ML Week11 Assignment

WANG TZU YI

November 18, 2025

1. Introduction

This project presents an innovative approach to enabling large language models (LLMs) to perform **visual reasoning** by transforming visual grid-based tasks into **textual sequence-learning tasks**. The core insight is that, while LLMs are naturally optimized for symbolic and linguistic patterns, many visual reasoning problems—such as those found in the ARC-AGI benchmark—can be encoded in a form amenable to linguistic processing. By converting grids, colors, and transformations into structured textual representations, the model learns to reason about visual concepts through language modeling.

This work represents the first step toward building a general reasoning system that unifies visual and linguistic patterns under a single computational framework.

2. Core Idea: Converting Visual Reasoning into Text

2.1 Grid-to-Text Representation

Each ARC task consists of one or more colored grids. In this project, a grid is converted into a tokenized textual array format. For example, a 3×3 grid is represented as a sequence of digits arranged in rows, such as:

```
[[2, 0, 0],
 [3, 1, 1],
 [2, 2, 0]]
```

This representation preserves the full structure of the visual pattern while allowing the LLM to interpret it using its existing sequence-processing capabilities.

2.2 Few-Shot Learning Structure

Using the `prepare_fine_tuning_dataset` function, each ARC task is rewritten into a few-shot prompt of the following format:

- Several **context examples**, each containing:
  - an input grid,
  - its corresponding output grid.
- A **query grid** that requires reasoning.
- The **target output grid** (for supervised training).

This structure encourages the model to infer a transformation rule from the examples and apply it to the query, mimicking human reasoning.

3. Techniques for Improving the Model's Reasoning Ability

3.1 Data Augmentation for Enhanced Generalization

Several systematic transformations are applied to each grid through `apply_color_swaps`, `add_rotations`, `add_mirrors`, and `add_shuffled`. These augmentations serve distinct reasoning purposes:

1. **Color Permutations**: Teach the model to focus on abstract patterns rather than specific colors.
2. **Geometric Transformations**: Rotations and reflections enhance spatial reasoning and symmetry recognition.
3. **Order Shuffling**: Prevents the model from overfitting to position bias, enforcing pattern consistency.

Together, these augmentations significantly broaden the model's ability to generalize across task variations.

3.2 Parameter-Efficient Fine-Tuning with LoRA

Using the `train.ipynb` workflow, the model is fine-tuned with the LoRA technique. Remarkably, only about **0.65%** of the model parameters are trained, yet this is sufficient for the model to learn transformation patterns typical of ARC reasoning tasks.

This demonstrates that LLMs already contain strong internal inductive biases for pattern recognition, and lightweight adaptations can unlock new modalities of reasoning.

3.3 Training Structure: Learning Patterns

Through the `build_rows_for_task` function, each task is decomposed into:

- **Context examples** for demonstrating the transformation.
- **A query instance** requiring reasoning.
- **The correct answer** in textual grid form.

The design mirrors human induction: observe a few examples, infer a rule, apply it.

4. Emergent Reasoning Capabilities

4.1 Abstract Pattern Recognition

The fine-tuned model exhibits the ability to recognize a wide range of abstract visual transformations:

- Symmetry and geometric reflection
- Rotational relationships
- Color-mapping and palette transformations
- Spatial relocation and grouping of shapes
- Logical composition of multiple visual rules

These abilities emerge *without* any explicit visual encoding architecture, purely from grid-to-text transformations.

4.2 Generalization Across Tasks

Using the multi-example structure processed by `_split_dictionary`, the model learns to:

- Infer rules from a small number of examples,
- Apply learned transformations to novel test cases,
- Handle variable input sizes and complex task structures,
- Maintain robustness across augmented examples.

4.3 Text-Based Visual Reasoning

The evaluation notebook `results.ipynb` confirms that the model can:

- Interpret visual patterns in text form,
- Infer visual transformations,
- Output the correct grid in textual numeric format,
- Successfully convert its reasoning back into structured visual predictions.

The `plot_prediction_comparison` tool visualizes predicted and ground-truth grids, allowing qualitative assessment of the model's reasoning accuracy.

5. Contributions and Innovations

This project demonstrates several novel contributions:

1. **Cross-modal Reasoning via Language**: Visual reasoning tasks are recast as symbolic sequence prediction, allowing LLMs to operate beyond natural language.
2. **Parameter-Efficient Training**: Only a small fraction of weights are updated, yet substantial reasoning ability emerges.
3. **Structured Data Augmentation**: A comprehensive augmentation pipeline enhances generalization and robustness.
4. **Interpretability**: The entire reasoning process is observable as a textual sequence, providing transparency uncommon in vision models.

6. Conclusion

This project demonstrates that large language models, when provided with appropriate representations and training strategies, are capable of performing non-linguistic reasoning tasks such as visual pattern inference. By converting ARC visual puzzles into structured text sequences and fine-tuning with highly efficient methods, we show that LLMs can acquire meaningful visual reasoning abilities without explicit visual encoders.

This constitutes an important first step toward my long-term goal: enabling AI systems to perform general-purpose reasoning across modalities. Through this work, I have shown that language models can extend beyond text and begin to solve structured, rule-based visual tasks, contributing a promising direction to the broader pursuit of AGI.