Secure Systems Group, Aalto University

Tommi Gröndahl, Yujia Guo

Can Transformers learn symbolic rules?

Symbolic rules

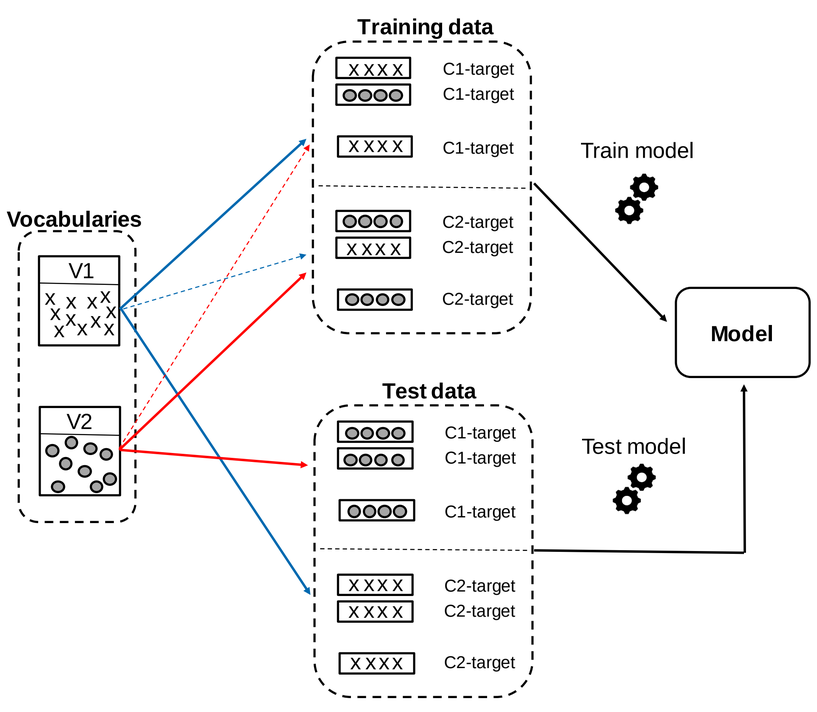
- Basis of classical computation (Turing machines, von Neumann architectures)
- Read/write operations on symbols in memory enacted by processor
- Allow systematic generalization
- Presence in DNNs unclear

Experiments

- Fine-tuning T5 (attention-based Transformer)
- Generalize rules across vocabularies V1/V2
- Zero-shot: train on V1, test on V2
- Vocabulary flip: train on V1 for class 1, train on V2 for class 2, test for converse
- MIX: ratio of training data in converse

Task		Example	
		Input	Target
seq2seq	copy/reverse	copy ab	ab
		reverse ab	ba
classification	copy/reverse detection	ab <\s> ab	copy
		ab <\s> ba	reverse
	palindrome detection	abba	1
		abaa	0
	repetition detection	aba	1
		abc	0

Table 1. Tasks studied in the experiments. |V1|=|V2|=10, |train|=8000, |eval|=2000, |test|=10000



Task		Zero-shot		Vocabulary flip	
		Eval	Test	Eval	Test
copy/reverse	copy	1.00	1.00	1.00	0.75
	reverse	1.00	0.97	1.00	0.72
copy/reverse detection	copy	1.00	1.00	1.00	0.00
	reverse	1.00	0.88	1.00	0.00
palindrome detection	palindrome	1.00	1.00	1.00	0.07
	non-palindrome	0.98	0.90	1.00	0.00
repetition detection	repetition	1.00	0.96	1.00	0.08
	no repetition	1.00	1.00	1.00	0.10

Table 2. T5 zero-shot and vocabulary flip performance (MIX=0).

Task	Class	MIX		
		0	0.01	0.1
copy/reverse	copy	0.75	1.00	1.00
	reverse	0.72	0.99	0.99
copy/reverse detection	copy	0.00	1.00	1.00
	reverse	0.00	0.94	1.00
palindrome detection	palindrome	0.07	0.09	0.99
	non-palindrome	0.00	0.02	0.91
repetition detection	repetition	0.08	0.27	0.30
	no repetition	0.10	0.12	0.10

Table 3. Impact of MIX on vocabulary flip performance (test set)

Results

- Strong zero-shot performance throughout
- Vocabulary flip markedly better on seq2seq
- Task difficulty based on MIX: copy/reverse > copy/reverse detection > palindrome > repetition

Discussion

- Zero-shot likely based on embedding similarities
- Seq2seq can use input itself as "external memory" via attention; classification cannot
- Repetition detection most challenging because:
  - does not allow simple heuristics
  - requires existential quantification

Conclusions

- Results explainable without model-internal symbolic rules
- Attention allows model to function as "processor" to input/output as external "memory"

Aalto University