Grounding Spatial Language in Perception by Combining Concepts in a Neural Dynamic Architecture

Daniel Sabinasz (daniel.sabinasz@ini.rub.de)
Mathis Richter (mathis.richter@ini.rub.de)
Jonas Lins (jonas.lins@ini.rub.de)
Gregor Schöner (gregor.schoner@ini.rub.de)

Institut für Neuroinformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany

Abstract

We present a neural dynamic architecture that grounds sentences in perception which combine multiple concepts through nested spatial relations. Grounding entails that the model gets features and relations as categorical inputs and matches them to objects in space-continuous neural maps which represent visual input. The architecture is based on the neural principles of dynamic field theory. It autonomously generates sequences of processing steps in continuous time, based solely on highly recurrent connectivity. Simulations of the architecture show that it can ground sentences of varying complexity. We thus address two major challenges in dealing with nested relations: how concepts may appear in multiple different relational roles within the same sentence, and how in such a scenario various grounding outcomes may be "tried out" in a form of hypothesis testing. We close by discussing empirical evidence for crucial assumptions and choices made when developing the architecture.

Keywords: language grounding; conceptual combination; neural network; dynamical system; dynamic field theory

Introduction

Human cognitive competences depend critically on the capacity to combine concepts through relations. Humans are capable of generating complex trains of thought to reach a conclusion. Humans understand and generate propositions that range in complexity from single words to deeply nested sentences. This competence is critical to establish joint attention to specific objects or events when humans communicate. Imagine you and your friend are standing on the top of a hill at night with a view of the city (Figure 1). To point your friend to your house, you say: "I live in the house to the right of the large house that is next to the big tree and below a star." You are using the complex combination of multiple spatial relations because the objects in the scene are relatively poor in visual features in this nightly scene. Your friend may *perceptually ground* your utterance by directing her attention to different objects in the sentence, until settling on the location of your house.

Through what kind of neural processes may the brain bring about the perceptual grounding of sets of phrases that invoke concepts and relations? Earlier modeling work proposed the component processes that enable the grounding of an individual relation between a pair of objects (e.g., "A above B", where "A" is the target and "B" the reference object; Lipinski, Schneegans, Sandamirskaya, Spencer, & Schöner, 2012; Richter, Lins, Schneegans, Sandamirskaya, & Schöner, 2014;

Figure 1: A scene for which multiple spatial relations may naturally be used to refer to an object.

Richter, Lins, & Schöner, 2017). In this paper, we extend that work to provide a neural process account for the perceptual grounding of sets of relations with varying degrees of complexity (e.g., "A above B and below C" or "A above the B that is below C"). The account is aligned with the notion of grounded cognition (Barsalou, 2008) in that the higher cognitive processes responsible for such grounding overlap with neural processes responsible for perception and visual cognition.

Rather than describing the outcome of the neural process in abstract computational terms such as symbol manipulation (Fodor & Pylyshyn, 1988), we take principles of neural processing seriously. As a result, we face challenges that are not visible at the abstract computational level. Principal among those is the fact that neural networks are not free to pass just any "information" from one processing step to another. The activation that synaptic connections pass from one population of neurons to the next does not include information about from where activation originated or what it is about. In neural networks, such information is instead implicit in the pattern of connectivity. For activation patterns to play different roles in a subsequent processing step, the patterns must actually reside in distinct neural populations with different sets of connections. For instance, activation patterns representing target objects reside in different sub-networks than activation patterns representing reference objects, so that the two sets of activation patterns play different roles in grounding a relation.

While grounding sequences of relations, neural process ac-

counts then face the problem that the role of an object may change along the way. In the example of Figure 1, the large house is the reference object in "the house to the right of the large house", while it is the target object in "the large house that is next to the big tree". In the course of processing these phrases sequentially, different activation patterns must represent the large house in these two different roles so that different paths of neural connectivity guide the next processing step.

