Improving Image and Sentence Matching with Multimodal Attention and Visual Attributes

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. 26, 2018
CRIPAC mainly focuses on the following research topics related to national public security.

- Biometrics
- Image and Video Analysis
- Big Data and Multi-modal Computing
- Content Security and Authentication
- Sensing and Information Acquisition

CRIPAC receives regular fundings from various Government departments or agencies. It is also supported by funds of R&D projects from many other national and international sources.

CRIPAC members publish widely in leading national and international journals and conferences such as IEEE Transactions on PAMI, IEEE Transactions on Image Processing, International Journal of Computer Vision, Pattern Recognition, Pattern Recognition Letters, ICCV, ECCV, CVPR, ACCV, ICPR, ICIP, etc.

http://cripac.ia.ac.cn/en/EN/volumn/home.shtml
Artificial Intelligence Laboratory

Researches on artificial intelligence and deep learning
Outline

1. Image and Sentence Matching
2. Related Work
3. Improved Image and Sentence Matching
   3.1 Context-modulated Multimodal Attention
   3.2 Joint Semantic Concepts and Order Learning
4. Future Directions
Outline

1. Image and Sentence Matching
2. Related Work
3. Improved Image and Sentence Matching
   3.1 Context-modulated Multimodal Attention
   3.2 Joint Semantic Concepts and Order Learning
4. Future Directions
Image and Sentence Matching

**Image-sentence retrieval**

- Until April, the Polish forces had been slowly but steadily advancing eastward.
- In southeastern Washington, a stretch of the river passes through the Hanford Site, established in 1943 as part of the Manhattan Project.
- Tropical Storm Edouard was the first of eight named storms to form in September 2002, the most such storms for any month in the Atlantic.

**Image caption**

- There are many kinds of vegetables.

**Image question answering**

- **GT Question**: What is the boy in green cap doing?
- **GT Answer**: He is playing skateboard.

The key challenge lies in how to well measure the cross-modal similarity.
Related Work


Mao et al., Deep Captioning with Multimodal Recurrent Neural Networks, ICLR, 2015.

Ma et al., Multimodal Convolutional Neural Networks for Matching Image and Sentence, ICCV, 2015.

Related Work

There are many kinds of vegetables

- Deep visual semantic embedding
  - Devise [1]
  - Order embedding [2]
  - Structure-preserving embedding [3]

- Deep canonical correlation analysis
  - Batch based learning [4]
  - Fisher vector on w2v [5]
  - Global + local correspondences [6]

Sentence only describes **partial salient** image content

Using **Global** image features might be inappropriate

---

Outline

1 Image and Sentence Matching
2 Related Work
3 Improved Image and Sentence Matching
   3.1 Context-modulated Multimodal Attention
   3.2 Joint Semantic Concepts and Order Learning
4 Future Directions
Motivation

There are many kinds of vegetables.

Association analysis

1. Image and sentence include much **redundant information**
2. Only **partial semantic instances** can be well associated
Instance-aware Image and Sentence Matching

Selectively attend to image-sentence instances (marked by colored boxes). Sequentially measure local similarities of pairwise instances, and fuse all the similarities to obtain the matching score.

The details at the $t$-th time step.
Details of LSTM at the $t$-th Timestep

- **Saliency probability of instance candidate:**

\[
\begin{align*}
    p_{t,i} &= \frac{e^{\hat{p}_{t,i}}}{\sum_{i=1}^{I} e^{\hat{p}_{t,i}}}, \quad \hat{p}_{t,i} = f_p(m, a_i, h_{t-1}), \\
    q_{t,j} &= \frac{e^{\hat{q}_{t,j}}}{\sum_{j=1}^{J} e^{\hat{q}_{t,j}}}, \quad \hat{q}_{t,j} = f_q(n, w_j, h_{t-1})
\end{align*}
\]

\[
    f_p(m, a_i, h_{t-1}) = w_p(\sigma(m W_m + b_m) + \sigma(a_i W_a + b_a) + \sigma(h_{t-1} W_h + b_h)) + b_p
\]

