

ChatPainter: Improving Text to Image Generation using Dialogue

Shikhar Sharma¹ Dendi Suhubdy^{2,3} Vincent Michalski^{2,3,1} Samira Ebrahimi Kahou¹ Yoshua Bengio^{2,3}

Abstract

Synthesizing realistic images from text descriptions on a dataset like Microsoft Common Objects in Context (MS COCO), where each image can contain several objects, is a challenging task. Prior work has used text captions to generate images. However, captions might not be informative enough to capture the entire image and insufficient for the model to be able to understand which objects in the images correspond to which words in the captions. We show that adding a dialogue that further describes the scene leads to significant improvement in the inception score and in the quality of generated images on the MS COCO dataset.

1. Introduction

Automatic generation of realistic images from text descriptions has numerous potential applications, for instance in image editing, in video games, or for accessibility. Spurred by the recent successes of Variational Autoencoders (VAEs) (Kingma & Welling, 2014) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Denton et al., 2015; Radford et al., 2016), there has been a lot of recent work and interest in the research community on image generation from text captions (Mansimov et al., 2016; Reed et al., 2016a; Zhang et al., 2017; Xu et al., 2017).



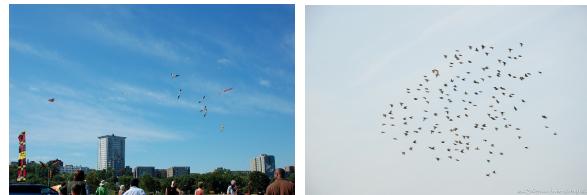
This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments

Figure 1. Caption and corresponding generated image from the StackGAN model (Zhang et al., 2017). Image reproduced with permission from authors.

Current state-of-the-art models are capable of generating

¹Microsoft Research, Montréal, Canada ²Université de Montréal, Montréal, Canada ³Montreal Institute for Learning Algorithms, Montréal, Canada. Correspondence to: Shikhar Sharma <shikhar.sharma@microsoft.com>.

realistic images on datasets of birds, flowers, room interiors, faces etc., but don't do very well on datasets like MS COCO (Lin et al., 2014) which contain several objects within a single image and where subjects are not always centred in the image. A caption for an image of a flower can usually describe most of the relevant details of the flower (see Figure 1). However, for the MS COCO dataset a caption might not contain all the relevant details about the foreground and the background. As can be seen from Figure 2, it is possible for two very similar MS COCO captions to correspond to very different images. Due to this complexity, a caption can be considered a noisy descriptor and due to the limited amount of image-caption paired data, the model might not always be able to understand which objects in the image correspond to which words in the caption. Previous work in the literature has found that conditioning on auxiliary data such as category labels (Mirza & Osindero, 2014), or object location and scale (Reed et al., 2016b) helps in improving the quality of generated images and in making them more interpretable.



(a) A flock of birds flying in a blue sky. (b) A flock of birds flying in an overcast sky

Figure 2. Two very different looking images can have similar captions in the COCO dataset and the captions also might not describe the image fully.

Sketch artists typically have a back and forth conversation with witnesses when they have to draw a person's sketch, where the artist asks for more details and draws the sketch while the witnesses provide requested details and feedback on the current state of the sketch. We hypothesize that conditioning on a similar conversation about a scene in addition to a caption would significantly improve the generated image's quality and we explore this idea in this paper. For this, we pair captions provided with the MS COCO dataset with dialogues from the Visual Dialog dataset (VisDial) (Das et al.,

2017). These dialogues were collected using a chat interface pairing two workers on Amazon Mechanical Turk (AMT). One of them was assigned the role of an ‘answerer’, who could see an MS COCO image together with its caption and had to answer questions about that image. The other was assigned the role of the ‘questioner’ and could see only the image’s caption. The questioner had to ask questions to be able to imagine the scene more clearly. Similar to a dialogue between a sketch artist who gradually refines an image and a witness describing a person, the VisDial dialogue turns iteratively add fine-grained details to a (mental) image.

In this paper,

- We use VisDial dialogues along with MS COCO captions to generate images. We show that this results in the generation of better quality images.
- We provide results indicating that our model obtains a higher inception score than the baseline StackGAN model which uses only captions.

Though we just demonstrate improvements over the StackGAN (Zhang et al., 2017) model in this paper, this additional dialogue module can be added to any caption-to-image-generation model and is an orthogonal contribution.

2. Related Work

In the past, variationally trained models have often been used for image generation. A significant drawback of these models has been that they tend to generate blurry images. Among latent variable based variationally trained models, Kingma et al. (2016) proposed inverse autoregressive flows, where they inverted the sequential data generation of an autoregressive model, which helped in parallelizing computation. They presented results on MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009) datasets. Recently, Cai et al. (2017) added residual blocks (He et al., 2016) and skip connections to their decoder and generated images in multiple stages. The initial components produce a coarse image and the final components refine the previously generated image. Their deep residual VAE was able to generate sharper images on MNIST and CelebA (Liu et al., 2015) compared to previous VAEs. Among tractable likelihood models, Van Den Oord et al. (2016) proposed the PixelRNN model which predicts pixels of an image sequentially along the rows or along the diagonal using fast two-dimensional recurrent layers. They achieved significantly improved log-likelihood scores on the MNIST, CIFAR-10 and ImageNet (Deng et al., 2009) datasets. Makhzani et al. (2016) proposed an adversarial autoencoder model where they use GANs to perform variational inference and achieve competitive performance on both generative and semi-supervised classification tasks.

