

Fine-Grained Image Generation from Bangla Text Description using Attentional Generative Adversarial Network

Md Aminul Haque Palash | Md Abdullah Al Nasim | Aditi Dhali |
Faria Afrin



Bangla Language to Image Generation

- Text-to-image generation refers to generating a visually realistic image that matches a given text description.
- We proposed a generative networks with stacked attention to generate images from low to high resolutions at multiple stages for Bangla text description.

এটি একটি বাদামী ডানা এবং একটি সাদা চঞ্চুযুক্ত একটি নীল পাখি

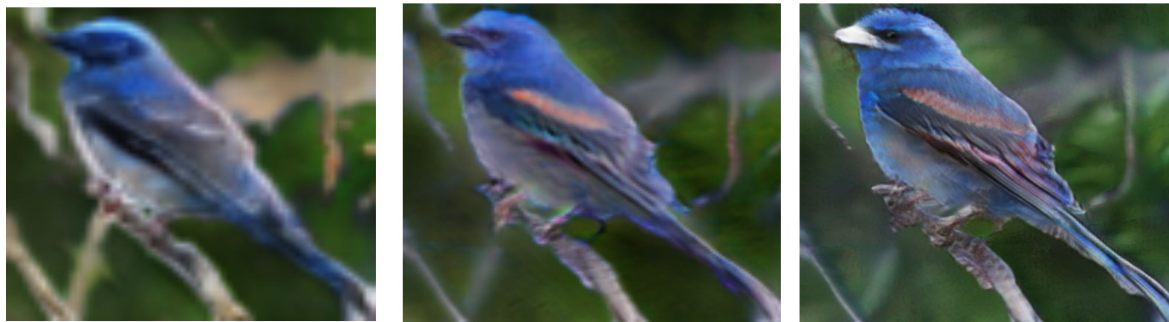
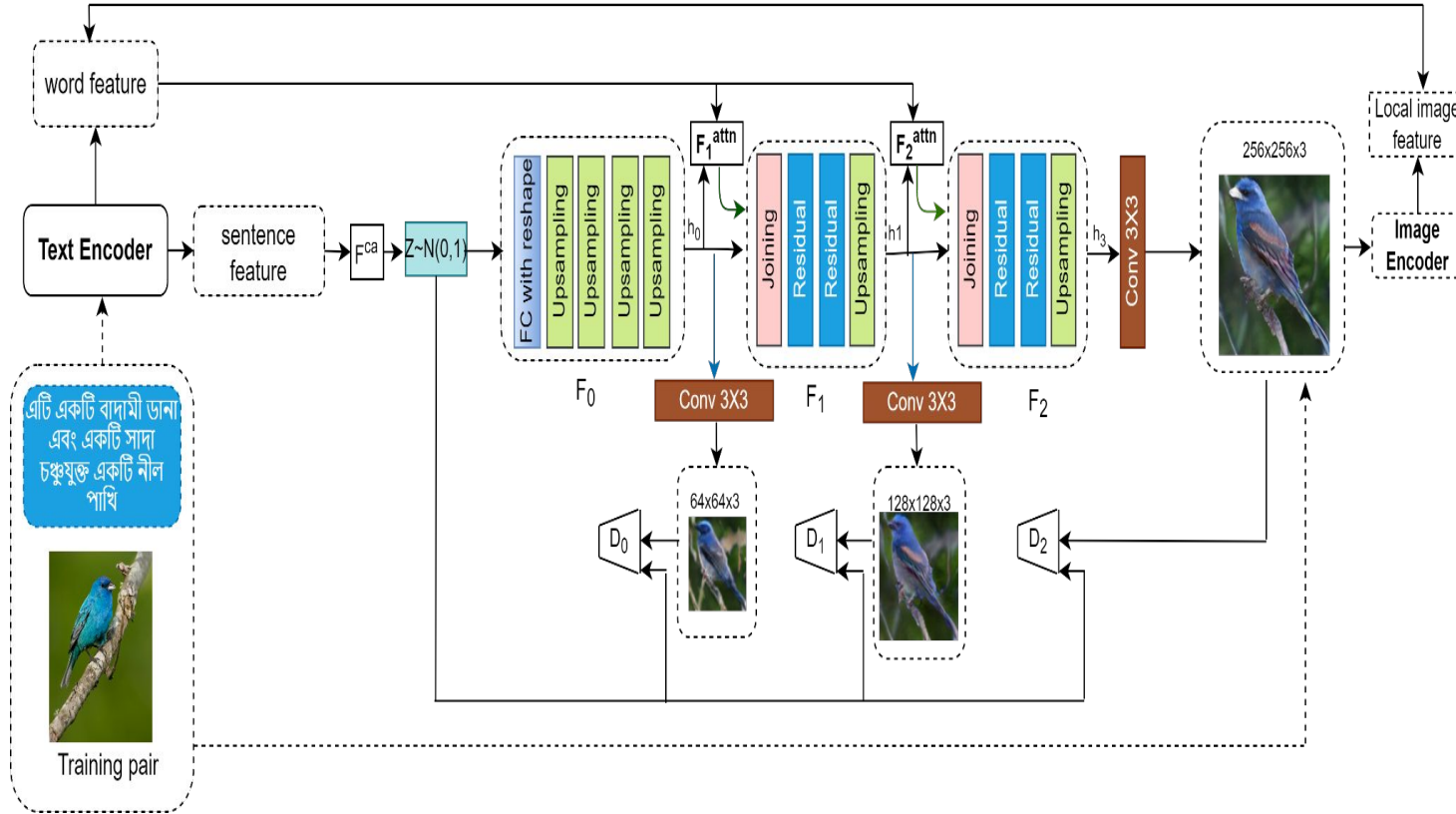


