Multimodal Memory Modelling for Video Captioning

Liang Wang & Yan Huang

Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)

Mar 28, 2018
Artificial Intelligence Laboratory
Researches on artificial intelligence and deep learning
Outline

- Introduction
- Model Description
- Experimental Results
- Conclusion
Outline

- Introduction
- Model Description
- Experimental Results
- Conclusion
Video Captioning

- Generate natural sentences to describe video content
  1. A man and a woman performing a musical.
  2. A teenage couple perform in an amateur musical.
  3. Dancers are playing a routine.
  4. People are dancing in a musical.

- Potential applications

- Challenges
  - Learning an effective mapping from visual sequence space to language space
  - The long-term visual-textual dependency modelling
Related Work

Language template-based approach


Related Work

- Sequence-to-Sequence learning-based approach


Pan et al. Jointly Modeling Embedding and Translation to Bridge Video and Language. CVPR16.
Motivation

- Recent work has pointed out that LSTM doesn’t work well when the sequence is long enough.
- Neural memory models have shown great potential to long-term dependency modelling, e.g., QA in NLP.
- Visual working memory is one of the key factors to guide eye movements.


Recent Related Work

Under review as a conference paper at ICLR 2017

**MEMORY-AUGMENTED ATTENTION MODELLING FOR VIDEOS**

Rasool Fakoor†, Abdel-rahman Mohamed,††, Margaret Mitchell††, Sing Bing Kang††, Pushmeet Kohli††
††Microsoft Research †University of Texas at Arlington ††Google
†rasool.fakoor@mavs.uta.edu, ††{asamir, singbing.kang, pkohli}@microsoft.com ††mmitchellai@google.com,

Recent Related Work

Multimodal Memory Modelling for Video Captioning

Junbo Wang¹  Wei Wang¹  Yan Huang¹  Liang Wang¹,²  Tieniu Tan¹,²
¹Center for Research on Intelligent Perception and Computing
    National Laboratory of Pattern Recognition
²Center for Excellence in Brain Science and Intelligence Technology
    Institute of Automation, Chinese Academy of Sciences
{junbo.wang, wangwei, yhuang, wangliang, tnt}@nlpr.ia.ac.cn

Abstract

Video captioning, which automatically translates video clips into natural language sentences, is a very important task in computer vision. By virtue of recent deep learning technologies, e.g., convolutional neural networks (CNNs) and recurrent neural networks (RNNs), video captioning has made great progress. However, learning an effective mapping from visual sequence space to language space is still a challenging problem. In this paper, we propose a Multimodal Memory Model ($M^3$) to describe videos, which builds a visual and textual shared memory to model the long-term visual-textual dependency and further guide global visual attention on described targets. Specifically, the proposed $M^3$ attaches an external memory to store and
Recent Related Work

Recurrent Memory Addressing for describing videos

Kumar Krishna Agrawal  Arnav Kumar Jain  Abhinav Agarwalla  Pabitra Mitra
Indian Institute of Technology Kharagpur
{kumarkrishna, arnavkj95, abhinavagarwalla, pabitra}@iitkgp.ac.in

Abstract

Deep Neural Network architectures with external memory components allow the model to perform inference and capture long term dependencies, by storing information ex-

KeyAddressing: $W_k h_t^k + W_d h_{t-1}$

Agrawal et al. Recurrent Memory Addressing for describing videos. arxiv16.
Captioning Framework

CNN-Based Video Encoder

2D/3D CNN

2D/3D CNN

2D/3D CNN

2D/3D CNN

Multimodal Memory

$v_1$ $v_2$ $v_3$ ... $v_n$

$\text{Attend}^{t+1}$ $a_1^{t+1}, a_2^{t+1}, ... a_n^{t+1} \sum_{i=1}^{n} a_i^{t+1} v_i$

$\text{Attend}^{t+2}$ $a_1^{t+2}, a_2^{t+2}, ... a_n^{t+2} \sum_{i=1}^{n} a_i^{t+2} v_i$

LSTM-Based Text Decoder

$LSTM^t$ $LSTM^{t+1}$ $LSTM^{t+2}$

#start $A$ man $is$

$\text{read}^{att}$ $\text{write}^{att}$ $\text{read}^{att}$ $\text{write}^{att}$ $\text{read}^{att}$ $\text{write}^{att}$

