Describing Images using Inferred Visual Dependency Representations

Desmond Elliot, Arjen P. de Vries

ACL 2015

Presented by:

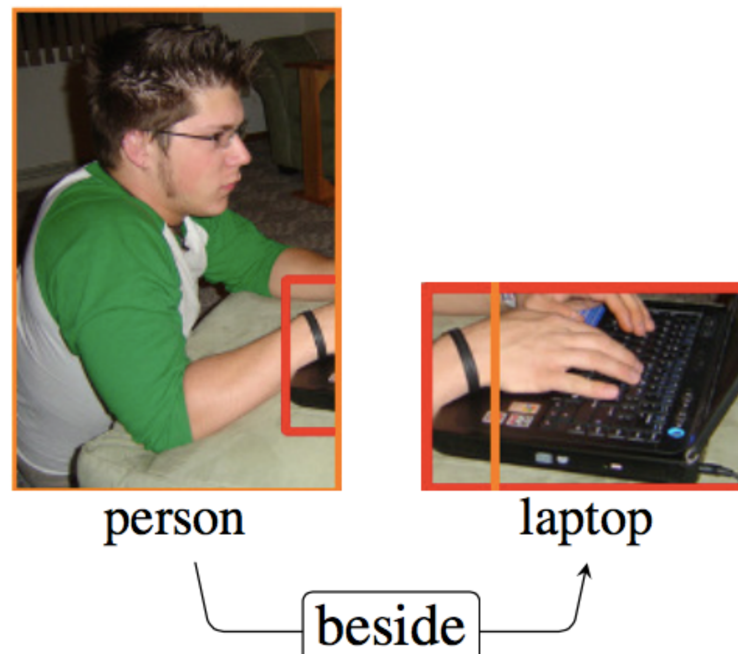
Sumit Maharjan

Tohoku University

Background

VDR (Visual Dependency Representation)

model spatial relationship between objects in an image



Objective and Motivation

Objective:

- Train a VDR without extensive human supervision
- Use VDR to generate image description

Motivation:

- Automatically generating literal description of images can help
  - Access to existing image
  - Information for visually impaired

Why VDR?

- Related with human cognition
- Spatial relationships between objects constrains image description

Related Works

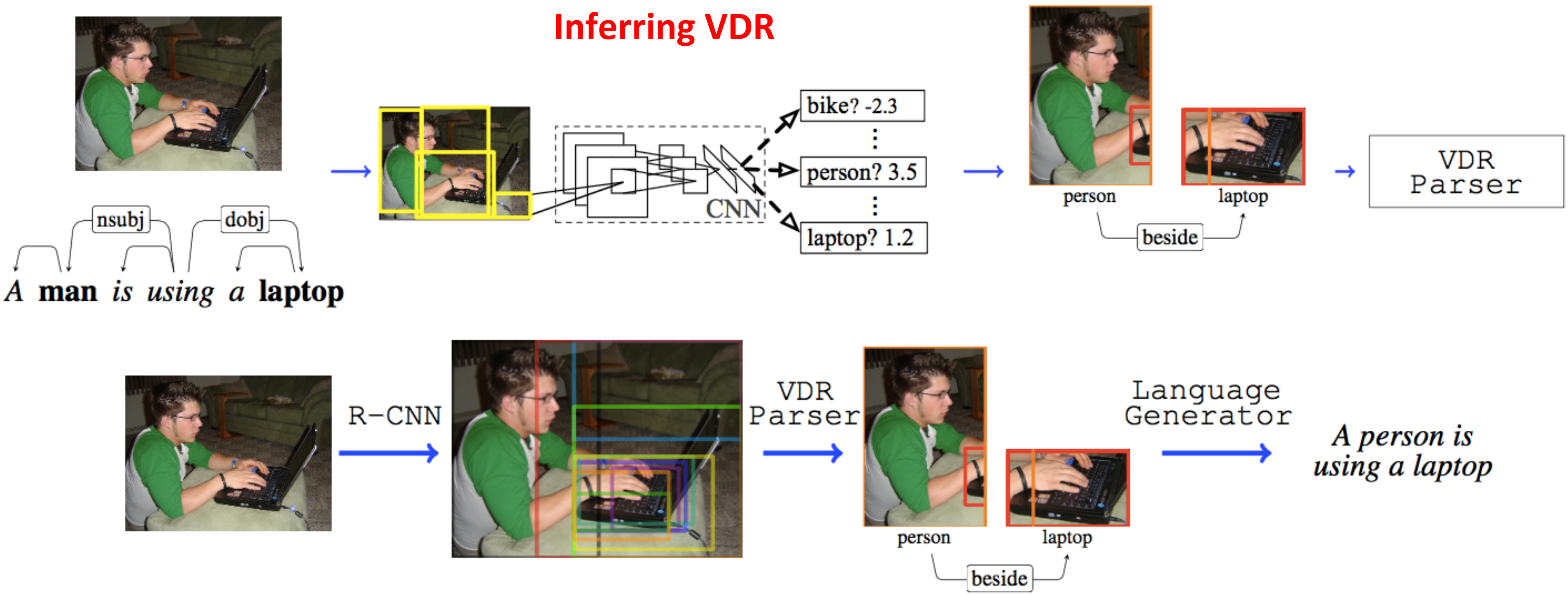
Different approaches

- *Spatial relationship* (Farhadi et al., 2010)
- *corpus-based relationships* (Yang et al., 2011)
- *spatial and visual attributes* (Kulkarni et al., 2011)
- *RNN and LSTM*
  (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Mao et al., 2015; Fang et al., 2015; Donahue et al., 2015; Lebret et al., 2015)
- VDR (Elliott and Keller, 2013)

