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 \(t_1, \ldots, t_4\) is shown below.

3. Architecture

4.1 Qualitative 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.

4.2 Quantitative Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GAN-INT-CLS</th>
<th>GAWWN</th>
<th>StackGAN</th>
<th>StackGAN++</th>
<th>DistillGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUB</td>
<td>2.88 ± .04</td>
<td>3.62 ± .07</td>
<td>3.70 ± .04</td>
<td>3.82 ± .06</td>
<td>3.92 ± .11</td>
</tr>
<tr>
<td>Oxford-102</td>
<td>2.66 ± .03</td>
<td>-</td>
<td>3.20 ± .01</td>
<td>-</td>
<td>3.41 ± .17</td>
</tr>
</tbody>
</table>

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.

Selected References