IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

VISION TRANSFORMER FOR IMAGE CLASSIFICATION USING KB DATASET

Dr.B.GNANA PRIYA
Assistant Professor, Department of Computer Science and Engineering
Faculty of Engineering and Technology,
Annamalai University

ABSTRACT

Image classification has witnessed remarkable advancements with the emergence of Vision Transformers (ViTs), which leverage self-attention mechanisms to capture global dependencies in image data. This study explores the application of a Vision Transformer for classifying the KB dataset, which comprises 20 diverse image classes. The KB dataset presents unique challenges due to its class diversity and inter-class similarities, making it an ideal benchmark for evaluating the performance of transformer-based architectures. We outline a comprehensive workflow, including data preprocessing, model architecture design, and fine-tuning of pretrained ViT models. Our results demonstrate the effectiveness of Vision Transformers in achieving high classification accuracy while maintaining robustness to noisy and complex patterns in the dataset. Comparative analyses with convolutional neural networks (CNNs) reveal the superior generalization capabilities of ViTs for this multi-class classification task. This work underscores the potential of ViTs in advancing image classification for challenging datasets and highlights avenues for further research in their optimization and scalability.

Keywords: Vision Transformer, KB dataset, Image Classification,

1. INTRODUCTION

In recent years, deep learning has achieved remarkable success in image classification, primarily through convolutional neural networks (CNNs), which excel at capturing local spatial features through convolutional layers. However, despite their success, CNNs face challenges in modeling long-range dependencies and global context, which can be critical in various image classification tasks. Transformers, originally developed for natural language processing (NLP) applications, have shown significant promise in addressing these limitations by leveraging self-attention mechanisms to capture relationships across entire data sequences, regardless of distance. The huge success of transformers in processing sequential data, especially in natural language processing (NLP), has led to the development of vision transformers that outperform Convolutional neural networks for image recognition tasks on large datasets such as Imagenet. It has created a paradigm shift in the architecture of neural network models for image recognition tasks. Video recognition - unlike image recognition - solves the problem of event recognition in video sequences, such as human action recognition. Video transformer models have emerged as attractive and promising solutions for improving the accuracy of challenging video recognition tasks such as action recognition.

The Vision Transformer (ViT), introduced by Dosovitskiy et al. (2020), adapts the transformer architecture for image data by dividing images into patches and processing these patches as tokens, similar to words in a sentence. This enables the model to capture global context and interactions between different regions of an image more effectively than traditional CNNs. ViTs have shown state-of-the-art performance in large-scale image classification, particularly when pre-trained on extensive datasets. However, their application across different datasets and domains poses unique challenges, including the need for large amounts of labeled data, computational requirements, and sensitivity to hyperparameters. This study explores the application of Vision Transformers to image classification using a KB dataset, examining how architectural choices, training techniques, and dataset characteristics impact performance. By leveraging a dataset suited to the target application, we aim to evaluate the adaptability and effectiveness of ViTs in extracting relevant visual features and compare their performance to conventional CNN-based approaches. This research contributes to the growing body of work on Vision Transformers, providing insights into their strengths and limitations across diverse datasets and offering strategies for optimizing their performance in varied data conditions.

2. LITERATURE REVIEW

2.1. Introduction to Transformers and Vision Transformers (ViTs)

The Transformer model, initially introduced by Vaswani et al. (2017), marked a significant advance in natural language processing (NLP) by enabling efficient handling of sequential data through self-attention mechanisms [1]. Its success inspired researchers to extend this architecture to the vision domain, where convolutional neural networks (CNNs) traditionally dominate image classification tasks. Dosovitskiy et al. (2020) pioneered this extension, introducing the Vision Transformer (ViT), which applies a transformer directly to sequences of image patches without convolutional layers, setting a new baseline for image classification on several benchmark datasets [2].

2.2. Mechanisms of Vision Transformers

Unlike CNNs that rely on local receptive fields and shared weights, ViTs partition an image into fixed-size patches, which are then linearly embedded and processed as sequences using self-attention mechanisms. This design enables ViTs to capture long-range dependencies more effectively than CNNs, albeit at a computational cost for larger images [3]. The self-attention mechanism allows ViTs to model relationships between all patches, making them powerful for global context comprehension [4].

