Thesis for Bachelor's Degree

Augmenting Few-Shot Deomonstrations with Large Language Model

Seo, Wongyu (서 원규 徐 源叫)

School of Electrical Engineering and Computer Science

Gwangju Institute of Science and Technology

2024

Augmenting Few-Shot Deomonstrations with Large Language Model

대형 언어 모델을 활용한 퓨샷 추론 문제의 데이터 증강

Augmenting Few-Shot Deomonstrations with Large Language Model

Advisor : Professor Kim, Sundong

by

Seo, Wongyu

School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology

A thesis submitted to the faculty of Gwangju Institute of Science and Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Electrical Engineering and Computer Science

Gwangju, Republic of Korea

2023. 12. 21.

Approved by

Professor Kim, Sundong

Committee Chair

Augmenting Few-Shot Deomonstrations with Large Language Model

Seo, Wongyu

Accepted in partial fulfillment of requirements
for the degree of Bachelor of Science

December. 21. 2023.

Committee Chair _____
Prof. Sundong Kim

Committee Member _____
Prof. Jeany Son

BS/EC
20185084

Seo, Wongyu (서 원규). Augmenting Few-Shot Deomonstrations with Large Language Model (대형 언어 모델을 활용한 퓨샷 추론 문제의 데이터 증강). School of Electrical Engineering and Computer Science. 2024. 11p. Advisor Prof. Kim, Sundong.

Abstract

The ARC (Abstraction and Reasoning Corpus) problems involve inferring logical relationships between problem inputs and outputs. Each ARC problem is characterized by distinct logical relationships. Consequently, for artificial intelligence to resolve ARC problems, it must comprehend the logical connections between inputs and outputs. In situations where the number of demonstrations is limited, many AI models struggle with logical relationship inference. Therefore, there is a need to generate additional demonstrations that form the basis for logical relationship inference. In this study, we aim to leverage the inferential capabilities of large language models to create supplementary demonstrations. We propose the method of "Problem Classification and Input Prediction" for data augmentation.

Contents

Chapter 1. Introduction

The ARC (Abstraction and Reasoning Corpus) dataset, released in 2020[**?**, **?**, paper1] is designed to evaluate the general intelligence of artificial intelligence. This dataset consists of problems that assess object inference and geometric abilities. Each problem, as shown in [Figure 1], is composed of around three demonstrations and a test problem. The task involves understanding the logical relationships in ARC problems through demonstration problems and predicting the output for the test problem. However, due to the limited number of demonstrations, computers have struggled in inferring logical relationships. Therefore, this research aims to discuss efficient methods for generating additional demonstrations that form the basis for logical relationship inference.

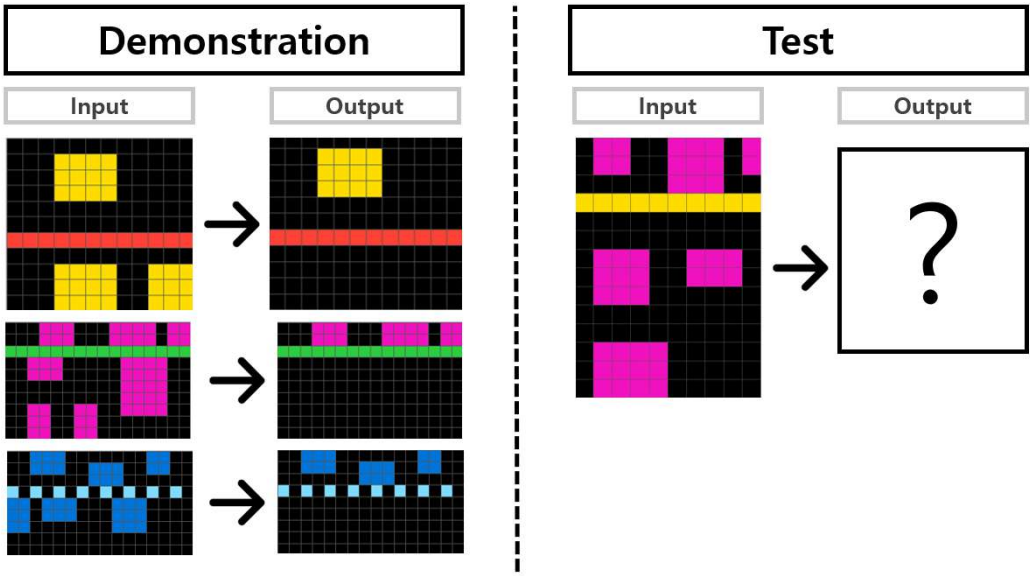

Figure 1.1: Problem of inferring rules based on input-output data from ARC demonstrations and predicting output for test inputs

