School of Informatics

Informatics Research Review
Towards Human-level Compositional Generalization
A Neuro-symbolic Perspective

Ege Ersü
s2124950@ed.ac.uk

Abstract

Humans are capable of understanding novel sentences that are composed of known parts. Lacking the compositional generalization skills of humans, deep neural networks are shown to be highly sample inefficient and inflexible when acquiring new concepts. To make the problem explicit, we present specific benchmarks that were designed to test neural language models for human level systematicity and productivity. Then, as viable solutions to this problem, we introduce multiple classes of models from the neuro-symbolic AI literature that were shown to achieve human-level systematic generalization. The review focuses on text-based, visually grounded and interactive domains where agents are required to understand highly compositional language.

Supervisor: Pavlos Andreadis

Date: January 2020

Contents

1 Introduction

Compositionality is a conceptually loaded term that is commonly brought up in both Machine Learning and Cognitive Science. It is usually argued to be a fundamental skill that human language users have, but something that modern Machine Learning systems are missing [1]. The phrase *Compositional Language of Thought* is commonly used to emphasize the property of human thinking that produces novels thoughts using known concepts, allowing us to generalize to novel situations in a flexible and data-efficient way [2].

A famous toy example is that once a person learns the concepts: "jump", "twice" and "again", they should be able to understand unseen sentences like "jump twice and jump again" without requiring any extra supervision. Due to the immense research space regarding such a conceptual skill, this review will only introduce the problem of compositionality in a tiny set of natural language understanding tasks and report classes of models that are experimentally shown to address the issue.

Lately, there has been considerable interest in testing neural language models in terms of compositional generalization. This was mostly in response to experiments showing that end-to-end neural architectures fail to generalize in a compositional manner, especially in sequence processing tasks where compositional rules must be learned from very little data [3, 4]. The problems of sample inefficiency and the inflexibility to perform rule-based generalizations has given the neuro-symbolic Artificial Intelligence community an opportunity to show that systematic-rule learning abilities can be achieved with symbolic approaches when combined with the flexible pattern recognition capacity of state-of-the-art neural models [5]. This review will introduce classes of **neuro-symbolic models** that have achieved great success in compositional language understanding, despite being highly underrepresented in language research and industry.

First, the review will conceptually cover the problem of compositionality, introducing two computationally meaningful aspects of compositionality that language models can be tested for. Then, it will introduce the distinction between two different paradigms that are trying to address the problem: end-to-end neural models and neuro-symbolic models. The review will then proceed to introduce relevant tasks that test for compositionality and corresponding neurosymoblic models that achieve close to perfect compositional generalization. This will be done for three general domains of language understanding: text-based, visually-grounded and interactive domains.

2 The Problem of Compositionality in Language Understanding

2.1 Defining Compositionality

To make unified progress on compositionality, the Machine Learning research community needs a set of benchmarks that can be followed to show that a model is capable of understanding compositional language. Currently most research groups are operating on separate definition, making it impossible to understand whether state-of-the-art models are truly making progress on compositionality.

The disagreements show that the field needs more conceptual analysis regarding this subject as opposed to constant empirical experimentation. This review will subscribe to two clearly defined aspects of compositionality: **systematicity & productivity** [6]. This categorization does not ignore the considerable amount of impactful research addressing compositionality, but allows us to identify which specific aspect of compositionality the authors are referring to.

The most studied aspect of compositionality is **systematicity**, the ability to understand novel sentences, given that they were produced by combining already known components [7, 8]. Assume a person knows the meanings of the words "twice", "then", "and" and "again". Now if that person learns the meaning of "jump", it should be able to understand previously unseen sentences like "jump twice and then jump again" or "jump and jump and jump...", or any other one of potentially infinite combinations of known components. Similarly, if a model fails to generalize to a novel combination of already known words, we say it does not have systematic compositionality. But to test for systematicity in a Machine Learning context, such a definition would not be enough. A language model that was trained on a large dataset might only be pattern-matching to already seen similar sentences, rather than showing *zero-shot generalization* by extracting the composition rules from the training set [3]. This has motivated researchers to prepare custom benchmarks where training and test sets are adjusted to leave out certain combinations of concepts to test for systematic generalization [3, 9, 10, 11].

