

CAGAN: TEXT-TO-IMAGE GENERATION WITH COMBINED ATTENTION GANS

Henning Schulze Dogucan Yaman Alexander Waibel

Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany

ABSTRACT

Generating images according to natural language descriptions is a challenging task. 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 improves the state of the art on the IS and FID on the CUB dataset and the FID 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, outperforming the state of the art on the CUB dataset, but generates unrealistic images through feature repetition.

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 [1]. 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 [2, 3, 4] yielding impressive results on limited domains. A majority of approaches are based on Generative Adversarial Networks (GANs) [5]. Zhang et al. introduced Stacked GANs [6] which consist of two GANs generating images in a low-to-high resolution fashion. The second generator receives the image encoding

this bird has wings that are grey and has a yellow belly



Fig. 1. Example results of the proposed CAGAN (SE) showing the low-to-high resolution image generation. The generated images are of 64x64, 128x128, and 256x256 resolutions respectively, bilinearly upsampled for better visualization.

of the first generator and the text embedding as input to correct defects and generate higher resolution images.

Inspired by previous 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 to obtain word and sentence features. Then, the images are generated in a coarse to fine fashion (see Figure 1) by feeding the encoded text features into a three stage GAN. Thereby, we utilise local-self attention mainly during the first stage of generation; word attention at the beginning of the second and the third generator; and squeeze-and-excitation attention throughout the second and the third generator. We use the publicly available CUB and COCO 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) and the Fréchet Inception Distance (FID).

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 attention.
- (2) We stabilise the training with spectral normalisation, 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 evalua-

tion metrics have to be viewed carefully by showing that evaluation metrics may react oppositely.

2. RELATED WORK

While there has been substantial work for years in the field of image-to-text translation, such as image caption generation [7], 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 [5].

Reed et al. [4] 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. [3] 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 [6] generating 256x256 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. [8] 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. [9] introduced a novel loss function and fine-grained word attention into the model.

Recently, a number of works built on Xu et al.’s [9] approach: Qiao et al. [10] introduced a semantic text regeneration and alignment module thereby learning text-to-image generation by redescription; Li et al. [11] added channel-wise attention to Xu et al.’s spatial word attention to generate shape-invariant images when changing text descriptions; Cai et al. [12] enhanced local details and global structures by attending to related features from relevant words and different visual regions; Yin et al. [13] focused on disentangling the semantic-related concepts and introduced a contrastive loss to strengthen the image-text correlation; and Zhu et al. [14] refined Xu et al.’s fine-grained word attention by dynamically selecting important words based on the content of an initial image. Instead of using multiple stages or multiple GANs, Li et al. [2] used one generator and three independent discriminators to generate multi-scale images conditioned on text in an adversarial manner. Qiao et al. [15] introduced LeicaGAN which adopts text-visual co-embeddings to convey the visual information needed for image 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} [6] 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 Figure 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:

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

Here, λ is a hyperparameter to balance the two terms. The first term is the GAN loss that jointly approximates conditional and unconditional distributions [8]. At the i^{th} stage, the generator G_i has a corresponding discriminator D_i . The adversarial loss for G_i is defined as:

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

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 Equation 1, L_{DAMSM} , is a fine-grained word-level image-text matching loss computed by the DAMSM [9]. 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 [16] with added perceptron layers to extract visual feature vectors for each subregion of the image and a global image vector.

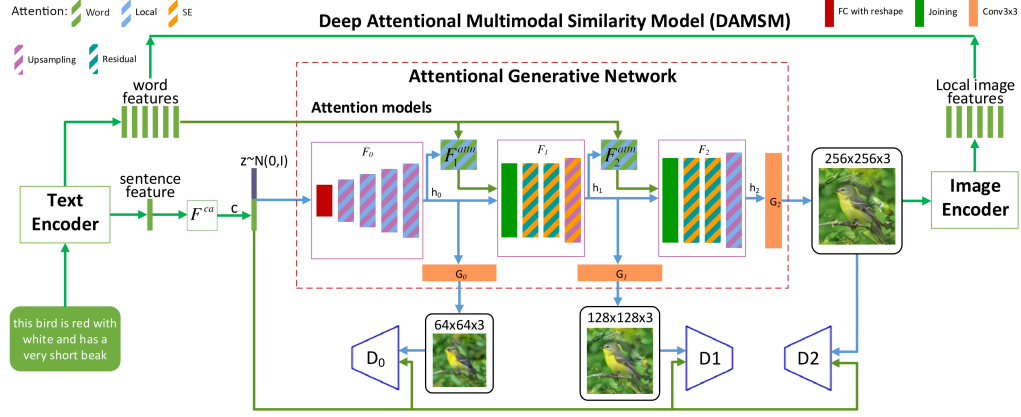


Fig. 2. The architecture of the proposed CAGAN L+SE with word, SE, and local attention. In CAGAN SE, local self-attention is removed from F_i^{attn} , downgrading F_i^{attn} to the regular word attention, and is replaced by SE attention in upsampling blocks.

