Compositional generalization in a deep seq2seq model by separating syntax and semantics

Jake Russin[1], Jason Jo[2], Randall C. O'Reilly[1], and Yoshua Bengio[2,3]

[1]Department of Psychology and Neuroscience , University of Colorado Boulder
{jacob.russin, randy.oreilly}@colorado.edu
[2]MILA, Université de Montréal
[3]CIFAR Senior Fellow

Abstract

Standard methods in deep learning for natural language processing fail to capture the compositional structure of human language that allows for systematic generalization outside of the training distribution. However, human learners readily generalize in this way, e.g. by applying known grammatical rules to novel words. Inspired by work in neuroscience suggesting separate brain systems for syntactic and semantic processing, we implement a modification to standard approaches in neural machine translation, imposing an analogous separation. The novel model, which we call Syntactic Attention, substantially outperforms standard methods in deep learning on the SCAN dataset, a compositional generalization task, without any hand-engineered features or additional supervision. Our work suggests that separating syntactic from semantic learning may be a useful heuristic for capturing compositional structure.

1 Introduction

A crucial property underlying the expressive power of human language is its systematicity (Lake et al., 2016; Fodor and Pylyshyn, 1988): syntactic or grammatical rules allow arbitrary elements to be combined in novel ways, making the number of sentences possible in a language to be exponential in the number of its basic elements. Recent work has shown that standard deep learning methods in natural language processing fail to capture this important property: when tested on unseen combinations of known elements, state-of-the-art models fail to generalize (Lake and Baroni, 2017; Loula et al., 2018; Bastings et al., 2018). It has been suggested that this failure represents a major deficiency of current deep learning models, especially when they are compared to human learners (Marcus, 2018; Lake et al., 2016).

A recently published dataset called SCAN (Lake and Baroni, 2017) (**S**implified version of the **C**omm**AI** **N**avigation tasks), tests compositional generalization in a sequence-to-sequence (seq2seq) setting by systematically holding out of the training set all inputs containing a basic primitive verb ("jump"), and testing on sequences containing that verb. Success on this difficult problem requires models to generalize knowledge gained about the other primitive verbs ("walk", "run" and "look") to the novel verb "jump," without having seen "jump" in any but the most basic context ("jump" $\rightarrow$ JUMP). It is trivial for human learners to generalize in this way (e.g. if I tell you that "dax" is a verb, you can generalize its usage to all kinds of constructions, like "dax twice and then dax again", without even knowing what the word means) (Lake and Baroni, 2017). However, standard seq2seq models fail miserably on this task, with the best-reported model (a gated recurrent unit augmented with an attention mechanism) achieving only 12.5% accuracy on the test set (Lake and Baroni, 2017; Bastings et al., 2018).

Perhaps these results present an opportunity to learn from research in neuroscience: the human brain must be implementing principles that allow humans to generalize systematically, but which are lacking in current deep learning models. One prominent idea from neuroscience research on language processing that may offer such a principle is that the brain contains partially separate systems for processing syntax and semantics. The part of the prefrontal cortex responsible for language production, called Broca's area, is thought to be important for parsing syntactic information, and applying selective attention in order to help a separate comprehension system interpret semantics (Thompson-Schill, 2004). Variants of this idea have been implemented in computational models of the prefrontal cortex (O'Reilly and Frank, 2006;

Kriete et al., 2013; Miller and Cohen, 2001), but have not been taken up in deep learning research.

In this paper, we present a simple implementation of this idea within the neural machine translation framework, and test our model on the SCAN dataset. Our novel model, which we call Syntactic Attention, encodes syntactic and semantic information in separate streams before producing output sequences. Preliminary experiments show that our novel architecture achieves substantially improved compositional generalization performance on the SCAN dataset.

2 SCAN dataset

The SCAN dataset is composed of sequences of commands that must be mapped to sequences of actions (Lake and Baroni, 2017). The dataset is generated from a simple finite phrase-structure grammar that allows for systematic splits of the training and testing data. These splits include the following:

- Simple split: training and testing data are split randomly
- Length split: training includes only shorter sequences
- Add primitive split: a primitive command (e.g. "turn left" or "jump") is held out of the training set, except in its most basic form (e.g. "jump" $\rightarrow$ JUMP)

Here we focus on the most difficult problem in the SCAN dataset, the add-jump split, where "jump" is held out of the training set. The best test accuracy reported in the original paper (Lake and Baroni, 2017), using standard seq2seq models, was 1.2%. More recent work, which tested more sophisticated seq2seq models, reports a test accuracy of 12.5% (a gated recurrent unit augmented with attention) on the "jump" split (Bastings et al., 2018). To our knowledge, this is the current best reported accuracy on this split of the SCAN dataset.