Previous work: Relied on gold-standard training annotation

This work: Automatically infer training examples

Method: Overview



Generating descriptions

Method: Inferring VDRs

- **Description:** Dependency parsing to extract *nsubj* and *dobj* candidates
  - Lemmatized and transformed to WordNet hypernym parent
- **Image:** R-CNN(Girshick et al., 2014) to detect objects in image [200 classes]
  - Outputs bounding box and Confidence score
- **Infer VDR** for the object pairs using spatial relations

Relation	Definition
Beside	The angle between the subject and the object is either between 315° and 45° or 135° and 225°.
Above	The angle between the subject and object is between 225° and 315°.
Below	The angle between the subject and object is between 45° and 135°.
On	More than 50% of the subject overlaps with the object.
Surrounds	More than 90% of the subject overlaps with the object.

Spatial Relations

Method: Inferring VDRs

A **boy** is using a **laptop**

(a) on

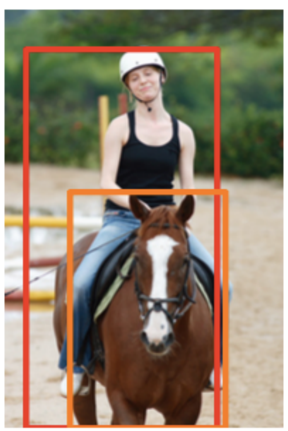
A **man** is riding a **bike**

(b) above

A **woman** is riding a **bike**

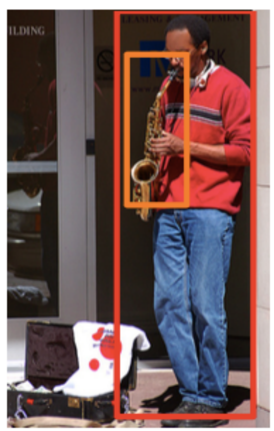
(c) surrounds

A **woman** is riding a **horse**



(d) surrounds

A **man** is playing a **sax**



(e) surrounds

A **man** is playing a **guitar**

(f) beside

The **woman** is wearing a **helmet**

(g) surrounds

Method: Generating Description

Language model

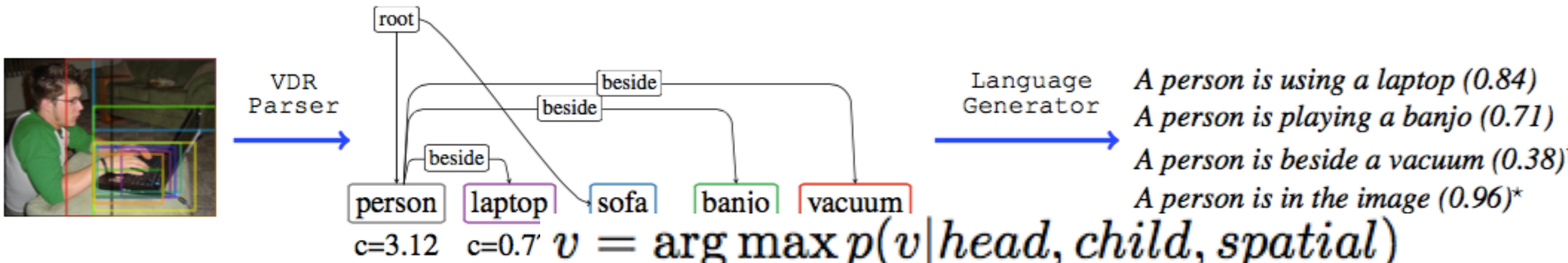
- **subjects, verbs, objects, and spatial relationships** from successfully constructed training examples
- Verb **stemmed** and inflected to **ing** using *morpha* and *morphg*
- *spatial relationship between the subject and object region is used to help constrain language generation to produce descriptions*

Method: Generating Description

- Description generated using template based model
- R-CNN detects gives top-N detected objects
- VDR Parser generates VDR structure for the detected objects
- All possible descriptions is generated using the template

Object detector output

Method: Generating Description



$$v = \arg\max_{v} p(v|head, child, spatial)$$

DT **head** is V DT **child**.

head *and* ***child****:*
objects from VDR

Verb selection

$$p(v|head,child, spatial) = p(v|head) \cdot p(child|v, head) \cdot p(spatial|child, v, head)$$

If relation can't be extracted

A/An **object** is in the image.