GANs have received attention recently because they produce sharper images (Goodfellow et al., 2014; Denton et al., 2015; Radford et al., 2016) compared to other generative models. They however suffer from several issues such as ‘mode collapse’ (i.e. the generator learns to generate samples from only a few modes of the distribution), lack of variation among generated images, and instabilities in the training procedure. Training GANs has generally required careful design of the model architecture and a balance between optimization of the generator and the discriminator. Arjovsky et al. (2017) proposed minimizing an approximation of the Earth Mover (Wasserstein) distance between the real and generated distributions. Their model, Wasserstein GAN (WGAN), is much more stable than previous approaches and reduces most of the aforementioned issues affecting GANs. Additionally, the reduction of the critic’s (the discriminator is called the ‘critic’ in this work) loss correlates with better sample quality which is a desirable property. Gulrajani et al. (2017) further improved upon these issues by removing weight clipping from the WGAN and instead adding a gradient penalty (WGAN-GP) for the gradient of the critic. Recently, Karras et al. (2018) have produced high-quality high resolution images (1024×1024) by progressively growing both the generator and the discriminator layer-by-layer and by using the WGAN-GP loss. Apart from faster training time, they found that by adding layers progressively, training stability is improved significantly. Among recent efforts to stabilize the training of GANs via noise-induced regularization, Roth et al. (2017) reduced several failure modes of GANs by penalizing a weighted gradient-norm of the discriminator. They observed stability improvements and better generalization performance. Miyato et al. (2018) proposed a spectral weight normalization technique for GANs in which they control the Lipschitz constant of the discriminator function resulting in global regularization of the discriminator. Gradient analysis of the spectrally normalized weights shows that their technique prevents layers from becoming sensitive in a single direction. This approach yields more complexity and variation in generated samples compared to previous weight normalization methods.

Apart from generating images directly from noise with GANs, there has been recent work on conditioning the generator or discriminator or both on additional information. Mirza & Osindero (2014) introduced the idea of conditioning both the generator and discriminator on extra information such as class labels. They ran experiments on both unimodal image data and multi-modal image-metadata-annotations data. Their experiments resulted in better Parzen window-based log-likelihood estimates for MNIST compared to unconditioned GANs. On the MIR Flickr 25 000 dataset (Huiskes & Lew, 2008), they generated metadata tags conditioned on images. Odena et al. (2017)

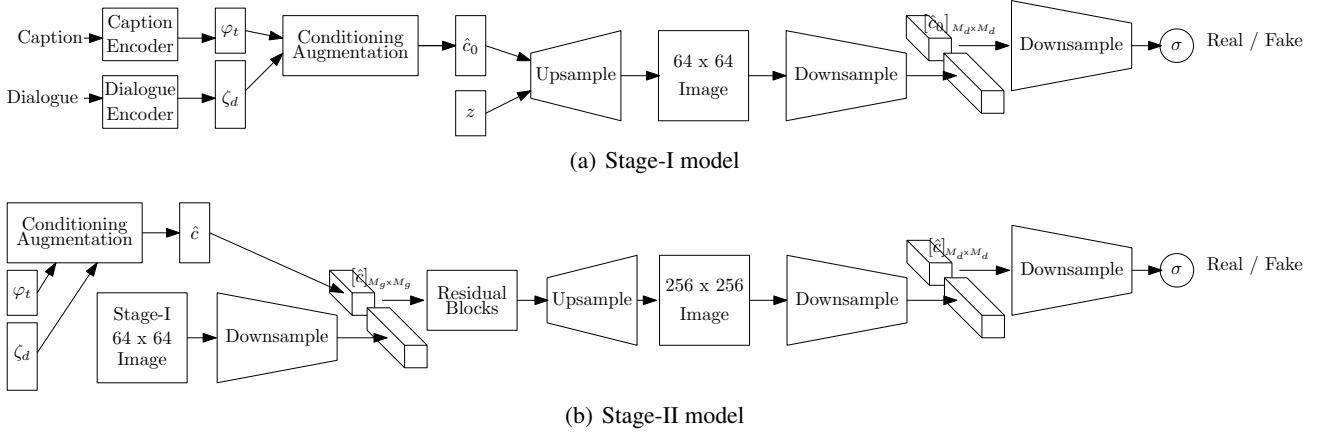


Figure 3. ChatPainter: (a) Stage-I of the model generates a 64×64 image conditioned on a caption and the corresponding dialogue. (b) Stage-II of the model generates a 256×256 image conditioned on Stage-I’s 64×64 generated image and the caption and corresponding dialogue.

conditioned their generator on both noise and class labels and their discriminator additionally classified images into classes. By explicitly making the generator and discriminator aware of class labels, they were able to generate better quality images on larger multi-class datasets and higher image variability compared to previous work. Their experiments also indicated that generating higher resolution images yielded higher discriminability.

