CAGAN: Text-To-Image Generation with Combined Attention Generative Adversarial Networks

Henning Schulze(✉), Dogucan Yaman, and Alexander Waibel

Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany
henning.schulze1@web.de, {dogucan.yaman,alexander.waibel}@kit.edu

Abstract. Generating images according to natural language descriptions is a challenging task. Prior research has mainly focused to enhance the quality of generation by investigating the use of spatial attention and/or textual attention thereby neglecting the relationship between channels. In this work, we propose the Combined Attention Generative Adversarial Network (CAGAN) to generate photo-realistic images according to textual descriptions. The proposed CAGAN utilises two attention models: word attention to draw different sub-regions conditioned on related words; and squeeze-and-excitation attention to capture non-linear interaction among channels. With spectral normalisation to stabilise training, our proposed CAGAN achieves state-of-the-art FID and comparative IS scores on the CUB dataset and on the more challenging COCO dataset. Furthermore, we demonstrate that judging a model by a single evaluation metric can be misleading by developing an additional model adding local self-attention which scores a higher IS than our other model, but generates unrealistic images through feature repetition.

Keywords: Text-to-image synthesis · Generative adversarial network (GAN) · Attention

1 Introduction

Generating images according to natural language descriptions spans a wide range of difficulty, from generating synthetic images to simple and highly complex real-world images. It has tremendous applications such as photo-editing, computer-aided design, and may be used to reduce the complexity of or even replace rendering engines [28]. Furthermore, good generative models involve learning new representations. These are useful for a variety of tasks, for example classification, clustering, or supporting transfer among tasks.

Although generating images highly related to the meanings embedded in a natural language description is a challenging task due to the gap between text and image modalities, there has been exciting recent progress in the field using numerous techniques and different inputs [3–5,12,18–21,29,38,39,45,46,

C. Bauckhage et al. (Eds.): DAGM GCPR 2021, LNCS 13024, pp. 392–404, 2021.
https://doi.org/10.1007/978-3-030-92659-5_25

Fig. 1. Example results of the proposed CAGAN (SE). The generated images are of 64×64, 128×128, and 256×256 resolutions respectively, bilinearly upsampled for visualization.

49] yielding impressive results on limited domains. A majority of approaches are based on Generative Adversarial Networks (GANs) [8]. Zhang et al. introduced Stacked GANs [47] which consist of two GANs generating images in a low-to-high resolution fashion. The second generator receives the image encoding of the first generator and the text embedding as input to correct defects and generate higher resolution images. Further research has mainly focused to enhance the quality of generation by investigating the use of spatial attention and/or textual attention thereby neglecting the relationship between channels.

In this work, we propose Combined Attention Generative Adversarial Network (CAGAN) that combines multiple attention models, thereby paying attention to word, channel, and spatial relationships. First, the network uses a deep bi-directional LSTM encoder [45] to obtain word and sentence features. Then, the images are generated in a coarse to fine fashion (see Fig. 1) by feeding the encoded text features into a three stage GAN. Thereby, we utilise local-self attention [27] mainly during the first stage of generation; word attention at the beginning of the second and the third generator; and squeeze-and-excitation attention [13] throughout the second and the third generator. We use the publicly available CUB [41] and COCO [22] datasets to conduct the experimental analysis. Our experiments show that our network generates images of similar quality as previous work while either advancing or competing with the state of the art on the Inception Score (IS) [35] and the Fréchet Inception Distance (FID) [11].

The main contributions of this paper are threefold:

(1) We incorporate multiple attention models, thereby reacting to subtle differences in the textual input with fine-grained word attention; modelling long-range dependencies with local self-attention; and capturing non-linear interaction among channels with squeeze-and-excitation (SE) attention. SE attention can learn to learn to use global information to selectively emphasise informative features and suppress less useful ones.

(2) We stabilise the training with spectral normalisation [24], which restricts the function space from which the discriminators are selected by bounding the Lipschitz norm and setting the spectral norm to a designated value.
(3) We demonstrate that improvements on single evaluation metrics have to be viewed carefully by showing that evaluation metrics may react oppositely.

The rest of the paper is organized as follows: In Sect. 2, we give a brief overview of the literature. In Sect. 3, we explain the presented approach in detail. In Sect. 4, we mention the employed datasets and experimental results. Then, we discuss the outcomes and we conclude the paper in Sect. 5.

