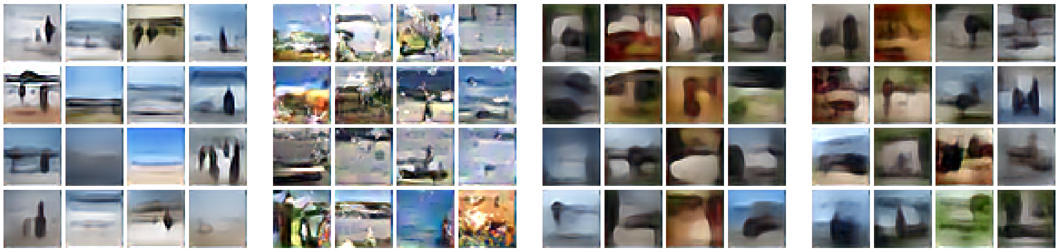
A rider on a blue motorcycle in the desert.

A rider on a blue motorcycle in the forest.

A surfer, a woman, and a child walk on the beach.

A surfer, a woman, and a child walk on the sun.

Figure 5: Examples of most attended words while changing the background in the caption.



Align DRAW

LAPGAN

Conv. VAE

Fully-conn. VAE

Figure 6: Four different models displaying results from sampling caption *A group of people walk on a beach with surf boards.*

As you can see on Figure 6, we generated several samples from prior of each of the current state-of-the-art generative models corresponding to the caption “A group of people walk on a beach with surf boards”. While all of the samples look sharp, the images generated by LAPGAN look more noisy and don’t have a very clear structure in them, whereas the images generated by variational models trained with L2 cost function have a watercolor effect on the images.

As for the quantitative comparison of different models, we first compare the performances of the model trained with variational methods. We rank the images in test set conditioned on the captions based on the variational lower bound of likelihood and then report the Precision-Recall metric to evaluate the quality of the generative model. As we expected, the quality of image retrieval using generative models is worse compared to the disriminative models that were specifically build for retrieval. To deal with large computational complexity of looping through each test image, we create a shortlist of hundred images including the correct one, based on the convolutional features of VGG like model trained on CIFAR dataset. Since there are “easy” images for which the model assigns high likelihood independent of the query caption, we look at the ratio of likehood of image conditioned on the sentence to likelihood of image conditioned on the mean sentence representation in the training set. We found that the reconstruction error $\mathcal{L}^x$ increased for the sharpened images that considerably hurt the retrieval results. Since sharpening changes the statistics of images, computing reconstruction error for each pixel is not necessarily a good metric.

Instead of calculating error per each pixel, we turn to the smarter metric, such as Structural Similarity Index (SSI), which incorporates luminace and contrast masking into the error calculation. Due to strong inter-dependencies of closer pixels, the metric is calculated on the small windows of the image. To calculate SSI, we sampled fifty images from prior of each generative model per each caption in the test set.

References

[1] Paul Smolensky. Information processing in dynamical systems: foundations of harmony theory. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1986.

[2] Ruslan Salakhutdinov et al. Deep boltzmann machines. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, pages 448–455, 2009.

[3] Geoffrey E. Hinton, et al. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

COCO (before sharpening)

	Image Search					Image Similarity
Model	**R@1**	**R@5**	**R@10**	**R@50**	**Med r**	**SSI**
LAPGAN	-	-	-	-	-	0.08
Fully-conn. VAE (L2 cost)	1.0	5.6	10.4	51.1	51	0.156
Conv. VAE (L2 cost)	1.0	5.9	10.9	50.8	50	**0.164**
Skipthought DRAW	2.0	11.2	18.9	63.3	36	0.157
Noalign DRAW	2.8	**14.1**	**23.1**	68.0	**31**	0.155
Align DRAW	**3.0**	14.0	22.9	**68.5**	**31**	0.156

[4] Diederik P. Kingma et al. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[5] Karol Gregor, et al. DRAW: A recurrent neural network for image generation. *CoRR*, abs/1502.04623, 2015.

[6] Andrej Karpathy et al. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

[7] Kelvin Xu, et al. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057, 2015.

[8] Ryan Kiros, et al. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[9] Ilya Sutskever, et al. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. 2014.

[10] Kyunghyun Cho, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1724–1734, 2014.

[11] Nitish Srivastava, et al. Unsupervised learning of video representations using LSTMs. *ICML*, 2015.

[12] Philip Bachman et al. Data generation as sequential decision making. *CoRR*, abs/1506.03504, 2015.

[13] D. Bahdanau, et al. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[14] Emily L. Denton, et al. Deep generative image models using a laplacian pyramid of adversarial networks. *CoRR*, abs/1506.05751, 2015.

[15] T.Y. Lin, et al. Microsoft COCO: Common objects in context. In *ECCV*, 2014.