A lot of work has been done in recent years to generate MS COCO images from captions. Mansimov et al. (2016) used a conditional DRAW (Gregor et al., 2015) model with soft attention over the words of the caption to generate images on the MS COCO dataset and then sharpened them with an adversarial network. However, their generated images were low resolution (32×32) and generated blob-like objects in most cases. Building upon other work in GANs, Nguyen et al. (2017) introduced a prior on the latent code used by their generator. They ran an optimization procedure to find the latent code which the generator takes as its input. This procedure maximized the activations of an image captioning network run over the generated image. This method produced high-quality and diverse images at high resolution (227×227). Reed et al. (2016a) trained a Deep Convolutional Generative Adversarial Network (DCGAN) with the generator and discriminator both conditioned on features from a character-level convolutional Recurrent Neural Network (RNN) encoder over the captions to generate visually-plausible 64×64 images. Dash et al. (2017) additionally trained their discriminator to classify images similar to Odena et al. (2017) and generated 128×128 images. Zhang et al. (2017) also conditioned their generator and discriminator on caption encodings but their StackGAN model generates images in multiple stages – stage-I gener-

ates a coarse low resolution (64×64) image and stage-II generates the final high resolution (256×256) image. This stacking of models resulted in generation of highly photo-realistic images, at higher resolutions compared to previous work, on datasets of flowers (Nilsback & Zisserman, 2008) and birds (Wah et al., 2011) and many good looking images on the MS COCO dataset as well. However, they did not train their two stages in an end-to-end fashion. Stage-I was trained to completion first and then Stage-II was trained. Very recently, Xu et al. (2017) increased the number of stages, trained them in an end-to-end fashion, added an attention mechanism over the captions, as well as added a novel attentional multimodal similarity model to guide the training loss, which resulted in significantly increased performance and the state-of-the-art inception score (Salimans et al., 2016) of 25.89 on the MS COCO dataset. Hong et al. (2018) first generated a semantic layout map of the objects in the image and then conditioned on the map and the caption to generate semantically meaningful 128×128 images.

3. Data

In our experiments, we used images and their captions from the MS COCO (Lin et al., 2014) dataset. MS COCO covers 91 categories of objects, grouped into 11 super-categories of objects such as *person and accessory*, *animal*, *vehicle*, etc. We use the ‘2014 Train’ set as our training set and the ‘2014 Val’ set as our test set. The train set consists of $\sim 80K$ images and includes five captions for each image. The test set consists of $\sim 40K$ images along with their captions.

We obtain dialogues for these images from the VisDial (Das et al., 2017) dataset. VisDial consists of 10 question-answer

Table 1. An example of the input data, the corresponding dataset image, and the image generated by our best ChatPainter model.

Input	Dataset image	Generated image
Caption: adult woman with yellow surfboard standing in water.		
Q: is the woman standing on the board?	A: no she is beside it.	
Q: how much of her is in the water?	A: up to her midsection.	
Q: what color is the board?	A: yellow.	
Q: is she wearing sunglasses?	A: no.	
Q: what about a wetsuit?	A: no she has on a bikini top.	
Q: what color is the top?	A: orange and white.	
Q: can you see any other surfers?	A: no.	
Q: is it sunny?	A: the sky isn't visible but it appears to be a nice day.	
Q: can you see any palm trees?	A: no.	
Q: what about mountains?	A: no.	

conversation turns per dialogue and has one dialogue for each of the MS COCO images. VisDial was collected by pairing two crowd-workers and having them talk about an image as described in Section 1. Hence, we have $\sim 80K$ dialogues for the training set and $\sim 40K$ for the test set.

4. Model

We build upon the StackGAN model introduced by Zhang et al. (2017). StackGAN generates an image in two stages where Stage-I generates a coarse 64×64 image and Stage-II generates a refined 256×256 image. We try to use the same notation everywhere as used in the original StackGAN paper. Our model ChatPainter's architecture is shown in Figure 3 and described below.

We generate caption embedding φ_t by encoding the captions with a pre-trained encoder¹ (Reed et al., 2016a). We generate dialogue embeddings ζ_d by two methods:

- **Non-recurrent encoder** We collapse the entire dialogue into a single string and encode it with a pre-trained Skip-Thought (Kiros et al., 2015) encoder².
- **Recurrent encoder** We generate Skip-Thought vectors for each turn of the dialogue and then encode them with a bidirectional LSTM-RNN (Graves & Schmidhuber, 2005; Hochreiter & Schmidhuber, 1997).

