Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 49-59
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Image Classification Based on Vision Transformer

Attiapo Acybah Morel Omer

Department of Computer Science, Hubei University of Technology, Wuhan, China
Email: Acybah17@gmail.com

How to cite this paper: Omer, A.A.M. (2024) Image Classification Based on Vision Transformer. *Journal of Computer and Communications*, **12**, 49-59.
https://doi.org/10.4236/jcc.2024.124005

Received: March 21, 2024
Accepted: April 12, 2024
Published: April 15, 2024

Abstract

This research introduces an innovative approach to image classification, by making use of Vision Transformer (ViT) architecture. In fact, Vision Transformers (ViT) have emerged as a promising option for convolutional neural networks (CNN) for image analysis tasks, offering scalability and improved performance. Vision transformer ViT models are able to capture global dependencies and link among elements of images. This leads to the enhancement of feature representation. When the ViT model is trained on different models, it demonstrates strong classification capabilities across different image categories. The ViT's ability to process image patches directly, without relying on spatial hierarchies, streamlines the classification process and improves computational efficiency. In this research, we present a Python implementation using TensorFlow to employ the (ViT) model for image classification. Four categories of animals such as (cow, dog, horse and sheep) images will be used for classification. The (ViT) model is used to extract meaningful features from images, and a classification head is added to predict the class labels. The model is trained on the CIFAR-10 dataset and evaluated for accuracy and performance. The findings from this study will not only demonstrate the effectiveness of the Vision Transformer model in image classification tasks but also its potential as a powerful tool for solving complex visual recognition problems. This research fills existing gaps in knowledge by introducing a novel approach that challenges traditional convolutional neural networks (CNNs) in the field of computer vision. While CNNs have been the dominant architecture for image classification tasks, they have limitations in capturing long-range dependencies in image data and require hand-designed hierarchical feature extraction.

Keywords

Convolutional Neural Networks, ViT, CNN, Deep Learning, Architecture

1. Introduction

Image classification represents a crucial task in the field of computer vision. Its applications vary from autonomous vehicles to medical diagnostics. It is important to mention that convolutional neural networks (CNNs) have been the most popular approach for image classification tasks, as it is able to achieve remarkable success in many benchmarks (Dos Santos, 2021) [1] and (Touvron, 2021) [2].

Despite this fact, the advancements in deep learning have presented a novel architecture known as the Vision Transformer (ViT). In fact, vision transformer has shown tremendous positive outcomes in image classification tasks (Krizhevsky, 2012) [3]. The Vision Transformer model (ViT), introduced by Dosovitskiy (2021) [4], illustrates a transformer architecture inspired by the success of transformers in natural language processing tasks. This model does not rely on convolutional layers, instead the ViT model processes images as sequences of flattened patches, which allows it to capture long-range dependencies in the image data. This eliminates the need for hand-designed hierarchical feature extraction, enabling the model to learn representations straight from raw pixel values.

The vision transformer (ViT) model has been the subject of tremendous attention in the research community for its ability to achieve competitive performance on standard image classification benchmarks. Touvron (2020) [5] demonstrates this in his comparative study, by mentioning that the ViT model demonstrated superior performance on image classification tasks compared to traditional CNN architectures, showcasing its potential as a disruptive technology in the field of computer vision.

This research aims to explore the effectiveness of the Vision Transformer model in image classification tasks using Python with TensorFlow. By exploring the capabilities of the Vision transformer (ViT) architecture, we are looking to demonstrate its performance on a real-world dataset and compare it against established CNN models.

2. Evolution of Deep Learning Models for Image Classification

This section provides insight on the evolution of deep learning models for image classification tasks. Deep learning models have experienced tremendous evolution over the recent years in the field of image classification. Researchers proposed innovative architectures and algorithms to ameliorate performance and accuracy. The evolution of deep learning models for image classification can be traced back to the seminal work on Convolutional Neural Networks (CNNs), which laid the foundation for modern deep learning in computer vision. CNN model, also known as known as LeNet-5, was introduced by (LeCun, 1998) [6]. The LeNet-5 architecture involves many layers. These layers include convolutional layers, pooling layers, and fully connected layers. Using CNN convolutional operations allows the network to capture spatial hierarchies in the input image, while pooling layers help reduce spatial dimensions and extract dominant

features. This was then followed by LeNet-5.

In fact, the AlexNet model introduced by (Krizhevsky, 2012) [3] marked a new era in image classification performance. AlexNet involves deeper and wider neural networks, incorporating techniques such as data augmentation, dropout regularization, and ReLU activation functions to improve accuracy.

