Systematically Grounding Language through Vision in a Deep, Recurrent Neural Network

Derek D. Monner and James A. Reggia

Department of Computer Science &
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA
{dmonner,reggia}@cs.umd.edu

Abstract. Human intelligence consists largely of the ability to recognize and exploit structural systematicity in the world, relating our senses simultaneously to each other and to our cognitive state. Language abilities, in particular, require a learned mapping between the linguistic input and one's internal model of the real world. In order to demonstrate that connectionist methods excel at this task, we teach a deep, recurrent neural network—a variant of the long short-term memory (LSTM)—to ground language in a micro-world. The network integrates two inputs—a visual scene and an auditory sentence—to produce the meaning of the sentence in the context of the scene. Crucially, the network exhibits strong systematicity, recovering appropriate meanings even for novel objects and descriptions. With its ability to exploit systematic structure across modalities, this network fulfills an important prerequisite of general machine intelligence.

Keywords: deep recurrent neural network, grounded language learning, systematicity, long short-term memory

1 Introduction

An essential aspect of human intelligence is the ability to recognize and exploit key structural relations between the different modalities of our experience, from our most basic senses all the way up to the most abstract of cognitive representations. Language is one of the clearest examples of the importance of recognizing cross-sensory structural relations. When learning the verb "give" in English, for example, children recognize the correspondence between what they see during give-events—generally a giver, a gift, and a recipient—and the three noun phrases they hear near "give" in the speech stream [1]. Language is also the domain of the most widely known litmus test for a general machine intelligence, the Turing Test [2]. Though Turing argued convincingly that this is the same test we unconsciously require of other humans on a daily basis, Searle famously questioned it in his Chinese Room thought experiment [3], from which he concluded that understanding cannot follow from symbol manipulation alone. Indeed, for a system to exhibit what we call "understanding", it needs to be able

to relate its symbols to something: to ground them in sensations of the external world [4]. Our view is that, without this level of language understanding, it is hard to believe that any system could pass the Turing Test.

With this in mind, we focus on the problem of learning grounded language as a step on the road to general machine intelligence. We present a deep, recurrent neural network—a variant of the long short-term memory LSTM [5, 6]—that learns a grounded version of a micro-language by relating it to a micro-world. We choose to use a neural network because neurobiology provides the only known working example of general intelligence. That said, our network is not meant to be a veridical model of any part of the human brain. However, to leave the door open to future extensions in that direction, we attempt to maintain a reasonable degree of neurobiological plausibility.

Our neural network experiences visual scenes and, upon hearing a sentence relating to a scene, reconstructs the meaning of the speaker in terms of the objects it sees. Stated a different way, the network identifies the intended referents and relations described in a natural language sentence. The network naturally learns to segment morphemes, words, and phrases in the auditory input; to construct, maintain, and query a working-memory representation of the visual scene; and to map singular and plural noun phrases onto one or more referents. Finally and most importantly, the network behaves systematically, generalizing not only to novel scene-sentence pairs, but to individual objects and descriptions never before seen or heard.

2 Background

2.1 Systematicity

For decades researchers have debated the question of what constitutes systematic behavior in neural networks. Hadley [7] introduced a graded definition of systematicity for language tasks based on the level of input novelty that a language processing system can correctly handle. Since we are primarily interested in the *grounding* of language, below we define levels of systematicity similar to Hadley's, but based explicitly on a system's ability to pick out appropriate referents for descriptions in a sentence:

1. *Weak* systematic grounding: The system can label familiar objects in novel scenes using familiar object descriptions.
2. *Categorical* systematic grounding: The system can label novel objects in a scene using familiar descriptions; this is tantamount to categorizing the new objects.
3. *Descriptive* systematic grounding: The system can use novel descriptions to label familiar objects in a scene.
4. *Strong* systematic grounding: The system can use novel descriptions to label objects it has never previously encountered.

