Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation

Strong AI Lab & LIU AI Lab, School of Computer Science, The University of Auckland

Authored by: Qiming Bao, Alex Yuxuan Peng, Tim Hartill, Neşet Özkan Tan, Zhenyun Deng, Michael Witbrock, Jiamou Liu

The 2nd International Joint Conference on Learning & Reasoning and 16th International Workshop on Neural-Symbolic Learning and Reasoning (IJCLR-NeSy 22)

11th October 2022

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Strong AI Lab & LIU AI Lab

- Strong AI Lab is led by Professor Michael Witbrock, at the intersection of machine learning, reasoning, and natural language understanding, with an additional focus on achieving the best social and civilisational impacts of increasingly powerful AI.
- LIU AI Lab is led by Dr. Jiamou Liu. We are an AI research group at the University of Auckland. We are engaged in artificial intelligence research and development from both the industrial and the academic sides. Our research interests cover a wide range of topics across the modern AI world, including deep learning, reinforcement learning, multi-agent systems, natural language processing, and complex network analysis.

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

https://www.ai.ac.nz/sail/
https://www.liuailab.org/

Symbolic Logic Programs

- ***Symbolic logic*** expresses logical statements and expressions in symbols and variables instead of natural language.
- An example of logic programs expressed in Prolog [1]

$$p(X) : -q(X).$$
$$q(a).$$

$p(X)$, where variables are notated in capital letters.

$q(a)$, where constants are in lower case.

[1] Programming in Prolog: Using the ISO standard, Clocksin, 2012

Symbolic Logic Programs

1: Facts

$e(l).$

$?e(l).\ 1$

$?i(d).\ 0$

2: Unification

$o(V, V).$

$?o(d, d).\ 1$

$?o(b, d).\ 0$

3: 1 Step

$p(X) : -q(X).$

$q(a).$

$?p(a).1$

$?p(b).0$

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Cingillioglu, N. et al., 2018. DeepLogic: Towards End-to-End Differentiable Logical Reasoning, AAAI-MAKE19.

Natural Language Reasoning

- In natural language reasoning, logical statements are expressed in natural language instead of symbols.
    - **The semantics of logic**, such as propositional logic and first-order logic.
    - **Diversity and flexibility of natural language**, such as polysemy, a paraphrase of sentences.
    - Reasoning obtain unknown information based on existing information.

      Deductive reasoning: Given premise and rules to derive the conclusion.

      Inductive reasoning: Given premise and conclusion to derive rules.

      Abductive reasoning: Given rules and conclusion to derive premise.

      More examples can be found in [1] and [2].

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.
[2] Young, N. et al. 2022. AbductionRules: Training Transformers to Explain Unexpected Inputs. The finding of ACL.

Example for Natural Language Reasoning

(Input Facts:) Alan is blue. Alan is rough. Alan is young.
Bob is big. Bob is round.
Charlie is big. Charlie is blue. Charlie is green.
Dave is green. Dave is rough.

(Input Rules:) Big people are rough.
If someone is young and round then they are kind.
If someone is round and big then they are blue.
All rough people are green.

Q1: Bob is green. True/false? **[Answer: T]**
Q2: Bob is kind. True/false? **[F]**
Q3: Dave is blue. True/false? **[F]**

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Clark, P., et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.

Example for Natural Language Reasoning

(Input Facts:) Alan is blue. Alan is rough. Alan is young.
Bob is big. Bob is round.
Charlie is big. Charlie is blue. Charlie is green.
Dave is green. Dave is rough.

(Input Rules:) Big people are rough.
If someone is young and round then they are kind.
If someone is round and big then they are blue.
All rough people are green.

Q1: Bob is green. True/false? **[Answer: T]**
Q2: Bob is kind. True/false? **[F]**
Q3: Dave is blue. True/false? **[F]**

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Clark, P., et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.

Example for Natural Language Reasoning

(Input Facts:) Alan is blue. Alan is rough. Alan is young.
Bob is big. Bob is round.
Charlie is big. Charlie is blue. Charlie is green.
Dave is green. Dave is rough.