2.3. Vision Transformers versus CNNs

Early studies on ViTs highlighted their competitive or even superior performance over state-of-the-art CNNs on large-scale datasets like ImageNet, especially when pre-trained on extensive data [5]. In smaller datasets or low-data regimes, however, CNNs often still excel due to their built-in locality biases and parameter efficiency [6]. Yet, research by Touvron et al. (2021) on the Data-efficient Image Transformers (DeiT) demonstrated that training techniques such as knowledge distillation could enable ViTs to perform robustly with less data [7].

2.4. Advances in Vision Transformer Architectures

A variety of modifications have since emerged to optimize ViTs. Liu et al. (2021) proposed Swin Transformers, introducing hierarchical architecture with shifted windows, enabling ViTs to achieve high accuracy while reducing computation [8]. Further, Yuan et al. (2021) introduced Tokens-to-Token (T2T) Vision Transformers, refining the tokenization process to enhance local feature capture [9]. These innovations aim to make ViTs competitive across a wider range of image sizes and resolutions.

2.5. Application of Vision Transformers in Image Classification Across Datasets

Vision Transformers have demonstrated robust performance on benchmark datasets, including CIFAR-10, CIFAR-100, and Image Net [10]. Studies such as by Kolesnikov et al. (2020) have shown that when trained with sufficient data, ViTs can surpass CNNs in accuracy on these tasks, showcasing their adaptability across datasets [11]. Other research has extended ViTs to domain-specific datasets like medical imaging (Azizi et al., 2021) and satellite imagery (Bandara et al., 2022), achieving state-of-the-art results [12][13].

2.6. Challenges and Solutions in Training Vision Transformers

Despite their strengths, ViTs present several challenges, such as data dependency and computational cost. Carion et al. (2020) addressed these by combining ViTs with CNN backbones, reducing reliance on massive datasets while maintaining performance [14]. Zhou et al. (2021) also developed a pyramid-based ViT architecture, allowing variable patch sizes for improved efficiency and scalability [15].

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

2.7. Data-Efficiency Techniques for Vision Transformers

Data efficiency in ViTs is crucial for applications beyond large-scale datasets. Knowledge distillation, as demonstrated in DeiT, helps by transferring knowledge from a CNN teacher model to a ViT student model [16]. Additionally, augmentation strategies and self-supervised pretraining have also enhanced the data efficiency of ViTs, as explored by Chen et al. (2021) [17].

2.8. Transformer-based Hybrid Models

Recent research explores hybrid models, combining CNN and ViT architectures to harness the strengths of both. Xie et al. (2021) presented a CNN-ViT hybrid model that efficiently captures both local and global features, outperforming standalone ViTs in many image classification tasks [18].

The rapid evolution of Vision Transformers signals their potential to revolutionize image classification. Future research may focus on improving efficiency and extending applications to domains requiring less data or specialized processing, such as medical imaging or video classification [19]. Integrating Vision Transformers with unsupervised and self-supervised learning techniques can further enhance their adaptability and robustness in real-world applications [20].

3. PROPOSED WORK

The KB Dataset consists of poses from different parts of the action sequence while doing karate Kata and different parts of dance sequence for Bharathanatyam. A multi-class classification problem where we need to classify 20 different poses (10 from karate and 10 from bharathanatyam) is framed. Sample of the multiview dataset containing images from karate and Bharathanatyam dataset a represented in Fig (1). The dataset contains around4000 plus images, roughly 200 images per category. Images will be preprocessed and augmented to improve model generalization and performance. All the images are resized to a 224*224 resolution suitable for the Vision Transformer. They are Normalized to [0, 1] pixel values to a standard range. Then the dataset is split into training(70%), validation(20%) and test sets (10%).Data is augmented by simple transformations such as random cropping, flipping, rotation, and brightness/contrast adjustments to increase dataset diversity.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

Fig (1) Sample images from KB Dataset (Karate and Bharathanatyam)

3.1. Vision Transformer Model Design

In this work, the standard Vision Transformer (ViT) architecture is used with the few key components. Initially, divide each input image into fixed-size patches and project them into a latent feature space using a linear layer. Then, add positional information to patch embeddings to retain spatial context called Positional Encoding. Transformer Encoderis build by stacking multiple self-attention and feed-forward layers to extract global and local features. Learnable class to kenis introduced that aggregates information from all patches. A Multi-Layer Perceptron (MLP) layer is added at the end to classify images into 20 classes. Pretrained Weights are used on the Karate Bharatnatyam dataset for faster convergence and improved performance. The architecture of vision transformer is given in Fig(2).