- **Instance representation:**

\[
    a'_t = \sum_{i=1}^{I} p_{t,i} a_i, \quad w'_t = \sum_{j=1}^{J} q_{t,j} w_j
\]
Local Similarity Measurement and Aggregation

Image-sentence instance representation:

\[ a'_t = \sum_{i=1}^{I} p_{t,i} a_i, \quad w'_t = \sum_{j=1}^{J} q_{t,j} w_j \]

- Measure their local similarity with a two-way MLP
- Feed into the current hidden state

Global similarity:

\[ s = w_{hs} (\sigma (W_{hh} h_t + b_h)) + b_s \]

- Aggregate all the similarities
- Measure local similarities at all timesteps

Detailed formulation of LSTM at the \( t \)-th timestep

\[
\begin{align*}
i_t &= \sigma (W_{si} s_t + W_{hi} h_{t-1} + b_i), \\
f_t &= \sigma (W_{sf} s_t + W_{hf} h_{t-1} + b_f), \\
c_t &= f_t \odot c_{t-1} + i_t \odot \text{tanh}(W_{se} s_t + W_{hc} h_{t-1} + b_c), \\
o_t &= \sigma (W_{so} s_t + W_{ho} h_{t-1} + b_o), \\
h_t &= o_t \odot \text{tanh}(c_t)
\end{align*}
\]
• **Structured objective function**
  – matched scores are larger than mismatched ones
    \[ \sum_{i,k} \max\{0, m - s_{ii} + s_{ik}\} + \max\{0, m - s_{ii} + s_{ki}\} \]

• **Pairwise doubly stochastic regularization**
  – constrain the sum of saliency values of an instance candidate at all timesteps to be 1
  – encourages the model to **pay equal attention to every instance** rather than a certain one
    \[ \lambda \left( \sum_i (1 - \sum_t p_{t,i}) + \sum_j (1 - \sum_t q_{t,j}) \right) \]

• **Optimize the objective using stochastic gradient descent**
Experimental Datasets

• Flickr 30k dataset
  - from the Flickr.com website
  - 31784 images
  - each image has 5 captions
  - use the public training, validation and testing splits, which contain 28000, 1000 and 1000 images, respectively

• Microsoft COCO dataset
  - 82783 images
  - each image has 5 captions
  - use the public training, validation and testing splits, with 82783, 4000 and 1000 images, respectively

1. A man in street racer armor is examining the tire of another racers motor bike.
2. The two racers drove the white bike down the road.
3. ......
Implementation Details

- **Evaluation criterions**
  - “R@1”, “R@5” and “R@10”, i.e., recall rates at the top 1, 5 and 10 results
  - “Med r” is the median rank of the first ground truth result
  - “Sum”:

\[
\text{Sum} = \underbrace{R@1 + R@5 + R@10 + R@1 + R@5 + R@10}_\text{Image annotation} \quad \underbrace{R@1 + R@5 + R@10 + R@1 + R@5 + R@10}_\text{Image retrieval}
\]

- **Feature extraction**

<table>
<thead>
<tr>
<th></th>
<th>Image</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global context</strong></td>
<td>the feature vector in “fc7” layer of the 19-layer VGG network</td>
<td>the last hidden state of a visual-semantic embedding framework</td>
</tr>
<tr>
<td><strong>Local representation</strong></td>
<td>512 feature maps (size: 14x14) in “conv5-4” layer</td>
<td>multiple hidden states of a bidirectional LSTM</td>
</tr>
</tbody>
</table>
Implementation Details

- Five variants of the proposed sm-LSTMs
  - **mean vector**: use mean instead of weighted sum vector
  - **attention**: use conventional attention scheme
  - **context**: use global context modulation
  - **ensemble**: sum multiple cross-modal similarity matrices

<table>
<thead>
<tr>
<th></th>
<th>Mean vector</th>
<th>Attention</th>
<th>Context</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>sm-LSTM-mean</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sm-LSTM-att</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sm-LSTM-ctx</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>sm-LSTM</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sm-LSTM*</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
### Table 1. Bidirectional image and sentence retrieval results on Flickr30k.

