UNIVERSITY OF GOTHENBURG

Grounding language in action and perception

Simon Dobnik

31 European Summer School in Logic,
Language and Information
12 August 2019, Riga Latvia

CLASP centre for linguistic theory and studies in probability

Previously

UNIVERSITY OF GOTHENBURG

- Theoretical questions related to language, action and perception
- Examples of image descriptions and challenges for their computational modelling
- Semantics of spatial language and computational models
- Experiment for vision and language
- Example of an image captioning system using DNNs

Previously

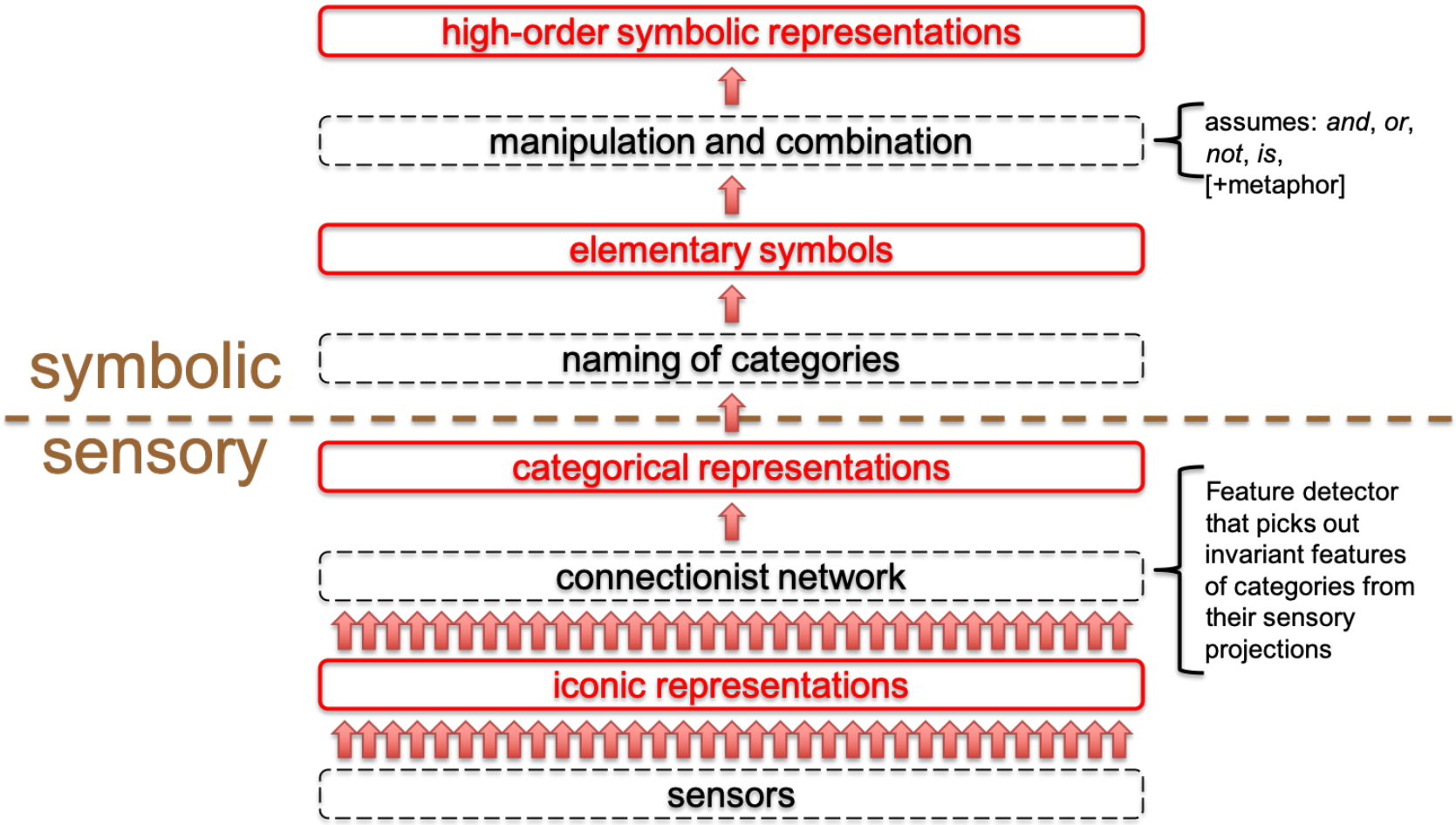
UNIVERSITY OF GOTHENBURG

- Theoretical questions related to language, action and perception
- Examples of image descriptions and challenges for their computational modelling
- Semantics of spatial language and computational models
- Experiment for vision and language
- Example of an image captioning system using DNNs
- TODAY: Computational systems for learning and generating grounded language

Outline

Words as classifiers

Perceptual grounding

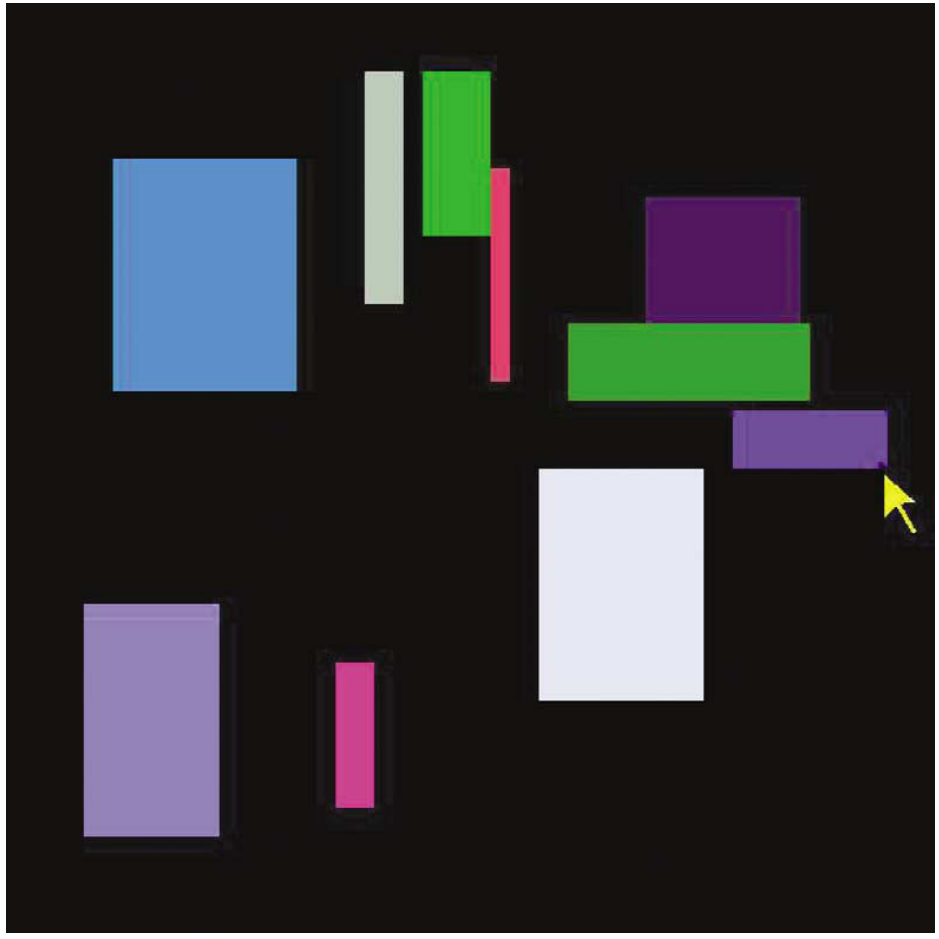


(Harnad, 1990) image from Kelleher (2010)

Words as classifiers

(Roy, 2002)

- A scene:



Training instances

- Vector of real-valued features representing the objects in the scene: r, g, b, hw_ratio, area, x, y mm_dimension
- Natural language descriptions
  - The pink square
  - The light blue square
  - The brightest green rectangle
  - The purple rectangle to the left of the pink square
  - The narrow purple rectangle below and to the right of the blue square

Learning

- Cluster words into classes (using probabilistic models):
  - Based on word distributions: words within a class co-occur infrequently with other words in that class
  - Their association with semantic features
  - A combination of both
- Statistical bi-gram model of classes (encodes word order constraints)

Generation

- For each bi-gram sequence of classes, for each class choose the most likely word given the target object:
  - the, the rectangle, the green rectangle, the large green rectangle, the large light green rectangle ...
- Estimate the fit of each description to the target object: the likelihood of a sequence of words to refer to the features of the object
- Contextual constraints and ambiguity of a description: $\psi(Q) = \mathit{fit}(x_{\mathit{target}}, Q) - \max_{\forall x \neq \mathit{target}} \mathit{fit}(x, Q)$
- Combine the scores from syntactic and contextual constraints with a weighted sum

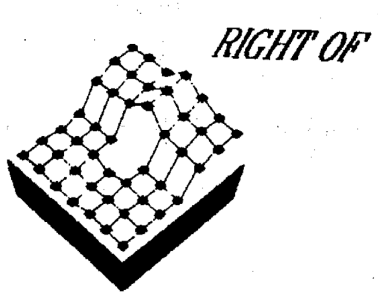
Generation

- For each bi-gram sequence of classes, for each class choose the most likely word given the target object:
  - the, the rectangle, the green rectangle, the large green rectangle, the large light green rectangle ...
- Estimate the fit of each description to the target object: the likelihood of a sequence of words to refer to the features of the object
- Contextual constraints and ambiguity of a description: $\psi(Q) = fit(x_{target}, Q) - \max_{\forall x \neq target} fit(x, Q)$
- Combine the scores from syntactic and contextual constraints with a weighted sum
- Evaluated by 3 human judges to select the target object given a description:
  - human-generated: 89.8%
  - machine-generated: 81.3%

Words as classifiers with robots

(Dobnik, 2009)

- Difficult to write a semantic model of spatial descriptions.

