Transformers in Vision: A Survey on Vision Transformers and Their Applications

By: Murad Tadesse - ID: GSR/9304/16

Addis Ababa Institute of Technology (AAiT)

21 June 2024

Introduction

- **Transformers in Vision:** Emerged as a powerful alternative to CNNs for various vision tasks.
- **Applications:** Image classification, object detection, segmentation, and generation.
- **Goals of the Survey:**
  1. Review the advancements in Vision Transformers (ViTs)
  2. Discuss architectural components and training techniques
  3. Compare ViTs with traditional CNNs
  4. Explore ViT applications
  5. Identify future research directions and challenges

Introduction (cont...)

Scope and Research Questions:

- What are the key architectural innovations in Vision Transformers?
- How do ViTs compare to CNNs in terms of performance and scalability?
- What are the primary applications of ViTs in computer vision?
- What challenges and opportunities exist for future research in this area?

Background

- **NLP Success:** The success of Transformers in NLP tasks led to exploring their applications in computer vision.
- **ViT Introduction:** Dosovitskiy et al. (2020) introduced the Vision Transformer (ViT) as a novel approach for image recognition tasks.
- **Key Features:**
  - Utilizes image patches as input tokens
  - Applies a standard Transformer encoder directly to sequences of image patches
  - Achieves competitive performance with state-of-the-art CNN models

Reference

Dosovitskiy, A., et al. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv preprint arXiv:2010.11929.

Vision Transformer (ViT)

- **Summary:** ViT marked a significant advancement in applying Transformer architecture to vision tasks.
- **Innovations:**
  - Self-attention mechanisms capture long-range dependencies in the input image.
  - Flexibility in handling variable-size inputs.
  - Scalability to higher resolutions and larger datasets.
- **Limitations:**
  - Requires large-scale pretraining data to achieve optimal performance.

Key Paper

Dosovitskiy, A., et al. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale."

Data-efficient Image Transformers (DeiT)

- **Summary:** DeiT addresses the data inefficiency of the original ViT model.
- **Innovations:**
  - Introduces a distillation token that learns from a CNN teacher model.
  - Achieves strong performance even with smaller datasets.
- **Limitations:**
  - Computationally intensive training process.

Key Paper

Touvron, H., et al. (2021). "Training data-efficient image transformers & distillation through attention." In Proceedings of the International Conference on Machine Learning (ICML).

Hierarchical Vision Transformers

- **Swin Transformer:** Introduces a hierarchical architecture with shifted windows.
- **Innovations:**
  - Efficient computation of local and global features.
  - Better handling of high-resolution images.
- **Limitations:**
  - High computational resource requirements.

Key Paper

Liu, Z., et al. (2021). "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows." In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Tokens-to-Token ViT (T2T-ViT)

- **Summary:** Utilizes a progressive tokenization process.
- **Innovations:**
  - Combines convolutional operations with Transformers.
  - Enhances feature extraction and representation.
- **Limitations:**
  - Complexity can impact scalability.

Key Paper

Yuan, L., et al. (2021). "Tokens-to-token vit: Training vision transformers from scratch on imagenet." In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Hybrid Architectures

- **Twins Model:**
  - Integrates a spatial transformer module and a depth-wise convolutional module.
  - Captures both long-range and local dependencies.
- **CeiT:**
  - Leverages convolutional inductive biases.
  - Balances between Transformers and CNNs for optimal performance.

Key Papers

Chu, X., et al. (2021). "Twins: Revisiting the Design of Spatial Attention in Vision Transformers." arXiv preprint arXiv:2104.13840.
Yuan, L., et al. (2021). "CeiT: Convolution-Enhanced Image Transformers." arXiv preprint arXiv:2103.11816.

Applications of Vision Transformers

- **Object Detection:**
  - DETR model (Dai et al., 2021) provides end-to-end object detection.
  - Simplifies the detection pipeline by predicting bounding boxes and class labels directly.
- **Semantic Segmentation:**
  - Segmenter model (Xie et al., 2021) directly applies transformers for segmentation tasks.
  - Achieves competitive performance without a CNN backbone.
- **Image Generation:**
  - VideoGPT model (Esser et al., 2021) uses hierarchical transformers for high-resolution video synthesis.

Key Papers

Dai, Z., et al. (2021). "UP-DETR: Unsupervised Pre-training for Object Detection with Transformers." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Limitations and Challenges

- **Data Efficiency:**
  - Vision Transformers require large-scale pretraining datasets for optimal performance.
  - Addressing this with data-efficient training strategies is crucial.
- **Computational Complexity:**
  - The self-attention mechanism is resource-intensive, especially for high-resolution images.
  - Innovations in efficient attention mechanisms are needed.
- **Inductive Biases:**
  - Lack of built-in biases like translation equivariance in CNNs.
  - Hybrid architectures can help balance biases.
- **Interpretability:**
  - Inner workings of Transformers are less intuitive compared to CNNs.
  - Developing interpretability techniques is essential.

Future Research Directions

- **Improving Data Efficiency:**
  - Explore self-supervised learning and transfer learning strategies.
  - Develop architectural refinements to reduce data requirements.
- **Enhancing Computational Efficiency:**
  - Investigate novel attention mechanisms and hardware-aware optimizations.
- **Addressing Interpretability:**
  - Advance attention visualization and feature attribution methods.
- **Exploring Hybrid Architectures:**
  - Combine strengths of Transformers and CNNs for optimal performance.
- **Expanding Application Scope:**
  - Apply Vision Transformers to novel areas such as multi-modal learning, video understanding, and 3D vision.

Conclusion

- **Summary of Key Points:**
  - Vision Transformers represent a significant advancement in computer vision.
  - They offer versatility and strong performance across various tasks.
- **Future Outlook:**
  - Promising research directions include improving data and computational efficiency, enhancing interpretability, and expanding applications.
  - Vision Transformers are poised to play a pivotal role in the future of computer vision and beyond.

Thank You!