Fluency-guided Cross-lingual Image Captioning

Weiyu Lan¹, Xirong Li¹, Jianfeng Dong²

¹ Renmin University of China
² Zhejiang University
Image Captioning

A person holding a book with a bird sitting on the book.

Una persona sostiene un libro con un pájaro que se sienta sobre el libro

一个人拿着一本书，有一只小鸟站在上面

책에 앉아 있는 새와 함께 책을 들고 있는 사람
Cross lingual image captioning

- Goal: To generate **relevant** and **fluent** captions in a target language with minimal human effort
Related Work

• Monolingual image captioning
  • Deep learning: Encoder + Decoder (CNN+RNN)

• Cross-lingual image captioning
  • Generate captions base on both image and captions in source language (Elliott et al., 2015)
  • Crowd sourcing to collect Japanese descriptions of the MSCOCO (Miyazaki and Shimizu, 2016)
  • Machine Translation (Li, ICMR2016)
Related Work

• Cross-lingual image captioning
  • Machine Translation (Li, ICMR2016)

A person holding a book with a bird sitting on the book

Captioning Model

Machine translation

Chinese Captioning Model

translated Chinese sentences

一个人拿着一本书，有一只小鸟站在上面

√

×
Machine-translated sentences are not fluent

A person holding a book with a bird sitting on the book.

拿着一本书和一只鸟坐在书上的人

Una persona que sostiene un libro con un pájaro sentado en el libro.

책 한 권을 쥔고 한 사람 한 마리가 책 위에 앉아 있다.
Our Approach

• **Fluency-guided** cross-lingual image captioning

![Image of a person holding a book with a bird sitting on it] (A person holding a book with a bird sitting on the book)

拿着一本书和一只鸟坐在书上的人

![Sentence](A small bird sitting on top of an open book)  

(A small bird sitting on top of an open book)  

一只小鸟坐在一本打开的书上

Captioning Model

Sentence

Image

Estimated Fluency

Not Fluent

Fluent
Sentence Fluency Estimation

- Binary classification
- Manual annotation
  - 8k sentences: fluent/not fluent
  - Less than 30% sentences are fluent
- LSTM based model
A four-way LSTM based classifier

A person holding a book with a bird sitting …

Fluency Estimation Model

1/4 \sum \text{Fluency Score}
# Sentence Fluency Estimation Results

<table>
<thead>
<tr>
<th>English sentences</th>
<th>Chinese sentences</th>
<th>Estimated fluency scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>The two large elephants are standing in the grass</td>
<td>两只大象正站在草地上</td>
<td>0.803</td>
</tr>
<tr>
<td>The young man in the blue shirt is playing tennis</td>
<td>穿蓝色衬衫的年轻人正在打网球</td>
<td>0.624</td>
</tr>
<tr>
<td>A group of people riding skis in their bathing suits</td>
<td>一群人在他们的沐浴骑滑雪服</td>
<td>0.117</td>
</tr>
<tr>
<td>A sports arena under a dome with snow on it</td>
<td>一个体育馆下一个圆顶下的雪在它</td>
<td>0.060</td>
</tr>
</tbody>
</table>
Image Captioning Model

- CNN + RNN framework [Vinyals, CVPR2015]
- Training loss is the sum of the negative log likelihoods of the next correct word at each step.
Fluency-Guided Training
Fluency-Guided Training
Strategy I: Fluency only

(A small bird sitting on top of an open book)
一只小鸟坐在一本打开的书上

(A person holding a book with a bird sitting on the book)
拿着一本书和一只鸟坐在书上的人

Captioning Model

0.9 Fluent

0.2 Not Fluent
Fluency-Guided Training
Strategy II: Rejection sampling

- Allow the sentences classified as not fluent to be used for training with a certain chance

(A small bird sitting on top of an open book)
一只小鸟坐在一本打开的书上

(A person holding a book with a bird sitting on the book)
拿着一本书和一只鸟坐在书上的人

0.9 Fluent

0.2 Not Fluent

Captioning Model

u ~ U (0, 0.5)
Fluency-Guided Training
Strategy III: Weighted loss

• Cost-sensitive learning

\[
\text{Weighted batch loss} = -\frac{1}{m} \left( \log p(s_1) + 0.2 \cdot \log p(s_2) + \cdots \right)
\]
Datasets and Experiments
Developing test set

• Manually translating sentences in test set as ground truth
  • Providing both English sentence and corresponding image
  • To eliminate ambiguity and translate referring to the image
Two Bilingual (English-Chinese) Datasets

- Extending Flickr8k and Flickr30k to bilingual version (English + Chinese)

Download: [https://github.com/li-xirong/cross-lingual-cap](https://github.com/li-xirong/cross-lingual-cap)

<table>
<thead>
<tr>
<th></th>
<th>Flickr8k-cn</th>
<th>Flickr30k-cn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Validation</td>
</tr>
<tr>
<td>Images</td>
<td>6,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Machine-translated</td>
<td>30,000</td>
<td>5,000</td>
</tr>
<tr>
<td>Chinese sentences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human-translated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese sentences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human-annotated</td>
<td>30,000</td>
<td>5,000</td>
</tr>
<tr>
<td>Chinese sentences</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments

• Baselines:
  1. Late translation[Li, ICMR2016]
  2. Late translation rerank
  3. Without fluency
  4. Manual Flickr8k-cn

• Proposed approaches:
  1. Fluency-only
  2. Rejection sampling
  3. Weighted loss

A person holding a book with a bird sitting on the book
A small bird sitting on top of an open book

Chinese Captioning Model

Machine translation

拿着一本书和一只鸟坐在书上的人 0.3
一只小鸟坐在一本打开的书上 0.8
Automatic Evaluation Results

- Late translation is not effective

Flickr8k-cn
- CIDEr
- ROUGE
- BLEU4

Flickr30k-cn
- CIDEr
- ROUGE
- BLEU4

Legend:
- Green: Late Translation
- Blue: Late Translation Rerank
- Yellow: Without fluency

√
Automatic Evaluation Results

- **Rejection sampling** and **Weighted loss** are able to preserve relevant information.
Human Evaluation

• Automatic evaluation is insufficient to guarantee the overall fluency

• Annotators rate the sentences using a Likert scale of 1 to 5 (higher is better) in two aspects, namely relevance and fluency
  • Sentences generated by distinct approaches are shown together
  • Sentences randomly shuffled before presenting to the annotators
Human Evaluation Results

- **Rejection sampling** achieves the best balance between relevance and fluency.

![Bar chart showing evaluation results for different methods and datasets.](chart.png)
Human Evaluation Results

- Rejection sampling achieves the best balance between relevance and fluency, without the need of manual written Chinese captions.
Conclusion

Fluency-guided framework

• Tackling cross-lingual image captioning with minimal manual annotation effort

• Capable of generating relevant and fluent captions in target language

https://github.com/li-xirong/cross-lingual-cap

bluey@ruc.edu.cn