Show and Tell: A Neural Image Caption Generator
Vinyals et al. (Google)

The IEEE Conference on Computer Vision and Pattern Recognition, 2015
The Problem

- **Image Caption Generation**
  - Automatically describe content of an image
  - Image $\rightarrow$ Natural Language
  - Computer Vision + NLP
  - Much more difficult than image classification/recognition
Background

- Success in image classification/recognition
- Close to human level performance
- Deep CNN’s, Big Datasets
- Image to fixed length vector
Background

- Machine Translation
- Language generating RNN’s
- Decoder-Encoder framework
- Maximize likelihood of target sentence
Idea

- Combine Vision CNN with Language RNN
- Deep CNN as encoder
- Language Generating RNN as decoder
- End to end model $I \rightarrow S$
- Maximize $p(S|I)$
The Model

Neural Image Caption (NIC)

- CNN: 22 layer GoogleNet
- LSTM for modeling
  \[ \log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1}) \]
- Word embedding \( W_e \)
Language LSTM

- Predicts next word in sentence
- Memory cell for longer memory
- $S_t$ one-hot vectors + START/END token
- $x_{t-1} = \text{CNN}(I)$, $x_t = W_e S_t$, $p_{t+1} = \text{LSTM}(x_t)$

\[
\begin{align*}
    i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1}) \\
    f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \\
    o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1}) \\
    c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) \\
    m_t &= o_t \odot c_t \\
    p_{t+1} &= \text{Softmax}(m_t)
\end{align*}
\]
Training

- Loss function \( L(I, S) = - \sum_{t=1}^{N} \log p_t(S_t) \)
- CNN pre-trained on ImageNet
- Minimize w.r.t. LSTM parameters, \( W_e \) and CNN top layer
- SGD on mini-batches
- Dropout and ensembling
- 512 dimensional embedding
Generation

- Give $x_{-1} = \text{CNN}(I)$
- $x_0 = W_e S_0$, $S_0$ START token
- Sample word $S_1$
- Feed $W_e S_1$ to LSTM
- BeamSearch, beam size 20
Results

- MSCOCO dataset: 80k train, 40k eval and test
- 5 human made captions per image
- M1-M5 human judgements

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human[^5^]</td>
<td>0.638</td>
<td>0.675</td>
<td>4.836</td>
<td>3.428</td>
<td>0.352</td>
</tr>
<tr>
<td>Google[^4^]</td>
<td>0.273</td>
<td>0.317</td>
<td>4.107</td>
<td>2.742</td>
<td>0.233</td>
</tr>
<tr>
<td>MSR[^11^]</td>
<td>0.268</td>
<td>0.322</td>
<td>4.137</td>
<td>2.662</td>
<td>0.234</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>CIDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIC</td>
<td>27.7</td>
<td>23.7</td>
<td>85.5</td>
</tr>
<tr>
<td>Random</td>
<td>4.6</td>
<td>9.0</td>
<td>5.1</td>
</tr>
<tr>
<td>Nearest Neighbor</td>
<td>9.9</td>
<td>15.7</td>
<td>36.5</td>
</tr>
<tr>
<td>Human</td>
<td>21.7</td>
<td>25.2</td>
<td>85.4</td>
</tr>
</tbody>
</table>

Table 1. Scores on the MSCOCO development set.
Results

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

A little girl in a pink hat is blowing bubbles.

A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.

A close up of a cat laying on a couch.

A red motorcycle parked on the side of the road.

A yellow school bus parked in a parking lot.
Results

- Improved Flickr8k, Flickr30k, PASCAL BLEU scores
- Need better evaluation metrics
- 80% of top-1 in training set
- 50% of top-15 in training set
- Similar diversity as human captions

| 1. A man throwing a frisbee in a park. |
| 2. A man holding a frisbee in his hand. |
| 3. A man standing in the grass with a frisbee. |

| 1. A close up of a sandwich on a plate. |
| 2. A close up of a plate of food with french fries. |
| 3. A white plate topped with a cut in half sandwich. |

| 1. A display case filled with lots of donuts. |
| 2. A display case filled with lots of cakes. |
| 3. A bakery display case filled with lots of donuts. |
Results

- Trained word embeddings $W_e$

<table>
<thead>
<tr>
<th>Word</th>
<th>Neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>van, cab, suv, vehicule, jeep</td>
</tr>
<tr>
<td>boy</td>
<td>toddler, gentleman, daughter, son</td>
</tr>
<tr>
<td>street</td>
<td>road, streets, highway, freeway</td>
</tr>
<tr>
<td>horse</td>
<td>pony, donkey, pig, goat, mule</td>
</tr>
<tr>
<td>computer</td>
<td>computers, pc, crt, chip, compute</td>
</tr>
</tbody>
</table>

- Captures semantics from the language data
- Independent of vocabulary size
Summary

- End-to-end model (Encoder-Decoder)
- Vision CNN + Language generating RNN
- Maximize likelihood of $S$ given $I$
- State of the art results on major datasets
- Datasets are limiting: Unsupervised approaches?