3 Syntactic attention

The Syntactic Attention model improves the compositional generalization capability of an existing attention mechanism (Bahdanau et al., 2014) by implementing two separate streams of information processing for syntax and semantics (see Figure 1). Here, by "semantics" we mean the information in each word in the input that determines its

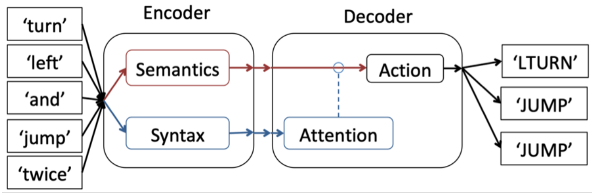

Figure 1: Syntactic Attention architecture. Syntactic and semantic information are maintained in separate streams.

meaning (in terms of target outputs), and by "syntax" we mean the information contained in the input sequence that should determine the *alignment* of input to target words. We describe the mechanisms of this separation in detail below, following the notation of (Bahdanau et al., 2014), where possible.

3.1 Encoder

The encoder produces two separate vector representations for each word in the input sequence. Unlike the previous attention model (Bahdanau et al., 2014)), we separately extract the semantic information from each word with a linear transformation:

$$m_j = W_m x_j$$

where W_m is a learned weight matrix that multiplies the one-hot encodings $\{x_1, ..., x_{T_x}\}$. Note that the semantic representation of each word does not contain any information about the other words in the sentence. As in the previous attention mechanism (Bahdanau et al., 2014), we use a bidirectional RNN (biRNN) to extract what we now interpret as the syntactic information from each word in the input sequence. The biRNN produces a vector for each word on the forward pass, $(\overrightarrow{h_1}, ..., \overrightarrow{h_{T_x}})$, and a vector for each word on the backward pass, $(\overleftarrow{h_1}, ..., \overleftarrow{h_{T_x}})$. The syntactic information (or "annotations" (Bahdanau et al., 2014)) of each word x_j is determined by the two vectors $\overrightarrow{h_{j-1}}$, $\overleftarrow{h_{j+1}}$ corresponding to the words surrounding it:

$$h_j = [\overrightarrow{h_{j-1}}; \overleftarrow{h_{j+1}}]$$

In all experiments, we used a bidirectional LSTM for this purpose. Note that because there is no sequence information in the semantic representations, all of the information required to parse (i.e. align) the input sequence correctly (e.g. phrase structure, modifying relationships, etc.) must be encoded by the biRNN.

3.2 Decoder

The decoder models the conditional probability of each target word given the input: $p(y_i|\mathbf{x})$, where y_i is the target translation and $\mathbf{x}$ is the whole input sequence. As in the previous model, we use an RNN to determine an attention distribution over the inputs at each time step (i.e. to align words in the input to the current target). However, our decoder diverges from this model in that the mapping from inputs to outputs is performed from a weighted average of the *semantic* representations of the input words:

$$d_i = \sum_{j=1}^{T_x} \alpha_{ij} m_j$$

$$p(y_i|\mathbf{x}) = f(d_i)$$

where f is parameterized by an arbitrary function, and the α_{ij} are the weights determined by the attention model. In all experiments we parameterized f with a linear function and a softmax nonlinearity. The attention weights are computed by a function measuring how well the syntactic information of a given word in the input sequence aligns with the current hidden state of the decoder RNN, s_i:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

where e_{ij} can be thought of as measuring the importance of a given input word x_j to the current target word y_i, and s_{i-1} is the current hidden state of the decoder RNN. Bahdanau et al. (2014) model the function a with a feedforward network, but following (Hudson and Manning, 2018), we choose to use a simple dot product:

$$a(s_{i-1}, h_j) = s_{i-1} \cdot h_j$$

, relying on the end-to-end backpropagation during training to allow the model to learn to make appropriate use of this function. Finally, the hidden state of the RNN is updated with the same weighted combination of the *syntactic* representations of the inputs:

$$s_i = g(s_{i-1}, c_i)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

where g is the decoder RNN, s_i is the current hidden state, and c_i can be thought of as the information in the attended words that can be used to determine what to attend to on the next time step. Again, in all experiments an LSTM was used.

