SYSTEMATIC GENERALIZATION EMERGES IN SEQ2SEQ MODELS WITH VARIABILITY IN DATA

Prabhu Prakash Kagitha
Akaike Technologies
Bengaluru, India
prabhu@akaiketech.com

ABSTRACT

Systematic generalization requires a model to generalize to novel combinations of already learned concepts and is a concern for many models that aim to learn and act efficiently in the world the way humans do. Tested for systematic generalization on the SCAN data set, it is shown that standard sequence to sequence (seq2seq) models cannot generalize to input-output pair (jump left, LTURN JUMP) after perfectly learning the pairs (walk left, LTURN WALK), (walk, WALK) and (jump, JUMP). In this work, we show that the standard seq2seq model (LSTM with attention) learns an invariant concept of a modifier (like 'left') if we increase the number of distinct primitives it operates on. This is analogous to human learning where empirical studies showed that stimulus variability helps in finding invariant structure (Gómez, 2002). With 300 distinct primitives, the model learned a subset of modifiers in the SCAN data set exhibiting systematic generalization. We then devised a surgery task to analyze the representations of the commands (like 'walk left') from the models trained with a different number of primitives and found that systematic generalization is strongly correlated with instance independent representations of modifiers. Surprisingly, models trained to learn modifiers on primitive variables (e.g. (jump left, LTURN JUMP)) showed generalization, though limited, to compound variables (e.g. ([jump left] left, LTURN [LTURN JUMP])) which is also human-like. Code used in the experiments is publicly available.[1]

1 INTRODUCTION AND RELATED WORK

Systematicity is one of the long-standing arguments against connectionist theories explaining cognition, as opposed to classical computational theories that have mental representations that are structurally composed and processes that are structure sensitive to explain systematicity (Fodor et al., 1988). More generally, this argument favors symbolic AI rather than connectionist approaches. See Minsky (1991) discussing the strengths and weaknesses of each program.

Lake & Baroni (2017) evaluated standard seq2seq RNN models, including GRU (Cho et al. (2014)), LSTM (Hochreiter & Schmidhuber (1997)) with/without attention (Bahdanau et al. (2014)), on an instruction following data set SCAN that they proposed. It is shown that all the standard seq2seq models fail in zero-shot generalization of novel combinations of concepts that are already learned. For example, models trained on the input-output pairs (walk left, LTURN WALK), (walk, WALK) and (jump, JUMP) failed at generalizing (jump left, LTURN JUMP).

Since this evaluation, it is shown that a compositional data augmentation strategy (Andreas, 2019) and use of CNNs instead of RNNs (Dessì & Baroni, 2019) could improve systematic generalization considerably. Russin et al. (2019) devised a novel attention mechanism which directly attends to the semantics separated from the syntax information i.e. at a particular time step, decoder attends to the meanings of the words across different positions in the input sequence rather than attending to all the sequential information. In Lake (2019), the meaning of a primitive is disassociated through meta learning by wrongly and randomly mapping their meanings in each task making the seq2seq

[1] https://github.com/prakashkagitha/Sys-Gen-with-Variability

model constrained to learn the right association through memory. This work effectively transforms the problem of a zero-shot generalization of novel combinations of concepts into a problem of associating the right meaning to a particular primitive through memory and thus learns an abstract concept of a variable. Recently, Gordon et al. (2020) proposed to solve systematic generalization in SCAN with RNNs equipped with the property of permutation equivariance.

Never before have been conducted a study to investigate the questions including, Do standard sequence to sequence models exhibit systematic generalization at all in any setting what so ever? If yes, what can we learn from this naturally emerging systematic generalization?

Section 2. shows that standard LSTM + Attention model exhibits systematic generalization in learning a subset of SCAN data set with 300 distinct primitives in the train set. Section 3. devises a surgery to show that the emergent systematic generalization is strongly correlated with the instance independent representations of the modifiers. Section 4. shows that the models trained to learn modifiers over primitive variables also showed generalization, though limited, to compound variables.

