Deformable patch-based-multi-layer perceptron mixer model for forest fire aerial image classification

Mittal, Payal; Sharma, Akashdeep; Singh, Raman

Published in:
Journal of Applied Remote Sensing

DOI:
10.1117/1.JRS.17.022203

Published: 07/11/2022

Document Version
Peer reviewed version

Mittal, P., Sharma, A., & Singh, R. (2022). Deformable patch-based-multi-layer perceptron mixer model for forest fire aerial image classification. *Journal of Applied Remote Sensing*, *17*(2), [022203]. https://doi.org/10.1117/1.JRS.17.022203

Deformable Patch based MLP Mixer Model for Forest Fire Aerial Image Classification

Payal Mittal,[a*] Akashdeep Sharma,[a] Raman Singh,[b]
[a]University Institute of Engineering and Technology, CSE Department, Sector 25, Chandigarh, India, 160014
[b]University of the West of Scotland, Cyber Security Department, Lanarkshire, United Kingdom, ML11 7AA

Abstract. Unmanned Aerial Vehicles (UAVs), equipped with mounting camera sensors, facilitate a wide domain of applications deployed in the real-time world. The situational awareness for applications such as search and rescue in case of wildfires, estimation of endangered flora and fauna and emergency responses have seen paradigm shift due to UAVs capability of accessing in remote and challenging areas such as forests. The last decade has seen tremendous growth in CNN based methods for object classification, detection and segmentation tasks. Recently emerged Attention-based Transformer models have been trying to achieve state-of-the art results in predicting images. This paper proposed a novel MLP-Mixer architecture for classification of burned piles in dense forests. MLP mixer architecture tries to eliminate the shortcomings of convolutions and attention by merging them to obtain good performance. The shallow learning of CNN layers and fixed-size patch embedding in transformers have been eliminated by introducing a new module of DePatch in proposed MLP mixer model which divides the input images in a deformable pattern to detect forest fires at an early stage. On the pile images dataset obtained by drones during a burning pile of debris in an Arizona pine forest, our suggested classification algorithm has been tested. The performance of the proposed model has been compared with transformer models.

Keywords: Aerial Image Classification; Convolutional Neural Networks; Transformers; MLP Mixer; Computer Vision; Deep Learning

*****First Author**,** E-mail: payalmittal6792@gmail.com

1. Introduction

The use of unmanned aerial vehicles (UAVs) as a remote sensing platform for a range of practical applications, such as traffic monitoring [1], wildfire detections [2], precision agriculture [3], and the processing of satellite data [4], has received a lot of attention recently. Due to its use in numerous civil and commercial applications, UAV growth has accelerated recently. The UAVs have characteristics of cost-effectiveness, high performance and low power consumption which make it possible to incorporate vision-based automatic applications [5]. Height of capturing, tilt camera angle, and light settings play a significant role in the information of aerial images. The quality of generated images is highly dependent on the altitude of the flying vehicle and the characteristics of the sensors used. Recent years have seen the emergence of high-performance deep learning-based categorization architectures. Applications for emergency response and catastrophe management can benefit from deep learning approaches to quickly retrieve vital information, improve response times in time-sensitive circumstances, and help in-the-loop decision-making processes. The challenges of densely located, occluded objects, noise and background clutter exist in aerial images [6]. There is a need for powerful classification algorithms for overcoming these challenges. CNNs became the mainstream for performing standard classification in computer vision. The effective deep learning classification architectures of VGG [7], GoogleNet [8], Inception [9], ResNet [10], DenseNet [11], Lite-HRNet [12], and EfficientNet [13] are based on architectural advances such as depth-wise and deformable convolutions. The CNNs continue to be the well-known foundational designs for computer vision applications, but

Transformer-like architectures have also shown great promise for unified modelling of vision and language. In addition to efficiently capturing long-range dependency within the sequence and extracting additional semantic data, freshly developed transformers outperform CNNs in this regard. In terms of research discoveries and logical knowledge, transformers in vision assist in bridging the gap between the NLP and computer vision communities. The transformers help to increase the receptive field deprived from the resolutions. The effect of receptive fields on

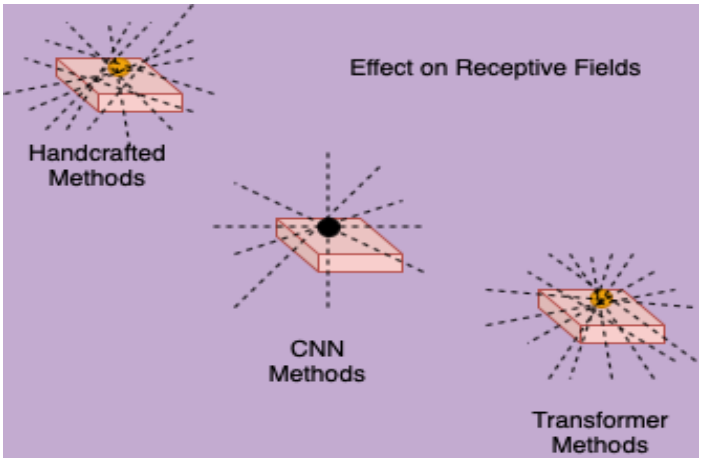


Fig. 1 Difference of receptive areas in handcrafted, CNN and Transformer based methods

