

Pattern Recognition is Not Enough: Representing Language, Action and Perception with Modular Neural Networks

Simon Dobnik¹ and John D. Kelleher²

¹Centre for Linguistic Theory and Studies in Probability (CLASP), FLoV, University of Gothenburg

²Technological University Dublin, Ireland

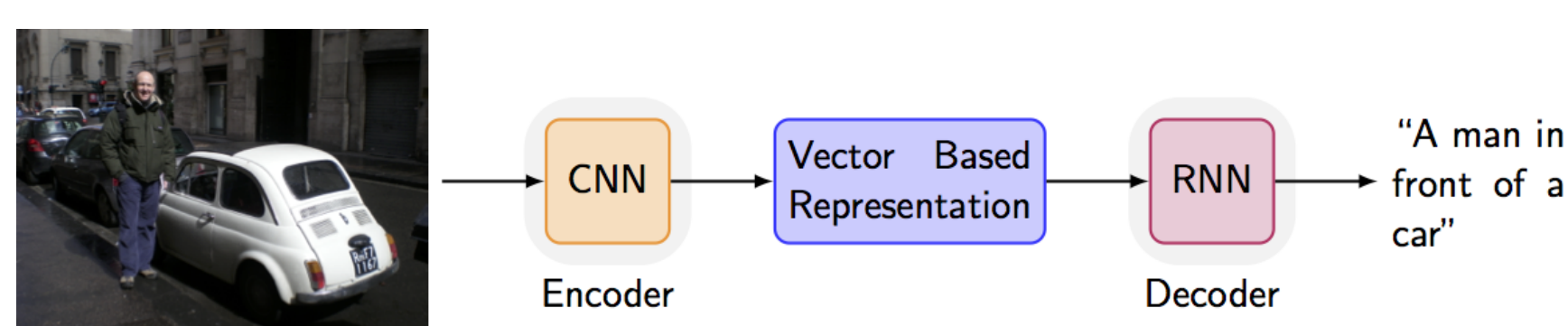
simon.dobnik@gu.se, john.d.kelleher@tudublin.ie

1 Aims

- Situated agents/robots need to refer to space
- Spatial descriptions:** “the chair is to the left and close to the table” and “go down the corridor until the large painting on your right, then turn left”
- Grounded in several modalities
- Shortcomings of DNN approaches when generating them
- We need a **modular approach** to DNNs
- Combines top down (**mechanistic**) and bottom up (**phenomenological**) approaches

2 Shortcomings of the current models

- DNNs are suited for learning **multi-modal representations**: discrete (words) and continuous (word embeddings and visual features)



- Generalised learning mechanisms that learn with relatively high-level (coarse) supervision through architecture design: **bottom-up** or **phenomenological approach**
- Pattern recognition is not enough**



“...without intuitive physics, intuitive psychology, compositionality, and causality.” (Lake et al., 2016)