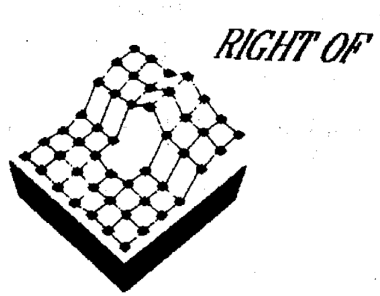


(Logan and Sadler, 1996)

Words as classifiers with robots

(Dobnik, 2009)

- Difficult to write a semantic model of spatial descriptions.



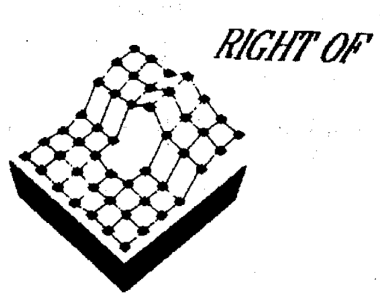
(Logan and Sadler, 1996)

- Difficult to pre-define a model of the world for a robot (cf. Shrdlu (Winograd, 1976))
- In SLAM (Dissanayake et al., 2001) the robot learns its environment incrementally through observations.

Words as classifiers with robots

(Dobnik, 2009)

- Difficult to write a semantic model of spatial descriptions.



(Logan and Sadler, 1996)

- Difficult to pre-define a model of the world for a robot (cf. Shrdlu (Winograd, 1976))
- In SLAM (Dissanayake et al., 2001) the robot learns its environment incrementally through observations.
- Let's learn language by combining observations of the environment with the ways humans describe it!

The robot

- Are robot's representations sufficient to learn spatial language?

The robot

UNIVERSITY OF GOTHENBURG

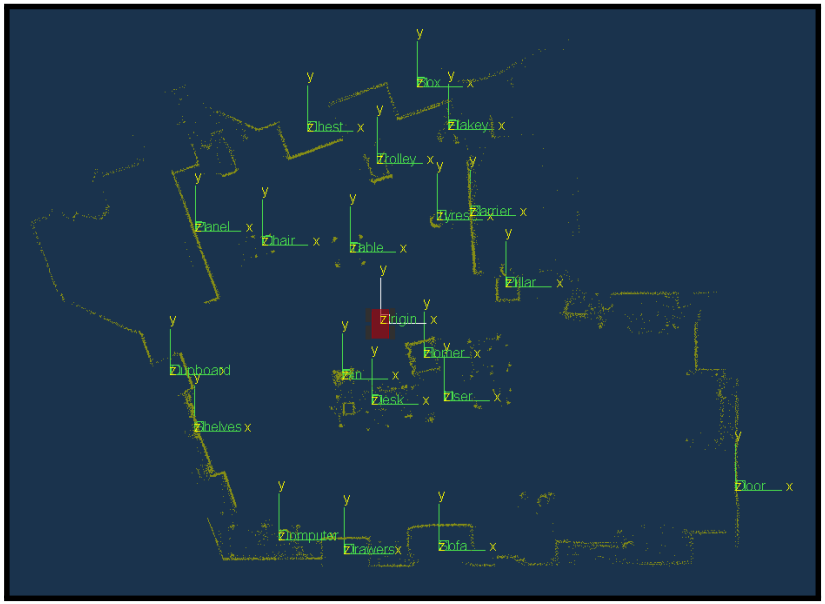
- Are robot's representations sufficient to learn spatial language?
- ATRV-JR mobile robot (iRobot) primarily used for tasks such as map building, localisation and navigation and runs the following components:
  - odometry component: provides information about the robot's motion, for example ⟨R-Heading⟩ and ⟨Speed⟩;
  - SLAM component: localises the robot on a 2-dimensional map consisting of a set of points relative to some random starting point, for example ⟨0.6234,0.2132⟩.

From sensors to spatial geometry

UNIVERSITY OF GOTHENBURG

Human

Robot



Film: 3D laser point cloud generation

CLASP centre for linguistic theory and studies in probability

Creation of learning examples

UNIVERSITY OF GOTHENBURG

- A robot an a human describer situated in a room.
- The robot is guided manually by another person.
- Human describers (4) freely generate descriptions from the perspective of the robot (speech recognition).
- All linguistic and non-linguistic observations are logged.

Linguistic descriptions

UNIVERSITY OF GOTHENBURG

Descriptions of robotic motion

- You're going forward slowly.
- Now you're turning right.

Descriptions of relations between objects

- The chair is to the left of you.
- The table is further away than the chair.

Supervised off-line learning

- Automatic extraction of useful information from datasets
- Matching observations in time . . . noise
- A lot of perceptual data for few descriptions

Supervised off-line learning

- Automatic extraction of useful information from datasets
- Matching observations in time . . . noise
- A lot of perceptual data for few descriptions
- Some extracted features:

R-Heading	Speed	Verb
0.001	0.234	moving
0.535	0.122	turning
0.123	-0.364	reversing
. . .		

LO_x	LO_y	REFO_x	REFO_y	Relation
0.632	0.536	0.001	0.321	to_the_right_of
0.212	0.447	0.346	0.342	to_the_right_of
0.573	0.731	0.564	0.632	near
. . .				

Applying the classifiers

UNIVERSITY OF GOTHENBURG

Two interactive systems:

- pDescriber: generates descriptions (of objects, of robot's motion)

CLASP centre for linguistic theory and studies in probability

Applying the classifiers

Two interactive systems:

- pDescriber: generates descriptions (of objects, of robot's motion)
- pDialogue: generates motion and answers user's questions

M Motion requests: Go forward slowly. Go forward right fast.

A Locating objects: Where is the table? - The table is to the left of the chair? Where are you? - I'm behind the sofa.

B Confirming object description: Is the table to the left of the chair? - Yes, the table is to the left of the chair. - No, the table is near the chair.

C Finding objects: What is to the left of the chair? - The pillars, the tyres and the wall are to the left of the chair.

D Referencing an object: What is the chair to the left of? - The chair is to the left of the table, the desk and the wall.

Was learning successful?

- Performance of the classifiers vs subjective opinion
- One context vs different contexts
- New room, 5 subjects (pDescriber), 13 subjects (pDialogue)
  - pDescriber: Is this a good description? - Yes/No.
  - pDialogue: How natural is the answer? - 1 to 5.

Question type	Accuracy (%)	Classifier	Accuracy (%)
pDescriber	59.28	relation	69.12
A	43.51	relation	69.12
B	54.17	relation	69.12
C	54.70	lo_x	48.80
		lo_y	72.80
D	56.92	refo_x	65.60
		refo_y	82.24
Mean	52.33		67.71

Words as classifiers + formal representations

Connecting grounded and formal representations

UNIVERSITY OF GOTHENBURG

- (Harnad, 1990): language in a domain of symbolic computation
- Not true but there are benefits of this view from the computational perspective
- Top-down filtering of bottom-up induced knowledge
- "Deeper" cognitively inspired representations vs representations induced from patterns
- Can examine and interpret the beliefs obtained
- Possibilities of using techniques for logical inference with machine learning methods
- The need for an over-arching framework for perception and language

(Dobnik and Kelleher, 2017)

NL, formal language and grounding

UNIVERSITY OF GOTHENBURG

Parsing natural language to a robot control language and grounding it in action and perception (Matuszek et al., 2012b)

go to	the	second	junction	and	go left
S/NP	NP/NP	NP/N	N	S\S/S	S
(move-to forward)	[null]	(do-n-times 2 x)	(until (junction current-loc) y)	(do-seq g f)	(turn-left)
		NP		S\S	
		(do-n-times 2 (until (junction current-loc) y))		(do-seq g turn-left)	
	NP				
	(do-n-times 2 (until (junction current-loc) y))				
S					
(do-n-times 2 (until (junction current-loc) (move-to forward)))					
S					
(do-seq (do-n-times 2 (until (junction current-loc) (move-to forward))) (turn-left))					

CLASP centre for linguistic theory and studies in probability

Type Theory with Records

UNIVERSITY OF GOTHENBURG

Type Theory with Records (Cooper, in prep; Dobnik et al., 2013; Larsson, 2015; Cooper et al., 2015; Dobnik and Cooper, 2017)

- meaning relative to agent
- judgements
- of situations, of speech events (and of neural events)
- meaning representations as record types (and a few basic types)
- types of perceptual readings to types of dialogue game-boards
- types are cognitive and intensional

Types of objects

(Dobnik and Cooper, 2017)

Perceptual domain

- [[34,24,48],[56,78,114]...]: *PointMap*
  PointMap $\sqsubseteq$ list(list(Real))

Conceptual domain

Types of objects

(Dobnik and Cooper, 2017)