3.2. Attention models

Local self-attention Similar to a convolution, local self-attention [17] 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 [18] 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 squeeze operation aggregates feature maps, typically created by a convolution, across the spatial dimension ($H \times W$) yielding a channel descriptor embedding the global distribution of channel-wise features. The proposed squeeze operation is mean-pooling across the entire spatial dimension of each channel. A following excitation operation 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. To limit model complexity and increase generalisation, a bottleneck is formed around the gating mechanism. The excitation operation computes per-channel modulation weights. These are applied to the original feature maps performing an adaptive recalibration.

4. EXPERIMENTS

Dataset The CUB dataset [19] 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 [20] 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 metrics The Inception Score (IS) [21] 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 [22] regarding the computation of the score itself and the usage of the score. The Fréchet Inception Distance (FID) [23] 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. 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 (<https://github.com/bioinf-jku/TTUR>) 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. [9] with a dimension of 256 for the word embedding. The sentence length was 18 for the CUB dataset and 12 for the COCO dataset.

Table 1. Fréchet Inception Distance (FID) and Inception Score (IS) of state-of-the-art models and our two CAGAN (SE and L+SE) models on the CUB and COCO datasets with a 256x256 image resolution. The unmarked scores are those reported in the original papers. Scores marked with † were calculated with a pre-trained model provided by the respective authors. \uparrow (\downarrow) means the higher (lower), the better.

Model	CUB dataset		COCO dataset	
	IS \uparrow	FID \downarrow	IS \uparrow	FID \downarrow
Real	25.52 \pm .09	0.00	37.97 \pm .88	0.00
[9]	4.36 \pm .04	47.76 †	25.89 \pm .47	31.05 †
[2]	4.38 \pm .05	-	-	-
[10]	4.56 \pm .05	-	26.47 \pm .41	-
[11]	4.58 \pm .09	49.18 †	24.06 \pm .60	-
[12]	4.59 \pm .07	-	-	-
[15]	4.62 \pm .06	-	-	-
[13]	4.67 \pm .09	-	35.69 \pm .50	-
[14]	4.75 \pm .07	43.20 †	30.49 \pm .57	22.84 †
[26]	-	-	27.88 \pm .12	23.29 †
[27]	-	-	52.73 \pm .61	49.92 †
SE	4.78 \pm .06	42.98	32.60 \pm .75	19.88
L+SE	4.96 \pm .05	61.06	33.89 \pm .69	27.40

All networks are trained using the Adam optimiser [25] 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 $\lambda = 50.0$.

Quantitative Results: As Table 1 shows, our model utilising squeeze-and-excitation (SE) attention outperforms the baseline AttnGAN [9] in both metrics 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 scores the highest IS and the highest FID on the CUB dataset and scores the best FID on the COCO dataset next to the third best IS. Our second model utilising SE attention and local self-attention shows better IS scores than our other model. With 4.96 ± 0.05 it outperforms all other models on the CUB dataset, improving the state of the art by $4.4\% \pm 2.6\%$. However, it generates completely unrealistic images through feature repetitions (see Figure 3) and has a major negative impact on the FID throughout training. This behaviour is similar to [27] on the COCO dataset and demonstrates that a single score can be misleading and thus the importance of reporting both scores.