We then concatenate the caption and dialogue embeddings and this is passed as input to the Conditioning Augmentation (CA) module. The CA module was introduced by Zhang et al. (2017) to produce latent variable inputs for the generator from the embeddings. They also proposed a regularization term to encourage smoothness over the conditioning manifold which we adapt for our additional dialogue embeddings:

$$D_{KL}(\mathcal{N}(\mu(\varphi_t, \zeta_d), \text{diag}(\sigma(\varphi_t, \zeta_d))) || \mathcal{N}(0, I)), \quad (1)$$

¹<https://github.com/reedscot/icml2016>

²<https://github.com/ryankiros/skip-thoughts>

where D_{KL} is the Kullback-Leibler divergence. In the CA module, a fully connected layer is applied over the input that generates μ and σ which are both N_g dimensional. The module samples ϵ from $\mathcal{N}(0, I)$. Finally, the conditioning variables \hat{c} are computed as

$$\hat{c} = \mu + \sigma \odot \epsilon, \quad (2)$$

where \odot is the element-wise multiplication operator. Thus, the conditioning variables \hat{c} are effectively samples from $\mathcal{N}(\mu(\varphi_t, \zeta_d), \text{diag}(\sigma(\varphi_t, \zeta_d)))$.

4.1. Stage-I

The conditioning variables for Stage-I, \hat{c}_0 , are concatenated with N_z -dimensional noise, z , drawn from a random normal distribution, p_z . The Stage-I generator upsamples this input representation to a $W_0 \times H_0$ image. This Stage-I image is expected to be blurry and a rough version of the final one. The discriminator downsamples this image to $M_d \times M_d \times N_{di}$. \hat{c}_0 is then spatially replicated to $M_d \times M_d \times N_d$ and concatenated with the downsampled representation. This is further downsampled to a scalar value between 0 and 1. The model is trained by alternating between maximizing \mathcal{L}_{D_0} and minimizing \mathcal{L}_{G_0} :

$$\begin{aligned} \mathcal{L}_{D_0} = & \mathbb{E}_{(I_0, t, d) \sim p_{data}} [\log D_0(I_0, \varphi_t, \zeta_d)] + \\ & \mathbb{E}_{z \sim p_z, (t, d) \sim p_{data}} [\log(1 - D_0(G_0(z, \hat{c}_0), \varphi_t, \zeta_d))], \end{aligned} \quad (3)$$

$$\begin{aligned} \mathcal{L}_{G_0} = & \mathbb{E}_{z \sim p_z, (t, d) \sim p_{data}} [\log(1 - D_0(G_0(z, \hat{c}_0), \varphi_t, \zeta_d))] \\ & + \lambda D_{KL}(\mathcal{N}(\mu(\varphi_t, \zeta_d), \text{diag}(\sigma(\varphi_t, \zeta_d))) || \mathcal{N}(0, I)), \end{aligned} \quad (4)$$

where I_0 is the real image, t is the text caption, d is the dialogue, p_{data} is the true data distribution, λ is the regularization coefficient, G_0 is the Stage-I generator, and D_0 is the Stage-I discriminator.

In our experiments, $N_z = 100$, $W_0 = 64$, $H_0 = 64$, $M_d = 4$, $N_{di} = 512$, $N_d = 128$, and $\lambda = 2$ – same as that in the StackGAN model.

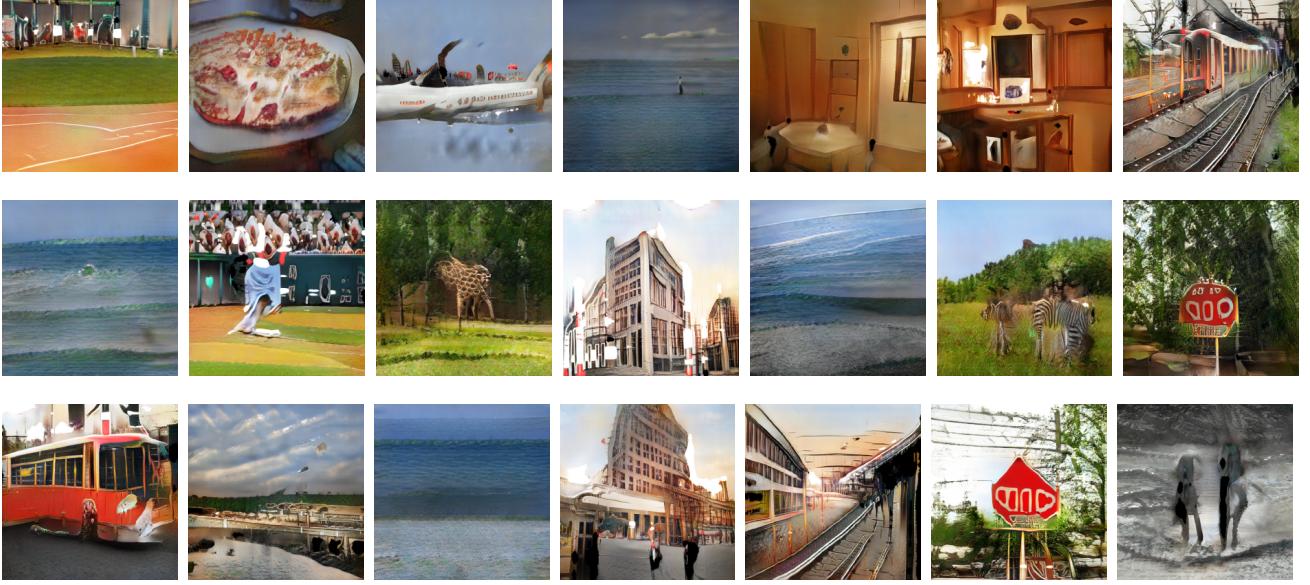


Figure 4. Example 256×256 images generated by our non-recurrent encoder ChatPainter model on the MS COCO test set. Best viewed in color. Images are cherry-picked from a larger random sample.

