Describe That GIF
A GIF Description Generator

Motivation
Finding the right GIF is hard. Despite their rising popularity, there is a surprising lack of scholarly work on animated GIFs in the computer vision community. We take a step towards improved GIF search and making GIFs more accessible by generating natural language descriptions of GIFs. We trained two models: CNN – LSTM and CNN - LSTM with attention and analyzed their performance on the TGF dataset.

Results
<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (last LSTM)</td>
<td>46.5</td>
<td>22.1</td>
<td>13.1</td>
<td>7.6</td>
<td>14.0</td>
<td>31.2</td>
<td>14.5</td>
</tr>
<tr>
<td>CNN-LSTM</td>
<td>46.5</td>
<td>22.1</td>
<td>13.1</td>
<td>7.6</td>
<td>14.0</td>
<td>31.2</td>
<td>14.5</td>
</tr>
<tr>
<td>CNN-LSTM with Attention</td>
<td>39.4</td>
<td>18.0</td>
<td>6.3</td>
<td>4.5</td>
<td>13.5</td>
<td>26.6</td>
<td>12.7</td>
</tr>
</tbody>
</table>

The CNN-LSTM model performed the best, outperforming the baseline model by 90.8% (6.9) on CIDER score, 8.8% (1.2) on METEOR score and 42% (2.2) on BLEU-4 score.

Data and Features
90,000 GIFs (from Tumblr) with descriptions
82% train, 5.5% dev, and 12.5% test
Training and validation set contains one reference description per GIF
Test set contains three reference descriptions per GIF

9.

an elephant appears to be jumping on a trampoline.

an elephant appears to be jumping on a trampoline.

an elephant appears to be jumping on a trampoline.

an elephant appears to be jumping on a trampoline.

Encoder:
pretrained ResNet-34 Kinetics CNN
- Splits GIF into individual frames and encoded every 10th frame
- Capped number of encodings at 10, resulting in a n x 5120 dimension vector which was our input into the LSTM models.

Models
CNN-LSTM Model

with sampling to minimize the cross entropy loss

CNN-LSTM Model with Attention
Attention Equations

Hyperparameters

Models

CNN-LSTM Model

with sampling to minimize the cross entropy loss

CNN-LSTM Model with Attention
Attention Equations

Hyperparameters

Future Work
To improve model performance:
Train the CNN-LSTM model on complete GIFs without the 10 frame cap
Encode the GIFs using a deeper CNN model like the ResNet-101 trained on the Kinetics dataset
Experiment with different types of attention and visualize what the model is attending to during decoding.

Analysis
The model produces good, complete, grammatically correct descriptions with few <unk> tokens

GT: three teenagers are dancing on a glass stage.
CL: a group of boys are dancing in rhythm.

...but it is not perfect

GT: a woman with glasses talked and moved her fingers
CL: a man is talking to someone else indoors.

...maybe far from perfect?

GT: a girl is riding a skateboard down the street.
CL: a woman holding a glass bottle and exhaling smoke.

References