<table>
<thead>
<tr>
<th>Method</th>
<th>Image Annotation</th>
<th>Image Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_{@1}$</td>
<td>$R_{@3}$</td>
</tr>
<tr>
<td>RVP (T+) [4]</td>
<td>12.1</td>
<td>27.8</td>
</tr>
<tr>
<td>Deep Fragment [13]</td>
<td>14.2</td>
<td>37.7</td>
</tr>
<tr>
<td>DCCA [34]</td>
<td>16.7</td>
<td>39.3</td>
</tr>
<tr>
<td>NICE [31]</td>
<td>17.0</td>
<td>-</td>
</tr>
<tr>
<td>DVSA (BRNN) [14]</td>
<td>22.2</td>
<td>48.2</td>
</tr>
<tr>
<td>MNLM [15]</td>
<td>23.0</td>
<td>50.7</td>
</tr>
<tr>
<td>LRCN [7]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>m-RNN [22]</td>
<td>35.4</td>
<td>63.8</td>
</tr>
<tr>
<td>FV* [17]</td>
<td>35.0</td>
<td>62.0</td>
</tr>
<tr>
<td>m-CNN* [21]</td>
<td>33.6</td>
<td>64.1</td>
</tr>
<tr>
<td>RTP+FV* [26]</td>
<td>37.4</td>
<td>63.1</td>
</tr>
<tr>
<td>RNN+FV* [19]</td>
<td>34.7</td>
<td>62.7</td>
</tr>
<tr>
<td>DSPE+FV* [32]</td>
<td>40.3</td>
<td>68.9</td>
</tr>
</tbody>
</table>

**Ours:**
- sm-LSTM-mean: 25.9
- sm-LSTM-att: 27.0
- sm-LSTM-ctx: 33.5
- sm-LSTM: 42.4
- sm-LSTM*: 42.5

### Table 2. Bidirectional image and sentence retrieval results on COCO.

<table>
<thead>
<tr>
<th>Method</th>
<th>Image Annotation</th>
<th>Image Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_{@1}$</td>
<td>$R_{@3}$</td>
</tr>
<tr>
<td>STD* [16]</td>
<td>33.8</td>
<td>67.7</td>
</tr>
<tr>
<td>m-RNN [22]</td>
<td>41.0</td>
<td>73.0</td>
</tr>
<tr>
<td>FV* [17]</td>
<td>39.4</td>
<td>67.9</td>
</tr>
<tr>
<td>DVSA [14]</td>
<td>38.4</td>
<td>69.9</td>
</tr>
<tr>
<td>MNLM [15]</td>
<td>43.4</td>
<td>75.7</td>
</tr>
<tr>
<td>m-CNN* [21]</td>
<td>42.8</td>
<td>73.1</td>
</tr>
<tr>
<td>RNN+FV* [19]</td>
<td>40.8</td>
<td>71.9</td>
</tr>
<tr>
<td>OME [30]</td>
<td>46.7</td>
<td>-</td>
</tr>
<tr>
<td>DSPE+FV* [32]</td>
<td>50.1</td>
<td>79.7</td>
</tr>
</tbody>
</table>

**Ours:**
- sm-LSTM-mean: 33.1
- sm-LSTM-att: 36.7
- sm-LSTM-ctx: 39.7
- sm-LSTM: 52.4
- sm-LSTM*: 53.2

---

[15] Kiros et al., Unifying visual-semantic embeddings with multimodal neural language models. TACL, 2015.
Analysis on Hyperparameters

Table 3. The impact of different numbers of timesteps on the Flick30k dataset.