We will demonstrate that the network presented in this paper exhibits strong systematic grounding of the language it learns. We next turn to previous models of grounded language learning with neural networks, illustrating the level of systematicity that each has demonstrated.

2.2 Past Neural Network Models of Grounded Language Learning

Feldman and colleagues [8] famously challenged the cognitive science community to create a model of language grounded in visual sensation, and many models have addressed this task during the last two decades. We wish, for reasons of neurobiological faithfulness, to focus only on those using purely connectionist methods, rather than hybrid connectionist-symbolic [9] or other types of models. While many of the following connectionist models are impressive, we argue that they fall short of achieving the systematicity required for mastery of language.

Riga, Cangelosi, and Greco [10] advanced a neural network model, utilizing both supervised and unsupervised components, that learned to describe two-dimensional images with combinations of words. The model shows evidence of categorical systematicity by recognizing and labeling novel images; however, there is no evidence that the model can use novel combinations of descriptors for a given image. The model is limited to scenes consisting of single objects and static noun-phrase-like bit-vector descriptions, having not been designed to handle natural language in the temporal domain or at the level of sentences.

Williams and Miikkulainen [11] presented the GLIDE model, consisting of two self-organizing maps [12] that learned visual and linguistic representations of the input and then mapped them to each other. Using subjective scoring to rate the appropriateness of the model's answers, the authors found it to perform poorly on novel scenes and descriptions compared to familiar ones, and thus we cannot conclude that it is strongly systematic in its grounding abilities.

Frank, Haselager, and Rooij [13] developed a model based on a Simple Recurrent Network (SRN) [14] that learned to map temporal sequences of words representing an event onto a "situation vector" designed to analogically represent the possible states of the world. The authors claimed that their model fulfills Hadley's [7] definitions for semantic systematicity. However, interpreting the outputs produced by the model was a complex task, and often led to puzzling situations where the network appeared to simultaneously entertain contradictory beliefs about the world. As such, the level of systematicity of its language grounding is at least questionable, although we find it to be the most impressive model to date.

3 Methods

3.1 Task Description

Our network learns to ground a natural micro-language—a subset of English—in terms of a micro-world. Given input streams representing a visual scene and an

auditory sentence, the network should combine these streams in order to create an output representation of the intended meaning of the speaker. By way of explaining the task, we will describe each of the streams of information that the network must integrate: the scene represented by the *visual stream*, the sentence represented by the *auditory stream*, and the grounded meaning represented by the *intention stream*.

Scenes, Objects, and the Visual Stream On each trial, the network receives a randomly generated scene as input. A scene consists of a collection of objects and their attributes, which include shape, color, and size. Scene objects are presented to the network as neural activity patterns, but for clarity in the text we denote scene objects in a fixed-width font enclosed in square brackets, as `[small blue pyramid]`. Each object is a combination of two neural activity patterns, the first consisting of a localist representation of the object's attribute values and the second being a localist unique identifier for the object. The latter allows the network to discriminate between objects that otherwise have identical attributes, allowing the scene to contain `[large red block 1]` and `[large red block 2]` simultaneously while allowing the network to transparently refer to either.

The neural activity patterns—representing the objects in the scene—are presented to the visual input layer of the network in a temporal sequence which we call the visual stream. During training, the network's visual pathway must learn to create distributed representations that can simultaneously encode several objects, maintaining the bindings between individual objects and their (likely overlapping) attributes.

Since it is not our intention to precisely model human visual sensation and perception, we do not concern ourselves with providing a retinotopic representation of the visual stream. Instead, we assume that something like our scene representation could be constructed by higher-level visual regions in response to sensory input. We present a scene's objects as a temporal sequence in part because it allows us to vary the number of objects presented while using the same weight set to process each.

