Syntax-Guided Transformers: Elevating Compositional Generalization and Grounding in Multimodal Environments

Danial Kamali, Parisa Kordjamshidi
Department of Computer Science and Engineering, Michigan State University
kamalida@msu.edu, kordjams@msu.edu

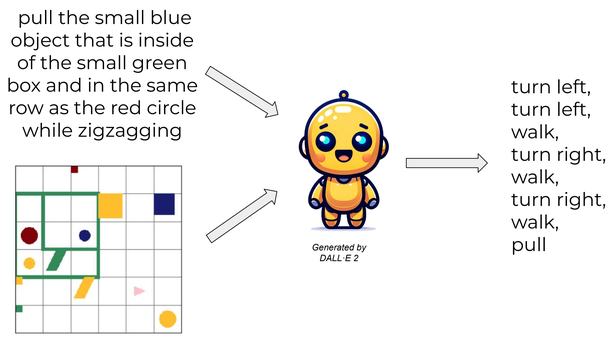
Highlights

- **Research questions**
  - How to help transformers generalize to higher reasoning depths in multi-modal grounding task?
  - Can syntactic structure of language help with compositional generalization of multi-modal transformers?
  - How do different parsing approaches influence compositional generalization capabilities?
- **Outcomes:**
  - We used attention masking guided by syntactic parsers to help compositional generalization and grounding.
  - Compared various syntactic parsing methods, assessing their impact.
  - Integrated weight sharing to alleviate the gradient vanishing issue caused by attention masking in transformer.

Problem Setting

Task (Object Grounding & Agent Navigation)

The goal is to comprehend and apply language commands in a multimodal setting.



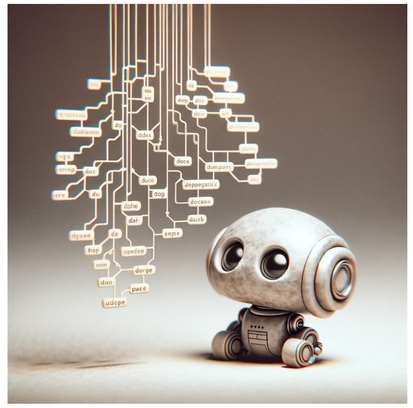
Compositional Learning Challenges

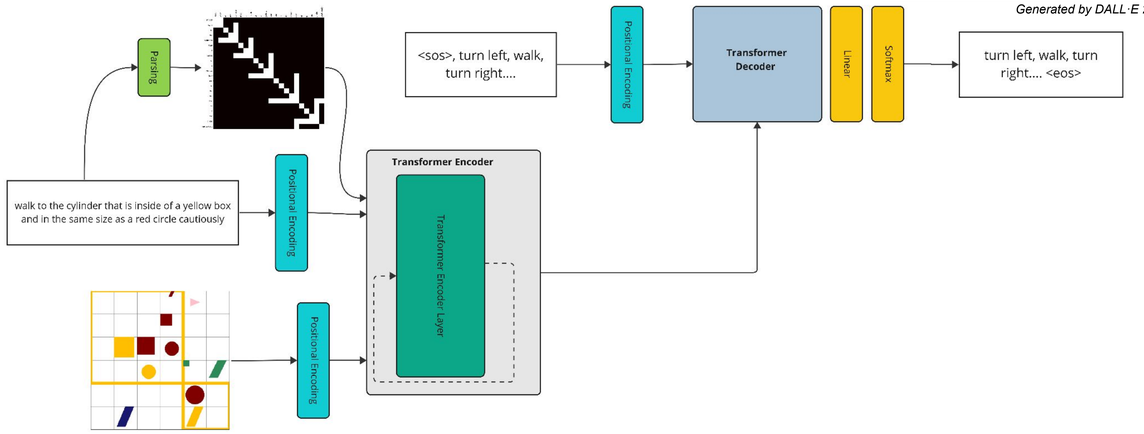
We evaluate compositional generalization capabilities, such as understanding and combining known words and concepts in novel ways **unseen in training**.

Split	Held-out Examples
Random	Random.
A1	*yellow square* referred with color & shape.
A2	*red square* referred in the command.
A3	*small cylinder* referred with size and shape
B1	co-occur of *small red circle* and *big blue square*.
B2	co-occur of *same size as* and *inside of* relations.
C1	Additional conjunction clause depth added to *2-relative-clause commands*.
C2	*2-relative-clause* command with *that is* instead of *and*.

ReaSCAN dataset test splits.

Dependency-parsing-guided Attention Masking along with Weight Sharing enhances structural generalization, while boosting efficiency in language to vision grounding.





Motivation

Syntactic Structure as a Key to Generalization:

- Utilizing readily available parsers to infer hints about the underlying syntactic structure.
- Removing connection instead of adding complexity

Efficacy With Weight Sharing:

- Addressing the backpropagation challenges in attention masking methods through weight sharing.
- Enhancing efficiency in model performance.

Method

Syntax-guided attention masking

Masking self-attention weights of tokens that are not syntactically related .

- **Dependency Parsing**: Represents relationship between tokens.
- **Constituency Parsing**: Represents hierarchical relationships among sentence parts.

Weight Sharing:

Sharing transformer encoder weights.