According to prior research[2] large language models like GPT-4.0 face difficulties in inferring the logical relationships in ARC problems. Considering this challenge, it is challenging to expect large language models to effectively solve ARC problems. Moreover, in situations requiring an understanding of complex inferences and precise operations, the performance of Transformer-based Large Language Models (LLM) is reported to sharply decline [paper3]. However, it has been observed that with appropriate use of prompting techniques, similar results to the answers of ARC problems can be obtained [paper2]. Additionally, research suggests that applying lenient evaluation criteria, where similar answers are also accepted as correct, can make large language models useful [paper3]. Therefore, this research aims to investigate the potential of augmenting ARC demonstrations through large language models.

Chapter 2. Background and Related work

2.1 Difference between Human and AI

The human brain, with its intricate network of neurons, remains one of the most sophisticated and adaptable computing systems known. Its ability to learn, reason, and process vast amounts of information is awe-inspiring. AI, on the other hand, represents our quest to impart machines with cognitive capabilities akin to human intelligence. Machine learning algorithms, inspired by the neural connections in the brain, enable AI systems to recognize patterns, make predictions, and improve their performance over time.

Most of the current artificial intelligence is often specialized in one special task. Many problems can be solved through such artificial intelligence, but there was a problem that not only a large amount of data was required to solve the problem, but only some tasks that could be solved using the data could be solved.

Humans exhibit a remarkable capacity for task-solving driven by intuition, creativity, and adaptability. Unlike AI, which often requires extensive datasets and explicit instructions, humans can generalize their knowledge and apply it to a wide range of tasks with minimal examples. The human brain leverages past experiences, innate understanding, and the ability to infer meaning from limited information.

Consider problem-solving scenarios where a person encounters a novel situation. The human mind draws upon a wealth of diverse experiences, allowing for quick adaptation and creative problem-solving. Humans can learn from a few examples, extrapolate patterns, and apply this knowledge to unforeseen challenges. This cognitive flexibility enables humans to navigate complex, ever-changing environments effectively.

In contrast, AI systems, particularly those powered by machine learning algorithms, often demand extensive datasets for effective task-solving. Training AI models requires exposing them to a plethora of examples to identify patterns and make accurate predictions. This data-centric approach is especially evident in narrow AI, where systems excel at specific tasks but struggle when faced with tasks outside their trained scope.

2.2 ARC and Intelligence

ARC can be thought of as a problem data set that evaluates computer intelligence, and I expect it to be called true artificial intelligence in that it evaluates intelligence rather than problem-solving ability. In particular, I think the ARC problem is different from artificial intelligence, which requires a large amount of data in that it needs to be solved after inferring the logical relationship through few number of examples.

Chapter 3. Methodology

3.1 Necessity to categorize ARC

In this research, we utilized a large language model to generate additional few-shot demonstration data for each ARC problem. To facilitate data generation, it was necessary to construct appropriate prompts that assist the inference capabilities of the large language model. To compose more precise prompts, we classified ARC problems by type and crafted prompts accordingly. Establishing rational criteria for problem type classification was challenging due to the diverse logical relationships present in each problem.

In this research, we conducted experiments by replacing traditional ARC problems with newly generated ones based on an already established classification system in the ConceptARC dataset [4]. ConceptARC consists of newly created ARC problems aligned with the classification system, encompassing a total of 16 types. Utilizing this classified typology, we formulated prompts to augment ARC demonstration data.

3.2 ARC Data Augmentation Method

3.2.1 The relationship between input and output

In ARC (Abstraction and Reasoning Corpus) often involves a one-to-many correspondence, as depicted in [Figure 2], where multiple inputs correspond to a single output. Understanding this one-to-many relationship is crucial for augmenting ARC demonstrations. Predicting inputs based on outputs allows the generation of additional input-output pairs, enabling the application of lenient evaluation criteria, as mentioned in Chapter 1, due to the diversity of possible answers.

