MULTILINGUAL IMAGE DESCRIPTION WITH NEURAL SEQUENCE MODELS

Desmond Elliott, Stella Frank, Eva Hasler
February 3, 2016
iV&L Net Working Group 3 Meeting
Stella Frank

Eva Hasler
VISION AND LANGUAGE RESEARCH

Grounded Semantics [Silberer and Lapata, 2014]

Image Description [Farhadi et al., 2010]

Video Description [Venugopalan et al., 2015]

Question-Answering [Gao et al., 2015]
A bike is leaning against a stone wall.

See Bernardi et al. [2016] for an overview of datasets, models, and evaluations.
• Extend image description generation to new languages
• Text-based image search in any language
• Localised alt-text generation on the Web
• Translate movie descriptions
HOW CAN WE EXPLOIT MULTILINGUAL MULTIMODAL CONTEXT?

Ein Rad steht neben der Mauer → A bicycle / wheel · · ·
Ein Rad steht neben der Mauer → A bicycle / wheel · · ·
<Image features> → A ? is leaning against the wall
HOW CAN WE EXPLOIT MULTILINGUAL MULTIMODAL CONTEXT?

Ein Rad steht neben der Mauer → A bicycle / wheel · · ·

<Image features> → A ? is leaning against the wall

Possible solutions:

• Collect more data
HOW CAN WE EXPLOIT MULTILINGUAL MULTIMODAL CONTEXT?

Ein Rad steht neben der Mauer → A bicycle / wheel ⋯
<Image features> → A ? is leaning against the wall

Possible solutions:

• Collect more data
• Exploit data in a different modality (images or video)
HOW CAN WE EXPLOIT MULTILINGUAL MULTIMODAL CONTEXT?

Ein Rad steht neben der Mauer → A bicycle / wheel · · ·
<Image features> → A ? is leaning against the wall

Possible solutions:

• Collect more data
• Exploit data in a different modality (images or video)
Let $t$ be the **target** language description, $s$ be the **source** language description and $i$ be the **image**.

1. **Multimodal Machine Translation**
   - Always given a source description and image
   - $\hat{t} = \arg\max_t p(t|i, s)$
Let $t$ be the target language description, $s$ be the source language description and $i$ be the image.

1. Multimodal Machine Translation
   - Always given a source description and image
   - $\hat{t} = \text{argmax}_t p(t|i, s)$

2. Crosslingual Image Description
   - Automatically generate a source language description
   - $\hat{t} = \text{argmax}_t p(t|i, \hat{s})$
MULTILINGUAL MULTIMODAL MODEL
MULTIMODAL LANGUAGE MODELS
[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]
MULTIMODAL LANGUAGE MODELS
[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]

\[ e_i = W_{he}w_i \]
MULTIMODAL LANGUAGE MODELS

[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]

\[
e_i = W_{he} w_i
\]

\[
h_i = f(W_{hh} h_{i-1} + e_i + \mathbb{1}(t = 0) \cdot W_{hv} v)
\]
MULTIMODAL LANGUAGE MODELS
[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]

\[ e_i = W_{he} w_i \]
\[ h_i = f(W_{hh} h_{i-1} + e_i + \mathbb{1}(t = 0) \cdot W_{hv} v) \]
\[ o_i = \text{softmax}(W_{oh} h_i) \]
Multimodal language models
[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]

\[
\begin{align*}
  e_i &= W_{he}w_i \\
  h_i &= f(W_{hh}h_{i-1} + e_i + 1(t = 0) \cdot W_{hv}v) \\
  o_i &= \text{softmax}(W_{oh}h_i) \\
  \text{Loss} &= - \sum_{n=1}^{N} \sum_{i=1}^{K} \log p(o_i)
\end{align*}
\]
• Initialise with image features and BOS token
- Initialise with image features and BOS token
- Feed sampled word into the next timestep
• Initialise with image features and BOS token
• Feed sampled word into the next timestep
• Decode until emit EOS token
SOURCE LANGUAGE ENCODER

\[ s_i \rightarrow \ldots \rightarrow s_{i+1} \]