$\text{read}^{dec}$ $\text{write}^{dec}$ $\text{read}^{dec}$ $\text{write}^{dec}$ $\text{read}^{dec}$ $\text{write}^{dec}$
CNN-Based Video Encoder

C3D

VGG-19

GoogLenet

Inception-3

Residual
Multimodal Memory Modelling

- Multimodal Memory
  - $N \times M$ matrix

1. Writing hidden representations to update memory

$$M_t(i) = M_{t-1}(i) \left[ 1 - w_{t}^{sw}(i) e_{t}^{sw} \right] + w_{t}^{sw}(i) a_{t}^{sw}$$

2. Reading updated memory for temporal attention

$$\sum_{i=1}^{N} w_{t}^{vr}(i) = 1, \quad 0 \leq w_{t}^{vr}(i) \leq 1 \quad r_{t}^{vr} = \sum_{i=1}^{N} w_{t}^{vr}(i) M_{t}(i)$$
Multimodal Memory Modelling

Temporal attention selection for video representation

\[ e_i^t = w^T \tanh (W_r r_t^{vr} + U_\alpha v_i + b_\alpha) \]

\[ \alpha_i^t = \exp \{ e_i^t \} / \sum_{j=1}^{n} \exp \{ e_j^t \} \]

\[ V_t = \sum_{i=1}^{n} \alpha_i^t v_i \]
Multimodal Memory Modelling

Multimodal Memory

- $N \times M$ matrix

③ Writing selected visual information to update memory

$$M_t(i) = M_t(i) \left[1 - w_t^{vw}(i) e_t^{vw}\right] + w_t^{vw}(i) a_t^{vw}$$

④ Reading the updated memory for LSTM-based language model

$$\sum_{i=1}^{N} w_t^{sr}(i) = 1, \ 0 \leq w_t^{sr}(i) \leq 1 \quad r_t^{sr} = \sum_{i=1}^{N} w_t^{sr}(i) M_t(i)$$
LSTM-Based Text Decoder

\[
i_t = \sigma(W_i E_{t-1} + U_i h_{t-1} + M_i r_t + b_i)
\]
\[
f_t = \sigma(W_f E_{t-1} + U_f h_{t-1} + M_f r_t + b_f)
\]
\[
o_t = \sigma(W_o E_{t-1} + U_o h_{t-1} + M_o r_t + b_o)
\]
\[
\tilde{c}_t = \phi(W_c E_{t-1} + U_c h_{t-1} + M_c r_t + b_c)
\]
\[
c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1}
\]
\[
h_t = o_t \odot \phi(c_t)
\]
Memory Addressing & Regularized Loss

**Content-Based Memory Addressing**

\[
K(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\| + \varepsilon} \quad z_t(i) = \beta_t K(k_t, M_t(i))
\]

\[
w_t(i) = \frac{\exp(z_t(i))}{\sum_{j=1}^{N} \exp(z_t(j))}
\]

**Regularized Loss**

\[
L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{t_i} \log \rho(y_j^i | y_1^i : j-1, x^i, \theta) + \lambda \|\theta\|^2
\]
Outline

- Introduction
- Model Description
- Experimental Results
- Conclusion
Implementation Details

- Variable-length sentences
  - A start tag and an end tag
- Beam search
  - beam size: 5
- LSTM-based decoder
  - visual hidden units: 1024, word embedding size: 468
- Memory matrix
  - memory size: (128, 512), GlorotUniform
  - read weight and write weight initial with OneHot
- Others
  - Minibatch: 64, optimization algorithm: ADADELTA
  - Dropout with rate of 0.5, gradient norm clipped (-10, 10)
## Experimental Results