3 Multi-dimensionality of meaning of spatial language



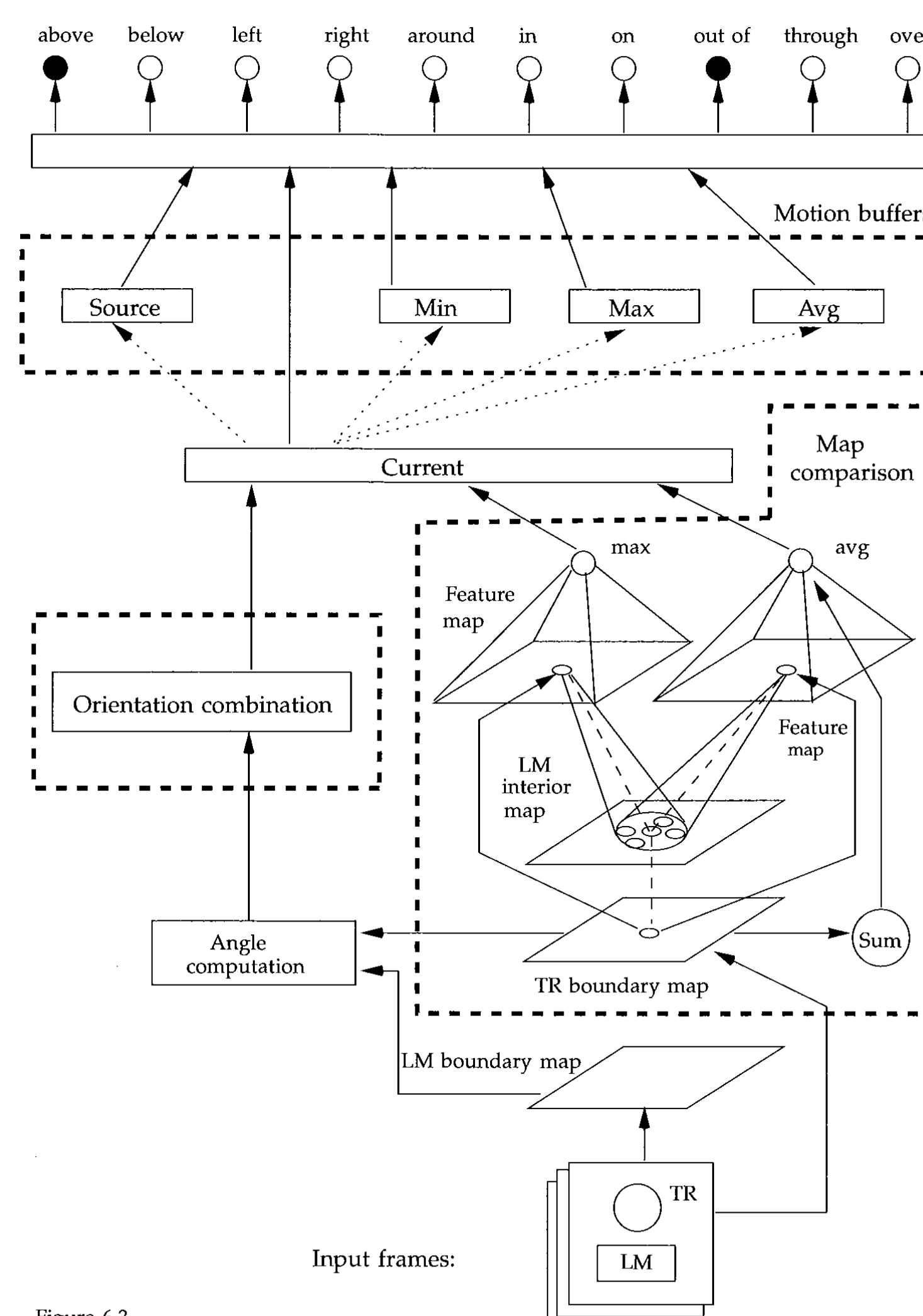
- Scene geometry
- Functional world knowledge about dynamic kinematic routines between objects
- Perspective
- Interaction between agents and with their environment
- A theory of how different factors in spatial language are integrated? (Herskovits, 1987; Coventry and Garrod, 2005)

4 Modular approaches

- Build a solution in a piece-wise manner and then integrate
- Deep learning is assisted with domain knowledge expressed as modules that are trained on data: a **top-down** or **mechanistic approach**