The second, less studied aspect of compositionality is **productivity**. Most Linguists argue for a generative view of language, claiming that infinitely many sentences can be generated from known components and rules [7]. To follow up on the previous example, the input sentence could get as long as "jump twice and then jump again and again and again and again". A language model equipped with the capacity of Productivity would then be able to understand exactly what is meant by such a sentence. Most practical Natural Language Processing tasks deal with finite sequences, so it is argued that language models should be tested for their ability to understand sentences that are longer than the ones encountered in their training set [6]. Although Productivity could be seen as a consequence of systematicity, in a practical setting it urges researchers to take sentence lengths into consideration when deciding on their training and test data splits.

2.2 End-to-end Architectures vs Neuro-Symbolic Architectures

It is generally argued that the popular success of modern Natural Language Understanding (NLU) systems are a consequence of two general ideas. The first idea is the distributional semantics hypothesis, which claims that words with similar contextual distributions will also have similar semantics. By utilizing large text corpora, models can then build these vectors as useful word representations for the task at hand. The second idea is to use large end-to-end neural networks (preferably combined with attention mechanisms) to process a sequence of words and produce a representation of the entire sentence. No matter how complex the compositional nature of a sentence might be, it is assumed that a network will able to produce a representation that fully captures the required compositional properties. Some popular models of this kind include Long-Short-Term memory networks that processes sentence tokens one by one using a recurrent neural network [12]; and more recent but highly popular Transformer models, which are built to perform massively distributed sequence processing with a complete reliance on attention mechanisms [13].

Neuro-symbolic models for Natural Language Understanding generally discard the second premise that I have introduced, claiming that fully end-to-end architectures might not have the capacity for systematicity & productivity; and that a compositional structure must be built into the model explicitly. Although the claim that connectionist networks cannot learn compositionality is usually introduced through Fodor & Pylyshyn's cognitive scientist perspective [8], there are many previous examples from within NLP research. For example Socher's work on recursive neural networks rely on collecting predefined syntactic parse trees of a sentence and then computing the representation recursively [14].

Neuro-symbolic models used in NLU tasks may still utilize the distributional semantics hypothesis to learn word representations and may also use deep learning to learn composition functions. But they additionally incorporate symbolic approaches to explicitly model the compositional nature of language. Neuro-symbolic AI has a huge scope since it uses tools coming from Linguistics, Programming languages and Logic. This review will only introduce a small subset of popular models that were shown to address the problem of compositional generalization in different domains.

3 Compositional Language Learning Tasks in Different Domains

3.1 Text-based Compositional Language Understanding

Currently most state-of-the-art NLP approaches start with a large pre-trained end-to-end Transformer model and then fine-tune it to perform a particular task. There is no doubt these models show great success in NLP benchmarks that are designed to test complex natural language understanding and they are heavily utilized in practical industry applications [15]. A recent line of research decided to take a step back from these successes and test these end-to-end neural models in terms of their compositional generalization skills.

A simple but popular dataset that is commonly used to test compositional language understanding is the **SCAN** dataset, which simulates an instruction following agent's understanding of compositional instructions in a completely text-based manner [3]. The dataset consists of instructions generated by a phrase-structure grammar and a corresponding sequence of actions. The task requires a sequence-to-sequence language model to first process the instruction sentence and then produce the corresponding sequence of actions. For example the instruction *turn left twice and jump* must be turned into to the action sequence [LEFT-TURN LEFT-TURN JUMP. The grammar can be used to generate highly composite instructions for testing. Lake and Baroni's experiments show that recurrent neural models are able to make zero-shot generalizations when the training and test sets are split randomly and when the model is allowed to

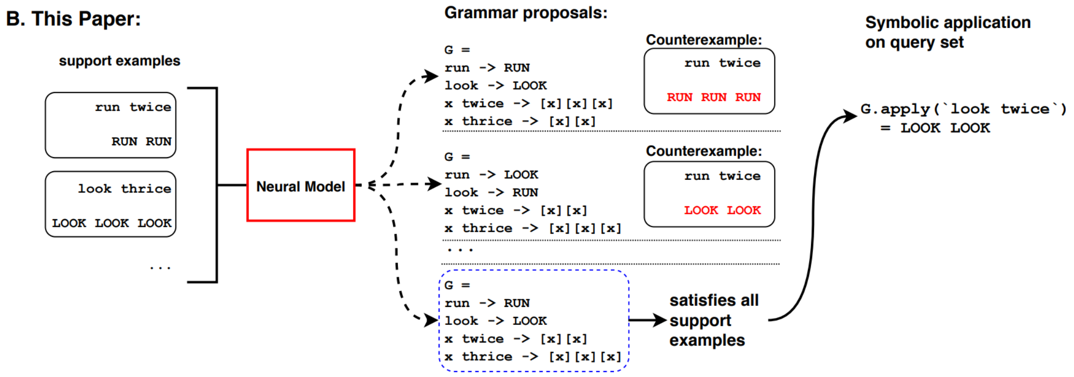


Figure 1: Neural Program Synthesis [5]