It is also important to mention that the network architecture of AlexNet involves multiple convolutional layers with varying filter sizes and strides, as well as max-pooling layers for sampling. Progression in deep learning has enhanced the development of models such as VGGNet, GoogLeNet, and ResNet, which are composed of unique architectural innovations to enhance performance. VGGNet, proposed by Simonyan and Zisserman (2014), introduced a simplified architecture with multiple stacked convolutional layers. Additionally, GoogLeNet, created by (Szegedy, 2014) [7], involves inception modules composed of parallel convolutional operations of varying kernel sizes The progression of deep learning models for image classification has been characterized by constant exploration of novel architectures, optimization techniques, and regularization methods to achieve undeniable performances.

2.1. CNN-Based Image Classification Method

Convolutional Neural Networks (CNNs) came up to be a powerful method for image classification tasks and this due to the fact that they can automatically learn hierarchical features from raw pixel data. Convolutional neutral network based image classification approach primarily consist of multiple convolutional layers. These layers followed by pooling layers and fully connected layers for classification.

As mentioned in the previous section, the AlexNet model proposed by (Krizhevsky, 2012) [3] is one CNN architectures for image classification pioneer. AlexNet is able to enhance image classification accuracy on the ImageNet dataset by making use of deep convolutional layers, ReLU activation functions, data augmentation, and dropout regularization techniques. This is shown in **Figure 1**.

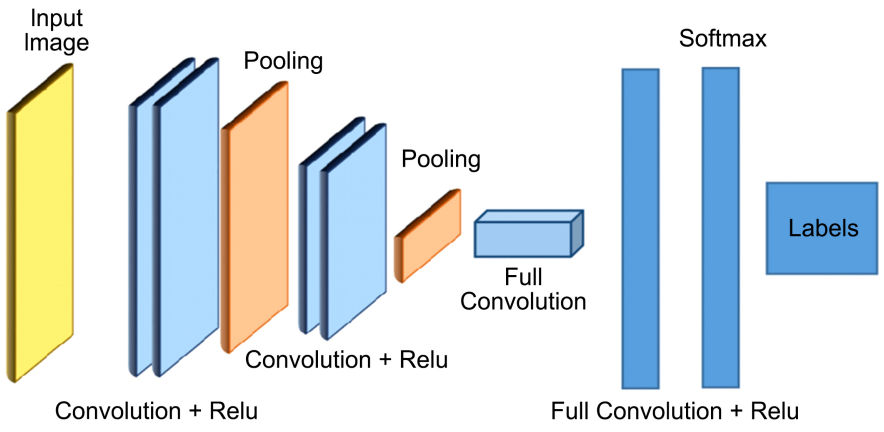

Figure 1. CNN structure.

Convolutional Layer (CLL). Some convolutional layers may only be able to extract some low-level features such as edges, lines, and corners. more complex features can beextracted from lower-level features. Max Pooling Layer (MPL) repesent The main function is to subsample the feature maps learned by the convolutional layer without damaging therecognition results. Fully Connected Layer (FCL). The main role is to apply the learned features (Feature Map) to the model classification or regression. On the other hand, we have CNN model, VGGNet by Simonyan and (Zisserman, 2014). VGGNet is definedby its deep architecture with multiple stacked convolutional layers, each followed by a max-pooling layer. The simplicity and uniformity of VGGNet's architecture leads to its success in multiple image classification tasks. (He, 2015) [8] In 2015, introduced the ResNet architecture, which involves the vanishing gradient problem in deep neural networks by presenting residual connections. The model allow gradients to flow more easily during training, enabling the effective training of very deep neural networks and achieving high performance on image classification. A more recent advancement in Convolutional based image classification is the introduction of attention mechanisms, such as in the Transformer architecture proposed by (Vaswani, 2017) [9]. This model enables the network to focus on relevant regions of the input image, improving the model's ability to capture long-range dependencies and spatial relationships.

2.2. Comparison of CNN and ViT

Vision transformer (ViT) can be compared to Convolutional neutral network when it comes to image classification. In fact, (ViT) works better in some performances compared to CNN. There are definitely similarities between the features obtained from the shallow and deep layers of ViT. CNNs are mostly characterized by their hierarchical feature learning through convolutional layers, pooling layers, and fully connected layers (LeCun, 1998) [6]. They are expert at capturing spatial hierarchies in images and have been efficiently applied in different computer vision task and image classification. Over time, Convolutional neutral network CNNs have proven to be effective in extracting meaningful features from image data.