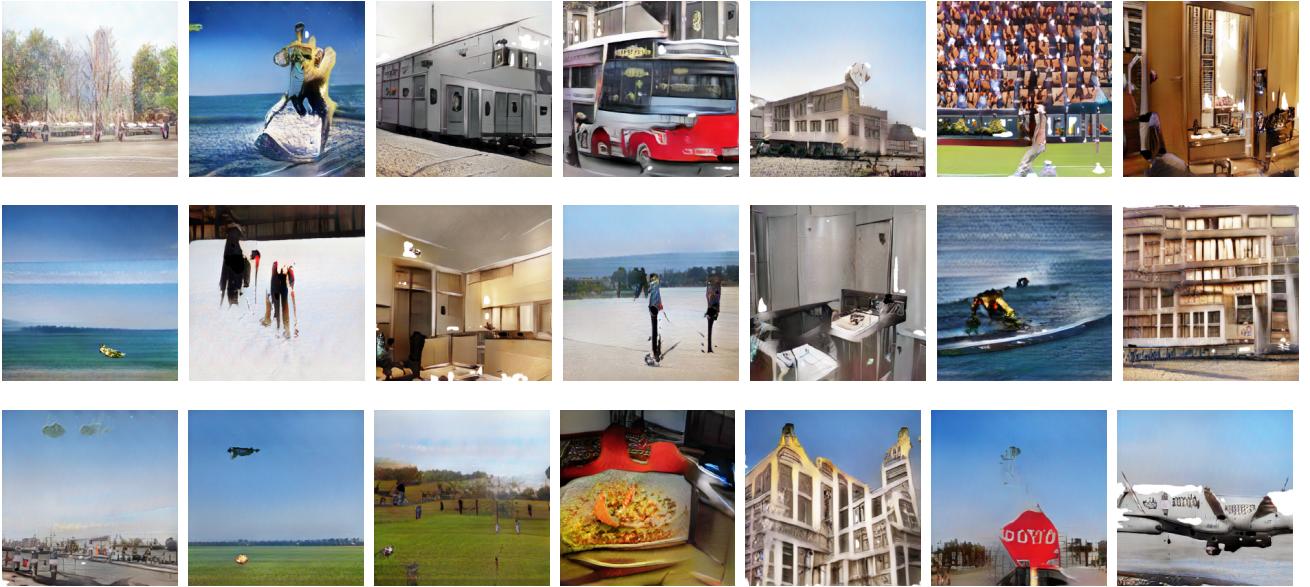


Figure 5. Example 256×256 images generated by our recurrent encoder ChatPainter model on the MS COCO test set. Best viewed in color. Images are cherry-picked from a larger random sample.

4.2. Stage-II

The Stage-II generator, G , first downsamples generated stage-I images to $M_g \times M_g \times N_{gi}$. The conditioning variables for Stage-II, \hat{c} , are generated and then spatially replicated to $M_g \times M_g \times N_g$ and finally concatenated to the downsampled image representation. For Stage-II training, in case of the recurrent dialogue encoder, the RNN weights

are copied from Stage-I and kept fixed. The concatenated input is passed through a series of residual blocks and is then upsampled to a $W \times D$ image. The Stage-II discriminator, D , downsamples the input image to $M_d \times M_d \times N_{di}$. \hat{c} is then spatially replicated to $M_d \times M_d \times N_d$ and concatenated with the downsampled representation which is further downsampled to a scalar value between 0 and 1. The Stage-II model is trained by alternating between maximizing \mathcal{L}_D

