UNIVERSITY OF GOTHENBURG

# L1: Introduction to Artificial Intelligence: Cognitive Systems

Simon Dobnik
Centre for Linguistic Theory and Studies in Probability (CLASP)
FLOV, University of Gothenburg

Gothenburg, November 11, 2018

CLASP centre for linguistic theory and studies in probability

# Practicalities

UNIVERSITY OF GOTHENBURG

- Course webpage: closed, open
- Schedule and topics
- Examination

CLASP centre for linguistic theory and studies in probability

# What this course is about?

UNIVERSITY OF GOTHENBURG

- Computational modelling of language, action and perception in relation to situated dialogue agents and image classification

# What this course is about?

UNIVERSITY OF GOTHENBURG

- Computational modelling of language, action and perception in relation to situated dialogue agents and image classification
- Relates to:
  - linguistics
  - cognitive science and psychology
  - computer science
  - computer vision
  - robotics
  - artificial intelligence

# What this course is about?

UNIVERSITY OF GOTHENBURG

- Computational modelling of language, action and perception in relation to situated dialogue agents and image classification
- Relates to:
  - linguistics
  - cognitive science and psychology
  - computer science
  - computer vision
  - robotics
  - artificial intelligence
- Embodied and situated language processing (LT2308)

# AICS and robotics

UNIVERSITY OF GOTHENBURG

- Spatial cognition and action represent the core of human cognition and behaviour.
- A robot that can make sense of the world and interact with humans is very useful: navigation systems, assistants to people with disabilities, robots on rescue missions, just for fun, etc.
- Having access to robot' sensors and actuators can give us a theoretical insight into language, spatial perception and action.

# AICS and computer vision

UNIVERSITY OF GOTHENBURG

- With social media there are a lot of images and videos available online
- Visual information closely linked to textual data, e.g. a newspaper article or a Facebook post
- Can we make sense of it?
  - Information retrieval
  - Navigation systems
  - Advertising
  - Security
- Generating images and video from text
  - Computer animation

# We will discuss three kinds of topics

UNIVERSITY OF GOTHENBURG

- Linguistics and psychology: how humans connect language, spatial perception, action?
- Formal computational systems: what kind of models and algorithms do we employ?
- Applications: what kind of problems do we want to solve?

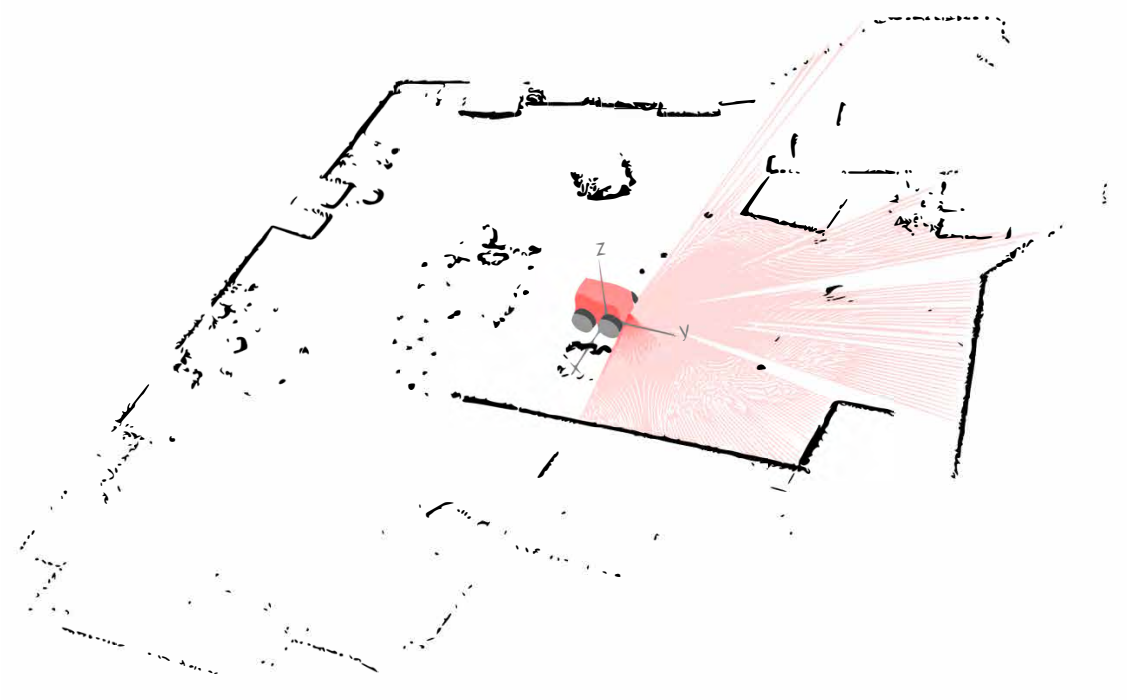
# How do we do it?