handcrafted, CNN and transformers based methods is observed in Fig. 1. The concentration of the receptive field in case of handcrafted methods is more when compared with CNN methods but prior methods need manual engineering. But, the transformers based methods contribute significantly to detect and classify objects in an effective manner. Transformer-based classification models employ a fixed-size patch embedding with the implicit presumption that images are appropriate for the fixed image split design. However, such strong patch splits may cause semantic disagreement between images and issues with the collapsing of local structures in an image. In order to address the discussed shortcomings of CNNs and transformers, MLP Mixer architecture is utilized for the categorization of fire photographs taken by drones during a burning pile of trash in an Arizona pine forest [14]. Only straightforward matrix multiplication operations, changes to the data layout, and scalar non-linearities are used in Mixer's architecture. The use of multi-layer perceptrons is repeated across either feature channels or spatial locations. To the best of our knowledge, the classification model we propose is the first MLP mixer that executes patch splitting in a data-specific manner. In this study, we create the DePatch deformable patch, which adaptably changes the scale and position of each patch. The DePatch module is easy to create and can be used as a plug-and-play module. Our study is focused by proposing a novel deep learning based classification model for aerial images. The proposed deep model contains a deformable patch for MLP based mixer model for object classification. The proposed study aims to identify forest wildfires by deep feature representation approach. Early wildfire detection is crucial because it has the potential to cause serious harm to ecosystems, residential areas, forests, and wildlife habitats in the past. Recent developments in aerial surveillance systems in particular can give operational troops and first responders more precise information on the behaviour of fires for improved fire management. The proposed research work aims to contribute to a modeling shift in achieving strong performance on visual recognition tasks. The major offerings of this paper include:

a) Proposing a novel deformable patch based MLP mixer model for aerial image classification

b) Proposed deep learning-based classification model is an amalgamation of CNN and transformer based methods
c) Proposed classification model has been evaluated on the pile imagery dataset collected by drones.

The organization of the paper is as follows: Section 2 describes the related work of deep learning object classification models based on CNNs and transformers. Section 3 presents the details of the proposed deep learning-based object classification model in an elaborate manner. Section 4 discusses training experimental setup in which the proposed model requirements and evaluation parameters are being analyzed. Section 4 also represents analysis of results obtained from the proposed classification model. The last section concludes the results achieved with focus on future directions.

2 Related Work

The growth of object detection algorithms and classification architectures is shown as a timeline diagram in Fig. 2. On a year-by-year time period, the figure depicts the evolution of various hand-crafted, deep learning, and transformer-based algorithms. Scale Invariant Feature Transform (SIFT) [15], Histogram of Gradients (HOG) [16] , and SURF [17] were among the manual feature descriptors that dominated prior to 2012, but since then, CNN-based object detection methods have emerged. A breakthrough happened in handcrafted-based object detection and classification methods when deep learning-based convolutional architecture AlexNet [18] performed significantly when compared with former approaches. In 2014, a powerful detector region-based CNN method was developed through the combination of region proposals with CNNs. By classifying object proposals using a deep convolution network, it obtains outstanding object identification accuracy. After the proposal of the RCNN method, advancement in deep learning-based object detectors began. After 2015, a number of powerful architectures such as VGG [7], Xception [19], DenseNet [11], EfficientNet [13] for image classification have been developed.

