Generating Image Descriptions with Gold Standard Visual Inputs: Motivation, Evaluation and Baselines

Josiah Wang & Robert Gaizauskas

Work presented as part of the Visual Sense (ViSen) consortium, funded by the D2K programme
Task: Generate a description...

... given labelled bounding boxes
This Presentation

• ImageCLEF 2015 image description generation task
• Fine-grained evaluation metric for content selection
• Introduce and evaluate baselines
Motivation

• Why gold standard visual input?
  – ‘Detect and generate’ methods – sensitive to noisy vision input
  – Assume ‘perfect’ input: Can we generate? How?

• Why fine-grained evaluation?
  – Human judgments are expensive, not scalable
  – Existing measures (BLEU, ROUGE, METEOR, CIDEr) are global
  – Better to evaluate tasks/ phases of NLG pipeline explicitly: referring expression, verb/event/attribute/spatial relation expression, content ordering, content selection
ImageCLEF

• Evaluation campaign since 2003
• Benchmark automatic image annotation & indexing for wide range of source images
• Part of CLEF initiative
  – Conference and Labs of the Evaluation Forum
  – Evaluation of information access systems since 2000
ImageCLEF 2015

- Four main tasks in 2015
  - **Scalable image annotation**
  - Medical classification
  - Medical clustering
  - Liver CT annotation
ImageCLEF 2015: Scalable Image Annotation

• Training set: 500k (image, webpage) pairs

• Subtask 1: Image annotation + localisation
  – For each 500k image, annotate + localise with 251 concepts

• Subtask 2: Image description generation
  – Noisy track: Generate descriptions for all 500k images
  – **Clean track:** Generate descriptions for 450 test images, given labelled bounding boxes
ImageCLEF 2015: Subtask 2 (Clean Track)

• Validation set (500 images from 500k)
  – Minimum 5 descriptions per image
    • Mean 9.5: Median: 8, Max: 51
  – Labelled bounding box annotations for 251 WordNet synsets
  – Correspondence annotation between text and bounding boxes


[Woman]² dressed in white with gold [boots]⁵ poses next to a police [car]³.


A [woman]² is leaning against a [car]³.


A blonde [woman]² wearing gold shiny [boots]⁵, a white [top]⁰ and short white skirt is leaning on a [car]³.
Content Selection Evaluation Metric

\[ P^I = \frac{1}{M} \sum_{m=1}^{M} \frac{|G_m^I \cap S^I|}{|S^I|} \]

bbox instances referenced in gold standard

\[ R^I = \frac{1}{M} \sum_{m=1}^{M} \frac{|G_m^I \cap S^I|}{|G_m^I|} \]

bbox instances referenced in generated sentence

\[ F^I = 2 \times \frac{P^I \times R^I}{P^I + R^I} \]

# gold standard descriptions for image I

- Final score: Average over all test images
Generating Descriptions: Baselines

• Visual Cues
  – Bounding box size (choose biggest first)
  – Bounding box position (choose most central object first)

• Threshold selected number of bounding boxes, $k$
Generating Descriptions: Baselines

• Textual Priors (from validation set)
  – Unigram
  – Bigram

A \texttt{[woman]} in a white \texttt{[dress]} and gold \texttt{[boots]} leaning on a \texttt{[car]}.

\begin{align*}
<\texttt{start}> & \rightarrow \texttt{woman} \\
\texttt{woman} & \rightarrow \texttt{dress} \\
\texttt{dress} & \rightarrow \texttt{boots} \\
\texttt{boots} & \rightarrow \texttt{car} \\
\texttt{car} & \rightarrow <\texttt{end}> 
\end{align*}
Generating Descriptions: Baselines

• Function Words
  – Random preposition/conjunction followed by optional ‘the’

\[ \text{[Woman]}^2 \text{ with [dress]}^0 \text{ and [boots]}^5 \text{ on the [car]}^3. \]
Generating Descriptions: Results (F-score)

![Graph showing the F-score results for generating descriptions as a function of k (maximum number of instances), with a random point indicated.](image-url)
Generating Descriptions: Results (F-score)

