Query-by-Example Image Retrieval using Visual Dependency Representations

Desmond Elliott, Victor Lavrenko, Frank Keller

University of Edinburgh

August 24, 2014
Query-by-Example Image Retrieval
Query-by-Example Image Retrieval

In this talk, similar means same action
Query-by-Example Image Retrieval
Query-by-Example Image Retrieval
Query-by-Example Image Retrieval
Typical Approaches to Image Retrieval

- Represent images as automatically extracted bag-of-visual-words (*visterms*)
  - SIFT, HoG, etc...

![Diagram of image retrieval approach]
Typical Approaches to Image Retrieval

- Represent images as automatically extracted bag-of-visual-words (visterms)
  - SIFT, HoG, etc...

- Large heterogenous data sets
  - Corel 5K (5K images)
  - CIFAR-10 (60K images)
  - TinyImages (100K images)
  - …
This Talk

- Represent images as annotated regions
  - Tighter connection to language than a *visterm*
This Talk

- Represent images as annotated regions
  - Tighter connection to language than a *visterm*

- Smaller data set: 341 images depicting **actions**
  - Explore the effect of action types on accuracy
This Talk

- Represent images as annotated regions
  - Tighter connection to language than a *visterm*

- Smaller data set: 341 images depicting **actions**
  - Explore the effect of action types on accuracy
  - Focus on encoding the **spatial** relationships between regions
Humans benefit from consistent spatial relationships
Biederman (1972)
Humans benefit from consistent spatial relationships
Biederman (1972)
Visual Dependency Representation (Elliott and Keller, 2013)

- Novel structured representation over image regions
  - Captures salient region-region relationships
  - Guided by the written description of the image
- Proved useful for describing actions in Elliott and Keller (2013)
- Inspired by dependency-syntact of language (Tesnière, 1953)
  - Tokens: image regions
  - Grammar: spatial relationships
Visual Dependency Representation (Elliott and Keller, 2013)

- Novel structured representation over image regions
  - Captures salient region-region relationships
  - Guided by the written description of the image
- Proved useful for describing actions in Elliott and Keller (2013)
- Inspired by dependency-syntax of language (Tesnière, 1953)
  - Tokens: image regions
  - Grammar: spatial relationships

A $\rightarrow$ on $\rightarrow$ B
A $\rightarrow$ surrounds $\rightarrow$ B
Visual Dependency Representation (Elliott and Keller, 2013)

- Novel structured representation over image regions
  - Captures salient region-region relationships
  - Guided by the written description of the image
- Proved useful for describing actions in Elliott and Keller (2013)
- Inspired by dependency-syntax of language (Tesnière, 1953)
  - Tokens: image regions
  - Grammar: spatial relationships

A $\rightarrow$ on B
A $\rightarrow$ surrounds B
A $\rightarrow$ beside B
A $\rightarrow$ opposite B
Visual Dependency Representation (Elliott and Keller, 2013)

- Novel structured representation over image regions
  - Captures salient region-region relationships
  - Guided by the written description of the image
- Proved useful for describing actions in Elliott and Keller (2013)
- Inspired by dependency-syntax of language (Tesnière, 1953)
  - Tokens: image regions
  - Grammar: spatial relationships

- $A \xrightarrow{\text{on}} B$
- $A \xrightarrow{\text{surrounds}} B$
- $A \xrightarrow{\text{beside}} B$
- $A \xrightarrow{\text{opposite}} B$
- $A \xrightarrow{\text{above}} B$
- $A \xrightarrow{\text{below}} B$
Visual Dependency Representation (Elliott and Keller, 2013)

- Novel structured representation over image regions
  - Captures salient region-region relationships
  - Guided by the written description of the image
- Proved useful for describing actions in Elliott and Keller (2013)
- Inspired by dependency-syntax of language (Tesnière, 1953)
  - Tokens: image regions
  - Grammar: spatial relationships

A on B
A surrounds B
A beside B
A opposite B
A above B
A below B
A infront B
A behind B
Gold Standard Example

