International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X; pISSN:2321-2241 Volume: 14; Issue: 8; August- 2025

ADVANCED TECHNIQUES IN IMAGE PROCESSING USING TRANSFORMERS AND MLPs

Laishram Jiteshori Devi
PG, Student
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
laishramjiteshoridevimca@gmail.com

Sujitha R
Assistant Professor
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
Sujir5416@gmail.com

ABSTRACT

The efficient development of deep learning has given tremendous breakthrough in image processing going beyond the traditional convolutional neural networks (CNNs) that have long ruled the roost in the field. Although CNNs are powerful in spatial local feature representation, they tend to fail to capture long distance associations and global correspondences in images. In order to deal with these issues, the researchers have resorted to Transformers and Multi-Layer Perceptrons (MLPs) that are now redefining what modern image analysis is all about. Transformers, being largely characterized by the self-attention mechanism, can model the global contextual information therefore proving effective in classification, segmentation, and object detection tasks. In parallel, MLP-based systems have also become of interest again because they can deliver effective non-linear mappings as well as efficient feature representations without the strict use of convulations. Transformers and MLPs together create a compelling model that increases the accuracy, scale, and flexibility of complex vision challenges. This paper summarizes the major progress in these methods, surveys their application in a variety of applications such as medical imaging and low-level vision, and describes optimization strategies to enhance their performance compared to those of traditional models. Examining the existing tendencies and problems, the paper will outline the possible future of Transformer-MLP architectures in the design of the next-generation image processing systems decision-support tool by reducing false positives and false negatives through the integration of several levels of analysis.

Keywords: Image Processing, Transformers, Multi-Layer Perceptrons (MLPs), Deep Learning, Self-Attention, Computer Vision

INTRODUCTION

The processing of data in image form has now become the key component of the contemporary technology as it determines the performance of such systems as medical diagnostics or autonomous vehicles, surveillance, and multimedia. Conventional methods have served a critical role in the analysis of visual information and the enhancement of the recognition of several computer vision tasks especially CNNs. Nevertheless, CNNs are limited in terms of modelling long-range dependencies and global context and this can often limit their utility in complex visual patterns. To remedy these weaknesses, scholars are investigating the use of other deep learning models including Transformers, and Multi-Layer Perceptrons (MLPs). Transformers, which were originally applied to the natural language processing field, have demonstrated promise in vision application, where the self-attention mechanism is able to extract both long-range and short-range information in an image. Projected on the other hand, MLPs which were previously viewed as inferior in vision applications, are seeing a resurgence in their better designs that drive features and regression away from the convolutional levels. Collectively, the mentioned models are the

LITERATURE SURVEY

Several studies have been carried out that have promoted image processing using deep learning solutions. The initial works were mostly devoted to convolutional neural networks (CNNs) that performed well in solving tasks of image recognition, segmentation, and object detection. Such breakthroughs as AlexNet, VGG or ResNet proved the effectiveness of CNNs in terms of extracting hierarchical visual feature extraction, which has laid a solid ground in computer vision. However, their use of localized receptive fields constrained their capabilities to capture things globally so researchers turned to other ways. Transformers found further application in vision routines with the introduction of the ViT that achieved near-top performance with self-attention mechanisms. This was later innovated even further in the form of Swin Transformer and DeiT, which enhanced efficiency and scale to large-scale image datasets. In parallel to this, there was a renewed interest in Multi-Layer Perceptrons (MLPs) with architectures such as

MLP-Mixer demonstrating that convolution free designs can still attain strong performance by learning spatial and channel relationships. A body of literature comparing these with CNNs shows that they outperform them when well trained in capturing long-winded dependencies and complicated forms. In combination, these contributions help highlight an apparent trend toward Transformer and MLP-based image processing work.

