Fusion of AraBERT and DF-GAN for Arabic Text to Image Generation

Mourad BAHANI (bahani.mourad@usmba.ac.ma)
Sidi Mohamed Ben Abdellah University National School of Applied Sciences
Aziza EL OUAAZIZI
Sidi Mohamed Ben Abdellah University National School of Applied Sciences
Khalil MAALMI
Sidi Mohamed Ben Abdellah University National School of Applied Sciences

Research Article

Keywords: Machine Learning, Deep Learning, Computer Vision, Generative Adversarial Networks, Text-to-Image Generation, Arabic Text Processing

Posted Date: May 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1533748/v2

Fusion of AraBERT and DF-GAN for Arabic Text to Image Generation

Mourad BAHANI[1*], Aziza EL OUAAZIZI[1] and Khalil MAALMI[1]

[1*]Artificial Intelligence, Data Sciences and Emerging Systems Laboratory (LIASSE), Sidi Mohamed Ben Abdellah University National School of Applied Sciences, Fez, Morocco.

*Corresponding author(s). E-mail(s): bahani.mourad@usmba.ac.ma;
Contributing authors: aziza.elouaazizi@usmba.ac.ma; k_maalmi@yahoo.com;

Abstract

Current AI systems have shown impressive results in the Automatic synthesis of high-resolution realistic images from texts descriptions. Specifically, Generative Adversarial Networks (GANs) as a powerful technology that utilizes computer vision tools to create two models, the Generator that generates realistic images and the discriminator that distinguishes whether the images synthesized are real or fake. Further, Most text-to-image generation frameworks leverage the power of GANs to generate realistic images conditioned with texts descriptions. In this paper, we fuse a sample and efficient text-to-image generation framework called DF-GAN and AraBERT architecture to generate images conditioned with Arabic texts descriptions. Firstly, we re-create new datasets matching the Arabic text-to-image generation task by applying super translation using the DeepL-Translator from English to Arabic on texts descriptions. Secondly, we leverage the power of AraBERT which is trained on billions of Arabic words to produce a strong sentence embedding, and we reduce that vector's dimension to match with DF-GAN shape. Thirdly, we inject the reduced sentence embedding into the DF-GAN framework to generate high-resolution realistic, and text-matching images conditioned with Arabic texts descriptions. Such as in previous work, we use CUB and Oxford-102 flowers as original datasets. Further, we measure our framework with FID and IS. Our framework

is the first that achieve much success in generating high-resolution realistic and text matching images conditioned with Arabic text.

Keywords: Machine Learning, Deep Learning, Computer Vision, Generative Adversarial Networks, Text-to-Image Generation, Arabic Text Processing.

1 Introduction

Generative Adversarial Networks (GANs) [6] have shown great success in the last few years in several applications, such as going from low to high resolution [25] and data augmentation [7]. Most text-to-image framework utilizes GANs to demonstrate remarkable results in generating high-resolution and realistic images conditioning with English text descriptions [11, 26–28, 31]. By stacking several pairs of generators and discriminators, stackGAN [28] utilizes two stages to generate realistic high-resolution (e.g., 256×256) images. AttnGAN [26] utilizes multiple stages to generate fine-grained details at images by paying attention to corresponding words description. MirrorGAN [15] proposes a text-to-image-to-text framework consisting of three models, the second model utilizes several pairs of generators and discriminators. Although significant progress has been made with previous frameworks, they are slow and not stable in training. Also, they need a long time with extensive computational resources to train. In contrast, DF-GANs [21] proposes a new framework consisting of one pair of generators and discriminators to generate high resolution and realistic images conditioned with an English text description. Also, it's not utilized extra-network to compute the image-text matching. However, DF-GANs does not use a big network compared with others [11, 26–28, 31], it's still slow and not stable in training. Moreover, it needs to pre-train a model to produce English sentence embedding. Furthermore, the CUB [24] and the oxford flowers [13] datasets respectively has 11788 and 8189 texts descriptions including training and test examples. Thus, it does not have enough data to train the model and produce a strong sentence embedding. On the other side, araBERT [1] follow the architecture of BERT [4], and is pre-trained on billions of Arabic words. Also, araBERT had shown great results in natural language understanding tasks. Such as Sentiment Analysis (SA) [19], Named Entity Recognition (NER) [23], and Question Answering (QA) [16]. Inspired by R. Robin [17] that use BERT to stable the training and reduce the complexity of bigGAN [2]. We fuse DF-GAN architecture with araBERT in order to synthesize realistic high-resolution and text-matching images conditioned with the Arabic text descriptions. We leverage the effectiveness of araBERT to produce Arabic sentence embedding from Arabic text descriptions without having to learn or fine-tune it. This proposition showed remarkable results in just a few epochs compared with the original framework DF-GAN [21]. To summarize, our paper contributions are as follows:

- we create new datasets matching with Arabic text-to-image generation task, by applying super translation using DeepL-Translator from English to Arabic on text descriptions of CUB dataset and Oxford-102 flower datasets.
- we leverage the effectiveness of araBERT to produce strong sentence embedding, and we reduce the dimension of the sentence vector to match with the DF-GAN input shape.
- we fuse araBERT and DF-GAN by injecting the sentence embedding produced by araBERT and reduced by training a fully connected layer to generator and discriminator.
- the experimental results on two challenging datasets prove the capability of our framework to generate high-resolution realistic and text matching images conditioned with an Arabic text description.

2 Related works

Generative Adversarial Networks (GANs) [6] Models have the purpose of Generating high-quality images conditioned with noise vectors, which are sampled from Normal or Gaussian distribution. Implemented with computer vision technology, GANs consist of a generator that generates realistic images, and a discriminator that trains to distinguish between the real and synthetic images. The generator and discriminator respectively train to minimize and maximize the distance between the real and synthetic images. However, GANs have shown a lot of success, it is still hard and not stable in training. Several works are proposed to solve the vanishing gradient and stabilize the training [9, 14] of GANs. Such as introducing Wasserstein distance [3, 29] and Gradient penalty [22]. Meanwhile, Most Text-to-Image generation frameworks leverage GANs architecture to increase performance. They have utilized multiple stages stacked with pairs of generators and discriminators. Such as stackGAN [28] that cast the problem into two stages. The first generates Low-resolution images features, taken as input noise vector and English text description through a sketch-refinement process. The second trains to generate realistic high-resolution (e.g., 256×256) images by using another pair conditioned with images resulted in stage-I and English text description. StackGAN-v2 [27] enhance the first architecture of stackGANs by stacking in each stage multiple generators and discriminators, which share their parameters in a tree-like structure to handle conditional and unconditional images distribution. This happens by introducing a new discriminator that computes conditional and unconditional loss. AttnGAN [26] proposes a new architecture that synthesizes fine-grained details in images by paying attention to corresponding words description. AttnGAN produces a low-resolution image in the first stage by using global sentence and noise vectors. Also, in the next stage, it utilizes the previously hidden image features and word features to produce new hidden image features and so on until achieving high-resolution images. MirrorGAN [15] proposes a text-to-image-to-text framework that consists of three models to reconstruct text descriptions from the generated images. Also, it proposes

the word sentence average embedding to ensure global semantic consistency between text and the generated images. DM-GAN [31] enhances initial images contents by proposing a dynamic memory module. Also, it selects the relevant text information considering the initial image content by utilizing a memory writing gate. And it uses a response gate to fuse the information read from the memories and the image features. However, the previously proposed architectures utilize a stack of generators discriminators pairs, they are not stable and need an expansive resource and a long time to train. Moreover, the images synthesized by G_n depend on G_0. If the first images are not synthesized well, they will reflect the others images. On the other hand, DF-GANs [21] comes with a new architecture that utilizes just one pair of generators and a discriminator. With the proposed architecture, DF-GAN succeed to synthesize realistic, high-resolution, and text-matching images conditioned with texts description. Also, without introducing an extra network, DF-GAN proposes a regularization in discriminator called Matching-aware zero-centered Gradient Penalty with the purpose to make the generator synthesize more realistic images matched with corresponding text. Although DF-GANs made great progress in their task and reduced cost and complexity. It uses a classic word embedding in the time of existing expert models. In other words, all these architectures use English text descriptions which are handled with several powerful technologies. In this paper, we use AraBERT [1] that follow the architecture of BERT [4] and achieve great success in natural language understanding with the Arabic language. We propose a new architecture that leverages the effectiveness of a pre-trained model called AraBERT and fused with DF-GAN, which is a sample and smart text-to-image generation architecture in order to synthesize images from Arabic text. We use Inception Score (IS) [18] and Fréchet Inception Distance (FID) [8] to measure the quality of generated images and the performance of our architecture.