Sentences, Phonemes, and the Auditory Stream After experiencing the visual stream, the network hears a sentence that describes some aspects of the scene. Sentences are generated from a simple, mildly context-sensitive grammar (Figure 1) that describes objects from the scene and relations between them. Using the grammar, a `[small blue pyramid]` could be described as a "small blue pyramid", a "blue pyramid", a "small pyramid", or simply a "pyramid". Notably, the grammar allows plural references to groups of objects, such that our pyramid from above might be grouped with a `[small green cylinder]` to be collectively described as the "small things" because of their one common attribute.

Each word in the sentence is transcribed into a sequence of phonemes; these sequences are then concatenated, creating a single uninterrupted sequence of

S	→	**NP VP**
NP	→	the [**Size**] [**Color**] **Shape**
VP	→	**Is Where** | **Is Color** | **Is Size**
Is	→	is | are *(as appropriate for subject)*
Where	→	on **NP** | under **NP** | near **NP**
Size	→	small | medium | large
Color	→	red | blue | green
Shape	→	things | pyramid | pyramids | block | blocks | cylinder | cylinders

Fig. 1. The grammar used to generate the sentences. Terminals begin with a lowercase letter while non-terminals are in boldface and begin with an uppercase letter. The symbol | separates alternative derivations, and terms in brackets are optional. The evaluation chosen for the **Is** nonterminal depends on the plurality of its subject.

phonemes representing the entire sentence. Each phoneme in such a sequence is input to the network as a neural activity pattern representing phonetic features [15]. Since we are not trying to model the entire auditory pathway, we take it as granted that feature-based phonetic representations similar to the ones used here are available at some level in the human auditory system.

These patterns—representing the phonemes in the sentence—are presented at the auditory input layer of the network as a temporal sequence which we call the auditory stream. During training, the auditory pathway must simultaneously learn to segment the auditory stream into morphemes and words, pay attention to the syntactic relations between these elements, and discover the cues that identify objects and relations.

Meanings, Predicates, and the Intention Stream After receiving both the visual and auditory streams, the network is tasked with constructing the sentence's meaning in the context of the scene. To do this, the network must generate a sequence of predicates—as activity patterns over its output layer—which we call the intention stream.

Each predicate in the intention stream corresponds to an attribute or relation mentioned in the sentence. We denote predicates using a fixed-width font enclosed in parentheses, distinguishing them from the square-bracketed visual objects. If a sentence refers to the visual object `[small red cylinder 2]` as "small cylinder", the network must produce the predicates `(small 2)` and `(cylinder 2)`, but not `(red 2)` since this attribute was not mentioned. If a sentence states that a "blue block" (referring to visual object 3) is "under" our small cylinder, the network must output the predicate `(under 3 2)`. It may be that some objects in the scene, or even most of them, are not referenced in the sentence that accompanies it. In this case, these objects can be considered distractor stimuli, and while they are present in the visual stream input, they are not included in the target intention stream.

After a training trial, the network is shown the target intention stream. Comparing this behavior to that of a human language learner, we must assume that

the learner can, at least sometimes, derive the speaker's meaning from other sources—a task at which language learners seem to excel [16]—and that this meaning is available in something resembling a propositional form.

A Complete Example Trial Figure 2 describes an input scene, consisting of four objects, and an input phoneme sequence for the sentence "The small pyramids are on the blue block". A correct intention stream for these inputs must contain predicates denoting the objects numbered 1 and 2 as the "small pyramids". The intention stream should indicate object 4 as the referent of "the blue block", containing predicates at the end of the sequence matching these two attributes with the appropriate object identifier. For the relation "on", the intention stream must contain a predicate representing the `(on)` relation, indicating that objects 1 and 2, the pyramids, are on object 4, the block.

Visual Stream	**Auditory Stream**	**Intention Stream**
`[small red pyramid 1]`	"The small pyramids are	`(small 1+2)`
`[large blue block 4]`	on the blue block"	`(pyramid 1+2)`
`[medium red cylinder 3]`		`(on 1+2 4)`
`[small green pyramid 2]`		`(blue 4)`
		`(block 4)`

Fig. 2. An example trial. Stream elements are depicted in human-readable form, but are presented to the network as sequences of neural activity patterns representing objects, phonemes, and predicates.