- Reduces parameters
- Helps with gradient vanishing

Results

Model	A1	A2	A3	B1	B2	C1	C2	Avg
LSTM*	50.4	14.7	50.9	52.2	39.4	49.7	25.7	40.40
GCN-LSTM	92.3	42.1	87.5	69.7	52.8	57.0	22.1	60.50
Transformer*	96.7	58.9	93.3	79.8	59.3	75.9	25.5	69.90
GroCoT	99.6	93.1	98.9	93.9	86.0	76.3	**27.3**	82.2
Constituency†	**99.75**±0.11	96.70±1.40	**99.68**±0.10	95.19±1.17	88.37±1.50	69.07±0.60	27.00±0.54	82.25±0.63
Dependency†	99.65±0.9	**97.37**±0.48	99.62±0.07	**95.46**±2.01	**90.15**±3.88	**92.55**±1.51	21.77±5.25	**85.22**±0.87

The result of our proposed model on the ReaSCAN dataset test splits. The results are an average of three runs. † denotes the models with masking. Models marked with * refer to the multimodal version of their implementation.

Ablation Study

W/S	Mask	A1	A2	A3	B1	B2	C1	C2	Avg
-	-	99.29±0.27	91.82±6.50	98.49±1.17	93.50±0.85	83.15±1.41	75.85±1.35	**25.03**±6.82	81.02±0.22
✓	-	99.68±0.22	97.09±1.72	99.64±0.20	94.86±0.77	81.49±4.27	66.30±6.65	21.66±1.83	80.10±1.08
-	Dep.	98.09±0.27	85.21±6.85	97.35±0.75	93.61±2.75	90.62±1.59	75.27±1.77	21.91±1.63	80.29±1.43
✓	Dep.	**99.65**±0.9	**97.37**±0.48	**99.62**±0.07	**95.46**±2.01	**90.15**±3.88	**92.55**±1.51	21.77±5.25	**85.22**±0.87

The ablation study result of our modifications on ReaSCAN dataset test splits. Results are reported on an average of three runs. We evaluate every combination of components from our best model. W/S stands for weight sharing, and the ✓shows the presence of the module. *Dep* in this table refers to the Dependency masking. We evaluate the model with or without dependency masking in the masking part.

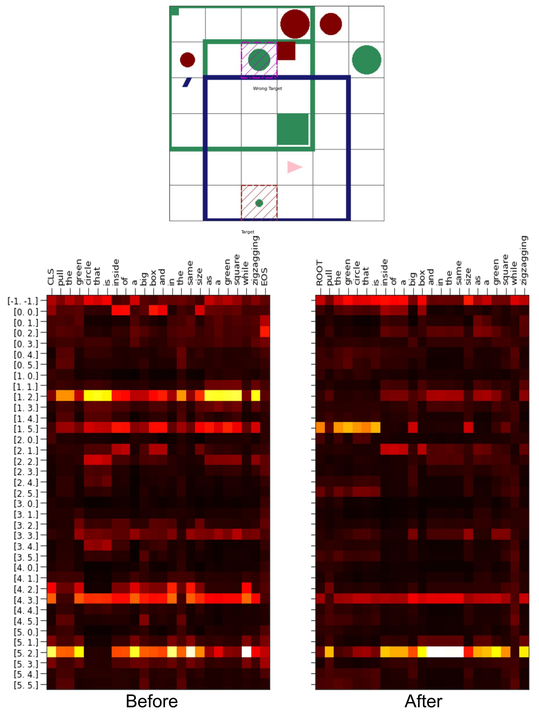
Efficacy Analysis

Model	#Parameters
Multimodal LSTM	74K
Multimodal Transformer	3M
GroCoT	4.6M
Dependency† (ours)	1.9M

Comparing model parameters: our model vs. current state-of-the-art models. Dependency† refers to the model with dependency parsing for attention masking.

Qualitative Analysis

Pull the green circle that is inside of a big box and in the same size as a green square while zigzagging



- In 86% of validation samples, the cross-attention module showed a significant focus on the target object after attention masking.
- Masking led to a sparser distribution of attention.
  - Rather than individual words focusing on every relevant cell, they now form compositional groups, focusing collectively on specific cells.

Take away messages

- Exploiting syntactic structure with weight sharing in Transformer encoders significantly improves generalization.
- Using Dependency parsing was more effective than constituency parsing.
- Using weight sharing with dependency parsing alleviates the backpropagation problem caused by attention masking.

REFERENCES

[1] Sikarwar, A., Patel, A., & Goyal, N. (2022). When can transformers ground and compose: Insights from compositional generalization benchmarks. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 648–669.
[2] Ruis, L., Andreas, J., Baroni, M., Bouchacourt, D., & Lake, B. M. (2020). A benchmark for systematic generalization in grounded language understanding.
[3] Gao, T., Huang, Q., & Mooney, R. (2020). Systematic generalization on gSCAN with language conditioned embedding. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp. 491–503.
[4] Qiu, Y., Zhang, J., & Zhou, J. (2021). Improving gradient-based adversarial training for text classification by contrastive learning and auto-encoder. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1698–1707.
[5] Kim, J., Ravikumar, P., Ainslie, J., & Ontañón, S. (2021). Improving compositional generalization in classification tasks via structure annotations.