Zero-Shot Image Generation by Distilling Concepts from Multiple Captions

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

K J Joseph · Arghya Pal · Vineeth N Balasubramanian
Department of Computer Science, Indian Institute of Technology, Hyderabad, India
cs17m18p100001@iith.ac.in

GitHub link

1. Overview

- We propose a generative model that will iteratively improve the quality of the generated image by making use of multiple captions about a single image.
- This is achieved by ensuring Cross-Caption Cycle Consistency between the captions and the intermediate image representations.
- Our experiments on Caltech-UCSD Birds dataset (CUB) and Oxford-102 Flowers dataset reveal that the proposed approach is able to generate plausible images even for classes with no training example.

2. Cross-Caption Cycle Consistency

- Cross-Caption Cycle Consistency ensures that the captions consumed at each time-step and the generated image features holds a cycle consistency.
- Cycle consistency across four captions $(\mathbf{t_1}, \cdots, \mathbf{t_4})$ is shown below.

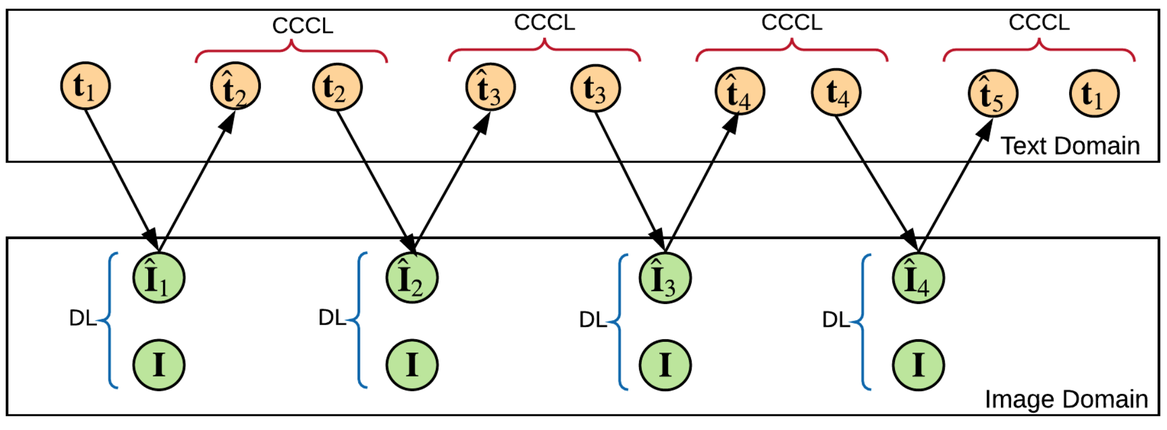


Figure 1. Illustration of Cross-Caption Cycle Consistency. A generator G converts $\mathbf{t}_i$ to an image $\hat{\mathbf{I}}_i$. Discriminator at each step forces $\hat{\mathbf{I}}_i$ to be realistic. A cross-caption consistency network converts $\hat{\mathbf{I}}_i$ back to a caption $(\hat{\mathbf{t}}_{i+1})$ and forces it to be close to $\mathbf{t}_{i+1}$. In the last step, $\hat{\mathbf{t}}_5$ is ensured to be consistent with the initial caption $\mathbf{t}_1$, hence completing a cycle. Meanwhile, the concepts in $\hat{\mathbf{I}}_i$ is incrementally improved.