3.2 Network Architecture

The neural network that learns our grounding task is a generalized long short-term memory (LSTM-g) [17], which is an extension of the long short-term memory (LSTM) network [5, 6]. LSTM uses stateful self-connected neural units called memory cells, which are allowed to have multiplicative input, output, and forget gates. LSTM-g is a formulation of LSTM that gains the ability to accommodate arbitrary multi-level network architectures without altering the learning rules.

Though the network is trained by gradient descent, and thus utilizes back-propagated error signals, we believe that the overall architecture is not as far removed from biological plausibility as one might expect. Specifically, it has been recently discovered that the gradient descent training method is essentially a convenient implementation of contrastive Hebbian learning [18], the latter being the main ingredient in biologically realistic neural training algorithms such as Leabra [19]. The fact that memory cells in LSTM maintain their state across time steps actually makes them resemble real, stateful neurons more closely than traditional stateless neural elements. Finally, the multiplicative functions of gate units in LSTM have close neurobiological correlates, and similar mechanisms have been used in models of the prefrontal cortex and basal ganglia [20].

The specific network architecture we use to learn our grounding task is depicted in Figure 3. Visual processing begins at the lower-right input layer and auditory processing at the lower-left, proceeding through one or two internal layers of self-recurrent LSTM memory cells, respectively, before integration at the final internal layer. We use two layers in the auditory pathway because the task involves multiple levels of auditory segmentation, with the first layer transforming phonemes into morphemes and words, which in the second layer become phrases. Previous experiments on learning ungrounded language representations [17] show that a two-layer pathway outperforms a single-layer pathway. To assist in the production of output sequences, the last internal layer has a recurrent connection from the previous time-step's output.

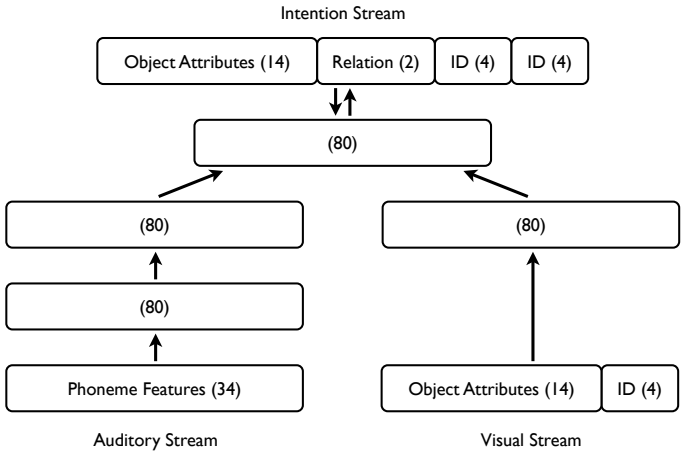

Fig. 3. The architecture of the network. Boxes represent layers of units (with number of units in parentheses) and arrows represent banks of trainable connection weights.

3.3 Experimental Evaluation

We train our network in four different ways, evaluating it on sets of test sentences that probe the different levels of grounding systematicity from Section 2.1. In what follows, an object or description is considered novel if it consists of a combination of features (e.g. `[large red pyramid]`) or words (e.g. "large red pyramid") that does not occur in the training set.

1. *Weak* condition: The set of scene-sentence pairs is partitioned at random with 10% reserved for testing. While test pairs are novel, the individual objects and descriptions are likely familiar to the network.
2. *Categorical* condition: One specific type of object is never present in scenes during training. The network is tested in situations where this novel object is given a familiar description.

3. *Descriptive* condition: One specific type of object, while allowed to be present in the scenes, is never described fully. We test the network on scenes containing this familiar object and sentences containing the full, novel description.
4. *Strong* condition: One type of object is never described *and* never appears in scenes. We test the network on inputs where this novel object appears and is referenced using a novel description.