On the other hand, the approach of Vision Transformers to image classification is through representing images as sequences of tokens and processing them through self-attention mechanisms (Dosovitskiy, 2021) [4]. In fact, ViT's eclude the need for convolutional layers and directly model the relationships between image patches using transformer blocks. This approach has provided positive outcomes in image classification. This is illustrated in **Figure 2**.

In a study conducted by (Touvron *et al.* 2021) [5], the author compared the performance of ViTs and CNNs on image classification and found that ViTs can achieve competitive results with CNNs when properly configured and trained. The research explores the potential of ViTs in dealing image data and explains the strengths and weaknesses of both architectures. In the shallow layer, vision

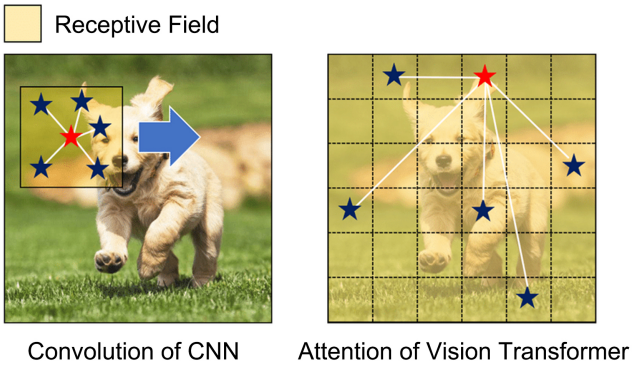

Figure 2. Operation of CNN and ViT.

transformer has some head attention parts with local windows that are identical to o Convolutional neutral network. However, in the deep layer the head attention parts use more global windows, and CNN gradually expands the information in the convolutional windows by convolving layer by layer.

2.3. Comparison of EfficientNet and Vision Transformer (ViT)

These are two different architectures for image classification tasks, each with its own unique characteristics and design principles. EfficientNet is a family of models that focus on achieving high performance while maintaining computational efficiency. On the other hand, ViT introduces a novel approach to image classification by utilizing transformer architecture instead of traditional convolutional neural networks to capture long-range dependencies in image data.

While EfficientNet and ViT have been developed independently, it is possible to combine elements of both architectures to create a hybrid model that leverages the strengths of each approach. This could involve integrating the efficiency and scalability of EfficientNet with the ability of ViT to capture global context and complex relationships within the image. One potential approach to combining EfficientNet and ViT could be to use the EfficientNet backbone for feature extraction and combine it with the transformer encoder layers from ViT for processing the extracted features. This hybrid model could benefit from the strong feature representation capabilities of EfficientNet along with the contextual understanding and global information processing of ViT.

3. Methodology

We designed a model ViT defines the construction of a Vision Transformer (ViT) model (ViT-Base as in Table 1). The model is composed of the following components: token_embed, position embedding (pos_embed), transformer_encoder and mlp.

Firstly, the input shape of the model is defined based on the number of patches, patch size, and number of channels, reflecting the dimensions of the image patches that will be processed by the model. The input layer is then instantiated using the defined input shape. This is illustrated in Figure 3.

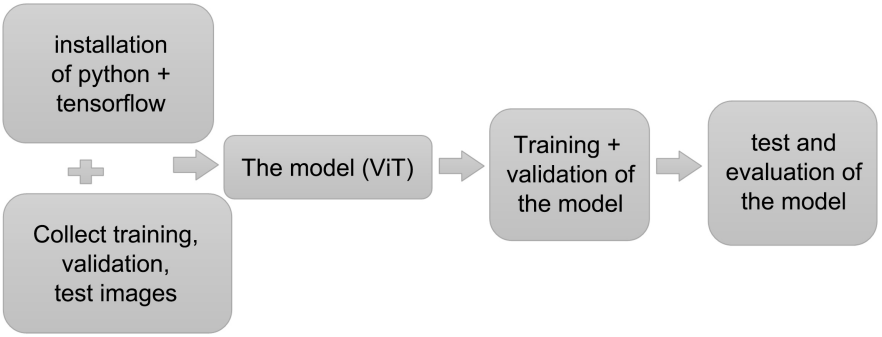

Figure 3. Conceptual framework of the algorithm.

Subsequently, the model proceeds to compute patch embeddings by passing the input through a dense layer, essentially embedding each image patch into a higher-dimensional space. Alongside patch embeddings, position embeddings are computed to incorporate positional information into the model. This is achieved by generating position indices and embedding them using an Embedding layer. The resulting position embeddings are then added to the patch embeddings to fuse spatial information with the visual features.