“A girl is using a laptop. She is sitting on a bed.”
Gold Standard Example

“A girl is using a laptop. She is sitting on a bed.”
Gold Standard Example

“A girl is using a laptop. She is sitting on a bed.”
Gold Standard Example

“A girl is using a laptop. She is sitting on a bed.”
Gold Standard Example

“A girl is using a laptop. She is sitting on a bed.”
Data

341 images
from PASCAL VOC
Action Recognition
gold action labels
Data: 341 Images

- 10 types of actions

- Reading
- Ride horse
- Phoning
- Ride bike
- Play instrument
Data: 341 Images

- 10 types of actions

Jumping  Running  Walking
Data: 341 Images

- 10 types of actions

- Jumping
- Running
- Walking
- Use computer
- Take photo
Data

341 images from PASCAL VOC Action Recognition with action labels

D1  D2  D3

3 descriptions/image
1. A teenage girl is using a laptop. She is sitting on a bed.
2. A girl is using a laptop. There is a lamp beside her.
3. A girl is using a computer. There is a picture behind her.
Data

341 images from PASCAL VOC Action Recognition with action labels

3 descriptions/image

Objects for 341 images
Data

341 images from PASCAL VOC Action Recognition with action labels

3 descriptions/image

Objects for 341 images

1,023 VDRs
Automatic VDR Prediction

- Framed as a dependency parsing task
  - MaltParser (Nivre et al., 2004) seems unsuitable because it is incremental
- Construct a complete graph between all regions using MSTParser (McDonald et al., 2005)
  - Remove all features that encode the linear order of the input
  - Extract features from the image regions
Automatic VDR Prediction

- Framed as a dependency parsing task
  - MaltParser (Nivre et al., 2004) seems unsuitable because it is incremental
- Construct a complete graph between all regions using MSTParser (McDonald et al., 2005)
  - Remove all features that encode the linear order of the input
  - Extract features from the image regions
Automatic VDR Prediction

- Framed as a dependency parsing task
  - MaltParser (Nivre et al., 2004) seems unsuitable because it is incremental
- Construct a complete graph between all regions using MSTParser (McDonald et al., 2005)
  - Remove all features that encode the linear order of the input
  - Extract features from the image regions
VDR Parsing Experiment

- **Task**
  - Predict VDR over region-annotated image

- **Data**
  - 1,023 VDR data set
  - 10 fold cross-validation

- **Evaluation**
  - Unlabelled/labelled directed attachment accuracy

- **Models**
  - **FLAT** is a bag-of-regions baseline
  - **VDR** uses only input features
  - **VDR+IMG** also uses visual features
VDR Parsing Results

<table>
<thead>
<tr>
<th>Directed Dependency Accuracy</th>
<th>Labelled</th>
<th>Unlabelled</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLAT</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>VDR</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>VDR+IMG</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

Chart showing the comparison of Directed Dependency Accuracy for Labelled and Unlabelled data with three different methods: FLAT, VDR, and VDR+IMG.
Query-by-Example Image Retrieval

- Given a query example, find images of the same action

- Matching function: cosine with \textit{tf-idf} weighting
Query-by-Example Image Retrieval

- Given a query example, find images of the same action

- Matching function: cosine with *tf-idf* weighting

- Evaluate with Mean Average Precision and Precision@10
  - Relevance means same action annotation
Query-by-Example Image Retrieval

- Given a query example, find images of the same action

- Matching function: cosine with tf-idf weighting

- Evaluate with Mean Average Precision and Precision@10
  - Relevance means same action annotation

- Models:
  - Bag-of-Regions representation
  - Visual Dependency Representation
  - Both use gold-standard object annotations
Bag-of-Regions Representation

\[ \cos(a, b) = \frac{a \cdot b}{\|a\| \|b\|} \]
Bag-of-Regions Representation

\[
\cos(\text{woman using computer}, \text{old man using computer})
\]
Bag-of-Regions Representation

\[
\cos \left( \mathbf{v}_1, \mathbf{v}_2 \right) = \mathbf{v}_1 \cdot \mathbf{v}_2
\]

\[
\langle \text{person}, \text{laptop} \rangle \cdot \langle \text{person}, \text{laptop} \rangle
\]
Bag-of-Regions Representation