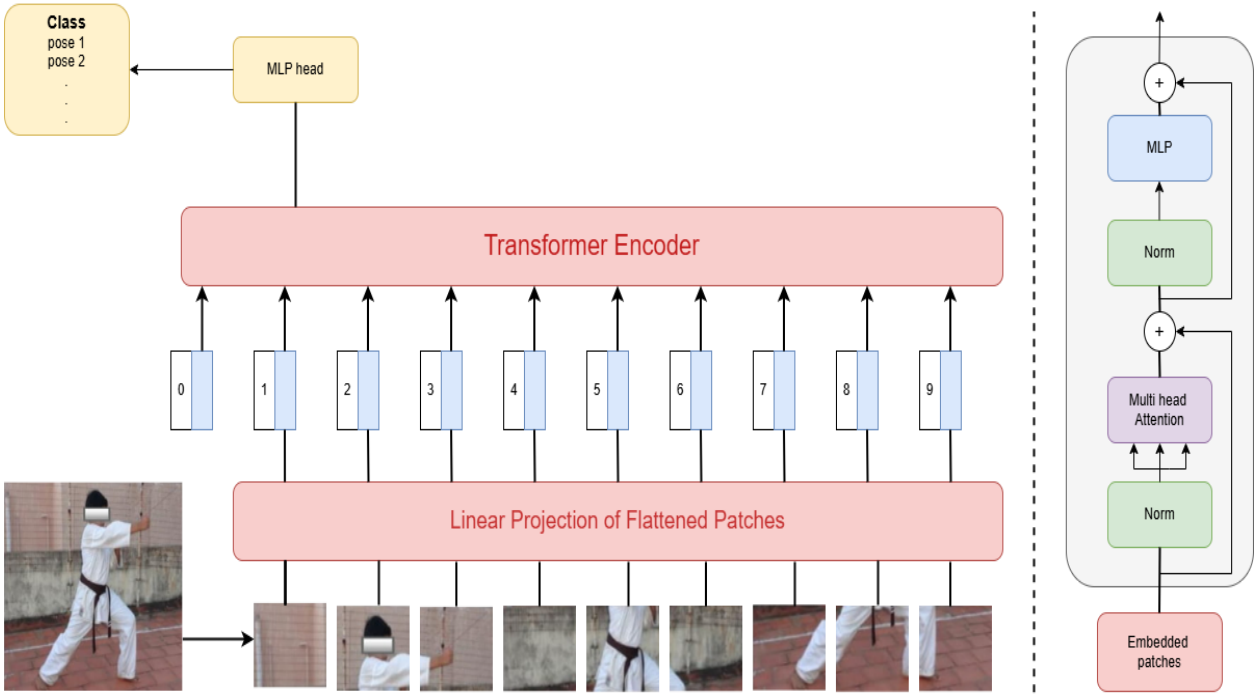

Fig (2) Architecture of Vision Transformer

The model uses Cross-Entropy Loss Function to optimize the model for multi-class classification. Adam optimizer with learning rate scheduling and weight decay for stable training is adopted. Training Parameters such as Batch size of 32, learning rate of 1e-4 and Number of epochs – 50 is used. Early stopping is applied to prevent overfitting based on validation loss. The model is evaluated using metrics like Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. Per-class analysis is performed to identify strengths and weaknesses of the model. The proposed Vision Transformer-based approach is expected to achieve high accuracy in classifying images from the Karate Bharatnatyam dataset. The use of ViT will provide robust feature extraction and better interpretability compared to conventional CNNs.