2 Naturally emerging systematic generalization

SCAN data set has 3 primitives in the train set ('walk', 'look', 'run'), 8 modifiers operating over each primitive ('left', 'right', 'apposite right', 'apposite left', 'around right', 'around left', 'twice', 'thrice') and 2 conjunctions ('and', 'after') to combine two action sequences thus produced. 'jump' primitive occurs only in the form ('jump', 'JUMP') combined with no modifier or conjunction in the train set. Lake & Baroni (2017) reports that LSTM + attn model (with 100 hidden units, attention and one layer) is the best model generalizing to combinations of 'jump' primitive with modifiers and conjunctions with about 2.5% accuracy. We use this best model in our experiments and as in Loula et al. (2018), we do not include 'turn' commands.

Investigating the drivers of systematicity in an instruction following situated agent, Hill et al. (2019), after training all the objects without negation (e.g. lift X) and a subset with negation (e.g. lift a not X), reported that the ability for an agent to understand negation as a modifier depends on the number of the subset of objects trained with negation. The agent trained on 100 different objects with negation achieved a good enough accuracy of 78%.

Hypothesizing that this natural driver of systematicity, which emerges with increasing the number of distinct entities an operator is operated on, is applicable to seq2seq learning as well, we make the standard seq2seq model learn all the modifiers which operate only on the primitives i.e, totalling 6 ('left', 'right', 'opposite left', 'opposite right', 'around left' and 'around right') with an increased number of distinct primitives in the train set on which these modifiers are operated.

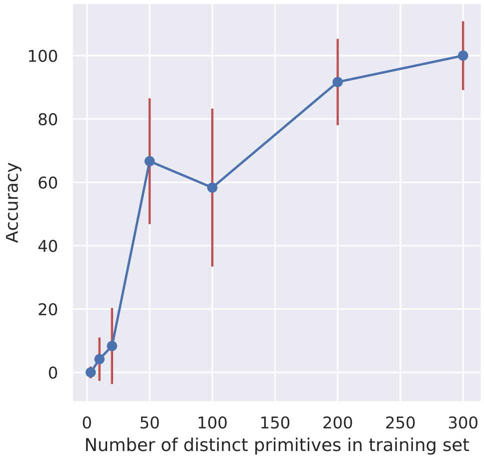

Figure 1: Generalization to new primitives. Median accuracy and standard deviation across five runs.

We trained the standard model (LSTM + attn) gradually increasing the number of distinct primitives starting from 3, the number of distinct primitives in the SCAN data set, up to 300. We tested all the models with four distinct primitives combined with 6 modifiers (totaling 24 data points) which occurred when training only in the form ('jump', 'JUMP'). See Figure 1. for the test accuracy of different models.

We trained the model to learn multipliers ('twice' and 'thrice') as well, which are operated over both primitive and compound variables, but the model managed to achieve only 40-50% accuracy at test even with 500 primitives. So, to go beyond the modifiers we studied here, engineering special inductive biases for systematic generalization might be the only way, at which the diverse work of Russin et al. (2019), Lake (2019), and Gordon et al. (2020) make strides.

3 Systematic generalization by learning instance independent representations of modifiers

Humans learn a modifier as a rule/formula which is operated on any compatible primitive (Lake et al., 2019). But the meaning of a modifier is independent of any primitive it had operated on i.e. the meaning of a modifier is instance independent.

Evaluating the models trained on relatively smaller number of distinct primitives, the output for ’jump around left’ is most of the times ’LTURN WALK LTURN WALK LTURN WALK LTURN WALK’, where ’WALK’ represents a primitive from the train set. We hypothesize that this is because the learned meaning of a modifier is instance dependent (here, depends on ’walk’, a primitive which the modifier has already operated on). Trained with around 20 distinct primitives, the model acknowledges the presence of a modifier (’around left’) with the right syntactic structure, thus producing the right direction commands, but, as it only has access to the meaning of modifier which is depended on one of the primitives it had already operated on, it produces a random/incorrect primitive where it supposed to.

