RADAR: A Reasoning-Guided Attribution Framework for Explainable Visual Data Analysis

Keywords: Mathematical Reasoning, Attribution, Multimodal Large Language Models

Data visualizations like charts are fundamental tools for quantitative analysis and decision-making across fields, requiring accurate interpretation and mathematical reasoning. The emergence of Multimodal Large Language Models (MLLMs) offers promising capabilities for automated visual data analysis, such as processing charts, answering questions, and generating summaries. However, they provide no visibility into which parts of the visual data informed their conclusions; this black-box nature poses significant challenges to real-world trust and adoption. We introduce RADAR, a semi-automatic approach to obtain a benchmark dataset comprising 17,819 diverse samples with charts, questions, reasoning steps, and attribution annotations. We also introduce a method that provides attribution for chart-based mathematical reasoning.

Our method addresses attribution through a two-stage pipeline using InternLM-XComposer2 model [1]. Given a chart-question-answer triple, we first generate step-by-step reasoning, then leverage these reasoning steps along with the chart, question, and answer to produce attribution bounding boxes for both the final answer and intermediate reasoning steps. The model architecture incorporates a vision encoder (CLIP ViT-Large) [2] that processes charts into a 35×35 grid (1225 visual tokens) and maps them to a shared 4096-dimensional embedding space with text from InternLM-2. We employ Partial LoRA (PLoRA) [3], which applies additional trainable parameters specifically to visual tokens while preserving the base 7B-parameter language model's capabilities.

We conducted experiments on three tasks: (i) Attribution based on Visual Question Answering (VQA), (ii) Attribution based on Visual Question Reasoning (VQR), and (iii) Answer generation based on feature attribution. We evaluated RADAR against three state-of-the-art MLLMs: GPT-4o, GPT-4v, and Claude 3.5 Sonnet using both zero-shot and few-shot prompting strategies. For attribution evaluation, we extend the Intersection Over Union (IOU) score to create the Multiple Box IOU Score, adapting the metric from single to multiple bounding boxes. For reasoning and answer evaluation, we employ BERTScore and Semantic Textual Similarity (STS).

Experimental results demonstrate that our reasoning-guided approach improves attribution accuracy by 15% compared to current MLLMs (For example, Claude 3.5 Sonnet). The generated reasoning achieves strong alignment with ground truth (BERTScore precision of 0.8928 and STS scores of ~0.74 across chart types). This advancement represents a significant step toward more interpretable and trustworthy chart analysis systems, enabling users to verify and understand model decisions through reasoning and attribution. Our framework provides a foundation for building more trustworthy and interpretable AI systems for mathematical reasoning tasks, paving the way for chart-based systems that can better explain their decision-making processes.

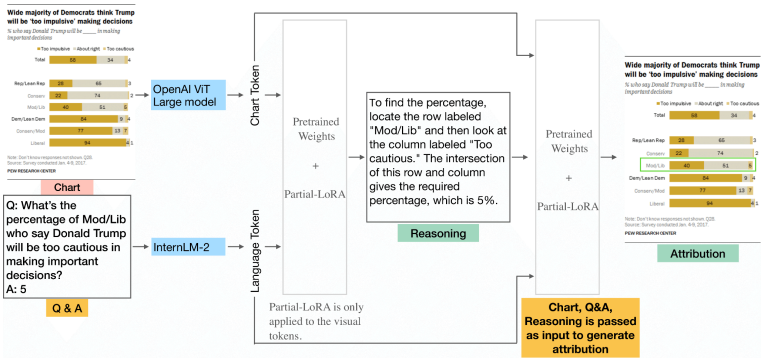

Figure: Overview of our proposed RADAR method for reasoning & attribution generation.

[1] Dong, Xiaoyi, et al. "Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model." arXiv preprint arXiv:2401.16420 (2024).

[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

[3] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.