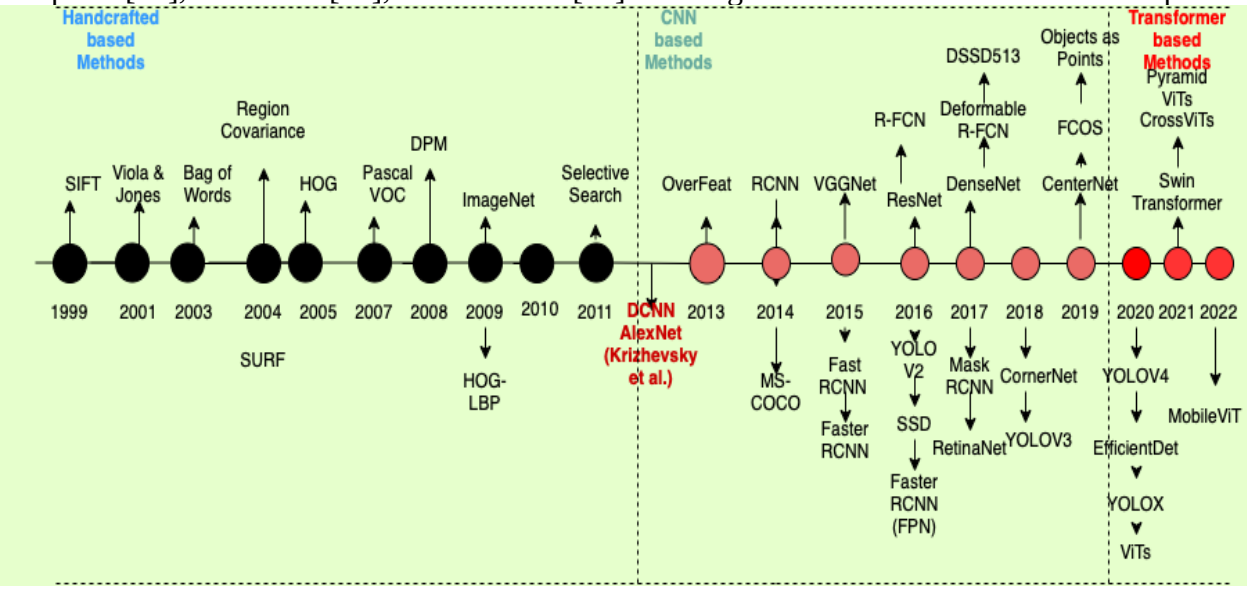


Fig. 2 Timeline of various object detection and classification algorithms

The history of computer vision demonstrates that the availability of larger datasets along with increased computational capacity frequently leads to a paradigm shift. Applications for computer vision are increasingly using CNNs and other deep learning methods. In comparison to CNN-based neural networks, the recently developed attention-based transformer models represent a

paradigm change for the middle of the 2020s [20]. When compared to CNNs, the recovered features from transformers can more accurately reflect long-range dependency within the sequence, and they also carry more semantic information.

In the next sections, we will be having a brief idea about self-attention based transformer based models. ViTs and Swin transformer based architectures help in improving the model interpretability as it relies on attention mechanisms which makes a prediction.

2.1 ViT based Classification Model

The ViT transformers for image classification were unveiled at the end of 2020. ViT carried on the time-consuming process of learning from unprocessed data while eradicating arbitrary visual features and inductive biases from models. The ViT transformer consists of three parts: a patch embedding module, multi-head self-attention blocks, and feed-forward multi-layer perceptrons. To develop features for image classification, the ViT divided an image into 16x16 patches and sent the image patch sequence through the transformer architecture. After the patch embedding module has turned the input image into a list of tokens, the network alternately stacks multi-head self-attention blocks and MLPs to get the final representation. The patch embedding module separates images into set sizes and positions before embedding a linear layer into each defined patch. With the use of substantial training data, ViT obtained outcomes comparable to those of conventional CNN designs. Despite great progress, most architectures still lost information since they divided the input image according to a preset pattern without taking the input's content or geometric variations into account [21].

2.2 Swin Transformer

Swin transformer functions as a general-purpose backbone for computer vision by converting traditional multi-head attention to shifted window attention based models. Swin Transformer's shift of the window partition between subsequent self-attention layers is a fundamental component of its design, and it greatly outperformed ViT and ResNeXt models with comparable latency on the tasks. Local multi-headed self-attention modules based on alternate shifting patch windows in succeeding blocks make up the Swin Transformer block. The Swin transformer design consists of a patch splitting module, similar to ViT, that divides an input RGB image into non-overlapping patches. Each patch is handled as a token, and its feature is configured as a concatenation of the RGB values of the individual pixels. The different transformer blocks with updated self-attention computation are applied to these patch tokens, which keep the token count, along with the linear embedding. The different Swin Transformer blocks are then applied to the patches in four stages, progressively reducing the number of patches to maintain hierarchical representation. As the network becomes deeper, patch merging layers reduces the number of tokens to provide a hierarchical representation. After concatenating the features of each set of two adjacent patches, the first patch combining layer applies a linear layer to the four-dimensional concatenated features. It can model information at various scales and has a linear computational complexity with respect to image size [22]. Swin Transformer obtained the best results on the MS-COCO dataset [23], although it utilizes a lot more parameters than convolutional models. Transformers offers a paradigm change away from CNN-based neural networks. Convolution may very likely be replaced by it in these tasks, even if it is still in the early stages of use in vision. The fixed scale of the tokens in the present transformer-based models precludes their usage in vision applications. Another distinction is that text passages have a significantly lower word resolution than pixels in graphics. Because the semantics of objects are destroyed by the Swin Transformers [22] and MobileViTs [24] models, which are still in development, they are still having trouble.

2.3 MLP Mixer

Transformers have the aforementioned drawbacks; thus, this paper incorporated the recently created MLP Mixer architecture, which comes from non-local blocks. Because it doesn't require self-attention or convolutional layers, the MLP-Mixer is a novel design in computer vision that varies from earlier, successful models like CNNs and transformers. The MLP-blocks in Mixer are designed to immediately process embeddings of these patches after converting images into a series of patches. The design of Mixer is influenced by more contemporary transformer-based systems. It utilizes conventional regularization and optimization methods, relies on token and channel-mixing MLPs, and is scalable to big data sets successfully. The competitive approach behind the MLP-Mixer architecture is built exclusively on multi-layer perceptrons rather than convolutions or self-attention. These perceptrons depend on simple matrix multiplication operations, adjustments to the data layout, and scalar non-linearities. They are repeatedly used across either feature channels or spatial locations. The MLP-Mixer architecture is based on multi-layer perceptrons and contains two different layers as described in Fig. 3. These two layers were made up of two MLPs: one that is applied individually to picture patches in order to mix location-specific features, and the other that is applied across patches in order to blend spatial data. Convolutions with small kernels of 11 are used in Mixer to transform convolutions into typical dense matrix multiplications (channel-mixing MLPs), which apply independently to each spatial point [25].