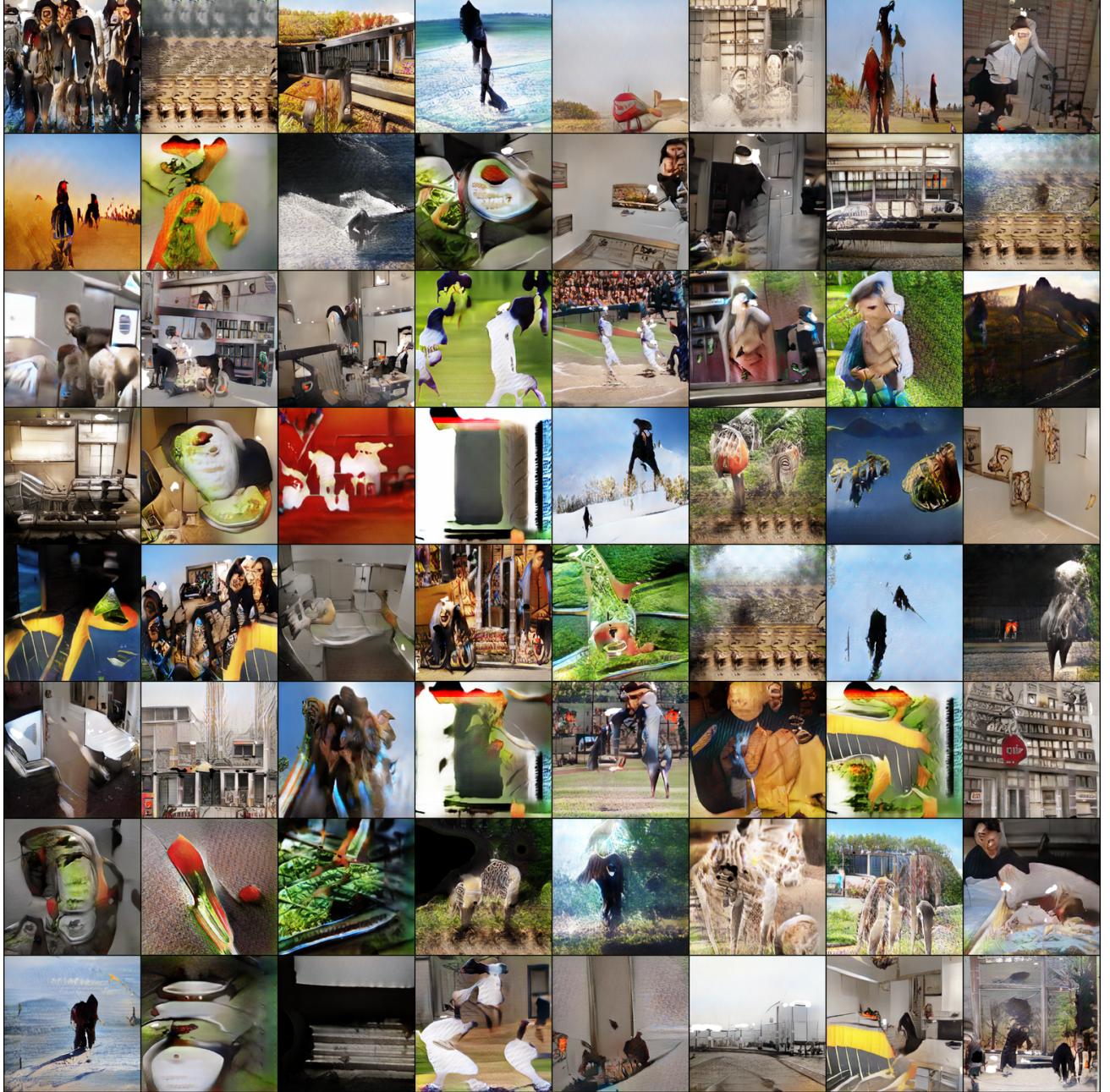


Figure 6. Example 256×256 images generated by our recurrent encoder ChatPainter model on the MS COCO test set. Best viewed in color. Images have been selected randomly.

and minimizing \mathcal{L}_G :

$$\begin{aligned} \mathcal{L}_D = & \mathbb{E}_{(I, t, d) \sim p_{data}} [\log D(I, \varphi_t, \zeta_d)] + \\ & \mathbb{E}_{s_0 \sim p_{G_0}, (t, d) \sim p_{data}} [\log(1 - D(G(s_0, \hat{c}), \varphi_t, \zeta_d))], \end{aligned} \quad (5)$$

$$\begin{aligned} \mathcal{L}_G = & \mathbb{E}_{s_0 \sim p_{G_0}, (t, d) \sim p_{data}} [\log(1 - D(G(s_0, \hat{c}), \varphi_t, \zeta_d))] \\ & + \lambda D_{KL}(\mathcal{N}(\mu(\varphi_t, \zeta_d), \text{diag}(\sigma(\varphi_t, \zeta_d))) || \mathcal{N}(0, I)), \end{aligned} \quad (6)$$

where I is the real image, and s_0 is the image generated from Stage-I.

In our experiments, $M_g = 16$, $N_{gi} = 512$, $N_g = 128$, $W = 256$, $D = 256$, $N_{di} = 512$, $N_d = 128$, and $\lambda = 2$ – same as that in the StackGAN model.

The architecture of the upsample, downsample and residual blocks, as shown in Figure 3 and as mentioned above in

Table 2. Inception scores for generated images on the MS COCO test set³.

Model	Inception Score
Reed et al. (2016a)	7.88 ± 0.07
StackGAN (Zhang et al., 2017)	8.45 ± 0.03
ChatPainter (non-recurrent)	9.43 ± 0.04
ChatPainter (recurrent)	9.74 ± 0.02
Hong et al. (2018)	11.46 ± 0.09
AttnGAN (Xu et al., 2017)	25.89 ± 0.47

the model details, is kept the same as that of the original StackGAN.