see composite instructions beforehand. But the accuracy drops drastically when the number of composite instructions in the training set is decreased and the model is forced to combine known components to understand the meaning a novel composite instruction. They also show that the model fails when the test set has composite commands that are longer than commands seen in the training set. This also shows that recurrent neural networks fail to learn productivity.

One Neuro-symbolic approach that has managed to achieve 100% accuracy on all configurations of the SCAN dataset is **Neural Program Synthesis** [5]. It is built on the criticism of end-to-end architectures, that they treat the problem as an over-general input-output mapping, which in turn requires vast amounts of data points to account for each kind of possible composition. Instead, the Neural Program Synthesis model assumes a distribution of rule systems that might have generated the data and learns to maximize the likelihood of the grammar that explains the training data. This model still uses neural LSTMs to process the input instruction and learns to produce a distribution over all possible grammars, as shown in Figure 1. The model then samples a grammar from the distribution and symbolically checks whether the training data matches it. Training is over once the proposed rule system is consistent with all prior data. Once the rule system is found, it can be applied to arbitrarily complex compositions and out of distribution data. This model fully acquires the systematicity and productivity aspects of compositionality, in context of the highly artificial SCAN dataset. But is not yet clear how Neural Program Synthesis can be expanded to more practical NLU data sets where rule systems are not as explicit.

There has been various other tests designed to test for the systematicity and productivity skills of neural language models using artificial languages, showing that all types of neural language models including Transformers fail to generalize to unseen compositions. When the same text-based data sets were used in Psycho-linguistics research, humans were shown to learn new concepts in a one-shot manner and manage to generalize to unseen compositions. It was found out that humans utilize three types of inductive biases to achieve this. The first bias **mutual exclusivity** happens when a person is presented with two objects. Once they are told that the first object is called a ”dax” and then asked which one is the ”zup”, they immediately assume that ”zup” is the second object. The second bias **one-to-one** assumes that each input symbol corresponds to exactly one output symbol. The third bias **iconic concatenation** refers to the preference for maintaining the order of input symbols in the order of the output symbols [16]. When end-to-end architectures are trained from scratch, they do not posses any of these biases and must learn all relations from scratch. Therefore it is commonly argued that building these inductive biases into NLU models are required to achieve human-level compositional language

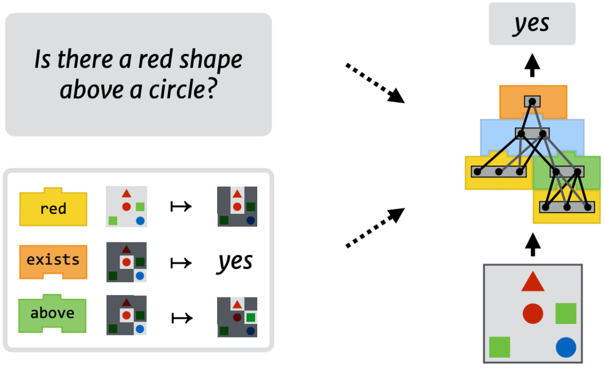


Figure 2: Neural Module Networks for VQA [18]

understanding [1, 10].

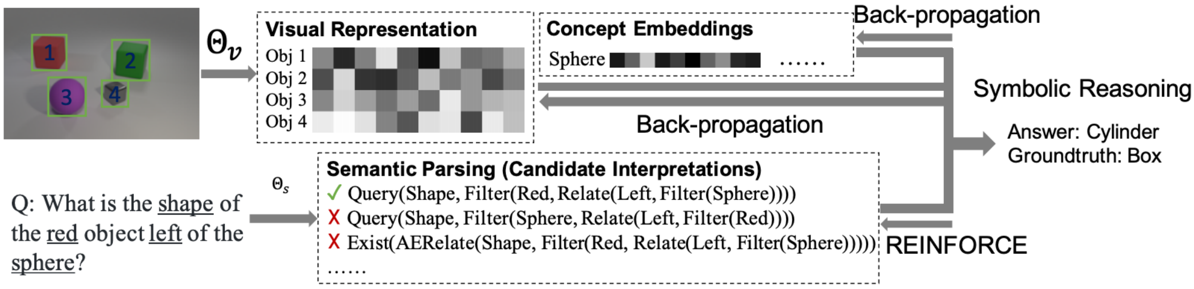
3.2 Visually Grounded Compositional Language Understanding

