Attention Driven Image Synthesis From Text Description

Anurag Paul, Harshul Gupta, Aashi Jain
Goal

Realistic Image Generation conditioned on the text descriptions with comparable Inception score to that of the SOTA results.

This bird has a black body and tail, a yellow head with black eyes, and a small black pointed beak.
Methods

This problem can be seen as an encoder-decoder problem.

- Text Encoder: Bidirectional LSTM, BERT
- Image Encoder: Pretrained InceptionV3
- Image Generator (Decoder): GAN, Transformer
Datasets

1. CUB Dataset: ~12K images of 200 bird classes with 10 captions/image with features like color, shape, etc.

2. Flickr 30 Dataset: ~30K image-caption corpus describing people involved in activities with 5 captions/image.
CUB Dataset examples
Some of the captions:

1. a medium sized bird that has a white belly and a very short stout bill
2. this is a black bird with a white eyering and a white belly and a orange bill
3. this is a bird with a white belly, black back and an orange beak.
Some of the captions:
1. A white and gray body bird with a regularly sized head in comparison to the body.
2. This bird has wings that are brown and has a white belly.
3. This bird has a metallic black color eye and small skinny tarsuses.
Flickr30k Dataset examples
Some of the captions:
1. A man sits in a chair while holding a large stuffed animal of a lion.
2. A man is sitting on a chair holding a large stuffed animal.
3. A man holds a large stuffed lion toy.
Proposed Solutions
- LSTM Text Encoder + CNN Image Encoder + Attention GAN with Image Self-Attention

- BERT Text Encoder + CNN Image Encoder + Attention GAN with Image Self Attention
Text-to-Image Generation Architecture
This bird is white and black in color with a black beak and black eye rings.
Implementation Details
Pretraining: Text Encoder

Bidirectional LSTM Model

**Word Features**: Semantic meaning of a word is obtained by concatenating both of its hidden states thereby generating a feature matrix of size: Batch x Embedding-Dim x Seq-Length

**Sentence Features**: Last hidden units of the Text-Encoder is used for generating the global sentence vector of size: Batch x Embedding-Dim
Pretraining: Image Encoder

**ImageNet pretrained Inception-v3 Model**: Learn features corresponding to different sub-regions of an image.

- **Image features** from mixed_6e layer of Inception network are converted to same feature space as that of the text encodings (i.e. 256) using a conv1x1 layer.

- **Global image features** from last average pooling layer is given to a FC layer to get same feature space as the global sentence vector.
Deep Attentional Multimodal Similarity Model (DAMSM)

Helps in matching an image-sentence pair based on an attention model between the image and the text.

**Words Loss:** We calculate similarity between $i^{th}$ word of the sentence and $j^{th}$ sub-region of the image. Hence it gives region-context vector for a word.

**Sentence Loss:** We calculate similarity between the global sentence vector and the global image vector.

DAMSM ensures that the word vectors generated are visually discriminative. Hence, word vectors of different colours are not clustered together.
Losses

**Generator Loss**: Generator tries to minimize the adversarial loss which is a combination of:

- **Unconditional Loss** determines whether the image is real or fake
- **Conditional Loss** determines whether the image and the sentence match or not

\[
\mathcal{L}_{G_i} = -\frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(D_i(\hat{x}_i))] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log(D_i(\hat{x}_i, \varepsilon))]
\]

- unconditional loss
- conditional loss
Losses

**Discriminator Loss**: sum of conditional as well as unconditional loss, where

- **Unconditional Loss** uses cross-entropy loss to classify input into real or fake.
- **Conditional Loss** determines if true image distribution and global sentence vector matches or not.

\[
\mathcal{L}_{D_i} = -\frac{1}{2} \mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i)] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log (1 - D_i(\hat{x}_i))] +
\]

unconditional loss

\[
-\frac{1}{2} \mathbb{E}_{x_i \sim p_{data_i}} [\log D_i(x_i, \bar{c})] - \frac{1}{2} \mathbb{E}_{\hat{x}_i \sim p_{G_i}} [\log (1 - D_i(\hat{x}_i, \bar{c}))],
\]

conditional loss
Overall Loss

\[ \mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM}, \quad \text{where} \quad \mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}. \]

\[ \mathcal{L}_{DAMSM} = \mathcal{L}_1^w + \mathcal{L}_2^w + \mathcal{L}_1^s + \mathcal{L}_2^s. \]
LSTM Text Encoder + CNN Image Encoder + Attn GAN with Self-Attn Layers
this bird is white and black in color with a black beak and black eye rings
LSTM Loss Plots for CUB Dataset
The **off white** belly throat and breast on the bird with tan wings and crown

This large **black** bird has large **black** wings with a **red** nape across the top of the head
Sample Results

this is a **black** bird with a long beak turned down at the end and a **huge red chest** that sticks out

the bird has a **yellow** belly small **black** bill and **striped** back
Sample Results

A small bird with a **grey** head and **black grey** feather and a **white throat** and belly.

this bird has wings that are **brown** and has a **striped belly**
Good Attempts

this bird has **large wings** is **black and white** in color with a **black head** has narrow black legs has a short tail and a long **orange pointed beak**

this bird has a **rust colored body** with a **black and white head** with wings that have a **shade of pink** in them
BERT Text Encoder + CNN
Image Encoder + AttnGan
with Self-Attn Layers
This bird has a yellow belly and breast with a short pointy bill.
BERT-Based Text Encoder
BERT Loss Plots for CUB Dataset

DAMSM(BERT) Sentence Loss

DAMSM(BERT) Word Loss
BERT Loss Plots for Flickr30K

**DAMSM(BERT) Sentence Loss**

- Train
- Validation

**DAMSM(BERT) Word Loss**

- Train
- Validation
This bird has a **yellow belly** and breast with a **short pointy** bill.

A **small** bird that is **grey, black and white**
Experimental Results and Analysis
Evaluation Metric: Inception Score

<table>
<thead>
<tr>
<th>Model Configuration</th>
<th>Inception Score</th>
<th>Epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attn GAN (state-of-the-art)</td>
<td>4.36</td>
<td>600</td>
</tr>
<tr>
<td>LSTM + Attn GAN + Self Attn</td>
<td>4.11</td>
<td>400</td>
</tr>
<tr>
<td>BERT + Attn GAN + Self Attn</td>
<td>4.23</td>
<td>~200</td>
</tr>
</tbody>
</table>
Inception Score Comparison over Epochs

Inception Score Comparison for the two Text Encoders

- LSTM-CUB
- BERT-CUB
Comparison

- With the limited resources, we were able to achieve an Inception score comparable to that of the SOTA on the test set.
- Self-Attention has helped in reducing the count of bird images with multiple heads and tails.
Analysis

- With BERT Text Encoder, we are able to train DAMSM Model in about 130 epochs compared to 500 with bidirectional LSTM

- Inception Score with BERT is generally higher when compared to with LSTM Encoder
Next steps

2-week plan:
● Further fine-tuning of the model
● Get results on Flickr30K dataset
● BERT Encoder with Vanilla Self Attention Architecture

4-week plan:
● Get results on COCO
● Get results for Image Transformers