2 Related Work

While there has been substantial work for years in the field of image-to-text translation, such as image caption generation [1,7,44], only recently the inverse problem came into focus: text-to-image generation. Generative image models require a deep understanding of spatial, visual, and semantic world knowledge. A majority of recent approaches are based on GANs [8].

Reed et al. [32] use a GAN with a direct text-to-image approach and have shown to generate images highly related to the text's meaning. Reed et al. [31] further developed this approach by conditioning the GAN additionally on object locations. Zhang et al. built on Reed et al.'s direct approach developing StackGAN [47] generating 256×256 photo-realistic images from detailed text descriptions. Although StackGAN yields remarkable results on specific domains, such as birds or flowers, it struggles when many objects and relationships are involved. Zhang et al. [48] improved StackGAN by arranging multiple generators and discriminators in a tree-like structure, allowing for more stable training behaviour by jointly approximating multiple distributions. Xu et al. [45] introduced a novel loss function and fine-grained word attention into the model.

Recently, a number of works built on Xu et al.'s [45] approach: Cheng et al. [5] employed spectral normalisation [24] and added global self-attention to the first generator; Qiao et al. [30] introduced a semantic text regeneration and alignment module thereby learning text-to-image generation by redescription; Li et al. [18] added channel-wise attention to Xu et al.'s spatial word attention to generate shape-invariant images when changing text descriptions; Cai et al. [3] enhanced local details and global structures by attending to related features from relevant words and different visual regions; Tan et al. [38] introduced image-level semantic consistency and utilised adaptive attention weights to differentiate keywords from unimportant words; Yin et al. [46] focused on disentangling the semantic-related concepts and introduced a contrasive loss to strengthen the image-text correlation; Zhu et al. [49] refined Xu et al.'s fine-grained word attention by dynamically selecting important words based on the content of an initial image; and Cheng et al. [4] enriched the given description with prior knowledge and then generated an image from the enriched multi-caption.

Instead of using multiple stages or multiple GANs, Li et al. [20] used one generator and three independent discriminators to generate multi-scale images

conditioned on text in an adversarial manner. Tao et al. [39] discarded the stacked architecture approach, proposing a GAN to directly synthesize images without extra networks. Johnson et al. [14] introduced a GAN that receives a scene graph consisting of objects and their relationships as input and generates complex images with many recognizable objects. However, the images are not photo-realistic. Qiao et al. [29] introduced LeicaGAN which adopts text-visual co-embeddings to convey the visual information needed for image generation.

Other approaches are based on autoencoders [6,36,42], autoregressive models [9,26,33], or other techniques such as generative image modelling using an RNN with spatial LSTM neurons [40]; multiple layers of convolution and deconvolution operators trained with Stochastic Gradient Variational Bayes [17]; a probabilistic programming language for scene understanding with fast general-purpose inference machinery [16]; and generative ConvNets [43].

We propose to expand the focus of attention to channel, word, and spatial relationships instead of a subset of these thereby enhancing the quality of generation.

3 The Framework of Combined Attention Generative Adversarial Networks

3.1 Combined Attention Generative Adversarial Networks

The proposed CAGAN utilises three attention models: word attention to draw different sub-regions conditioned on related words, local self-attention to model long-range dependencies, and squeeze-and-excitation attention to capture non-linear interaction among channels.

The attentional generative model consists of three generators, which receive image feature vectors as input and generate images of small-to-large scales. First, a deep bidirectional LSTM encoder encodes the input sentence into a global sentence vector s and a word matrix. Conditioning augmentation F^{CA} [47] converts the sentence vector into the conditioning vector. A first network receives the conditioning vector and noise, sampled from a standard normal distribution, as input and computes the first image feature vector. Each generator is a simple 3x3 convolutional layer that receives the image feature vector as input to compute an image. The remaining image feature vectors are computed by networks receiving the previous image feature vector and the result of the i^{th} attentional model F_i^{attn} (see Fig. 2), which uses the word matrix computed by the text encoder.