4 Experiments

After an informal hyperparameter search, our best model used an LSTM with 2 layers and 200 hidden units in the encoder, an LSTM with 1 layer and 400 hidden units in the decoder, and a dropout rate of 0.5. Following (Bastings et al., 2018) we use early stopping by validating on a 20% held out sample of the training set. Each model was trained for 200,000 iterations with a batch size of 1, using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001. Each architecture was trained 5 times with different random seeds for initialization, to measure variability in results. All experiments were implemented in PyTorch.

4.1 Results

The Syntactic Attention model achieves state-of-the-art performance on the key compositional generalization task of the SCAN dataset (see table 1). The model shows strong compositional generalization performance, attaining a mean accuracy of 91.1% (median = 98.5%) on the test set of the add-jump split. This represents a substantial improvement over the best previously reported result on this task (Bastings et al., 2018), and does so without any hand-engineered features or additional supervision.

4.2 Ablations

To test our hypothesis that compositional generalization requires a separation between syntax (i.e. sequential information used for alignment), and semantics (i.e. the mapping from individual source words to individual targets), we conducted two experiments:

- Sequential semantics. An additional biLSTM was used to process the semantics of the sentence:

  $$m_j = [\overrightarrow{m_j}; \overleftarrow{m_j}]$$

  , where $\overrightarrow{m_j}$ and $\overleftarrow{m_j}$ are the vectors produced for the source word x_j by a biLSTM on the forward and backward passes, respectively.

- Syntactic action. Syntactic information was allowed to directly influence the output at

Model	**Simple**	**Add turn left**	**Add jump**
GRU + attn	100.0 $\pm$ 0.0	59.1 $\pm$ 16.8	12.5 $\pm$ 6.6
GRU + attn - dep	100.0 $\pm$ 0.0	90.8 $\pm$ 3.6	0.7 $\pm$ 0.4
Syntactic Attention	100.0$\pm$ 0.0	98.3 $\pm$ 0.03	**91.1** $\pm$ **10.3**

Table 1: Mean accuracy (%) and standard deviations on test splits of SCAN dataset. Best reported models (Bastings et al., 2018) compared to Syntactic Attention model.

each time step in the decoder:

$$p(y_i|\mathbf{x}) = f([d_i; c_i])$$

, where again f is parameterized with a linear function and a softmax output nonlinearity.

The results of the ablation experiments are shown in table 2. These results confirmed our hypothesis. Performance on the test set was worse when the strict separation between syntax and semantics was violated, but interestingly, this was worse in the case where semantics were processed with an additional biLSTM (adding syntactic/sequential information to the semantic representations).

5 Related work

Our model integrates ideas from computational and cognitive neuroscience (Thompson-Schill, 2004; O'Reilly and Frank, 2006; Kriete et al., 2013; Miller and Cohen, 2001), into the neural machine translation framework. This work has been influenced by previous successes in neural machine translation (Bahdanau et al., 2014), as well as previous work studying compositionality in neural networks in other settings, such as visual-question answering (Hudson and Manning, 2018).

5.1 Neuroscience

It has long been thought that humans possess specialized cognitive machinery for learning the syntactic or grammatical structure of language (Chomsky, 1957). One of the most foundational findings in all of cognitive neuroscience was that lesions to a particular area of frontal cortex, called Broca's area, results in a profound deficit in the fluid production of sentences, while leaving comprehension relatively intact (Thompson-Schill, 2004). Broca's area was subsequently found to be important for parsing syntactically complex sentences, leading some to conclude that it is important for syntactic processing in general (Caramazza and Zurif, 1976). A more recent view adds nuance to this idea and situates the functioning of Broca's area within the context of prefrontal cortex in general, noting that it may simply be an area of prefrontal cortex specialized for language (Thompson-Schill, 2004). The prefrontal cortex is known to be important for cognitive control, or the active maintenance of top-down attentional signals that bias processing in other areas of the brain (Miller and Cohen, 2001). In this framework, Broca's area can be thought of as an area of prefrontal cortex specialized for language, and responsible for selectively attending to linguistic representations housed in other areas of the brain (Thompson-Schill, 2004). The prefrontal cortex has received much attention from computational neuroscientists (Miller and Cohen, 2001; O'Reilly and Frank, 2006), and one model even showed a capacity for compositional generalization (Kriete et al., 2013). The current work was heavily inspired by these key principles from cognitive and computational neuroscience research on Broca's area and the prefrontal cortex.