(Input Rules:) Big people are rough.
If someone is young and round then they are kind.
If someone is round and big then they are blue.
All rough people are green.

Q1: Bob is green. True/false? **[Answer: T]**
Q2: Bob is kind. True/false? **[F]**
Q3: Dave is blue. True/false? **[F]**

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

[1] Clark, P., et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.

Example for Natural Language Reasoning

(Input Facts:) Alan is blue. Alan is rough. Alan is young.
Bob is big. Bob is round.
Charlie is big. Charlie is blue. Charlie is green.
Dave is green. Dave is rough.

(Input Rules:) Big people are rough.
If someone is young and round then they are kind.
If someone is round and big then they are blue.
All rough people are green.

Q1: Bob is green. True/false? **[Answer: T]**
Q2: Bob is kind. True/false? **[F]**
Q3: Dave is blue. True/false? **[F]**

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

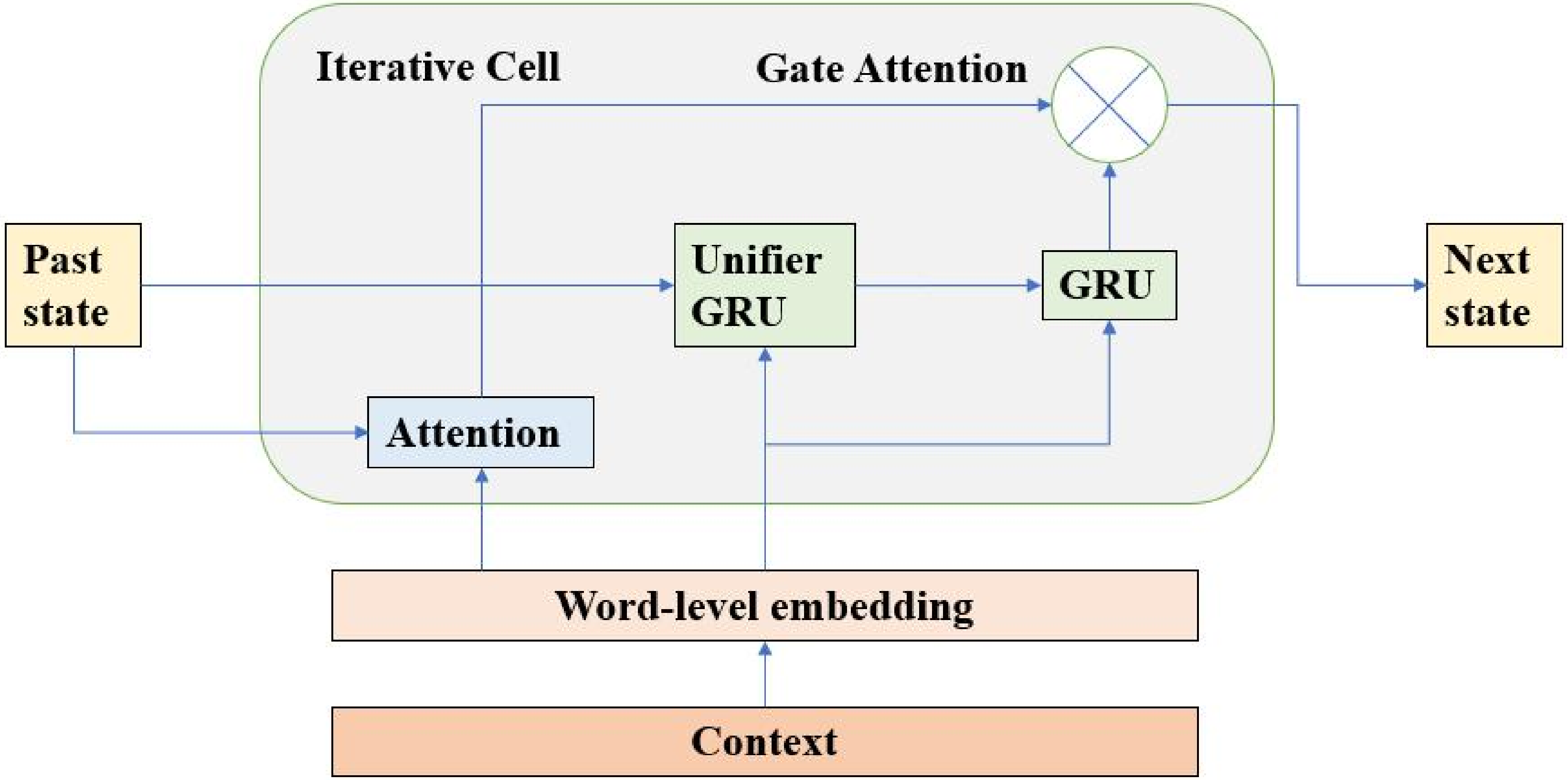
[1] Clark, P., et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.