A second challenge arises when different possible grounding outcomes must be "tried out" in a form of hypothesis testing. For example, if there are multiple large houses in the scene, the model may select one candidate as reference object and then try to ground the rest of the description. If the wrong large house was initially selected, that choice must be rejected and a new candidate selected, a process easily conceived of in terms of algorithms that manipulate symbols (as suggested in the verbal description we give here), but challenging to conceive of in terms of activation patterns in neural networks. Here we extend an earlier proposal for such hypothesis testing to sequences of relations (Richter et al., 2014).

Our neural process account is based on dynamic field theory (DFT; Schöner, Spencer, & the DFT Research Group, 2015), a framework for using strongly recurrent neural networks to understand embodied cognition. Grounding happens by generating activation patterns in neural fields, populations of neurons that receive input from the visual surface. The interface to language is a set of neural nodes that represent feature concepts like RED, relation concepts like TO THE LEFT OF, and grounding instructions that correspond to different grammatical roles. We begin by introducing the relevant concepts of DFT, then provide an overview over the proposed neural architecture, and finally demonstrate the competence of the model in a few exemplary simulations.

Methods: DFT

In DFT, the activity of populations of neurons is captured by dynamic neural fields (or maps), whose activation, $u(x,t)$, is defined over continuous feature dimensions, x, and evolves in continuous time, t, according to:

$$\tau \dot{u}(x,t) = -u(x,t) + h + \int g(u(x',t)) k(x-x') \, dx' + s(x,t)$$

τ is the time scale and $h < 0$ is the resting level. The integral captures interaction within the field, excitatory ($k(x-x') > 0$) over short distances, $x-x'$, and inhibitory ($k(x-x') < 0$) over large distances. The sigmoid transfer function, $g(u(x,t))$, makes this a nonlinear integro-differential equation. Input, $s(x,t)$, from the sensory surface reflects the forward connectivity that makes that the field is sensitive to dimensions, x. Input may also come from other fields within an architecture. Zero-dimensional fields are dynamic neural nodes that respond categorically to input.

At small levels of input, the sub-threshold state, $u(x,t) = h + s(x,t) < 0$, is stable until it reaches the detection instability when interaction engages. The activation then switches to supra-threshold peak solutions that are self-stabilized by interaction. A single peak may result for selective fields, a small number of peaks may co-exist for different parameter settings. Peaks may remain stable when inducing input is removed, a simple model of working memory.

Peaks are the units of representation in DFT. This is because only supra-threshold activation is passed on to downstream neural processes as all connections entail the sigmoid threshold function, and the only supra-threshold patterns of activation are peaks. The stability of peaks makes that when multiple fields and nodes are coupled in a neural dynamic architecture, the dynamics remains largely invariant.

Higher-dimensional fields afford additional functions such as biased competition (when a top-down input is localized along one and constant along other dimensions, peaks pop up where bottom-up inputs overlap with top-down inputs), binding, and coordinate transforms (see chapters 5, 7, 8, and 9 of Schöner et al. (2015)).

Architecture

The neural dynamic architecture is depicted in Figure 2. We provide a survey over the functions of its components.

Object representation: perception and mental map

The model receives sensory input from a vision sensor (camera) or an image. Visual pre-processing extracts feature values that provide input to three three-dimensional fields (bottom right in Figure 2), the *color/space perception field*, *orientation/space perception field*, and *shape/space perception field*. They all share the two spatial dimensions of the visual surface. The third dimension reflects the extracted feature, where shape is computed as similarity to a template. The three fields can be regarded as representing the population activation of retinotopic maps in the visual cortex. As such, they comprise a simplified model of perception.

Paired with each perception field is a *mental map* field, defined over the same dimensions but tuned to support sustained activation. These fields keep previously grounded object representations in working memory in order to allow referring to them in later processing steps.