To compute word attention, the word vectors are converted into a common semantic space. For each subregion of the image a word-context vector is computed, dynamically representing word vectors that are relevant to the subregion of the image, i.e., indicating the weight the word attention model attends to the l^{th} word when generating a subregion. The final objective function of the attentional generative network is defined as:

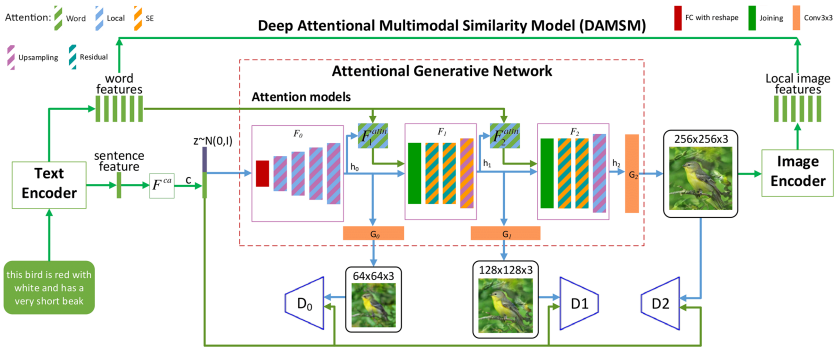

Fig. 2. The architecture of the proposed CAGAN with word, SE, and local attention. When omitting local attention, local attention is removed from the F_n^{attn} networks. In the upsampling blocks it is replaced by SE attention.

$$L = L_G + \lambda L_{\text{DAMSM}}, \quad where\ L_G = \sum_{i=0}^{m-1} L_{G_i}. \tag{1}$$

Here, λ is a hyperparameter to balance the two terms. The first term is the GAN loss that jointly approximates conditional and unconditional distributions [48]. At the i^{th} stage, the generator G_i has a corresponding discriminator D_i. The adversarial loss for G_i is defined as:

$$L_{G_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{y}_i \sim P_{G_i}}\left[log(D_i(\hat{y}_i))\right]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{y}_i \sim P_{G_i}}\left[log(D_i(\hat{y}_i, s))\right]}_{\text{conditional loss}}, \tag{2}$$

where $\hat{y}_i$ are the generated images. The unconditional loss determines whether the image is real or fake while the conditional loss determines whether the image and the sentence match or not. Alternately to the training of G_i, each discriminator D_i is trained to classify the input into the class of real or fake by minimizing the cross-entropy loss.

The second term of Eq. 1, L_{DAMSM}, is a fine-grained word-level image-text matching loss computed by the DAMSM [45]. The DAMSM learns two neural networks that map subregions of the image and words of the sentence to a common semantic space, thus measuring the image-text similarity at the word level to compute a fine-grained loss for image generation. The image encoder prior to the DAMSM is built upon a pretrained Inception-v3 model [37] with added perceptron layers to extract visual feature vectors for each subregion of the image and a global image vector.

3.2 Attention Models

Local Self-attention. Similar to a convolution, local self-attention [27] extracts a local region of pixels $ab \in \mathcal{N}_k(i,j)$ for each pixel x_{ij} and a given spatial extent k. An output pixel y_{ij} computes as follows:

$$y_{ij} = \sum_{a,b \in \mathcal{N}_k(i,j)} \text{softmax}_{ab}(q_{ij}^T k_{ab}) v_{ab}. \quad (3)$$

$q_{ij} = W_Q x_{ij}$ denotes the queries, $k_{ab} = W_K x_{ab}$ the keys, and $v_{ab} = W_V x_{ab}$ the values, each obtained via linear transformations W of the pixel ij and their neighbourhood pixels. The advantage over a simple convolution is that each pixel value is aggregated with a convex convolution of value vectors with mixing weights (softmax_{ab}) parametrised by content interactions.

Squeeze-and-Excitation (SE) Attention. Instead of focusing on the spatial component of CNNs, SE attention [13] aims to improve the channel component by explicitly modelling interdependencies among channels via channel-wise weighting. Thus, they can be interpreted as a light-weight self-attention function on channels.

First, a transformation, which is typically a convolution, outputs the feature map U. Because convolutions use local receptive fields, each entry of U is unaware of contextual information outside its region. A corresponding SE-block addresses this issue by performing a feature recalibration.

A squeeze operation aggregates the feature maps of U across the spatial dimension ($H \times W$) yielding a channel descriptor. The proposed squeeze operation is mean-pooling across the entire spatial dimension of each channel. The resulting channel descriptor serves as an embedding of the global distribution of channel-wise features.

