Image Pivoting for Learning Multilingual Multimodal Representations

Spandana Gella, Rico Sennrich, Frank Keller, Mirella Lapata
Semantic Image Search

Given a database of images and a natural language query:
Retrieve images that are relevant to the query.

A dog is chasing a mouse
Semantic Image Search

Given a database of images and a natural language query:
Retrieve images that are relevant to the query.

Ein Hund jagt eine Maus
Motivation

Multi-modal Datasets and Applications: English!

- Image Description (MSCOCO)
- Visual Question Answering (VQA), Visual Dialog (VisDial)
Motivation

- Can we learn multilingual multimodal representations using image as bridge between languages?
- Are these representations better than monolingual multimodal representations?
Multi30k Dataset (Elliott et al., 2015)

- 31k images from Flickr30k
- Each image has 5 human written descriptions in English and German
- English and German descriptions are not parallel
- 29k training, 1k dev, 1k test
1) Two professional men’s soccer players playing soccer.
2) Two men playing soccer on a field.
3) Two soccer players on a green field play with a soccer ball.

1) Zwei Fussballer zweier Mannschaften jagen auf dem Spielfeld im Freien dem Ball hinterher.
2) Zwei männer kämpfen um einen fussball.
3) Szenen eines fussballspieles.
Models

**Baseline**: Visual Semantic Embeddings \(\text{(Vendrov et al., 2016)}\)

**PIVOT**: Image as pivot between the descriptions in both the languages

**PARALLEL**: PIVOT + image descriptions in both languages to be closer to each other
Visual Semantic Embeddings

\[ \sum_{(c,i)} \left( \sum_{c'} \max\{0, \alpha - S(c, i) + S(c', i)\} + \sum_{i'} \max\{0, \alpha - S(c, i) + S(c, i')\} \right) \]

\[ f_i(i) = W_i \cdot \text{CNN}(i) \]
\[ f_c(c) = \text{RNN}(c) \]

Two men playing soccer on a field

Joint Space

Women with headdresses are dancing
Visual Semantic Embeddings

Two men playing soccer on a field

Joint Space

Women with headdresses are dancing

\[ \sum_{(c,i)} \phi(c, i, c', i') \]
PIVOT Model

Two men playing soccer on a field

Zwei männer kämpfen einen fussball

Women with headdresses are dancing

Frauen in Kostümen posieren in einem Raum

\[ \sum_{(c_1,i)} \phi(c_1, i, c'_1, i') + \sum_{(c_2,i)} \phi(c_2, i, c'_2, i') \]
PARALLEL Model

Two men playing soccer on a field

Women with headdresses are dancing

Zwei männer kämpfen einen fussball

Frauen in Kostümen posieren in einem Raum

\[ \sum_{(c_1,i)} \phi(c_1, i, c'_1, i') + \sum_{(c_2,i)} \phi(c_2, i, c'_2, i') + \sum_{(c_1,c_2)} \phi(c_1, c_2, c'_1, c'_2) \]
## Results

<table>
<thead>
<tr>
<th></th>
<th>English Text to Image</th>
<th>German Text to Image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@10</td>
</tr>
<tr>
<td>Vendrov et al. (2016)</td>
<td>25.8</td>
<td>67.8</td>
</tr>
<tr>
<td>PIVOT</td>
<td>26.2</td>
<td><strong>68.4</strong></td>
</tr>
<tr>
<td>PARALLEL</td>
<td><strong>27.1</strong></td>
<td>66.9</td>
</tr>
</tbody>
</table>
Ein alter Mann mit blauer Jacke und Mütze wechselt einen Reifen
An old man with a blue jacket and a hat is changing a tire

Vendrov et al. (2016)

PIVOT
zwei junge Männer trommeln auf der Straße
Two young men are playing the drums on the street.

Vendrov et al. (2016)

PARALLEL
## Results

<table>
<thead>
<tr>
<th>Image</th>
<th>Descriptions</th>
<th>Image Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td>2 Menschen auf der Straße mit Megafon</td>
<td>141 37 6</td>
</tr>
<tr>
<td></td>
<td>two people in blue shirts are outside with a bullhorn</td>
<td>85 7 3</td>
</tr>
<tr>
<td><img src="image2.png" alt="Image" /></td>
<td>ein Verkäufer mit weißem Hut und blauem Hemd, verkauft Kartoffeln oder ähnliches an Männer und Frauen</td>
<td>36 1 3</td>
</tr>
<tr>
<td></td>
<td>at an outdoor market, a small group of people stoop to buy potatoes from a street vendor, who has his goods laid out on the ground</td>
<td>24 2 2</td>
</tr>
</tbody>
</table>
Conclusion

- Image pivoting across languages helps in learning better multimodal representations
- German ranking benefits most from multilingual signal
Conclusion

- Image pivoting across languages helps in learning better multimodal representations
- German ranking benefits most from multilingual signal

<table>
<thead>
<tr>
<th>Language</th>
<th>Types</th>
<th>Tokens</th>
<th>Avg.L</th>
<th>Singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>22.8k</td>
<td>1.8M</td>
<td>12.3</td>
<td>9.2k</td>
</tr>
<tr>
<td>German</td>
<td>46.1k</td>
<td>1.4M</td>
<td>9.6</td>
<td>26.5k</td>
</tr>
</tbody>
</table>
Conclusion

▶ Image pivoting across languages helps in learning better multimodal representations
▶ German ranking benefits most from multilingual signal

Future Work:

▶ Monolingual image descriptions
▶ Parallel corpora between languages
▶ Sub-word units and Regional Image Representations
▶ Pivoting for tri-modal applications (speech, text, vision)

Thank You!
Bibliography I


Two people riding a colorfully decorated bicycle

Vendrov et al. (2016)

PIVOT
Semantic Textual Similarity

Given two sentences $s_1$ and $s_2$, determine how similar the sentences are on a scale 0–5

- MSR Video description pairs, Image Description pairs
- Evaluation: Pearson Correlation Coefficient

<table>
<thead>
<tr>
<th>Model</th>
<th>Video</th>
<th>STS-2014</th>
<th>STS-2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW Baseline</td>
<td>29.9</td>
<td>51.3</td>
<td>60.4</td>
</tr>
<tr>
<td>STS Best</td>
<td>87.3</td>
<td>83.4</td>
<td>86.4</td>
</tr>
<tr>
<td>VSE Baseline</td>
<td>82.2</td>
<td>84.1</td>
<td>90.4</td>
</tr>
<tr>
<td>PIVOT</td>
<td>83.1</td>
<td>83.8</td>
<td>90.3</td>
</tr>
<tr>
<td>PARALLEL</td>
<td>84.6</td>
<td>84.5</td>
<td>91.5</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>Image to Text English</th>
<th></th>
<th>Image to Text German</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@10</td>
<td>Mr</td>
</tr>
<tr>
<td>Vendrov et al. (2016)</td>
<td>34.8</td>
<td>74.8</td>
<td>3</td>
</tr>
<tr>
<td>Pivot</td>
<td>33.8</td>
<td>75.2</td>
<td>3</td>
</tr>
<tr>
<td>Parallel</td>
<td>31.5</td>
<td>74.7</td>
<td>3</td>
</tr>
</tbody>
</table>
## STS Results

<table>
<thead>
<tr>
<th>S1</th>
<th>S2</th>
<th>GT</th>
<th>Pred</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black bird standing on concrete.</td>
<td>Blue bird standing on green grass.</td>
<td>1.0</td>
<td>4.2</td>
</tr>
<tr>
<td>Two zebras are playing.</td>
<td>Zebras are socializing.</td>
<td>4.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Three goats are being rounded up by a dog.</td>
<td>Three goats are chased by a dog.</td>
<td>4.6</td>
<td>4.5</td>
</tr>
<tr>
<td>A man is folding paper.</td>
<td>A woman is slicing a pepper.</td>
<td>0.6</td>
<td>0.6</td>
</tr>
</tbody>
</table>
MT errors

Google Translate

Translate

Sania is serving a tennis ball

Sania dient einem Tennisball
State-of-the-art NMT English-German

Source: people **riding** horses on the beach
Translation: **die leute fahren** pferde am strand
Reference: **personen reiten** auf pferden am strand
Related Work

- Multilingual Representation Learning (Chandar et al., 2013; Schwenk et al., 2017)
- Multilingual Image Description Generation (Elliott et al., 2015)
- Multi-modal Machine Translation (Specia et al., 2016)
- Multilingual Multimodal Representations (Rajendran et al., 2016; Calixto et al., 2017)
- Visual Semantic Embeddings (Socher et al., 2014; Kiros et al., 2015)
## Multi30k statistics

<table>
<thead>
<tr>
<th>Language</th>
<th>Types</th>
<th>Tokens</th>
<th>Avg.L</th>
<th>Singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>22.8k</td>
<td>1.8M</td>
<td>12.3</td>
<td>9.2k</td>
</tr>
<tr>
<td>German</td>
<td>46.1k</td>
<td>1.4M</td>
<td>9.6</td>
<td>26.5k</td>
</tr>
</tbody>
</table>
Visual Semantic Embeddings

Given a dataset of image and description pairs \((c, i)\) : Learn a joint space \(\mathbb{R}^N\) such that \(i\) and \(c\) are closer to each other.

\[
\sum_{(c,i)} \left( \sum_{c'} \max\{0, \alpha - S(c, i) + S(c', i)\} + \sum_{i'} \max\{0, \alpha - S(c, i) + S(c, i')\} \right)
\]

\((1)\)

\[
S(c, i) = \begin{cases} 
  f_c(c) \cdot f_i(i) & \text{VSE (Kiros et al., 2015)} \\
  -E(f_i(i), f_c(c)) & \text{OE (Vendrov et al., 2016)}
\end{cases}
\]

\[
f_i(i) = W_i \cdot \text{CNN}(i)
\]

\[
f_c(c) = \text{RNN}(c)
\]
<table>
<thead>
<tr>
<th>System</th>
<th>En BLEU ↑</th>
<th>Meteor ↑</th>
<th>TER ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elliott et al. (2015)</td>
<td>–</td>
<td><strong>15.8</strong></td>
<td><strong>31.2</strong></td>
</tr>
<tr>
<td>PIVOT N</td>
<td><strong>5.4</strong></td>
<td>20.2</td>
<td>89.7</td>
</tr>
<tr>
<td>PARALLEL N</td>
<td><strong>6.1</strong></td>
<td>20.0</td>
<td>92.0</td>
</tr>
<tr>
<td>PIVOT Y</td>
<td>11.3</td>
<td>27.0</td>
<td><strong>74.3</strong></td>
</tr>
<tr>
<td>PARALLEL Y</td>
<td>11.4</td>
<td>27.8</td>
<td>74.9</td>
</tr>
</tbody>
</table>

Results on WMT’16 MMT Task2; Column En denote whether English descriptions provided at test are used to re-rank the German descriptions or not.