Perceptual domain

- [[34,24,48],[56,78,114]. . .]: *PointMap*
  PointMap $\sqsubseteq$ list(list(Real))
- Object detection function
  $(Pointmap \rightarrow \text{set}(\begin{bmatrix} \text{reg} & : & Pointmap \\ \text{pfun} & : & (Ind \rightarrow Type) \end{bmatrix}))$
  pfun $= \lambda x{:}Ind.\text{chair}(x)$

Conceptual domain

Types of objects

(Dobnik and Cooper, 2017)

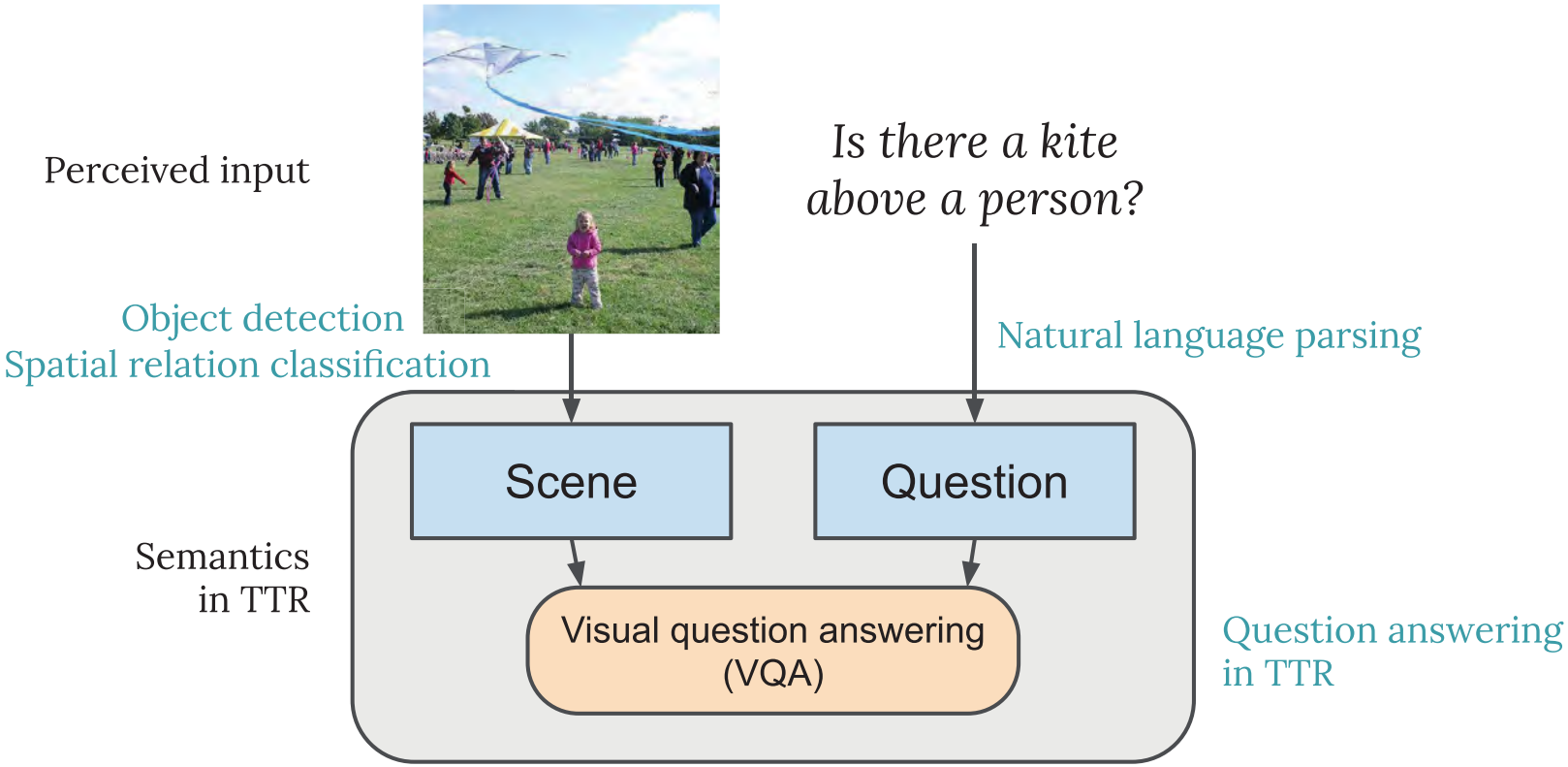
Perceptual domain

- [[34,24,48],[56,78,114]. . .]: *PointMap*
  PointMap $\sqsubseteq$ list(list(Real))
- Object detection function
  $(\mathit{Pointmap} \rightarrow \text{set}(\begin{bmatrix} \text{reg} & : & \mathit{Pointmap} \\ \text{pfun} & : & (\mathit{Ind} \rightarrow \mathit{Type}) \end{bmatrix}))$
  pfun $= \lambda x{:}\mathit{Ind}.\text{chair}(x)$
- Individuation function
  $\lambda r{:}\begin{bmatrix}\text{reg}{:}\mathit{Pointmap} \\ \text{pfun}{:}(\mathit{Ind} \rightarrow \mathit{Type})\end{bmatrix} . \begin{bmatrix} \text{a} & : & \mathit{Ind} \\ \text{loc} & : & \text{location(a, } r.\text{reg)} \\ \text{c} & : & r.\text{pfun(a)} \end{bmatrix}$

Conceptual domain

ImageTTR

(Matsson, Dobnik, and Larsson, 2019): pyTTR for Visual Question Answering (VQA)



https://github.com/arildm/imagettr

Question answering

A polar question as a subtype check

Scene ⊑ *Question*

$$
\begin{bmatrix}
\mathrm{loc}_0 : \begin{bmatrix} \mathrm{w} : \mathit{Int} \\ \mathrm{cy} : \mathit{Int} \\ \mathrm{cx} : \mathit{Int} \\ \mathrm{h} : \mathit{Int} \end{bmatrix} \begin{bmatrix} \mathrm{w} = 422 \\ \mathrm{cy} = 355 \\ \mathrm{cx} = 435 \\ \mathrm{h} = 242 \end{bmatrix} \\
\mathrm{y}_4 : \mathit{Ind}_{a_0} \\
\mathrm{cp}_1 : \langle \lambda a : \mathit{Ind} \,.\, \mathrm{person}(a), [x_9] \rangle \\
\mathrm{cr}_3 : \langle \lambda a : \mathit{Ind} \,.\, \lambda b : \mathit{Ind} \,.\, \mathrm{right}(a,b), [y_4, x_{11}] \rangle \\
\mathrm{x}_{11} : \mathit{Ind}_{a_2} \\
\mathrm{cp}_2 : \langle \lambda a : \mathit{Ind} \,.\, \mathrm{dog}(a), [x_{11}] \rangle \\
\mathrm{cr}_6 : \langle \lambda a : \mathit{Ind} \,.\, \lambda b : \mathit{Ind} \,.\, \mathrm{above}(a,b), [x_9, y_4] \rangle \\
\mathrm{x}_9 : \mathit{Ind}_{a_1} \\
\mathrm{cl}_1 : \langle \lambda k : \mathit{Ind} \,.\, \lambda l : \begin{bmatrix} \mathrm{w} : \mathit{Int} \\ \mathrm{cy} : \mathit{Int} \\ \mathrm{cx} : \mathit{Int} \\ \mathrm{h} : \mathit{Int} \end{bmatrix} \,.\, \mathrm{location}(k,l), [x_9, \mathit{loc}_1] \rangle \\
\mathrm{cr}_5 : \langle \lambda a : \mathit{Ind} \,.\, \lambda b : \mathit{Ind} \,.\, \mathrm{right}(a,b), [x_9, x_{11}] \rangle \\
\mathrm{cr}_0 : \langle \lambda a : \mathit{Ind} \,.\, \lambda b : \mathit{Ind} \,.\, \mathrm{below}(a,b), [y_4, x_9] \rangle \\
\begin{bmatrix} \mathrm{w} : \mathit{Int} \end{bmatrix}
\end{bmatrix}
$$

?

$$
\sqsubseteq
$$

$$
\begin{bmatrix}
\mathrm{x} : \mathit{Ind} \\
\mathrm{y} : \mathit{Ind} \\
\mathrm{c}_0 : \mathrm{kite}(\mathrm{x}) \\
\mathrm{c}_1 : \mathrm{person}(\mathrm{y}) \\
\mathrm{c}_2 : \mathrm{above}(\mathrm{x}, \mathrm{y})
\end{bmatrix}
$$

Grounding a language model

Grounding bottom-up without formal representations

(Ghanimifard and Dobnik, 2017)

- Compositionality as a strength of formal grammars
- A probabilistic/neural language model learns the associations between words in a sequence
- Is bottom-up grounding compositional?
  - Composed strings of words
  - Composed perceptual representations
- Grounded language model
  $Pr(w_{1:T}|c) = \prod_{t=1}^{T} Pr(w_t|w_{1:t}, c)$
  where a word sequence $w_{1:T}$ is a description of an image with the visual features c.