Fig-1: Experiment with our proposed model outcomes. Result displays from low-to high-resolution images

Bangla text to image generation in Recent Literature

- Some research works have been done on text to image generation.
 - AttnGAN[1]
 - MirrorGAN[2]
 - StackGAN[3]
 - CycleGAN[4] and
 - DualGAN[5]
- All of them used generation adversarial machism in their core architecture and used english language
- There is no work for bangla language to generate images from visual text description.

Deep Attentional Multimodal Similarity Model (DAMSM)



Proposed Method: Workflow

fig: Followed workflow for our proposed model

Bangla Attentional Generative Network

- Takes multi-level conditions (global-level sentence feature and fine-grained word features) as input.
- Generated images from low-to-high resolutions at multiple stages.
- In the first stage
 - based on the sentence feature, the image with basic color and shape is generated by F0
 - hidden features h_0 are decoded from the sentence feature.
- The following stages, attention models are built
 - for each region feature of previous generated image, compute its word-context vector.
 - concatenate previous image region features (e.g h_0) and word-context vectors to generate higher resolution.

Experiments

- Datasets

Dataset	CUB-2011	
	train	test
# samples	8,500	2,933
captions/image	10	10



- Images of 200 species of birds along with 10 captions for each image.











Experiments (cont.)

- Evaluation metrics
 - Inception Score reflects the quality and diversity of the generated images.
 - Frechet Inception Distance (FID) used to ensure that the images created are of high quality and uniformity

Dataset	Inception ↑	FID ↓
CUB	$3.58 \pm .06$	41.08

CUB attention maps

এই ছোট পাখির একটি সাদা পেট এবং একটি হলুদ পাখা সাথে একটি কালো মাথা রয়েছে

৪:সাদা	১১:কালো	৮:হলুদ	০:এই	১০:একটি
				
০:এই	২:পাখি	৫:পেট	১২:মাথা	১১:কালো
				

Examples - generated images by our model

এটি একটি বাদামী ডানা এবং একটি সাদা চঞ্চুযুক্ত
একটি নীল পাখি



এই ছোট পাখিটি হলুদ বুক সাদা পিঠের সাথে
বাদামী মিশ্রণ এবং সাদা মাথা এবং কালো চঞ্চু



ছোট থেকে মাঝারি কালো এবং হলুদ পাখি লম্বা
কালো টারসাস এবং মাঝারি কালো চঞ্চু



এই ছোট পাখির একটি লম্বা পাতলা বিল এবং হলুদ
গলা স্তন এবং পেট বাদামী এবং সাদা ডানায়ুক্ত



Generated images

Real images



Qualitative evaluation - generalization Ability

- Change some most attended words in the text description

এই ছোট পাখির একটি **সাদা** পেট এবং একটি **হলুদ** পাখা সাথে একটি **কালো** মাথা রয়েছে



এই ছোট পাখির একটি **কালো** পেট এবং একটি **কালো** পাখা সাথে একটি **কালো** মাথা রয়েছে



এই ছোট পাখির একটি **হলুদ** পেট এবং একটি **নীল** পাখা সাথে একটি **লাল** মাথা রয়েছে



Qualitative analysis - generalization Ability

- Novel images (failure cases) generated by our attentional generative model for bangla text description.



Concluding Remarks

- Our proposed model can successfully generate images from bangla text description in most of the cases.
- Although there is some limitation and drawbacks on the performance of our model than other model with english dataset.
- We're planning to improve our model performance by taking various initiatives
 - Improving the quality of bangla text description
 - integrating the scale-specific control from StyleGAN
 - Using transformer based language model as text encoder

References

- [1] S. Naveen, M. R. Kiran, M. Indupriya, T. Manikanta, and P. Sudeep, “Transformer models for enhancing atnngan based text to image generation,” Image and Vision Computing, p. 104284, 2021
- [2] T. Qiao, J. Zhang, D. Xu, and D. Tao, “Mirrorgan: Learning text-toimage generation by redescription,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1505–1514.
- [3] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, “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.
- [4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.
- [5] Z. Yi, H. Zhang, P. Tan, and M. Gong, “Dualgan: Unsupervised dual learning for image-to-image translation,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2849–2857.

Thank You!

We are open to all relevant queries.
