Revisit Systematic Generalization via Meaningful Learning

Ning Shi, Boxin Wang, Wei Wang, Xiangyu Liu, Zhouhan Lin
ning.shi@ualberta.ca, boxinw2@illinois.edu, {luyang.ww,eason.lxy}@alibaba-inc.com, lin.zhouhan@gmail.com

Introduction

Humans can systematically generalize to novel compositions of existing concepts. Recent studies argue that neural networks appear inherently ineffective in such cognitive capacity, leading to a *pessimistic* view and a lack of attention to *optimistic* results.

In contrast, the successful one-shot generalization in the turn-left experiment on the Simplified CommAI Navigation (SCAN) task reveals the potential of seq2seq recurrent networks in controlled environments (Lake and Baroni, 2018).

Question by Lake and Baroni (2018) on page 8:

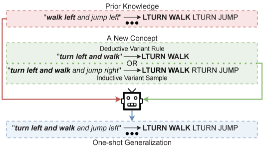
"What are, precisely, the generalization mechanisms that subtend the networks' success in these experiments?"

Meaningful Learning

In educational psychology, *meaningful learning* refers to learning new concepts by relating them to old ones (Ausubel, 1963).

On the contrary, *rote learning* stands for learning new concepts without the consideration of relationships.

One-shot Example



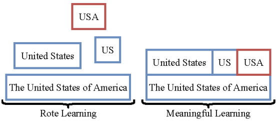
An example of the one-shot compositional generalization from the old concept **"walk left"** to the new one **"turn left and walk"** in SCAN.

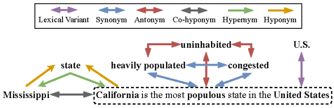
From - "**walk left** and jump left"
To - "**turn left and walk** and jump left"

Old concept - "walk left"
New concept - "turn left and walk"

Connect both by pointing to **LTURN WALK**:

- Inductive Learning
- Deductuive Learning





Inductive Learning

Inductive learning is a *bottom-up* approach from the more specific to the more general. In grammar teaching, inductive learning is a rule-discovery approach starting with the presentation of specific examples from which a general rule can be inferred.

Deductive Learning

Deductive Learning, the opposite of inductive learning, is a *top-down* approach from the more general to the more specific. As a rule-driven approach, teaching in a deductive manner often begins with presenting a general rule followed by specific examples in practice where the rule is applied

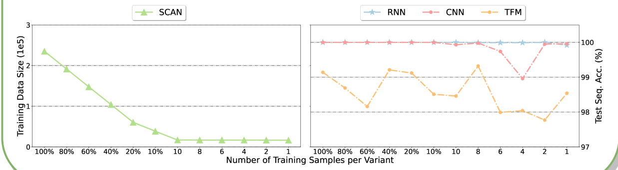
Data	Primitive	Variant	#Variants	Prompt
SCAN	*jump*	*jump_0*	10	*[concept] twice*
GEO	*new york city*	*houston city*	39	*how many people in [concept]*
	mississippi rivier	*red rivier*	9	*how long is [concept]*
	dc	*kansas*	49	*where is [concept]*
	dover	*salem*	8	*what states capital is [concept]*
ADV	*a history of american film*	*advanced ai techniques*	5/424	*who teaches [concept] ?*
	aaron magid	*cargo*	5/492	*does [concept] give upper-level courses ?*
	aaptis	*survmeth*	5/1720	*name core courses for [concept] .*
	100	*171*	5/1895	*can undergrads take [concept] ?*