[Graph showing F-score changes with k (maximum number of instances). The graph includes two lines: one for bbox pos and one for bbox size. The F-score peaks at a certain k value for bbox pos, while bbox size maintains a relatively flat line. The random line is shown as a comparison.]
Generating Descriptions: Results (F-score)

The diagram shows the results of generating descriptions using different methods: bigram, unigram, and random. The x-axis represents the maximum number of instances (k), and the y-axis represents the F-score. The graph illustrates the performance of each method as the number of instances increases.
Generating Descriptions: Results (Precision)

![Graph showing precision vs. k (maximum number of instances) for different features: unigram, bigram, bbox pos, bbox size, and random. The graph indicates that bigram feature performs the best, followed by unigram, bbox size, bbox pos, and random.]
Generating Descriptions: Results (Recall)

- **bbox pos**
- **bbox size**

Graph showing the recall ($R$) as a function of $k$ (maximum number of instances) for different methods:
- **unigram**
- **bigram**
- **random**
- random [F: 0.43]
  - [Wall]$^4$ inside [door]$^3$ around the [bicycle]$^0$.

- bbox position [F: 0.79]

- bbox size [F: 0.79]
  - [Bicycle]$^0$ on [floor]$^1$ with [wall]$^2$.

- unigram [F: 0.34]
  - [Table]$^7$ in the [wall]$^4$ around [wall]$^2$.

- bigram [F: 0.03]
  - [Table]$^7$ near [door]$^3$. 
• random [F: 0.05]
  – [Park] behind [wheel] underneath the [window].

• bbox position [F: 0.59]
  – [Park] on the [car] below [river].

• bbox size [F: 0.44]
  – [Park] behind the [car] against the [tree].

• unigram [F: 0.42]
  – [Tree] beneath [car] by [window].

• bigram [F: 0.71]
  – [Car] inside [flag] underneath the [flag].
• random [F: 0.39]
  – [Vest] at [hat] behind the [picture].

• bbox position [F: 0.49]
  – [Picture] on [man] beside the [train].

• bbox size [F: 0.49]

• unigram [F: 0.77]

• bigram [F: 0.77]
  – [Man] around the [hat] along the [book].
Generating Descriptions: Combined

- Combining Text + Visual Priors
  - Re-rank by average rank from two systems

<table>
<thead>
<tr>
<th>Bigram</th>
<th>Bbox Position</th>
<th>Average Rank</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>dress</td>
<td>2</td>
<td>(2+3)/2 = 2.5</td>
<td>2</td>
</tr>
<tr>
<td>wheel</td>
<td>- (6)</td>
<td>(6+4)/2 = 5</td>
<td>5</td>
</tr>
<tr>
<td>woman</td>
<td>1</td>
<td>(1+1)/2 = 1</td>
<td>1</td>
</tr>
<tr>
<td>car</td>
<td>4</td>
<td>(4+2)/2 = 3</td>
<td>3</td>
</tr>
<tr>
<td>hair</td>
<td>- (6)</td>
<td>(6+6)/2 = 6</td>
<td>7</td>
</tr>
<tr>
<td>boot</td>
<td>3</td>
<td>(3+7)/2 = 5</td>
<td>4</td>
</tr>
<tr>
<td>sign</td>
<td>- (6)</td>
<td>(6+5)/2 = 5.5</td>
<td>6</td>
</tr>
</tbody>
</table>

[Woman]² in [dress]⁰ by [car]³. (for k=3)
Generating Descriptions: Combined
Discussion

• We presented the sentence generation task of ImageCLEF
  – Proposed content selection evaluation metric and baselines
• Challenges
  – Bounding box annotations may not be informative enough
  – Suitability of fine-grained metrics
• Future work
  – Fine-grained metrics
    • Content ordering, referring expressions, verbs/predicates/prepositions
  – Generation of image descriptions
    • Stronger cues (co-occurrences, spatial relations)
Generating Image Descriptions with Gold Standard Visual Inputs: Motivation, Evaluation and Baselines

Josiah Wang & Robert Gaizauskas

Work presented as part of the Visual Sense (ViSen) consortium, funded by the CHIST-ERA D2K programme