UNIVERSITY OF GOTHENBURG

1. 
2. $\forall x \forall y[\text{supports}(y,x) \wedge \text{contiguous}(\text{surface}(x), \text{surface}(y)) \rightarrow \text{on}_1(x,y)]$
3. *The newspaper is on the table*

CLASP centre for linguistic theory and studies in probability

# How do we do it?

1. 

2. $\forall x \forall y[\text{supports}(y,x) \wedge \text{contiguous}(\text{surface}(x), \text{surface}(y)) \rightarrow \text{on}_1(x,y)]$
3. *The newspaper is on the table*

# Some key questions for AICS

- How does natural language interact with the physical world through action and perception?
- How a situated agent can make sense of the world/assign meaning in which it is located?
- How a situated agent can make sense of the conversation with other situated agents?
- How to mediate between perceptual sensory data (real numbers) and symbolic representations of language?
- How to deal with constantly changing world - learn from experience?

# What do we need?

Here are some examples from the Flickr8k corpus (Rashtchian et al., 2010). Each image is followed by five descriptions. The descriptions were made by human annotators using crowd-sourcing with Amazon Mechanical Turk, one description per person per image.

The spatial relations that I would like you to focus on are highlighted. Think about the problems we need to solve to connect words (describing spatial relations) with images.

- A man is riding on a red motorcycle.
- A motorcycle driver dressed in orange gear swerves to the right.
- A motorcyclist on a red speed bike leans into a sharp turn.
- Motorcyclist crouches low as he rounds a turn.
- This person is on a red motorcycle.

- A baseball is recoiling from an action taken on a treated field watched by others.
- A baseball player on a playing field springs into action.
- A baseball player standing on the mound.
- A Philadelphia Phillie pitcher on the pitchers mound with his left leg up behind him.

- A big black and brown dog plays outdoors.
- A black and tan dog leaps over the green grass.
- A brown and black dog runs on the grass outdoors in front of a sidewalk.
- A dog runs.
- A German shepherd jumps left on patchy grass.

# What do we need? - Simon's summary

- Theoretical background
  - How language relates to perception
  - World knowledge about objects (rules of physics, conceptual information about objects, interaction between objects)
  - Perspective
  - Information fusion
- Representations and algorithms
  - Computational models of language and perception
  - Representation of action in time
  - Process image and detect objects
  - Process image and detect relations between objects
  - Relate whole sentences to images
  - NLU and LNG
- Applications
  - Generating image descriptions
  - Visual question answering

# Situated meaning

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

- Model-theoretic semantics
  - model
  - interpretation function
  - types $e$ and $t$ and function types
- TTR
  - meaning relative to agent
  - 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
  - judgements
  - of situations, of speech events and of neural events

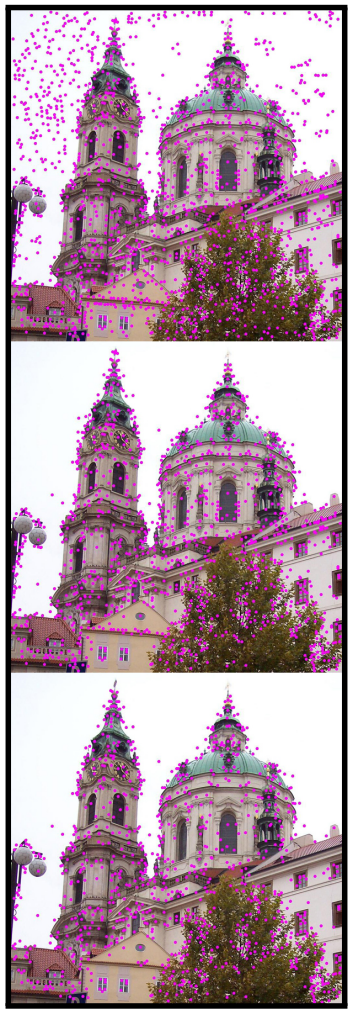
# Grounding

Harnad (1990) and Roy (2002)

- Language, sensors, motors and learning
- Connecting symbolic and "connectionist" models
- Types of representations:
  - Iconic representations
  - Categorical representations
  - Higher level symbolic representations: compositional structure