**Microsoft Video Description Dataset**

<table>
<thead>
<tr>
<th>Method</th>
<th>B@1</th>
<th>B@2</th>
<th>B@3</th>
<th>B@4</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGM [30]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>13.68%</td>
<td>23.90%</td>
</tr>
<tr>
<td>LSTM-YT [33]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>33.29%</td>
<td>29.07%</td>
</tr>
<tr>
<td>SA [41]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>40.28%</td>
<td>29.00%</td>
</tr>
<tr>
<td>S2VT [32]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>29.2%</td>
</tr>
<tr>
<td>LSTM-E [23]</td>
<td>74.9%</td>
<td>60.9%</td>
<td>50.6%</td>
<td>40.2%</td>
<td>29.5%</td>
</tr>
<tr>
<td>p-RNN [42]</td>
<td>77.3%</td>
<td>64.5%</td>
<td>54.6%</td>
<td>44.3%</td>
<td>31.1%</td>
</tr>
<tr>
<td>HRNE [22]</td>
<td>79.2%</td>
<td>66.3%</td>
<td>55.1%</td>
<td>43.8%</td>
<td>33.1%</td>
</tr>
<tr>
<td>BGRCN [2]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>49.63%</td>
<td>31.7%</td>
</tr>
<tr>
<td>M³-c3d</td>
<td>77.30%</td>
<td>68.20%</td>
<td>56.30%</td>
<td>45.50%</td>
<td>29.91%</td>
</tr>
<tr>
<td>M³-vgg19</td>
<td>77.70%</td>
<td>67.50%</td>
<td>58.90%</td>
<td>49.60%</td>
<td>30.09%</td>
</tr>
<tr>
<td>M³-google</td>
<td>79.05%</td>
<td>68.74%</td>
<td>60.00%</td>
<td>51.17%</td>
<td>31.47%</td>
</tr>
<tr>
<td>M³-res</td>
<td>80.80%</td>
<td>69.90%</td>
<td>60.40%</td>
<td>49.32%</td>
<td>31.10%</td>
</tr>
<tr>
<td>M³-inv3</td>
<td><strong>81.56%</strong></td>
<td><strong>71.39%</strong></td>
<td><strong>62.34%</strong></td>
<td><strong>52.02%</strong></td>
<td><strong>32.18%</strong></td>
</tr>
<tr>
<td>Method</td>
<td>B@1</td>
<td>B@2</td>
<td>B@3</td>
<td>B@4</td>
<td>METEOR</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>---------</td>
</tr>
<tr>
<td>SA-G-3C [41]</td>
<td></td>
<td></td>
<td></td>
<td>41.92%</td>
<td>29.60%</td>
</tr>
<tr>
<td>S2VT-rgb-flow [32]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>29.8%</td>
</tr>
<tr>
<td>LSTM-E-VC [23]</td>
<td>78.8%</td>
<td>66.0%</td>
<td>55.4%</td>
<td>45.3%</td>
<td>31.0%</td>
</tr>
<tr>
<td>p-RNN-VC [42]</td>
<td>81.5%</td>
<td>70.4%</td>
<td>60.4%</td>
<td>49.9%</td>
<td>32.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>B@1</th>
<th>B@2</th>
<th>B@3</th>
<th>B@4</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>M^{3}-VC</td>
<td>81.90%</td>
<td>71.26%</td>
<td>62.08%</td>
<td>51.78%</td>
<td>32.49%</td>
</tr>
<tr>
<td>M^{3}-IC</td>
<td><strong>82.45%</strong></td>
<td><strong>72.43%</strong></td>
<td><strong>62.78%</strong></td>
<td><strong>52.82%</strong></td>
<td><strong>33.31%</strong></td>
</tr>
</tbody>
</table>

**CVPR2017**

Table 1. Experiment results on the Youtube2Text Dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>Feat.</th>
<th>Fine</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-Encoder</td>
<td>0.404</td>
<td>0.295</td>
<td>0.515</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>GoogLeNet-Encoder</td>
<td>0.427</td>
<td>0.303</td>
<td>0.534</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>t-KeyAddressing (G)</td>
<td>0.436</td>
<td>0.308</td>
<td>0.545</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>m-KeyAddressing (Memory LSTM) (G)</td>
<td><strong>0.457</strong></td>
<td><strong>0.319</strong></td>
<td><strong>0.573</strong></td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

**ICLR2017**

Table 2: Video captioning evaluation on the test set of 670 videos in MSVD.

<table>
<thead>
<tr>
<th>Method</th>
<th>METEOR</th>
<th>BLEU@1</th>
<th>BLEU@2</th>
<th>BLEU@3</th>
<th>BLEU@4</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venugopalan et al. (2015b)</td>
<td>27.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venugopalan et al. (2015a)</td>
<td>29.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pan et al. (2016b)</td>
<td>29.5</td>
<td>74.9</td>
<td>60.9</td>
<td>50.6</td>
<td>40.2</td>
<td></td>
</tr>
<tr>
<td>Yu et al. (2016)</td>
<td>31.10</td>
<td>77.30</td>
<td>64.50</td>
<td>54.60</td>
<td>44.30</td>
<td></td>
</tr>
<tr>
<td>Pan et al. (2016a)</td>
<td><strong>33.10</strong></td>
<td>79.20</td>
<td>66.30</td>
<td>55.10</td>
<td>43.80</td>
<td></td>
</tr>
<tr>
<td><strong>Our Model</strong></td>
<td><strong>31.80</strong></td>
<td><strong>79.40</strong></td>
<td><strong>67.10</strong></td>
<td><strong>56.80</strong></td>
<td><strong>46.10</strong></td>
<td><strong>62.70</strong></td>
</tr>
</tbody>
</table>
Experimental Results

