Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)

http://www.stat.ucla.edu/~junhua.mao/m-RNN.html

Junhua Mao\textsuperscript{1,2}, Wei Xu\textsuperscript{1}, Yi Yang\textsuperscript{1}, Jiang Wang\textsuperscript{1}, Zhiheng Huang\textsuperscript{1}, Alan Yuille\textsuperscript{2}

\textsuperscript{1}Baidu Research \hspace{1cm} \textsuperscript{2}University of California, Los Angeles

- a close up of a bowl of food on a table
- a train is traveling down the tracks in a city
- a pizza sitting on top of a table next to a box of pizza
Abstract

• Three Tasks:
  ▪ Image caption generation
  ▪ Image retrieval (given query sentence)
  ▪ Sentence retrieval (given query image)

• One model (m-RNN):
  ▪ A deep Recurrent NN (RNN) for the sentences
  ▪ A deep Convolutional NN (CNN) for the images
  ▪ A multimodal layer connects the first two components

• State-of-the-art Performance:
  ▪ For three tasks
  ▪ On four datasets: IAPR TC-12 [Grubinger et al. 06′], Flickr 8K [Rashtchian et al. 10′], Flickr 30K [Young et al. 14′] and MS COCO [Lin et al. 14′]
The m-RNN Model

$w_{\text{start}}, w_{1}, ..., w_{L}$ is the sentence description of the image

$w_{\text{start}}, w_{\text{end}}$ is the start and end sign of the sentence
The m-RNN Model

Detailed calculation for recurrent $r(t)$ and multimodal layer $m(t)$

- $r(t) = f(U_r \cdot r(t-1) + w(t))$, $w(t)$ is the activation of embedding layer II for the word $w_t$
- $m(t) = g(V_w \cdot w(t) + V_r \cdot r(t) + V_I \cdot I)$, $I$ is the image representation
- "+" here means element-wise plus
The m-RNN Model

Layers: Embedding I Embedding II Recurrent Multimodal SoftMax

The m-RNN model for one word

Non-linear activation functions:
- For the recurrent layer: ReLU [Nair and Hinton 10'] \( f(x) = \max(0, x) \)
- For the word embedding layers and the multimodal layer: S-tanh [LeCun et.al 12']: \( g(x) = 1.72 \cdot \tanh\left(\frac{2}{3} x\right) \)
The m-RNN Model

The m-RNN model for one word

The output of the trained model:

\[ P(w_n|w_{1:n-1}, I) \]
Application

• Image caption generation:
  o Begin with the start sign $w_{start}$
  o Sample next word from $P(w_n|w_{1:n-1}, I)$
  o Repeat until the model generates the end sign $w_{end}$
Application

• Image caption generation:
  o Begin with the start sign $w_{start}$
  o Sample next word from $P(w_n|w_{1:n-1}, I)$
  o Repeat until the model generates the end sign $w_{end}$

• Image retrieval given query sentence:
  o Ranking score: $P(w_{1:L}^Q|I^D) = \prod_{n=2}^{L} P(w_n^Q|w_{1:n-1}^Q, I^D)$
  o Output the top ranked images
Application

• Image caption generation:
  o Begin with the start sign $w_{\text{start}}$
  o Sample next word from $P(w_n|w_{1:n-1}, I)$
  o Repeat until the model generates the end sign $w_{\text{end}}$

• Image retrieval given query sentence:
  o Ranking score: $P(w_{1:L}^Q|I^D) = \prod_{n=2}^{L} P(w_n^Q|w_{1:n-1}^Q, I^D)$
  o Output the top ranked images

• Sentence retrieval given query image:
  o Problem: Some sentences have high probability for any image query
  o Solution: **Normalize** the probability. $I'$ are images sampled from the training set:
    \[
    P(w_{1:L}^D|I^Q) \cdot P(w_{1:L}^D) = \sum_{I'} P(w_{1:L}^D|I') \cdot P(I')
    \]
Application

• Image caption generation:
  o Begin with the start sign $w_{start}$
  o Sample next word from $P(w_n|w_{1:n-1}, I)$
  o Repeat until the model generates the end sign $w_{end}$

• Image retrieval given query sentence:
  o Ranking score: $P(w^Q_{1:L}|I^D) = \prod_{n=2}^{L} P(w_n^Q|w_{1:n-1}^Q, I^D)$
  o Output the top ranked images

• Sentence retrieval given query image:
  o **Problem**: Some sentences have high probability for any image query
  o **Solution**: **Normalize** the probability. $I'$ are images sampled from the training set:
    \[
    P(w^D_{1:L}|I^Q) \frac{P(w^D_{1:L})}{P(w^D_{1:L})} = \sum_{I'} P(w^D_{1:L}|I') \cdot P(I')
    \]
  o Equivalent to using a ranking score: $P(I^Q|w^D_{1:L}) = \frac{P(w^D_{1:L}|I^Q) \cdot P(I^Q)}{P(w^D_{1:L})}$
# Experiment: Retrieval