3 Approach

In this section, we will explain the detail of BERT, ArabBERT, and DF-GAN.

3.1 BERT and ArabBERT

BERT [4] is a deep bidirectional representation model that trains over unlabeled text by jointly conditioning both left on right contexts in all layers. BERT is consist of two models. The first is BERTbase [4] composed of 12 layers with 768 hidden size and 12 self-attention heads with Total Parameters=340M. The second is BERTlarge [4] composed of 24 layers with 1028 hidden size and 16 self-attention with Total Parameters=110M. BERT has as input sentence or pair of sentences. Also, BERT architecture was trained by masking usually 15% of the input tokens at random, then predicting those masked tokens in the evaluation part. Further, BERT trains on English text extracted from BooksCorpus (800M words) [30] and Wikipedia (2,500M words) to produce a powerful English word and sentence Embedding. As a result,

the model demonstrates remarkable results in several natural language understanding tasks. Such as, question answering (QA) [16], sentiment analysis (SA) [19], and named entity recognition (NER) [23]. On the other side, AraBERT [1] follow the BERT architecture and handle the Arabic language, which is a morphological and rich language with relatively few resources and less discovered syntax compared with English. However, Arabic Wikipedia Dumps are small compared to the English one, araBERT train on craped Arabic news websites for articles and publicly available large Arabic corpora. Such as the 1.5 billion words Arabic Corpus [5], which is extracted from ten major news sources covering 8 countries, and the Open Source International Arabic News Corpus [32] that consists of great than three million articles (close to 1B tokens) from 31 news sources in 24 Arab countries. The final size of the pre-training dataset is 70 million sentences. Therefore, araBERT achieved a state-of-the-art result on several Arabic natural language understanding tasks, such as QA, SA, and NER. Inspired by those results, we utilize araBERT to produce a strong Arabic sentence embedding. We reduce sentence vector dimension by training a fully connected layer to both generator and discriminator to match with DF-GAN [21] input shape.

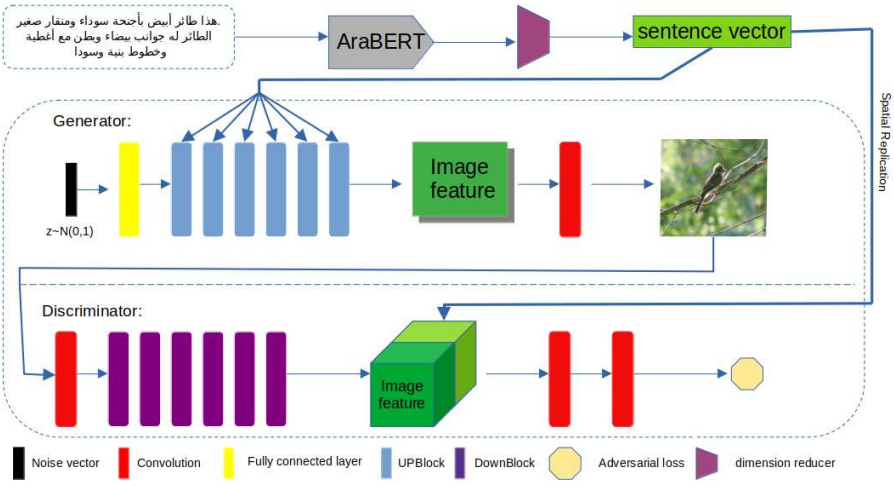

Fig. 1 The architecture of the proposed Fusion of AraBERT and DF-GAN that utilized to generate high resolution (256×256) realistic Images conditioned with the Arabic texts descriptions.

3.2 Deep Fusion GANs (DF-GAN)

The purpose of Deep Fusion DF-GAN [21] is to generate high-resolution realistic and text images matching images given a noise vector and text descriptions. This architecture consists of one pair of generators and discriminators. The

generator has as input the sentence vector that is encoded by araBERT [1], and a noise vector sampled from the Gaussian distribution. In order to achieve the generated images diversity, the noise vector is injected into a fully connected layer, the output reshaped to (-1, 4, 4), then apply a series of blocks to upsample the image features. The block is consisted of upsampling layer, residual block, and DFBlock to fuse deeply the text and image features during the image generation task. Finally, image features are converted into images by adding a convolution layer. DFBlock [21] is composed of a series of Affine Transformations, ReLU layers, and convolution layers. Affine Transformations manipulates visual feature maps conditioned on natural language descriptions by adopting two one-hidden-layer multi-layer perceptrons MLPs in order to predict the language-conditioned channel-wise shifting parameters β and scaling parameters γ from sentence vector e, respectively:

$$\gamma = MLP_1(e), \qquad \beta = MLP_2(e) \tag{1}$$

the final equation of affine transformation is as follows :

$$AFF(x_i \backslash e) = \gamma_i . x_i + \beta_i \tag{2}$$

Where function AFF is Affine transformation. x_i is the i^{th} channel of visual feature maps. e is the sentence vector. β_i and γ_i are the shifting and scaling parameter respectively for the i^{th} channel of visual feature maps. DF-GAN second part is a discriminator that is composed of several DownBlocks and convolution layers. DownBlocks consist of downsampling and residual block. Further, the images are converted into feature maps, and by a series of DownBlocks the output is downsampled and the image feature is concatenated with replicated sentence vector. The discriminator promotes the generator to synthesize higher quality and text-image semantic consistency images by distinguishing real samples from generated samples. Also, an adversarial loss will be predicted by the discriminator in order to evaluate the semantic consistency and visual realism of inputs. FD-GANs applies the hinge loss [12] to stabilize the training process. The discriminator computes the loss of synthetic images with matching text, synthetic images with mismatched text, real images with mismatched text, and real images with matching text. The generator loss is the matching of generated data and matching text. Furthermore, Matching-Aware Gradient Penalty (MA-GP) proposed by [21] to apply the gradient penalty on the target data point, which is real images with the matching sentences. The whole formulation of models losses with MA-GP is as follows:

$$\begin{aligned} L_D \; &= \mathbb{E}_{x \sim \mathbb{P}_r}[min(0, -1 + D(x, e)] \\ &\quad -(1/2)\mathbb{E}_{G(z) \sim \mathbb{P}_g}[min(0, -1 - D(G(z), e)] \\ &\quad -(1/2)\mathbb{E}_{x \sim \mathbb{P}_{mis}}[min(0, -1 - D(x, e))] \\ &\quad +k\mathbb{E}_{x \sim \mathbb{P}_r}[\|\nabla_x D(x, e)\|^p + \|\nabla_e D(x, e)\|^p] \\ L_G \; &= -\mathbb{E}_{G(z) \sim \mathbb{P}_g} D(G(z), e) \end{aligned} \tag{3}$$

Where e is the sentence vector. z is the noise vector sampled from the Gaussian distribution. p and k are two hyper-parameters to balance the effectiveness of the gradient penalty. $\mathbb{P}_g$, $\mathbb{P}_r$, $\mathbb{P}_{mis}$ denote, respectively, the generated, real, and mismatching data distribution. MA-GP helps the generator to generate text-matching and realistic images from given text descriptions without employing extra networks to compute the text-image semantic similarity.

3.3 Fusion of AraBERT in DF-GAN

AraBERT [1] has applied and shown a potential result in eleven natural language understanding. Such as QA, NER, and SA. As shown in Fig. 1, we utilize AraBERT as a powerful architecture pre-trained on billion of Arabic words to produce sentence embedding of each text description without fine-tuning. We reduce the dimension of the sentence vector to achieve consistency with its input shape, by training a fully connected layer to both generator and discriminator on DF-GAN. On the other hand, compared with previous architectures [15, 27, 31], DF-GAN successes to synthesize high-resolution text-matching and realistic images given text descriptions with just one pair of a discriminator and generator. As we mentioned previously, the generator has two inputs, the noise vector, and the sentence vector. As shown in Fig. 1, we inject the reduced sentence vector on overall UPBlocks. Specifically, in DFBlocks to ensure the text-image consistency in generated images. Furthermore, in order to compute adversarial loss and evaluate the visual and semantic consistency of inputs, we concatenate replicated sentences with produced images features by DownBlock. Our architecture demonstrates its ability to generate high-resolution, text-matching, and realistic images given Arabic text descriptions.

4 Experiments

In this section, we introduce the datasets, training details, and evaluation metrics, then our Method variants quantitatively and qualitatively.

