Compositionality as Directional Consistency in Sequential Neural Networks

Najoung Kim
Department of Cognitive Science
Johns Hopkins University
n.kim@jhu.edu

Tal Linzen
Department of Cognitive Science
Johns Hopkins University
tal.linzen@jhu.edu

Abstract

Sequential neural networks have shown success on a variety of natural language tasks, but through what internal mechanisms they achieve systematic compositionality crucial to language understanding is still an open question. In particular, gated networks such as Gated Recurrent Units (GRUs) are known to significantly outperform Simple Recurrent Neural Networks (SRNs). We conduct an exploratory study comparing the abilities of SRNs and GRUs to make compositional generalizations, using adjective semantics as testing ground. Our results demonstrate that GRUs are able to generalize more systematically than SRNs. On analyzing the learned representations, we find that GRUs encode the compositional contribution of adjectives as more directionally consistent. This consistency correlates with generalization accuracy within GRU models, suggesting that directional invariance is an effective strategy for deriving more compositionally generalizable representations.

1 Introduction

The impressive performance of neural network models in natural language processing (NLP), a domain in which symbolic representations are traditionally viewed as indispensable, raises the question of how these models represent (or approximate) symbolic compositionality. Among sequential neural networks, gated models such as Long Short-Term Memory (LSTMs) [1] and Gated Recurrent Units (GRUs) [2] dominate Simple Recurrent Neural Networks (SRNs) in a range of sequence modeling tasks [3] including language [4], and achieve better zero-shot compositional generalizations [5, 6]. In this paper, we conduct an exploratory study comparing SRNs and GRUs in their ability to make compositional semantic generalizations, using adjective semantics as our testing ground. We furthermore investigate whether compositionally useful invariance can manifest as geometric regularity in the sequence representations encoded by these networks, similarly to [7] where consistent vector offsets denote the same linguistic relation between pairs of words in the embedding space.

1.1 Related work

Our work shares motivation with recent works on neural network analyses aiming to 'open the black box' [8], especially regarding the issue of compositionality [5, 9, 10, 11]. Prevalent approaches for analyzing neural models for NLP include auxiliary classifiers, challenge sets, and adversarial perturbation, which target specific linguistic properties [12]. Although such methods serve as useful probes for gauging what the models are capable (or incapable) of, they provide limited insights into their operational behavior. We directly analyze the properties of model-internal representations, along the lines of [13, 14]. We employ Natural Language Inference (NLI) [15] as a task format, similarly to [16, 17, 18], and also draw from works that use synthetic datasets for conducting focused evaluations of various linguistic phenomena [14, 19, 20, 21]. Our work shares topical interests with NLP literature on semantic compositionality [22, 23], logical reasoning [24, 25], and adjective semantics [26].

Preprint. Under review.

2 Methodology

2.1 Dataset for testing compositional semantic generalization

We design a task that tests a model's capacity to make compositional semantic generalizations adopting the widely-used NLI [15] setup. The input consists of a premise-hypothesis (p/h) pair, and the task is to predict the entailment relationship between p/h. We use a binary version of this task, where the labels are {*entailed, not entailed*}, to reduce theoretical semantic complications (c.f., [27]). Solving our task is contingent upon correctly understanding the effect of adjectives on the entailment pattern between p/h. The dataset consists of training, development and generalization sets, where the generalization set contains classes of examples not shown during training (zero-shot), but are such that we expect a model that makes human-like compositional generalizations to be able to solve. In particular, we target two types of generalization patterns: (1) generalization to unseen sequence types, and (2) generalization from complex to simpler compositional forms. See Table 1 for examples.

Table 1: Example p/h pairs from the dataset. → denotes 'entails' and ↛ denotes 'does not entail' (not all template types are listed, due to space constraints).

