

Comparative Analysis of YOLO11, Vision Transformer (ViT), EfficientNet-V2, and Swin Transformer

Surya Ashok¹

¹*Department of Computer Science, College of Applied Science Chelakkara*

Abstract - Computer vision has emerged as a major research area in artificial intelligence because of its ability to automate visual understanding tasks. Recent deep learning models have improved the performance of image classification, object detection, and recognition systems. This paper presents a comparative study of four modern computer vision architectures: YOLO11, Vision Transformer (ViT), EfficientNet-V2, and Swin Transformer. The comparison focuses on architecture structure, processing speed, computational requirements, scalability, and practical applications. A Python OpenCV implementation is also included to demonstrate a real-time comparison system. The study highlights the advantages and limitations of each model and identifies their suitability for different computer vision applications.

Key Words: Computer Vision YOLO11 Vision Transformer EfficientNet-V2 OpenCV

1. INTRODUCTION

Computer Vision is an important field of Artificial Intelligence that enables computers to analyze, process, and understand images and videos similar to human vision. The development of deep learning techniques has significantly improved the performance of computer vision systems in tasks such as image classification, object detection, facial recognition, segmentation, and video analysis. Traditional computer vision systems relied heavily on manually designed feature extraction techniques. However, the introduction of Convolutional Neural Networks (CNNs) transformed the field by enabling automatic feature learning from large datasets. CNN-based architectures became widely used because of their strong spatial feature extraction capability and efficient image processing performance.

Recently, Transformer-based architectures have gained significant attention in computer vision research. Vision Transformer (ViT) introduced self-attention mechanisms to process image patches and capture global contextual relationships within images. Swin Transformer further improved this approach by introducing hierarchical feature learning and shifted-window attention mechanisms, reducing computational complexity while improving scalability and performance.

Real-time object detection has become one of the most important applications of computer vision. Models such as YOLO11 are specifically designed for fast and accurate object detection in real-time environments. These architectures are widely used in autonomous vehicles, smart surveillance systems, robotics, traffic monitoring, healthcare, and industrial automation because of their low latency and high-speed inference capability.

This research paper presents a comparative analysis of four advanced computer vision architectures: YOLO11, Vision Transformer (ViT), EfficientNet-V2, and Swin Transformer. The study compares these models based on architecture design, speed, accuracy, computational requirements, scalability, and real-time performance. The objective of this research is to understand the advantages and limitations of each model and identify their suitability for different computer vision applications.

2. METHODOLOGY

The comparative study was conducted by analyzing the architecture design, performance characteristics, and real-time capabilities of the selected models. YOLO11 was evaluated as a real-time object detection architecture optimized for speed and detection accuracy. Vision Transformer was studied for its self-attention mechanism and ability to process global contextual information. EfficientNet-V2 was analyzed for parameter efficiency, training speed, and lightweight deployment capability. Swin Transformer was evaluated for its hierarchical architecture and shifted-window attention mechanism that improves scalability and computational efficiency.

The models were compared using factors such as inference speed, computational complexity, scalability, and practical applications.

2.1 YOLO11

YOLO11 is a modern real-time object detection architecture developed to improve detection speed and accuracy. It uses an optimized Convolutional Neural Network (CNN) structure for efficient feature extraction and multi-scale object detection. The model is capable of detecting objects in images and videos with very low latency, making it suitable for surveillance systems, autonomous vehicles, and robotics applications. YOLO11 supports real-time inference, improved small-object detection, and efficient deployment on edge devices. Its architecture focuses on balancing speed and accuracy while reducing computational overhead.

2.2 Vision Transformer (ViT)

Vision Transformer (ViT) is a Transformer-based computer vision architecture that processes images as sequences of patches instead of using traditional convolution operations. The model uses self-attention mechanisms to capture long-range dependencies and global contextual information from images. ViT demonstrates strong performance in image classification and large-scale visual recognition tasks. It provides excellent scalability and contextual understanding, especially when trained on large datasets. However, the model requires high computational resources and large amounts of training data for optimal performance.