Research Gap

- Existing models, including DeepLogic and other RNN-based baseline models, have room for improvement in their reasoning abilities over natural language.
- We found existing models are not good at out-of-distribution (OOD) generalisation, in three scenarios:
  - When the model is trained on data with shallow reasoning depths and tested on data with deeper reasoning depths.
  - When the model is trained on synthetically generated data and tested on data paraphrased by human.
  - When the model is trained on unshuffled data and tested on shuffled data.
- Existing multi-step deductive reasoning datasets like PARARULES and CONCEPTRULE V1 and V2 have unbalanced distributions over the reasoning depths. Only a small portion of the datasets require deep reasoning (2<=Depth<=5).

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Model Overview



[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Word-level Embedding

- The input to the network consists of a context and a statement.
- The input sequence is represented using GloVe [1] word embeddings.
- The concatenated representations of context and statement will be fed into the gated recurrent unit (GRU).

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Pennington, et al., 2014. Glove: Global vectors for word representation, EMNLP.

Iteration

- The iteration process is from the DeepLogic [1]. The iteration step consists of attending to the rules, computing a new state using each rule and the old state.
- To apply a rule, we use another recurrent neural network called the **inner GRU unifier** that processes every literal of a given rule. The inner GRU unifier needs to learn unification between **variables** and **constants** as well as how each rule interacts with the current state.

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Cingillioglu, N. et al., 2018. DeepLogic: Towards End-to-End Differentiable Logical Reasoning, AAAI-MAKE19.

Gate Attention

- Dynamic Memory Network+ [1] achieved 100% test accuracy by using gate attention on bAbI deductive reasoning task (Task-15), which gave us the idea of integrating Gate Attention into DeepLogic. GRU can use gate attention to update the internal state.

THE UNIVERSITY OF AUCKLAND Te Whare Wananga o Tamaki Makaurau NEW ZEALAND

[1] Xiong, C., et al., 2016. Dynamic Memory Networks for Visual and Textual Question Answering, ICML.

Established Baselines - RNNs & PLM

- We have three baseline models that we borrowed from the bAbI task leaderboard. We also set DeepLogic as one of the baseline methods, and then we have a Transformer-based model RoBERTa-Large as a baseline model. We use glove.6B.zip [4] as the word vector representation for the RNN-based models.
  - Long short-term memory (LSTM, 1997) [1] (The baseline method on bAbI dataset),
  - Dynamic Memory Network (DMN, 2016) [2] (One of the first paper use Attention in the memory network),
  - Memory Attention Control networks (MAC, 2018) [3] (A classical method from memory network).

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Hochreiter, et al. 1997. Long short-term memory,
[2] Kumar, et al. 2016. Ask me anything: Dynamic memory networks for natural language processing, ICML
[3] Hudson, et al. 2018. Compositional attention networks for machine reasoning, ICLR.
[4] Pennington, et al. 2014. Glove: Global vectors for word representation, EMNLP.
[5] Liu, Y. et al., 2019. Roberta: A robustly optimized bert pretraining approach. arxiv.