A following excitation operation F_{ex} aims to capture channel-wise dependencies, specifically non-linear interaction among channels and non-mutually exclusive relationships. The latter allows multiple channels to be emphasized. The excitation operation is a simple self-gating operation with a sigmoid activation function:

$$F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \quad (4)$$

where δ refers to the ReLU activation function, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. To limit model complexity and increase generalisation, a bottleneck is formed around the gating mechanism: a Fully Connected (FC) layer reduces the dimensionality by a factor of r. A second FC layer restores the dimensionality after the gating operation. The authors recommend an r of 16 for a good balance between accuracy and complexity (∼10% parameter increase on ResNet-50 [10]). Ideally, r should be tuned for the intended architecture.

The excitation operation F_{ex} computes per-channel modulation weights. These are applied to the feature maps U performing an adaptive recalibration.

4 Experiments

Dataset. We employed CUB [41] and COCO [22] datasets for the experiments. The CUB dataset [41] consists of 8855 train and 2933 test images. To perform

evaluation, one image per caption in the test set is computed since each image has ten captions. The COCO dataset [22] with the 2014 split consists of 82783 train and 40504 test images. We randomly sample 30000 captions from the test set for the evaluation.

Evaluation Metric. In this work, we utilized the Inception Score (IS) [35] and The Fréchet Inception Distance (FID) [11] to evaluate the performance of proposed method. The IS [35] is a quantitative metric to evaluate generated images. It measures two properties: highly classifiable and diverse with respect to class labels. Although the IS is the most widely used metric in text-to-image generation, it has several issues [2,25,34] regarding the computation of the score itself and the usage of the score. According to the authors of [2] it: "fails to provide useful guidance when comparing models".

The FID [11] views features as a continuous multivariate Gaussian and computes a distance in the feature space between the real data and the generated data. A lower FID implies a closer distance between the generated image distribution and the real image distribution. The FID is consistent with human judgment and more consistent to noise than the IS [11] although it has a slight bias [23]. Please note that there is some inconsistency in how the FID is calculated in prior work, originating from different pre-processing techniques that significantly impact the score. We use the official implementation[1] of the FID. To ensure a consistent calculation of all of our evaluation metrics, we replace the generic Inception v3 network with the pre-trained Inception v3 network we used for computing the IS of the corresponding dataset. We re-calculate the FID scores of papers with an official model to provide a fair comparison.

Implementation Detail. We employ spectral normalisation [24], a weight normalisation technique to stabilise the training of the discriminator, during training. To compute the semantic embedding for text descriptions, we employ a pre-trained bi-direction LSTM encoder by Xu et al. [45] with a dimension of 256 for the word embedding. The sentence length was 18 for the CUB dataset and 12 for the COCO dataset.

All networks are trained using the Adam optimiser [15] with a batch size of 20, a learning rate of 0.0002, and $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We train for 600 epochs on the CUB and for 200 epochs on the COCO dataset. For the model utilising squeeze-and-excitation attention we use $r = 1$, and $\lambda = 0.1$ and $\lambda = 50.0$, respectively for the CUB and the COCO dataset. For the model utilising local self-attention as well we use $r = 4$, and $\lambda = 5.0$ and $\lambda = 50.0$.

4.1 Results

Quantitative Results. As Table 1 and Fig. 3 show, our model utilising squeeze-and-excitation attention outperforms the baseline AttnGAN [45] in both metrics

[1] https://github.com/bioinf-jku/TTUR.

Table 1. Fréchet Inception Distance (FID) and Inception Score (IS) of state-of-the-art models and our two CAGAN models on the CUB and COCO dataset with a 256×256 image resolution. The unmarked scores are those reported in the original papers, note that the reported FID scores may be inconsistent (see Sect. 4 Evaluation Metric). Scores marked with † were re-calculated by us using the pre-trained model provided by the respective authors. ↑ (↓) means the higher (lower), the better.