Artificial dataset

- Average acceptability/probability scores over locations (Logan and Sadler, 1996)
- $freq(w_{1:T}, c) = n \times Pr(w_{1:T}, c)$
- Artificial composition
  - Simple language with connectives

    $g_\wedge : (v_i, v_j) \rightarrow [v_i, \text{“and”}, v_j]$

    $g_\vee : (v_i, v_j) \rightarrow [\text{“}either\text{”}, v_i, \text{“or”}, v_j]$

    $g_\neg : v \rightarrow [\text{“not”}, v]$
  - Language with “distractor” words

    $g_1 : (v*) \rightarrow [v*]$

    $g_2 : (v*) \rightarrow [\text{“}it\text{”}, \text{“}is\text{”}, v*]$

    $g_3 : (v*) \rightarrow [\text{“}it\text{”}, \text{“}is\text{”}, v*, \text{“}the\text{”}, \text{“}box\text{”}]$

    $g_4 : (v*) \rightarrow [\text{“}the\text{”}, \text{“}ball\text{”}, \text{“}is\text{”}, v*, \text{“}the\text{”}, \text{“}box\text{”}]$

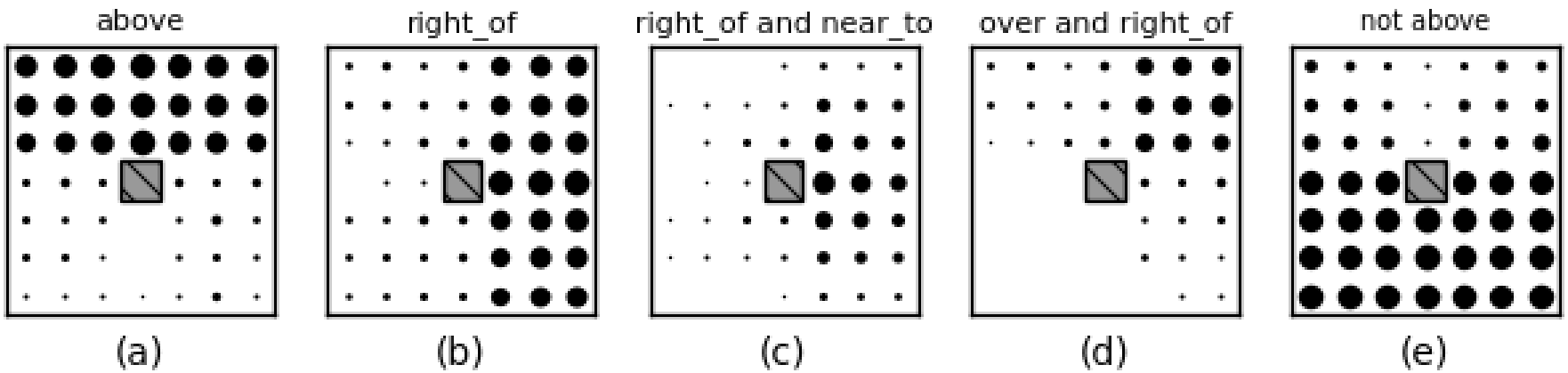
    $g_5 : (v*) \rightarrow [\text{“}the\text{”}, \text{“}object\text{”}, \text{“}is\text{”}, v*, \text{“}the\text{”}, \text{“}box\text{”}]$
  - Locations

    $\hat{s}_{g_\wedge(v_i,v_j),c} = s_{v_i,c} \times s_{v_j,c}$

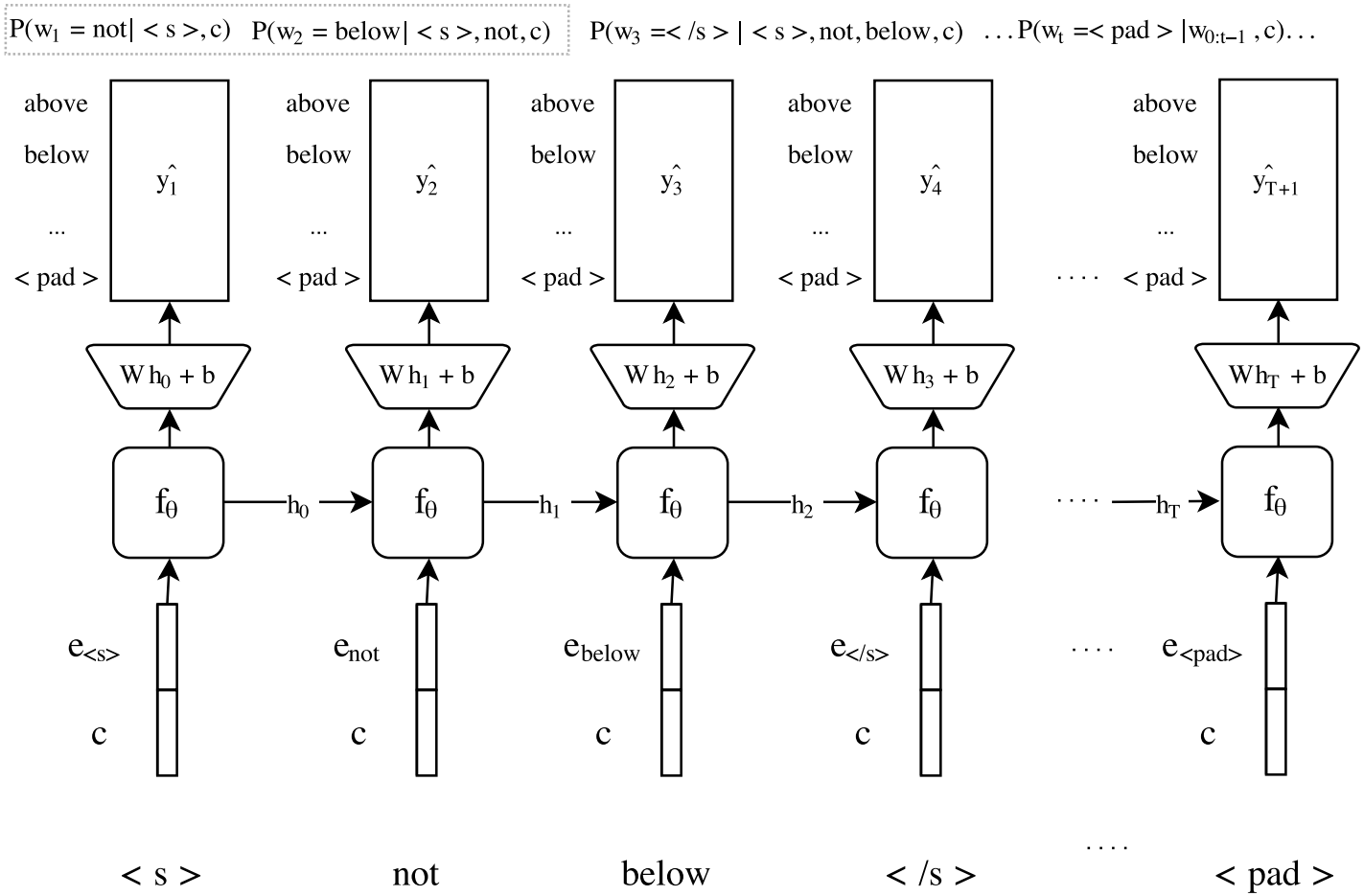
    $\hat{s}_{g_\vee(v_i,v_j),c} = s_{v_i,c} + s_{v_j,c} - s_{v_i,c} \times s_{v_j,c}$

    $\hat{s}_{g_\neg(v),c} = 1 - s_{v,c}$

Generated templates



Grounded neural language model



Evaluation: composition

The probability of the generated sequence for a particular location is a judgement score

$$
\begin{aligned}
T_{w_{1:T}} &= \{Score_{w_{1:T},c}\}_{c \in L} \\
\hat{T}_{w_{1:T}} &= \{Pr(w_{1:T}|c)\}_{c \in L} \\
\rho(T_{w_{1:T}}, \hat{T}_{w_{1:T}}) & \quad \text{Spearman's rank correlation coefficient}
\end{aligned}
$$

Evaluation: composition

Performance on different datasets

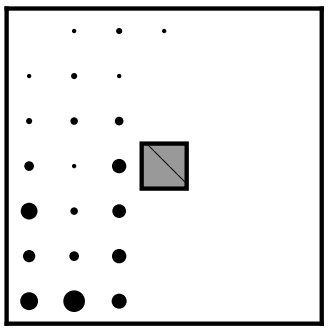
Descriptions	-0%	+Distractors	-10%	-20%	-30%	-80%
One word	0.92	0.91				
Phrases	0.83	0.84				
AND-phrases*	0.87	0.85	0.84	0.80	0.78	0.53
OR-phrases*	0.79	0.80	0.74	0.73	0.69	0.38
NEG-phrases*	0.72	0.82				

