Jump to better conclusions: SCAN both left and right

Joost Bastings[1] Marco Baroni[2] Jason Weston[2]
Kyunghyun Cho[2,3,4] Douwe Kiela[2]

[1]ILLC, University of Amsterdam [2]Facebook AI Research
[3]New York University [4]CIFAR Global Scholar

Objective

Testing the **strong generalization** capability of recurrent **sequence-to-sequence** models

Strong Generalization

- If you knew how to **dax**, could you **dax twice**?
- Humans can infer the meaning of a **phrase** from its **constituent parts** (compositionality)
- There is a **long debate** [1] about whether neural networks can do the same

What is SCAN?

- SCAN [2] was designed to **test** strong generalization in neural nets
- SCAN maps **commands** to **actions**:

jump
JUMP

turn around left
LTURN LTURN LTURN LTURN

jump thrice and turn left twice
JUMP JUMP JUMP LTURN LTURN

jump opposite left after walk twice
WALK WALK LTURN LTURN JUMP

Phrase-Structure Grammar

- **Generates** the data set (commands):

$C \rightarrow S$ and S | S after S | S
$S \rightarrow V$ twice | V thrice | V
$V \rightarrow D_{[1]}$ opposite $D_{[2]}$ | $D_{[1]}$ around $D_{[2]}$ | D | U
$D \rightarrow U$ left | U right | turn left | turn right
$U \rightarrow$ walk | look | run | jump

- **20910** unambiguous commands

Interpretation Functions

- **Map commands** to **actions**, e.g.:

$$[\![\text{jump}]\!] = \text{JUMP}$$
$$[\![u \text{ around left }]\!] = \text{LTURN } [\![u]\!] \text{ LTURN } [\![u]\!] \text{ LTURN } [\![u]\!] \text{ LTURN } [\![u]\!]$$
$$[\![x_1 \text{ after } x_2]\!] = [\![x_2]\!] \; [\![x_1]\!]$$

Summary

- SCAN [2] was designed to test **strong generalization** in neural networks
- Good SCAN accuracy does **not** always mean strong generalization
- We propose **NACS** to remedy this
- NACS introduces **non-determinism** and **target-side dependencies** and is a more realistic benchmark

The SCAN Tasks

1. map a **random subset** of the data (‘**simple**’)
2. map **longer** action sequences than those seen during training (‘**length**’)
3. map commands that **compose a word in novel ways** that was only seen in isolation during training (‘**primitive**’)

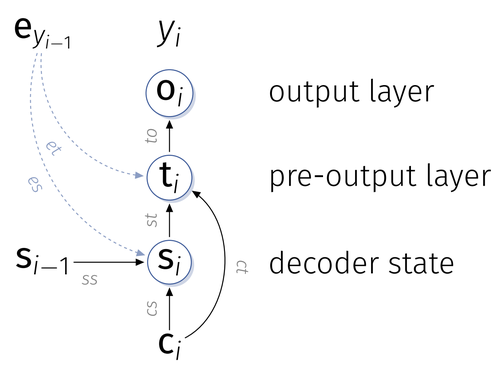
The Problem with SCAN

- **Limited** target-side dependencies
- **Deterministic** mapping
- This allows SCAN to be solved with **simple strategies** (e.g. advancing a pointer)

The Solution: NACS

- NACS maps **actions** to **commands** instead
- **Strong** target-side dependencies
- **Non-deterministic** mapping

Decoder with Attention

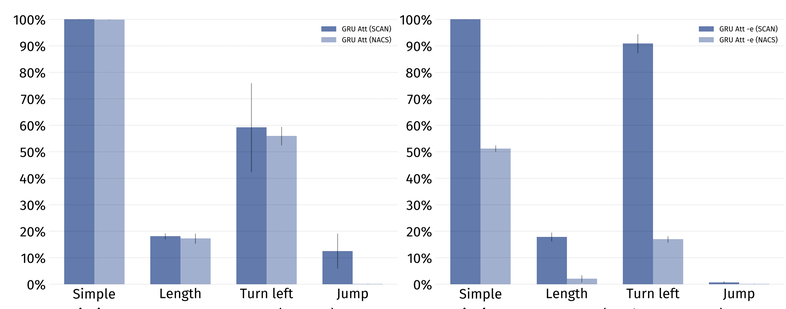


- We remove the es and et connections in an ablation experiment

Experiments

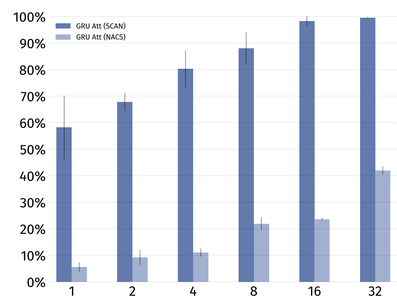
- Zero-shot generalization on all tasks
- `github.com/facebookresearch/NACS`

Results

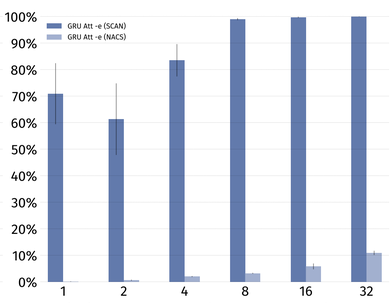


(a) SCAN vs NACS (GRU)

(b) Ablation (w/o es, et)



(c) Jump + extra commands



(d) Ablation (w/o es, et)

Discussion

- SCAN does **not** necessarily indicate generalization performance: it is possible to score 100%, while failing on NACS
- NACS also requires systematicity, but is **harder** because of its target-side dependencies

Conclusion

- Benchmarks for strong generalization need to take **non-determinism** and **target-side dependencies** into account
- NACS performance is more in line with **realistic** scenarios e.g. Machine Translation
- We confirm the **mixed picture** on strong generalization of seq-to-seq models

References

[1] Jerry A Fodor and Zenon W Pylyshyn. “Connectionism and cognitive architecture: A critical analysis”. In: *Cognition* 28.1-2 (1988).

[2] Brenden M. Lake and Marco Baroni. “Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks”. In: *ICML* (2018).

Work done while first author interned at Facebook AI Research, NY.