Table 1 CUB and Oxford-102 Statistics

Datasets	Train	Test
CUB	9414	2374
Oxford-102	7034	1155

4.1 Datasets

As in previous works [26, 27, 31], we evaluate the proposed model on two challenging datasets. The CUB [24] and the Oxford-102 flowers [13] datasets. We present in Table. 1 the number of examples in train and test on both datasets. The CUB dataset contains 11788 images belonging to 200 classes. the Oxford-102 dataset contains 8,189 images of flowers from 102 different categories. CUB and Oxford-102 images have 10 texts descriptions, we translate those

Fig. 2 The generated images using our framework on the CUB test set conditioned with the Arabic texts descriptions.

Fig. 3 The generated images using our framework on the oxford-102 test set conditioned with the Arabic texts description.

texts from English to Arabic language using the DeepL Translator model and reconstruct new Arabic text descriptions.

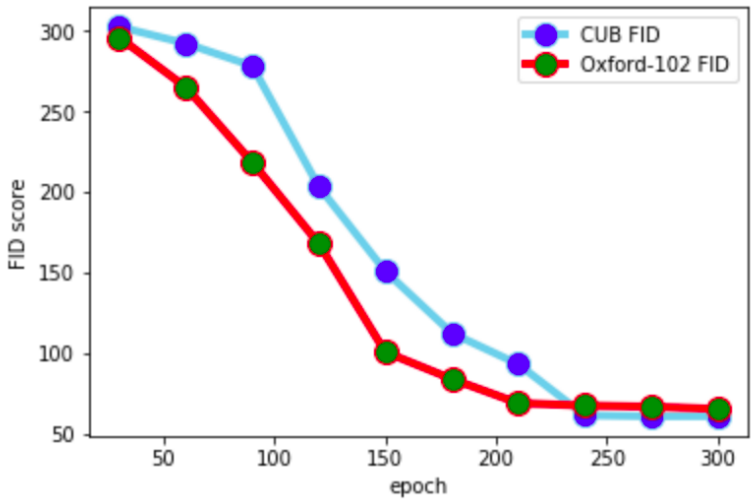

Fig. 4 Decreasing of FID During the training of our Framework on CUB and Oxford-102 datasets.

4.2 Training Details

We utilize Adam[10] as optimizer with $\beta_1 = 0.01$ and $\beta_2 = 0.9$. The learning rate is set to 0.0005 for the discriminator and 0.0002 for the generator. We train our modules on both CUB and Oxford-102 flowers with 300 epochs to prevent the over-fitting.

4.3 Evaluation Details

Following DF-GAN and previous works [26–28], we utilize the Inception Score (IS) [18] and Frechet Inception Distance (FID) [8] to evaluate the performance of our framework. IS computes the Kullback-Leibler (KL) divergence between conditional distribution $p(y \mid x)$ and marginal distribution $p(y)$, the Inception Score is formulated as:

$$I = exp(\mathbb{E}_x D_{KL}(p(y \mid x) \| p(y))), \tag{4}$$

Where x is a generated image and y is the image label predicted by a pre-trained Inception v3 network [20]. Higher IS means a higher quality of the generated images and each image clearly belongs to a specific class.
FID is another metric that computes the Frechet distance between the real images and the synthetic images distributions. Also, FID uses the pre-trained Inception v3 network in order to predict the feature space images. The FID equation is as follows:

$$\begin{aligned} F(r, g) &= \|\mu_r - \mu_g\|^2 + \\ &\quad trace(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})), \end{aligned} \tag{5}$$

Where r and g are real and generated data, μ_r, μ_g and Σ_r, Σ_g are the means and covariance of real and generated data distribution, respectively. Contrary to IS, lower FID means more realistic images, and the distributions look similar.

Table 2 Inception Score and FID in CUB test set

	IS score	FID score
AttnGAN	4.36	23.98
Stack++	3.70	51.89
Our architecture	3.51	55.96

Table 3 Inception Score and FID in Oxford-102 test set

	IS	FID score
Stack++	3.20	55.28
AttnGAN	-	-
Our architecture	3.06	59.45

4.4 Quantitative Evaluation

In this section, we compare our architecture achievement using FID and SI with previous work [26, 27] achievement. However, the previous works generate images conditioned with English text description, which is handled with a lot of technology. Meanwhile, our framework utilizes Arabic text description, which is a challenging and meaningful language. As shown in Table. 2 and Table. 3 we set the IS and FID achieved on both test sets using the previous and our framework. Our achievements are 60.96 on FID and 3.21 on SI, while StackGAN++ [27] achieve 51.89 in FID and 3.70 in SI on the CUB test set. Furthermore, Our achievements are 65.45 on the FID score and 3.01 on SI, while StackGAN++ achieves 55.28 in FID and 3.20 in SI on the Oxford-102 test set. Our result is close to the StackGAN++ result despite the difference between language and the difficulty of Arabic text processing. Therefore, the quantitative comparisons show remarkable success and the ability to synthesize realistic and high-resolution images conditioned with an Arabic text description.