Understanding compositional language is not only a required skill for text-based NLP tasks, but also for multi-modal tasks where concepts and relations are visually grounded on an image [17]. Due to the immense scope of such tasks, the review will be limited to Visual Question Answering (VQA) tasks where models have to jointly reason about an image and a question about that image to produce the correct answer. The review will be further narrowed down to simpler synthetic data-sets, where questions are about a small set of physical objects, their properties and spatial relations. The goal of such experiments is not just to fit multi-modal models to specific real-world data-sets, but to make sure that models are capable of reasoning about all possible object combinations, despite being trained on a very small subset of such combinations. [17].

One such synthetic dataset is Spatial Queries On Object Pairs (SQOOP), built to test grounded NLU systems in terms of systematicity. It consists of 64x64 grids where each box may or may not have one of a large set of symbols. These images are paired with questions such as *"Is there a letter C right of a letter A?"* where questions are only composed of the relations LEFT-OF, RIGHT-OF, ABOVE, BELOW. The main challenge is not learning complicated object representations, but showing systematic generalization to unseen objects and unseen relations between such objects. This is forced by creating various difficulties of data splits, the most difficult requiring one-shot learning. This dataset does not test for productivity since it assumes a constant question length.

Several fully end-to-end neural architectures that used to be state-of-the-art on practical VQA tasks are tested to show poor systematic generalization on the SQOOP data splits. [19]. In comparison, a type of neuro-symbolic model called **Neural Module Network** achieves close to zero error. Neural Module Networks first parse a question into its dependency tree and then use that to construct a custom neural network for that specific question. The model achieves this by working with small neural networks called *modules* that are unique to each concept. For example the question *"Is there a red shape above a circle?"* will first retrieve the modules for RED, SHAPE, ABOVE, CIRCLE, EXISTS and then assemble them to produce a complete network as shown in Figure 2.

Recently there has been research on trying to bring in more popular methods from Deep Learn-



Θ_v
Visual Representation
Obj 1
Obj 2
Obj 3
Obj 4
Concept Embeddings
Sphere ……
Back-propagation
Symbolic Reasoning
Back-propagation
Answer: Cylinder
Groundtruth: Box
Θ_s
Q: What is the shape of the red object left of the sphere?
Semantic Parsing (Candidate Interpretations)
✓ Query(Shape, Filter(Red, Relate(Left, Filter(Sphere))))
X Query(Shape, Filter(Sphere, Relate(Left, Filter(Red))))
X Exist(AERelate(Shape, Filter(Red, Relate(Left, Filter(Sphere)))))
……
REINFORCE

Figure 3: Neuro-symbolic Concept Learner [22]

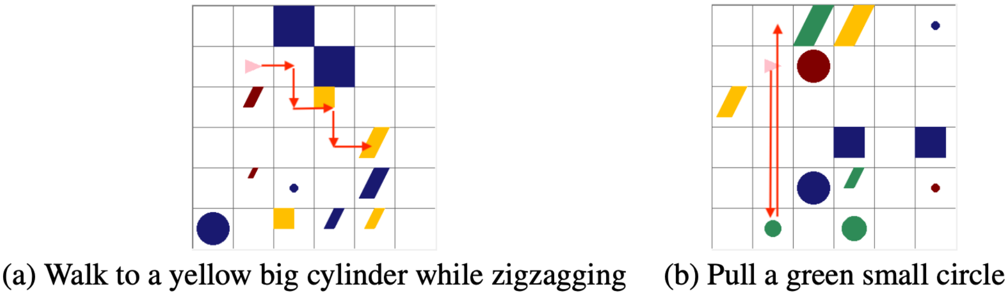
ing to this line of models. For example a more complicated variant utilizes the AND module to compute an element-wise maximum for two input attention maps and the FIND module to perform trainable convolutions on the input attention map [20]. Experiments on SQOOP & and another VQA dataset called CLEVR are utilized to argue that tree-based structures provide an advantage for systematic generalization [21]. Another advantage is that Neural Module Networks are built to acquire new concepts due to their modular structure and are much more interpretable than end-to-end architectures.

