Show and tell: A Neural Image Caption Generator

SHUANGFEI FAN
Framework

A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.
Framework
Objective

- Loss for each training pair:

\[ L(I, S) = - \sum_{t=1}^{N} \log p_t(S_t) \]

- Optimization (SGD):

\[ \theta^* = \arg \max_{\theta} \sum_{(I,S)} \log p(S|I; \theta) \]
## Performance (BLEU-1 scores)

<table>
<thead>
<tr>
<th>Approach</th>
<th>PASCAL (xfer)</th>
<th>Flickr 30k</th>
<th>Flickr 8k</th>
<th>SBU</th>
<th>MSCOCO (BLEU-4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Im2Text [18]</td>
<td>25</td>
<td></td>
<td>48</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>TreeTalk [14]</td>
<td></td>
<td>55</td>
<td>58</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>BabyTalk [3]</td>
<td></td>
<td>56</td>
<td>51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tri5Sem [16]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m-RNN [27]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MNLM [29]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SOTA</strong></td>
<td>25</td>
<td>56</td>
<td>58</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td><strong>NIC</strong></td>
<td>59</td>
<td>66</td>
<td>63</td>
<td>28</td>
<td>27.7</td>
</tr>
<tr>
<td><strong>Human</strong></td>
<td>69</td>
<td>68</td>
<td>70</td>
<td></td>
<td>21.7</td>
</tr>
</tbody>
</table>
System Set-up

- OS: Ubuntu 16.4
- GPU with CUDA
- Platform: Tensorflow
- Dependencies
  - Bazel (build tool)
  - Numpy
  - NLTK (Natural Language Toolkit)
- Trained for 36 hours (467102 steps), compared to training for 1000000 steps (1-2 weeks)
Dataset

- Microsoft COCO

Ground truth:
- A person on a surfboard rides on a wave
- A person is surfing on a wave in the ocean
- A person on a surfboard riding a wave
- A person in black wetsuit riding a yellow surfboard on water
- A person in a black shirt riding on a surfboard

Setup
- Training set: 82783
- Test set: 4k from validation set
Good results for images in test set
Results of images from validation dataset

Captions generated by NIC model:
0) a man riding a wave on top of a surfboard (p=0.052550)
1) a person riding a surf board on a wave (p=0.017121)
2) a man riding a wave on a surfboard in the ocean (p=0.007918)

Ground truth:
1. A person on a surfboard rides on a wave
2. A person is surfing on a wave in the ocean
3. A person on a surfboard riding a wave
4. A person in black wetsuit riding a yellow surfboard on water
5. A person in a black shirt riding on a surfboard
Results of images from validation dataset

Captions generated by NIC model:
0) a bunch of items that are on a table. (p=0.000370)
1) a pile of luggage sitting on top of a table. (p=0.000222)
2) a bunch of items that are on a bed. (p=0.000127)

Ground truth:
1. A table with cell phones and clutter on top of it
2. Hiking equipment and gear neatly spread out on a table.
3. An image of books and other electronics on the table.
4. A table with hiking supplies such as lotions and equipment.
5. Many travel items have been placed neatly on a desk.
Bad results for images in test set
Results of images from validation dataset

Captions generated by NIC model:
0) a baseball player swinging a bat at a ball (p=0.009450)
   1) a baseball player holding a bat on top of a field. (p=0.003202)
   2) a baseball player swinging a bat on a field (p=0.002669)

Ground truth:
1. A baseball game is in action at the plate.
2. A group of men in a field playing baseball.
3. A man with a baseball bat that is standing in the dirt.
4. The Umpire, catcher, and batter playing baseball, just as batter swung his bat.
5. there is a baseball game on and a player has swung for the ball.

Weakness: only focus on one object
Captions generated by NIC model:
0) a black and white cat sitting on a window sill . (p=0.000052)
1) a black and white cat sitting on top of a table .
   (p=0.000042)
2) a black and white cat sitting on top of a window sill .
   (p=0.000041)

Ground truth:
1. A cat is laying on the other side of a cactus.
2. a cat laydoing by a window close to a plant
3. The cat is laying down by the cactus in the sun.
4. Grey cat sitting in front of a green cactus
5. A cat behind a cactus, in front of a window.