5.2 Neural machine translation

In neural machine translation, a neural network is trained end-to-end to produce outputs in the target language from inputs in the source language (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014). Much of this work uses an encoder-decoder framework, where one recurrent neural network (RNN) is used to encode the source sentence, and then a decoder neural network decodes the representations given by the encoder to produce the words in the target sentence (Sutskever et al., 2014). Earlier work attempted to encode the source sentence into a single fixed-length vector (the final hidden state of the encoder RNN), but it was subsequently shown that better performance could be achieved by encoding each word in the source, and using an attention mechanism to *align* these encodings with each target word during the decoding process (Bahdanau et al., 2014). The current work builds directly on this attention

Model (ablation)	**Add jump**
Sequential semantics	42.3 ± 32.7
Syntactic action	88.7 ± 14.2
Syntactic Attention	91.1 ± 10.3

Table 2: Mean accuracy (%) and standard deviations on test splits of SCAN dataset. Ablations compared to full Syntactic Attention model.

model, while incorporating a separation between syntactic and semantic information streams.

5.3 Compositionality

The principle of compositionality has recently regained the attention of deep learning researchers (Lake et al., 2016; Lake and Baroni, 2017; Battaglia et al., 2018; Johnson et al., 2016) . In particular, the issue has been explored in the visual-question answering (VQA) setting (Hudson and Manning, 2018; Johnson et al., 2016; Perez et al., 2017; Hu et al., 2017; Santoro et al., 2017; Yi et al., 2018). Many of the successful models in this setting learn hand-coded operations (Hu et al., 2017), use highly specialized components (Hudson and Manning, 2018; Santoro et al., 2017), or use additional supervision (Hu et al., 2017; Yi et al., 2018). In contrast, our model uses standard recurrent networks and simply imposes the additional constraint that syntactic and semantic information are processed in separate streams. However, this work has greatly influenced the design and implementation of our model.

6 Conclusion

The Syntactic Attention model incorporates ideas from cognitive and computational neuroscience into the neural machine translation framework, and produces the kind of systematic generalization thought to be a key component of human language-learning and intelligence. The key feature of the architecture is the separation of sequential information used for alignment (syntax) from information used for mapping individual inputs to outputs (semantics). This separation allows the model to generalize the usage of a word with known syntax to all of its valid grammatical constructions. The success of this approach points to the untapped potential of incorporating ideas from cognitive science and neuroscience into modern appraches in deep learning and artificial intelligence (Marblestone et al., 2016).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*.

Joost Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. 2018. Jump to better conclusions: SCAN both left and right. *arXiv:1809.04640 [cs]*.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261 [cs, stat]*.

Alfonso Caramazza and Edgar B. Zurif. 1976. Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language*, 3(4):572–582.

N. Chomsky, editor. 1957. *Syntactic Structures*. Mouton & Co., The Hague.

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

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to Reason: End-to-End Module Networks for Visual Question Answering. *arXiv:1704.05526 [cs]*.

Drew A. Hudson and Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. *arXiv:1803.03067 [cs]*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. *arXiv:1612.06890 [cs]*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

T. Kriete, D. C. Noelle, J. D. Cohen, and R. C. O'Reilly. 2013. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, 110(41):16390–16395.

Brenden M. Lake and Marco Baroni. 2017. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv:1711.00350 [cs]*.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2016. Building machines that learn and think like people. *arXiv:1604.00289 [cs, stat]*.

João Loula, Marco Baroni, and Brenden M. Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. *arXiv:1807.07545 [cs]*.

Adam H. Marblestone, Greg Wayne, and Konrad P. Kording. 2016. Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience*, 10.

Gary Marcus. 2018. Deep learning: A critical appraisal.

E. K. Miller and J. D. Cohen. 2001. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202.

R. C. O'Reilly and Michael J. Frank. 2006. Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2017. FiLM: Visual Reasoning with a General Conditioning Layer. *arXiv:1709.07871 [cs, stat]*.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *arXiv:1706.01427 [cs]*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *arXiv:1409.3215 [cs]*.

Sharon L. Thompson-Schill. 2004. Dissecting the language organ : A new look at the role of Broca ' s area in language processing.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. 2018. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. *arXiv:1810.02338 [cs]*.