# Don't Mention the Shoe! A Learning to Rank Approach to Content Selection for Image Description Generation

Josiah Wang & Robert Gaizauskas

*Department of Computer Science, University of Sheffield, UK*


## Task

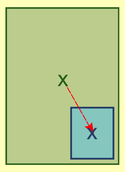
Select the bounding box instance(s) most likely to be mentioned by humans in the corresponding image description

[Learn from annotated human descriptions]

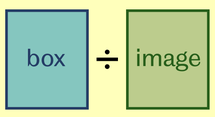
There is a woman[1] sitting on the floor.
A lady[1] sitting in the street with a snake[0] in a basket[11] .
A old Indian lady[1] is displaying a black snake[0] and other items .
An Indian lady[1] with a snake[0] .
A woman[1] in a red dress[5] sitting in the street .
A woman[1] in brightly colored clothing[5] is sitting with a snake[0] in a basket[11].


## Features


**Bounding box**




distance from image centre




area relative to image

**Image region**

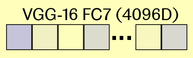
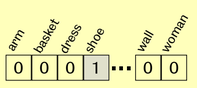
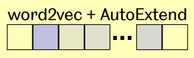



image embedding

**Concept label**




one-hot vector




synset embedding

## Rankers

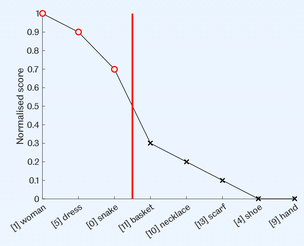
Rank objects based on features or feature combinations

Different rankers:

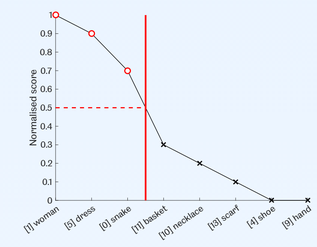
- **cascent:** Coordinate Ascent (listwise)
- **svmrank:** Ranking SVM (pairwise)
- **rforest:** Random Forests (pointwise)

## Stopping Criteria

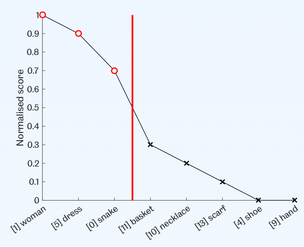
**fixed:** retain top 3 objects



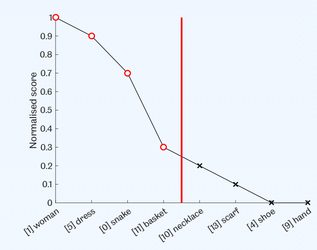

**absolute:** retain objects with score > 0.5




**relative1:** retain objects before largest difference in score




**relative2:** retain objects before largest difference in score + the object immediately after




## Results

Results of concatenating all features (please refer to paper for other results):

| | Stopping Criterion | Precision | Recall | $F_1$-score |
|---|---|---|---|---|
| **cascent** | fixed (top 3) | 0.59 ± 0.22 | 0.56 ± 0.23 | 0.55 ± 0.20 |
| | fixed (top 4) | 0.50 ± 0.20 | 0.63 ± 0.22 | 0.54 ± 0.17 |
| | absolute | 0.42 ± 0.22 | 0.72 ± 0.22 | 0.49 ± 0.17 |
| | relative1 | 0.72 ± 0.33 | 0.57 ± 0.29 | 0.53 ± 0.22 |
| | relative2 | 0.56 ± 0.25 | 0.66 ± 0.26 | 0.54 ± 0.20 |
| **svmrank** | fixed (top 3) | 0.60 ± 0.20 | 0.59 ± 0.22 | 0.57 ± 0.18 |
| | fixed (top 4) | 0.53 ± 0.18 | 0.68 ± 0.21 | 0.58 ± 0.16 |
| | absolute | 0.43 ± 0.20 | 0.80 ± 0.19 | 0.52 ± 0.15 |
| | relative1 | 0.67 ± 0.31 | 0.61 ± 0.29 | 0.53 ± 0.19 |
| | relative2 | 0.55 ± 0.25 | 0.70 ± 0.25 | 0.55 ± 0.18 |
| **rforest** | fixed (top 3) | 0.69 ± 0.18 | 0.68 ± 0.21 | 0.66 ± 0.16 |
| | fixed (top 4) | 0.60 ± 0.17 | 0.76 ± 0.19 | 0.65 ± 0.14 |
| | absolute | 0.84 ± 0.19 | 0.64 ± 0.21 | **0.70 ± 0.16** |
| | relative1 | 0.89 ± 0.18 | 0.57 ± 0.23 | 0.66 ± 0.18 |
| | relative2 | 0.71 ± 0.18 | 0.69 ± 0.21 | 0.68 ± 0.17 |
| **Human** | | 0.77 ± 0.11 | 0.77 ± 0.11 | 0.74 ± 0.12 |

## Discussion

- Random forests ranker performs best (new state-of-the-art)
  - Surprising: Random forests (pointwise) assumes bounding boxes are independent
  - Ranking SVM and Coordinate Ascent consider other bounding boxes as context
  - Perhaps random forests is a strong classifier?
- Stopping criteria:
  - **absolute**: depends on ranking algorithm
  - **relative1** precision > **relative2** precision
  - **relative2** recall > **relative1** precision
- Features:
  - Concept label features are more salient than bounding box or image features
  - Text (synset) embedding > one-hot vectors
  - Bounding box size > bounding box distance from centre
  - Image embedding + bounding box size slightly better than image embedding alone

## Future Work

- Stronger features
- Automatically gather larger noisy datasets: reduce annotation

## Acknowledgements

Supported by the chist-era Visual Sense (ViSen) project (EPSRC EP/K019082/1).