A Benchmark for Systematic Generalization in Grounded Language Understanding

gSCAN

Laura Ruis *Jacob Andreas* *Marco Baroni* *Diane Bouchacourt* *Brenden Lake*

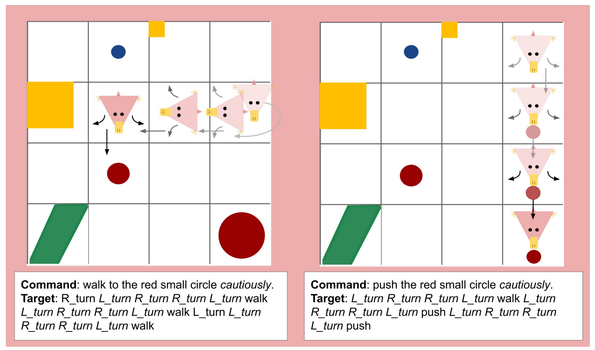
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

If a human knows the meaning of the word *'small'*, they can pick out the *'small wampimuk'* among larger ones, even if they have never encountered wampimuks before. If they know what it means to *'walk cautiously'*, they probably know what it means to *'cycle cautiously'* through a busy intersection. This is because the meaning of words like *'small'* and *'cautiously'* compose systematically. This endows humans with **data efficiency** and **out-of-distribution generalization.** Previous benchmarks that test for this skill (i.e., **systematic generalization**) only do so in a limited or ungrounded context.

The Benchmark

gSCAN tests for **linguistic, rule-based** generalization in **eight** different challenges. Grounding lets us go beyond previous work and allows us to explicitly disentangle failures of systematicity due to perception, sentence understanding, and word grounding. gSCAN poses a multi-modal sequence-to-sequence supervised learning task.



Command: walk to the red small circle *cautiously*.
Target: R_turn *L_turn R_turn R_turn L_turn* walk *L_turn R_turn R_turn L_turn* walk L_turn *L_turn R_turn R_turn L_turn* walk

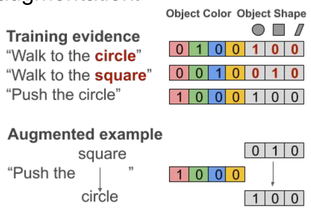
Command: push the red small circle *cautiously*.
Target: *L_turn R_turn R_turn L_turn* walk *L_turn R_turn R_turn L_turn* push *L_turn R_turn R_turn L_turn* push

In the two examples above the same determiner phrase *'the red small circle'* has different reference objects and demands different action sequences. Can agents learn that which object is correctly picked by the 'small'-modifier is dependent on the world state? Can agents learn to *'push cautiously'* when they know how to *'walk cautiously'*?

The Models

Baseline. Multi-modal NN. Bi-LSTM encoder for synthetic language input command, CNN for symbolic world state encoding, LSTM decoder with double attention over the input command and world state.

GECA. The same multi-modal model but extended with SOTA compositional data augmentation.



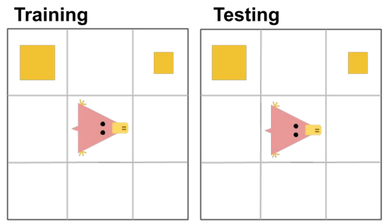
GECA identifies evidence for interchangeable situations in the input command concatenated with the target object vector. In the image on the LHS you can see how it identifies that **circle** can be exchanged by **square**.

Compositional Generalization

7 Challenges from 1 Training Set

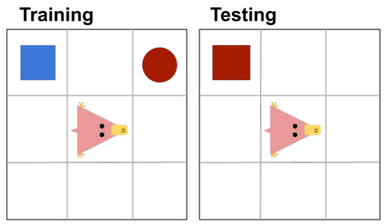
Split A: Sanity check random test set i.i.d. as training set. Can the agent do the task without systematic differences?

Split B. During training the agent is familiarized with the target object yellow square, but it is only ever referred to without the color. Can it learn to generalize to the yellow square being referred to with its color?