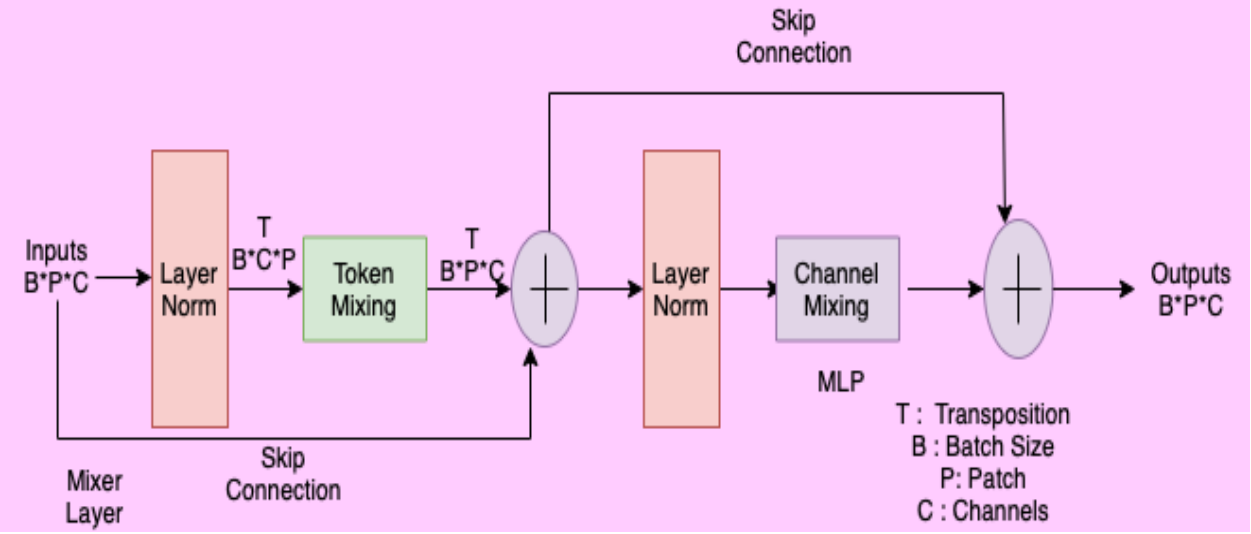


Fig.3 Mixer architecture containing token-mixing and channel-mixing MLP

As a result, geographic information cannot be aggregated; as a workaround, dense matrix multiplications, or token-mixing MLPs, are performed to each feature across all spatial locations. The competitive scores that the MLP mixer model achieved on benchmarks for image classification served as the basis for the proposed classification method. When trained on large datasets or using contemporary regularization techniques, these models have pre-training and inference costs that are comparable to those of cutting-edge models. The MLP-Mixer Architecture uses skip-connections and normalization layers, and each layer accepts an input of the same size. Matrix multiplications are then applied to the patches and features input table. As opposed to ViTs, Mixer uses a token-mixing layer to combine spatial information rather than position embeddings. MLPs are sensitive to the order of the input tokens and use a conventional classification head with a global average pooling layer, followed by a linear classifier.

3 De-Patch based MLP Mixer Architecture

The existing CNN and transformer models for object classification pose serious challenges to aerial images. The transformer-based methods utilized a fixed-size patch embedding which might destroy the semantics of aerial objects. Further, the hard patch splits of CNNs brought two problems related to collapse of local structures and having semantic inconsistency across aerial images. Because scale-variance items can be seen in a variety of aerial photographs, it is difficult to capture the entire object-related local structure in a 16 * 16 regular patch [26]. The same item may appear differently geometrically in many aerial photographs, depending on the scale, rotation, etc. The fixed method of picture splitting may capture contradictory information for the same object in various photographs. These updated patches run the risk of erasing semantic data, which would reduce classification accuracy. To address the aforementioned difficulties, we suggest DePatch, a novel module in the MLP mixer architecture that learns to adaptively partition the images into patches with varying positions and scales in a data-driven way as opposed to using predetermined fixed patches. The semantics in patches may be effectively preserved using our suggested classification architecture by integrating DePatch into MLP mixer design.

