CP-GAN: Caption to Image using Generative Adversarial Network

Saurabh Shinde (BE17F06F053) , Adil Ansari (BE17F06F055), Gaurav Ramole (BE17F06F046), Abhishek Reveskar (BE17F06F048)

Guided by: Mrs. K.R.Sarode

Information Technology , Government college of Engineering Aurangabad

Introduction

- **Background**
- Generative Adversarial Networks (GANs) are powerful generative models introduced by (Goodfellow et al.) and can be trained on as little data as a single image
- In this project, a caption-to-image generation framework is proposed which translates text descriptions to photo-realistic images , thus allowing written languages to potentially benefit from this technology..
- Main aim is to generate realistic and caption-consistent images from given natural language descriptions.
- Due to its practical applications, caption-to-image synthesis is active research topic .

Objectives

- To increase the performance of the model and quality of images generation and to reduce modality gap between caption and image which is not achieved in previous frameworks.
- Conditioned on the caption, the proposed CPGAN model synthesizes images that are semantically consistent with the corresponding captions.
- To reduce computational cost

Methodology & Study Area

- Given an image-caption pair, we need to find a common space for both modalities, so that we can minimize the modality gap and obtain visually-grounded embedding.
- We developed target aware discriminator that significantly improves the image quality and caption-image semantic consistency, and helps to convergence of generator without adding more networks.
- Developed CP-GAN which generated synthesizes images and help to find caption and image features more effectively.
- To reduce noise in the caption which affect the performace we add extra layer which reduce noise and provide generator with less noise data which help to get quality image of 256 *256 resolution.
- Data augmentation helps to stablize the model

Model Architectures

Image Encoder

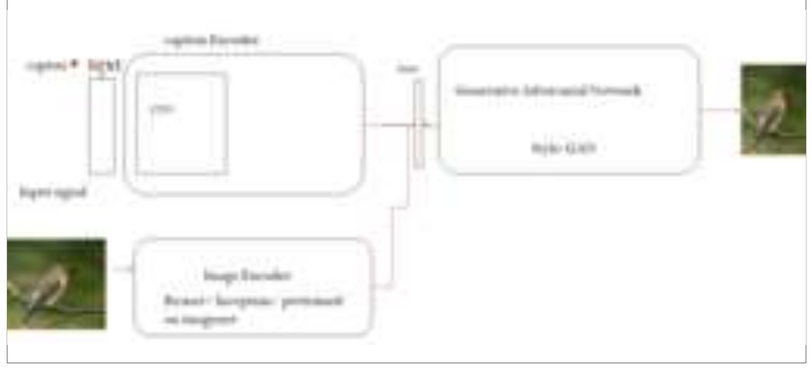
- We have used Inception-v3 architecture model pretrained on image-net to extract visual features.
- to extract visual features a layer will be employed to convert the visual feature to a common space of visual and caption embedding.

Caption Encoder

- Inspired by the character-based text embedding architecture and caption-visual cross-modal embedding learning , a multi-layer CNN will be used.
- First Sentence features are transformed into word features
- . Conventional generator predicts an initial image with a rough shape and few details from sentence feature
- This stage can be repeated multiple times to retrieve more precise information and generate a high-resolution image with more fine-grained details.

GAN Model

- Our method considers the interaction between each word and the whole generated image
- StackGAN generate photo-realistic high-resolution images with two stages.
- Stage-I GAN sketches the primitive shape and basic colors of the object conditioned on the given text description, and draws the background layout from a random noise vector, yielding. a low-resolution image. Stage-II GAN corrects defects in the low-resolution image from Stage-I and completes details of the object by reading the text description again, producing a high-resolution photo-realistic image.



Result

Models	CUB-IS ↑	CUB-FID ↓
AttnGAN	4.36	23.98
DM-GAN	4.75	16.09
CP-GAN (Proposed)	7.39	10.82

, .Dataset and Evaluation metrics

- Datasets CUB commonly-used dataset in the field of Text to Image.
- CUB is a fine-grained bird dataset that contains 11,788 bird images belonging to 200 categories.
- Each image in both datasets has 10 text descriptions collected .
- Evaluation metrics will be used are inception score and Frechet inception distance

Conclusion

we proposed a novel caption to image Generative Adversarial Networks (CP-GAN) for caption-to-image generation tasks which is able to directly synthesize more realistic and text-image semantic consistent images without stacking architecture and extra networks. Moreover, our target-aware discriminator composed significantly improves the image quality and caption mage semantic consistency, and accelerate the convergence of generator. Result show that our proposed CP-GAN significantly outperforms other model on the CUB dataset

References

- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014
- Wang, Xinsheng and Qiao, Tingting and Zhu, Jihua and Hanjalic, Alan and Scharenborg, Odette, “S2IGAN: Speech-to-Image Generation via Adversarial Learning”
- Yuchuan Gou, Qiancheng Wu, Minghao Li, Bo Gong, and Mei Han. Segattngan: Text to image generation with segmentation attention. arXiv preprint arXiv:2005.12444, 2020
- Z. {He} and W. {Zuo} and M. {Kan} and S. {Shan} and X. {Chen}, AttGAN: Facial Attribute Editing by Only Changing What You Want, 2019
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Advances in neural information processing systems, pages 2234–2242, 2016
- Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 5907– 5915, 2017
- Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE TPAMI, 41(8):1947–1962, 2018
- Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for textto-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5802– 5810, 2019.
- Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In CVPR, pages 5657– 5666, 2018.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In CVPR, pages 2818–2826, 2016.