Adaptive Pre-training of Language Models for Better Logical Reasoning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Logical reasoning of text is an important ability that requires understanding the logical information present in the text and reasoning through them to infer new conclusions. Prior works on improving the logical reasoning ability of language models require complex processing of training data (e.g., aligning symbolic knowledge to text), yielding task-specific data augmentation solutions that restrict the learning of general logical reasoning skills. In this work, we propose AERIE, an adaptively pre-trained language model that has improved logical reasoning abilities. We select a subset of Wikipedia, based on a set of logical inference keywords, for continued pretraining of a language model. We use two self-supervised loss functions: a modified masked language modeling loss where only specific parts-of-speech words, that would likely require more reasoning than basic language understanding, are masked, and a sentence classification loss that teaches the model to distinguish between entailment and contradiction types of sentences. The proposed training paradigm is both simple and generalizable across tasks. We demonstrate the effectiveness of AERIE by comparing it with prior baselines on two logical reasoning datasets. AERIE performs comparably on ReClor and outperforms baselines on LogiQA.

1 Introduction

Logical reasoning is an important ability of humans that helps us in making rational decisions based on some known information. Recently, logical reasoning of text has seen an increasing focus as it is a fundamental skill required to solve any downstream task that requires machine reading [Yu et al., 2020, Liu et al., 2021]. In these datasets, the model needs to understand a given context, reason about a question, and then select the correct answer from a set of options. With the advent of large pre-trained language models (PLMs) in NLP [Devlin et al., 2019, Radford et al., 2019, Raffel et al., 2020], understanding and improving the logical reasoning abilities of these models has become even more important as these are increasingly being used across a wide variety of real-world tasks.

There have been some recent works on improving the logical reasoning abilities of PLMs [Wang et al., 2022, Ouyang et al., 2022, Jiao et al., 2022]. These works typically generate a dataset containing symbolic structures such as logical graphs from text, logical contrast sets, etc., and then train the LM using custom loss objectives to learn logical reasoning abilities. While the performance improvements achieved by these methods are encouraging, the proposed solutions generally require complex data processing to generate the additional structural information (graphs, contrast data, etc.) required for logical reasoning. Further, the loss functions proposed in these works are very specifically designed in accordance to their respective data augmentation technique, and widely differs from the typical masked language modeling loss used for LM pretraining [Devlin et al., 2019]. These complex

Submitted to the Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

processing steps usually require task-specific design choices, which are not necessarily learning generalizable logical reasoning ability that is reusable across different task formats. Also, it is unclear if these specific inductive biases are indeed essential for improving the logical reasoning abilities in language models, or a simpler approach is sufficient.

Prior works [Gururangan et al., 2020] have shown that continued domain-adaptive pretraining of PLMs lead to performance gains on downstream tasks. Inspired by this, we propose AERIE, a continued pretraining-based approach to inject logical reasoning abilities in language models. To gather a dataset that can teach logical reasoning, we use a set of keywords to select a subset of the Wikipedia, such that every sentence in the subset contains at least one of the keywords. These keywords are chosen such that the sentences containing the keywords are more likely to elicit reasoning when filling out masked tokens. We note that in contrast to previous works [Gururangan et al., 2020], our method only requires selecting sentences from Wikipedia, eliminating the need for extra domain-specific corpus. Secondly, we restrict the type of tokens being masked from *any* random token, to only specific types of tokens based on the parts-of-speech of the word. This choice is again based on increasing the likelihood of using logical reasoning to predict the masked word. Lastly, we add a sentence-level classification loss to predict if the reasoning in the sentence conveys an entailment or a contradiction. This enables the model to understand the differences between these two types of logical reasoning.

To test AERIE, we evaluate it on two downstream logical reasoning tasks: ReClor and LogiQA, and compare it with other baselines. We achieve state-of-the-art performance on LogiQA and comparable performance on ReClor. This demonstrates that our simple approach is generalizable to different datasets and enables the PLM to learn logical reasoning abilities.

2 Problem Statement

In this work, we study the problem of using logical reasoning to solve the task of multiple choice question answering based on a given context. Formally, for a given context C, question Q, and a list of K candidate answers $A = \{A_1, \dots A_K\}$, the task is to select the correct answer A_y, where $y \in [1, K]$. Getting to the right answer typically requires reasoning logically through the context and then selecting the best answer for the question. Evaluation of a model is based on the accuracy metric.

3 Method

In this section, we describe the details of our proposed approach. In AERIE, we use a keyword-based dataset selection strategy to collect a dataset of reasoning-related sentences called IMPLICATION (§3.1) and then continue training a pretrained model checkpoint using two loss functions jointly (§3.2). This model is then finetuned on the training dataset of each task separately.