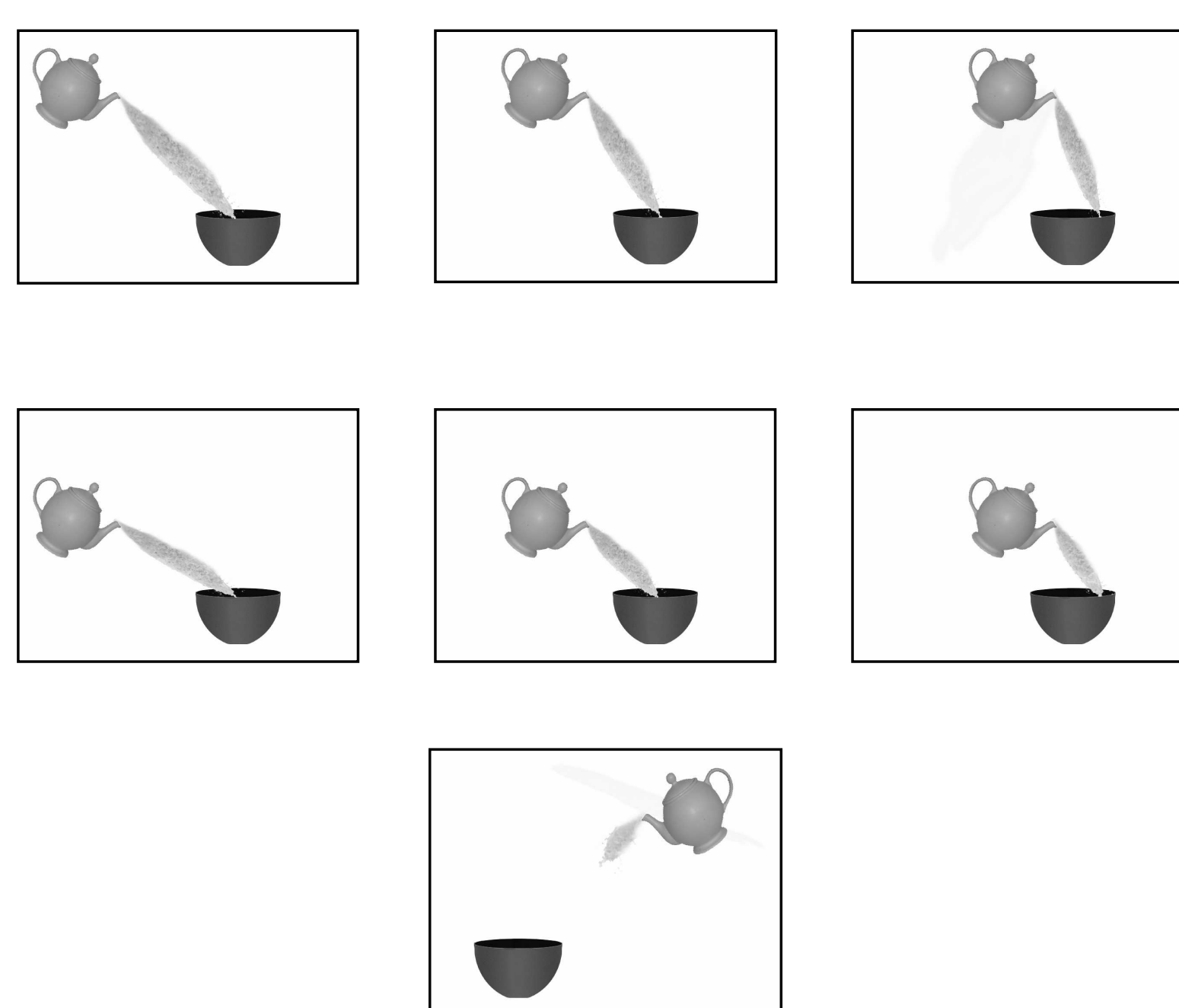
5 Promising architectures

- (Regier, 1996): **constrained connectionist network**, captures geometric factors and paths of object motion to predict a description



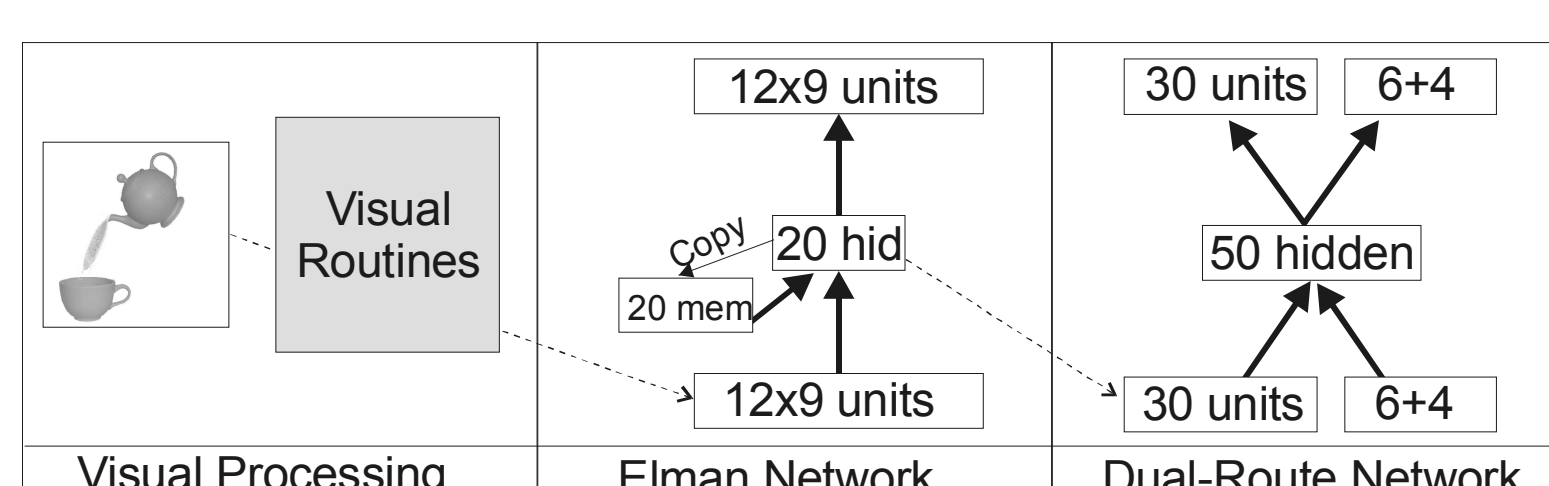
- Coventry et al. (2005): **interconnected networks**