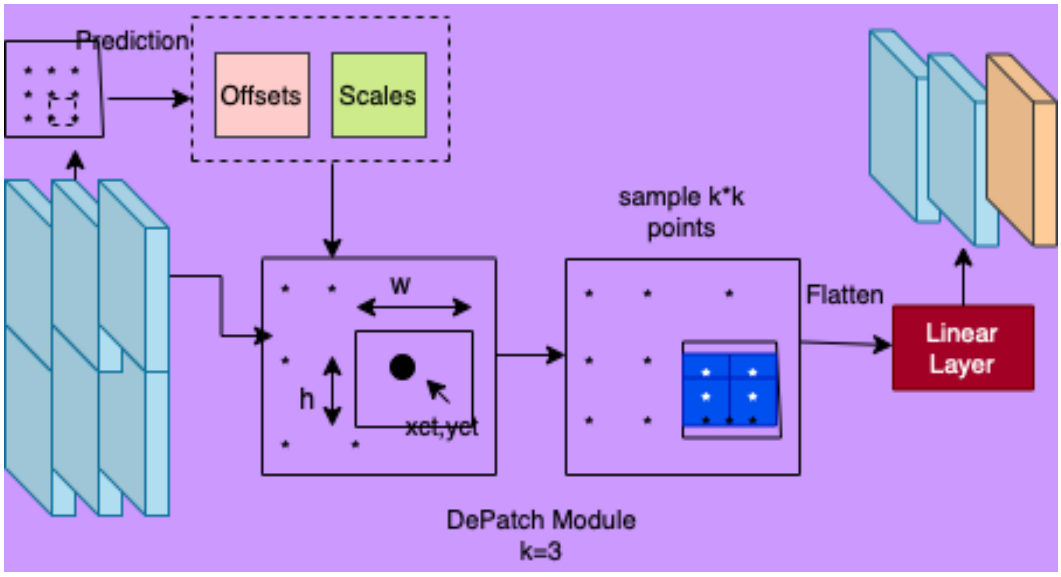


Fig. 4 DePatch module with offsets and scales within local features

The proposed DePatch based MLP Mixer architecture is shown in Fig. 4. The rectangle region is immutable for each patch, as illustrated in the figure, because the coordinates (x_{ct}, y_{ct}), and size s of the patch are fixed. The interior pixels of the patch being used directly depict its feature. We relax these requirements to build our deformable patch embedding module, DePatch, which can better locate key structures and handle geometric deformation. Based on the contents of the input, projected parameters include the position and size of each patch. We estimate an offset (x, y) that will allow the location to move away from the original center. In terms of scale, all we do is swap out the fixed patch size s for the predictable s_h and s_w. In this way, we can determine a new rectangle region, and denote its left-top corner as (x_1, y_1) and right- bottom corner as (x_2, y_2). We emphasize that (δx, δy, s_w, s_h) can be different even for patches in a single image as shown in eq. 1 and 2:

$$x_1 = x_{ct} + \delta x - \frac{S_w}{2}, y_1 = y_{ct} + \delta y - \frac{S_h}{2} \quad (1)$$

$$x_2 = x_{ct} + \delta x - \frac{S_w}{2}, y_2 = y_{ct} + \delta y - \frac{S_h}{2} \quad (2)$$

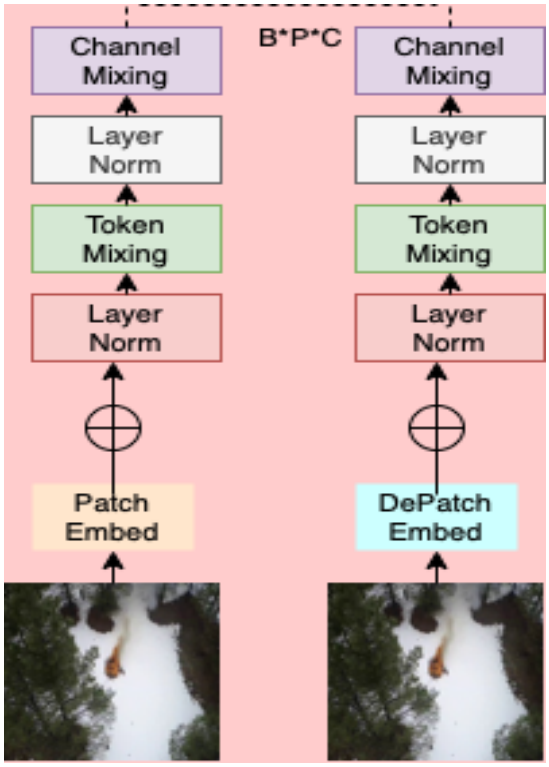


Fig. 5 Left: Original patch based MLP architecture. Right: Modified architecture equipped with DePatch module

The architectural details of the proposed DePatch based MLP Mixer have been discussed. To retain semantics in the aerial photos, we integrated the DePatch module into the MLP Mixer framework. The MLP mixer architecture, as depicted in Fig. 5, receives the DePatch module as an input. The locations of the token and channel mixers are the same as in the original MLP mixer. When compared to CNNs and transformers, the performance of the suggested classification model shows that MLP Mixer is a superior option for aerial data. Our suggested classification approach is capable of handling tasks that demand pixel-level predictions without the use of transformers, as well as photos with significantly higher resolution. In this study, we use the suggested categorization model to locate slash piles that have been burned over the winter in high-elevation forests in the Southwest. The proposed model is able to combat challenges of aerial image classification such as lack of context information and imbalance of fore-ground and background training examples.

4 Experimental Analysis

Our proposed DePatch based MLP Mixer architecture performed pile aerial imagery classification in an effective manner. In this section, details about the aerial pile imagery dataset, training methodology and results for the proposed framework have been discussed.

4.1 Aerial Pile Imagery Dataset

The proposed classification technique has been evaluated on a fire image dataset collected through drones during a burning piled detritus in Arizona pines forest. This dataset contains annotated drone-based images and videos shot from infrared cameras for executing fire related detection, classification and segmentation problems. The fire based classification and segmentation studies can be evaluated on this dataset. For "Fire" vs. "Non-Fire," a total of 39,375 labelled frames were used in the training phase, and 8,617 frames were used for the test data. Early wildfire detection is crucial because it has the potential to cause serious harm to ecosystems, residential areas, forests, and wildlife habitats in the past. These troubling statistics spur scientists to look for fresh approaches to early fire detection and classification. A deep classification model can learn features more effectively by being trained to execute fire picture classification tasks.