Sentence scoring

$$score(head, v,child, spatial) = p(v|head, child, spatial) \cdot sgm(head) \cdot sgm(child)$$

Experiments

Task: generation of natural language description of an image

Models to compare with

- **MIDGE** (Mitchell et al., 2012) [tree-substitution grammar and discrete object detections]
- **BRNN** (Karpathy and Fei-Fei, 2015) [multimodal deep neural network]

Evaluation Measures

- Meteor (Denkowski and Lavie, 2011)
- BLEU4 (Papineni et al., 2002),

Experiments : Data sets

Data sets

- Pascal1K
  - 1,000 images
  - sampled from the PASCAL Object Detection Challenge data set (Everingham et al., 2010)
  - each image has five descriptions collected from Mechanical Turk
  - Has a wide variety of subject matter
- VLT2K
  - 2,424 images
  - trainval 2011 portion of the PASCAL Action Recognition Challenge each image paired with three descriptions collected from Mechanical Turk

80% training, 10% validation, 10% test

Experiments: Results

- Performance of VDR depends on type of images
- Difference in Metoer and BLEU

	VLT2K		Pascal1K	
	Meteor	BLEU	Meteor	BLEU
VDR	16.0	14.8	7.4	9.0
BRNN	18.6	23.7	12.6	16.0
-genders	16.6	17.4	12.1	15.1
MIDGE	5.5	8.2	3.6	9.1
Human	26.4	23.3	21.7	20.6

Experiments: Results

VDR is better

VDR: A person is playing a saxophone.
BRNN: A man is playing a guitar

VDR: A person is playing a guitar.
BRNN: A man is jumping off a cliff

VDR: A person is playing a drum.
BRNN: A man is standing on a

BRNN is better

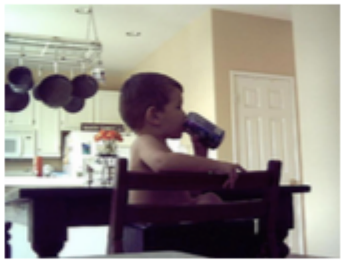
VDR: A person is using a computer.
BRNN: A man is jumping on a trampoline

VDR: A person is riding a horse.
BRNN: A group of people riding horses

VDR: A person is below sunglasses.
BRNN: A man is reading a book

Experiments: Results

Equally good



VDR: A person is sitting a table.
BRNN: A man is sitting on a chair

VDR: A person is using a laptop.
BRNN: A man is using a computer

VDR: A person is riding a horse.
BRNN: A man is riding a horse

Equally bad

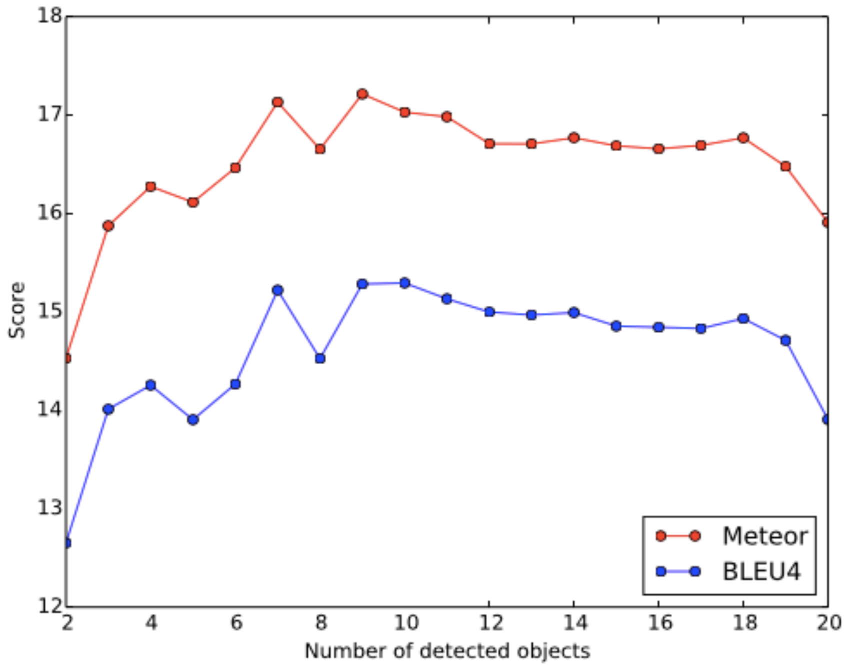
VDR: A person is holding a microphone.
BRNN: A man is taking a picture

VDR: A person is driving a car.
BRNN: A man is sitting on a phone

VDR: A person is driving a car.
BRNN: A man is riding a bike

Experiments: No of detected objects

- Improvements are seen until eight objects
  - *good descriptions do not always need the most confident detections*
- quality of the descriptions does not significantly decrease with an increased number of detected objects
  - *model formulation appropriately discards unsuitable detections*



Discussion and Conclusion

- Infers useful and reliable Visual Dependency Representations of images without expensive human supervision
- Uses these to generate image descriptions
- One of the main problem is detector's accuracy
- Changing the language model to n-gram might generate better/richer descriptions
- Quality of the generated text largely depended on the data set (better in people performing actions)
- Transferring model improved in the diverse data set