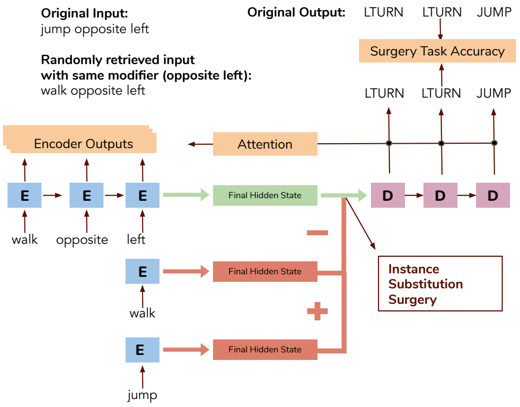

Figure 2: Instance Substitution Surgery and surgery task accuracy.

Quantitatively, model trained on 20 distinct primitives would get 100% accuracy if we don’t account for the correctness of the primitives. 87% for 3 primitives. This character is observed for models trained on SCAN data set as well, with 3 distinct primitives. We found that 73 percent of predictions (5638 of 7706) at test only made the mistake of producing a wrong primitive, with directions and length perfect. (Analysis is on the best model of five runs, to show the ability to learn.)

We deduce that models learn the syntactic correctness of a modifier’s usage with just a few distinct primitives but can’t afford to tease out a rule/formula for a modifier, by having instance independent representation, as it hadn’t seen enough primitives operated over by a modifier yet. A considerable part of achieving systematic generalization is just achieving the instance independent representations for modifiers.

We quantify the instance independence of modifier representations with a surgery task which is a variant of the standard inference. First, to process an original data point in the test set (for e.g, ’jump around left’), a data point from the train set is retrieved with the same modifier (e.g. ’walk around left’). Then, the encoder processes the retrieved data point. But, at the point of transferring encoder’s last hidden state to the decoder we do Instance Substitution Surgery(ISS) which subtracts the representation of retrieved primitive and adds the representation of the original primitive. And the model is expected to produce the action sequence as if it had processed the original data point. Figure 2 explains ISS and Figure 3 shows the surgery accuracy.

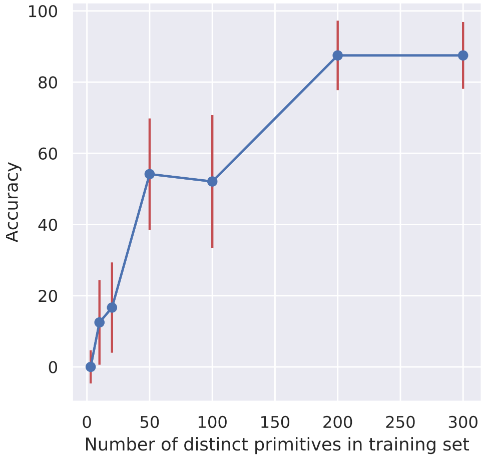

Figure 3: Median surgery task accuracy with standard deviation of different models across five runs and 10 encoder initializations.

This surgery is not a required condition for the model to exhibit instance independent representations for modifiers because of the fact that this surgery needs the encoder to represent commands as an additive composition of concepts. But a Pearson coefficient of 0.99 between accuracy when testing for systematic generalization and the surgery task accuracy of different models validates both the statements that the encoder representations do exhibit the property of composing concepts by addition and that the instance independent representation of modifiers is a good proxy to quantify systematic generalization.

Although, the current instantiation of Instance Substitution Surgery(ISS) doesn't scale when there are more than one primitive or modifier in the command (e.g. 'jump twice and walk left'). May this be the case, but we conjecture that instance independent representation of modifiers might still be one of the properties that correlates strongly with systematic generalization.

4 Systematic generalization of modifiers over compound variables (limited)

Surprisingly, we saw another richer aspect of generalization in models exhibiting natural systematic generalization. The models trained to learn modifiers over primitive variables also generalized to compound variables in some cases.