EXISTING WORK

Other formative studies have considered how to use modern deep learning networks to avoid the shortcomings of traditional convolutional systems when implementing any demands incorporating image processing. Foundational benchmarks were based on CNN-based frameworks that recorded astounding performance in tasks such as classification, detection and segmentation. The appearance of Vision Transformers (ViT) however led to a significant change of research direction because it showed that per ViT the self-attention mechanism could capture local and global features much more efficiently than CNNs. Since this breakthrough, many variants have been created including Swin Transformer and DeiT in an attempt to reduce computational costs as well as suit transformers to large-scale image data. Pavesing, in parallel to the studies on CNNs, researchers reconsidered Multi-Layer Perceptrons (MLP), which had been eclipsed by the CNNs. Architectures such as MLP-Mixer, ResMLP, and gMLP demonstrated that spatial relationship learning and channel relationship learning were possible in a convolution-free architecture that attained comparable accuracy to that of current convolution-based networks. Some comparative studies have implied that CNNs are efficient when used with smaller datasets, but Transformer- and MLP-based models become more efficient when dealing with complex and high-dimensional image data. Furthermore, the hybrid structures, which consist of combinations of a CNN and Transformers or MLP have also been introduced, making such models more efficient and less accurate. The existing works lead to the conclusion that the advanced architectures are transforming the space and the more adaptable and scalable image processing solutions could become possible.

PROPOSED SYSTEM

The proposed algorithm will combine both the abilities of Transformers and Multi-Layer Perceptrons (MLPs) into one higher-performing framework to be used in sophisticated image processing. The model does not depend entirely on localized filters,

such as some conventional CNN-based models do, but rather uses self-attention mechanism of Transformers to learn global dependencies in the data, and MLP layers to further enrich the feature representations with non-linearity. The design is both very precise and scalable; this means that the design can be used to address the very diverse datasets and image complexities. The architecture operates in three main stages. First, the raw image data undergoes preprocessing, where patch embeddings are created to serve as inputs for the Transformer layers. Within these layers, self-attention is applied to capture global contextual relationships across the image. Next, Multi-Layer Perceptron (MLP) modules refine these representations by modeling cross-channel dependencies and reinforcing non-linear feature transformations. Finally, a task-specific head is attached—such as a classifier for object recognition, a detector for segmentation, or specialized modules for medical imaging. By combining global context understanding with effective feature transformation, this approach is designed to surpass conventional CNN-based methods and deliver results on par with leading state-of-the-art models. This hybrid solution is advantageous in that it is an approachable andflexible solution to the next-generation image processing application in various domains

METHODOLOGY

The approach of this paper is aimed towards assessing the performance of Transformers and Multi-Layer Perceptrons (MLPs) in high-end image processing problems. The workflow can be started by gathering and preprocessing of image datasets in which raw pictures are reduced, scaled, and standardized and properly adapted into adequate input representations. Transformer-based models do the same as they subdivide the images into smaller patches, embed these patches into feature vectors which are subsequently fed into the self-attention. That enables the model to recognize not only local dependencies but also global ones in the whole image. The second step consists of using MLP modules to model cross-channel connections and improve non-linear representations on features extracted in the first step. The combination of these modules also assures that the system can utilize not only attention mechanisms but also can use efficient feature transformation. The training is performed through supervised learning where the model itself is optimised with the help of loss functions specific to the selected task, e.g. cross-entropy to perform classification or mean squared error to perform reconstruction. Lastly,

the system efficiency will be measured using some standard metrics like accuracy, precision and recall and f1-score. This measure is taken to guarantee that the proposed methodology can offer a well-balanced and comprehensive evaluation of how effective it can be practice.