- Microsoft Research-Video to Text Dataset
  - the largest dataset in terms of sentence and vocabulary, 10,000 video clips and 200,000 sentences
  - each video is labelled with about 20 sentences
  - training set (6513), validation set (497), and test set (2990)

<table>
<thead>
<tr>
<th>Method</th>
<th>B @ 1</th>
<th>B @ 2</th>
<th>B @ 3</th>
<th>B @ 4</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA-V [41]</td>
<td>67.82%</td>
<td>55.41%</td>
<td>42.90%</td>
<td>34.73%</td>
<td>23.11%</td>
</tr>
<tr>
<td>SA-C [41]</td>
<td>68.90%</td>
<td>57.50%</td>
<td>47.00%</td>
<td>37.40%</td>
<td>24.80%</td>
</tr>
<tr>
<td>SA-VC [41]</td>
<td>72.20%</td>
<td>58.90%</td>
<td>46.80%</td>
<td>35.90%</td>
<td>24.90%</td>
</tr>
<tr>
<td>M³-V</td>
<td>70.20%</td>
<td>56.60%</td>
<td>44.80%</td>
<td>35.00%</td>
<td>24.60%</td>
</tr>
<tr>
<td>M³-C</td>
<td>77.20%</td>
<td>61.30%</td>
<td>47.20%</td>
<td>35.10%</td>
<td>25.70%</td>
</tr>
<tr>
<td>M³-VC</td>
<td>73.60%</td>
<td>59.30%</td>
<td>48.26%</td>
<td>38.13%</td>
<td>26.58%</td>
</tr>
</tbody>
</table>
## Experimental Results

### Description Generation

<table>
<thead>
<tr>
<th>SA: Yao et al. ICCV 2015</th>
<th>M³: Our model</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td><strong>Generated Sentence:</strong></td>
<td><strong>Reference Sentence:</strong></td>
</tr>
<tr>
<td>SA: a man is riding a bike</td>
<td>1. a man is riding a motorcycle</td>
</tr>
<tr>
<td>M³: a man is riding a motorcycle</td>
<td>2. a man rides a motorcycle on a beach</td>
</tr>
<tr>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td><strong>Generated Sentence:</strong></td>
<td><strong>Reference Sentence:</strong></td>
</tr>
<tr>
<td>SA: a man is playing a guitar</td>
<td>1. a man is petting two dogs</td>
</tr>
<tr>
<td>M³: a man is playing with a dog</td>
<td>2. a man pets some dogs</td>
</tr>
<tr>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td><strong>Generated Sentence:</strong></td>
<td><strong>Reference Sentence:</strong></td>
</tr>
<tr>
<td>SA: men are playing soccer ball</td>
<td>1. a basketball game is in play</td>
</tr>
<tr>
<td>M³: people are playing basketball</td>
<td>2. two teams playing basketball</td>
</tr>
<tr>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td><strong>Generated Sentence:</strong></td>
<td><strong>Reference Sentence:</strong></td>
</tr>
<tr>
<td>SA: a man is riding a horse</td>
<td>1. a woman is riding her horse</td>
</tr>
<tr>
<td>M³: a woman is riding a horse</td>
<td>2. a woman riding in horse competition</td>
</tr>
<tr>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td><strong>Generated Sentence:</strong></td>
<td><strong>Reference Sentence:</strong></td>
</tr>
<tr>
<td>SA: a woman is peeling a fish</td>
<td>1. a woman is peeling shrimp</td>
</tr>
<tr>
<td>M³: a woman is peeling a shrimp</td>
<td>2. the lady peeled the shrimp</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Model Description
- Experimental Results
- Conclusion
Conclusion & Future Work

Textual/Visual/Attribute Memory

Working Memory, Baddeley et al.
Acknowledgement

NVAIL

Artificial Intelligence Laboratory

Sponsor excellent hardware resources
Thank you!

(Q/A)

wangliang@nlpr.ia.ac.cn