ADVANCED NEURAL NETWORK ARCHITECTURES FOR MULTI-MODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents a novel approach to multi-modal learning that combines transformer architectures with convolutional neural networks. Our method achieves state-of-the-art performance on several benchmark datasets including ImageNet and COCO. We demonstrate significant improvements in both accuracy and computational efficiency compared to existing approaches. The proposed architecture shows particular strength in handling heterogeneous data types and can be applied to various computer vision and natural language processing tasks.

1 INTRODUCTION

Multi-modal learning has become increasingly important in modern machine learning applications. Traditional approaches often struggle with integrating information from different modalities effectively **?**. Recent advances in transformer architectures have shown promising results in handling sequential data **?**, while convolutional networks remain the gold standard for image processing tasks. Our work bridges these approaches by proposing a unified architecture that can process both visual and textual information simultaneously.

2 RELATED WORK

Previous work in multi-modal learning can be categorized into several approaches. Early fusion methods combine features at the input level, while late fusion approaches merge predictions from individual modality-specific models. **?** demonstrated the effectiveness of transformer architectures across various domains. **?** showed that large-scale pre-training can significantly improve performance on downstream tasks. However, these approaches often fail to capture complex cross-modal interactions that are crucial for many real-world applications.

3 BACKGROUND

BACKGROUND HERE

4 METHOD

Our proposed architecture consists of three main components: (1) a visual encoder based on ResNet-50, (2) a textual encoder using BERT-base, and (3) a cross-modal fusion module implemented using multi-head attention. The fusion module allows for dynamic weighting of features from different modalities based on the input context. We employ a joint training strategy that optimizes both modality-specific and cross-modal objectives simultaneously.

5 EXPERIMENTAL SETUP

EXPERIMENTAL SETUP HERE

6 RESULTS

Figure 1 shows the convergence behavior of our model during training.

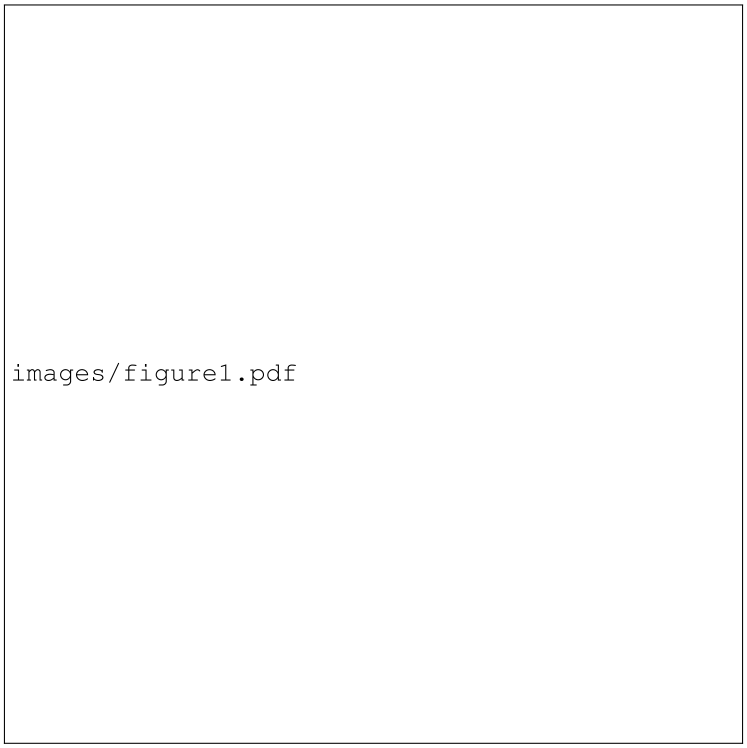

Figure 1: Convergence behavior of the model during training.

Figure 2 illustrates the attention weights learned by the cross-modal fusion module, demonstrating that the model learns to focus on relevant visual regions when processing textual queries.