Data	Primitive	Semantic Links	Variant	Concept Rule: Primitive Rule	Concept Rule: Variant Rule
SCAN	*jump*	Lexical Variant	*jump_0*	*jump* → JUMP	*jump_0* → JUMP
	look		*look_0*	*look* → LOOK	*look_0* → LOOK
	run		*run_0*	*run* → RUN	*run_0* → RUN
	walk		*walk_0*	*walk* → WALK	*walk_0* → WALK
GEO	*new york city*	Co-hyponym	*houston city*	*new york city* → CITY_NAME	*houston city* → CITY_NAME
	mississippi rivier		*red rivier*	*mississippi rivier* → RIVER_NAME	*red rivier* → RIVER_NAME
	dc		*kansas*	*dc* → STATE_NAME	*kansas* → STATE_NAME
	dover		*salem*	*dover* → CAPITAL_NAME	*salem* → CAPITAL_NAME
ADV	*a history of american film*	Co-hyponym	*advanced ai techniques*	*a history of american film* → TOPIC	*advanced ai techniques* → TOPIC
	aaron magid		*cargo*	*aaron magid* → INSTRUCTOR	*cargo* → INSTRUCTOR
	aaptis		*survmeth*	*aaptis* → DEPARTMENT	*survmeth* → DEPARTMENT
	100		*171*	*100* → NUMBER	*171* → NUMBER

Systematic Generalization

Setup - we treat concepts in the initial data set as primitives and generate variant samples and rules accordingly. Next, we mix them up and construct a seq2seq task after a random split. We repeatedly train and evaluate models but slowly decrease the number of times they see each variant until one-shot learning.

Results - we observe there is *hardly* a performance drop for three representative model structures. This evidences that, with *semantic linking*, even canonical neural networks can generalize systematically to new concepts and compositions.



Semantic Linking Injection

We increase the difficulty of compositional generalization by excluding from the training set the primitive samples for inductive learning (left) and primitive rules for deductive learning (right).

Data	Model	Token Acc.% Standard	Token Acc.% Difficult	Token Acc.% Challenging	Seq. Acc.% Standard	Seq. Acc.% Difficult	Seq. Acc.% Challenging
SCAN	RNN	99.99 ± 0.03	99.89 ± 0.19	99.96 ± 0.02	99.95 ± 0.08	99.85 ± 0.08	99.80 ± 0.31
	CNN	99.96 ± 0.08	99.76 ± 0.54	98.89 ± 2.44	99.85 ± 0.34	99.52 ± 1.07	97.57 ± 5.24
	TFM	98.91 ± 0.78	98.90 ± 1.10	98.76 ± 0.85	97.35 ± 1.62	96.86 ± 2.64	96.38 ± 2.81
GEO	RNN	75.71 ± 8.42	75.69 ± 6.12	73.46 ± 3.05	44.95 ± 14.69	43.27 ± 13.47	36.77 ± 5.60
	CNN	87.99 ± 2.67	79.51 ± 6.03	77.40 ± 2.48	69.46 ± 5.78	51.20 ± 8.64	48.58 ± 3.40
	TFM	75.37 ± 7.84	75.11 ± 4.88	68.41 ± 4.76	45.93 ± 12.42	44.59 ± 9.76	36.93 ± 7.47
ADV	RNN	58.61 ± 6.18	59.74 ± 5.67	58.11 ± 5.82	36.18 ± 5.75	35.69 ± 6.05	35.45 ± 6.69
	CNN	57.83 ± 7.55	54.05 ± 5.74	53.66 ± 2.57	45.08 ± 9.32	42.14 ± 6.90	41.37 ± 4.04
	TFM	53.43 ± 2.80	51.51 ± 4.50	49.17 ± 2.58	42.59 ± 3.65	41.28 ± 4.35	38.88 ± 2.68

Data	Model	Token Acc.% Standard	Token Acc.% Difficult	Seq. Acc.% Standard	Seq. Acc.% Difficult
SCAN	RNN	99.48 ± 0.71	98.70 ± 0.92	98.27 ± 2.38	95.39 ± 2.72
	CNN	99.99 ± 0.01	98.59 ± 3.10	99.96 ± 0.03	96.66 ± 7.27
	TFM	96.90 ± 1.78	96.68 ± 2.21	91.94 ± 4.04	91.26 ± 5.80
GEO	RNN	54.44 ± 7.15	39.71 ± 18.38	13.61 ± 7.08	7.76 ± 5.34
	CNN	41.86 ± 3.38	41.07 ± 7.48	4.85 ± 4.66	4.04 ± 2.18
	TFM	67.02 ± 6.91	65.97 ± 5.17	36.38 ± 10.08	31.57 ± 7.42
ADV	RNN	36.50 ± 7.66	36.42 ± 7.39	12.84 ± 4.31	12.66 ± 5.19
	CNN	43.51 ± 11.31	35.34 ± 14.68	32.33 ± 12.93	23.58 ± 16.04
	TFM	56.82 ± 3.79	53.33 ± 3.85	47.43 ± 3.71	43.24 ± 5.14