# Image classification

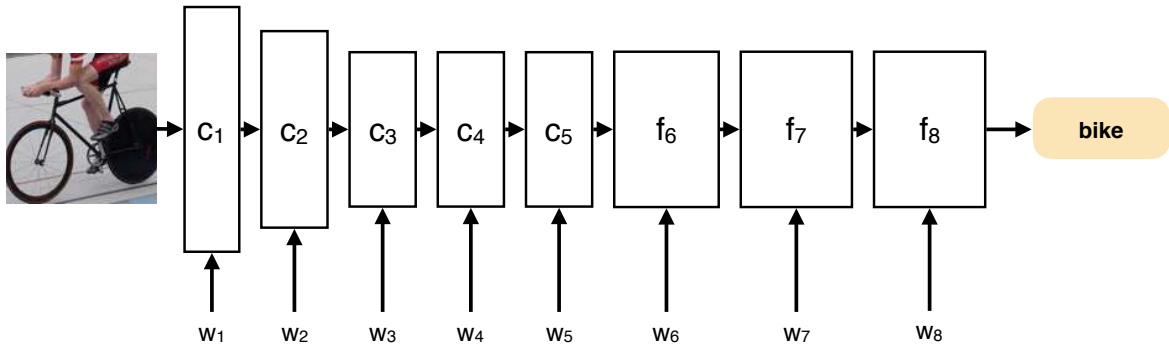
- Scale Invariant Image Transform (SIFT) features
  Lowe (1999)
- Creating visual words



# Learning features with deep learning

UNIVERSITY OF GOTHENBURG

## Review

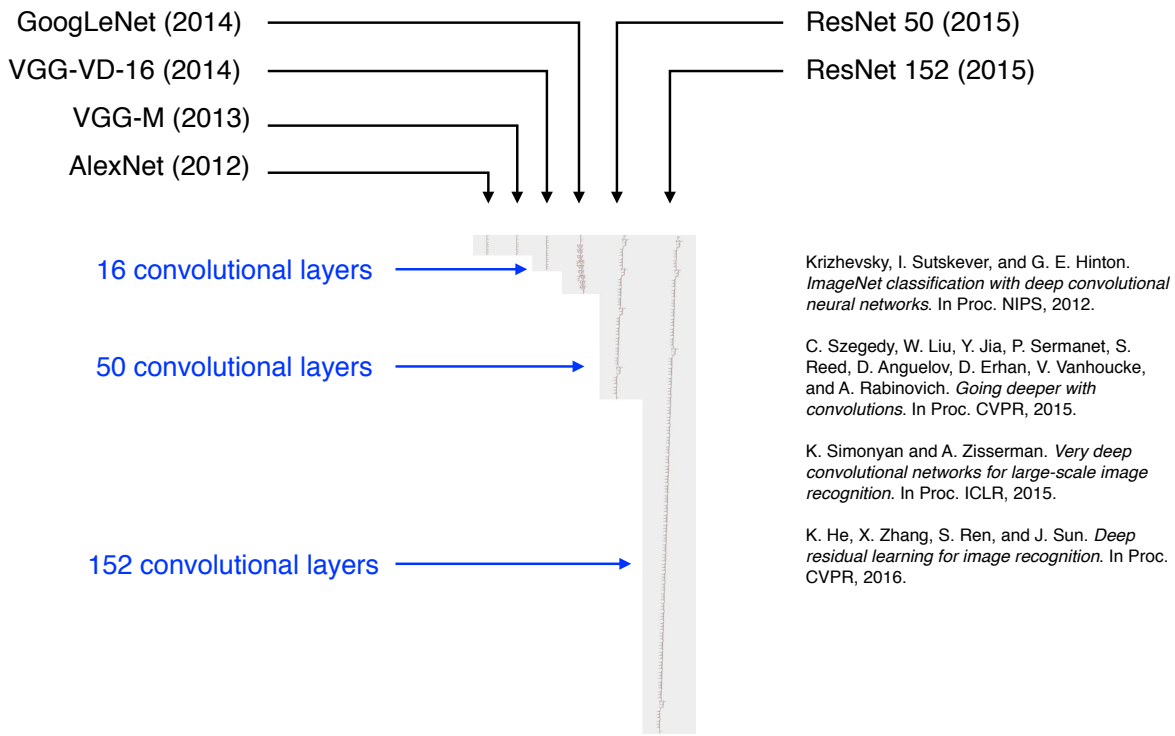



A. Krizhevsky, I. Sutskever, and G. E. Hinton. *Imagenet classification with deep convolutional neural networks*. In Proc. NIPS, 2012.

Slide from Vedaldi (2016)

# How deep is enough?






Krizhevsky, I. Sutskever, and G. E. Hinton. *ImageNet classification with deep convolutional neural networks.* In Proc. NIPS, 2012.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. *Going deeper with convolutions.* In Proc. CVPR, 2015.

K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition.* In Proc. ICLR, 2015.