Model	CUB dataset		COCO dataset	
	IS↑	FID↓	IS↑	FID↓
Real data	$25.52 \pm .09$	0.00	$37.97 \pm .88$	0.00
AttnGAN [45]	$4.36 \pm .04$	$47.76^{\dagger}$	$25.89 \pm .47$	$31.05^{\dagger}$
PPAN [20]	$4.38 \pm .05$	-	-	-
HAGAN [5]	$4.43 \pm .03$	-	-	-
MirrorGAN [30]	$4.56 \pm .05$	-	$26.47 \pm .41$	-
ControlGAN [18]	$4.58 \pm .09$	$49.18^{\dagger}$	$24.06 \pm .60$	-
DualAttn-GAN [3]	$4.59 \pm .07$	-	-	-
LeicaGAN [29]	$4.62 \pm .06$	-	-	-
SEGAN [38]	$4.67 \pm .04$	-	$27.86 \pm .31$	32.28
SD-GAN [46]	$4.67 \pm .09$	-	$35.69 \pm .50$	-
DM-GAN [49]	$4.75 \pm .07$	$43.20^{\dagger}$	$30.49 \pm .57$	$22.84^{\dagger}$
DF-GAN [39]	5.10	–	–	21.42
RiFeGAN [4]	$\mathbf{5.23 \pm .09}$	–	31.70	-
Obj-GAN [19]	-	-	$30.29 \pm .33$	25.64
OP-GAN [12]	-	-	$27.88 \pm .12$	$23.29^{\dagger}$
CPGAN [21]	-	-	$\mathbf{52.73 \pm .61}$	$49.92^{\dagger}$
CAGAN_SE (ours)	$4.78 \pm .06$	**42.98**	$32.60 \pm .75$	**19.88**
CAGAN_L+SE (ours)	$4.96 \pm .05$	61.06	$33.89 \pm .69$	27.40

on both datasets. The IS is improved by $9.6\% \pm 2.4\%$ and $25.9\% \pm 5.3\%$ and the FID by 10.0% and 36.0% on the CUB and the COCO dataset, respectively. Our approach also achieves the best FID on both datasets though not all listed models could be fairly compared (see Sect. 4 Evaluation Metric).

Our second model, utilising squeeze-and-excitation attention and local self-attention, shows better IS scores than our other model. However, it generates completely unrealistic images through feature repetitions (see Fig. 4) and has a major negative impact on the FID throughout training (see Fig. 3). This behaviour is similar to [21] on the COCO dataset and demonstrates that a single score can be misleading and thus the importance of reporting both scores.

In summary, according to the experimental results, our proposed CAGAN achieved state-of-the-art results on both the CUB dataset and COCO dataset based on the FID metric and comparative results on the IS metric. All these results indicate how our CAGAN model is effective for the text-to-image generation task.

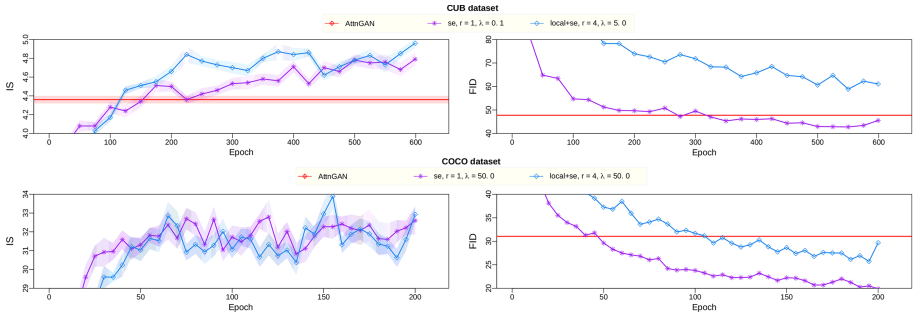

Fig. 3. IS and FID of the AttnGAN [45], our model utilising squeeze-and-excitation attention, and our model utilising squeeze-and-excitation attention and local self-attention on the CUB and the COCO dataset. The IS of the AttnGAN is the reported score and the FID was re-evaluated using the official model. The IS of the AttnGAN on the COCO dataset is with $25.89 \pm .47$ significantly lower than our models. We omitted the score to highlight the distinctions between our two models.

Qualitative Results: Figure 4 shows images generated by our models and by several other models [12,45,49] on the CUB dataset and on the more challenging COCO dataset. On the CUB dataset, our model utilising SE attention generates images of vivid details (see 1^{st}, 4^{th}, 5^{th}, and 6^{th} row), demonstrating a strong text-image correlation (see 3^{th}, 4^{th}, and 5^{th} row), avoiding feature repetitions (see double beak, DM-GAN $6th$ row), and managing the difficult scene (see 7^{th} row) best. Cut-off artefacts occur in all presented models.