	Training/Development sets	
Template	Example	
$Adj_1\ Adj_2\ N \rightarrow Adj_2\ N$	Mary is a tall American lawyer. → Mary is an American lawyer.	
$Adj_1\ Adj_2\ N \nrightarrow Adj_2\ N$	Mary is a former American lawyer. ↛ Mary is an American lawyer.	
$Adj_1\ Adj_2\ N \rightarrow N$	Mary is a tall American lawyer. → Mary is a lawyer.	
$Adj_1\ Adj_2\ N \nrightarrow N$	Mary is a former American lawyer. ↛ Mary is a lawyer.	
	Generalization set	
Template	Example	
$Adj_1\ Adj_2\ N \rightarrow Adj_2\ N$	Mary is a tall former lawyer. → Mary is a former lawyer.	(unseen)
$Adj_1\ Adj_2\ N \nrightarrow Adj_1\ N$	Mary is a tall former lawyer ↛ Mary is a tall lawyer.	(unseen)
$Adj\ N \rightarrow N$	Mary is a tall lawyer. → Mary is a lawyer.	(complex to simple)
$Adj\ N \nrightarrow N$	Mary is a former lawyer. ↛ Mary is a lawyer.	(complex to simple)

Training set. We use adjective classes, namely subsective and nonsubsective adjectives, that give rise to different entailments [28, 29]. With subsective adjectives, *Adj N* entails *N* (e.g., *tall president → president*), whereas with nonsubsective adjectives, *Adj N* does not entail *N* (e.g., *fake president ↛ president*). We exploit this distinction to generate an NLI dataset where the entailment relation depends on the understanding of the adjective semantics. The training set is constructed so that there are no cases where the subsectivity of a single adjective with respect to the modified noun is transparently given. That is, there are no input pairs such as *John is a former teacher ↛ John is a teacher* or *John is a tall teacher → John is a teacher*); the training set strictly contains cases where the premise uses nested forms (i.e., $Adj_1\ Adj_2\ N$).

Generalization set. The generalization set tests for the following two types of generalizations, which we would plausibly expect from a model that has successfully learned the semantic contribution of individual adjectives from the training set.

- **Generalization to unseen sequence.** The generalization set contains adjective sequences that are not seen during training (but the individual adjectives are). For instance, *tall American* and *former American* both appear in the training set, but *tall former* is unseen.
- **Generalization from complex to simple form.** The set also contains examples that require teasing apart the individual contributions of each adjective meaning, which is not explicitly given in the training/development sets. For instance, *tall x → x*, but *former x ↛ x*.

Generation. We use templates *Subj is a* $Adj_1\ Adj_2\ N$ →/↛ {*Subj is a* $Adj_1\ N$, *Subj is a* $Adj_2\ N$} to generate training data, using 12 different subsective adjectives (half in Adj_1 position and half in Adj_2 position) and 4 nonsubsective adjectives (only seen in Adj_1 position in training). We use 9 different noun phrases that can appear in $Subj$ position, which can be either one or two words long to keep the length of the whole sequence variable (e.g., *Mary, my dad*). We use 10 different nouns that appear in the N position, which are single words that are potentially modified by the adjectives (e.g., *president, student*). We also add two trivial cases: (1) self-entailment ($X \rightarrow X$), and (2) non-entailment of subject-mismatched p/h (e.g., *x is a z* $\nrightarrow$ *y is a z*). 23,400 unique pairs are generated through this process, 15% of which are used as a development set ($|train| = 19,890, |dev| = 3,510$). For the generalization set, we use the same templates but with nonsubsectives adjectives in Adj_2 position in the premise, for generating the unseen sequence cases. A new template *Subj is a Adj N* $\rightarrow$/$\nrightarrow$ *Subj is a N* is used for the complex to simple form generalization cases. This process yields $|test| = 15,120$.

2.2 Geometric measures

We test the hypothesis that geometric consistency could be a way of systematically encoding invariance of the compositional semantic contribution of adjectives. For instance, we would expect an adjective such as $former$ to have some common meaning shared across different linguistic contexts it appears in, rather than carrying an idiosyncratic meaning in every use. One way this context-invariant semantics could be captured is by projecting the representation in a similar direction whenever the model encounters $former$. This is not the only possible way; the global meaning could be represented via displacement by the same amount (i.e., same magnitude regardless of direction). We consider both possibilities: the direction and magnitude consistency of vector offsets. The consistency of a given word w is defined as follows. For all sentences in the test set that contain w, take the last hidden state h_n of their encoding. Then take a version of each sentence with w removed, and take the last hidden state h'_{n-1} of that sentence. The vector offset of the ith sentence that contains w is defined as $o_i = h_n - h'_{n-1}$, where n is the length of sentence i. Then the directional consistency θ_w of a word w is defined as the average pairwise cosine similarity for all o (Eq. 1), and magnitude consistency ι_w is defined as the average pairwise absolute difference in Euclidean norms for all o (Eq. 2), where N is the total number of sentences containing w.