Another class of Neuro-symbolic models that has been successful in understanding visually grounded compositional questions is the **Neuro-symbolic Concept Learner (NSCL)** [22, 23]. In VQA tasks, the standard Deep Learning approach is to process both the image and the question using an end-to-end neural architecture and rely on the network for both compositional language understanding and visuospatial reasoning [19]. Instead, the NSCL divides the pipeline into three interconnected modules as in Figure 3. The high-level goal is to use neural networks for pattern recognition and symbolic programs to reason about objects and concepts.

First, the visual perception module produces an object-based representation of the scene using a pre-trained Convolutional Neural Network. Then, this object-based representation is processed by concept-specific neural networks to produce what the authors call *concept embeddings*. For example visual concepts like SHAPE or RED are still represented as vectors, and are adjusted during training. Based on the concepts, the network computes a final vector for each object, and these vectors are concatenated to produce a final representation of the scene. To process the highly compositional question, the model runs a semantic parsing module to turn the question into a symbolic program. Finally, this symbolic program is executed on the scene representation to produce the answer. All of these modules are jointly trained.

The authors argue that the model is human-like because it jointly learns words, visual concepts and semantic parsing through natural supervision. A variant of the model that also explicitly models *meta-concepts* achieves close to 100% accuracy on the CLEVR dataset, while still having a highly interpretable structure [23]. Despite the advantages, the parser assumes a domain-specific language, so using the NSCL in a different domain requires further work on designing a semantic parser and adding new concepts.

Advocates for completely end-to-end neural architectures have criticized the model for being extremely domain specific and requiring additional human labor to construct semantic parsers and to decide on which concepts will be needed for the task at hand. A recently published model that has a completely Transformer based architecture has outperformed the NSCL on similar VQA tasks such as CLEVRER & CATER [24, 25]. The jury is still out on whether neuro-symbolic models have a future in visually grounded compositional language understanding or



(a) Walk to a yellow big cylinder while zigzagging (b) Pull a green small circle

Figure 4: Grounded Instruction Following [11]

if end-to-end architectures are capable of having better compositional generalization skills.

3.3 Compositional Language Understanding in Interactive Environments

Lately, there has been a self-critical argument made by a crowd of pioneers in NLP:

> Language understanding research is held back by a failure to relate language to the physical world it describes and to the social interactions it facilitates. Despite the incredible effectiveness of language processing models to tackle tasks after being trained on text alone, successful linguistic communication relies on a shared experience of the world. It is this shared experience that makes utterances meaningful. [26].

In Philosophy, the necessity of a relation between words and the world has been identified as the *Symbol Grounding Problem* [27]. Developmental Psychology has also shown that babies learn language by utilizing multi-modal perception, incremental learning, acting and experimenting with objects in a physical environment, receiving guidance from adults and utilizing symbolic structures [28]. There is a strong case to be made that the compositional language understanding problem in machine learning may **not** be solved without training language-understanding agents in physical or virtual grounded environments [26].

To be able to study and train such agents, there has been a recent push to create virtual environments and benchmarks where agents have to utilize linguistic supervision to complete tasks such as instruction following. A simple example is given in Figure 4 where the agent has to learn actions that correspond to instructions composed of previously seen concepts like *walk, yellow, big, while, zigzagging* [11].

Methods from Reinforcement Learning (RL) and NLP are currently being combined to jointly deal with the highly compositional and relational nature of such environments and instructions [29]. But due to the scope of this review, we will only cover neuro-symbolic models that help models in terms of compositional systematicity and productivity in understanding and following grounded instructions.