## Table 1. Retrieval results on Flickr 30K and MS COCO

<table>
<thead>
<tr>
<th></th>
<th>Sentence Retrieval (Image to Text)</th>
<th>Image Retrieval (Text to Image)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td><strong>Flickr30K</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>DeepFE-RCNN (Karpathy et al. 14’)</td>
<td>16.4</td>
<td>40.2</td>
</tr>
<tr>
<td>RVR (Chen &amp; Zitnick 14’)</td>
<td>12.1</td>
<td>27.8</td>
</tr>
<tr>
<td>MNLM-AlexNet (Kiros et al. 14’)</td>
<td>14.8</td>
<td>39.2</td>
</tr>
<tr>
<td>MNLM-VggNet (Kiros et al. 14’)</td>
<td>23.0</td>
<td>50.7</td>
</tr>
<tr>
<td>NIC (Vinyals et al. 14’)</td>
<td>17.0</td>
<td>56.0</td>
</tr>
<tr>
<td>LRCN (Donahue et al. 14’)</td>
<td>14.0</td>
<td>34.9</td>
</tr>
<tr>
<td>DeepVS-RCNN (Karpathy et al. 14’)</td>
<td>22.2</td>
<td>48.2</td>
</tr>
<tr>
<td>Ours-m-RNN-AlexNet</td>
<td>18.4</td>
<td>40.2</td>
</tr>
<tr>
<td>Ours-m-RNN-VggNet</td>
<td>35.4</td>
<td>63.8</td>
</tr>
<tr>
<td><strong>MS COCO</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>DeepVS-RCNN (Karpathy et al. 14’)</td>
<td>29.4</td>
<td>62.0</td>
</tr>
<tr>
<td>Ours-m-RNN-VggNet</td>
<td>41.0</td>
<td>73.0</td>
</tr>
</tbody>
</table>

R@K: The recall rate of the groundtruth among the top K retrieved candidates

Med r: Median rank of the top-ranked retrieved groundtruth

(*) Results reported on 04/10/2015. The deadline for our camera ready submission.
## Experiment: Captioning

Table 2. Caption generation results on Flickr 30K and MS COCO

<table>
<thead>
<tr>
<th>Model</th>
<th>Flickr30K</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>MS COCO</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PERP</td>
<td>B-1</td>
<td>B-2</td>
<td>B-3</td>
<td>B-4</td>
<td>PERP</td>
<td>B-1</td>
<td>B-2</td>
<td>B-3</td>
<td>B-4</td>
</tr>
<tr>
<td>RVR (Chen &amp; Zitnick 14’)</td>
<td>-</td>
<td>0.47</td>
<td>0.21</td>
<td>0.09</td>
<td>0.13</td>
<td>0.53</td>
<td>0.28</td>
<td>0.15</td>
<td></td>
<td>0.19</td>
</tr>
<tr>
<td>DeepVS-AlexNet (Karpathy et al. 14’)</td>
<td>-</td>
<td>0.47</td>
<td>0.21</td>
<td>0.09</td>
<td>0.13</td>
<td>0.53</td>
<td>0.28</td>
<td>0.15</td>
<td></td>
<td>0.19</td>
</tr>
<tr>
<td>DeepVS-VggNet (Karpathy et al. 14’)</td>
<td>21.20</td>
<td>0.50</td>
<td>0.30</td>
<td>0.15</td>
<td>0.19</td>
<td>19.64</td>
<td>0.57</td>
<td>0.37</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>NIC (Vinyals et al. 14’)</td>
<td>-</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>0.67</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LRCN (Donahue et al. 14’)</td>
<td>-</td>
<td>0.59</td>
<td>0.39</td>
<td>0.25</td>
<td>0.16</td>
<td>-</td>
<td>0.63</td>
<td>0.44</td>
<td>0.31</td>
<td>0.21</td>
</tr>
<tr>
<td>DMSM (Fang et al. 14’)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.21</td>
</tr>
<tr>
<td>Ours-m-RNN-AlexNet</td>
<td>35.11</td>
<td>0.54</td>
<td>0.36</td>
<td>0.23</td>
<td>0.15</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours-m-RNN-VggNet</td>
<td>20.72</td>
<td>0.60</td>
<td>0.41</td>
<td>0.28</td>
<td>0.19</td>
<td>13.60</td>
<td>0.67</td>
<td>0.49</td>
<td>0.34</td>
<td>0.24</td>
</tr>
</tbody>
</table>