$$\theta_w = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{o_i \cdot o_j}{\|o_i\| \|o_j\|}}{N} \qquad (1)$$

$$\iota_w = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \big| \|o_i\| - \|o_j\| \big|}{N} \qquad (2)$$

3 Experiments

Models. We use a Siamese recurrent classifier architecture similar to [6], which uses the same recurrent network to encode p and h, and passes the concatenated encodings (we use the last hidden state) of the two sentences to the classification layer as in [15]. We use AllenNLP [30] to implement our models. For the recurrent units, we test SRNs and GRUs with a single hidden layer. The input dimension is fixed to 300, and the hidden dimension of the recurrent units is varied between $h = \{8, 16, 32, 64, 128, 256, 512\}$. Word embeddings are initialized using Xavier initalization, the default setting in AllenNLP. The classifier is a single feedforward layer with linear activation followed by a softmax, which takes $2h$ dimensional inputs.

Training. We train models on the entailment dataset for a maximum of 50 epochs using stochastic gradient descent (learning rate=0.1, batch size=16), early stopping when the development set accuracy does not improve for 5 epochs. In practice, most models reached peak development accuracy within 10 epochs. We run each model with the same hyperparameters with 10 different random initializations.

Results. Both SRN and GRU models were able to learn the train/development sets perfectly, with small variations depending on random restarts (SRN: 0.96 (± 0.05) (train), 0.98 (± 0.03) (dev), GRU: 0.99 (± 0.01) (train), 0.9999 (± 0.0001) (dev)). However, SRN and GRU models significantly differed in their generalization accuracy (Mann-Whitney $U = 3231, p < .001$)—GRU models on average achieved near-perfect accuracy (0.97), whereas SRNs did not (0.69). No single SRN model generalized perfectly (highest accuracy = 0.87).

GRU models encoded adjectives' compositional contributions with higher directional consistency ($U = 3025, p < .001, |\Delta| = 0.29$) (see Figure 1 for an illustration), and also with smaller difference in magnitude ($U = 119, p < .001, |\Delta| = 1.47$). Within GRU models, we found a significant correlation between generalization set accuracy and directional consistency of adjective encodings (Pearson's

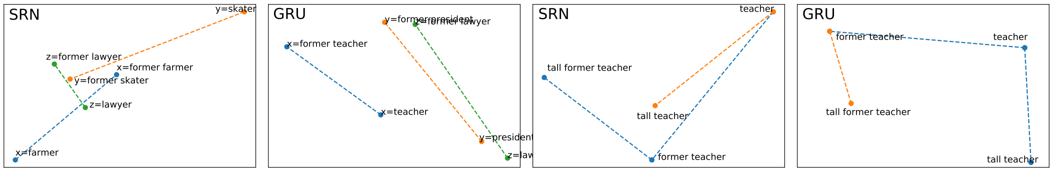

Figure 1: Directional consistency of adjectives in SRN and GRU models.

$r = 0.54, p < .001$), but not between accuracy and magnitude consistency ($r = -0.00, p = 0.99$). Within SRNs, we observe an inverse correlation between directional consistency of adjectives and accuracy ($r = -0.69, p < .001$). This negative effect was largely driven by a cluster of models that had below majority-class accuracy (< 0.69) (see Figure 3, far left). The inverse correlation no longer holds if we exclude models in this cluster ($r = 0.32, p > .05$ after multiple-comparisons correction).

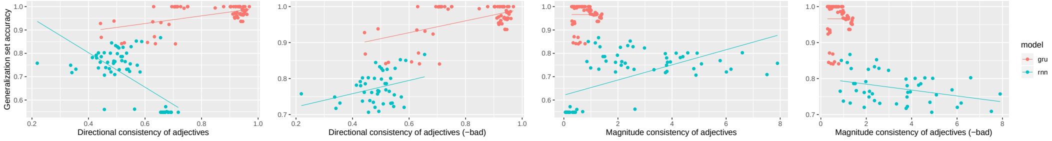

Figure 2: Accuracy plotted against consistency measures with the line of best fit by model group. Additional plots are shown for data excluding models with accuracy below majority-class.