3.1 Dataset Selection

PLMs are typically trained on the data from the internet which helps them in learning the language model and then they are finetuned on specific downstream datasets to specialize on a task [Devlin et al., 2019, Radford et al., 2018, Raffel et al., 2020]. We hypothesize that using a training data that contains more reasoning related sentences, rather than generic internet data, should help in improving the logical reasoning abilities of the PLM. Although creating such a dataset can be a challenging task in itself, in AERIE, we explore a simple and intuitive way to curate a set of such sentences. First, we select logical keywords that are generally encountered in sentences with some implication. Broadly, we categorize these keywords into two types:

- **Positive implication (Entailment)**: These keywords are typically present in sentences where the reason entails the inference. We consider the following keywords in this category: “therefore”, “accordingly”, “so”, “thus”, “consequently”, “hence”, “thence”, “and so”, “for this reason”, “in consequence”, “on account of”, “on the ”grounds”, “since”, “therefrom”, “thereupon”, “to that end”, “whence”, and “wherefore”.
- **Negative implication (Contradiction)**: In this category, the keywords are usually present in sentences where the reason contradicts the inference. Here, we consider the following

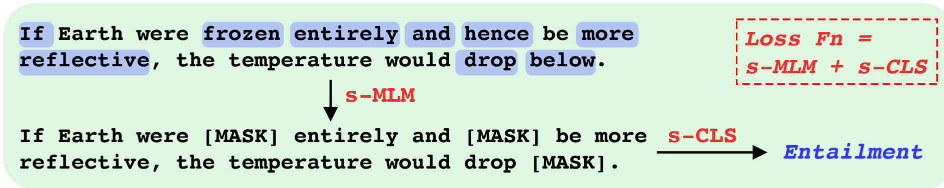

Figure 1: **Loss Functions in AERIE.** The s-MLM loss masks tokens from a specific set of POS tags (candidate tokens highlighted in blue), instead of any random token. The s-CLS loss classifies the masked sentence into one of two categories: entailment or contradiction. The overall loss function is the sum of both loss functions.

keywords: "but", "although", "however", "nevertheless", "on the other hand", "still", "though", and "yet".

Next, we select sentences that contain at least one of the keywords. Specifically, we filter sentences from Wikipedia [1] such that they contain at least one of the keywords. We name this filtered version of the Wikipedia as IMPLICATION. While this keyword-based filtering does not necessarily ensure that the sentence has an implication statement, it increases the chances of such logically rich sentence being present in the training set.

3.2 Loss Function Design

Selective masked language modeling loss (s-MLM) This a modified version of the masked language modeling (MLM) loss used in BERT [Devlin et al., 2019]. In the MLM loss, tokens in a sentence are masked at random and the model learns to predict the masked tokens. While this helps in learning a good language model, we hypothesize that not all masked tokens require similar degree of reasoning to predict them. For example, most prepositions in a sentence are generally governed by the English grammar. In contrast, some specific parts-of-speech (POS) tags such as adverbs require more reasoning to predict the right token. Thus, in s-MLM, we mask out tokens that belong to a specific set of POS tags. In AERIE, we mask tokens from the following POS tags [Honnibal and Montani, 2017]: "ADJ", "ADV", "CONJ", "CCONJ", "PART", "SCONJ", and "VERB".

Sentence classification loss (s-CLS) In addition to s-MLM, we add another auxiliary loss function that predicts whether a sentence contains reasoning that entails or contradicts the inference. To predict if a sentence is related to a positive or negative implications, a model would require strong logical reasoning abilities. The labels for this loss is bootstrapped using the simple heuristic of whether the specific type of keyword is present in the sentence. We note that although the keyword can be a direct feature that can be used to predict the label, on average the keyword would be masked more often due to our selective masking policy, leading to teaching the model some logical semantics.

4 Experimental Setup

Following prior works [Jiao et al., 2022], we evaluate AERIE on two logical reasoning datasets: ReClor [Yu et al., 2020] and LogiQA [Liu et al., 2021]. Both the datasets are reading comprehension style datasets, where the metric is the accuracy of the model in selecting the right answer for a given context and question pair. We compare AERIE with three prominent baselines: LRReasoner [Wang et al., 2022], Focal Reasoner [Ouyang et al., 2022], and MERIt [Jiao et al., 2022]. All these baselines train a PLM using some additional data to improve logical reasoning abilities.

5 Results

Overall Results We use RoBERTa-Large pretrained checkpoints as the starting point for AERIE and all the baselines. In Table 1, we compare the performance of our method with the baselines on the two logical reasoning datasets. Overall, we observe that AERIE performs at par on ReClor and outperforms all baselines on LogiQA.

Model	ReClor		LogiQA	
	Dev	Test	Dev	Test
RoBERTa	62.6	55.6	35	35.3
DAGN	65.2	58.2	35.5	38.7
DAGN (Aug)	65.8	58.3	36.9	39.3
LRReasoner	64.7	62.4	38.1	40.6
Focal Reasoner	**66.8**	58.9	41.0	40.3
MERIt	**66.8**	**59.6**	40.0	38.9
AERIE	**66.8**	57.6	**41.6**	**42.1**

Table 1: Comparison of AERIE with other baselines on ReClor and LogiQA.