<table>
<thead>
<tr>
<th>$T = 1$</th>
<th>$T = 2$</th>
<th>$T = 3$</th>
<th>$T = 4$</th>
<th>$T = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Annotation</td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@1</td>
</tr>
<tr>
<td>$T = 1$</td>
<td>38.8</td>
<td>65.7</td>
<td>76.8</td>
<td>28.0</td>
</tr>
<tr>
<td>$T = 2$</td>
<td>38.0</td>
<td>68.9</td>
<td>77.9</td>
<td>28.1</td>
</tr>
<tr>
<td>$T = 3$</td>
<td><strong>42.4</strong></td>
<td><strong>67.5</strong></td>
<td><strong>79.9</strong></td>
<td><strong>28.2</strong></td>
</tr>
<tr>
<td>$T = 4$</td>
<td>38.2</td>
<td>67.6</td>
<td>78.5</td>
<td>27.5</td>
</tr>
<tr>
<td>$T = 5$</td>
<td>38.1</td>
<td>68.2</td>
<td>78.4</td>
<td>28.1</td>
</tr>
</tbody>
</table>

$T$: the number of timesteps in the sm-LSTM.

Table 4. The impact of different values of the balancing parameter on the Flick30k dataset.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Image Annotation</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 0$</td>
<td>37.9</td>
<td>65.8</td>
<td>77.7</td>
<td>27.2</td>
<td>55.4</td>
<td>67.6</td>
<td>27.2</td>
<td>55.4</td>
<td>67.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 1$</td>
<td>38.0</td>
<td>66.2</td>
<td>77.8</td>
<td>27.4</td>
<td>55.6</td>
<td>67.7</td>
<td>27.4</td>
<td>55.6</td>
<td>67.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 10$</td>
<td>38.4</td>
<td>67.4</td>
<td>77.7</td>
<td>27.5</td>
<td>56.1</td>
<td>67.6</td>
<td>27.5</td>
<td>56.1</td>
<td>67.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 100$</td>
<td><strong>42.4</strong></td>
<td><strong>67.5</strong></td>
<td><strong>79.9</strong></td>
<td><strong>28.2</strong></td>
<td><strong>57.0</strong></td>
<td><strong>68.4</strong></td>
<td><strong>28.2</strong></td>
<td><strong>57.0</strong></td>
<td><strong>68.4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 1000$</td>
<td>40.2</td>
<td>67.1</td>
<td>78.6</td>
<td>27.8</td>
<td>56.9</td>
<td>67.9</td>
<td>27.8</td>
<td>56.9</td>
<td>67.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\lambda$: the balancing parameter between the structured objective and regularization.
Table 5. **Attended image instances** at three different timesteps.
Instance-aware Saliency Maps

Figure 2. Visualization of **attended image and sentence instances** at three different timesteps.

- **A airplane that is flying in the sky.**
  - sky, flying
  - airplane, flying
  - airplane, sky

- **A stop sign and a no access sign in the grass.**
  - sign, stop
  - grass, access
  - grass, sign

- **A man and his son in the grass flying a kite.**
  - men, kite
  - grass, man
  - kite, flying

- **People are getting onto a city bus.**
  - people, getting
  - city, bus

- **A group of people are observing planes at an air show.**
  - group, people
  - planes, air

- **Children in a park two are sitting on a play giraffe.**
  - giraffe, play
  - children, play
  - park, children
Conclusion

- **selectively process redundant information** with context-modulated attention
- **gradually accumulate salient information** with multi-modal LSTM-RNN

For more details, please refer to the following paper:

Outline

1. Image and Sentence Matching
2. Related Work
3. Improved Image and Sentence Matching
   3.1 Context-modulated Multimodal Attention
   3.2 Joint Semantic Concepts and Order Learning
4. Future Directions
Only Instances Are Not Enough

- **Semantic concepts:**
  - **Objects:**
    - cheetah
    - gazelle
    - grass
  - **Properties:**
    - quick
    - young
    - green
  - **Actions:**
    - chasing
    - running

- **Semantic order:**
  cheetah chasing gazelle grass ...