Qualitative Results: Figure 3 shows images generated by our models and by several other models [26, 14, 9] on the CUB dataset. Our model utilising se attention generates

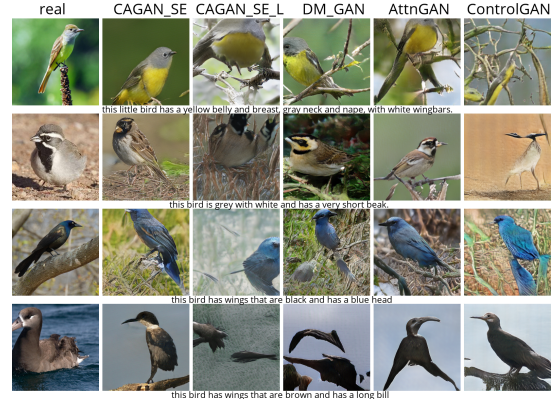


Fig. 3. Comparison of images of the CUB dataset generated by our models, state-of-the-art models and the real images.

images of vivid details (see 1st, 2nd, and 3rd row), demonstrating a strong text-image correlation (see 1st and 2nd row), avoiding feature repetitions (see double beak, DM-GAN 3rd row), and managing the difficult scene (see 4th row) well. 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, presuming this ensues high image 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.

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 improves the state of the art on the IS and FID on the CUB dataset and the FID 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 which scores a higher IS, outperforming the state of the art on the CUB dataset, but generates unrealistic images through feature repetition.

6. REFERENCES

- [1] M. Pharr, W. Jakob, and G. Humphreys, *Physically based rendering: From theory to implementation*, Morgan Kaufmann, 2016.
- [2] Z. Li, M. Wu, J. Zheng, and H. Yu, “Perceptual adversarial networks with a feature pyramid for image translation,” *IEEE CG&A*, vol. 39, no. 4, pp. 68–77, 2019.
- [3] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, “Learning what and where to draw,” in *NIPS*, 2016, pp. 217–225.
- [4] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” in *ICML*, 2016, vol. 48 of *JMLR Workshop and Conference Proceedings*, pp. 1060–1069.
- [5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative adversarial nets,” in *NIPS*, 2014, pp. 2672–2680.
- [6] H. Zhang, T. Xu, and H. Li, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” in *ICCV*, 2017, pp. 5908–5916.
- [7] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *ICML*, 2015, vol. 37 of *JMLR Workshop and Conference Proceedings*, pp. 2048–2057.
- [8] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, “Stackgan++: Realistic image synthesis with stacked generative adversarial networks,” *CoRR*, vol. abs/1710.10916, 2017.
- [9] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, “Attngan: Fine-grained text to image generation with attentional generative adversarial networks,” in *CVPR*, 2018, pp. 1316–1324.
- [10] T. Qiao, J. Zhang, D. Xu, and D. Tao, “Mirrorgan: Learning text-to-image generation by redescription,” in *CVPR*, 2019, pp. 1505–1514.
- [11] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, “Controllable text-to-image generation,” in *NIPS*, 2019, pp. 2063–2073.
- [12] Y. Cai, X. Wang, Z. Yu, F. Li, P. Xu, Y. Li, and L. Li, “Dualattn-gan: Text to image synthesis with dual attentional generative adversarial network,” *IEEE Access*, vol. 7, pp. 183706–183716, 2019.
- [13] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, “Semantics disentangling for text-to-image generation,” in *CVPR*, 2019, pp. 2327–2336.
- [14] M. Zhu, P. Pan, W. Chen, and Y. Yang, “DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis,” in *CVPR*, 2019, pp. 5802–5810.
- [15] T. Qiao, J. Zhang, D. Xu, and D. Tao, “Learn, imagine and create: Text-to-image generation from prior knowledge,” in *NIPS*, 2019, pp. 885–895.
- [16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *CVPR*, 2016, pp. 2818–2826.
- [17] N. Parmar, P. Ramachandran, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-alone self-attention in vision models,” in *NIPS*, 2019, pp. 68–80.
- [18] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *CVPR*, 2018, pp. 7132–7141.
- [19] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset. technical report cns-tr-2011-001,” 2011.
- [20] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” in *ECCV* (5), 2014, vol. 8693 of *LNCS*, pp. 740–755.
- [21] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” in *NIPS*, 2016, pp. 2226–2234.
- [22] S. T. Barratt and R. Sharma, “A note on the inception score,” *CoRR*, vol. abs/1801.01973, 2018.
- [23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *NIPS*, 2017, pp. 6626–6637.
- [24] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” in *ICLR, Conference Track Proceedings*, 2018.
- [25] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR (Poster)*, 2015.
- [26] T. Hinz, S. Heinrich, and S. Wermter, “Semantic object accuracy for generative text-to-image synthesis,” *CoRR*, vol. abs/1910.13321, 2019.
- [27] J. Liang, W. Pei, and F. Lu, “CPGAN: full-spectrum content-parsing generative adversarial networks for text-to-image synthesis,” *CoRR*, vol. abs/1912.08562, 2019.