3. Architecture

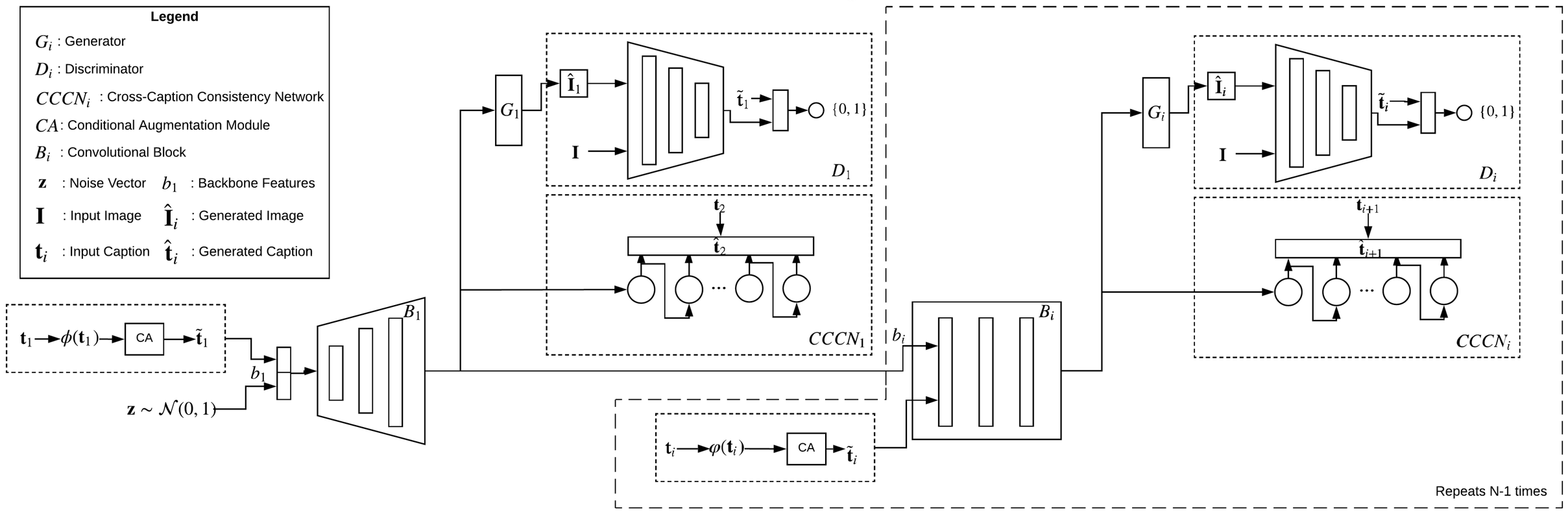


Figure 2. The proposed architecture. A vector representation of the first caption and a noise vector is fused together and passed through a set of up-sampling layers in the initial convolution block $\mathbf{B_1}$. The first Generator ($\mathbf{G_1}$) branches off from here, while the activations from $\mathbf{B_1}$ are fused with the next text embedding, followed by a set of residual convolutional layers to grow the backbone network. Those features that are passed to the generator is also passed to the CCCN, where it is converted to a caption and its consistency with the next caption is ensured. Images generated by each $\mathbf{G}_i$ is passed on to the corresponding $\mathbf{D}_i$.

4.1 Qualitative Results

- This bird is blue with black tail feathers and has a very short beak.
- A colorful bird that has blue over its head, neck and back, along with grey secondaries and blue wingbars.
- Vibrant blue and black bird, mostly blue except for outer rectrices and primaries.

- This bird is brown and white in color with a small brown beak and brown eye rings.
- A small, light brown bird with a white belly and breast, black spots on its wings, and black eyes.
- This small brown bird has a small gray and yellow beak with light tan breast.

- This flower is white and yellow in color, with petals that are striped.
- This flower has petals that are white with purple lines
- The flower shown has thick yellow and white flowers as its main feature.

Figure 3. The top row shows three zero-shot generations and the corresponding captions consumed in the process. The first two images belong to Indigo Bunting, Tree Sparrow class of CUB dataset and the last image belongs to Peruvian Lily class of Flowers dataset. None of these classes were used while training the model. The bottom row shows some random samples of generated images. All the images are generated for classes that are unseen while training the model.

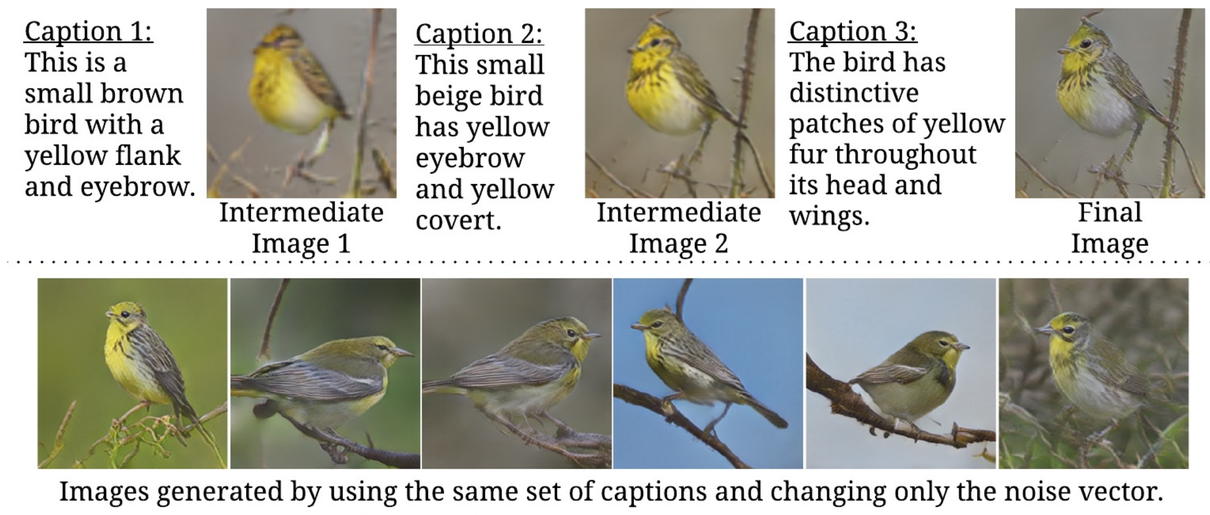
4.2 Quantitative Results

Dataset	GAN-INT-CLS	GAWWN	StackGAN	StackGAN++	DistillGAN
CUB	2.88 ± .04	3.62 ± .07	3.70 ± .04	3.82 ± .06	**3.92 ± .11**
Oxford-102	2.66 ± .03	-	3.20 ± .01	-	**3.41 ± .17**

Table 1. Comparison with other text-to-image synthesis methods.

4.3 Additional Results

- The progressive improvement in the quality of the image, after consuming each caption is captured in the first row.
- The birds in the second row are generated by changing the noise vector used to condition the GAN, while keeping the input text the same. This generates images with the same bird but in various poses and backgrounds.



Images generated by using the same set of captions and changing only the noise vector.

Figure 4. Additional qualitative results.

Selected References

[1] Zhang, Han et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *ICCV*, 2017.

[2] Zhang, Han et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *arXiv: 1710.10916*, 2017.

Towards Learning with Limited Labels Workshop, International Conference on Machine Learning (ICML), Stockholm, Sweden, July 2018.