Generating Images from Captions with Attention

Elman Mansimov
Emilio Parisotto
Jimmy Ba
Ruslan Salakhutdinov

ICLR 2016
Motivation

• To simplify the image modelling task
  • Captions contain more information about the image.
  • Although you need to learn language model.
• To better understand model generalization
  • Create textual descriptions of completely new scenes not seen at training time.
Novel Compositions

A **stop sign** is flying in blue skies.

A **pale yellow school bus** is flying in blue skies.

A **herd of elephants** flying in blue skies.

A **large commercial airplane** flying in blue skies.
General Idea

- Part of the sequence-to-sequence framework. (Sutskever et al. 2014; Cho et al. 2014; Srivastava et al. 2015)

- Caption is represented as a sequence of consecutive words.

- Image is represented as a sequence of patches drawn on canvas.

- Also need to figure out where to put generated patches on canvas.
Language Model
(Bidirectional RNN)

- Forward LSTM reads sentence from left to right
- Backward LSTM reads sentence from right to left
- Sentence representation is average of hidden states

Image Model
(DRAW: Variational Recurrent Auto-encoder with Visual Attention)

- At each step model produces $p \times p$ patch.
- It gets transformed into $h \times w$ canvas using two arrays of 1D filter banks ($h \times p$ and $w \times p$ respectively).
- Mean and variance of latent variables depend on the previous hidden states of generative RNN.

Gregor et. al. 2015
Model is trained to maximize variational lower bound

\[
\mathcal{L} = \mathbb{E}_{Q(Z_{1:T} \mid y, x)} \left[ \log p(x \mid y, Z_{1:T}) - \sum_{t=2}^{T} D_{KL} \left( Q(Z_t \mid Z_{1:t-1}, y, x) \parallel P(Z_t \mid Z_{1:t-1}, y) \right) \right] - D_{KL} \left( Q(Z_1 \mid x) \parallel P(Z_1) \right)
\]

Kingma et. al. 2014, Rezende et. al. 2014
Compute alignment between words and generated patches

\[ e^t_j = v^\top \tanh(U h^\text{lang}_j + W h^\text{gen}_{t-1} + b) \]

\[ \alpha^t_j = \frac{\exp(e^t_j)}{\sum_{j=1}^{N} \exp(e^t_j)} \]

Bahdanau et. al. 2015
Sharpening

• Another network trained to generate edges sharpens the generated samples.

• Instead is trained to fool separate network that discriminates between real and fake samples.

• Doesn’t have reconstruction cost and gets sharp edges.

Goodfellow et. al. 2014, Denton et. al. 2015
Complete Model
Main Dataset (Microsoft COCO)

- Contains ~83k images
- Each image has 5 captions
- Standard benchmark dataset for recent image captioning systems

Lin et. al. 2014
A yellow school bus parked in a parking lot.

A red school bus parked in a parking lot.

A green school bus parked in a parking lot.

A blue school bus parked in a parking lot.
Flipping Backgrounds

A very large commercial plane flying in clear skies.

A very large commercial plane flying in rainy skies.

A herd of elephants walking across a dry grass field.

A herd of elephants walking across a green grass field.
The decadent chocolate dessert is on the table.

A bowl of bananas is on the table.

A vintage photo of a cat.

A vintage photo of a dog.
Examples of Alignment

A rider on the blue motorcycle in the desert.

A rider on the blue motorcycle in the forest.

A surfer, a woman, and a child walk on the beach.

A surfer, a woman, and a child walk on the sun.
A very large commercial plane flying in clear skies.

A large airplane flying through a blue sky.

A stop sign is flying in blue skies.

A picture of a building with a blue sky.

A toilet seat sits open in the grass field.

A window that is in front of a mirror.

with Jamie Kiros (Xu et al. 2015)
More Results
(Image Retrieval and Image Similarity)

<table>
<thead>
<tr>
<th>Model</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@50</th>
<th>Med r</th>
<th>SSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAPGAN</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.08</td>
</tr>
<tr>
<td>Fully-Conn VAE</td>
<td>1.0</td>
<td>6.6</td>
<td>12.0</td>
<td>53.4</td>
<td>47</td>
<td>0.156</td>
</tr>
<tr>
<td>Conv-Deconv VAE</td>
<td>1.0</td>
<td>6.5</td>
<td>12.0</td>
<td>52.9</td>
<td>48</td>
<td>0.164</td>
</tr>
<tr>
<td>skipthoughtDRAW</td>
<td>2.0</td>
<td>11.2</td>
<td>18.9</td>
<td>63.3</td>
<td>36</td>
<td>0.157</td>
</tr>
<tr>
<td>noalignDRAW</td>
<td>2.8</td>
<td>14.1</td>
<td>23.1</td>
<td>68.0</td>
<td>31</td>
<td>0.155</td>
</tr>
<tr>
<td>alignDRAW</td>
<td>3.0</td>
<td>14.0</td>
<td>22.9</td>
<td>68.5</td>
<td>31</td>
<td>0.156</td>
</tr>
</tbody>
</table>
## Lower Bound of Log-Likelihood in Nats

<table>
<thead>
<tr>
<th>Model</th>
<th>Train</th>
<th>Test</th>
<th>Test (after sharpening)</th>
</tr>
</thead>
<tbody>
<tr>
<td>skipthoughtDRAW</td>
<td>-1794.29</td>
<td>-1791.37</td>
<td>-2045.84</td>
</tr>
<tr>
<td>noalignDRAW</td>
<td>-1792.14</td>
<td>-1791.15</td>
<td>-2051.07</td>
</tr>
<tr>
<td>alignDRAW</td>
<td>-1792.15</td>
<td>-1791.53</td>
<td>-2042.31</td>
</tr>
</tbody>
</table>
Qualitative Comparison

Our Model

Conv-Deconv VAE

LAPGAN

Fully-Connected VAE

A group of people walk on a beach with surf boards
Conclusions

• Samples from our generative model are okay; but aren’t great.

• Mostly because the model is underfitting.

• The model generalizes to captions describing novel scenarios that are not seen in the dataset.

• Key factor, treat image generation as computer graphics. Learn what to generate and where to place it.
Thank You!


Code: https://github.com/emansim/text2image