CONCEPTRULE vs CONCEPTRULE V2

(*Input Context:*) Book is not located in bed.
Bed is located in loft.
Loft is located in city.
City is located in fast-food restaurant.
Question 1: Book is located in loft. True/False? [**Answer: T**]
Question 2: Bed is located in city. True/False? [**Answer: T**]
Question 3: Book is located in bed. True/False? [**Answer: F**]

(*Input Context:*) Book is not located in bed.
Bed is located in loft.
Loft is located in city.
City is located in fast-food restaurant.
Question 1: Book is not located in bed. True/false? [**Answer: T**] [**Depth: 0**]
Question 2: Book is not located in loft. True/false? [**Answer: T**] [**Depth: 1**]
Question 3: Book is not located in city. True/false? [**Answer: T**] [**Depth: 2**]

Hartill, T., CONCEPTRULE, https://drive.google.com/file/d/1lxoAvtcvqVCYiO8e3tENnrTQ1NNVtpjs/view
Hartill, T., CONCEPTRULE V2, https://drive.google.com/file/d/1IOCbW8bfZxj1RIzKDxn8xKg99XyYNj7z/view

Dataset Description

Table 2

Information about the datasets used in this paper. PARARULES has less number of examples that require deep reasoning steps. CONCEPTRULES V2 does not consider reasoning depths greater than 3. The train, dev and test set are already splitted by the author of each dataset.

Dataset		Depth=0	Depth=1	Depth=2	Depth=3	Depth=4	Depth=5
PARARULES	Train	290435	157440	75131	48010	9443	7325
	Dev	41559	22276	10833	6959	1334	1038
	Test	83119	45067	21496	13741	2691	2086
PARARULE-Plus	Train	-	-	89952	90016	90010	90022
	Dev	-	-	16204	16154	16150	16150
	Test	-	-	2708	2694	2704	2692
CONCEPTRULES V2 (full)	Train	2074360	1310622	873748	436874	-	-
	Dev	115148	72810	48540	24270	-	-
	Test	115468	72810	48540	24270	-	-
CONCEPTRULES V2 (simplified)	Train	131646	74136	49424	24712	-	-
	Dev	7166	4116	2744	1372	-	-
	Test	7362	4116	2744	1372	-	-

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Dataset Description

Table 3
The entity types and relation types for CONCEPTRULES V1 (simplified/full), CONCEPTRULES V2 (simplified/full), PARARULES, and our PARARULE-Plus.

Dataset	#Entity	#Relation	Shuffled Rules	Depth Tag	Derivable	NAF
CONCEPTRULES V1 (simplified)	385	7	No	No	Yes	Yes
CONCEPTRULES V1 (full)	4048	24	Yes	No	Yes	No
CONCEPTRULES V2 (simplified)	385	7	No	Yes	Yes	Yes
CONCEPTRULES V2 (full)	4048	24	Yes	Yes	Yes	Yes
PARARULES	19	4	No	Yes	Yes	Yes
PARARULE-Plus	71	8	No	Yes	Yes	Yes

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

A Sample for Negation as Failure (NAF)

(*Input Facts:*) The bear visits the lion.
The tiger likes the cat.
The cat does not like the bear.
The lion likes the tiger.
(*Input Rules:*) If someone sees the lion then the lion is kind.
If the tiger visits the lion and someone does not see the tiger then the tiger visits the bear.
If someone likes the bear and they like the tiger then the bear visits the tiger.
If someone is not round then they like the cat.
If someone visits the lion then they are blue.
If someone visits the bear and they do not see the lion then they visit the tiger.
If someone is cold and they do not visit the lion then the lion visits the tiger.
If someone visits the tiger and they are green then the tiger likes the cat.
Question 1: The bear likes the cat. True/false? [**Answer: T**]
Question 2: The bear is round. True/false? [**F**]
Question 3: The bear is not round. True/false? [**T**]

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

[1] Peter Clark, et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.