Table 2: Pearson's correlation between consistency measures and generalization set accuracy. The p-values are adjusted using Holm-Sidak correction. ($^{*} = p < .05, ^{**} = p < .01, ^{***} = p < .001$)

Model	n	Accuracy	Corr(acc, dir.)			Corr(acc, magn.)		
			All	Adj.	N	All	Adj.	N
SRN	70	0.69 (±0.11)	0.21	-0.69^{***}	0.44^{**}	0.59^{***}	0.59^{***}	0.53^{***}
GRU	70	0.97 (±0.04)	0.52^{***}	0.54^{***}	-0.12	-0.10	-0.00	-0.41^{*}

Our findings can be summarized as follows. SRNs and GRUs both could learn the training data perfectly, but their capacity to make systematic generalizations differed greatly. GRUs encoded the contribution of adjectives to the sentence in a more geometrically consistent manner, with respect to both direction and magnitude. Within GRU models (but not within SRNs), more directional consistency positively correlated with higher generalization accuracy.

4 Conclusion

We investigated the difference between SRNs and GRUs in their capacity to make compositional semantic generalizations. Preliminary results suggest that SRNs and GRUs employ qualitatively different approaches for solving the same task, and the strategy GRUs adopt proves more effective for making systematic generalizations. Furthermore, we observe that the representations GRUs develop display more geometric regularity across different linguistic contexts, measured by the average direction and magnitude consistency of the compositional contributions of adjective modification. Directional regularity in particular seems to facilitate systematic generalization for GRUs, suggested by the significant within-GRU correlation between directional consistency and generalization accuracy.

What is the nature of the architectural bias that gives rise to this discrepancy? One insight can be drawn from [31], which makes an empirical remark about the importance of a forget gate. We could speculate that the 'forgetting' mechanism encourages models to discard contextual information (if it is useful to do so), biasing models towards developing more globally invariant representations of lexical items. Exploring this hypothesis further would be an interesting follow-up, elucidating the roles of different architectural components in representing compositionality. More broadly, we plan to investigate whether we could inject bias into the models for learning more compositionally generalizable respresentations, and extend the scope of our work to more naturalistic datasets.

References

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

[2] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

[3] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[4] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, Aug 2018.

[5] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2879–2888, 2018.

[6] Mathijs Mul and Willem Zuidema. Siamese recurrent networks learn first-order logic reasoning and exhibit zero-shot compositional generalization. *arXiv preprint arXiv:1906.00180*, 2019.

[7] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[8] Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes. Proceedings of the 2019 ACL workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019.

[9] Jacob Andreas. Measuring compositionality in representation learning. In *International Conference on Learning Representations*, 2019.

[10] Joris Baan, Jana Leible, Mitja Nikolaus, David Rau, Dennis Ulmer, Tim Baumgärtner, Dieuwke Hupkes, and Elia Bruni. On the realization of compositionality in neural networks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 127–137, Florence, Italy, August 2019. Association for Computational Linguistics.

[11] Kris Korrel, Dieuwke Hupkes, Verna Dankers, and Elia Bruni. Transcoding compositionally: Using attention to find more generalizable solutions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 1–11, Florence, Italy, August 2019. Association for Computational Linguistics.

[12] Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics (TACL)*, 7:49–72, 2019.

[13] Brendan Whitaker, Denis Newman-Griffis, Aparajita Haldar, Hakan Ferhatosmanoglu, and Eric Fosler-Lussier. Characterizing the impact of geometric properties of word embeddings on task performance. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 8–17, Minneapolis, USA, June 2019. Association for Computational Linguistics.

[14] Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[16] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[17] Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[18] Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[19] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[20] Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[21] Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. Studying the inductive biases of RNNs with synthetic variations of natural languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3532–3542, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[22] Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 12–21, Beijing, China, July 2015. Association for Computational Linguistics.

[23] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[24] Richard Evans, David Saxton, David Amos, Pushmeet Kohli, and Edward Grefenstette. Can neural networks understand logical entailment? In *International Conference on Learning Representations*, 2018.

[25] Sara Veldhoen and Willem Zuidema. Can neural networks learn logical reasoning? *CLASP Papers in Computational Linguistics*, page 34, 2017.

[26] Ellie Pavlick and Chris Callison-Burch. Most “babies" are “little" and most “problems" are “huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2164–2173, 2016.

[27] Barbara Partee. Are there privative adjectives? In *Conference on the Philosophy of Terry Parsons, University of Massachusetts, Amherst*, 2003.

[28] Barbara Partee. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360, 1995.

[29] Neha Nayak, Mark Kowarsky, Gabor Angeli, and Christopher D Manning. A dictionary of nonsubsective adjectives. Technical report, 2014.

[30] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[31] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning*, pages 2342–2350, 2015.