We devised a test set containing the simplest modifier operating over the simplest compound variable, which are commands in the pattern of 'primitive {left/right} {left/right}'. Models showed drastically different behavior towards different type of commands consisting compound variables. Models showed an increase in the accuracy, with the number of distinct primitives in the train set, for commands in the form 'primitive left left' and 'primitive right right' (see Figure 4), but showed zero or near zero accuracy for commands in the form 'primitive left right' and 'primitive right left'. Reason for this poor performance could be that the decoder needs to generate 'LTURN' after 'RTURN' and vice versa which it never did in the train set and also didn't observe many variations of it, as opposed to primitives, to learn it systematically.

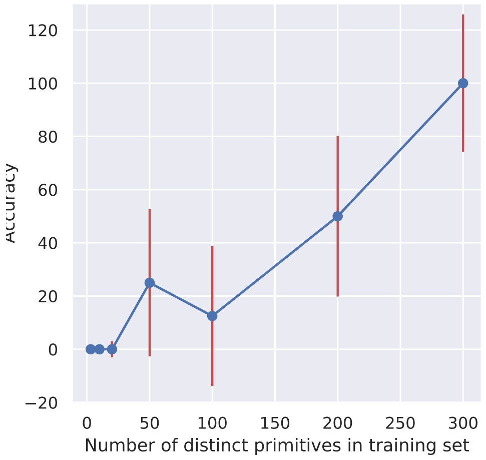

Figure 4: Median accuracy and standard deviation on commands of form 'primitive left left' and 'primitive right right' of different models across five runs and 10 encoder initializations.

We evaluated models with syntactic attention (Russin et al., 2019) and models trained with meta seq2seq learning (Lake, 2019) on compound variables. Models trained with 3 primitives (4 for meta seq2seq learning model) which solved the entire SCAN data set successfully showed no sign of generalization over compound variables. The more serious issue that we hold against both of these models is that even when the number of distinct primitives is 300 the generalization accuracy over compound variables is zero. Our concern is that these inductive biases on top of a standard model suppressed the aspects of generalization that a standard model already has. (Analysis on both the models contains at least 3 different runs that are perfect at primitive test set).

5 Conclusions and future work

Naturally emerging systematic generalization of a standard seq2seq model with variability in data commends instance independent representation of modifiers, an insight that could direct sensible engineering of inductive biases to enable systematic generalization that go beyond the SCAN data set.

Outside of the seq2seq learning task, systematic generalization is a concern for many tasks where we need instance independence like language understanding, abstract and analogical reasoning, semantic scene analysis etc.

References

Jacob Andreas. Good-enough compositional data augmentation. *arXiv preprint arXiv:1904.09545*, 2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Roberto Dessì and Marco Baroni. Cnns found to jump around more skillfully than rnns: Compositional generalization in seq2seq convolutional networks. *arXiv preprint arXiv:1905.08527*, 2019.

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

Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. Permutation equivariant models for compositional generalization in language. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SylVNerFvr`.

Rebecca L. Gómez. Variability and detection of invariant structure. *Psychological Science*, 13 (5):431–436, 2002. doi: 10.1111/1467-9280.00476. URL `https://doi.org/10.1111/1467-9280.00476`. PMID: 12219809.

Felix Hill, Andrew Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. Environmental drivers of systematicity and generalization in a situated agent. *arXiv preprint arXiv:1910.00571*, 2019.

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

Brenden M Lake. Compositional generalization through meta sequence-to-sequence learning. In *Advances in Neural Information Processing Systems*, pp. 9788–9798, 2019.

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

Brenden M Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*, 2019.

Joao Loula, Marco Baroni, and Brenden M Lake. Rearranging the familiar: Testing compositional generalization in recurrent networks. *arXiv preprint arXiv:1807.07545*, 2018.

Marvin L Minsky. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI magazine*, 12(2):34–34, 1991.

Jake Russin, Jason Jo, and Randall C O'Reilly. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *arXiv preprint arXiv:1904.09708*, 2019.