The activation in the perceptual fields have an impact on the attentional system that can be gated (depicted as "gates" in Figure 2). Every perceptual field gives input to a respective perceptual gating field, defined over the same dimensions. These gating fields receive additional inhibitory input from the *from mental map node*, such that they only pass activation on to the attentional system if that node is inactive. Analogously, every mental map field gives input to a mental map gating field that also receives excitatory input from the *from mental map node*, allowing the mental map fields to pass activation on to the attentional system if the *from mental map node* is active. This way, the attentional system is flexible to operate on input from either the perceptual system or from working memory representations of objects attended to in the past.
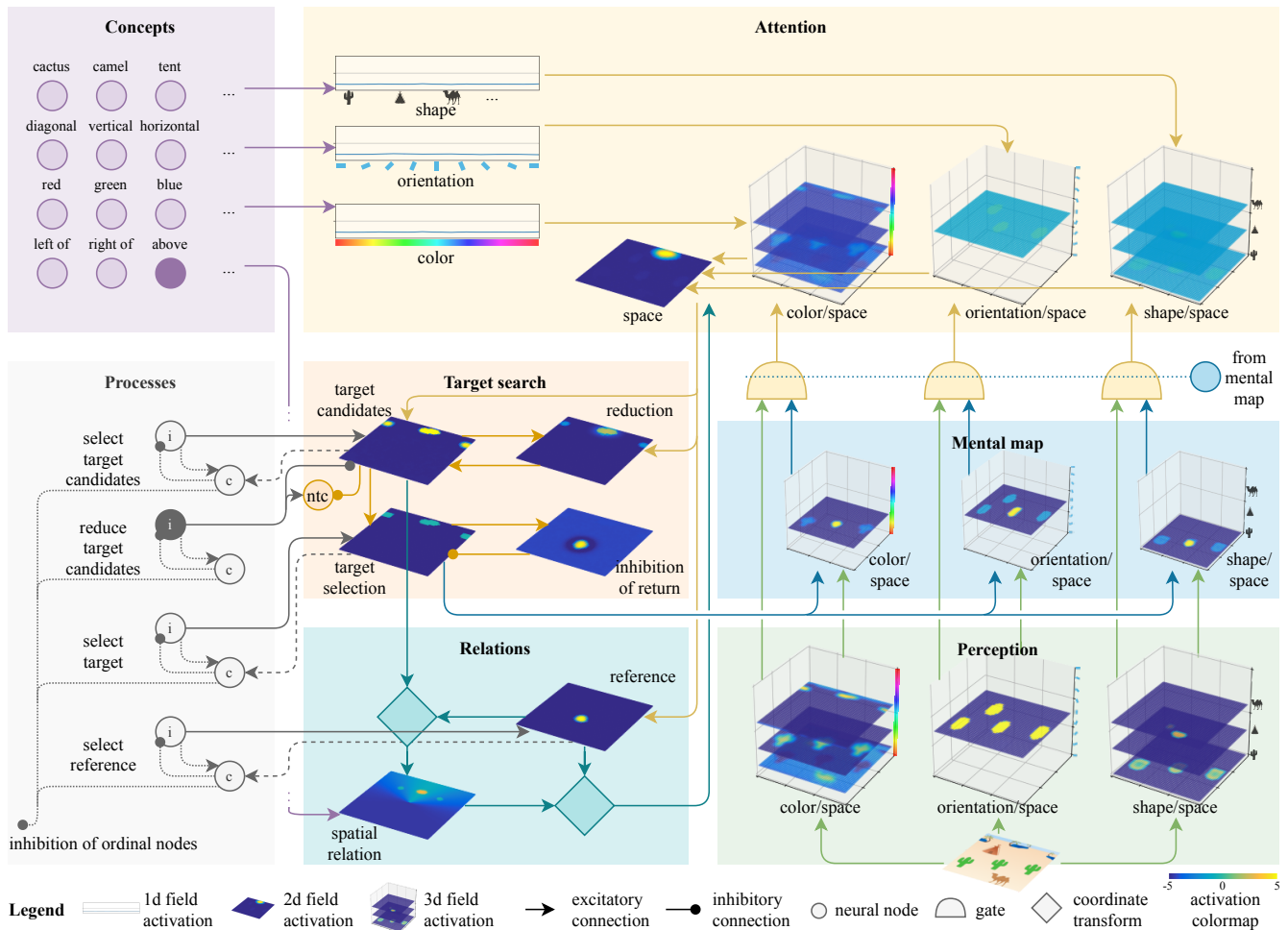
Figure 2: The architecture for grounding combinations of concepts. Only the main structure is visualized.