Experiment Result

Table 4
We use GloVe [16] as the word vector representation. We use PARARULES with all depths as the training set for all models and then test them on examples with different reasoning depths (D). Comparison among our IMA-GloVe-GA, IMA-GloVe, MAC-GloVe, DMN-GloVe, IMASM-GloVe, LSTM-GloVe, and RoBERTa-Large on PARARULES test sets with different reasoning depths.

Train ↓; Test →	D=1	D=2	D=3	D≤3	D≤3+NatLang	D≤5	D≤5+NatLang
IMA-GloVe	0.861	0.853	0.830	0.842	0.810	0.792	0.705
MAC-GloVe	0.792	0.776	0.750	0.763	0.737	0.701	0.652
DMN-GloVe	0.846	0.843	0.817	0.827	0.789	0.779	0.666
IMASM-GloVe	0.864	0.855	0.824	0.838	0.801	0.782	0.608
LSTM-GloVe	0.500	0.500	0.500	0.499	0.499	0.500	0.500
IMA-GloVe-GA	**0.950**	**0.943**	**0.919**	**0.927**	**0.883**	**0.879**	**0.741**
RoBERTa-Large	**0.986**	**0.985**	**0.977**	**0.979**	**0.972**	**0.967**	**0.949**

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Experiment Result

Table 5

IMA-GloVe, IMA-GloVe-GA, and RoBERTa-Large trained on CONCEPTRULES V1 (simplified / full) and tested on different test sets. Rules in CONCEPTRULES V1 Simplified are not shuffled, while CONCEPTRULES V1 full contains randomly shuffled rules. CONCEPTRULES V1 full has larger number of relations and entities than CONCEPTRULES V1 simplified.

Model	Train set	Test accuracy (Simplified Test set)	Test accuracy (Full Test set)
IMA-GloVe	Simplified	0.994	0.729
	Full	0.844	**0.997**
IMA-GloVe-GA	Simplified	**0.998**	**0.747**
	Full	0.851	**0.999**
RoBERTa-Large	Simplified	**0.997**	0.503
	Full	**0.927**	0.995

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Experiment Result

Table 6

IMA-GloVe, IMA-GloVe-GA, and RoBERTa-Large trained on CONCEPTRULES V2 (full) and tested on test sets that require different depths of reasoning.

Model	Test set	Mod1 Depth=1	Mod2 Depth=2	Mod3 Depth=3	Mod01 Depth≤1	Mod012 Depth≤2	Mod0123 Depth≤3
IMA-GloVe	Depth=1	**0.999**	**0.998**	**0.990**	**0.997**	**0.998**	**0.997**
	Depth=2	**0.998**	**0.999**	**0.988**	**0.995**	**0.998**	**0.997**
	Depth=3	**0.997**	0.998	0.981	**0.991**	0.996	**0.997**
IMA-GloVe-GA	Depth=1	0.993	0.996	0.987	0.987	0.991	**0.997**
	Depth=2	0.993	**0.999**	0.974	0.986	0.991	0.995
	Depth=3	0.988	**1**	**0.994**	0.989	**0.997**	0.994
RoBERTa-Large	Depth=1	0.998	0.975	0.831	0.995	0.975	0.971
	Depth=2	0.997	0.972	0.885	0.993	0.972	0.965
	Depth=3	0.987	0.951	0.984	0.988	0.951	0.936

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Experiment Result

Table 7

RoBERTa-Large trained on PARARULES with different reasoning depths and tested on test sets that require different depths of reasoning. A bold number indicates the highest accuracy in a test set.