*Contains single words and their negations

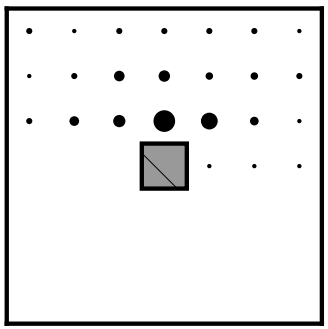
Some examples of new compositions

-50%

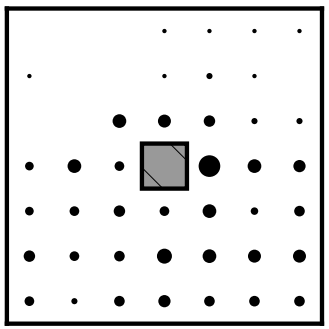
above and left_of



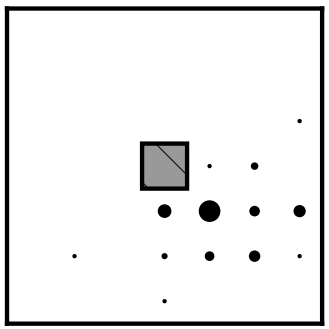
next_to and above



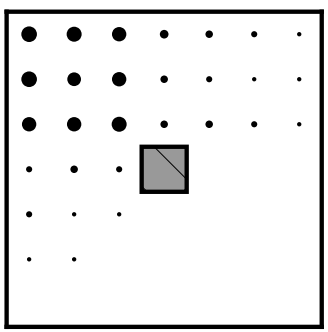
either far_from or next_to



below and right_of

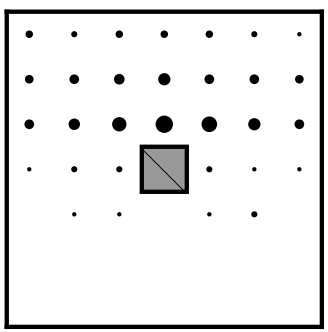


above and left_of



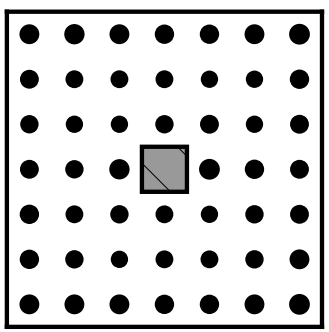
$\rho = 0.50$

next_to and above



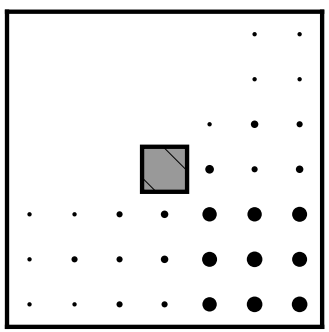
$\rho = 0.91$

either far_from or next_to



$\rho = -0.05$

below and right_of



$\rho = 0.80$

Evaluation: decomposition

Trained on all descriptions with some removed

Removed	-0%	-10%	-20%	-30%	-40%
–AND-phrases*	0.83	0.86	0.80	0.77	0.81
–NEG-phrases*	0.83	0.83	0.64	0.59	0.43
–OR-phrases*	0.83	0.73	0.78	0.68	0.69
–One word	0.92	0.90	0.90	0.84	0.87

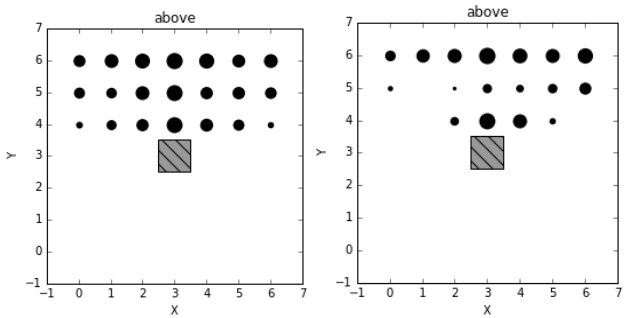
* also excludes single words and their negation

Evaluation: decomposition

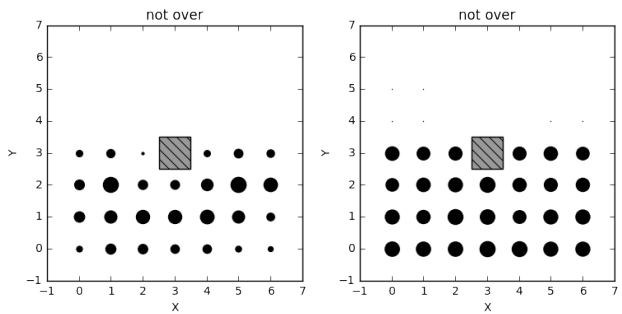
Trained on all descriptions with some removed

Removed	-0%	-10%	-20%	-30%	-40%
–AND-phrases*	0.83	0.86	0.80	0.77	0.81
–NEG-phrases*	0.83	0.83	0.64	0.59	0.43
–OR-phrases*	0.83	0.73	0.78	0.68	0.69
–One word	0.92	0.90	0.90	0.84	0.87

* also excludes single words and their negation



above, $\rho = 0.8$



not over, $\rho = 0.77$

Tutorial #2

CLASP centre for linguistic theory and studies in probability

Tutorial #2: Learning to compose

by Mehdi Ghanimifard

In this tutorial we look at the code that was used for the paper. You can:

- adjust the descriptions and the spatial templates used to generate the artificial dataset
- remove individual words or composed phrases from the training dataset
- train and evaluate the models on them

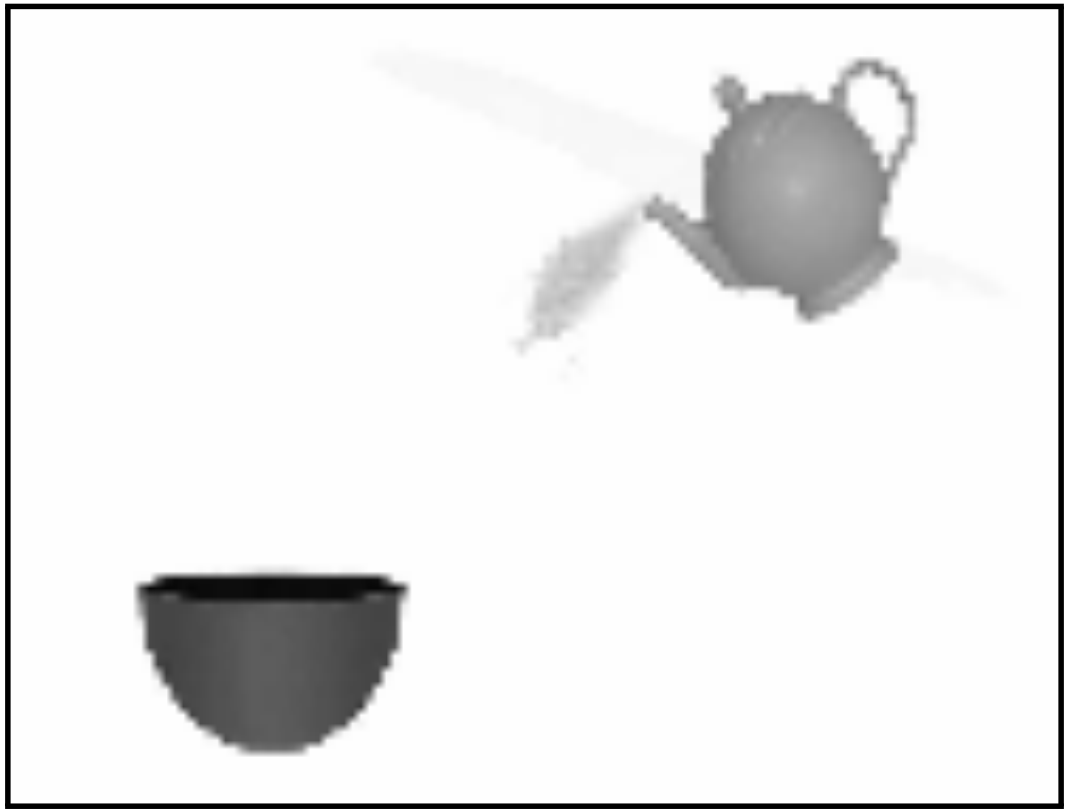
Code on Github

The power of a language model

Interacting objects

UNIVERSITY OF GOTHENBURG

Coventry et al. (2005)



Degrees of contributions

The umbrella is over/above the man.

Coventry et al. (2001)

Extracting knowledge about object interaction

UNIVERSITY OF GOTHENBURG

- Encoded in the language model, cf. the success of distributional semantics
- Use the predictions as a filter in description generation
- Predict the bias of a spatial relation to functional or geometric knowledge:
  - A functional spatial relation is more selective of their target and landmark objects
  - A geometric relation will occur with any kind of objects.