The most recent and extensive benchmark that tests for systematic and productive generalization is the grounded version of the SCAN dataset: gSCAN [11]. In this task a grounded agent has to learn to process a composite instruction and act accordingly in a 2D grid world. Most instructions are novel compositions of already seen actions, properties and objects in training, as depicted in Figure 5. In addition to understanding text-based composite instructions, the

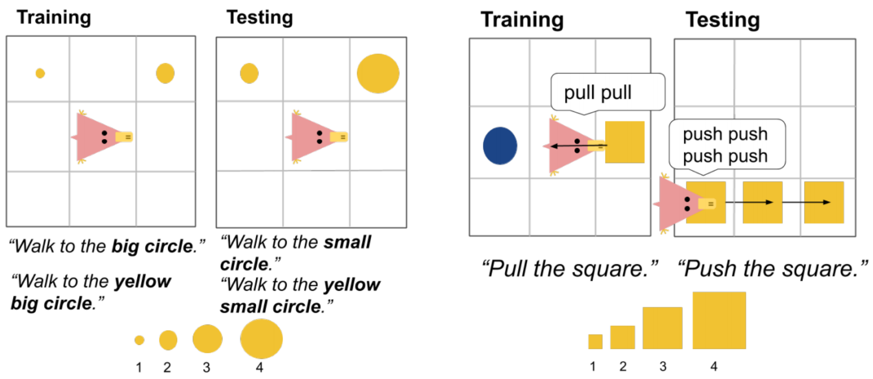

Training
Testing
“Walk to the big circle.”
“Walk to the yellow big circle.”
“Walk to the small circle.”
“Walk to the yellow small circle.”
1
2
3
4
Training
Testing
pull pull
push push push push
“Pull the square.”
“Push the square.”
1
2
3
4

Figure 5: Generalizing from calling an object “big” to calling it “small” and generalizing from pulling to pushing a heavy square. [11]

agent also has to figure out what object the ”small cylinder” refers to among many cylinders. It further has to deal with the interactive nature of the environment as both the agent and objects may move around after certain actions. The benchmark has varying difficulties and data splits to test for generalization in systematicity and productivity.

Like the SCAN dataset, neural models achieve low accuracy on splits that require compositional generalization. Although neuro-symbolic approaches manage to improve on the baseline results, there are currently no models that achieve human-level performance on difficult levels. The state-of-the-art model is a type of **Compositional Network** built on the principle that ”compositional structure of networks should reflect the compositional structure of the problem” [30]. The model first uses a constituency parser to get the parse tree of an instruction. Then, it assembles a custom neural network by following the structure of the parse tree and replacing each node with a Recurrent Neural Network corresponding to that individual word or function. Although the final model is technically a non-symbolic neural network, the compositionality is explicitly built into it according to a symbolic parse tree. There is considerable overlap with these kinds of compositional networks and neural module networks. While end-to-end architectures *wait* for compositionality to emerge, compositional networks assume that the data is going to be compositional in a certain way.

4 Summary & Conclusion

Humans can easily understand novel composite sentences as long as they are composed of already known concepts. This ability is called compositional generalization and allows language users to generate and understand novel sentences in a systematic and productive way. As opposed to humans, end-to-end neural models that are commonly used in NLP tasks are shown to be extremely data inefficient and inflexible in terms of these capacities. While a human can learn to use a new concept from a single example, neural architectures cannot generalize without seeing a huge amount of composite use cases. This causes modern NLP systems to not be able to acquire new concepts as easy as humans do. Also in practical settings, it results in AI systems not being able to follow instructions or answer questions that have a highly compositional structure.

Although compositional generalization is required for all kinds of communicative tasks, it further requires additional capabilities under different domains. While a text-based system only has access to sequences of words, a visually grounded NLU system has to figure out the referents of each word and additionally perform visuo-spatial reasoning. Language understanding in an interactive domain is even more difficult since agents and objects might move around, or their properties might change due to the actions of agents. Although each domain has its own additional issues, it is shown that neural models fail to show human-level compositional generalization in all three domains.

It must also be noted that benchmarks that were specifically designed to test for compositional language learning are not directly comparable to other practical data sets. The commonality between benchmarks like SCAN, SQOOP and gSCAN are that they enable testing based on systematicity and productivity. The systematic aspect of compositionality is tested via custom data splits that train the model on very few combinations of concepts and test the model on unseen compositions of seen concepts. The productive aspect of compositionality is tested by designing the dataset to have shorter training sentences and longer compositional testing sentences, such that it can only be solved with human-like productive generalization.