- **Matched sentence:**
  A quick cheetah is chasing a young gazelle on grass.

- **“Semantic concepts”** include separated instances and their descriptive properties
- Different **“semantic orders”** lead to various semantic meanings
Joint Semantic Concept and Order Learning

- **Improve image representations** by learning semantic concepts and then organizing them in the correct semantic order

- Multi-regional multilabel CNN for **concept prediction**
- Global context modulation & sentence generation for **order learning**
Semantic Concept Prediction

- **Process** the existing dataset, **select** the desired concept, and **reduce** the size of vocabulary
- **Learn** a multi-label CNN and perform **testing** in a multi-region way

A couple of giraffes eating out of basket

\[\downarrow\]

A, couple, of, giraffes, eating, out, of, basket

\[\downarrow\]

A, couple, of, giraffes, eating, out, of, basket

\[\Leftrightarrow\]

giraffe, eating, basket

Multi-regional multi-label CNN [1]

Use Global Context as Reference

- Directly learn the semantic order is very difficult!

➢ The global context indicates **the spatial relation** of concepts
➢ Selectively **balance the importance** of concept and context

Groundtruth sentence:
*a couple of giraffes eating out of a basket*

Relations among concepts

Separated semantic concepts

Global context

Local regions
Use Sentence Generation as Supervision

- A straightforward approach is to directly generate a sentence based on the predicted concepts, but not working

➢ Regard the fused context and concepts as image representation
➢ use the groundtruth semantic order to supervise it during sentence generation

\[ L = L_{mat} + \lambda \times L_{gen} \]

\[ \sum_{ik} \max \{0, m - s_{ii} + s_{ik}\} + \max \{0, m - s_{ii} + s_{ki}\} - \sum_{t} \log P(w_t|w_{t-1}, w_{t-2}, \ldots, w_0, x, p) \]
Use Sentence Generation as Supervision

➢ Regard the **fused context and concepts** as image representation
➢ Use the **groundtruth semantic order** to supervise it during sentence generation

\[
L = L_{mat} + \lambda \times L_{gen}
\]

\[
\sum_{i,k} \max \{0, m - s_{ii} + s_{ik}\} + \max \{0, m - s_{ii} + s_{ki}\} - \sum_t \log P(w_t|w_{t-1}, w_{t-2}, \cdots, w_0, x, p)
\]
Table 6. The experimental settings of ablation models.

<table>
<thead>
<tr>
<th></th>
<th>1-crop</th>
<th>10-crop</th>
<th>context</th>
<th>concept</th>
<th>sum</th>
<th>gate</th>
<th>sentence</th>
<th>generation</th>
<th>sampling</th>
<th>shared</th>
<th>non-shared</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctx (1-crop)</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ctx</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ctx + sen</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>ctx + gen (S)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ctx + gen (E)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ctx + gen</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cnp</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>cnp + gen</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>cnp + ctx (C)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>cnp + ctx</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cnp + ctx + gen</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **“10-crop”**: crop 10 regions from images
- **“concept”** and **“context”**: use semantic concepts and context
- **“sum”** and **“gate”**: combine semantic concepts and context via feature concatenation and gated unit, respectively
- **“sentence”, “generation” and “sampling”**: image captioning, sentence generation and scheduled sampling
- **“share”** and **“non-shared”**: two word embedding matrices
Evaluation of Ablation Models