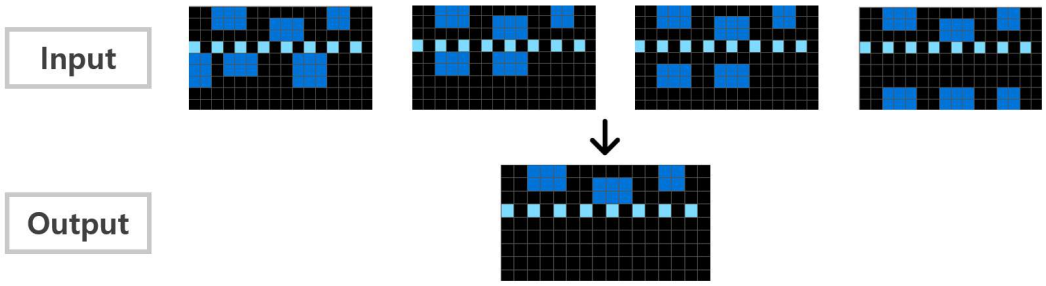

Figure 3.1: Prorepresents an Above and Below problem, where the lower part is removed based on the cyan dashed line. This problem exhibits a one-to-many correspondence, where multiple inputs correspond to a single output

3.2.2 Prompt for helping Logical Relationship Inference

We utilized a model from the GPT (Generative Pre-trained Transformer) series [5], a class of large language models, to aid in the inference of logical relationships in ARC problems. Currently, models

in the GPT series face challenges in deducing the logical relationships in ARC problems. Therefore, it was necessary to craft additional prompts to assist large language models in deducing these logical relationships.

Following the types introduced in Chapter 2.1, as exemplified by ConceptARC, we tailored prompts accordingly. These prompts were designed to assist in predicting inputs from outputs, offering a form of inverse transformation method (output → input). For instance, for the Above and Below type, a prompt like "Carefully examine above and below the horizontal baseline, then apply the observed changes" was provided in English. Similarly, for the Center type, the prompt suggested, "Check if something in the center has moved or been removed. You can verify through demonstrations."

3.2.3 Demonstration Augmentation Process

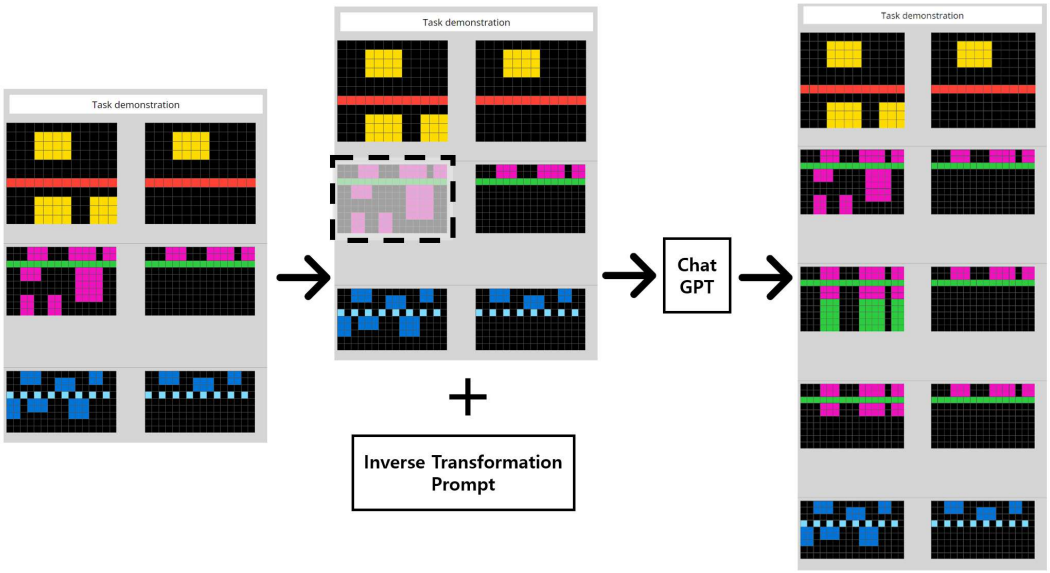

Figure 3.2: Demonstration data augmentation process. process to augment second demonstration