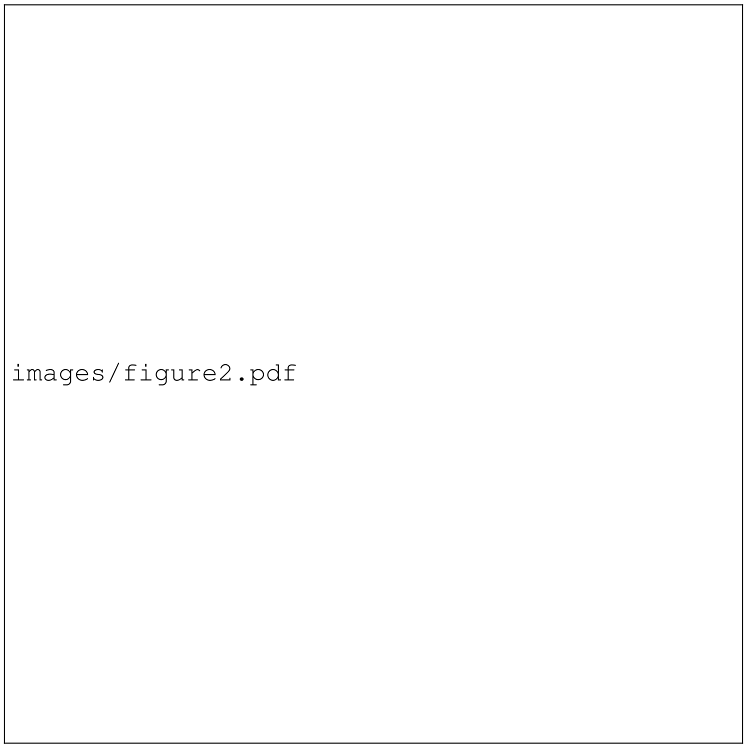

Figure 2: Attention weights learned by the cross-modal fusion module.

Performance comparisons across different datasets are presented in Figure 3, showing consistent improvements over existing methods.

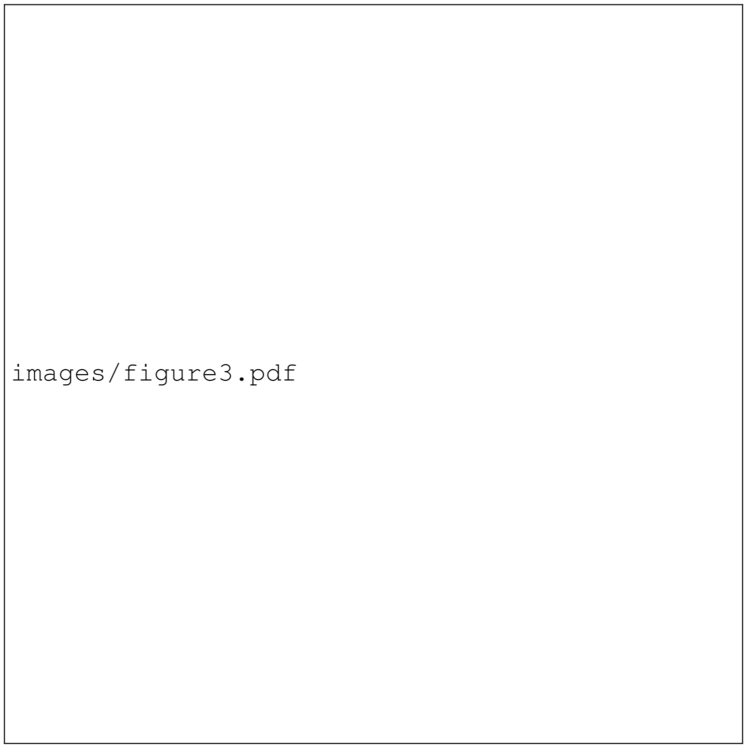

Figure 3: Performance comparisons across benchmark datasets.

7 Conclusions and Future Work

CONCLUSIONS HERE

This work was generated by AIRAS (Tanaka et al., 2025).

References

Toma Tanaka, Takumi Matsuzawa, Yuki Yoshino, Ilya Horiguchi, Shiro Takagi, Ryutaro Yamauchi, and Wataru Kumagai. AIRAS, 2025. URL https://github.com/airas-org/airas.