4.2 Training Methodology

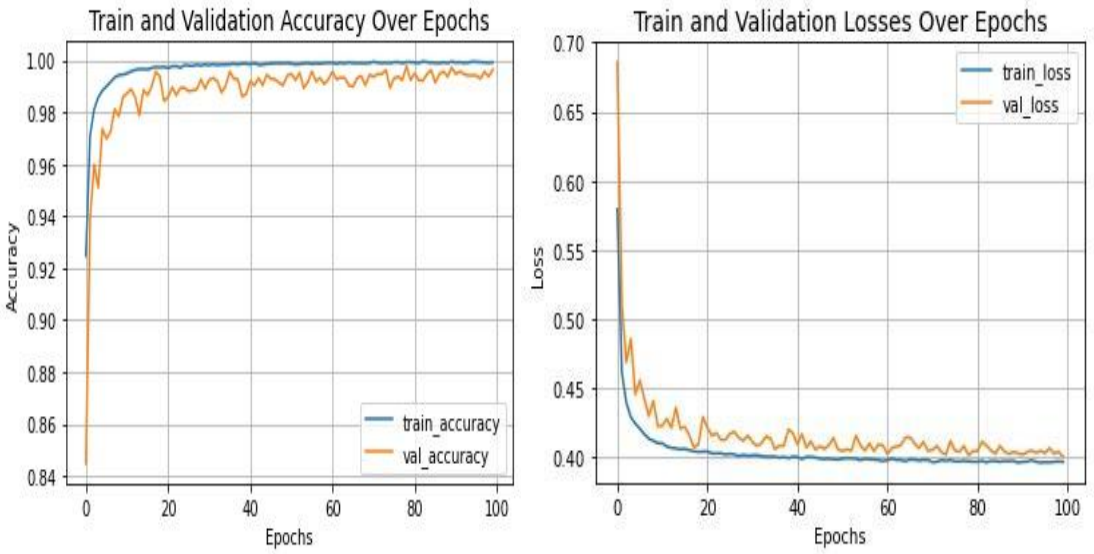
The large-scale images dataset ImageNet has been deployed for initial training of feature extractor for the proposed classification technique. The ImageNet dataset consists of 1.28M images belonging to 1000 categories. We performed fine-tuning on a fire images dataset to detect wildfires at an early stage. The dataset images are resized into 224 × 224 for training of proposed classification technique. To broaden the variety of categorized items, advanced data augmentation techniques like Mix up, CutMix, label smoothing, and Rand-Augment were used. The suggested classification method has been trained using a 32-person batch size over 100 iterations. There is no sign of convergence after 100 epochs. The training approach used the optimizer RMSProp with a weight decay for non-bias parameters of 0.05 and an initial learning rate of 1 103. For fair comparison, all of these settings are maintained with the MLP Mixer architecture.

4.3 Evaluation Parameters

The popular classification-related evaluation metrics had been used to evaluate the effectiveness of the suggested deep learning-based classification technique in order to effectively depict the findings. While recall provides the percentage of True Positives (TP) from the whole quantity of TP and False Negatives (FN), the metric precision refers to the TP fraction from the total sum of TP and False Positives (FP). The true predictions from class one with the highest probability make up the accuracy score.