4.3. Training details

Similar to StackGAN, we use a matching-aware discriminator (Reed et al., 2016a), that is trained using “real” pairs consisting of a real image together with matching caption and dialogue, and “fake” pairs that consist either of a real image together with another images’s caption and dialogue or a generated image with the corresponding caption and dialogue. We train both stages for 800 epochs using the Adam optimizer (Kingma & Ba, 2015). The initial learning rate for all experiments is 0.0002. We decay the learning rate to half of its previous value after every 50 epochs. For Stage-I, we use a batch size of 384 and for Stage-II, we use a batch size of 64. In case of the recurrent dialogue encoder, the hidden dimension of the RNN is set to 1024. The implementation is based on PyTorch (Paszke et al., 2017) and we trained the models on a machine with 4 NVIDIA Tesla P40s.

5. Results

Table 1 shows the corresponding caption, dialogue inputs, and the test set image for an image generated by our best ChatPainter model. We present some of the more realistic images generated by our non-recurrent encoder ChatPainter in Figure 4, and by our recurrent encoder ChatPainter in Figure 5. For fairness of comparison, we also present a random sample of the images generated by our recurrent encoder ChatPainter on the MS COCO dataset in Figure 6. As seen from these figures, the model is able to generate close-to-realistic images for some of the caption and dialogue inputs though not very realistic ones for most.

We report inception scores on the images generated from our models in Table 2 and compare with other recent models. For computing inception score, we use the Inception

³The two best-performing methods were released while writing this manuscript and we will evaluate the effect of our scheme using these methods as base architecture in future work.

v3 model pretrained on ImageNet available with PyTorch. We then generate images for the 40k test set and use 10 random splits of 30k images each. We report the mean and standard deviation across these splits. We see that the ChatPainter model, which is conditioned on additional dialogue information, gets higher inception score than the StackGAN model just conditioned on captions. Also, the recurrent version of ChatPainter gets higher inception score than the non-recurrent version. This is likely due to it learning better encoding of the dialogues as the Skip-Thought encoder isn’t trained with very long sentences, which is the case in the non-recurrent version where we collapse the dialogue in a single string.

6. Discussion and Future Work

In this paper, apart from conditioning on image captions, we additionally conditioned the ChatPainter model on publicly available dialogue data and obtained significant improvement in inception score on the MS COCO dataset. While many of the generated 256×256 images look quite realistic, the StackGAN family of models (including ChatPainter) has several limitations and also exhibits some of the issues other GANs also suffer from. The StackGAN family is able to generate photo-realistic images easily on restricted-domain datasets such as those on flowers and birds but on MS COCO, it is able to generate images that exhibit strong global consistency but does not produce recognizable objects in many cases. The current training loss formulation also makes it susceptible to mode collapse. Training the model with dialogue data is also not very stable. Recent improvements in the literature such as training with the WGAN-GP loss can help mitigate these issues to some extent. Using an auxiliary loss for the discriminator by doing object recognition or caption generation from the generated image should also lead to improvements as has been observed in prior work on other image generation tasks. The non-end-to-end training also leads to longer training time and loss of information which can be improved upon by growing the model progressively layer-by-layer as done by Karras et al. (2018).

An interesting research direction we wish to explore further is to generate an image at each turn of the conversation (or modify the previous time-step’s image) using dialogues as a feedback mechanism. The datasets we use in this paper neither have separate images for each turn of the dialogue nor is the dialogue dependent on multiple images. In the sketch-artist scenario discussed in Section 1, the sketch artist would make several changes to the image as the conversation progresses and the future conversation also would depend on the image at that point in the conversation. However, no such publicly available dataset exists yet to the best of our knowledge and we plan to collect such a dataset

soon. The recently announced dataset CoDraw (Kim et al., 2017) contains dialogues about clip art drawings, where intermediate images are updated after every dialogue turn. At the time of publication of this work, CoDraw had not yet been publicly released and if the intermediate images for this dataset are released, that would be a useful contribution for dialogue-to-image-generation research. Image generation guided by dialogue has tremendous potential in the areas of image editing, video games, digital art, accessibility, etc., and is a promising future research direction in our opinion.

Acknowledgements

We would like to acknowledge Amjad Almahairi, Kuan-Chieh Wang, and Philip Bachman for helpful discussions on GANs and Alex Marino for help with reviewing the generated images. We would also like to thank the authors of StackGAN for releasing their *PyTorch* code which we built upon. This research was enabled in part by support provided by WestGrid and Compute Canada.

