Attention Mechanisms and Their Impact on Computer Vision Models

By: Mohammed Siraj

Addis Ababa Institute of Technology

June 21 2024

Contents

Introduction

Introduction

- Overview of attention mechanisms in deep learning.
- Definition and importance of attention in computer vision.
- Scope and objectives of the survey.

Background

Background

- Early attention mechanisms in NLP and their adaptation to vision tasks.
- Key concepts: self-attention, multi-head attention, and scaled dot-product attention.
- Advantages of attention mechanisms: capturing long-range dependencies, flexibility in handling variable-sized inputs.

Literature Review

Key Paper: Vision Transformer (ViT) by Dosovitskiy et al.

- Introduced the Vision Transformer (ViT) architecture.
- Showed competitive performance on ImageNet.
- **Gap:** Requires large-scale pretraining datasets.

Key Paper: DeiT by Touvron et al.

- Proposed data-efficient training strategies for ViTs.
- Leveraged knowledge distillation from CNNs.
- **Gap:** Still computationally intensive despite improvements.

Key Paper: DETR by Carion et al.

- Introduced an end-to-end object detection model using Transformers.
- Eliminated the need for hand-crafted anchor boxes.
- **Gap:** Requires longer training times compared to traditional detectors.

Key Paper: EfficientDet by Tan et al.

- Integrated efficient attention mechanisms into object detection.
- Achieved state-of-the-art performance with reduced computational cost.
- **Gap:** Limited by the complexity of attention modules for very high-resolution images.

Key Paper: SETR by Zheng et al.

- Applied Transformers to semantic segmentation tasks.
- Demonstrated competitive performance on standard benchmarks.
- **Gap:** High computational overhead and memory usage.

Key Paper: CCNet by Huang et al.

- Proposed a criss-cross attention mechanism.
- Improved efficiency in capturing contextual information.
- **Gap:** Struggles with very large and diverse datasets.

Key Paper: TimeSformer by Bertasius et al.

- Introduced a Transformer-based model for video classification.
- Captured spatiotemporal dependencies effectively.
- **Gap:** Requires extensive computational resources for training.

Key Paper: ViViT by Arnab et al.

- Developed a video vision Transformer for video action recognition.
- Achieved state-of-the-art results on multiple benchmarks.
- **Gap:** High memory consumption during training.

Discussion

Discussion

- Impact of attention mechanisms on the performance and efficiency of vision models.
- Key advantages: flexibility, scalability, and ability to capture long-range dependencies.
- Challenges: computational complexity, data efficiency, and interpretability.
- Promising directions: efficient attention mechanisms, hybrid architectures, and explainable AI.

Conclusion

Conclusion

- Summary of key findings on attention mechanisms in computer vision.
- Importance of addressing current challenges for further advancements.
- Future research directions: improving data and computational efficiency, enhancing interpretability, and exploring new applications.

Thank you!