Sparse Multimodal Model Alignment and Reasoning Accuracy with Expert Scaling

Assignee Research

May 29, 2026

Abstract

The multimedia community has shown a significant interest in perceiving and representing the physical world with multimodal pretrained neural network models, and among them, the visual-language pertaining (VLP) is, currently, the most captivating topic. However, there have been few endeavors dedicated to the exploration of 1) whether essential linguistic knowledge (e.g., semantics and syntax) can be extracted during VLP, and 2) how such linguistic knowledge impact or enhance the multimodal alignment. In response, here we aim to elucidate the impact of comprehensive linguistic knowledge,

1 Introduction

This paper examines: Can Linguistic Knowledge Improve Multimodal Alignment in Vision-Language Pretraining?. Research question: How does the alignment score (e.g., via RLHF or DPO) of a sparse multimodal model change as the number of experts increases on the VQAv2 benchmark, and does this correlate with improved reasoning accuracy on OK-VQA?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

11 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Fine-tuning BERT on GLUE may overlook word order information.	×	0.07
Word order information is not important during pretraining of large language models.	×	0.06
Transformers use co-occurrence statistics of content words to predict next words for long-range contexts.	×	0.04
BERT is insensitive to negation factors.	×	0.02
VLP models show low sensitivity of multimodal alignment to word order.	×	0.12
Contextual information in images affects model's understanding of text.	×	0.06
Pretraining models emphasize textual information during inference.	×	0.03
Text encoder feature representations are more influenced by visual features.	×	0.02
Vision language models have poor perception of object quantity information.	×	0.05
MLLMs face limitations in recognizing complex visual content.	×	0.04

References

- http://arxiv.org/abs/2407.17856v4
- http://arxiv.org/abs/2308.12898v2
- http://arxiv.org/abs/2312.11456v4