References

- Arjovsky, Martin, Chintala, Soumith, and Bottou, Léon. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.
- Cai, Lei, Gao, Hongyang, and Ji, Shuiwang. Multi-stage variational auto-encoders for coarse-to-fine image generation. *CoRR*, abs/1705.07202, 2017.
- Das, Abhishek, Kottur, Satwik, Gupta, Khushi, Singh, Avi, Yadav, Deshraj, Moura, José MF, Parikh, Devi, and Batra, Dhruv. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- Dash, Ayushman, Gamboa, John Cristian Borges, Ahmed, Sheraz, Liwicki, Marcus, and Afzal, Muhammad Zeshan. TAC-GAN - text conditioned auxiliary classifier generative adversarial network. *CoRR*, abs/1703.06412, 2017.
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- Denton, Emily L., Chintala, Soumith, szlam, arthur, and Fergus, Rob. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems* 28. 2015.
- Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in Neural Information Processing Systems* 27. 2014.
- Graves, Alex and Schmidhuber, Jürgen. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6), 2005.
- Gregor, Karol, Danihelka, Ivo, Graves, Alex, Rezende, Danilo, and Wierstra, Daan. Draw: A recurrent neural network for image generation. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.
- Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, Dumoulin, Vincent, and Courville, Aaron C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems* 30. 2017.
- He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8), 1997.
- Hong, Seunghoon, Yang, Dingdong, Choi, Jongwook, and Lee, Honglak. Inferring semantic layout for hierarchical text-to-image synthesis. *CoRR*, abs/1801.05091, 2018.
- Huiskes, Mark J. and Lew, Michael S. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, 2008.
- Karras, Tero, Aila, Timo, Laine, Samuli, and Lehtinen, Jaakko. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations*, 2018.
- Kim, Jin-Hwa, Parikh, Devi, Batra, Dhruv, Zhang, Byoung-Tak, and Tian, Yuandong. Codraw: Visual dialog for collaborative drawing. *CoRR*, abs/1712.05558, 2017.
- Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.
- Kingma, Diederik P. and Welling, Max. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2014.
- Kingma, Diederik P, Salimans, Tim, Jozefowicz, Rafal, Chen, Xi, Sutskever, Ilya, and Welling, Max. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems* 29. 2016.

- Kiros, Ryan, Zhu, Yukun, Salakhutdinov, Ruslan R, Zemel, Richard, Urtasun, Raquel, Torralba, Antonio, and Fidler, Sanja. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28*. 2015.
- Krizhevsky, Alex. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.
- LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.
- Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C Lawrence. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.
- Liu, Ziwei, Luo, Ping, Wang, Xiaogang, and Tang, Xiaoou. Deep learning face attributes in the wild. In *The IEEE International Conference on Computer Vision*, 2015.
- Makhzani, Alireza, Shlens, Jonathon, Jaitly, Navdeep, and Goodfellow, Ian. Adversarial autoencoders. In *International Conference on Learning Representations*, 2016.
- Mansimov, Elman, Parisotto, Emilio, Ba, Lei Jimmy, and Salakhutdinov, Ruslan. Generating images from captions with attention. In *Proceedings of the International Conference on Learning Representations*, 2016.
- Mirza, Mehdi and Osindero, Simon. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.
- Miyato, Takeru, Kataoka, Toshiki, Koyama, Masanori, and Yoshida, Yuichi. Spectral normalization for generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2018.
- Nguyen, Anh, Clune, Jeff, Bengio, Yoshua, Dosovitskiy, Alexey, and Yosinski, Jason. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- Nilsback, M-E. and Zisserman, A. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.
- Odena, Augustus, Olah, Christopher, and Shlens, Jonathon. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.
- Paszke, Adam, Gross, Sam, Chintala, Soumith, Chanan, Gregory, Yang, Edward, DeVito, Zachary, Lin, Zeming, Desmaison, Alban, Antiga, Luca, and Lerer, Adam. Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*, 2017.
- Radford, Alec, Metz, Luke, and Chintala, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2016.
- Reed, Scott, Akata, Zeynep, Yan, Xinchen, Logeswaran, Lajanugen, Schiele, Bernt, and Lee, Honglak. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016a.
- Reed, Scott E, Akata, Zeynep, Mohan, Santosh, Tenka, Samuel, Schiele, Bernt, and Lee, Honglak. Learning what and where to draw. In *Advances in Neural Information Processing Systems 29*. 2016b.
- Roth, Kevin, Lucchi, Aurelien, Nowozin, Sebastian, and Hofmann, Thomas. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems 30*. 2017.
- Salimans, Tim, Goodfellow, Ian, Zaremba, Wojciech, Cheung, Vicki, Radford, Alec, Chen, Xi, and Chen, Xi. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29*. 2016.
- Van Den Oord, Aäron, Kalchbrenner, Nal, and Kavukcuoglu, Koray. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.
- Wah, Catherine, Branson, Steve, Welinder, Peter, Perona, Pietro, and Belongie, Serge. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- Xu, Tao, Zhang, Pengchuan, Huang, Qiuyuan, Zhang, Han, Gan, Zhe, Huang, Xiaolei, and He, Xiaodong. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485, 2017.
- Zhang, Han, Xu, Tao, Li, Hongsheng, Zhang, Shaoting, Wang, Xiaogang, Huang, Xiaolei, and Metaxas, Dimitris N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *The IEEE International Conference on Computer Vision*, 2017.