\[
\cos(\text{using computer}, \text{using computer}) = \frac{\langle \text{person}, \text{laptop} \rangle \cdot \langle \text{person}, \text{laptop} \rangle}{||\text{person}, \text{laptop}, \ldots|| \cdot ||\text{person}, \text{laptop}, \ldots||}
\]
Bag-of-Regions Representation

\[ \cos(\text{using computer}, \text{playing instrument}) \]
Bag-of-Regions Representation

\[
\cos(\text{playing instrument}, \text{using computer}) = \frac{\langle \text{person}, \text{laptop} \rangle \cdot \langle \text{person}, \text{laptop} \rangle}{\| \text{person}, \text{laptop} \| \cdot \| \text{person}, \text{laptop} \|}
\]
Visual Dependency Representation

- How to compare two trees?

Decompose all edges into bigrams and trigrams
Visual Dependency Representation

- How to compare two trees?
  - Decompose all edges into bigrams and trigrams

![Diagram with nodes and edges representing visual dependency]

18/25
How to compare two trees?
- Decompose all edges into bigrams and trigrams
Visual Dependency Representation

\[ \cos(\text{using computer, using computer}) = \frac{19}{25} \]
Visual Dependency Representation

\[
\text{cos} \left( \begin{array}{c}
\text{beside}
\end{array} \right) \cdot \begin{array}{c}
\text{beside}
\end{array} = \langle \text{Girl Laptop} \rangle \cdot \langle \text{Man Laptop} \rangle
\]
Visual Dependency Representation

\[
\cos(\text{using computer}, \text{using computer}) = \langle \text{Girl Laptop} \rangle \cdot \langle \text{Man Laptop} \rangle
\]
Visual Dependency Representation

\[ \cos(\text{using computer}, \text{playing instrument}) = \frac{19}{25} \]
Visual Dependency Representation

\[
\cos(\text{using computer}, \text{playing instrument}) = \langle \rangle
\]
Visual Dependency Representation

\[ \cos(\text{using computer}, \text{playing instrument}) = \]

\[ \begin{array}{c}
\text{beside} \\
\text{above}
\end{array} \begin{array}{c}
\text{beside}
\end{array} \]

\begin{align*}
\text{Girl, Laptop, Girl, Bed} & \cdots \\
\text{Man, Trumpet} & \cdots
\end{align*}
### Results

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag-of-Regions</td>
<td>0.467</td>
<td>0.415</td>
</tr>
</tbody>
</table>

VDR is significantly better than Bag-of-Regions at \( p < 0.01 \).
Results

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag-of-Regions</td>
<td>0.467</td>
<td>0.415</td>
</tr>
<tr>
<td>VDR</td>
<td>0.508*</td>
<td>0.451*</td>
</tr>
</tbody>
</table>

*: significantly better than Bag-of-Regions at $p < 0.01$
Transitive actions

![Graph showing Precision vs Recall for VDR and Bag-of-Regions](image-url)

- **VDR**
- **Bag-of-Regions**

The graph illustrates the performance of VDR and Bag-of-Regions in terms of Precision and Recall for transitive actions.
Intransitive actions
“Light” actions - use computer / take photo

![Graph showing precision and recall for use computer and take photo actions]
“Light” actions - use computer / take photo

Recall
0.0
0.2
0.4
0.6
0.8
1.0Precision
use computer
take photo

Recall
0.0
0.2
0.4
0.6
0.8
1.0Precision
use computer
take photo

Precision

0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Recall

use computer
take photo
Conclusions

- VDR increases the accuracy of query-by-example image retrieval compared to a bag-of-regions baseline
- Improvement depends on the type of action:
  - Most pronounced for transitive verbs
  - Least pronounced when no object is required for the action
- Future work:
  - Scaling to larger data sets
  - Different matching paradigms, e.g. RankSVM
  - Explore the effect of other languages on actions
Questions?

- VDRParser: http://github.com/elliotttd/vdrparser
- Data: http://homepages.inf.ed.ac.uk/s0128959/dataset/
- http://homepages.inf.ed.ac.uk/s0128959/
- d.elliott@ed.ac.uk // @delliott
References