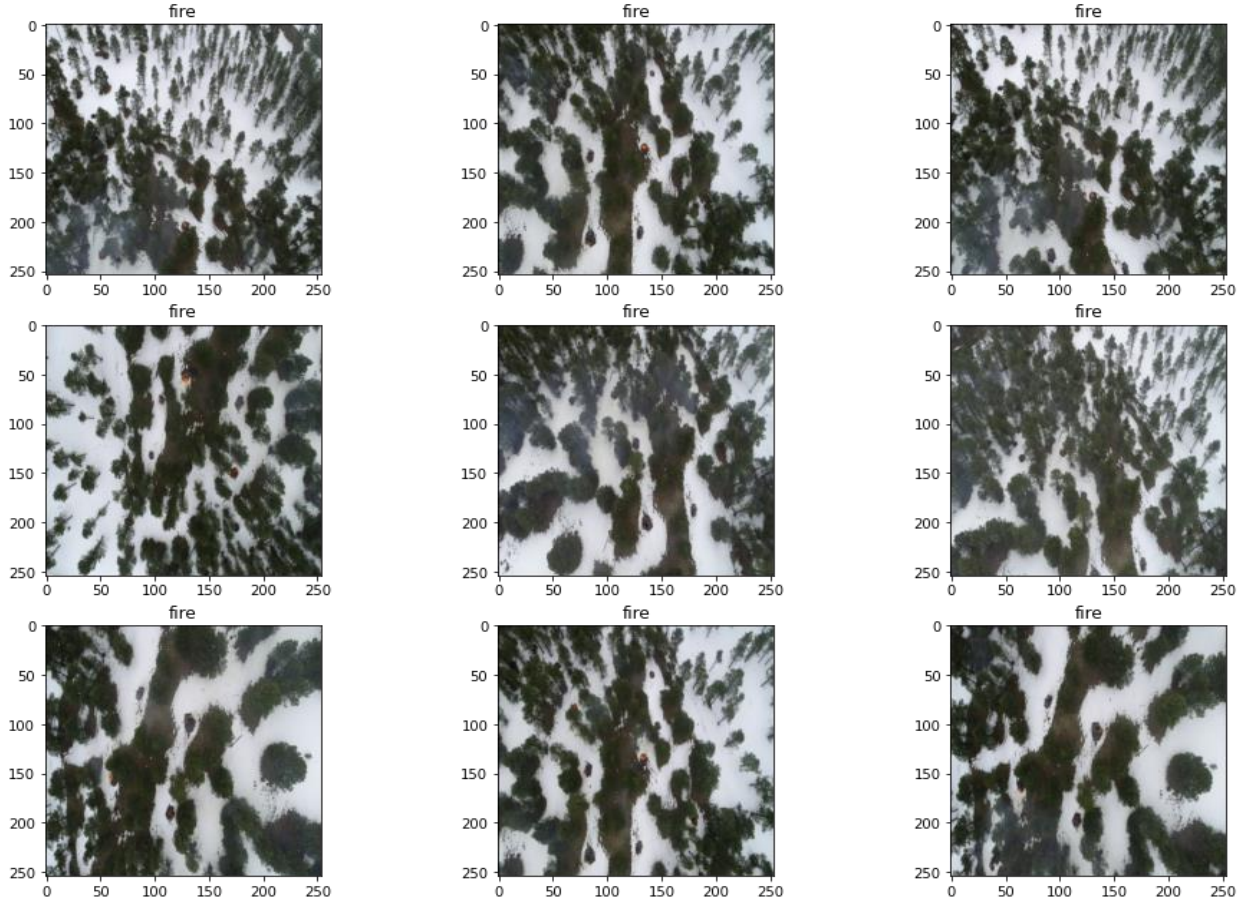
5 Results

In this section, an analysis of the proposed optimized classification technique is provided by considering the accuracy of chosen fire based aerial dataset. The plots related to accuracy and loss for the proposed classification technique have been mentioned. The training accuracy value started from 0.925 and goes up to 1.00 over 100 epochs. The validation accuracy values initially showed bumpy behavior for starting 30 epochs, after which it started converging with the training values. The starting value for the validation accuracy was 0.85 and goes up to 0.99. The training was stopped as no further convergence took place. The training loss value started from 0.85 and goes up to 0.40 over 100 epochs. The validation loss values initially showed a steep change in behavior from .70 to 0.48 value in starting epochs, after which it started converging with the training loss values. The settled value for the validation loss was 0.48 and goes up to 0.41 for 100 epochs. The training was stopped as no further convergence took place. The predictions for the proposed classification technique have been illustrated in Figure. The fire class was properly recognized by DePatch-based MLP mixer technique.



285

286 **Fig. 6** Accuracy and loss plots for the proposed classification model



287

288 **Fig. 7** Correct Predictions for the proposed classification model

289

5.1 Comparison with Other Approaches

To compare the proposed classification technique, the promising transformer models of Vision and Swin transformers have been chosen, shown in Table 1. ViTs used learning from unprocessed data and inductive biases from models. By limiting self-attention computation to non-overlapping local windows while simultaneously allowing for cross-window connections, Swim Hierarchical Transformers were constructed with shifted windows and improved efficiency. These transformer models lacked speed and accuracy while exhausting more hardware resources. The ViT-B/16 version obtained an accuracy score of 46.18 while exhausted 86.4 million parameters and 55.5 G FLOPs. The better version of transformers i.e. Swin Transformers had also attained 61.35 and 82.48 as top-1 and top-5 accuracy scores. The classification results of Swin Transformers were better than vision transformers but could not hold better position in front of our proposed classification model. Transformers' subpar performance is caused by the absence of inductive bias and fixed post-training weights. The drawbacks of transformers include the fact that they can only compute global self-attention for non-overlapping local windows, the difficulty of pixel-level predictions for transformers, and the fact that the computational complexity of their self-attention is quadratic to image size. The input image size for vanilla MLP Mixer model is 224*224 and utilized 19 million parameters through which it achieved top-1 accuracy 74.67 and 87.54 top-5 accuracy. The best performance was recorded by our proposed DePatch based MLP mixer model which outperformed employed other classification models. The proposed classification technique achieved top-1 accuracy of 77.23 and top-5 accuracy of 93.45 which outperformed employed transformer models. The number of employed parameters were also minimum in case of DePatch-based MLP mixer model.

Table 1 Comparison of the proposed classification model with other models

Model	Resolution	Param	FLOPs	Top-1 (%)	Top-5 (%)
ViT-B/16	384*384	86.4M	55.5G	46.18	73.63
Swin	224*224	29M	4.5G	61.35	82.48
MLP Mixer	224*224	19M	2.2G	74.67	87.54
DePatch based MLP (ours)	224*224	18M	2.0G	77.23	93.45

A new module of DePatch had been proposed in MLP mixer model which divides the input images in a deformable pattern to detect forest fires at an early stage. This DePatch module in the chosen MLP mixer based classification model incorporated the awareness of input images and geometric variations. The improved classification model for forest pile burn images of Arizona forest has been proposed which can further be extended for performing multitude of applications. These aerial applications provide the methods for performing smart agriculture, defense missions and industry related activities.