Attention

Attentional selection of objects occurs through feature attention fields (top right box in Figure 2), defined over the features shape, orientation, and color. These fields are coupled to feature/space attention fields, which combine the respective feature dimension with two visual spatial dimensions. These fields are also coupled to the *space attention field*, which is defined over visual space only.

When a particular feature value is attended, a peak in the corresponding feature field provides slice input into the matching feature/space attention field, highlighting attended features. Additional input reflecting object representations may enter these fields either from perceptual fields or the mental map fields (see above). When that input matches the location of the slice input along the feature dimension, peaks may form. This is how objects with particular features are brought into the attentional foreground.

This part of the architecture models covert attention in the visual cortex. It is a placeholder for a more comprehensive model of visual search (Grieben et al., 2020) that can also accommodate gaze shifts (Schneegans, Spencer, Schöner,

Hwang, & Hollingworth, 2014).

Concepts

Feature concepts (e.g., the color concept RED) are represented by neural nodes (top left in Figure 2). The perceptual meaning of a concept is instantiated in the pattern of synaptic connections between its node and a feature attention field (Richter et al., 2014). Concepts of spatial relations (e.g., LEFT OF) are represented by neural nodes and their synaptic connections to the *spatial relation field*.

In the simulations reported in this paper, the synaptic weights are fixed. For features, the synaptic weight pattern is modeled as a Gaussian centered on a prototypical feature value. For spatial relations, the synaptic weight pattern is modeled based on empirical data (Logan & Sadler, 1996) through a Gaussian in polar coordinates, centered on the angle of the direction of that relation. The synaptic nature of the coupling between concept nodes and sensorimotor fields enables learning concepts by repeatedly presenting examples and altering the synaptic weights with Hebb's rule (Hebb, 1949).

Target search

The fields depicted in the central box of Figure 2 enable the system to hold possible candidates for the target object in working memory and refine the selection of candidates through a sequence of processing steps.

The *target candidates field* holds peaks at the spatial locations of all objects that, at the current stage of processing, are still viable candidates for the target object. Input from the *space attention field* determines where peaks may arise. The *target candidates field* is in the dynamic regime in which peaks can be self-sustained in the absence of localized input, so that it acts as a working memory of candidate objects.

The *reduction field* holds peaks at the locations of all target candidates that presently receive spatial attention. It forms peaks at locations at which inputs from the *space attention field* and from the *target candidates field* overlap. Unless supported through input from the *reduction field*, peaks in the *target candidates field* may decay due to inhibitory input. This enables the model to iteratively eliminate all target candidates that do not receive spatial attention.

The *target selection field* receives input from the *target candidates field*. It operates in the selective dynamic regime in which only a single localized peak may form. Excitatory input into this field controls when such a peak is formed and selection thus takes place. Whenever a selection is made, the representation of the selected object is kept in memory in the mental map fields as well as in the *inhibition of return field*.

Relations

When grounding a target object that stands in a given relation to a reference object, the target candidates have to be reduced to those consistent with that relation (bottom middle box in Figure 2). For example, to ground "a house below the star", the candidate houses have to be reduced to those that are below the star.

The spatial location of the reference object is represented by a peak in the *reference field* that receives input from the *space attention field*, is in the self-sustained dynamic regime, and can be brought into the selective regime by excitatory input.

To apply a spatial relational concept to a reference object, the spatial locations of target candidates are transformed (depicted by a diamond shape in Figure 2) into a reference frame that is centered on the location of the reference object (see Schneegans and Schöner (2012) for the neural dynamics of coordinate transforms). The result feeds into the *spatial relation field*, forming sub-threshold bumps of activation there. That field receives additional sub-threshold input from any activated spatial relation concept node with its characteristic spatial pattern. Only where that pattern overlaps with input from target candidates does the *spatial relation field* form a peak. The reverse coordinate transform converts the peaks' locations back into the reference frame of the visual array and projects them onto the *space attention field*, effectively enabling the model to eliminate all target candidates inconsistent with the relation.