We train 10 fresh networks in each of the above conditions. Individual units in different layers are connected with a probability of 0.7, leading to networks with approximately 60 thousand trainable weights. The learning rate is 0.01. Each network is allowed to train on 3 million randomly selected scene-sentence pairs from its training set.

For each training trial, we generate random scenes consisting of two, three, or four distinct objects, with uniform probability. We then use the grammar to generate a random sentence describing the scene. Over half a million distinct scenes are possible, each giving rise to, on average, 36 possible grammatical sentences. Since inputs, especially simple ones, are often repeated, the network sees a very small fraction of the input space during training. For each pair of test inputs, the network must produce the correct intention stream, consisting of a temporal sequence of 2 to 7 predicates.

4 Results

Figure 4 compares network accuracy across conditions. The ten networks in the *weak* condition produced correct meanings for, on average, 95% of novel scene-sentence pairs, while those in the *categorical*, *descriptive*, and *strong* conditions were 93%, 93%, and 97% accurate, respectively. The networks clearly pass all of our systematicity tests on the grounding task.

Comparing the conditions, we observed a significant difference in performance only between the *descriptive* and *strong* conditions on a Welch two-sample t-test ($t \approx -3.2$, $df \approx 17.5$, $p < 0.01$). We think this has to with an (intentional) asymmetry in the *descriptive* condition's training set. The network, observing 27 different visual objects but only 26 complete auditory object descriptions, is slightly impaired by this structural asymmetry. By contrast, in the *strong* condition, the scenes and sentences maintain their structural symmetry, with 26 visual objects corresponding to 26 complete auditory descriptions.

It is worth noting that the network had far more trouble with the grounding part of the task—that is, selecting the referents for the various object descriptions—than it had with parsing the linguistic descriptions themselves. When scoring on accurate recognition of linguistic descriptions and ignoring referents, trained networks produced, on average, less than one error per 1000 sentences. While trained networks produced the correct referent for 98% of noun phrases—with their accuracy varying inversely with the number of objects in the scene and the number of referents in the sentence, as one might expect—it also took them much longer to reach this accuracy level. A typical network required only the first quarter of its training time to reach ceiling when recognizing

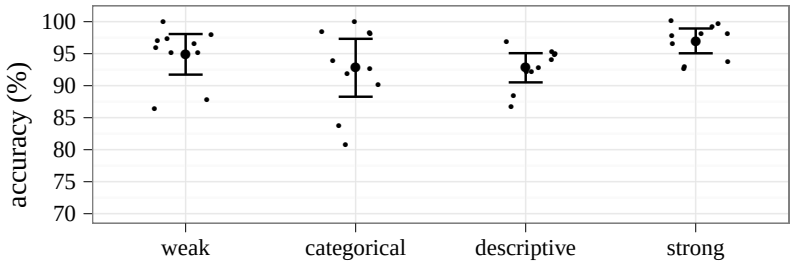

Fig. 4. The percentage of completely correct intention streams recovered from random samples of 100 test-set sentences, averaged over 10 trials in each of the conditions. The small dots represent the performance of individual networks in a condition, and large dots represent overall condition means. The error bars denote the 95% confidence intervals on the condition means.

linguistic descriptions, at which point it was identifying referents correctly only 80% of the time, a figure which slowly improved for the duration of training. That referents are so much more difficult to identify than object attributes and relations only underscores the difficulty of the language grounding task.

The network appears to scale well to larger scenes, with overall accuracy decreasing nominally as we increase the maximum number of objects in the scene to five or six—causing the number of possible scenes to exceed 400 million—while keeping the number of trainable weights constant.

5 Discussion

The results in the previous section demonstrate that the network uses grounded language systematically. We are currently analyzing the network's learned internal representations in hopes of providing a detailed explanation of how they support this systematic behavior. A key question will be whether these learned representations can be viewed as classical symbols (in some meaningful sense of the term) or are of a fundamentally different nature.

