Hierarchical Cross-Attention Integration in PaLI with EfficientViT for Image-Text Matching Recall

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the integration of a hierarchical cross-attention mechanism in PaLI with EfficientViT impact the Image-Text Matching (ITM) recall metrics on the ECCV Caption dataset compared to the original. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO. Research question: How does the integration of a hierarchical cross-attention mechanism in PaLI with EfficientViT impact the Image-Text Matching (ITM) recall metrics on the ECCV Caption dataset compared to the original PaLI architecture?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.0/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.0/10.

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.

References

- http://arxiv.org/abs/2204.03359v5
- http://arxiv.org/abs/2407.20114v3
- http://arxiv.org/abs/2209.06794v4