[1] https://huggingface.co/datasets/wikipedia

Ablation Studies To study the effect of using IMPLICATION for continued pretraining along with the proposed loss functions, we first create RANDOM, a random subset of Wikipedia of similar size as that of IMPLICATION, and also consider using the standard masked language modeling (MLM) loss Devlin et al. [2019], where any token can be masked at random. The results of the ablation are shown in Table 2. We observe that using the IMPLICATION dataset leads to consistent improvements on both datasets, when compared to RANDOM dataset. Additionally, we find that both the S-MLM and S-CLS loss lead to improvements over MLM loss. Thus, this empirically justifies our choice of the dataset and loss functions proposed here.

6 Related Works

Reasoning in natural language has been a prevalent problem in NLP. In recent years, logical reasoning in text has seen an increasing focus. ReClor [Yu et al., 2020] and LogiQA [Liu et al., 2021] are reading comprehension style datasets focused on questions that require reasoning using information from a given context. Wang et al. [2022] proposed LRReasoner, which parses symbolic logical structures from the training data of ReClor for data augmentation using logical context extensions. Ouyang et al. [2022] constructed logical graphs using the chain of facts present in a task instance, and used GNNs to reason on the graph. Jiao et al. [2022] proposed MERIt, that used Wikipedia to generate sentence pairs for contrastive learning that are logically related, and trained the PLM using contrastive loss. Both LRReasoner and Focal Reasoner use data augmentation that are specific to the task being solved, making the pretraining process specific to the downstream dataset, and thus not generalizable across tasks. While MERIt addresses this issue by using Wikipedia to generate logical graphs, their contrastive loss formulation requires counterfactual data augmentation, that potentially distorts the factual knowledge present in the pretrained model. In contrast to prior works, we propose a simple continued pretraining strategy using minor modifications of standard masked language modeling loss [Devlin et al., 2019] and sentence classification loss to improve the logical reasoning ability of language models. Our approach is simple to integrate during pretraining, and is generalizable across tasks.

Continued Training Setup	ReClor	LogiQA
RoBERTa (RANDOM w/ MLM)	60.2	35.0
RoBERTa (RANDOM w/ S-MLM)	63.8	36.4
RoBERTa (IMPLICATION w/ MLM)	64.8	36.6
RoBERTa (IMPLICATION w/ S-MLM)	65.4	41.5
RoBERTa (IMPLICATION w/ S-MLM + S-CLS)	66.8	41.6

Table 2: Effect of IMPLICATION dataset and the loss functions on the validation performance of ReClor and LogiQA.

In a related line of work, a set of works [Clark et al., 2020, Saha et al., 2020, Tafjord et al., 2021, Sanyal et al., 2022b] used synthetically generated data to show that PLMs can perform complex deductive reasoning to predict entailment of a given hypothesis. While progress on these datasets are encouraging, some recent works have questioned if models are indeed robustly learning to perform logical reasoning Sanyal et al. [2022a].

7 Conclusion

In this paper, we proposed AERIE, an adaptive pre-trained language model with logical reasoning abilities. We use a subset of Wikipedia sentences for continued pretraining of the model using two self-supervised loss functions. The choice of the training dataset and loss functions are guided by the objective to include more reasoning related sentences and training signals, respectively. Through experiments on two logical reasoning datasets and ablation studies, we demonstrate the effectiveness of our proposed approach. Overall, we show that AERIE is a generalized solution to improving logical reasoning in language models.

References

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL `https://aclanthology.org/2020.acl-main.740`.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Fangkai Jiao, Yangyang Guo, Xuemeng Song, and Liqiang Nie. MERIt: Meta-Path Guided Contrastive Learning for Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3496–3509, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.276. URL `https://aclanthology.org/2022.findings-acl.276`.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Siru Ouyang, Zhuosheng Zhang, and hai zhao. Fact-driven logical reasoning, 2022. URL `https://openreview.net/forum?id=gKWxifgJVP`.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Swarnadeep Saha, Sayan Ghosh, Shashank Srivastava, and Mohit Bansal. PRover: Proof generation for interpretable reasoning over rules. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 122–136, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.9. URL `https://aclanthology.org/2020.emnlp-main.9`.

Soumya Sanyal, Zeyi Liao, and Xiang Ren. Robustlr: Evaluating robustness to logical perturbation in deductive reasoning, 2022a. URL `https://arxiv.org/abs/2205.12598`.

Soumya Sanyal, Harman Singh, and Xiang Ren. FaiRR: Faithful and robust deductive reasoning over natural language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1075–1093, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.77. URL `https://aclanthology.org/2022.acl-long.77`.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL `https://aclanthology.org/2021.findings-acl.317`.

Siyuan Wang, Wanjun Zhong, Duyu Tang, Zhongyu Wei, Zhihao Fan, Daxin Jiang, Ming Zhou, and Nan Duan. Logic-driven context extension and data augmentation for logical reasoning of text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1619–1629, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.127. URL `https://aclanthology.org/2022.findings-acl.127`.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*, April 2020.