While we suggest that our network provides one the best demonstrations of strongly systematic, grounded language learning to date, we realize that it is still a long way from a general machine intelligence. However, we believe that it provides compelling evidence that connectionist methods excel at something essential to general intelligence: the ability to recognize and exploit structural systematicity in the environment across sensory modalities, relating the senses simultaneously to each other and to what we might call the internal, cognitive world. We are convinced that what we colloquially refer to as "intelligence" consists largely of the ability to discover systematicity, whether at the most basic level of our senses or at the highest levels of cognitive abstraction. Our

hope is that future work in this vein will shed light on how human intelligence is implemented *in vivo* while simultaneously bringing us closer to recreating it *in silico*.

References

1. Goldberg, A.E.: Learning Linguistic Patterns. The Psychology of Learning and Motivation 47, 33–63 (2006)
2. Turing, A.M.: Computing Machinery and Intelligence. Mind 59(236), 433–460 (1950)
3. Searle, J.R.: Minds, Brains, and Programs. Behavioral and Brain Sciences 3(03), 417–457 (1980)
4. Harnad, S.: The Symbol Grounding Problem. Physica D: Nonlinear Phenomena 42(1-3), 335–346 (1990)
5. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation 9(8), 1735–1780 (1997)
6. Gers, F.A., Cummins, F.: Learning to Forget: Continual Prediction with LSTM. Neural Computation 12(10), 2451–2471 (2000)
7. Hadley, R.: Systematicity in Connectionist Language Learning. Mind & Language (1994)
8. Feldman, J.A., Lakoff, G., Stolcke, A., Weber, S.H.: Miniature Language Acquisition: A Touchstone for Cognitive Science. In: Proceedings of the Annual Conference of the Cognitive Science Society, pp. 686–693 (1990)
9. Hadley, R.F., Cardei, V.C.: Language Acquisition from Sparse Input Without Error Feedback. Neural Networks 12(2), 217–235 (1999)
10. Riga, T., Cangelosi, A., Greco, A.: Symbol Grounding Transfer with Hybrid Self-Organizing/Supervised Neural Networks. In: Proceedings of the International Joint Conference on Neural Networks, pp. 2865–2869 (2004)
11. Williams, P., Miikkulainen, R.: Grounding Language in Descriptions of Scenes. In: Proceedings of the 8th Annual Conference of the Cognitive Science Society, pp. 2381–2386 (2006)
12. Kohonen, T.: The Self-Organizing Map. Proceedings of the IEEE 78, 1464–1480 (1990)
13. Frank, S.L., Haselager, W.F.G., van Rooij, I.: Connectionist Semantic Systematicity. Cognition 110(3), 358–79 (2009)
14. Elman, J.L.: Finding Structure in Time. Cognitive Science 14, 179–211 (1990)
15. Weems, S.A., Reggia, J.A.: Simulating Single Word Processing in the Classic Aphasia Syndromes Based on the Wernicke-Lichtheim-Geschwind Theory. Brain and Language 98(3), 291–309 (2006)
16. Tomasello, M.: Constructing a Language: A Usage-Based Theory of Language Acquisition. Harvard University Press (2003)
17. Monner, D.D., Reggia, J.A.: A Generalized LSTM-Like Training Algorithm for Second-Order Recurrent Neural Networks. Submitted (2011)
18. Xie, X., Seung, H.S.: Equivalence of Backpropagation and Contrastive Hebbian Learning in a Layered Network. Neural Computation 15(2), 441–54 (2003)
19. O'Reilly, R.C.: Generalization in Interactive Networks: The Benefits of Inhibitory Competition and Hebbian Learning. Neural Computation 13(6), 1199–1241 (2001)
20. O'Reilly, R.C., Frank, M.J.: Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. Neural Computation 18(2), 283–328 (2006)