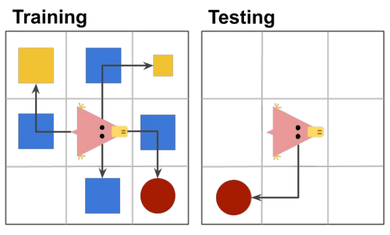
*"Walk to the **big square**."*
*"Walk to the **small square**."*
*"Push the **small square**."*
*"Walk to the **big yellow square**."*
*"Pull the **small yellow square**."*

Split C. During training the agent is familiarized with objects like 'blue square' and 'red circle', but never with 'red squares'. Can it at test time compositionally combine known meaning?

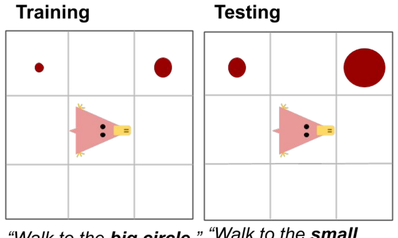


*"Walk to the **blue square**."*
*"Walk to the **red circle**."*
*"Push the **red circle**."*
*"Walk to the **red square**."*
*"Push the **red square**."*
*"Pull the **small red square**."*

Split D. During training the agent learns to walk in all directions except the south-west. Can it generalize to walking to the south-west by combining known commands?



Split E. During training the small circle of size 2 is only ever referred to as the 'big circle', or the 'circle'. Can the agent generalize to that same circle being called 'the small circle'?

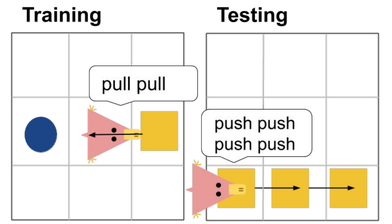


*"Walk to the **big circle**."*
*"Walk to the **red big circle**."*
*"Walk to the **small circle**."*
*"Walk to the **red small circle**."*

Length Generalization

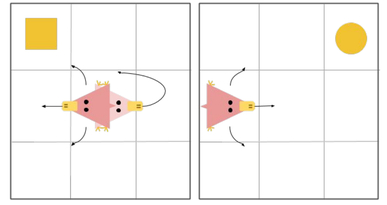
Split I. During training the agent only needs to generate commands with less than 16 actions sequences, at test time it needs to generalize to familiar command and world state combinations that require longer action sequences (i.e., require walking a larger distance over the grid.)

Split F. During training the agent learns how to pull the square of size 3 (which requires 2 pull actions per grid cell). At test time the agent needs to zero-shot generalize to pushing the square of size 3 (also requiring 2 push actions per grid cell).



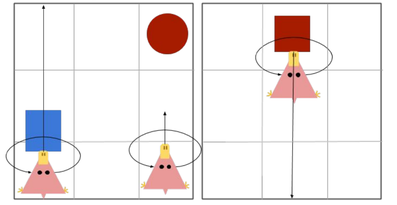
"Pull the square." *"Push the square."*

Split G. At training time the agent only sees ***k*** examples of how to do something 'cautiously', at test time it needs to generalize this adverb to novel world states.



"Walk to the square cautiously."
"Walk to the circle cautiously."

Split H. During training the agent learns how to push while spinning and how to walk while spinning in many situations. At test time the agent needs to generalize to pull while spinning.



"Push the blue square while spinning."
"Walk to the red circle while spinning."
"Pull the red square while spinning."

Conclusion

The baseline fails on all challenges except split A and F, GECA fails on all but split A, C, and F. These failures show advances are needed in neural architectures for compositional learning. We hope gSCAN facilitates progress in compositional learning.

Models & experiments code
https://github.com/LauraRuis/multimodal_seq2seq_gSCAN

Benchmark generation code
https://github.com/LauraRuis/groundedSCAN

Now with **RL mode**! Check the GitHub