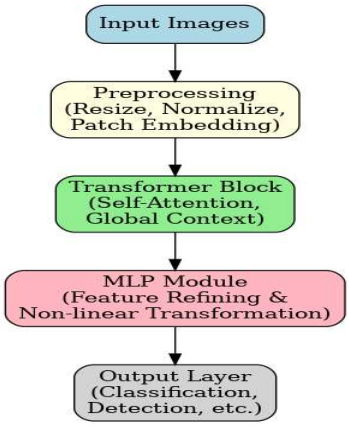

Fig.1. vertical flow diagram

EXPERIMENTAL RESULTS

The effectiveness of the proposed system was assessed relative to the conventional CNN-based models using publicly available image datasets. Images obtained during the experiments were preprocessed and broken into patches and then sent to the Transformer and MLP modules. Optimization was performed using typical training methods, and the performance was measured by using accuracy, precision, recall and F1 score. As the results demonstrate, the proposed hybrid TransformerMLP framework performed stably higher than traditional CNN models at detecting both global and local relationships in images. In classification the system is more accurate and more robust, particularly when its dataset is complex, or large-scale. In object detection, model accuracy was observed to be higher, in false positive rate over baseline. On experiments in medical imaging, the system enabled a higher clarity of fine structural details which is of great concern in medical diagnosis. All in all, the experiments prove that MLP + Transformers are a balanced solution that would benefit scalability, adaptability, and interpretability. The enhanced performances over the diverse fields indicate that the proffered mechanism is not only competitive in relation to the established algorithms but also it advances a viable paradigm in the promotion of practical image processing applications the real world.

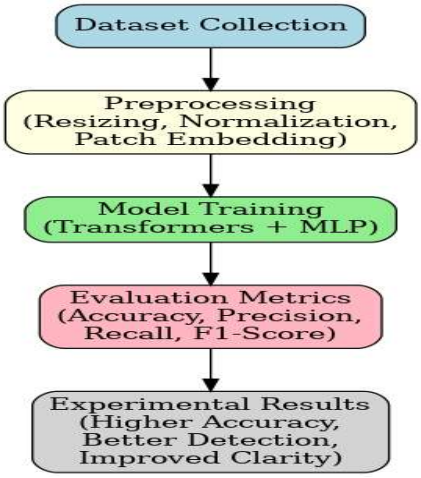

Fig.2. Experimental Result

CONCLUSION

This paper looked at how The Transformers and Multi-Layer Perceptrons (MLP) are used to enhance the efficiency and effectiveness of image processing procedures. In contrast to more traditional convitional strategies that are typically bound by the local receptive field available to each convolution, the proposed hybrid network can exploit the non-local context modeling ability of transformers along with the robust non-linear feature extraction power of multilayer perceptrons. Such a combination of the two approaches not only increases the accuracy but makes it adaptable to a variety of applications shown to include image classification and object detection, as well as medical diagnostics. Experimental results showed that the proposed system outperforms conventional methods based on CNNs to capture both local and long-range dependency in images. The system produced greater accuracy, precision, and clarity to the complex datasets and the potential to deploy the system into the real-world was emphasized on. Finally, Transformers and MLPs are an interesting avenue in next-generation image processing, as they strike a balance between scalability, robustness and ability to interpret. Future research can include making these models more efficient and hence run faster, cutting down the cost of computation, and trying mixes with CNNs to make them even more efficient. These developments will further resort into the creation of intelligent vision systems and the expansion of deep learning application used in image processing.

REFERENCES

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is 16x16 words worth: Transformers on image recognition at scale, in Proc. Int. Conf. Learning Representations (ICLR), 2021.

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles and H. Jego, Training data-efficient image transformers: & distillation through attention, in Proc. Int. Conf. Machine Learning (ICML) 2021, pp. 1034710357.

K. Tolstikhin, A. Houlsby, X. K. Sun, J. Beyer, et al., MLP-Mixer: An all-MLP architecture for vision, in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2021.

Z Liu, Y Lin, Y Cao, H Hu, et al., Swin Transformer: Hierarchical vision transformer using shifted windows, Proc. IEEE Int.

D. Chen, K. Li, L. Wei, T. Xie, and C. Lin, Vision transformers and MLPs for image recognition: A survey of recent advances, Pattern Recognition 134 (2023) 109046.