Table 7. Comparison results by **ablation models** on the Flickr30k and MSCOCO datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Flickr30k dataset</th>
<th>MSCOCO dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image Annotation</td>
<td>Image Retrieval</td>
</tr>
<tr>
<td></td>
<td>R@1  R@5 R@10</td>
<td>R@1  R@5 R@10</td>
</tr>
<tr>
<td>ctx (1-crop)</td>
<td>29.8 58.4 70.5</td>
<td>22.0 47.9 59.3</td>
</tr>
<tr>
<td>ctx</td>
<td>33.8 63.7 75.9</td>
<td>26.3 55.4 67.6</td>
</tr>
<tr>
<td>ctx + sen</td>
<td>22.8 48.6 60.8</td>
<td>19.1 46.0 59.7</td>
</tr>
<tr>
<td>ctx + gen (S)</td>
<td>34.4 64.5 77.0</td>
<td>27.1 56.3 68.3</td>
</tr>
<tr>
<td>ctx + gen (E)</td>
<td>35.5 63.7 75.9</td>
<td>27.4 55.9 67.5</td>
</tr>
<tr>
<td>ctx + gen</td>
<td>35.6 66.3 76.9</td>
<td>27.9 56.8 68.2</td>
</tr>
<tr>
<td>cnp</td>
<td>30.9 60.9 72.4</td>
<td>23.1 52.5 64.8</td>
</tr>
<tr>
<td>cnp + gen</td>
<td>31.5 61.7 74.5</td>
<td>25.0 53.4 64.9</td>
</tr>
<tr>
<td>cnp + ctx (C)</td>
<td>39.9 71.2 81.3</td>
<td>31.4 61.7 72.8</td>
</tr>
<tr>
<td>cnp + ctx</td>
<td>42.4 72.9 81.5</td>
<td>32.4 63.5 73.9</td>
</tr>
<tr>
<td><strong>cnp + ctx + gen</strong></td>
<td><strong>44.2 74.1 83.6</strong></td>
<td><strong>32.8 64.3 74.9</strong></td>
</tr>
</tbody>
</table>

Table 8. Comparison results of **balancing parameter** $\lambda$ on the Microsoft COCO dataset.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Flickr30k dataset</th>
<th>MSCOCO dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image Annotation</td>
<td>Image Retrieval</td>
</tr>
<tr>
<td></td>
<td>R@1  R@5 R@10</td>
<td>R@1  R@5 R@10</td>
</tr>
<tr>
<td>0</td>
<td>42.4 72.9 81.5</td>
<td>32.4 63.5 73.9</td>
</tr>
<tr>
<td>0.01</td>
<td>43.1 72.8 83.5</td>
<td>32.8 63.2 73.6</td>
</tr>
<tr>
<td>1</td>
<td><strong>44.2 74.1 83.6</strong></td>
<td><strong>32.8 64.3 74.9</strong></td>
</tr>
<tr>
<td>100</td>
<td>42.3 73.8 83.1</td>
<td>32.5 63.3 74.0</td>
</tr>
</tbody>
</table>
Comparison with State-of-the-art Methods

Table 9. Bidirectional image and sentence retrieval results on two datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>VGGNet</th>
<th>ResNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Flickr30k dataset</td>
<td>MSCOCO dataset</td>
</tr>
<tr>
<td></td>
<td>Image Annotation</td>
<td>Image Retrieval</td>
</tr>
<tr>
<td>VGGNet</td>
<td>Image Annotation</td>
<td>Image Retrieval</td>
</tr>
<tr>
<td>VGGNet</td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>ResNet</td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>VGGNet</td>
<td>35.4</td>
<td>63.8</td>
</tr>
<tr>
<td>ResNet</td>
<td>35.0</td>
<td>62.0</td>
</tr>
<tr>
<td>VGGNet</td>
<td>22.2</td>
<td>48.2</td>
</tr>
<tr>
<td>ResNet</td>
<td>23.0</td>
<td>50.7</td>
</tr>
<tr>
<td>m-CNN [20]</td>
<td>33.6</td>
<td>64.1</td>
</tr>
<tr>
<td>ResNet</td>
<td>34.7</td>
<td>62.7</td>
</tr>
<tr>
<td>VQA [18]</td>
<td>33.9</td>
<td>62.5</td>
</tr>
<tr>
<td>ResNet</td>
<td>37.4</td>
<td>63.1</td>
</tr>
<tr>
<td>RTP [26]</td>
<td>40.3</td>
<td>68.9</td>
</tr>
<tr>
<td>ResNet</td>
<td>42.5</td>
<td>71.9</td>
</tr>
<tr>
<td>2WayNet [4]</td>
<td>49.8</td>
<td>67.5</td>
</tr>
<tr>
<td>ResNet</td>
<td>41.4</td>
<td>73.5</td>
</tr>
<tr>
<td>DAN [23]</td>
<td>41.3</td>
<td>69.0</td>
</tr>
<tr>
<td>ResNet</td>
<td>47.6</td>
<td>77.4</td>
</tr>
<tr>
<td>VSE++ [5]</td>
<td>55.0</td>
<td>81.8</td>
</tr>
<tr>
<td>ResNet</td>
<td>52.9</td>
<td>79.1</td>
</tr>
<tr>
<td>Ours</td>
<td>44.2</td>
<td>74.1</td>
</tr>
<tr>
<td>Ours (Res)</td>
<td>55.5</td>
<td>82.0</td>
</tr>
</tbody>
</table>
Table 10. Results of image annotation by 3 ablation models.