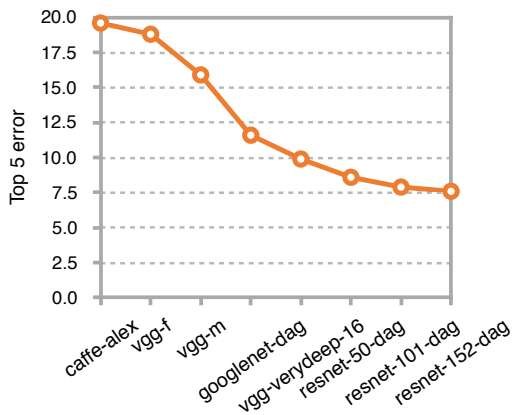
K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition.* In Proc. CVPR, 2016.

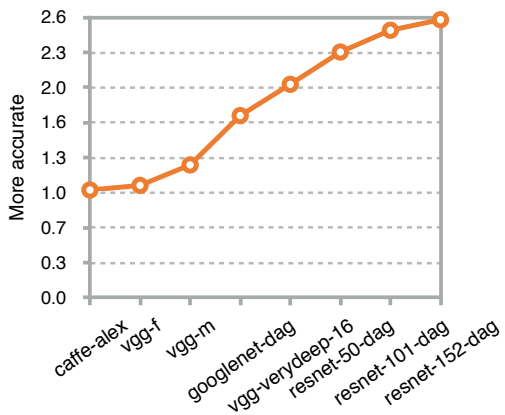
Slide from Vedaldi (2016)

# Accuracy



**3 × more accurate in 3 years**



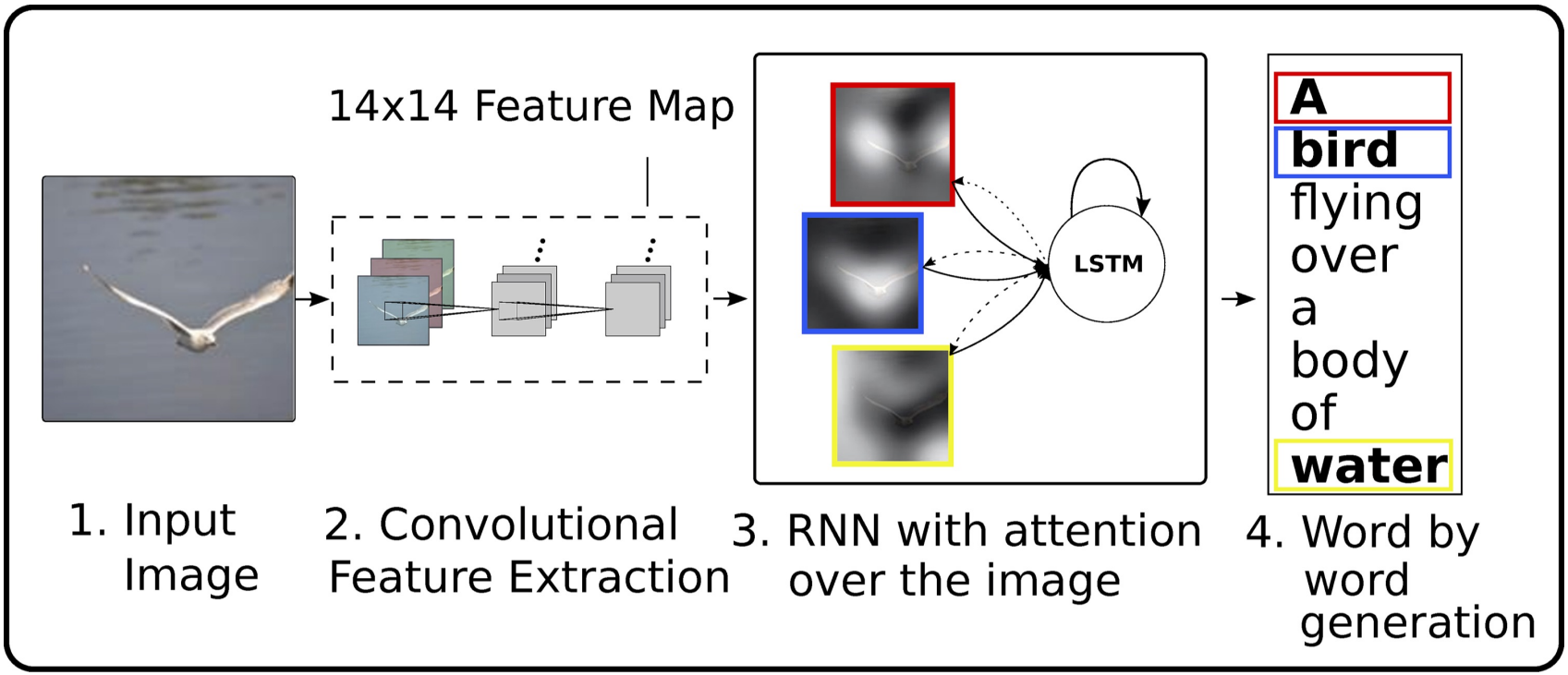

Slide from Vedaldi (2016)