Model	Test set	Mod012 (Depth≤2)	Mod0123 (Depth≤3)	Mod0123Nat (Depth≤3+NatLang)	Mod012345 (Depth≤5)
RoBERTa-Large	Depth=0	**0.971**	0.946	0.968	0.953
	Depth=1	**0.943**	0.907	0.933	0.909
	Depth=2	**0.933**	0.902	0.932	0.902
	Depth=3	0.562	0.902	**0.926**	0.907
	Depth=4	0.481	0.863	**0.904**	0.888
	Depth=5	0.452	0.856	0.916	**0.933**
	NatLang	0.573	0.579	**0.962**	0.594

THE UNIVERSITY OF AUCKLAND
Te Whare Wananga o Tamaki Makaurau
NEW ZEALAND

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

Experiment Result

Table 8

RoBERTa-Large is fine-tuned on examples with different depths from PARARULES and also the entire PARARULE-Plus(PPT), and then is evaluated on test sets that require different depths of reasoning. The yellow background indicates improvement on accuracy after adding our PARARULE-Plus in the training process.

Model	Test set	Mod012 (Depth≤2+PPT)	Mod0123 (Depth≤3+PPT)	Mod0123Nat (Depth≤3+NatLang+PPT)	Mod012345 (Depth≤5+PPT)
RoBERTa-Large	Depth=0	0.946	0.901	0.965	**0.963 (+0.010)**
	Depth=1	0.877	0.847	**0.937 (+0.004)**	0.881
	Depth=2	0.868	0.873	**0.927**	0.839
	Depth=3	0.771 (+0.209)	0.862	**0.904**	0.826
	Depth=4	0.675 (+0.194)	0.852	**0.897**	0.832
	Depth=5	0.661 (+0.209)	0.888 (+0.032)	0.923 (+0.007)	**0.934 (+0.001)**
	NatLang	0.557	0.593 (+0.014)	**0.970 (+0.008)**	0.649 (+0.055)

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

[1] Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.

References

- Bao, Q. et al., 2022. Multi-Step Deductive Reasoning Over Natural Language: An Empirical Study on Out-of-Distribution Generalisation. IJCLR-NeSy.
- Brown, T. et al. 2020. Language Models are Few-Shot Learners. NIPS 2020.
- Chung, J. et al. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, NIPS 2014.
- Cingillioglu, N. et al., 2018. DeepLogic: Towards End-to-End Differentiable Logical Reasoning, AAAI-MAKE19.
- Clark, P., et al. 2020. Transformers as Soft Reasoners over Language, IJCAI 2020.
- Devlin, J. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2018.
- Hartill, T., CONCEPTRULE, https://drive.google.com/file/d/1lxoAvtcvqVCYiO8e3tENnrTQ1NNVtpjs/view
- Hartill, T., CONCEPTRULE V2, https://drive.google.com/file/d/1lOCbW8bfZxj1RIzKDxn8xKg99XyYNj7z/view
- Hochreiter, S. et al. 1997. LONG SHORT-TERM MEMORY. Neural computation, 9(8), 1735-1780.
- Hudson, et al. 2018. Compositional attention networks for machine reasoning, ICLR.

References

- J J Hopfield et al. Neural networks and physical systems with emergent collective computational abilities. PNAS 1982.
- Kumar, et al. 2016. Ask me anything: Dynamic memory networks for natural language processing, ICML
- Liu, Y. et al., 2019. Roberta: A robustly optimized bert pretraining approach. arxiv.
- Mikolov, T. et al. 2013. Efficient estimation of word representations in vector space.
- Young, N. et al. 2022. AbductionRules: Training Transformers to Explain Unexpected Inputs. The finding of ACL.
- Pennington, et al., 2014. Glove: Global vectors for word representation, EMNLP.
- Programming in Prolog: Using the ISO standard, Clocksin, 2012
- Radford, A. 2019. Language Models are Unsupervised Multitask Learners. OpenAI blog.
- Xiong, C., et al., 2016. Dynamic Memory Networks for Visual and Textual Question Answering, ICML.
- Vaswani, A. et al. 2017. Attention is all you need. NIPS.

Thank You!

Welcome to visit our lab!

Strong AI Lab

https://www.ai.ac.nz/sail/

LIU AI Lab

https://www.liuailab.org/

qiming_bao

Qiming Bao

Qiming (Bill) Bao

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Paper

GitHub Repo