Learning a Recurrent Visual Representation for Image Caption Generation

Xinlei Chen and C. Lawrence Zitnick
Presented by Yu-Cheng Lin
CVPR 2015

Most slides from Xinlei Chen
From Labels to Texts

- Airplane
- Runway

From PASCAL VOC Challenge
TIME FOR HALLUCINATION!
A girl
A girl sits
A girl sits on a tire swing
A girl sits on a tire swing with a dog.
Can my computer learn something like that?

“A girl”
Model the sequence: memory

“A girl sits”
A simple model that memorizes ...
Un-roll in Time: Recurrent
Predicting the next

- Recurrent Neural Network based Language Model (RNNLM) [Mikolov et.al. 2010]
Now, back to our task

• Learn a model between images and sentences, so that:
  – Given an image, it can retrieve/\textit{generate} a description
  – Given a description, it can retrieve/\textit{hallucinate} an image
Generating Image Descriptions

• EASY, Objective Function: $\Pi t \left[ P(w \downarrow t | I, H \downarrow t ) \right]$
Provide more supervision!
Solution: Split into 2!

Memorizes the semantics of the entire image

But no information flow from \( s \) to \( i \)!

Memorizes the visualizable parts of the sentence so far
Our Model (Per Stage)

\[ \Pi_{t} \left[ P(w_{t}, I | H_{t}) \right] \]
Visual memory

Semantic memory

\[ \prod_t [P(w_t | I, H^t)] \]
Image generation

\[ \Pi\downarrow t [P(I|H\downarrow t)] \]

Visual memory

Semantic memory

A girl sits
### Summary: Objective Function

- **Language:**

  \[ \pi_t[P(w_t|I,H_t)] \]

- **Image:**

  \[ \pi_t[P(I|H_t)] \]

- **Together:**

  \[ \pi_t[P(I,w_t|H_t)] \]
How does it work?

• Test-bed: Clip-art Dataset
  
  • Jenny is catching the ball;
  • Mike is kicking the ball;
  • The table is next to the tree.

– Sentence (Vertical): concatenation of the 3 simple sentences

– Image (Horizontal): 82 binary variable indicating whether the clip-art/attribute occurs
Image Feature Generation (Baseball)
Real images: plug and play

- **Datasets**
  - PASCAL
  - Flickr 8K/30K
  - MS COCO

- **Image Representation**
  - Whole frame 4096D CNN features

- **Text Preprocessing**
  - Tokenization by Stanford CoreNLP
  - Lowercase all letters
Sentence Generation

• Procedure
  – Sample sentence length, n
  – Using language model, sample 100 sentences of length n
  – Select sentence with best likelihood

Length Distribution on COCO
Qualitative Results

PC: A plate with a bowl of soup and a cup of coffee on a table.

GT: A cup of coffee and a sandwich on a plate.
Qualitative Results

PC: An old brick building with a clock on top of it.

GT: A black and tan clock on a brown brick building.
Qualitative Results

PC: A cat sitting on top of a bed with a laptop.

GT: A dog on top of a person so you cannot see the person.
Qualitative Results

PC: A bunch of stuffed animals sitting next to each other.

GT: An overhead view of a city cross walk on a rainy day features an array of colorful umbrellas.
Qualitative Results

PC: A laptop computer sitting on top of a desk with a laptop.

GT: An open laptop computer sitting on top of a white table.
## Quantitative

<table>
<thead>
<tr>
<th></th>
<th>Flickr 8K</th>
<th></th>
<th>Flickr 30K</th>
<th></th>
<th>MS COCO Val</th>
<th></th>
<th>MS COCO Test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PPL</td>
<td>BLEU</td>
<td>METEOR</td>
<td>PPL</td>
<td>BLEU</td>
<td>METEOR</td>
<td>PPL</td>
<td>BLEU</td>
</tr>
<tr>
<td>RNN</td>
<td>17.5</td>
<td>4.5</td>
<td>10.3</td>
<td>23.0</td>
<td>6.3</td>
<td>10.7</td>
<td>16.9</td>
<td>4.7</td>
</tr>
<tr>
<td>RNN+IF</td>
<td>16.5</td>
<td>11.9</td>
<td>16.2</td>
<td>20.8</td>
<td>11.3</td>
<td>14.3</td>
<td>13.3</td>
<td>16.3</td>
</tr>
<tr>
<td>RNN+IF+FT</td>
<td>16.0</td>
<td>12.0</td>
<td>16.3</td>
<td>20.5</td>
<td>11.6</td>
<td>14.6</td>
<td>12.9</td>
<td>17.0</td>
</tr>
<tr>
<td>RNN+VGG</td>
<td>15.2</td>
<td>12.4</td>
<td>16.7</td>
<td>20.0</td>
<td>11.9</td>
<td>15.0</td>
<td>12.6</td>
<td>18.4</td>
</tr>
<tr>
<td>Our Approach</td>
<td>16.1</td>
<td>12.2</td>
<td>16.6</td>
<td>20.0</td>
<td>11.3</td>
<td>14.6</td>
<td>12.6</td>
<td>16.3</td>
</tr>
<tr>
<td>Our Approach+FT</td>
<td>15.8</td>
<td>12.4</td>
<td>16.7</td>
<td>19.5</td>
<td>11.6</td>
<td>14.7</td>
<td>12.0</td>
<td>16.8</td>
</tr>
<tr>
<td>Our Approach+VGG</td>
<td>15.1</td>
<td>13.1</td>
<td>16.9</td>
<td>19.1</td>
<td>12.0</td>
<td>15.2</td>
<td>11.6</td>
<td>18.8</td>
</tr>
<tr>
<td>Human</td>
<td>-</td>
<td>20.6</td>
<td>25.5</td>
<td>-</td>
<td>18.9</td>
<td>22.9</td>
<td>-</td>
<td>19.2</td>
</tr>
</tbody>
</table>
Conclusions

• CNN makes things work
• Language Models are nice to explore for language generation
• Try to reconstruct the image is helpful
Future Work

• Capture more details (better feature localization)
• Replace RNN with LSTM
THANKS! QUESTIONS?