# Generating image descriptions

Elliott and Keller (2013) and Mitchell et al. (2013)

- What is the best description for an image?
- Visual focus
- Knowledge about the world
- Generating referring expressions Dale and Reiter (1995): include maximally distinguishable descriptions
- But there is over-specification (red)

# . . . with deep learning




(Xu et al., 2015)

# Visual question answering, I

## Learning to compose neural networks (Andreas et al., 2016)

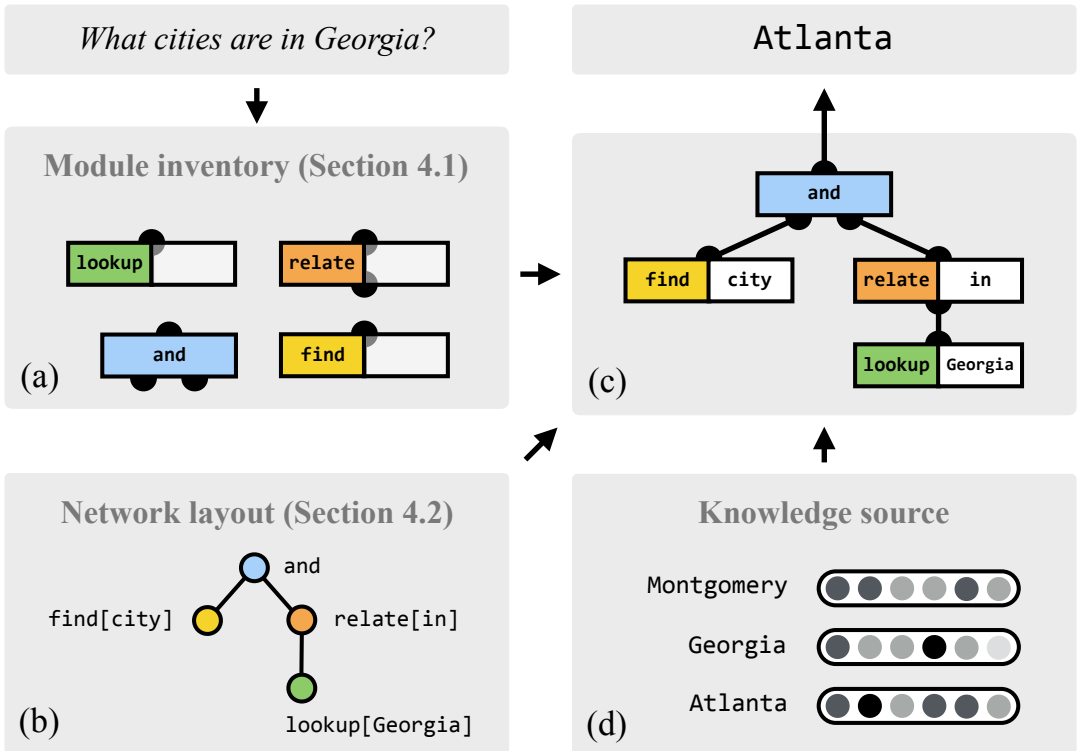



Figure 1: A learned syntactic analysis (a) is used to assemble a collection of neural modules (b) into a deep neural network (c), and applied to a world representation (d) to produce an answer.