One class of models that provide such systematic and productive flexibility are neuro-symbolic models that use neural networks as pattern recognition engines and symbolic approaches to handle compositional reasoning. For text-based tasks we introduced Neural Program Synthesis models that are able to model the underlying rule structure of a dataset. Once the rule structure is learned, the model is able to understand every possible novel composition, displaying perfect systematic generalization. For visually-grounded tasks, we introduced Neural Module Networks that parse a question and then use that parse tree to assemble a neural network node by node, which is specific to that question. We have also introduced the Neuro-symbolic concept learner, which turns questions into symbolic programs and executes them against a scene representation. For interactive tasks, we introduced Compositional Networks that also assemble an instruction-specific neural architecture based on the instruction's constituency parse.

Neuro-symbolic approaches are still under heavy criticism despite being able to generalize better systematically and productively compared to end-to-end neural architectures. One major criticism is that they only work well on simple benchmarks that they were designed to solve. One major next step would be to start experimenting with these models in more practical tasks where pattern recognition plays a larger goal due to the complexity of natural language. Another related criticism is that almost all neuro-symbolic models make use of modules, parsers or rule systems that are mostly hand-crafted. For example it is not clear how a module network would be able to learn and store a module for each concept and function in natural language.

It seems to be the case that while end-to-end architectures are strong pattern recognition engines that can capture the complexity of natural language, they suffer when they have to systematically generalize and understand new composite utterances. The converse is true for neuro-symbolic models: although they show great systematic generalization, it is harder for them to adapt to the complexity of real world. Although the jury is still out whether neuro-symbolic models are the future of NLP systems, one safe conclusion is that if we want models that can generalize compositionally from only a few examples, there needs to be certain cognitive biases in place. This means further research in cognitive science is required to build the most viable cognitive models for human compositional reasoning. As better models of human compositional reasoning become available, it will undoubtedly influence machine learning models that are designed to solve similar problems.

References

[1] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people, 2016.

[2] Steven Piantadosi, Joshua Tenenbaum, and Noah Goodman. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123, 04 2016.

[3] Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks, 2018.

[4] Brenden M. Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions, 2019.

[5] Maxwell I. Nye, Armando Solar-Lezama, Joshua B. Tenenbaum, and Brenden M. Lake. Learning compositional rules via neural program synthesis, 2020.

[6] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: how do neural networks generalise?, 2020.

[7] Noam Chomsky. Syntactic structures (the hague: Mouton, 1957). *Review of Verbal Behavior by BF Skinner, Language*, 35:26–58, 1957.

[8] J. Fodor and Z. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71, 1988.

[9] Jasmijn Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. Jump to better conclusions: Scan both left and right, 2020.

[10] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. The omniglot challenge: a 3-year progress report, 2019.

[11] Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. A benchmark for systematic generalization in grounded language understanding, 2020.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

[14] Richard Socher, Cliff Lin, Andrew Ng, and Christopher Manning. Parsing natural scenes and natural language with recursive neural networks. pages 129–136, 01 2011.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[16] Laura E. de Ruiter, Anna L. Theakston, Silke Brandt, and Elena V.M. Lieven. Iconicity affects children's comprehension of complex sentences: The role of semantics, clause order, input and individual differences. *Cognition*, 171:202 – 224, 2018.

[17] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron C. Courville. Systematic generalization: What is required and can it be learned? *CoRR*, abs/1811.12889, 2018.

[18] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Deep compositional question answering with neural module networks. *CoRR*, abs/1511.02799, 2015.

[19] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017.

[20] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. *CoRR*, abs/1704.05526, 2017.

[21] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, abs/1612.06890, 2016.

[22] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *CoRR*, abs/1904.12584, 2019.

[23] Chi Han, Jiayuan Mao, Chuang Gan, Joshua B. Tenenbaum, and Jiajun Wu. Visual concept-metaconcept learning, 2020.

[24] David Ding, Felix Hill, Adam Santoro, and Matt Botvinick. Object-based attention for spatio-temporal reasoning: Outperforming neuro-symbolic models with flexible distributed architectures. *arXiv preprint arXiv:2012.08508*, 2020.

[25] Kexin Yi*, Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020.

[26] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language, 2020.

[27] Stevan Harnad. The symbol grounding problem. *CoRR*, cs.AI/9906002, 1999.

[28] L. Smith and M. Gasser. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11(1-2):13–29, 2005.

[29] Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N. Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. A survey of reinforcement learning informed by natural language. *CoRR*, abs/1906.03926, 2019.

[30] Yen-Ling Kuo, Boris Katz, and Andrei Barbu. Compositional networks enable systematic generalization for grounded language understanding, 2020.