(Dobnik and Kelleher, 2013, 2014)

Corpora of image descriptions

a yellow building with white columns in the background; two palm trees in front of the house; cars parked in front of the house; a woman and a child are walking over the square;

Choosing a relation

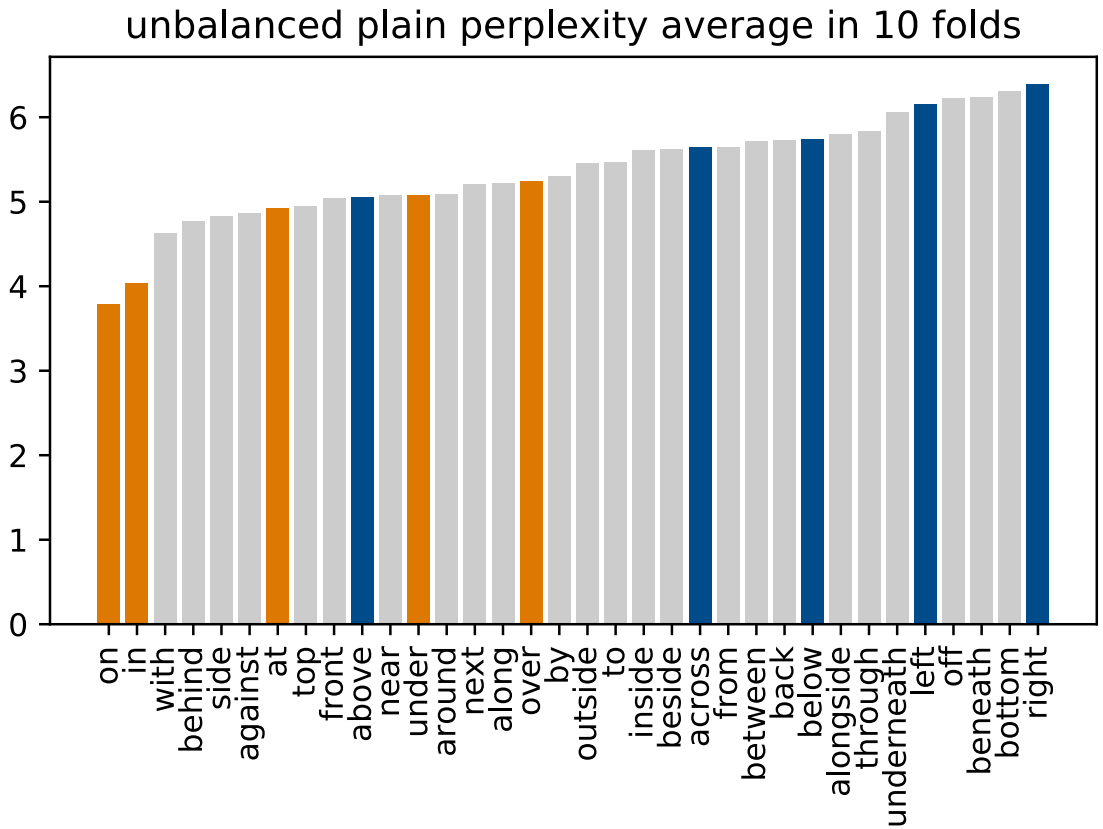
FG	Prep	$-2log\lambda$	H_2 vs. H_1
people*square	on	655.66*	2.37×10^{142}
people*square	in	133.63*	1.04×10^{29}
people*square	at	1.81	2.47
people*umbrella	with	16.06*	3076.878
boy*umbrella	under	12.16*	436.788
table*umbrella	under	9.39*	109.447
child*umbrella	under	8.35*	65.006
sculpture*umbrella	with	6.88*	31.25
woman*umbrella	with	6.83*	30.428
woman*umbrella	under	6.78*	29.592
girl*umbrella	with	4.59*	9.921
man*umbrella	with	2.29	3.15
child*umbrella	with	1.53	2.153

*: $p < 0.05$

(Normalised) entropy and object variation

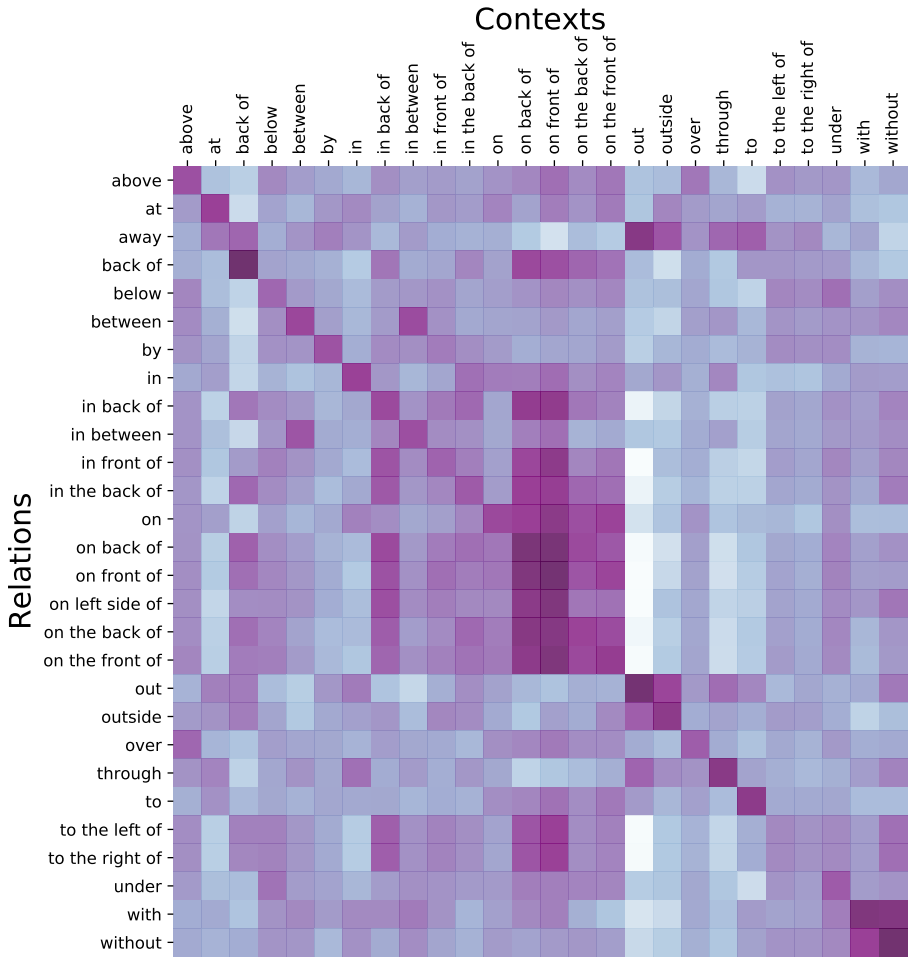
#	Preposition	FG-Types	Tokens	Norm FG ent
1	on_left_side_of	5	31	0.35448
2	underneath	31	74	0.65535
3	in	7584	34846	0.6714
4	onto	49	86	0.79109
5	down	83	142	0.81099
6	over	440	736	0.83106
7	at	1393	2726	0.83148
8	on_top_of	61	87	0.83409
9	against	50	68	0.85171
10	on	4897	10085	0.852
11	on_side_of	46	63	0.87644
...	...	...	...	...
15	on_back_of	9	11	0.89489
16	through	179	245	0.89738
17	in_front_of	1278	1938	0.90998
...	...	...	...	...
22	under	167	220	0.92096
23	above	145	190	0.9228
...	...	...	...	...
26	below	13	14	0.96248
...	...	...	...	...

Neural language models and perplexity



(Dobnik, Ghanimifard, and Kelleher, 2018)

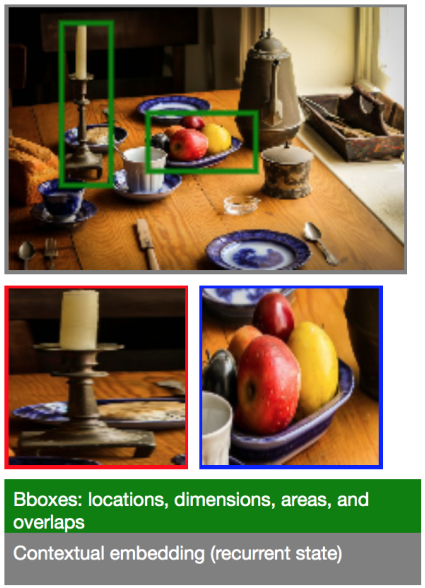
Perplexities by NLMs as vectors

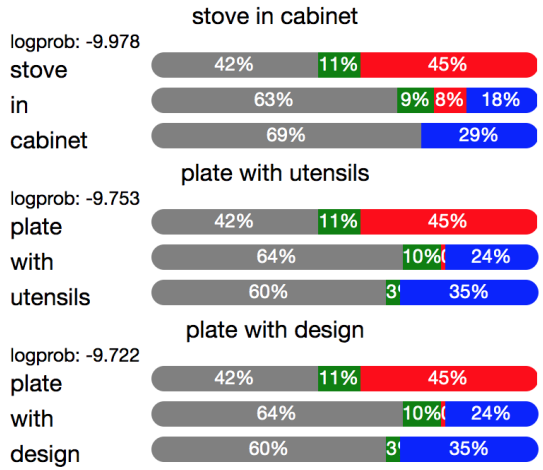


(Ghanimifard and Dobnik, 2019)

When language model takes over

Demo from yesterday





(Ghanimifard and Dobnik, 2018b)

On the linguistic bias of vision and language datasets (Agrawal et al., 2017)

Tutorial #3

CLASP centre for linguistic theory and studies in probability

Tutorial #3: bottom-up grounding of vision and language

by Mehdi Ghanimifard

In this tutorial we look at the extension of the code from the previous tutorial that replaces simple spatial locations with visual features that are also trained from the data. It will examine

- how the visual features are trained, represented and used in the model
- how objects are detected
- how image descriptions are generated
- limitations and further work in generating image descriptions

Code on Github

Summary

Summary

- Grounding words classifiers
- Integration of grounded words with a language model
  - probabilistic
  - rule-based
- End-to-end grounding with a grounded neural language model
  - Compositionality?
  - Information encoded in language alone
- End-to-end grounding of visual features and a language model

References I

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2017. Don't just assume; look and answer: Overcoming priors for visual question answering. *arXiv*, arXiv:1712.00377 [cs.CV]:1–15.