The architecture described here implements the spatial relations TO THE LEFT OF, TO THE RIGHT OF, ABOVE, and BELOW, all in the viewer centered reference frame. Previous work has presented similar architectures implementing an object-centered reference frame, as well as movement relations, such as TOWARD (Richter et al., 2017).

Generating sequences of processing steps

The interface between language and perceptual grounding is a neural dynamical system that generates sequences of processing steps. We assume here, for now, that the language pre-processing system, which we do not model, specifies the serial order of processing steps required to ground the language input. Our architecture activates these processing steps in a sequence that unfolds autonomously in time. Each step in a sequence is represented by an ordinal node, whose connectivity to other ordinal nodes enables their sequential activation (not shown in Figure 2; see Sandamirskaya & Schöner, 2010). Each ordinal node projects onto a processing step (bottom left box in Figure 2), represented by an intention node (labeled "i") that activates the step and a Condition-of-Satisfaction (CoS) node (labeled "c") that signals the completion of the step (Richter, Sandamirskaya, & Schöner, 2012). Activation of the CoS nodes also triggers the transition to the next ordinal node and, hence, the next processing step.

We propose four processing steps that together enable the grounding of sentences comprising multiple concepts in nested relations.

The *select target candidates intention node* is activated in conjunction with a feature concept node that represents a cued feature value of the target. It initiates the grounding of a new target by boosting the *target candidates field* such that it forms peaks at locations at which objects with matching feature values receive spatial attention. When peaks have formed, the CoS node is activated.

The *reduce target candidates intention node* is activated in conjunction with a feature concept node or, in case a reference object has previously been selected, a spatial relation concept node. It inhibits the *target candidates field*, causing all target candidate peaks to decay that do not receive excitatory support from the *reduction field*, i.e., that do not receive spatial attention. The CoS node associated with this intention node becomes active by default after a fixed time determined by its time constant.

The *select target intention node* boosts the *target selection field*, bringing it into a dynamic regime in which it can form a single peak based on input from the *target candidates field*. When a peak has formed, the CoS node is activated.

The *select reference intention node* boosts the *reference field*, bringing it into a dynamic regime in which it can form a single peak, which then activates the CoS node. The selection of where in the field a peak arises is guided by input from the *space attention field*. When activated in conjunction with the *from mental map node*, an object from the mental map instead

of the perceptual input is attended to and selected as reference object.

We will demonstrate in the results section how activating these four processing steps sequentially enables grounding sentences of varying complexity.

Hypothesis testing Grounding combinations of relations may require a form of hypothesis testing: The selection of the target of one relation may depend on the outcome of the grounding of another relation, which cannot be stated without selecting the first target. Initial selection decisions must then be made, but potentially also be reversed later. Reversing a decision requires a neural representation of the failure to find a suitable target or reference object. This is achieved through the *no target candidates node* (labeled "ntc" in Figure 2). It is excited by the *reduce target candidates intention node* and inhibited by the *target candidates field*. The node becomes active if and only if the *reduce target candidates intention node* is active while there is no peak in the *target candidates field*, reflecting that all target candidates have been eliminated.

When the *no target candidates node* becomes active, its global inhibitory projection onto the mental maps resets activation in these, and the sequence generation mechanism is re-initiated. The *inhibition of return field* retains a memory of all objects previously selected as targets. It is defined over visual space and is in the dynamic regime of self-sustained peaks. By providing localized inhibitory input to the *target selection field*, it biases the competition in favor of objects that have not previously been selected.

Results

The architecture was implemented and simulated in the graphical programming framework *cedar* (Lomp, Zibner, Richter, Ranó, & Schöner, 2013), which solves what is essentially one large integro-differential equation numerically. Simulation runs differ only in perceptual input and linguistic description. The latter is supplied by the user by setting connections between the ordinal nodes of the sequence generation system and the process and concept nodes. Parameters of the model remained unchanged across simulations.

Table 1: Exemplary simulation results.