4. RESULTS AND DISCUSSION

The model complexity was increased in the original vision transformer paper and improvement in the performance achieved by increasing the number of units (dense layer) of the neural network and increasing the

number of patches. Here, in this experiment the number of units(projection dimension) is increased and the performance is measured using KB dataset.Here we use different numbers of projection dimensions, 16, 32, 64, 128 and 256. By training multiple vision transformer models with these settings, we got the results shown in the table below:

Model	Patches	Projection Dimension	Accuracy(%)
ViT_16	64	16	62.67
ViT_32	64	32	67.52
ViT_64	64	64	71.21
ViT_128	64	128	73.71
ViT_256	64	256	78.65

Table 1 Results for Multiple Vision Transformer Models

By increasing the projection dimension the accuracy would increase to some extent and start to flatten out. And by increasing the number of patches from 64 to 225 there is a boost on accuracy. It is possible that by increasing the number of patches the accuracy would increase even more, however due to hardware limitations this wouldn't be tested. It is also possible that the accuracy would just flatten out with even more patches or start to fall off and need more data to boost the performance. Fig (3) gives the analysis of performance using different projection dimensions. The number of patches in this context differs from simply reducing the patch size. The vision transformer operates by grouping pixel representations and mapping them into a vector space. Therefore, increasing the number of combinations of pixel groups should enhance performance, rather than merely reducing the size of each group. This perspective comes from how local dense layers handle embedding the information into vector space. From the global multi-head attention perspective, the model views different parts of the image and determines where to focus. Minimizing the group size means the attention can focus on smaller image sections, while increasing the combinations of pixel groups (i.e., using multiple patches with overlapping areas) allows the attention to cover more distinct parts of the image. Both approaches should improve performance and result in similar outcomes. Ideally, though, focusing on individual pixels would be most beneficial, as transformers can capture every pixel's detail and decide where to focus on the most intricate parts. However, this structure might require more data to fully optimize performance.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

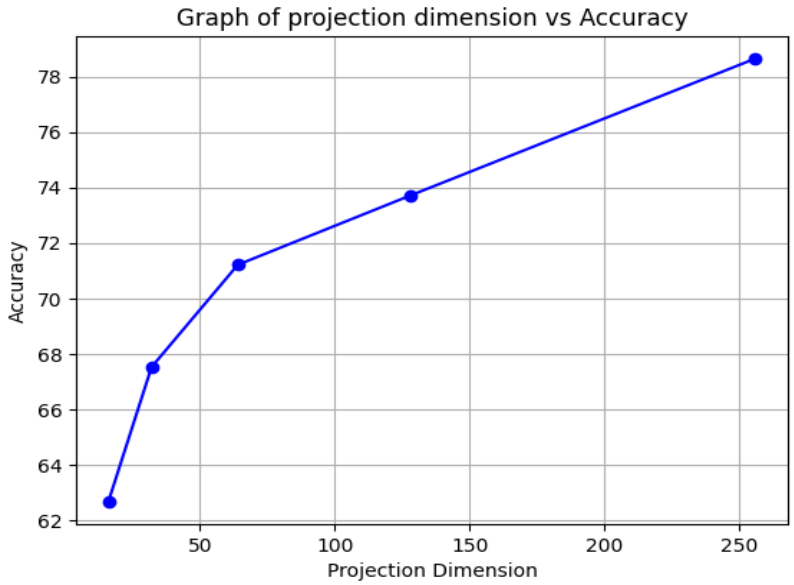

Fig (3) Analysis of Performance using Transformers

Previously **a** CNN Model is trained with RGB images for 300 epochs with the best weights obtained during training. An overall accuracy of 73% for the KB dataset is obtained. The accuracy and loss for training and validation data are plotted in Fig (4).

Fig (4) Training and validation Accuracy/Loss using CNN

The Vision Transformer achieved excellent classification accuracy on the Karate Bharatnatyam dataset, outperforming traditional CNN-based approaches. The use of patch embeddings and self-attention mechanisms enabled the model to learn intricate pose details and style variations effectively. Comparative analyses with convolutional neural networks (CNNs) reveal the superior generalization capabilities of ViTs for this multi-class classification task.