Proof of Concept

Model	IWSLT'14 En-De BLEU	IWSLT'14 En-De SacreBLEU	IWSLT'14 De-En BLEU	IWSLT'14 De-En SacreBLEU	IWSLT'15 En-Fr BLEU	IWSLT'15 En-Fr SacreBLEU	IWSLT'15 Fr-En BLEU	IWSLT'15 Fr-En SacreBLEU
Baselines								
LSTM (Luong et al., 2015)	24.98	24.88	30.18	32.62	38.06	42.93	37.34	39.36
Transformer (Vaswani et al., 2017)	28.95	28.85	35.24	37.60	41.82	46.41	40.45	42.61
Dynamic Conv. (Wu et al., 2019)	27.39	27.28	33.33	35.54	40.41	45.32	39.61	41.42
+Vocabulary Augmentation								
LSTM (Luong et al., 2015)	$25.35\uparrow_{0.37}$	$25.38\uparrow_{0.50}$	$30.99\uparrow_{0.81}$	$33.63\uparrow_{1.01}$	$38.32\uparrow_{0.26}$	$43.30\uparrow_{0.37}$	$37.77\uparrow_{0.43}$	$39.83\uparrow_{0.47}$
Transformer (Vaswani et al., 2017)	$29.40\uparrow_{0.45}$	$29.29\uparrow_{0.44}$	$35.72\uparrow_{0.48}$	$38.07\uparrow_{0.47}$	$42.19\uparrow_{0.37}$	$46.68\uparrow_{0.27}$	$41.04\uparrow_{0.59}$	$43.15\uparrow_{0.54}$
Dynamic Conv. (Wu et al., 2019)	$27.60\uparrow_{0.21}$	$27.50\uparrow_{0.22}$	$33.62\uparrow_{0.29}$	$36.00\uparrow_{0.46}$	$40.87\uparrow_{0.46}$	$45.95\uparrow_{0.63}$	$39.95\uparrow_{0.34}$	$41.86\uparrow_{0.44}$

Model	Geography Train Token Acc.%	Geography Train Seq. Acc.%	Geography Test Token Acc.%	Geography Test Seq. Acc.%	Advising Train Token Acc.%	Advising Train Seq. Acc.%	Advising Test Token Acc.%	Advising Test Seq. Acc.%
Baselines								
RNN	89.05	17.39	69.81	9.68	92.22	3.64	60.41	6.11
CNN	98.45	70.74	78.44	55.91	99.74	81.62	81.74	51.13
TFM	99.45	84.95	80.24	49.82	99.68	76.90	78.51	29.67
+Entity Augmentation								
RNN	87.47	29.96	$72.39\uparrow_{2.58}$	$15.05\uparrow_{5.37}$	88.82	30.97	$71.17\uparrow_{10.76}$	$16.06\uparrow_{9.95}$
CNN	97.54	76.03	$80.32\uparrow_{1.88}$	$60.93\uparrow_{5.02}$	99.65	87.01	$84.50\uparrow_{2.76}$	$56.02\uparrow_{4.89}$
TFM	99.30	85.73	$81.09\uparrow_{0.85}$	$54.84\uparrow_{5.02}$	99.57	86.94	$84.26\uparrow_{5.75}$	$35.08\uparrow_{5.41}$

arXiv

@BlackboxNLP

This work was supported by Shining Lab, Learnable, Inc., and Alibaba Group.

GitHub