Robin Cooper. in prep. Type theory and language: from perception to linguistic communication. Draft at `https://sites.google.com/site/typetheorywithrecords/drafts`.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2015. Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology — LiLT*, 10(4):1–43.

Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, volume 3343 of *Lecture Notes in Computer Science*, pages 98–110. Springer Berlin Heidelberg.

References II

Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the apprehension of Over, Under, Above and Below. *Journal of Memory and Language*, 44(3):376–398.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

M. W. M. G Dissanayake, P. M. Newman, H. F. Durrant-Whyte, S. Clark, and M. Csorba. 2001. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotic and Automation*, 17(3):229–241.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik and Robin Cooper. 2017. Interfacing language, spatial perception and cognition in Type Theory with Records. *Journal of Language Modelling*, 5(2):273–301.

References III

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2013. Modelling language, action, and perception in type theory with records. In Denys Duchier and Yannick Parmentier, editors, *Constraint Solving and Language Processing - 7th International Workshop on Constraint Solving and Language Processing, CSLP 2012, Orleans, France, September 13-14, 2012. Revised Selected Papers*, number 8114 in Publications on Logic, Language and Information (FoLLI). Springer, Berlin, Heidelberg.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

References IV

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference*, pages 1–6, Berlin, Germany.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third V&L Net Workshop on Vision and Language*, pages 33–37, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

Simon Dobnik and John D. Kelleher. 2017. Modular mechanistic networks: On bridging mechanistic and phenomenological models with deep neural networks in natural language processing. In *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017), Gothenburg, 12–13 June 2017*, volume 1 of *CLASP Papers in Computational Linguistics*, pages 1–11, Gothenburg, Sweden. Department of Philosophy, Linguistics and Theory of Science

References V

(FLOV), University of Gothenburg, CLASP, Centre for Language and Studies in Probability.

Raquel Fernández. 2013. Rethinking overspecification in terms of incremental processing. In *Proceedings of the PRE-CogSci 2013 Workshop on the Production of Referring Expressions*.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics*, pages 1–12, Montpellier, France. Association for Computational Linguistics.

Mehdi Ghanimifard and Simon Dobnik. 2018a. Knowing when to look for what and where: Evaluating generation of spatial descriptions with adaptive attention. In *Computer Vision – ECCV 2018 Workshops. ECCV 2018*, volume 11132 of *Lecture Notes in Computer Science (LNCS)*, pages 1–9, Proceedings of the Workshop on Shortcomings in Vision and Language (SiVL), ECCV 2018, Munich, Germany. Springer, Cham.

References VI

Mehdi Ghanimifard and Simon Dobnik. 2018b. Visual grounding of spatial relations in recurrent neural language models. In *Proceedings of the 3rd Workshop on Models and Representations in Spatial Cognition (MRSC-3) at 11th International Conference on Spatial Cognition 2018*, pages 1–7, Tübingen, Germany. https://dobnik.net/simon/events/mrsc-3/#programme.

Mehdi Ghanimifard and Simon Dobnik. 2019. What a neural language model tells us about spatial relations. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP) at NAACL-HLT 2019*, pages 1–13, Minneapolis, Minnesota, USA. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Association for Computational Linguistics.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42(1–3):335–346.

References VII

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

John D. Kelleher. 2010. How to preposition a robot: A case-study in symbol grounding. Presentation at the University College Dublin Research Seminars.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017), Gothenburg, 12 –13 June*, volume 1 of *CLASP Papers in Computational Linguistics*, pages 41–52, Gothenburg, Sweden. Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, CLASP, Centre for Language and Studies in Probability.

Emiel Krahmer and Kees van Deemter. 2011. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

References VIII

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. IEEE.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv*, arXiv:1612.01887 [cs.CV]:1–10.

Arild Matsson, Simon Dobnik, and Staffan Larsson. 2019. ImageTTR: Grounding type theory with records in image classification for visual question answering. In *Proceedings of the IWCS 2019 Workshop on Computing Semantics with Types, Frames and Related Structures*, pages 55–64, Gothenburg, Sweden. Association for Computational Linguistics.

References IX

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012a. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, Edinburgh, Scotland.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. 2012b. Learning to parse natural language commands to a robot control system. In *Proceedings of the 13th International Symposium on Experimental Robotics (ISER)*.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2013. Generating expressions that refer to visible objects. In *HLT-NAACL*, pages 1174–1184.

Deb Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer speech and language*, 16(3):353–385.

Andrea Vedaldi. 2016. Convolutional networks for computer vision applications. IV&L summer school on vision and language, Malta. http://www.robots.ox.ac.uk/∼vedaldi/assets/teach/vedaldi16deepcv.pdf.

References X

Terry Winograd. 1976. *Understanding Natural Language*. Edinburgh University Press.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv*, arXiv:1502.03044 [cs.LG]:1–22.

Referring to what matters

From (Fernández, 2013)

Generating referring expressions (GRE)

UNIVERSITY OF GOTHENBURG

The Incremental Algorithm, Dale and Reiter (1995)

- Each target object is associated with certain properties, e.g. Colour, Type and Position
- Each target object is assigned a set of distractor objects having some property in common
- An unambiguous referring expression matches the target object but none of the distractors
- Properties are assigned a preference order ⟨Colour, Type, Position⟩ based on their salience in that domain
- Add properties in this order to the description plan iff a property has a discriminatory power (reduces the set of distractor objects) at that point in the order: "the red"
- Stop if the description is uniquely identifies the target: "the red chair"

Not just visual properties

Other factors may influence the amount of information speakers include in a referring description:

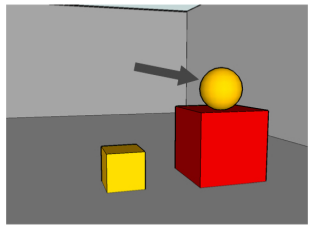
- Complexity of the domain: the number of properties
- Cardinality of the target: plural targets are more likely to be over-specified
- Cross-linguistic differences
- Dialogue context and intent

Further reading on GRE: (Krahmer and van Deemter, 2011)

Rule-based generation (Mitchell et al., 2013) I

- The datasets

TUNA corpus



GRE3D3 corpus

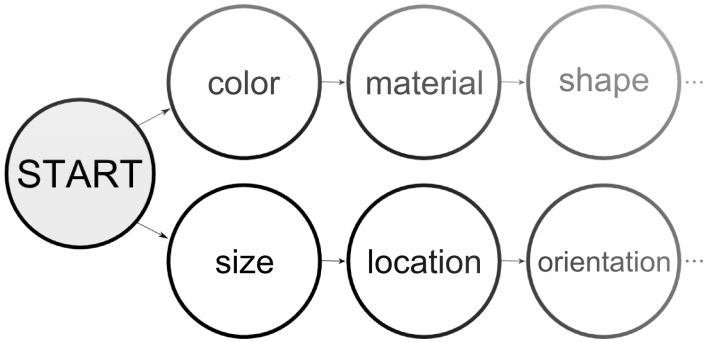
- Represented as properties

tg	color:yellow	size:(63,63)	type:ball	loc:right-hand
lm	color:red	size:(345,345)	type:cube	loc:right-hand
obj3	color:yellow	size:(70,70)	type:cube	loc:left-hand

obj1	colour:grey	size:(454,454)	type:desk	loc:(3,1)	ori:front
obj2	colour:blue	size:(454,454)	type:desk	loc:(2,1)	ori:front
obj3	colour:red	size:(454,454)	type:desk	loc:(3,2)	ori:back
obj4	colour:green	size:(254,254)	type:desk	loc:(4,1)	ori:left
obj5	colour:blue	size:(454,454)	type:fan	loc:(1,1)	ori:front
obj6	colour:red	size:(454,454)	type:fan	loc:(5,1)	ori:back
obj7	colour:green	size:(254,254)	type:fan	loc:(2,2)	ori:left

Rule-based generation (Mitchell et al., 2013) II

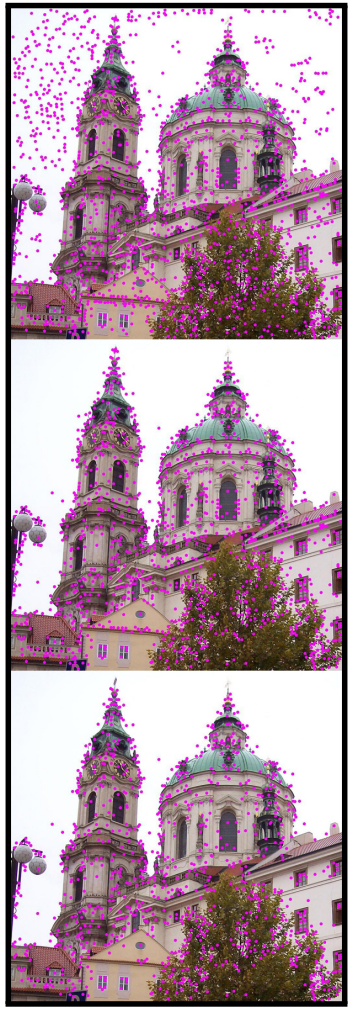
- Two categories of properties:
  absolute ("what") and relative ("where")