CNN
\[ h_i = f(W_{hh}h_{i-1} + e_i + \mathbb{1}(t = 0) \cdot W_{hv}v + \mathbb{1}(t = 0) \cdot W_{hs}s) \]
• Each model trained towards its own objective, unlike Sequence-to-Sequence Learning [Sutskever et al., 2014]
  • CNN: object recognition
  • Source LM: source language generation
  • Target LM: target language generation

• MMLM learns task-specific representations given transferred inputs
  • e.g. Target-LM with multimodal source features vs. separate visual and source features
  • Easily work on new languages with fixed input representations
EXPERIMENTS
• Generate description in target language

• Measures\(^1\): Meteor, BLEU, Perplexity

1. Multimodal Machine Translation
   • Always given a source description and image

2. Crosslingual Image Description
   • Given an image, automatically generate a source description with a source MLM
     • and pass encoded textual features to a target LM
     • and pass encoded visual+textual features to a target LM

\(^1\)See Elliott and Keller [2014] and Vedantam et al. [2015] for more details on measuring image description quality
1. a yellow building with white columns in the background
2. two palm trees in front of the house

- 17,655 training / 1,962 testing
- Up to five semantically diverse descriptions / image
  - We use only the first description
- Descriptions translated from English to German
• Models are built using Keras library
• Adam optimiser [Kingma and Ba, 2014]
• Mini-batches of 100 examples
• Dropout over word, visual, and source features ($p = 0.5$)
• LSTM with 256-D memory cell [Hochreiter and Schmidhuber, 1997]
• 4096-D visual features from 15th layer of VGG-16 CNN [Simonyan and Zisserman, 2015]
• 256-D source feature vectors
• 256-D word embedding features
• Vocabulary size German: 2,374, English: 1,763 (UNK<3)
Schulkinder sitzen in einem Klassenzimmer
MODELS: SOURCE LM $\rightarrow$ TARGET LM

Source Model: CNN $\rightarrow$ RNN

Target Model: RNN $\rightarrow$ RNN

Children sitting in a classroom

Schulkinder sitzen in einem Klassenzimmer
MODELS: SOURCE MLM → TARGET LM

Source Model

Target Model

CNN → RNN

children sitting in a classroom

Schulkinder sitzen in einem Klassenzimmer
RESULTS
Trained and evaluated over all references.
First non-English image description results.
CROSSLINGUAL IMAGE DESCRIPTION RESULTS

Source descriptions automatically generated by Source-MLM
MLM: a man is standing on a grey rock in the foreground

De Ref: bergsteiger klettern auf einen sehr steilen eishang

MLM-MLM: tourists are climbing up a snowy slope

\(^4\)Thousands of examples from all models at http://staff.fnwi.uva.nl/d.elliott/GroundedTranslation/
De MLM:
ein mann und eine frau stehen
an einem sandstrand mit dem
meer im hintergrund

En MLM: a man with a black
jacket and a black jacket is
standing in a brown rocky
desert landscape

En MLM-LM:
a man and a woman are
standing in a reed boat on a
lake
VISUALISING THE EFFECT OF TRANSFERING FEATURES

- t-SNE plots of the LSTM memory cell at $t=0$
- MLM $\rightarrow$ MLM: closer to pictures of snow!
• How well does this generalise to other languages?
• Attention-based Image Description [Xu et al., 2015]
• Compare with target-side translation retrieval with multimodal features [Hitschler and Riezler, 2016]
• Human judgements of generated descriptions
• Larger datasets (Shared Task at WMT16!)
• Multilingual video description, other tasks . . .
CONCLUSIONS

• Multilingual Image Description is a natural extension of Image Description

• MMLM transfers multimodal features between languages

• Transferring multilingual multimodal representations between languages improves image description quality

• Code: http://github.com/elliottd/GroundedTranslation
APPENDICES
## Complete English Results

