Our AttnGAN

A novel attentional generative network

- Progressively generate low-to-high resolution images with m generators
- Attention model term
  - For each region feature of previous generated image, query its most relevant words.
  - Synthesizes fine-grained details at different sub-regions of the image by paying attention to the relevant words in the natural language description.
- The final objective function
  \[
  L = \sum_{i=0}^{m-1} L_{GAN} + \lambda L_{DAMSM}
  \]

- Text encoder (LSTM) extracts word features \( c_1, c_2, \ldots, c_T \)
- Image encoder (CNN) extracts image region features \( v_1, v_2, \ldots, v_k \)
- Attention mechanism: for the \( i \)-th word, compute its region-context vector \( c_i \),
  \[
  c_i = \sum_{j=1}^{k} a_j v_j, \quad a_j = \frac{\exp(y_j z_i)}{\sum_j \exp(y_j z_i)}
  \]
  - \( z_i \) is the dot product between features of the \( i \)-th word and the \( j \)-th image region

- The similarity between the image (Q) and the sentence (D)
  \[
  R(Q, D) = \log \left( \sum_{i=1}^{n} \exp \left( y_j R(c_i, v_j) \right) \right)^{1/\gamma}
  \]
  - \( R(c_i, v_j) \) is the cosine similarity between \( c_i \) and \( v_j \)
- The negative log posterior probability that the images are matched with their ground truth text descriptions
  \[
  L_{DAMSM} = -\sum_{i=0}^{M} \log P(D_i | Q), \quad P(D_i | Q) = \frac{\exp(y_j R(c_i, v_j))}{\sum_j \exp \left( y_j R(c_i, v_j) \right)}
  \]
  - \( M \) is the number of training pairs
  - \( y_j, a_1, a_2 \) and \( z_i \) are hyper-parameters
  - The \( L_{DAMSM} \) provides a fine-grained image-text matching loss for training the generator

A Deep Attentional Multimodal Similarity Model (DAMSM)

\[
\begin{align*}
\text{Image Encoder} & \quad \rightarrow \quad \text{Text Encoder} \\
\text{Attentional Generative Network} & \quad \rightarrow \quad \text{Local image features} \\
\text{Word features} & \quad \rightarrow \quad \mathcal{L}_{DAMSM}
\end{align*}
\]

This bird is red with white and has a very short beak

\[
\begin{align*}
\text{Encoder} & \quad \rightarrow \quad \text{Text} \\
\text{Attention Maps} & \quad \rightarrow \quad \text{Attention Maps}
\end{align*}
\]

\[
\begin{align*}
\text{Dataset} & \quad \rightarrow \quad \text{Inception Score} \\
\text{CUB} & \quad 2.88 \pm 0.04 \\
\text{COCO} & \quad 3.82 \pm 0.07 \\
\text{Vanilla DCGAN} & \quad 3.70 \pm 0.04 \\
\text{Attention Maps on CUB} & \quad 3.82 \pm 0.06 \\
\text{Our AttnGAN} & \quad 4.36 \pm 0.03 \\
\end{align*}
\]

Results

- The DAMSM loss is important
- Stacking more attention models helps
- Attention maps on CUB (left) and COCO (right)
- Novel images on CUB (left) and COCO (right)
- Compare with state-of-the-art
- Generalize the proposed attention mechanisms to DCGAN framework