Figure 3.2 illustrates the process of augmenting an demonstration for the Above and Below problem. We fed the input-output of the problem to a large language model as a 2D array. Only the output of the second demonstration was shown, while the rest (in this case, the first and third demonstrations) displayed input-output pairs. For the inference of the logical relationship in the ARC problem, we employed the reverse transformation prompt specific to the Above and Below type, which was "Carefully examine above and below the horizontal baseline, then apply the changes." Using this information, we inferred the input for the second demonstration, generated new input-output pairs, and repeated this process for the remaining demonstrations (in this case, the first and third demonstrations).

3.2.4 Definition of Augmented Data

Let's denote a specific ARC problem as T. T consists of n demonstrations d_1, d_2, d_3, ... , d_n . Each demonstration d is composed of input x and output y. Every demonstration d belonging to T has a unique solution f that satisfies (x y). Let's call the set of demonstrations generated for problem

T using the method introduced in section 2.2 as T_G. If an element d in T_G can be resolved by f, then d is considered a valid augmented demonstration.

Chapter 4. Experiment

4.1 Experimental Result

The experiment was conducted using GPT-4.0 32k with a temperature setting of 1.0 for augmentation. While there were cases, as depicted in [Figure 4], where augmentation was appropriately performed, it is notable, as indicated in [Table 1], that instances of inaccurately predicting inputs occurred frequently. The cases of such inaccuracies will be analyzed in the following section.