Our model incorporating local self-attention fails to produce realistic looking image, despite scoring higher ISs than the AttnGAN and our model utilising SE attention. Instead, it draws repetitive features manifesting in the form of multiple birds, drawn out birds, multiple heads, or strange patterns. The drawn features mostly match the textual descriptions. This provides a possible explanation why the model has a high IS despite scoring poorly on the FID: the IS cares mainly about the images being highly classifiable and diverse. Thereby, it presumes that highly classifiable images are of high quality. Our network demonstrates that high classify-ability and diversity and therefore a high IS can be achieved through completely unrealistic, repetitive features of the correct bird class. This is further evidence that improvements solely based on the IS have to be viewed sceptically.

On the more challenging COCO dataset, our model utilising SE attention demonstrates semantic understanding by drawing features that resemble the object, for example, the brown-white pattern of a giraffe (1^{st} row), umbrellas (4^{th} row), and traffic lights (5^{th} row). Furthermore, our model draws distinct shapes for the bathroom (2^{nd} row), broccoli (3^{rd} row), and is the only one that properly approximates a tower building with a clock (7^{th} row). Generally speaking, the results on the COCO dataset are not as realistic and robust as on the CUB dataset. We attribute this to the more complex scenes coupled with more abstract descriptions that focus rather on the category of objects

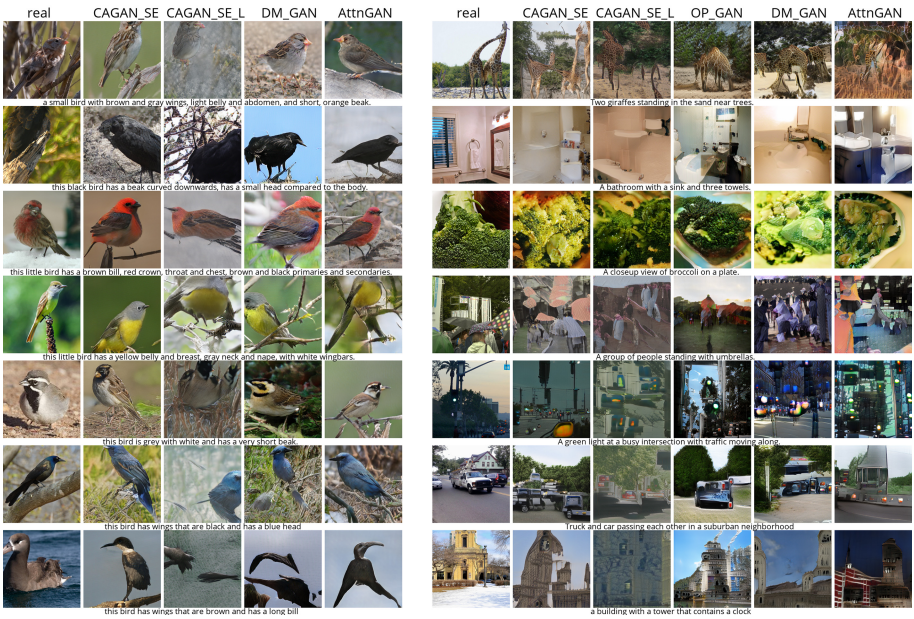

Fig. 4. Comparison of images generated by our models (CAGAN_SE and CAGAN_SE_L) with images generated by other current models [12,45,49] on the CUB dataset (left) and on the more challenging COCO dataset (right).