Weakness: hard to recognize from part of the object (cactus)
Results of images from validation dataset

Captions generated by NIC model:
0) a green street sign sitting on the side of a road. (p=0.000968)
1) a green street sign sitting on top of a pole. (p=0.000204)
2) a green street sign sitting on top of a metal pole. (p=0.000186)

Ground truth:
1. A street post with three different street signs on it
2. The pole has three different street names on it
3. Three street signs that are in a residential neighborhood
4. A street sign with three signs mounted on it
5. A street sign has three different streets on it

Weakness: can't reconstructs 3D information
Captions generated by NIC model:
0) a group of people sitting at a table eating pizza .
   (p=0.000217)
1) a group of people sitting at a table with pizza .
   (p=0.000203)
2) a group of people sitting at a table with pizza and drinks .
   (p=0.000037)

Ground truth:
1. A table with cell phones and clutter on top of it
2. Hiking equipment and gear neatly spread out on a table.
3. an image of books and other electronics on the table
4. A table with hiking supplies such as lotions and equipment.
5. Many travel items have been placed neatly on a desk.

Weakness: hard to retrieve 3D information
Good results for images in my phone
Results of other images

Captions generated by NIC model:

0) a group of people standing under an umbrella . (p=0.000384)
1) a group of people standing under an umbrella (p=0.000194)
2) a woman holding an umbrella in front of a building . (p=0.000086)
Results of other images

Captions generated by NIC model:

0) a white cat sitting on top of a table. (p=0.000252)
1) a white cat sitting on top of a white table. (p=0.000164)
2) a white cat is sitting on a table (p=0.000067)
Results of other images

Captions generated by NIC model:

0) a cat laying on top of a bed . (p=0.000902)
1) a cat laying on top of a bed next to a window . (p=0.000307)
2) a cat laying on top of a bed next to pillows . (p=0.000167)
Results of other images

Captions generated by NIC model:

0) a cat laying on top of a pair of shoes . (p=0.001615)
1) a cat laying on top of a pair of shoes (p=0.000254)
2) a cat laying on a rug next to a shoe . (p=0.000197)
Bad results for images in my phone
Results of other images

Captions generated by NIC model:

0) a small dog is sitting on a table . (p=0.000013)
1) a small dog is sitting on a table (p=0.000012)
2) a small dog is sitting on a chair (p=0.000007)
Results of other images

Captions generated by NIC model:

0) a wooden bench sitting in front of a building . (p=0.000209)
1) a wooden bench sitting in the middle of a park . (p=0.000082)
2) a wooden bench sitting in front of a brick wall . (p=0.000070)
Results of other images

Captions generated by NIC model:

0) a computer mouse sitting on top of a desk . (p=0.000116)

1) a computer mouse sitting on top of a wooden desk . (p=0.000078)

2) a computer mouse sitting on top of a table . (p=0.000074)
Results of other images

Captions generated by NIC model:

0) a cat laying on top of a suitcase. (p=0.000289)
1) a cat laying on top of a red chair. (p=0.000137)
2) a cat laying on top of a suitcase on a bed. (p=0.000063)
Results of other images

Captions generated by NIC model:

0) a cat laying on top of a pair of shoes . (p=0.000895)
1) a cat is laying on top of a suitcase . (p=0.000334)
2) a cat laying on top of a blue suitcase . (p=0.000257)
Results of other images

Captions generated by NIC model:

0) a cat sitting on the floor under an umbrella . (p=0.000150)
1) a cat sitting on the floor under an umbrella (p=0.000082)
2) a cat sitting on the floor under an open umbrella . (p=0.000025)
Conclusion

- 3D information
  - Retrieve 3D information from images during encoding procedure
- Focus on multiple objects or scenes instead of just one
  - multi-label classification model
- Zero-shot learning: Generate captions for unknown objects or scenes (oscilloscope)
  - Using attributes
  - Hard to recognize from part of the object