6 Conclusion

This paper proposed an aerial scene classification of forest fire situations from drone images using a novel multi-layer perceptron based network model. The hard patch split of CNNs brought two problems related to collapse of local structures and having semantic inconsistency across images. Transformers would be unable to provide pixel-level predictions on high-resolution aerial photos because the computational difficulty of its self-attention scales quadratically with the size of the image. In MLP Mixer model, multi-layer perceptron blocks turn pictures into a series of patches

and process embeddings of these patches directly. To efficiently analyze massive amounts of data, it relied on token and channel-mixing MLPs as well as conventional regularization and optimization approaches. In order to maintain the semantics of the aerial images, DePatch module had been included into the MLP Mixer framework. Instead of utilising pre-determined fixed patches, this DePatch module adaptively divides the photos into patches with varying positions and scales. A drone-collected fire image dataset from a smouldering pile of debris in an Arizona pine forest was used to evaluate the suggested deep learning-based categorization technique. The suggested classification method outperforms transformer models and the standard MLP-Mixer model with top-1 accuracy of 77.23 and top-5 accuracy of 93.45. The DePatch-based MLP Mixer model likewise used the fewest possible parameters.

Acknowledgements

We acknowledge DIC, Panjab University Chandigarh for funding a workstation that enabled us to perform experiments on NVIDIA TITAN XP GPUs. This research work has been done under UGC NET JRF scholarship, New Delhi, India.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] J.S. Zhang, J. Cao, and B. Mao, "Application of deep learning and unmanned aerial vehicle technology in traffic flow monitoring," in *Machine Learning and Cybernetics (ICMLC), 2017 International Conference on*, 2017, vol. 1, pp. 189–194.

[2] S. Treneska and B. R. Stojkoska, "Wildfire detection from UAV collected images using transfer learning."

[3] Y.-C. C. and P.Y. W. Pei-Chun Chen, "Imaging Using Unmanned Aerial Vehicles for Agriculture Land Use Classification," *Agriculture*, vol. 10, no. 9, pp. 1–14, 2020.

[4] M. Kraft, M. Piechocki, B. Ptak, and K. Walas, "Autonomous, Onboard Vision-Based Trash and Litter Detection in Low Altitude Aerial Images Collected by an Unmanned Aerial Vehicle," *Remote Sens.*, vol. 13, no. 5, p. 965, 2021.

[5] P. Q. and T. P. E. Wang, W. Shu, J. Zhu, S. Xu, "Low-altitude UAV Recognition and Classification Algorithm Based on Machine Learning," in *IEEE 16th Conference on Industrial Electronics and Applications (ICIEA)*, 2021, pp. 1136–1141.

[6] X. Zhang, E. Izquierdo, and K. Chandramouli, "Dense and Small Object Detection in UAV Vision Based on Cascade Network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, p. 0.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv Prepr. arXiv1409.1556*, 2014.

[8] M. Al-Qizwini, I. Barjasteh, H. Al-Qassab, and H. Radha, "Deep learning algorithm for autonomous driving using googlenet," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 89–96.

[9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on*

Artificial Intelligence, 2017.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[11] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv Prepr. arXiv1404.1869*, 2014.
[12] C. Yu *et al.*, "Lite-hrnet: A lightweight high-resolution network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10440–10450.
[13] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105–6114.
[14] A. Shamsoshoara, F. Afghah, A. Razi, L. Zheng, P. Z. Fulé, and E. Blasch, "Aerial imagery pile burn detection using deep learning: The FLAME dataset," *Comput. Networks*, vol. 193, p. 108001, 2021.
[15] H. Zhou, Y. Yuan, and C. Shi, "Object tracking using SIFT features and mean shift," *Comput. Vis. image Underst.*, vol. 113, no. 3, pp. 345–352, 2009.
[16] Y. Pang, Y. Yuan, X. Li, and J. Pan, "Efficient HOG human detection," *Signal Processing*, vol. 91, no. 4, pp. 773–781, 2011.
[17] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*, 2006, pp. 404–417.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[19] F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[20] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 782–791.
[21] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv Prepr. arXiv2010.11929*, 2020.
[22] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.
[23] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.
[24] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv Prepr. arXiv2110.02178*, 2021.
[25] I. O. Tolstikhin *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 24261–24272, 2021.
[26] Z. Chen *et al.*, "Dpt: Deformable patch-based transformer for visual recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2899–2907.

Biographies

Payal is awarded with Junior Research Fellow (JRF) award from University Grant Commission (UGC), New Delhi in 2016. She has published various papers in international journals and conferences. Presently, she is pursuing full time Ph.D. in Computer Science and Engineering at University Institute of Engineering. & Technology, Panjab University, Chandigarh, India. Her research interests include unmanned aerial vehicles, machine and deep learning based image processing.

Akashdeep is currently working as an assistant professor in Computer Science and Engineering at University Institute of Engineering. & Technology, Panjab University, Chandigarh, India. He has chaired 5+ sessions in national and international conferences. He has published 22 research papers in international journals and conferences. He has 12+ years of teaching and research experience. His research interests include wireless networks, soft computing and video analytics, moving object detection and tracking, traffic sensing and classification.

Raman Singh is working as Lecturer in Cyber Security, University of the West of Scotland. He has completed Ph.D. from Institute of Engineering and Technology, Panjab University Chandigarh. He has published 14 research papers in international journals and has 6 years of teaching and research experience. He is a Microsoft Certified Technology Specialist (MCTS) and TechNet Certified Technology Expert. His area of interest includes Intrusion Detection, Network Security, Cyber Security, Autonomous Driving and Machine Learning.