Addis Ababa Institute of Technology

ATTENTION MECHANISMS AND THEIR IMPACT ON COMPUTER VISION MODELS

Mohammed Sraj (GSR/5161/16)

20/06/2024

ATTENTION MECHANISMS AND THEIR IMPACT ON COMPUTER VISION MODELS

Mohammed Sraj GSR/5161/16
Addis Ababa Institute of Technology

ABSTRACT

The rapid advancements in transformer-based architectures, which have demonstrated remarkable success in natural language processing, have inspired their exploration and application in the field of computer vision. This collection of papers showcases the versatility of transformers in addressing a variety of challenging visual tasks, ranging from image classification and video understanding to object detection.

At the core of these efforts is the idea of treating visual data, such as images and videos, as sequences of tokens or patches, which can then be effectively processed using transformer-based models. This paradigm shift from the traditionally dominant convolutional neural network (CNN) approaches has enabled transformers to capture long-range dependencies and global relationships within visual data, leading to state-of-the-art performance on numerous benchmarks.

The research papers in this collection present novel transformer-based architectures tailored for specific visual tasks, such as pure transformer models for image classification and the adaptation of transformers to the video domain for action recognition. Furthermore, the research also explores techniques to improve the data-efficiency of transformer-based vision models, which is crucial for their practical deployment, especially in scenarios with limited training data.

Collectively, these works demonstrate the remarkable potential of transformers in computer vision, showcasing their ability to outperform or complement conventional CNN-based approaches across a diverse range of applications. The insights and advancements presented in this collection pave the way for the widespread adoption of transformer architectures and their continued evolution in the field of visual understanding.

1 INTRODUCTION

Over the past decade, convolutional neural networks (CNNs) have been the dominant force in computer vision, achieving remarkable breakthroughs across a wide range of tasks. However, the research community has recently witnessed a significant shift, with the rising prominence of Transformer architectures, which have historically been the driving force behind transformative advancements in natural language processing (NLP).

The papers examined in this literature review illustrate the ongoing exploration and successful application of Transformer-based models to various computer vision problems, challenging the long-standing supremacy of CNNs. Researchers are investigating whether pure attention-based approaches can outperform CNN-based methods, particularly in the realm of video understanding tasks such as action recognition.

Beyond video understanding, the literature also delves into the use of Transformer architectures for fundamental computer vision tasks, such as semantic segmentation. Interestingly, some studies have reframed these tasks as sequence-to-sequence problems, rather than adhering to the traditional dense prediction paradigm, highlighting the versatility of Transformer models.

Innovations in attention mechanisms, such as the introduction of Criss-Cross Attention, have also been explored, demonstrating significant improvements over previous CNN-centric approaches to semantic segmentation. Furthermore, the literature examines hybrid architectures that combine the strengths of CNN backbones with Transformer-based detection heads for object detection, showcasing the potential for synergistic integration of these distinct paradigms.

A crucial aspect of the reviewed literature is the emphasis on improving the data efficiency of Transformer-based models for image recognition tasks. Achieving strong performance with limited training data is a key challenge, and the research community has made notable strides in addressing this concern, paving the way for more widespread adoption of Transformer-based vision models. Studies have even shown that a standard Transformer architecture, when pre-trained on large-scale datasets, can be applied directly to image classification tasks, outperforming CNN-based approaches. This breakthrough underscores the significant potential of Transformer models to reshape the landscape of computer vision research and applications.

Overall, the papers examined in this literature review illustrate the growing prominence and success of Transformer-based architectures in the computer vision domain, challenging the long-held dominance of CNNs and opening up new avenues for innovative research and practical applications.

2 Literature Review

2.1 Image Classification

Recent advancements in Transformer architectures have shown promising results in computer vision tasks, moving beyond the traditional convolutional neural network (CNN) approach. These Transformer-based models have demonstrated the potential to outperform state-of-the-art CNN models, especially when trained on large-scale datasets.

One of the key innovations is the Vision Transformer (ViT) model [1], which takes a fundamentally different approach to processing images compared to CNNs. ViT divides an image into fixed-size patches, linearly embeds each patch, and then feeds the resulting sequence of tokens into a standard Transformer encoder. By leveraging the self-attention mechanism of Transformers, ViT is able to capture long-range dependencies in the image data, rather than relying on the local receptive fields and translation equivariance properties of CNNs. The authors show that when ViT is pre-trained on large datasets like ImageNet-21k or JFT-300M, it can achieve excellent performance on a variety of image recognition benchmarks, outperforming state-of-the-art ConvNet models while being more computationally efficient.