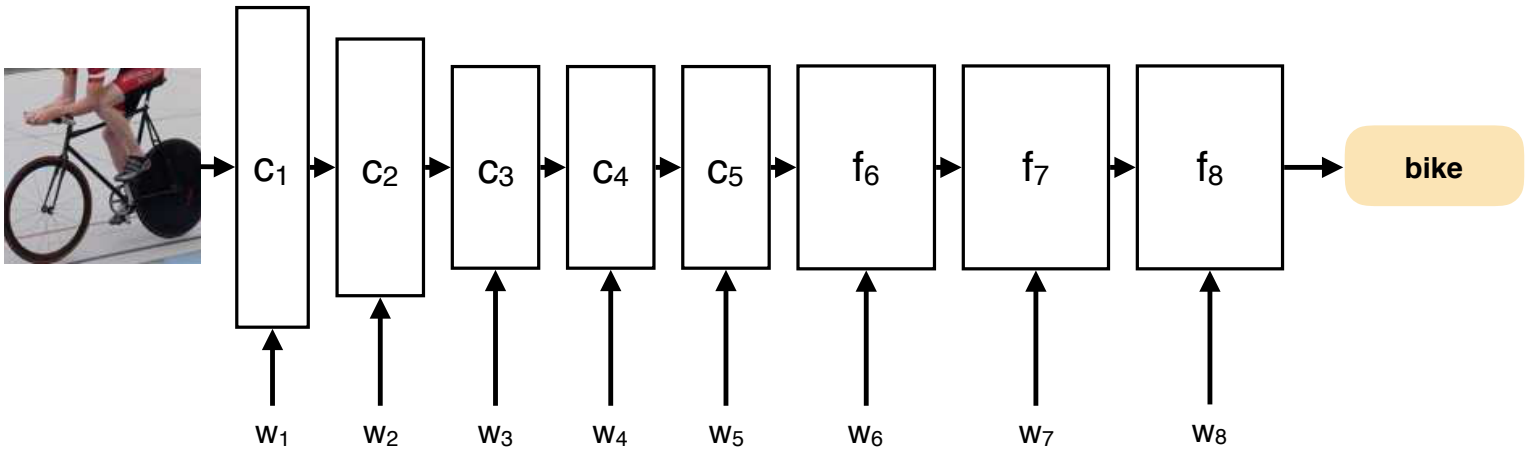
- Process through this graph of attributes, calculating the likelihood of generating a property based on its prior α_{att} and a description length penalty γ
  $f(A \cup \{x\}) = \gamma\alpha_{att}$
- Add the property if $f > n$ where n is random $0 \leq n \leq 1$
- Scan through objects; if there are more objects like the referent object, generate properties that distinguish them constrained by γ

Engineering visual features

- Scale Invariant Image Transform (SIFT) features (Lowe, 1999)
- Creating visual words



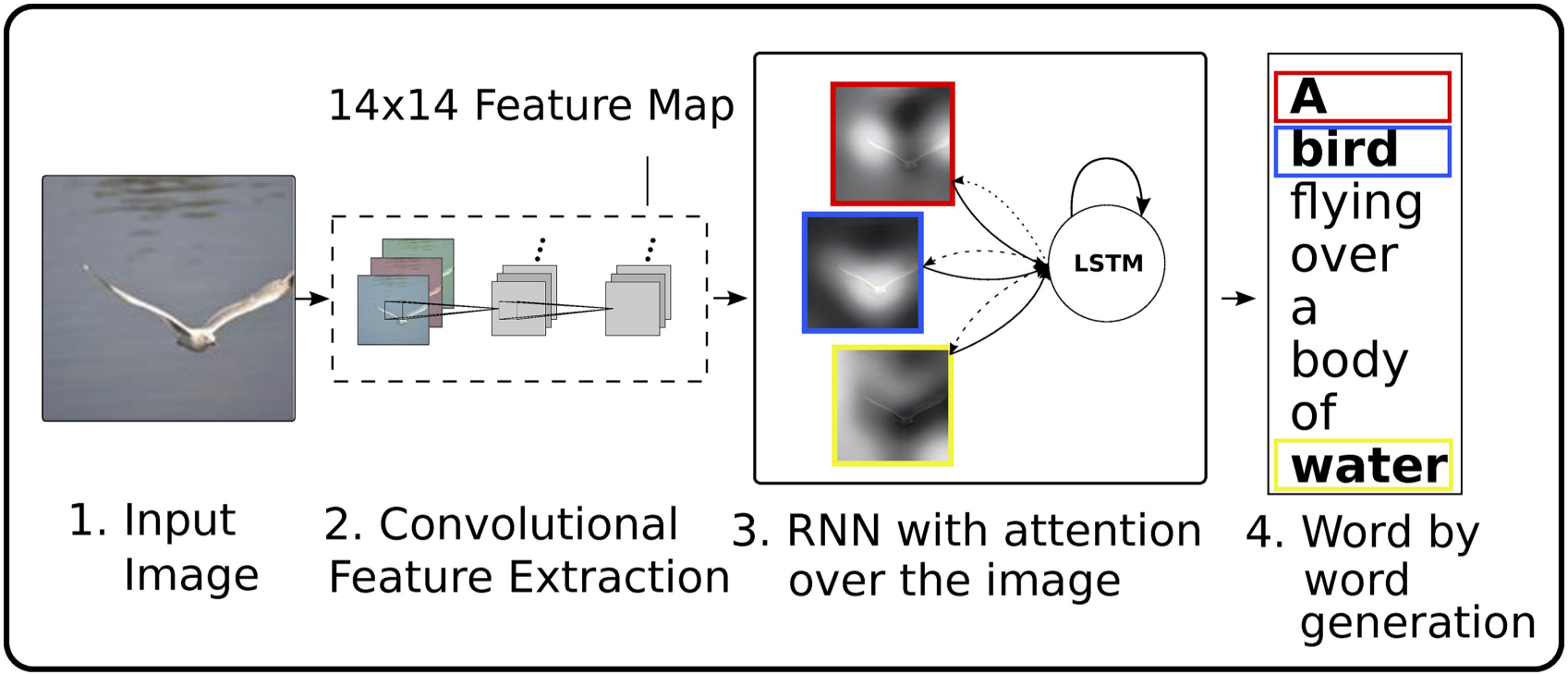
Learning visual features with CNNs



A. Krizhevsky, I. Sutskever, and G. E. Hinton. *Imagenet classif cation with deep convolutional neural networks*. In Proc. NIPS, 2012.

Image from Vedaldi (2016)

Generation of image descriptions with deep learning



(Xu et al., 2015)

Attention

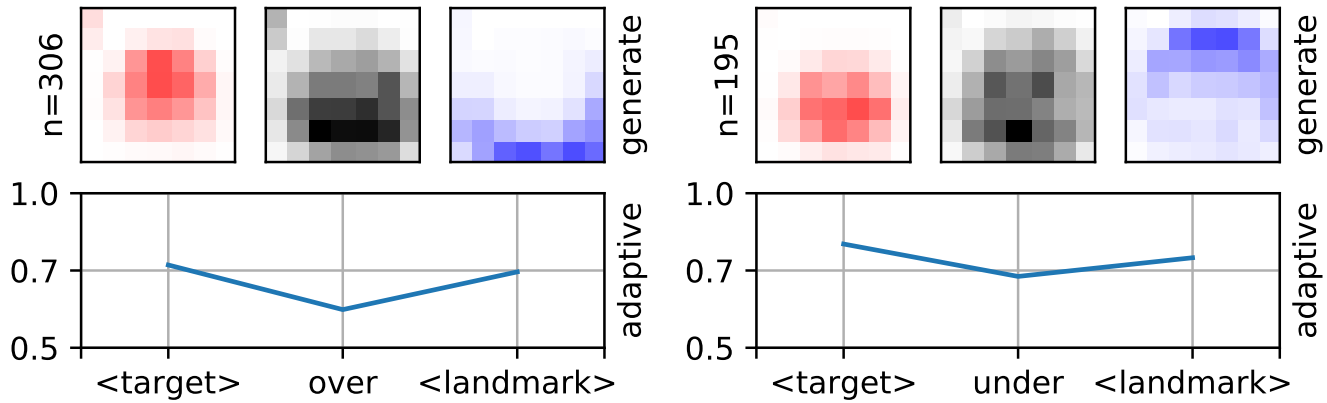
(Xu et al., 2015; Lu et al., 2016)

- Align visual features with words
- Combine the image features V and the hidden state h_t of the LSTM through a single layer followed by a softmax
  $z_t = w_h^T tanh(W_v V + (W_g h_t)\mathbb{1}^T)$
  $\alpha_t = softmax(z_t)$
- $c_t = \sum_{i=1}^{k} \alpha_{ti} v_i$
  - c_t: attended visual features at time t
  - i: a region of k regions of an image
  - v_i: visual representation of a region i
  - α_{ti}: the attentional weight to the region i
- c_t and h_t are combined to predict the next word y_{t+1}
- Adaptive attention (Lu et al., 2016):
  $\hat{c}_t = \beta_t s_t + (1 - \beta_t) c_t$
  s_t is obtained from the memory state of the language model

Attention and different description types

Kelleher and Dobnik (2017); Ghanimifard and Dobnik (2018a)

- Attention as spatial templates?



- Visual attention is high in general: higher with objects than relations
- Spatial relations depend more on the language model
- Spatial relations are attended in less focused way: not geometric relation