– Dynamic visual scenes containing three objects: a teapot pouring tea into a cup



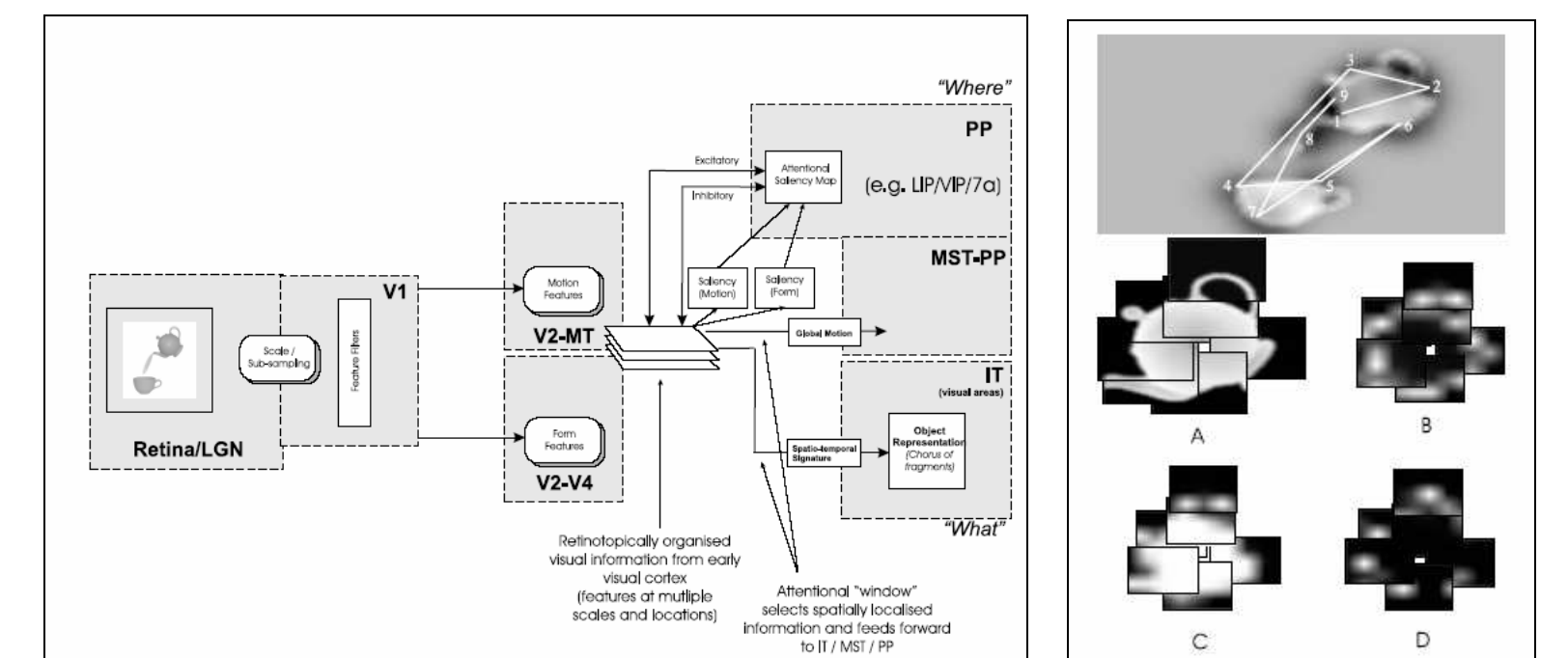
– **Geometric arrangement** (6 locations) vs **function of objects** (tea reaches the container, misses the container, no tea), degrees of pouring

– For each temporal snapshot of the scene, optimise the appropriateness score of a spatial description obtained in subject experiments



– Transfer learning: modules trained independently but are connected to encode representations

– **Object recognition:** a neurally inspired vision processing module that deals with detection of objects (“what”) and motion (“where”) of objects from image sequences using an attention mechanism