Building on the success of ViT, researchers have explored ways to improve the data efficiency of Transformer-based models for image classification [2]. Techniques such as multi-scale vision Transformers, distillation through attention, and task-adaptive prediction heads have enabled these models to achieve strong performance on ImageNet and other benchmarks, even when trained on relatively small datasets compared to the original ViT work. These advancements suggest that Transformer architectures can be made more accessible and applicable to a wider range of computer vision problems.

2.2 Object Detection

Beyond image classification, Transformer models have also shown promise in object detection tasks. The DEtection TRansformer (DETR) [3] formulates object detection as a direct set prediction problem, eliminating the need for many components of traditional object detectors, such as anchor generation, non-maximum suppression, and complex backbones. DETR uses a Transformer encoder-decoder to directly predict a set of detected objects, achieving competitive performance on standard benchmarks like COCO. This demonstrates the potential of Transformer models for various computer vision tasks beyond just image classification

2.3 SEMANTIC SEGMENTATION

Semantic segmentation, the task of assigning semantic class labels to each pixel in an image, is a fundamental problem in computer vision with wide-ranging applications. Over the years, researchers have proposed numerous approaches to address this challenge, leveraging the advancements in deep learning.

One prominent line of work has focused on utilizing fully convolutional network (FCN) architectures for semantic segmentation. FCNs process the input image in a patchwise fashion, generating pixel-wise predictions. While effective, FCNs can be limited in their ability to capture global contextual information, which is crucial for accurate semantic segmentation.

To overcome this limitation, the work by Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transform [4] proposes to reframe semantic segmentation as a sequence-to-sequence problem. The authors treat the image as a sequence of patches and apply transformer-based models, which have shown great success in natural language processing tasks, to capture the global dependencies between patches. By modeling these global contextual cues, the proposed approach can achieve state-of-the-art performance on benchmark datasets.

Another line of research has focused on developing efficient and effective attention mechanisms to aggregate contextual information. The work titled CCNet: Criss-Cross Attention for Semantic Segmentation [5] introduces the Criss-Cross Network (CCNet), which utilizes a novel criss-cross attention module. This module harvests contextual information from all pixels on the criss-cross path of a given pixel, using a recurrent operation to capture full-image dependencies. Importantly, the criss-cross attention mechanism is much more GPU memory-efficient and computationally lighter than previous attention-based approaches, such as the non-local block.

Another relevant work is the Pyramid Attention Network (PAN) [6], which proposes a pyramid attention module to capture multi-scale contextual information. PAN constructs a pyramid of features with different receptive fields and uses attention mechanisms to selectively aggregate relevant contextual cues at each scale. This allows the model to effectively leverage both local and global context for improved semantic segmentation performance.

Overall, these works highlight the critical importance of contextual information for semantic segmentation and demonstrate different approaches to modeling and leveraging such information. The sequence-to-sequence perspective, the criss-cross attention mechanism, and the pyramid attention module represent important advancements in the ongoing efforts to push the boundaries of pixel-wise visual understanding.

2.4 VIDEO UNDERSTANDING

The increasing success of attention-based Transformer models in natural language processing has inspired researchers to explore their application in computer vision tasks, including video understanding. Bertasius and Wang's work investigates the potential of using space-time attention mechanisms for video understanding [7]. The authors argue that attention mechanisms are well-suited for modeling the long-range dependencies that are a key challenge for convolutional neural network (CNN) architectures in video processing.

Building on this motivation, Arnab et al. present ViViT, a pure Transformer-based model for video classification tasks [8]. To handle the large number of spatio-temporal tokens encountered in video data, the authors develop several efficient variants of their Transformer-based architecture that factorize the spatial and temporal dimensions. Additionally, they demonstrate effective regularization techniques and the leveraging of pre-trained image models to enable training on relatively small video datasets, overcoming a known challenge with Transformer-based models. Through thorough ablation studies, the authors identify the best design choices for their Transformer-based architecture. The proposed ViViT model achieves state-of-the-art results on multiple standard video classification benchmarks, outperforming prior methods based on 3D convolutional networks.