Linguistic description	Image	Target
the red horizontal rectangle		
the red below a green		
the red below the green and above the blue		
the blue below the red below the green		

We tested scenarios of varying complexity that differ in the number of features per object, the pattern of relations between objects, and the number of distractors (Table 1). The model handles the grounding of phrases about a single object with multiple features, phrases with a relation between two objects, phrases with multiple relations, as well as phrases with nested relations. In all cases, the model successfully finds an object matching the linguistic description.

Figure 3 shows detailed activation time courses of relevant nodes and fields generated while grounding the linguistic description "There is a cactus below the tent and above the camel. Find the blue object above that cactus." in the presence of the visual input depicted in the upper left of that figure.

At t_1, the sequence generation system has activated the *select target candidates intention node* and the *cactus node*, causing the locations of the three cacti to be memorized in the *target candidates field*. At t_2, the sequence generation system has activated the *select reference intention node* and the *tent node*, causing the location of the tent to be memorized in the *reference field*. It has also activated the *below node*, causing the *spatial relation field* to form peaks on the relative positions of the two target candidates that are below the tent. At t_3, the sequence generation system has activated the *reduce target candidates intention node*, which by t_4 has caused the target candidates to be reduced to those cacti that receive spatial attention (reflected in the *space attention field* and the *reduction field*), i.e., the cacti below the tent. At t_5, the sequence generation mechanism has activated the *select reference intention node* and the *camel node*, causing the camel to be stored in the *reference field*. It has also activated the *above node*, causing the *spatial relation field* to form peaks at the relative position of the target candidate that is above the camel. At t_6, the sequence generation mechanism has activated the *reduce target candidates intention node*, which by t_7 has caused the target candidates to be reduced to the cactus above the camel. At t_8, the sequence generation mechanism has activated the *select target intention node*, causing the cactus to be selected in the *target selection field* and to be stored in the mental map. At t_9, the sequence generation mechanism has activated the *select target candidates intention node* and the *blue node*, causing the three blue objects to be stored in the *target candidates field*. At t_{10}, the sequence generation mechanism has activated the *select reference intention node* in conjunction with the *from mental map node* (not in Figure 3), as well as the *cactus node*. This causes the previously grounded cactus from the mental map to be attended to, and to be stored in the *reference field*, effectively allowing the grounding of the blue target object to refer to the reference cactus that was the target of a previous grounding step. It has also activated the *above node*, causing the *spatial relation field* to form peaks on the relative position of the target candidate that is above the reference cactus. At t_{11}, the sequence generation mechanism has activated the *reduce target candidates intention node*, which by t_{12} has caused the target candidates to be reduced to the blue object which is above

