Attention-Augmented Contrastive Loss for Zero-Shot Cross-Modal Retrieval in Flickr30K

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of integrating attention mechanisms with contrastive loss on zero-shot cross-modal retrieval accuracy for image-text pairs in Flickr30K. 18 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: End-to-End Cross-Modality Retrieval with CCA Projections and Pairwise Ranking Loss. Research question: What is the impact of integrating attention mechanisms with contrastive loss on zero-shot cross-modal retrieval accuracy for image-text pairs in Flickr30K?.

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 4.4/10.

3 Results

14 papers retrieved. 18 claims extracted; 2 independently verified. Quality review score: 4.4/10.

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
The embedding spaces should account for low-level cross-modal correlations.	×	0.09
The embedding spaces should enable semantic abstraction.	×	0.05
DCCA ensures highly correlated latent representations.	×	0.03
The optimization of pairwise ranking losses yields embedding spaces that are useful for retrieval.	×	0.14
DCCA is designed to maximize correlation.	×	0.08
DCCA does not allow to use loss formulations specialized for the task at hand.	×	0.02
The proposed method performs better than DCCA and models using pairwise ranking loss alone, especially when little train	✓	0.19
DCCA defines an objective optimizing a dual-view neural network such that its two views will be maximally correlated.	×	0.04
Pairwise ranking losses are loss functions to optimize a dual-view neural network such that its two views are well-suite	✓	0.18
The proposed approach boosts optimization of a pairwise ranking loss based on cosine distance by placing a special-purpo	×	0.13
The objective of CCA is to find two matrices A* and B* composed of k paired column vectors Aj and Bj that project x and	×	0.05
The proposed method (CCAL-Lrank) achieves R@1 of 31.6, R@5 of 61.0, R@10 of 72.2, MR of 3.0, and MRR of 45.0 for image-t	×	0.06
The proposed method (CCAL-Lrank) achieves R@1 of 32.0, R@5 of 59.2, R@10 of 70.4, MR of 3.2, and MRR of 44.8 for text-to	×	0.10
The dataset contains 18,046 validation and 16,042 test audio-sheet-music pairs.	×	0.06
The proposed method uses pre-trained ImageNet features and relatively shallow fully connected text-feature processing ne	×	0.03
The convolutional networks are learned entirely from scratch.	×	0.02
The architecture is a VGG-style network consisting of sequences of 3 $\times$ 3 convolution stacks followed by 2 $\times$ 2 max4poolin	×	0.01
The final building block is a 1$\times$1 convolution having k feature maps followed by global average pooling.	×	0.01

References

- http://arxiv.org/abs/2410.12595v1
- http://arxiv.org/abs/2308.15273v1
- http://arxiv.org/abs/1705.06979v2