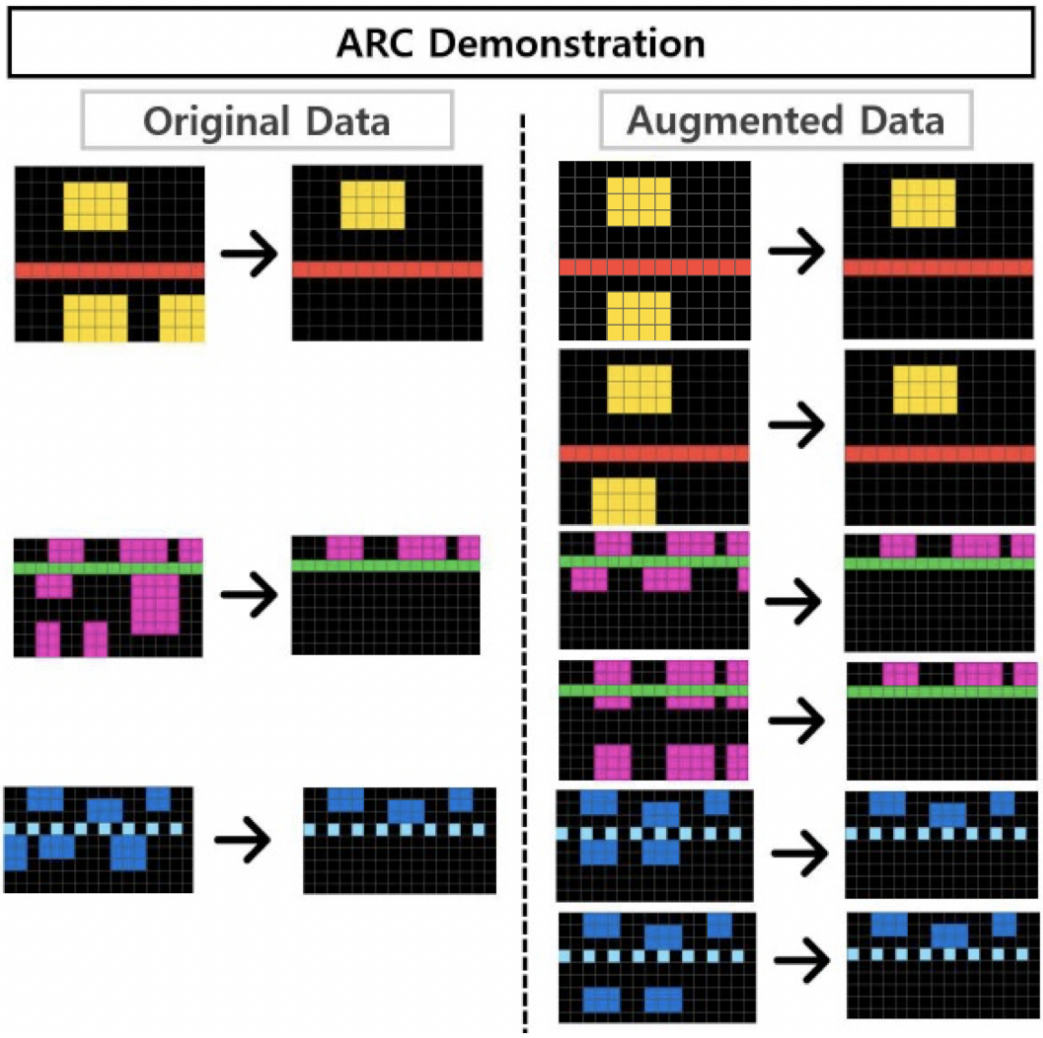

Figure 4.1: corresponds to cases where the demonstrations in the ARC question were appropriately augmented.

Since the current amount may be sufficiently modified through the number of augmentation, the prompt quality and the augmented ratio may be checked through the difference between the ratio of original data and valid data for each current category.
In addition, since an experiment for classifying ARC is currently underway in the laboratory, it is

Table 4.1: The quantity of generated data and valid data

Category	Original Data	Valid Data	Data
Above Below	24	34	58
Center	30	35	65
Clean Up	23	83	106
Complete Shape	21	37	58
Copy	23	4	27
Count	27	29	56
Extend To Boundary	29	8	37
Extract Objects	23	21	44
Filled Not Filled	29	29	58
Horizontal Vertical	25	7	32
Inside Outside	29	24	53
Move To Boundary	25	12	37
Order	21	26	47
Same Different	33	76	109
Top Bottom 2D	34	59	93
Top Bottom 3D	31	25	56
Total	427	509	936

considered necessary to evaluate how useful the augmented ARC data is through this.

4.2 Experimental Analysis

Figure 4.2 As evident from the diagram, despite being of the same Complete Shape type, the problem-solving approaches for the input-output pairs vary significantly. For the left problem (Complete Shape 1), a suitable reverse prompt might be "Remove a portion of the object corresponding to the symmetry in all directions." On the other hand, for the right problem (Complete Shape 5), a prompt like "Change one part with a different color in the 2 x 2 square to black" is needed. Attempting to create a universal prompt to describe such diverse prompts proved challenging and abstracting the prompts couldn't adequately explain the reverse transformation methods.

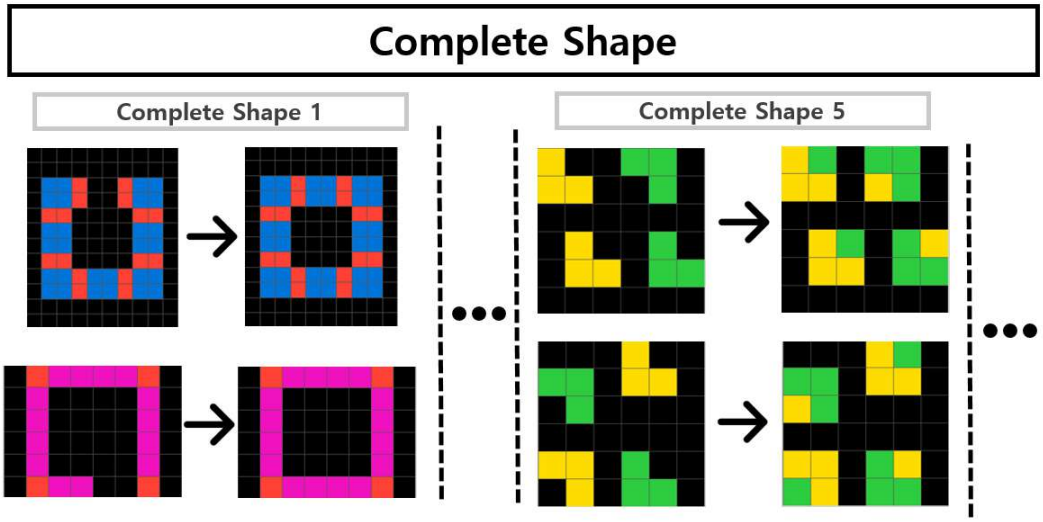

Figure 4.2: corresponds to cases where the demonstrations in the ARC question were appropriately augmented.