# Visual question answering, II

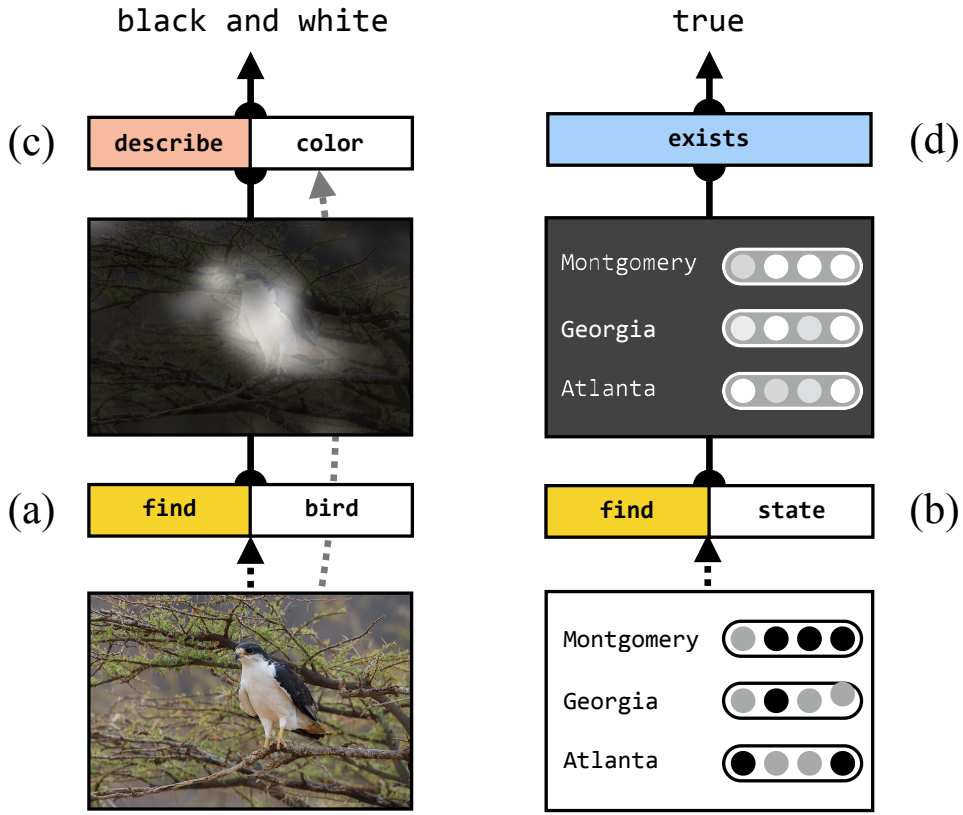
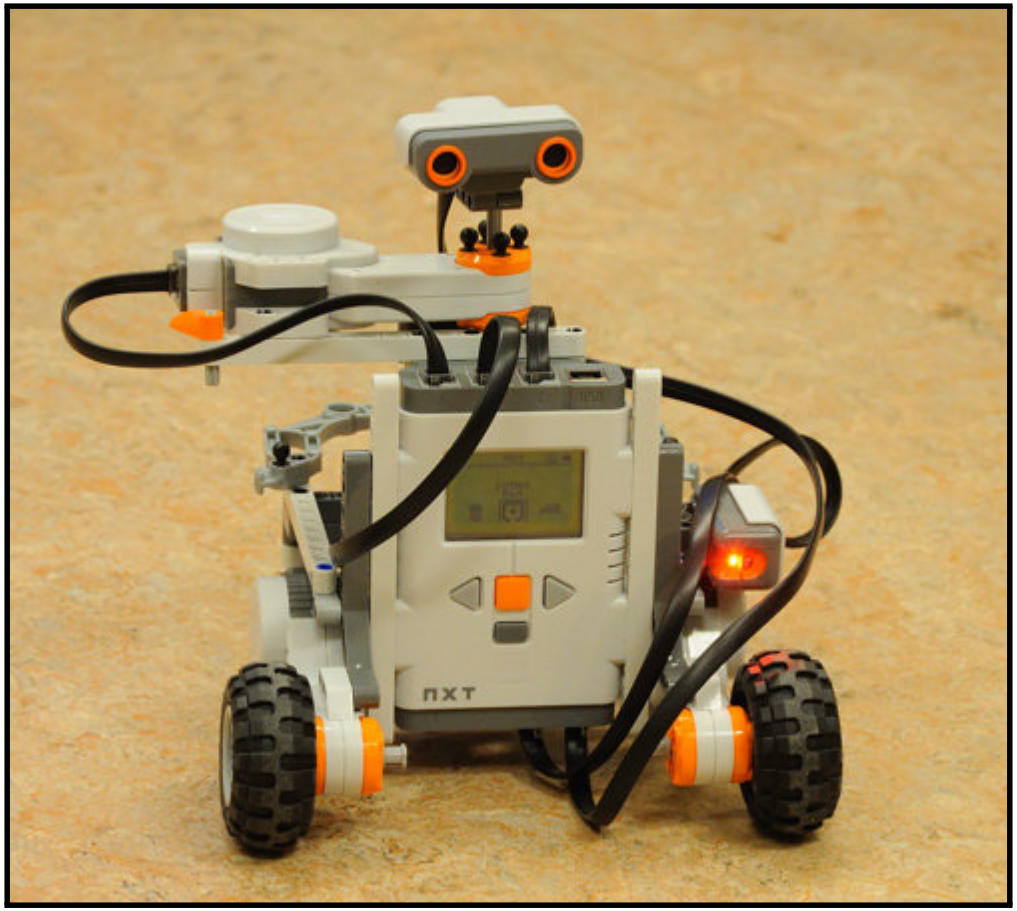



Figure 2: Simple neural module networks, corresponding to the questions What color is the bird? and Are there any states?