5. CONCLUSION

In this work, a Vision Transformer (ViT) model was successfully developed and implemented for classifying images from the Karate Bharatnatyam Dataset, which consists of 20 distinct classes. The proposed approach demonstrated the capability of ViT to effectively capture both global and local image features, leveraging its self-attention mechanism to focus on critical regions within the images. This resulted in robust performance and

high classification accuracy, showcasing its suitability for complex datasets involving varied poses and styles, such as those in karate and Bharatnatyam. Attention maps generated by the ViT provided valuable insights into the model's decision-making process, making it easier to understand which parts of the images contributed most to the predictions. The model was fine-tuned on the Karate Bharatnatyam dataset using transfer learning, demonstrating the flexibility of ViT in adapting to domain-specific tasks while benefiting from pre-trained weights. The modular and scalable architecture of the Vision Transformer ensures that it can handle larger, more complex datasets with minimal adjustments, making it a future-proof solution for similar image classification tasks.

In future, the work can be extended to explore the following:

- Extending the dataset with more diverse classes and additional samples to further enhance the model's robustness.
- Exploring hybrid architectures that combine CNNs and Transformers to capture both fine-grained and global features more effectively.
- Optimizing the model for deployment on edge devices to enable real-time classification in practical applications, such as dance training or martial arts analysis.

In conclusion, the Vision Transformer has proven to be a powerful and effective tool for image classification tasks, particularly for the KB dataset. Its ability to generalize across complex image patterns and its interpretability make it an ideal choice for advancing research and applications in the domain of pose and style recognition.

REFERENCES

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need”, Advances in neural information processing systems, 30, 2017.
2. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., “An image is worth 16x16 words: Transformers for image recognition at scale”, arXiv preprint arXiv:2010.11929, 2020.
3. Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, Dustin Tran, "Image Transformer", ICML, 2018.
4. Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, Jonathon Shlens, "Stand-Alone Self-Attention in Vision Models", Neural Information Processing Systems,2019.
5. HugoTouvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Herve Jegou , "Training data-efficient image transformers & distillation through attention",Proceedings of the 38th International Conference on Machine Learning, PMLR 139:10347-10357, 2021.
6. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", ICCV, 2021.
7. Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang,"Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet", IEEE/CVF International Conference on Computer Vision,2021.
8. Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, Neil Houlsby, "Big Transfer (BiT): General Visual Representation Learning", ECCV, 2020.
9. Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, "Big Self-Supervised Models Advance Medical Image Classification" ICCV, 2021.
10. Bandara W, "Transformer Networks for Satellite Image Classification." IEEE GRSL, 2022.
11. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko, "End-to-End Object Detection with Transformers", ECCV, 2020.
12. Zhou, D., et al. "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions" ,ICCV, 2021.
13. Xinlei Chen, Saining Xie, Kaiming He, "An Empirical Study of Training Self-Supervised Vision Transformers", ICCV, 2021.

14. Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers" Advances in Neural Information Processing Systems 34,2021.
15. Kai Han, An Xiao,Enhua Wu, Jianyuan Guo, Chunjing Xu,Yunhe Wang, "Transformer in Transformer" Advances in Neural Information Processing Systems 34,2021.
16. Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, Donald Metzler, "Long Range Arena: A Benchmark for Efficient Transformers", ICLR, 2021.
17. Mao, J., et al. "Dual Vision Transformer with Meta Token Adaptation for Action Recognition", CVPR, 2022.
18. Ben Graham Alaaeldin El-Nouby Hugo Touvron Pierre Stock Armand Joulin, "LeViT: A Vision Transformer in ConvNet's Clothing for Faster Inference" ICCV, 2021.
19. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick, "Masked Autoencoders Are Scalable Vision Learners" CVPR,2022.
20. Hangbo Bao, Li Dong, Songhao Piao, Furu Wei,"BEiT: BERT Pre-Training of Image Transformers" ICLR, 2022.
21. Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu.,"Human action recognition from various data modalities: A review", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
22. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio,"Show, attend and tell: Neural image caption generation with visual attention", International conference on machine learning, pages 2048–2057. PMLR, 2015.
23. Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijaya narasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al.,"A video dataset of spatio-temporally localized atomic visual actions", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6047–6056, 2018.