<table>
<thead>
<tr>
<th>Query</th>
<th>Retrieved top-4 relevant sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>ctx</strong></td>
</tr>
</tbody>
</table>
| ![Image 1](image1.png) | 1. a dinner table with various plates of food and a glass of water on the table  
2. a table top with some plates of food on it  
3. a table set for three with food and wine  
4. a dinner table with three plates of gourmet hamburgers | 1. a meal is being displayed on a table  
2. a table with bowls of grains and fruit and a hand with a plate  
3. a table top with some plates of food on it  
4. a person holding a bowl of oats next to bowls of other condiments | 1. a person holding a bowl of oats next to bowls of other condiments  
2. a meal is being displayed on a table  
3. a table with bowls of grains and fruit and a hand with a plate  
4. a table that has some food on it |
| ![Image 2](image2.png) | 1. a man riding a skateboard up the side of a ramp  
2. a man riding a skateboard up the side of a ramp  
3. a man at a skate park with his foot on the side of the skateboard  
4. a man on a skateboard performing a trick | 1. its a cloudy night for a ride on the motorcycle  
2. a motorcyclist surveys the sunlit road into the horizon  
3. a close up of a person riding a motorcycle on a long empty road  
4. a photo taken from a car looking at a skateboarder on the side of the road  | 1. a motorcyclist surveys the sunlit road into the horizon  
2. a close up of a person riding a motorcycle on a long empty road  
3. a photo taken from a car looking at a skateboarder on the side of the road  
4. its a cloudy night for a ride on the motorcycle |
| ![Image 3](image3.png) | 1. a couple of giraffes look around the ground in the zoo  
2. two giraffe standing near brick building  
3. a pair of giraffes standing around in their enclosure  
4. two giraffes roaming around an enclosed area on a sunny day | 1. a pair of giraffes standing around in their enclosure  
2. a couple of giraffes eating hay from a trough  
3. two giraffes that are eating from a basket  
4. two giraffes stand and eat food out of a basket | 1. a couple of giraffes eating hay from a trough  
2. a couple of giraffes eating out of a basket  
3. two giraffes stand and eat food out of a basket  
4. a couple of giraffes reach for a basket |

Note: groundtruth matched sentences are marked as **red**, while some sentences sharing similar meanings as groundtruths are marked as **underline**.
Conclusion

- **multi-regional multi-label CNN** for semantic concept prediction
- **gated fusion unit, and joint matching-generation learning** for semantic order learning

For more details, please refer to the following paper:

Outline

1 Image and Sentence Matching
2 Related Work
3 Improved Image and Sentence Matching
   3.1 Context-modulated Multimodal Attention
   3.2 Joint Semantic Concepts and Order Learning
4 Future Directions
• Understand vision jointly with language and speech in a unified framework
Acknowledgement

NVAIL
Artificial Intelligence Laboratory
Sponsor excellent hardware resources
THANK YOU

yhuang@nlpr.ia.ac.cn