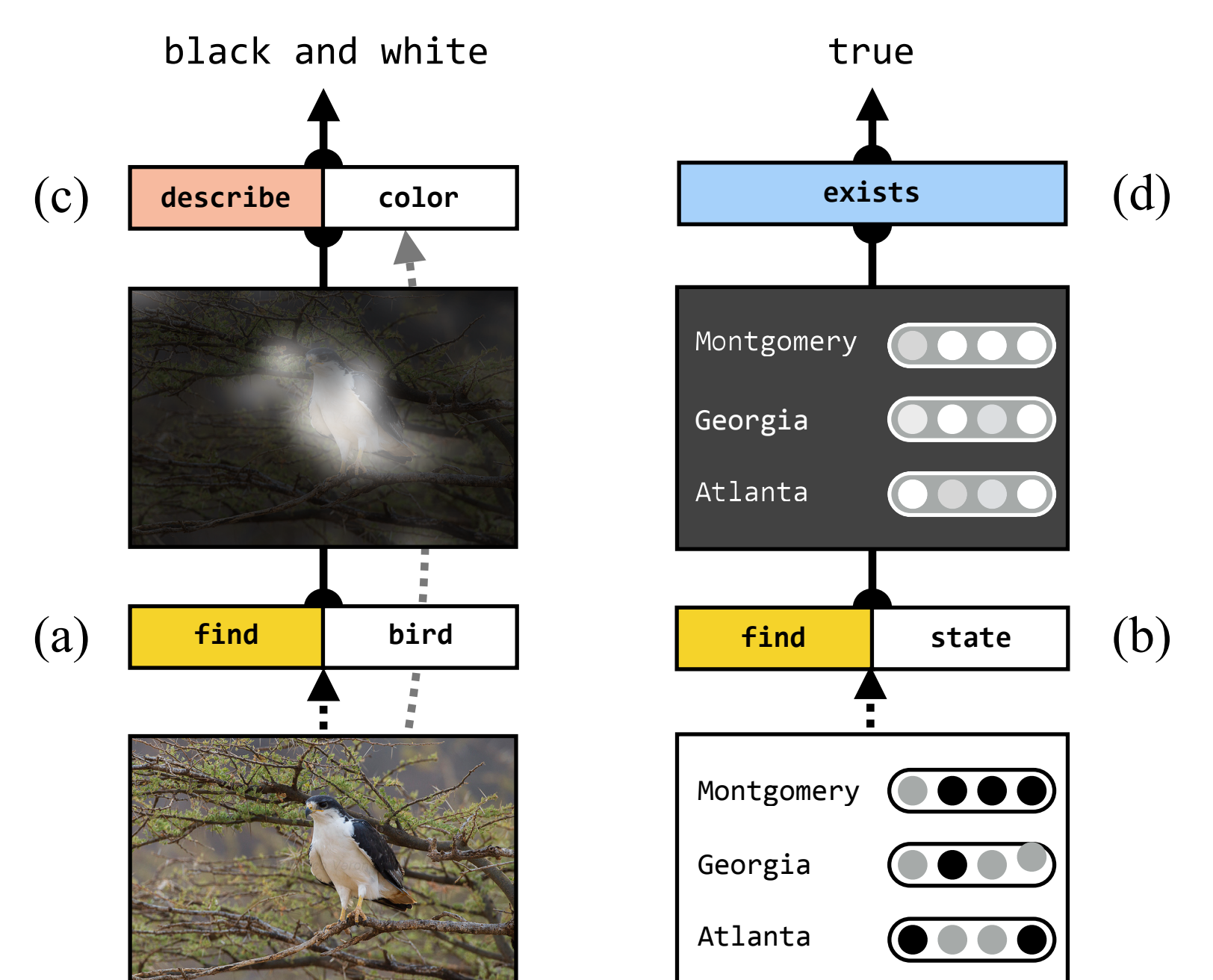


– **Interaction of objects:** an Elman recurrent network that learns the dynamics of the attended objects in the scene over time

– **Vision and language fusion:** integration of the grounded visual information (30) with language (6 object names and 4 prepositions) to predict the same visual data, 6 object names, and ratings for 4 prepositions

- Andreas et al. (2016): **sequencing the modules**

– Visual question answering: associate a question and visual/database representation with an answer by finding a sequence of trainable neural modules using reinforcement learning



6 Conclusions and future work

- DNNs allow for a great flexibility in combining top-down specification (hand-designed structures and rules) and data driven approaches
- Can be modularised to specialise for a particular task
- Modules can be pre-trained (even on a different dataset) and used as feature encoders
- Good at information fusion
- Well-suited for modelling spatial language
- Scale the existing neural spatial language models to a large corpus of image descriptions (Krishna et al., 2017)
 - distortion of object appearance and geometry by perspective at which an image was taken
 - not all spatial configurations of an object pair in a temporal sequence are there
 - different configurations may appear similar
 - no direct human judgements scores
 - bias to particular kinds of objects and interactions
- Extend the modalities of (Coventry et al., 2005), e.g. referential games (Lazaridou et al., 2016)