This body of work provides a comprehensive investigation of adapting Transformer-based models for video understanding, demonstrating their strong potential to surpass conventional 3D CNN approaches in this domain. The factorization strategies employed by Arnab et al. to handle the long sequences of spatio-temporal tokens in video data, as well as the effective regularization and

pre-training techniques, represent important advancements in applying Transformer architectures to video understanding tasks. The state-of-the-art results achieved by ViViT on various benchmarks further validate the effectiveness of Transformer-based models for video classification, opening up new directions for future research in this area.

3 CONCLUSION

The papers reviewed in this literature showcase the impressive progress of Transformer-based architectures in the field of computer vision. The studies highlight the versatility of Transformers, which have traditionally been associated with natural language processing (NLP), and their ability to tackle a diverse array of computer vision tasks.

A key emphasis of the literature is the exploration of pure attention-based Transformer models for video understanding, particularly in the context of action recognition. The paper "Is Space-Time Attention All You Need for Video Understanding?" investigates whether Transformer-based approaches can outperform CNN-based methods, providing valuable insights into the comparative strengths and limitations of these two paradigms. The successful application of Transformer-based models, as demonstrated in the "ViViT: A Video Vision Transformer" paper, suggests that attention-centric architectures hold significant promise for advancing the field of video understanding.

Beyond video-centric tasks, the literature also explores the use of Transformers for fundamental computer vision problems, such as semantic segmentation. The paper "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers" presents a novel formulation of these tasks as sequence-to-sequence problems, highlighting the versatility of Transformer architectures and their ability to adapt to various problem formulations. The introduction of innovative attention mechanisms, like the "Criss-Cross Attention" presented in the corresponding paper, showcases the ongoing efforts to enhance the performance of Transformer-based models for specific vision tasks, such as semantic segmentation. These advancements underscore the research community's commitment to unlocking the full potential of Transformers in computer vision.

Gaps: While the reviewed papers demonstrate the impressive capabilities of Transformer-based models in computer vision, there are several gaps that warrant further investigation:

- Comparative performance evaluation: The literature primarily focuses on the performance of Transformers in specific tasks, but a more comprehensive comparison of Transformer and CNN-based models across a wider range of computer vision applications, particularly in real-world, high-stakes scenarios, is needed to fully understand their relative strengths and limitations.
- Interpret-ability and explain-ability: The inner workings of Transformer models can be opaque, making it challenging to understand the underlying mechanisms and inductive biases that enable their success in certain domains. Exploring methods to improve the interpret-ability and explain-ability of Transformer-based vision models could enhance their trustworthiness and facilitate further advancements.
- Data efficiency: As highlighted in the Training data-efficient image transformers" paper, improving the data efficiency of Transformer-based models is crucial for their widespread adoption and practical deployment, especially in resource-constrained settings. Addressing this challenge represents an important area for future research.
- Integrating Transformers with other paradigms: The literature suggests that the future of computer vision may involve the strategic combination of Transformer-based approaches with other paradigms, such as CNNs, as demonstrated in the "EfficientDet: Scalable and Efficient Object Detection" paper. Exploring novel hybrid architectures could lead to further performance gains and synergies.

By addressing these gaps, the research community can unlock the full potential of Transformer-based models and drive the continued advancement of computer vision, leveraging the unique strengths of this emerging paradigm.

4 REFERENCES

1. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.
2. H. Touvron et al., "Training data-efficient image transformers and distillation through attention," ICML, 2021.
3. N. Carion et al., "End-to-End Object Detection with Transformers," ECCV, 2020.
4. S. Zheng et al., "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers," CVPR, 2021
5. Z. Huang et al., "CCNet: Criss-Cross Attention for Semantic Segmentation," ICCV, 2019
6. H. Li et al., "Pyramid Attention Network for Semantic Segmentation," AAAI, 2019
7. G. Bertasius and J. Wang, "Learning Temporal Attention in Dilated Convolutional Networks for Action Recognition," CVPR, 2022
8. A. Rnab et al., "ViViT: A Video Vision Transformer," ICCV, 2021
9. A. Vaswani et al., "Attention is All You Need," NeurIPS, 2017