4.5 Qualitative Evaluation

In this section, we will analyze the visual result shown in Fig. 2 and Fig. 3. As can be seen in Fig. 2, Our architecture is able to achieve our goals on the CUB test set. As shown in overall columns except for the last image, our framework is successful in ensuring the diversity and fidelity, and realism of generated images. However, we utilized a sentence vector, and as shown in two images in Fig. 3 our proposed architecture is able to catch meaningful words

and synthesize realistic images with multiple objects related to that specific word. Such as "flowers" and "flower" in Arabic. That means our framework is able to distinguish between singular and plural words.

4.6 Uncompleted Generated Images

In Fig. 4 the FID score between 250 and 300 epochs seems constant. That is means the modules achieve stable training. Despite the FID score still being higher, it is close to some previous works. On the other hand, as shown in Fig. 2 and Fig. 3, we can see some incomplete images generated on CUB and Oxford test sets. Specifically, on the last images in both figures. we suppose that's due to early stop learning and some contradictory text description. Also, the underestimate and exaggeration of colors descriptions. Therefore, this hard task needs consistent texts descriptions from professional sources. Eventually, with those challenging datasets and this shortcoming, our framework is the first one that showed a lot of success in Arabic text to images generation.

5 Conclusion

In this paper, we propose a powerful architecture with the purpose to generate high-resolution realistic, and text-image matching images conditioned with an Arabic text description. Also, In order to achieve this goal, we apply the DeepL Translator to translate the text description from English to Arabic on CUB and Oxford-102 flowers datasets. Further, we reduce the sentences vector dimension produced by AraBERT to match with FD-GANs input shape. Furthermore, we fuse AraBERT and DF-GAN by injecting the sentence embedding vector into the DF-GAN generator and discriminator. Our approach demonstrates remarkable result compared with stackGAN++ which use English text description. The experience showed that our approach achieved 60.96 and 65.45 on the FID score, and 3.21 and 3.01 on SI in CUB and Oxford-102 datasets respectively. However, Arabic text is complicated compared with English text. Our framework is the first one that showed a lot of success in Arabic text to images generation.

6 Declarations section

6.1 Ethical Approval and Consent to participate

Not Applicable

6.2 Consent for publication

Not Applicable

6.3 Availability of supporting data

Not Applicable

6.4 Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

6.5 Funding

No funding was received to assist with the preparation of this manuscript.

6.6 Authors' contributions

All authors contributed to the study conception, building the architecture, and data collection. the manuscript was Written by [BAHANI Mourad], and all authors read, reviewed, and approved the final manuscript.

6.7 Acknowledgments

Not Applicable

6.8 Authors' information

Corresponding author
Name : BAHANI Mourad
Affiliation : National School of Applied Sciences of Fez
Address : MOROCCO
Email : bahani.mourad@usmba.ac.ma
Phone : +212697335540

Back up contact
Name : Aziza EL OUAAZIZI
Affiliation : National School of Applied Sciences of Fez
Address : MOROCCO
Email : aziza.elouaazizi@usmba.ac.ma
Phone : +212661537529

Name : Khalil MAALMI
Affiliation : National School of Applied Sciences of Fez
Address : MOROCCO
Email : k maalmi@yahoo.com
Phone : +212661257242

References

[1] Antoun W, Baly F, Hajj H, Antoun W, Baly F, Hajj H, 2021. Arabic BERT https://arxiv.org/abs/2003.00104v4.

[2] Brock A, Donahue J, and Simonyan K, 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In 7th International Conference on Learning Representations, ICLR.

[3] Brock A, Donahue J, and Simonyan K, 2019. Large scale gan training for high fidelity natural image synthesis, in International Conference on Learning Representations.

[4] Devlin J, Chang M.W, Lee K, and Toutanova K, 2018. Bert: Pre-training of deep bidirectional trans- formers for language understanding. arXiv preprint arXiv:1810.04805.