2.3 EfficientNet-V2

EfficientNet-V2 is a CNN-based deep learning architecture designed for efficient training and reduced computational complexity. The model introduces compound scaling and fused-MBConv layers to improve both training speed and accuracy. EfficientNet-V2 achieves high performance using fewer parameters compared to many traditional CNN models. It is highly suitable for mobile devices, embedded systems, and lightweight image classification tasks. The architecture focuses on maintaining a balance between efficiency, accuracy, and hardware optimization.

2.4 Swin Transformer

Swin Transformer is a hierarchical Transformer architecture designed to improve scalability and computational efficiency in computer vision tasks. It uses shifted window attention mechanisms to process images efficiently while maintaining strong contextual understanding. Unlike standard Vision Transformers, Swin Transformer builds hierarchical feature maps similar to CNNs, making it suitable for object detection, segmentation, and image recognition tasks. The architecture reduces computational complexity while providing excellent performance across multiple vision applications.

3. COMPARISON ANALYSIS

The comparison of YOLO11, Vision Transformer (ViT), EfficientNet-V2, and Swin Transformer shows significant differences in architecture design, computational efficiency, speed, and application suitability. YOLO11 is primarily designed for real-time object detection and achieves very high inference speed with good detection accuracy. The architecture is optimized for low latency and efficient deployment in surveillance systems, autonomous vehicles, and robotics applications. Its lightweight detection pipeline makes it suitable for edge devices and real-time environments.

Vision Transformer (ViT) differs from traditional CNN-based models by using self-attention mechanisms to process image patches. The model provides strong contextual understanding and excellent classification accuracy on large-scale datasets. However, ViT requires high computational resources and large training datasets, making it less suitable for low-resource systems and real-time applications. Its architecture focuses more on global feature learning rather than computational efficiency.

EfficientNet-V2 is a CNN-based architecture developed to improve parameter efficiency and reduce training time. The model achieves high image classification accuracy with lower computational complexity compared to many deep learning models. EfficientNet-V2 is highly suitable for mobile devices, embedded systems, and lightweight applications because of its efficient resource utilization and faster training process.

Swin Transformer combines Transformer-based attention mechanisms with hierarchical feature extraction similar to CNN architectures. The shifted-window attention mechanism reduces computational complexity while maintaining strong contextual learning capability. Swin Transformer performs effectively in image classification,

segmentation, and object detection tasks. The architecture provides better scalability and balanced performance compared to standard Vision Transformers.

Overall, YOLO11 is most suitable for real-time detection tasks, EfficientNet-V2 is ideal for lightweight classification systems, while ViT and Swin Transformer are more effective for large-scale visual understanding and advanced computer vision applications. The comparison highlights that each architecture has unique strengths depending on the application requirements, computational resources, and performance objectives.

4. CONCLUSIONS

This paper compared four modern computer vision architectures: YOLO11, Vision Transformer (ViT), EfficientNet-V2, and Swin Transformer. The study showed that YOLO11 performs efficiently in real-time object detection, while EfficientNet-V2 provides lightweight and computationally efficient image classification. ViT and Swin Transformer demonstrated strong contextual understanding and scalability through attention mechanisms. The comparison highlights that CNN-based models are suitable for real-time and lightweight applications, whereas Transformer-based architectures are more effective for large-scale visual understanding tasks. Future computer vision systems are expected to combine the advantages of both CNN and Transformer architectures for improved performance and efficiency.

REFERENCES

1. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified Real-Time Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Press, Las Vegas (2016) 779–788
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations. Springer, Vienna (2021) 1–21
3. Tan, M., Le, Q.: EfficientNetV2: Smaller Models and Faster Training. In: Proceedings of the International Conference on Machine Learning. PMLR, Virtual Conference (2021) 10096–10106
4. Liu, Z., Lin, Y., Cao, Y., Hu, H.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: Proceedings of the IEEE International

- Conference on Computer Vision. IEEE Press, Montreal (2021) 9992–10002
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Press, Las Vegas (2016) 770–778
6. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. Int. Conf. Learn. Represent. 1 (2015) 1–14
7. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149
8. Howard, A., Sandler, M., Chu, G., Chen, L.: Searching for MobileNetV3. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE Press, Seoul (2019) 1314–1324
9. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet Classification with Deep Convolutional Neural Networks. Commun. ACM 60 (2017) 84–90
- Bochkovskiy, A., Wang, C., Liao, H.: YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv Preprint arXiv:2004.10934 (2020) 1–17