Concepts

cactus	tent		camel		blue	cactus	
	below		above			above	

t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_{10} t_{11} t_{12} t_{13}

Active processes

select target candidates
reduce target candidates
select reference
select target

t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_{10} t_{11} t_{12} t_{13}

Field activations

space attention

target candidates

target selection

reference

spatial relation

reduction

target selection

s./s. mental map

·5   0   5
activation colormap

t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_{10} t_{11} t_{12} t_{13}
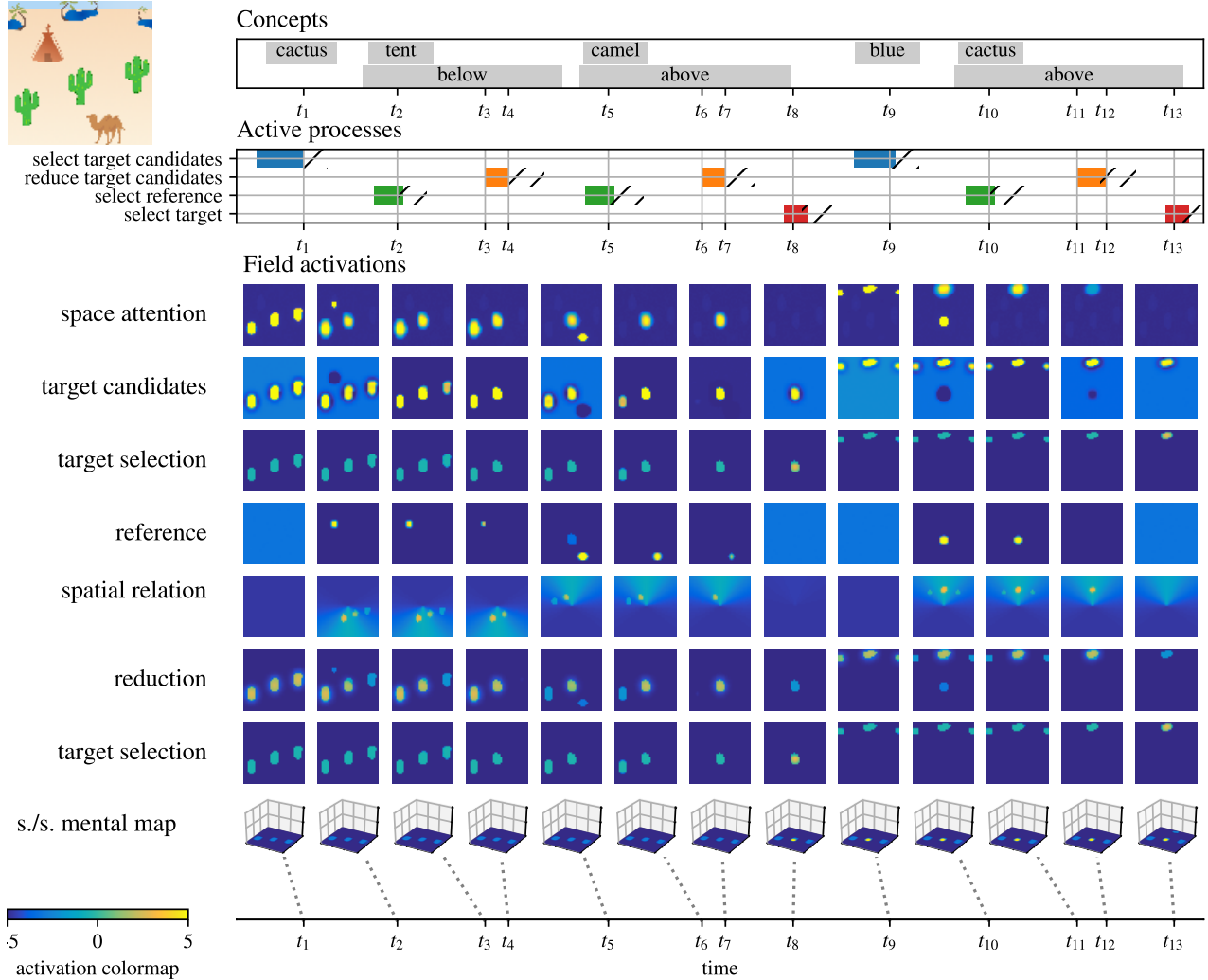time

Figure 3: Activation snapshots of parts of the architecture as it grounds a description in a visual scene (top left). Top panels: horizontal bars show time ranges during which nodes are active (gray bars: concept nodes, colored bars: intention nodes, striped bars: CoS nodes). Below: field activations are shown as color coded snapshots at discrete moments in time ($t_1, \ldots, t_{13}$).

the cactus. At t_{13}, the sequence generation mechanism has activated the *select target intention node*, causing the target object to be selected in the *target selection field*.

Discussion

We have presented a neural dynamic architecture that grounds spatial language in perception by sequentially combining concepts. The model is a single dynamical system that evolves in continuous time, and therefore solves the grounding task without algorithmic control. We addressed the two major challenges that this entails: First, the model allows for a single word to appear in different grammatical roles by coactivating its concept node with different process and parameter nodes, it allows the corresponding object representations to appear in different relational roles by representing them in distinct neural populations, and it allows to establish reference to previously grounded objects by directing atten-

tion to the mental map instead of the perceptual input. Second, the model allows for hypothesis testing by deactivating prior choices when no match is detected and using inhibition of return to avoid activation of the same hypothesis again. Beyond the simulations reported, the model can ground more complex constructions that are combinations of the phrases listed in Table 1.