<table>
<thead>
<tr>
<th></th>
<th>BLEU4</th>
<th>Meteor</th>
<th>PPLX</th>
</tr>
</thead>
<tbody>
<tr>
<td>En MLM</td>
<td>14.2 ± 0.3</td>
<td>15.4 ± 0.2</td>
<td>6.7 ± 0.0</td>
</tr>
<tr>
<td>De LM → En LM</td>
<td>21.3 ± 0.5</td>
<td>19.6 ± 0.2</td>
<td>6.0 ± 0.1</td>
</tr>
<tr>
<td>Mao et al. [2015]</td>
<td>20.8</td>
<td>—</td>
<td>6.92</td>
</tr>
<tr>
<td>De MLM → En MLM</td>
<td>18.0 ± 0.3</td>
<td>18.0 ± 0.2</td>
<td>6.3 ± 0.1</td>
</tr>
<tr>
<td>De LM → En MLM</td>
<td>17.3 ± 0.5</td>
<td>17.6 ± 0.5</td>
<td>6.3 ± 0.0</td>
</tr>
<tr>
<td>De MLM → En LM</td>
<td>23.1 ± 0.1</td>
<td>20.9 ± 0.0</td>
<td>5.7 ± 0.1</td>
</tr>
</tbody>
</table>
## Complete German Results

<table>
<thead>
<tr>
<th></th>
<th>BLEU4</th>
<th>Meteor</th>
<th>PPLX</th>
</tr>
</thead>
<tbody>
<tr>
<td>De MLM</td>
<td>9.5 ± 0.2</td>
<td>20.4 ± 0.2</td>
<td>10.35 ± 0.1</td>
</tr>
<tr>
<td>En LM → De LM</td>
<td><strong>17.8 ± 0.7</strong></td>
<td><strong>29.9 ± 0.5</strong></td>
<td><strong>8.95 ± 0.4</strong></td>
</tr>
<tr>
<td>En MLM → De MLM</td>
<td>11.4 ± 0.7</td>
<td>23.2 ± 0.9</td>
<td>9.69 ± 0.1</td>
</tr>
<tr>
<td>En LM → De MLM</td>
<td>12.1 ± 0.5</td>
<td>24.0 ± 0.3</td>
<td>10.2 ± 0.7</td>
</tr>
<tr>
<td>En MLM → De LM</td>
<td><strong>17.0 ± 0.3</strong></td>
<td><strong>29.2 ± 0.2</strong></td>
<td><strong>8.84 ± 0.3</strong></td>
</tr>
</tbody>
</table>
RNN ARCHITECTURE: LONG-SHORT TERM MEMORY
[HÖCHREITER AND SCHMIDHUBER, 1997]

Credit: Christopher Olah
Effect of optimisation method

<table>
<thead>
<tr>
<th>Optimisation Method</th>
<th>Epoch</th>
<th>Val Loss</th>
<th>BLEU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdaGrad</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AdaDelta</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADAM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSprop</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The diagrams show the comparison of Val Loss and BLEU4 scores for different optimisation methods over epochs.
EFFECT OF RNN TYPE

- **Val loss**
  - LSTM | $h|=256$
  - GRU | $h|=256$

- **BLEU4**
  - LSTM | $h|=256$
  - GRU | $h|=256$
EFFECT OF HIDDEN STATE SIZE

![Graphs showing the effect of hidden state size on validation loss and BLEU score.]
EFFECT OF DECOMPOUNDING GERMAN WORDS

![Graphs showing the effect of decomposing German words.

- Log Perplexity graph with the following legend:
  - UNK0 (100%) (10.79)
  - UNK1 (59.1%) (11.11)
  - UNK2 (45.5%) (10.44)
  - UNK3 (38.1%) (10.09)

- BLEU4 graph with the following legend:
  - UNK0 (100%) (10.92)
  - UNK1 (59.1%) (11.47)
  - UNK2 (45.5%) (12.59)
  - UNK3 (38.1%) (12.09)
EFFECT OF UNK THRESHOLD

Log Perplexity vs Epoch

- UNK0 (100%) (10.62)
- UNK1 (59.1%) (10.03)
- UNK2 (45.5%) (9.56)
- UNK3 (38.1%) (9.38)

BLEU4 vs Epoch

- UNK0 (100%) (11.68)
- UNK1 (59.1%) (11.21)
- UNK2 (45.5%) (11.92)
- UNK3 (38.1%) (11.85)
• Trained to predict 1000 object labels
• Over 1m training images
• Visual features transferred from the penultimate layer
VISUALISING CNN FILTERS

conv1_1: a few of the 64 filters

conv2_1: a few of the 128 filters

Credit: François Chollet
**VISUALISING CNN FILTERS**

**conv3_1:** a few of the 256 filters

**conv4_1:** a few of the 512 filters

Credit: François Chollet
conv5_1: a few of the 512 filters

Credit: François Chollet


