(Project Proposal - Group 23)

Enhancing DSR-LM's abilities to produce robust outputs

Abhinav Gupta (agupta67@usc.edu)
Akshita Kapur (kapuraks@usc.edu)
Dhruvam Zaveri (dzaveri@usc.edu)
Hrishikesh Thakur (hthakur@usc.edu)
Shreyas Shrawage (shrawage@usc.edu)

Abstract

This document presents DSR-LM, a framework that enhances logical reasoning in language models by integrating them with a symbolic reasoning module and explores ways that can enhance its abilities to produce explainable and robust outputs. DSR-LM can extract relationships from text, apply logical rules for deductions, and autonomously learn and refine these rules, ensuring logical consistency through a semantic loss and integrity constraints. The proposed enhancements, including **Interpretable Outputs**, **Transfer Learning**, **Few-shot Learning**, and **External Knowledge incorporation**, aim to improve the system's transparency, adaptability, and accuracy, making DSR-LM a versatile tool in natural language processing.

1 Introduction

Large Language Models (LLMs), such as ChatGPT, Gemini, and several others as we see today, have become very good at solving dynamic user problems in real time. The functionality of such problems comes from their knowledge and training based on millions and billions of examples fed to them during training. However, contrary to popular belief, we, as professionals of Natural Language Processing, know that it is no magic. The Generative Pre-Trained Transformer used in such LLMs functions on an algorithm that predicts one word at a time. Most of the time, the answers make perfect sense when talking about common topics and general understanding. However, the model's answers become inaccurate as we delve deeper into niche domains and ask specific questions.

In the realm of natural language processing, tackling complex applications presents a dual challenge. On one front lies the intricate richness, nuances, and vast lexicon inherent in natural language. Meanwhile, on the other front, the endeavor demands proficiency in employing logical connectives, navigating lengthy reasoning chains, and harnessing domain-specific knowledge to derive logical conclusions.

DSR-LM (Differentiable Symbolic Reasoning for Language Models) is a framework that enhances logical reasoning in language models by integrating pre-trained models with a symbolic reasoning module. It extracts relationships from text, uses logical rules to make deductions, and can learn and fine-tune these rules automatically. The framework also introduces semantic loss and integrity constraints to ensure logical consistency. With the use of a reasoning engine and a two-part loss function, DSR-LM can be trained to balance deduction loss and semantic loss, making it adaptable to various tasks and domains while providing interpretable and logical deductions based on text input.

In the proposed enhancements to DSR-LM, we introduce interpretable outputs to make the system's reasoning process transparent and understandable to humans, allowing users to see the logical steps leading to a conclusion. Transfer Learning and Few-shot Learning will be applied to enable the system to learn new reasoning rules and adapt to new tasks with minimal training, increasing its flexibility and efficiency. Incorporating External Knowledge will allow the system to integrate additional information from outside sources, enhancing its ability to make accurate and informed deductions using a broader range of information in its reasoning process.

2 Literature review

2.1 DSR-LM

DSR-LM (Differentiable Symbolic Reasoning for Language Models) combines pre-trained language models with symbolic reasoning to enhance logical reasoning capabilities. It features a relation extraction module and a differentiable symbolic inference

Duration	Planned activities
2 weeks	Rigorous literature review and reading more papers
1 week	Finding and finalizing dataset
3 days	EDA and data cleaning
4 days	Testing various models to best suit our implementation
2-3 weeks	Implementation
1 week	Further optimizations (if possible/applicable) and report generation

Table 1: Timeline for the proposed project

module for applying logical rules. The framework can learn and fine-tune logic rules, providing interpretable outputs and the ability to incorporate external knowledge. DSR-LM has shown significant improvements in reasoning accuracy and is adaptable to various tasks and domains, making it a valuable tool in natural language processing and artificial intelligence. (Zhang, 2023)

2.2 Interpretable Outputs

Producing Interpretable Output means that the system's decisions can be understood by humans. In the context of DSR-LM, it means that the reasoning process and the rules it uses to make deductions can be explained in a way that makes sense to us. For example, if the system deduces that "Alice is Bob's niece," it can show the logical steps it took to reach that conclusion, making it easier for us to trust and understand its decision-making process. (Murdoch, 2019)

2.3 Transfer Learning and Few-shot Learning

Transfer learning is a technique where a model trained on one task is used for another related task.(Pan, 2010) Few-shot learning is a method where a model learns to perform a task with only a few examples. In DSR-LM, these concepts are applied to help the system learn new reasoning rules and adapt to new tasks with minimal additional training. This makes the system more flexible and efficient, as it can quickly adapt to new problems without needing a lot of new data. (Wang, 2020)

2.4 Incorporating External Knowledge

This refers to the ability of the system to integrate information from outside sources into its reasoning process. In DSR-LM, this means that the system can take into account additional rules or facts that are not explicitly mentioned in the input text but are relevant to the reasoning task. This helps the system make more accurate and informed deductions, as it can use a broader range of information in its reasoning. (Paulheim, 2017)

3 Timeline

Table 1 shows the prospective activities and tentative duration required for executing those activities within a span of eight weeks.

References

Singh C. Kumbier K. Abbasi-Asl R. Yu B. Murdoch, W. J. 2019. Interpretable machine learning: definitions, methods, and applications. *National Academy of Sciences*.

Yang Q. Pan, S. J. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*.

H. Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web, 8(3), 489-508*.

Yao Q. Kwok J. T. Ni L. M. Wang, Y. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*.

Huang J. Li Z. Naik M. Xing E Zhang, H. 2023. Improved logical reasoning of language models via differentiable symbolic programming.