[5] El-Khair, I. A. 2016. 1.5 billion words arabic corpus. arXiv preprint, arXiv:1611.04033.

[6] Goodfellow I, Abadie P. J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y, Generative adversarial nets. In Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[7] Henrique Kiyoiti dos Santos Tanaka F, Aranha C, 2019. Data augmentation using gans. Proceedings of Machine Learning Research XXX, 1:16.

[8] Heusel M, Ramsauer H, Unterthiner T, Nessler B, and Hochreiter S, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in Advances in neural information processing systems, 2017, pp. 6626–6637.

[9] Hochreiter S, Bengio Y, Frasconi P, Schmidhuber J, 2001. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-term Dependencies, A Field Guide to Dynamical Recurrent Neural Networks. IEEE Press, New Jersey.

[10] Kingma P, Ba D.J, 2015. Adam: A method for stochastic optimization, in International Conference on Learning Representations.

[11] Li W, Zhang P, Zhang L, Huang Q, He X, Lyu S, Gao J, Object-driven text-to-image synthesis via adversarial training. In Proceed- ings of the IEEE Conference on Computer Vision and Pattern Recognition,2019, pp. 12174–12182.

[12] Lim J, Ye H.J.C, 2017. Geometric gan, arXiv preprint, arXiv:1705.02894.

[13] Nilsback M, Zisserman A, Automated Flower Classification over a Large Number of Classes. ICVGlP 2008: 722729,2008.

[14] Pascanu R, Mikolov T, Bengio Y, 2013. On the difficulty of training recurrent neural networks. JMLR: W and CP, In: Proceedings of the 30th International Conference on Machine Learning, 28, Atlanta, Georgia, USA.

[15] Qiao T, Zhang J, Xu D, and Tao D, Mirrorgan: Learning text- to-image generation by redescription, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1505–1514.

[16] Rajpurkar P, Zhang J, Lopyrev K, and Liang P, Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2383–2392.

[17] Robin R, Patrick E, Bjorn O, 2020. Network-to-Network Translation with Conditional Invertible Neural Networks arXiv:2005.13580v2.

[18] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, and Chen X, Improved techniques for training gans, in Advances in neural information processing systems, 2016, pp. 2234–2242.

[19] Socher R, Perelygin A, Wu J, Chuang J, Manning D.C, Ng A, and Potts C, Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1631–1642.

[20] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z, Rethinking the inception architecture for computer vision, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[21] Tao M, Tang H, Wu S, Sebe N, Jing X, Wu F, Bao B, 2020. Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis, arXiv preprint, arXiv:2008.05865v2.

[22] Thanh-Tung H, Tran T, Venkatesh S, 2019. Improving generalization and stability of generative adversarial networks, in International Conference on Learning Representations.

[23] Tjong Kim Sang E.F, and Meulder F.D, 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In CoNLL.

[24] Wah C, Branson S, Welinder P, Perona P, and Belongie S, 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR2011-001, California Institute of Technology.

[25] Wu H, Zheng S, Zhang J, Huang K, GP-GAN: towards Realistic High-Resolution Image Blending. In: Proceedings of the 27th ACM International Conference on Multimedia, ser. MM, 2019, pp. 2487–2495. Association for Computing Machinery, New York.

[26] Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X, Attngan: Fine- grained text to image generation with attentional generative adversarial networks, In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1316–1324.

[27] Zhang H, Xu T, Li H, Zhang S, gang Wang X, Huang X, and Metaxas D.N, 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE TPAMI, 41(8):1947–1962.

[28] Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, and Metaxas D.N, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision, 2017, pp. 5907– 5915.

[29] Zhang H, Goodfellow I, Metaxas D, and Odena A, Self-attention generative adversarial networks, in International Conference on Machine Learning, 2019, pp. 7354–7363.

[30] Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, 2015, pp. 19–27.

[31] Zhu M. P, Chen W, Yang Y, Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5802–5810, https://arxiv.org/abs/1904.01310

[32] Zeroual I, Goldhahn D, Eckart T, Lakhouaja A, 2019. OSIAN: Open source international Arabic news corpus - preparation and integration into the CLARINinfrastructure. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 175–182, Florence, Italy, August, Association for Computational Linguistics.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- snjnl.cls
- pdftex.sty
- snarticle.tex
- snvancouver.bst
- pdftex.sty
- snarticle.tex
- snjnl.cls
- snvancouver.bst