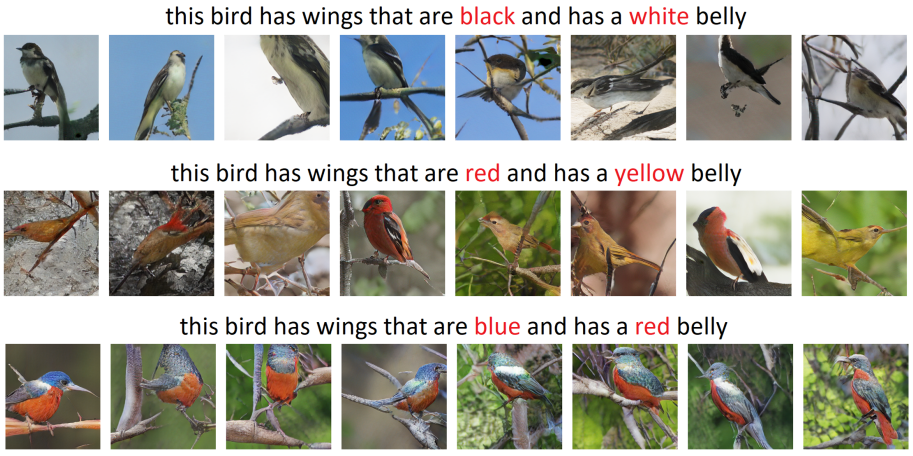

Fig. 5. Example results of our SE attention model with $r = 1, \lambda = 0.1$ trained on the CUB dataset while changing some most attended, in the sense of word attention, words in the text descriptions.

than detailed descriptions. In addition, although there are a large number of categories, each category only has comparatively few examples thereby further increasing the difficulty for text-to-image-generation.

For our SE attention model we further test its generalisation ability by testing how sensitive the outputs are to changes in the most attended, in the sense of word attention, words in the text descriptions (see Fig. 5). The test is similar to the one performed on the AttnGAN [45]. The results illustrate that adding SE attention and spectral normalisation do not harm the generalisation ability of the network: the images are altered according to the changes in the input sentences, showing that the network retains its ability to react to subtle semantic differences in the text descriptions.

5 Conclusion

In this paper, we propose the Combined Attention Generative Adversarial Network (CAGAN) to generate photo-realistic images according to textual descriptions. We utilise attention models such as, word attention to draw different sub-regions conditioned on related words; squeeze-and-excitation attention to capture non-linear interaction among channels; and local self-attention to model long-range dependencies. With spectral normalisation to stabilise training, our proposed CAGAN achieves state-of-the-art FID and comparative IS scores on the CUB dataset and on the more challenging COCO dataset. Furthermore, we demonstrate that judging a model by a single evaluation metric can be misleading by developing an additional model adding local self-attention which scores a higher IS than our other model, but generates unrealistic images through feature repetition.

References

1. Bai, S., An, S.: A survey on automatic image caption generation. Neurocomputing **311**, 291–304 (2018)
2. Barratt, S.T., Sharma, R.: A note on the inception score. CoRR abs/1801.01973 (2018)
3. Cai, Y., et al.: Dualattn-GAN: text to image synthesis with dual attentional generative adversarial network. IEEE Access **7**, 183706–183716 (2019)
4. Cheng, J., Wu, F., Tian, Y., Wang, L., Tao, D.: RiFeGAN: rich feature generation for text-to-image synthesis from prior knowledge. In: CVPR, pp. 10908–10917 (2020)
5. Cheng, Q., Gu, X.: Hybrid attention driven text-to-image synthesis via generative adversarial networks. In: Tetko, I.V., Kůrková, V., Karpov, P., Theis, F. (eds.) ICANN 2019. LNCS, vol. 11731, pp. 483–495. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30493-5_47
6. Dorta, G., Vicente, S., Agapito, L., Campbell, N.D.F., Prince, S., Simpson, I.: Laplacian pyramid of conditional variational autoencoders. In: CVMP, pp. 7:1–7:9 (2017)