# What do we need? II




http://www.youtube.com/watch?v=6afrMnEmXFI

# What do we need? III

https://www.youtube.com/watch?v=AsEgaka6tH0

de Graaf (2016); Dobnik and de Graaf (2017)

# What do we need? - Simon's summary

UNIVERSITY OF GOTHENBURG

- Theoretical background
  - Incrementality of representations
  - Theory of pragmatics + interaction
  - Human-machine interaction
- Representations and algorithms
  - Dialogue management
  - Dialogue games and interactional strategies
  - Incremental machine learning
  - ML from a few examples (one-shot learning, active learning)
  - Dialogue games as optimisation in ML: reinforcement learning
  - Middleware for robotics
- Applications
  - Personal assistants

# Learning language with robots

Dobnik (2006), Dobnik (2009), Kruijff et al. (2007), Zender et al. (2008), Lauria et al. (2001) and Lauria et al. (2002)

- Processes required
- Temporal processing
- Information flow
- Information fusion
- Increased abstraction of representations
- Middle-ware for the robot: ROS

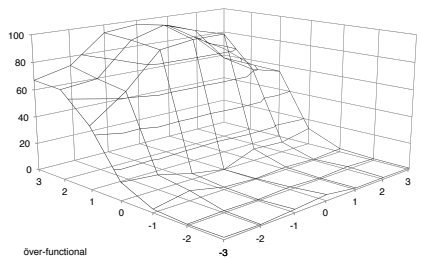
# Learning through dialogue interaction

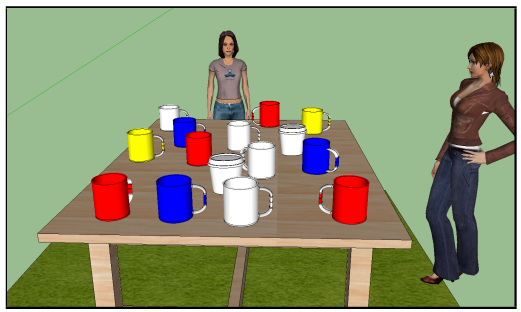
Skočaj et al. (2010), Skočaj et al. (2011), Steels and Loetzsch (2009), Steels and Belpaeme (2005), Steels and Baillie (2003), Schütte et al. (2015), Dobnik and de Graaf (2017)

- Situated agents interact with their environment and other situated agents
- Interaction is accommodation, i.e. learning.
- A robot is lost: why not just ask!
- Instruction is one-shot learning
- Relation to data-driven learning from several observations

# Language and space

Logan and Sadler (1996), Dobnik and Åstbom (2017), Regier and Carlson (2001), Coventry and Garrod (2005), Kelleher and Costello (2009), Dobnik, Howes, and Kelleher (2015)



# What we need? IV

https://www.youtube.com/watch?v=8hxIVpWf5x8

Skantze (2016)

# Gestures and emotions

Expressing meaning with our body

- Conversational resources
- Non-verbal cues and information
- . . . but not any kind of movement and prosody.
- Coordination in conversation:
  - understanding and misunderstanding
  - turn-taking
  - topic progression
  - empathy
  - sarcasm
  - attitude
  - mood
  - . . .

# References I


Andreas, J., M. Rohrbach, T. Darrell, and D. Klein (2016, June 12-17). Learning to compose neural networks for question answering. In *Proceedings of NAACL-HLT 2016*, San Diego, California, pp. 1545–1554. Association for Computational Linguistics.

Cooper, R. (in prep). Type theory and language: from perception to linguistic communication. Draft of book chapters available from https://sites.google.com/site/typetheorywithrecords/drafts.

Cooper, R., S. Dobnik, S. Lappin, and S. Larsson (2015, November). Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology — LiLT 10*(4), 1–43.

Coventry, K. and S. Garrod (2005). Spatial prepositions and the functional geometric framework. towards a classification of extra-geometric influences. In L. A. Carlson and E. v. d. Zee (Eds.), *Functional features in language and space: insights from perception, categorization, and development*, Volume 2, pp. 149–162. Oxford University Press.

# References II


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

de Graaf, E. (2016, June, 8th). Learning objects and spatial relations with Kinect. Master's thesis, Department of Philosophy, Linguistics and Theory of Science. University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik, examiner: Richard Johansson, opponent: Lorena Llozhi.

Dobnik, S. (2006, March 8–9). Learning spatial referential words with mobile robots. In *Proceedings of the 9th Annual CLUK Research Colloquium*, Milton Keynes, United Kingdom, pp. 1–8. The Open University.

Dobnik, S. (2009, September 4). *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.

# References III


Dobnik, S. and A. Åstbom (2017, August 15–17). (Perceptual) grounding as interaction. In V. Petukhova and Y. Tian (Eds.), *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken, Germany, pp. 17–26.