The inference result from Prompt

Despite the application of the Negative Prompting technique to prevent the direct duplication of examples, instances were observed, as illustrated in the [Figure 6] , where the input of the example's input-output pair was manipulated to generate the output.

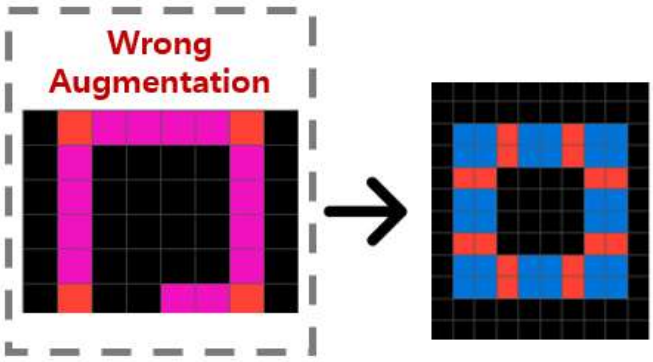

Figure 4.3: An example of augmenting Complete Shape 1, where the inference is made using the example input

Wrong inference from LLM

In cases where the conventional method represented by f cannot resolve the issue with d, it was possible to find situations where human intervention is necessary for selecting a solution until the emergence of a model capable of verifying whether the augmented example d, enhanced by f, can be resolved.

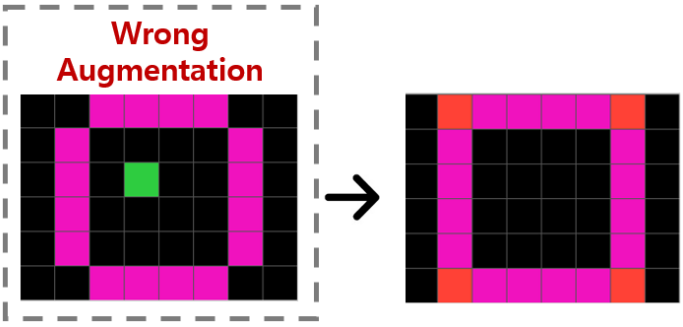

Figure 4.4: represents an incorrect inference result for the Complete Shape 1 problem. In this case, it is not possible to infer the colors of the corners of the square based on the input image

Impossible to Augment because of one-to-one correspondence

Our research was based on the one-to-many relationship between input and output pairs for data augmentation. However, in cases where the input-output relationship is one-to-one, as shown in [Figure 8], it is not feasible to augment examples using the method proposed in this research.