7. Fang, H., et al.: From captions to visual concepts and back. In: CVPR, pp. 1473–1482 (2015)
8. Goodfellow, I.J., et al.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)
9. Gupta, T., Schwenk, D., Farhadi, A., Hoiem, D., Kembhavi, A.: Imagine this! Scripts to compositions to videos. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11212, pp. 610–626. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01237-3_37
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS, pp. 6626–6637 (2017)
12. Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text-to-image synthesis. CoRR abs/1910.13321 (2019)
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018)
14. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: CVPR, pp. 1219–1228 (2018)
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (Poster) (2015)
16. Kulkarni, T.D., Kohli, P., Tenenbaum, J.B., Mansinghka, V.K.: Picture: a probabilistic programming language for scene perception. In: CVPR, pp. 4390–4399 (2015)
17. Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.B.: Deep convolutional inverse graphics network. In: NIPS, pp. 2539–2547 (2015)
18. Li, B., Qi, X., Lukasiewicz, T., Torr, P.H.S.: Controllable text-to-image generation. In: NIPS, pp. 2063–2073 (2019)
19. Li, W., et al.: Object-driven text-to-image synthesis via adversarial training. In: CVPR, pp. 12174–12182 (2019)
20. Li, Z., Wu, M., Zheng, J., Yu, H.: Perceptual adversarial networks with a feature pyramid for image translation. IEEE CG&A **39**(4), 68–77 (2019)
21. Liang, J., Pei, W., Lu, F.: CPGAN: full-spectrum content-parsing generative adversarial networks for text-to-image synthesis. CoRR abs/1912.08562 (2019)
22. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
23. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are GANs created equal? A large-scale study. In: NIPS, pp. 698–707 (2018)
24. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR, Conference Track Proceedings. OpenReview.net (2018)
25. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: Proceedings of Machine Learning Research, ICML, vol. 70, pp. 2642–2651. PMLR (2017)
26. van den Oord, A., Kalchbrenner, N., Espeholt, L., Kavukcuoglu, K., Vinyals, O., Graves, A.: Conditional image generation with PixelCNN decoders. In: NIPS, pp. 4790–4798 (2016)
27. Parmar, N., Ramachandran, P., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. In: NIPS, pp. 68–80 (2019)
28. Pharr, M., Jakob, W., Humphreys, G.: Physically Based Rendering: From Theory to Implementation. Morgan Kaufmann, Burlington (2016)

29. Qiao, T., Zhang, J., Xu, D., Tao, D.: Learn, imagine and create: text-to-image generation from prior knowledge. In: NIPS, pp. 885–895 (2019)
30. Qiao, T., Zhang, J., Xu, D., Tao, D.: MirrorGAN: learning text-to-image generation by redescription. In: CVPR, pp. 1505–1514 (2019)
31. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: NIPS, pp. 217–225 (2016)
32. Reed, S.E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: JMLR Workshop and Conference Proceedings, ICML, vol. 48, pp. 1060–1069 (2016)
33. Reed, S.E., et al.: Parallel multiscale autoregressive density estimation. In: Proceedings of Machine Learning Research, ICML, vol. 70, pp. 2912–2921 (2017)
34. Rosca, M., Lakshminarayanan, B., Warde-Farley, D., Mohamed, S.: Variational approaches for auto-encoding generative adversarial networks. CoRR abs/1706.04987 (2017)
35. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: NIPS, pp. 2226–2234 (2016)
36. Snell, J., Ridgeway, K., Liao, R., Roads, B.D., Mozer, M.C., Zemel, R.S.: Learning to generate images with perceptual similarity metrics. In: ICIP, pp. 4277–4281 (2017)
37. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR, pp. 2818–2826 (2016)
38. Tan, H., Liu, X., Li, X., Zhang, Y., Yin, B.: Semantics-enhanced adversarial nets for text-to-image synthesis. In: ICCV, pp. 10500–10509 (2019)
39. Tao, M., Tang, H., Wu, S., Sebe, N., Wu, F., Jing, X.: DF-GAN: deep fusion generative adversarial networks for text-to-image synthesis. CoRR abs/2008.05865 (2020)
40. Theis, L., Bethge, M.: Generative image modeling using spatial LSTMs. In: NIPS, pp. 1927–1935 (2015)
41. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD birds-200-2011 dataset. Technical report CNS-TR-2011-001 (2011)
42. Wu, J., Tenenbaum, J.B., Kohli, P.: Neural scene de-rendering. In: CVPR (2017)
43. Xie, J., Lu, Y., Zhu, S., Wu, Y.N.: A theory of generative convnet. In: JMLR Workshop and Conference Proceedings, ICML, vol. 48, pp. 2635–2644 (2016)
44. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: JMLR Workshop and Conference Proceedings, ICML, vol. 37, pp. 2048–2057 (2015)
45. Xu, T., et al.: AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: CVPR, pp. 1316–1324 (2018)
46. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. In: CVPR, pp. 2327–2336 (2019)
47. Zhang, H., Xu, T., Li, H.: StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV, pp. 5908–5916 (2017)
48. Zhang, H., et al.: StackGAN++: realistic image synthesis with stacked generative adversarial networks. CoRR abs/1710.10916 (2017)
49. Zhu, M., Pan, P., Chen, W., Yang, Y.: DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In: CVPR, pp. 5802–5810 (2019)