### B-K: BLEU-K score

\[ \text{BLEU-K} = \frac{1}{L} \sum_{n=1}^{L} \log_2 P(w_n|w_{1:n-1}, I) \]

### PERP: Perplexity

\[ \text{PERP} = \text{PerplexityL} = \log_2 P(w_{1:L}|I) = \frac{1}{L} \sum_{n=1}^{L} \log_2 P(w_n|w_{1:n-1}, I) \]

(*) Results reported on 04/10/2015. The deadline for our camera ready submission.
## Experiment: Captioning

### Table 4. Results on the MS COCO test set

<table>
<thead>
<tr>
<th></th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>CIDEr</th>
<th>ROUGE&lt;sub&gt;L&lt;/sub&gt;</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-c5 (***)</td>
<td>0.663</td>
<td>0.469</td>
<td>0.321</td>
<td>0.217</td>
<td>0.854</td>
<td>0.484</td>
<td>0.252</td>
</tr>
<tr>
<td>m-RNN-c5</td>
<td>0.668</td>
<td>0.488</td>
<td>0.342</td>
<td>0.239</td>
<td>0.729</td>
<td>0.489</td>
<td>0.221</td>
</tr>
<tr>
<td>m-RNN-beam-c5</td>
<td>0.680</td>
<td>0.506</td>
<td>0.369</td>
<td>0.272</td>
<td>0.791</td>
<td>0.499</td>
<td>0.225</td>
</tr>
<tr>
<td>Human-c40 (**)</td>
<td>0.880</td>
<td>0.744</td>
<td>0.603</td>
<td>0.471</td>
<td>0.910</td>
<td>0.626</td>
<td>0.335</td>
</tr>
<tr>
<td>m-RNN-c40</td>
<td>0.845</td>
<td>0.730</td>
<td>0.598</td>
<td>0.473</td>
<td>0.740</td>
<td>0.616</td>
<td>0.291</td>
</tr>
<tr>
<td>m-RNN-beam-c40</td>
<td>0.865</td>
<td>0.760</td>
<td>0.641</td>
<td>0.529</td>
<td>0.789</td>
<td>0.640</td>
<td>0.304</td>
</tr>
</tbody>
</table>

- **c5 and c40:** evaluated using 5 and 40 reference sentences respectively.
- "beam" means that we generate a set of candidate sentences, and then selects the best one. (beam search)

---

(**) Provided in https://www.codalab.org/competitions/3221#results

(***) We evaluate it on the MS COCO evaluation server: https://www.codalab.org/competitions/3221
Discussion
Discussion

Other language: Chinese

一个年轻的男孩坐在长椅上。
一列火车在轨道上行驶。
一辆双层巴士停在一个城市街道上。

We acknowledge Haoyuan Gao and Zhiheng Huang from Baidu Research for designing the Chinese image captioning system.
Discussion

Other language: Chinese

一个年轻的男孩坐在长椅上。
A young boy sitting on a bench.

一列火车在轨道上行驶。
A train running on the track.

一辆双层巴士停在一个城市街道上。
A double decker bus stop on a city street.

We acknowledge Haoyuan Gao and Zhiheng Huang from Baidu Research for designing the Chinese image captioning system.
Can we design a system that learns to describe new visual concepts from a few examples?
Can we design a system that learns to describe new visual concepts from a few examples?


- Efficiently enlarge the vocabulary
- Needs only *a few images* with only *a few minutes*
- Datasets for evaluation
For more details, please visit the project page: 
http://www.stat.ucla.edu/~junhua.mao/m-RNN.html

The updated version of our paper: http://arxiv.org/abs/1412.6632

The novel visual concept learning paper: http://arxiv.org/abs/1504.06692

a young girl brushing his teeth with a toothbrush

a group of people flying kites in a field

a man is doing a trick on a skateboard
Performance comparison with different word-embedding configuration
Appendix

Performance comparison with different image representation input methods

<table>
<thead>
<tr>
<th></th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>m-RNN</td>
<td>0.600</td>
<td>0.412</td>
<td>0.278</td>
<td>0.187</td>
</tr>
<tr>
<td>m-RNN-visInRnn</td>
<td>0.466</td>
<td>0.267</td>
<td>0.157</td>
<td>0.101</td>
</tr>
<tr>
<td>m-RNN-visInRnn-both</td>
<td>0.546</td>
<td>0.333</td>
<td>0.191</td>
<td>0.120</td>
</tr>
<tr>
<td>m-RNN-visInRnn-both-shared</td>
<td>0.478</td>
<td>0.279</td>
<td>0.171</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td>MNLM</td>
<td>NIC</td>
<td>LRCN</td>
<td>RVR</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td>300</td>
<td>512</td>
<td>1000 (x4)</td>
<td>100</td>
</tr>
<tr>
<td><strong>LSTM</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>