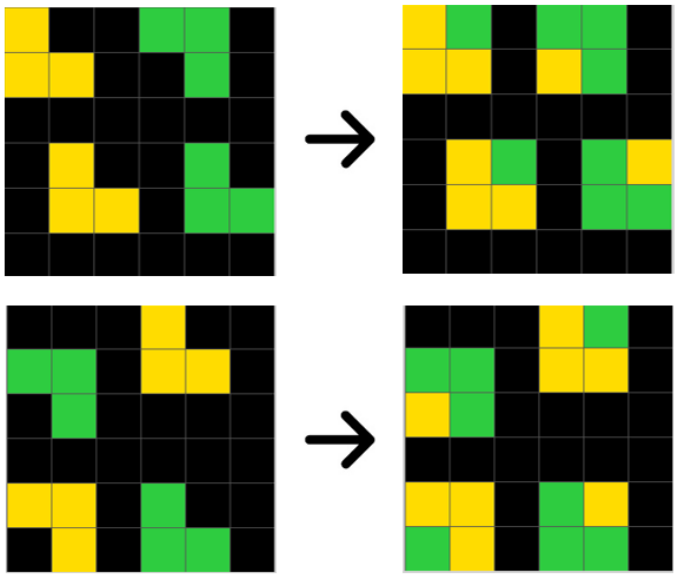

Figure 4.5: Complete Shape 5th Problem

Chapter 5. Conclusions and Forthcoming Research

This research confirmed the feasibility of data augmentation through large language models for ARC problems with a one-to-many correspondence between input and output. If improvements are made to the augmentation method using prompts and large language models, it is believed that more diverse data can be obtained. In the future, the goal is to go beyond ConceptARC and augment examples for all ARC problems. To achieve this, improvements are needed in three areas. Firstly, in this research, the usability of generated data was determined through human intervention. However, manual interventions run the risk of being subjective depending on the classifier. Therefore, future research should focus on designing models that can automatically filter and identify cases where large language models make incorrect inferences using clear criteria, reducing the need for manual intervention. Secondly, to apply the proposed method to ARC problems, it is necessary to establish a clear classification system for these problems. Hence, further research, such as representation learning for ARC problem classification, is needed. Lastly, the current research has the limitation that the output for each question is fixed during augmentation. Future research should explore methods for generating new outputs, implying the possibility of creating a broader range of examples with new input-output pairs.

References

1. Francois Chollet, “On the Measure of Intelligence”, arXiv:1911.01547, 2019.

2. Xu, Yudong, et al. ”LLMs and the Abstraction and Reasoning Corpus: Successes, Failures, and the Importance of Object-based Representations.” arXiv preprint arXiv:2305.18354 (2023).

3. Dziri, Nouha, et al. ”Faith and Fate: Limits of Transformers on Compositionality.” arXiv preprint arXiv:2305.18654 (2023).

4. Moskvichev, Arseny, Victor Vikram Odouard, and Melanie Mitchell. ”The ConceptARC Benchmark: Evaluating Understanding and Generalization in the ARC Domain.” arXiv preprint arXiv:2305.07141 (2023).

5. OpenAI, “GPT-4 Technical Report”, arXiv:2303.0877, 2023.

6. Oppenlaender, Jonas, Rhema Linder, and Johanna Silvennoinen. ”Prompting ai art: An investigation into the creative skill of prompt engineering.” arXiv preprint arXiv:2303.13534(2023).

Summary

Augmenting Few-Shot Deomonstrations with Large Language Model

현재 ARC 문제는 적은 예제의 수로 인해 입력과 출력 간의 논리적 관계를 찾는 것이 힘듭니다. 현재의 인공지능을 최대한 활용하고, 또 이를 발전시켜 나가면서 ARC 문제에 접근하고자 합니다. 따라서 본 연구의 경우는 ARC 문제를 풀기 위해서 대형 언어 모델(Chat-GPT)을 활용하여 예제의 수를 증강하는 것에 초점을 맞추고 있습니다.

인공지능의 발전은 지난 몇년 동안 놀라운 속도로 발전됐습니다. 특히 자연어 처리, 컴퓨터 비전에 집중된 경향이 있었습니다. 하지만, 자연어 처리 및 컴퓨터 처리는 인공 일반 지능(AGI)와는 다른 부분에 초점을 맞추고 있었다고 생각합니다. 이러한 인공지능은 머신러닝을 바탕으로 데이터에서 패턴을 학습하고, 그와 유사한 문제를 해결할 수 있는 능력이 크게 향상되어지만, 주어진 데이터에서 벗어난 문제를 해결하는 것에는 어려움을 겪고 있었습니다. 하지만, 딥러닝과 신경망의 발전은 인간의 뇌 시냅스를 모방하면서 점차 발전해나가고 있습니다. 우리는 이러한 발전을 바탕으로 ARC와 같은 퓨샷 추론 문제를 해결할 수 있기를 바랍니다.

감 사 의 글

논문 작성을 처음 접해서 방황하고 헤매던 제게 조언을 아끼지 않으신 김선동 지도교수님께 감사드립니다. 또한 귀한 시간을 내어 논문 심사를 봐주신 손진희 교수님께 감사드립니다. 이 외에도 제게 도움을 주신 모든 분들께 감사드립니다.

약 력

이　　름: 서원규
생 년 월 일: 2000년 5월 24일
출 생 지: 부산광역시
주　　소: 부산광역시 수영구 광안해변로326번길 31

학　　력

2016. 3. – 2018. 2.　부산일과학고등학교
2018. 2. – 2024. 2.　광주과학기술원 전기전자컴퓨터공학부 (학사)