The assumption that the grounding of the individual objects proceeds sequentially, one object at a time, is supported by empirical evidence. Logan (1994) found that the time it takes to ground a relation between two objects increases proportionally with the number of distractors that are distinguished from the two objects only by their spatial relation. This suggests that discriminating object pairs based on their spatial relation requires selective spatial attention, and that the consideration of different candidate pairs proceeds sequentially. Franconeri, Scimeca, Roth, Helseth, and Kahn

(2012) review further evidence that objects are attended to individually and sequentially in search tasks involving spatial relations. Additionally, Holcombe, Linares, and Vaziri-Pashkam (2011) found that applying spatial relations requires selective attention to the objects, which can only happen sequentially. Further support comes from the fact that language grounding usually proceeds in real time as a sequential linguistic representation is processed, i.e., people raise attention to the objects one by one as they are mentioned in a discourse (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

The assumption that the grounding proceeds by selecting a set of target candidates through bottom-up attention and subsequently iteratively eliminating them is also supported by empirical evidence. In a series of experiments on the concurrent discrimination of different features (color, shape, and motion) reported by Lee, Koch, and Braun (1999), it was found that different discriminations draw on the same limited attentional capacities. Thus, attending to one feature value comes at the expense of not being able to attend to another feature value, even across modalities. This makes it plausible that attention to the feature values proceeds sequentially. Moreover, Burigo and Knoeferle (2015) review attentional studies during spoken language comprehension. In these studies, it is found that upon processing a noun phrase, the words in that phrase are processed in an incremental fashion and constrain spatial attention to relevant target candidates: Initially, spatial attention is not directed. Upon hearing the first word specifying a feature value, spatial attention is divided between all objects with that feature value. Subsequently, upon hearing each new word denoting a feature value, attention narrows down to all objects that have all feature values mentioned so far.

Our current research examines the limits of the complexity the model can handle and asks if those limits may be reflected in human performance. Furthermore, we examine the interface between the model and language input.

References

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Burigo, M., & Knoeferle, P. (2015). Visual attention during spatial language comprehension. *PLoS ONE*, *10*(1), e0115758. doi: 10.1371/journal.pone.0115758

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227.

Grieben, R., Tekülve, J., Zibner, S. K., Lins, J., Schneegans, S., & Schöner, G. (2020). Scene memory and spatial inhibition in visual search. *Attention, Perception, & Psychophysics*, 1–24.

Hebb, D. (1949). *The organization of behavior*. New York, NY: Wiley Sons.

Holcombe, A. O., Linares, D., & Vaziri-Pashkam, M. (2011). Perceiving spatial relations via attentional tracking and shifting. *Current Biology*, *21*(13), 1135–1139.

Lee, D. K., Koch, C., & Braun, J. (1999). Attentional capacity is undifferentiated: Concurrent discrimination of form, color, and motion. *Perception & Psychophysics*, *61*(7), 1241–1255.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neuro-behavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *38*(6), 1490–1511.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1015-1036.

Logan, G. D., & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (pp. 493–529). Cambridge, MA: MIT Press.

Lomp, O., Zibner, S. K. U., Richter, M., Ranó, I., & Schöner, G. (2013). A software framework for cognition, embodiment, dynamics, and autonomy in robotics: cedar. In *International Conference on Artificial Neural Networks* (pp. 475–482).

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schöner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cog. Sci. Society* (pp. 2847–2852). Austin, TX: Cognitive Science Society.

Richter, M., Lins, J., & Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. *Topics in Cognitive Science*, *9*(1), 35–47.

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2457–2464). New York, NY: IEEE.

Sandamirskaya, Y., & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179.

Schneegans, S., & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109.

Schneegans, S., Spencer, J. P., Schöner, G., Hwang, S., & Hollingworth, A. (2014). Dynamic interactions between visual working memory and saccade target selection. *Journal of Vision*, *14*(11), 427–432.

Schöner, G., Spencer, J. P., & the DFT Research Group. (2015). *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York, NY: Oxford University Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.