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

Dobnik, S., R. Cooper, and S. Larsson (2013). Modelling language, action, and perception in type theory with records. In D. Duchier and Y. Parmentier (Eds.), *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). Berlin, Heidelberg: Springer.

# References IV


Dobnik, S. and E. de Graaf (2017, 22–24 May). KILLE: a framework for situated agents for learning language through interaction. In J. Tiedemann and N. Tahmasebi (Eds.), *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, Gothenburg, Sweden, pp. 162–171. Northern European Association for Language Technology (NEALT): Association for Computational Linguistics.

Dobnik, S., C. Howes, and J. D. Kelleher (2015, 24–26th August). Changing perspective: Local alignment of reference frames in dialogue. In C. Howes and S. Larsson (Eds.), *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, Gothenburg, Sweden, pp. 24–32.

Elliott, D. and F. Keller (2013). Image description using visual dependency representations. In *EMNLP*, Volume 13, pp. 1292–1302.

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

# References V


Karpathy, A. and L. Fei-Fei (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137.

Kelleher, J. D. and F. J. Costello (2009). Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics 35*(2), 271–306.

Kruijff, G.-J. M., H. Zender, P. Jensfelt, and H. I. Christensen (2007). Situated dialogue and spatial organization: what, where... and why? *International Journal of Advanced Robotic Systems 4*(1), 125–138. Special issue on human and robot interactive communication.

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

Lauria, S., G. Bugmann, T. Kyriacou, J. Bos, and E. Klein (2001, September/October). Training personal robots using natural language instruction. *IEEE Intelligent Systems 16*, 38–45.

Lauria, S., G. Bugmann, T. Kyriacou, and E. Klein (2002). Mobile robot programming using natural language. *Robotics and Autonomous Systems 38*(3–4), 171–181.

# References VI


Logan, G. D. and D. D. Sadler (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. A. Peterson, L. Nadel, and M. F. Garrett (Eds.), *Language and Space*, pp. 493–530. Cambridge, MA: MIT Press.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, Volume 2, pp. 1150–1157. IEEE.

Mitchell, M., K. van Deemter, and E. Reiter (2013). Generating expressions that refer to visible objects. In *HLT-NAACL*, pp. 1174–1184.

Rashtchian, C., P. Young, M. Hodosh, and J. Hockenmaier (2010, 6 June). Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with Amazon's Mechanical Turk*, Los Angeles, CA. North American Chapter of the Association for Computational Linguistics (NAACL).

# References VII


Regier, T. and L. A. Carlson (2001). Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General 130*(2), 273–298.

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

Schütte, N., J. Kelleher, and B. Mac Namee (2015). Reformulation strategies of repeated references in the context of robot perception errors in situated dialogue. In *Proceedings of the Workshop on Spatial Reasoning and Interaction for Real-World Robotics at the International Conference on Intelligent Robots and Systems (IROS-2015)*, pp. 4–11.

Skantze, G. (2016). Real-time coordination in human-robot interaction using face and voice. *AI Magazine 37*(4), 19–31.

Skočaj, D., M. Janiček, M. Kristan, G.-J. M. Kruijff, A. Leonardis, P. Lison, A. Vrečko, and M. Zillich (2010). A basic cognitive system for interactive continuous learning of visual concepts. In *ICRA 2010 workshop ICAIR - Interactive Communication for Autonomous Intelligent Robots*, Anchorage, AK, USA, pp. 30–36.

# References VIII


Skočaj, D., M. Kristan, A. Vrečko, M. Mahnič, M. Janíček, G.-J. M. Kruijff, M. Hanheide, N. Hawes, T. Keller, M. Zillich, and K. Zhou (2011, 25-30 September). A system for interactive learning in dialogue with a tutor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2011*, San Francisco, CA, USA.

Steels, L. and J.-C. Baillie (2003, May). Shared grounding of event descriptions by autonomous robots. *Robotics and Autonomous Systems 43*(2–3), 163–173.

Steels, L. and T. Belpaeme (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences 28*(4), 469–489.

Steels, L. and M. Loetzsch (2009). Perspective alignment in spatial language. In K. R. Coventry, T. Tenbrink, and J. A. Bateman (Eds.), *Spatial Language and Dialogue*. Oxford University Press.

Vedaldi, A. (2016, March). 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 IX

Xu, K., J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio (2015, February 11). Show, attend and tell: Neural image caption generation with visual attention. *arXiv arXiv:1502.03044 [cs.LG]*, 1–22.

Zender, H., O. Martínez-Mozos, P. Jensfelt, G.-J. M. Kruijff, and W. Burgard (2008, June). Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems 56*(6), 493–502. Special issue "From sensors to human spatial concepts".