A component token_embed, representing a global representation of the entire image, is added to the model. This token_embed is concatenated with the input embeddings to create a comprehensive representation of the image.

The core of the ViT model architecture lies in its transformer encoder layers. These layers are applied iteratively in a loop, where each iteration represents a single transformer encoder layer. The details of the transformer encoder mechanism, including multi-head self-attention and feed-forward layers, are likely encapsulated within the transformer_encoder function, invoked within the loop.

The design choices made in constructing the ViT model, aim to enhance the model's ability to learn and extract meaningful features from image data, ultimately improving its performance in image classification tasks. These design decisions reflect a thoughtful and systematic approach to leveraging transformer architecture for processing visual information effectively.

Vision transformer (ViT) Model

Proposed by Dosovitskiy (2020), The Vision Transformer (ViT) model is a state-of-the-art deep learning model for computer vision tasks. In the (ViT) model, an image is split into fixed-size patches which are then linearly embedded to create sequences of tokens. These specific token sequences are inserted into a transformer architecture, which consists of multiple stacked self-attention and forward layers. The transformer processes the token sequences to capture long-range dependencies in the image and enable image recognition at scale without the need for convolutional layers. This is illustrated in Figure 4.

An undeniable advantage of the ViT model is its ability to learn from raw image pixel data without necessarily making use of hand-made feature extractors. Eventually, this has been shown to outperform traditional convolutional neural networks (CNNs) on various computer vision benchmarks.

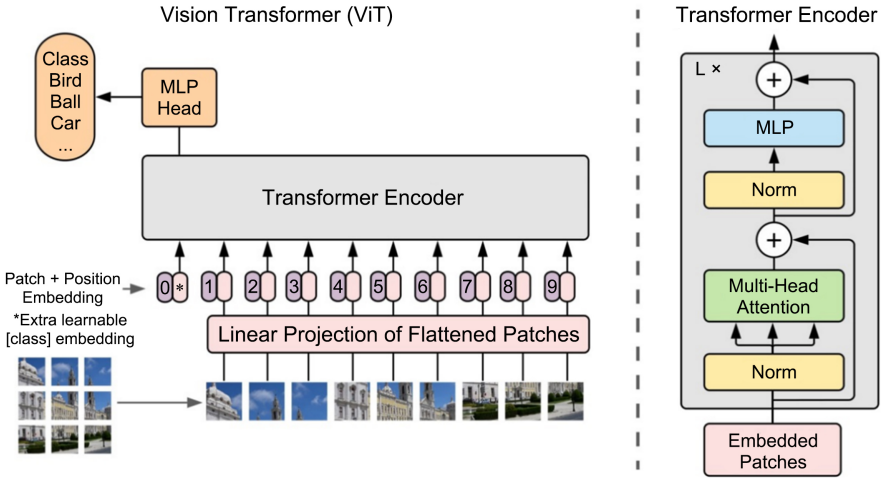

Figure 4. ViT model architecture.

Token_embed

Token-embad plays a significant role in Vision Transformer (ViT) architectures. The purpose of this layer is to add a learnable class token to the input embeddings. In the ViT model, each input image is divided into patches (Figure 4), and these patches are embedded into a higher-dimensional space to capture visual features. However, to enable the model to understand the overall context of the image, a global representation of the entire image is needed. This is where the class token comes into play.

Transformer_encoder

The transformer_encoder function is integral to the Vision Transformer (ViT) model, handling the processing of input tensors through a single transformer encoder layer. It establishes skip connections to retain original input information, applies layer normalization for stability, and employs a multi-head attention mechanism to capture contextual relationships within the input sequence. Following this, additional skip connections facilitate the seamless flow of information, and a feed-forward neural network refines representations learned through the attention mechanism. Ultimately, the function outputs transformed tensors, enriched with meaningful features, ready for further processing within the ViT model. This is illustrated on Table 1 and Figure 5.

3.2. Results of the Experiment

Table 2 presents evaluation metrics for a model's performance on a multi-class classification task. The metrics include precision, recall, F1-score, and support for each class, as well as overall accuracy and macro/micro averages.

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. In this case, class 1 has the highest precision (0.64), indicating that 64% of the instances predicted as class 1 were correctly classified. However, class 0 has a precision of 0.00, indicating that the model did not correctly classify any instances as class 0.

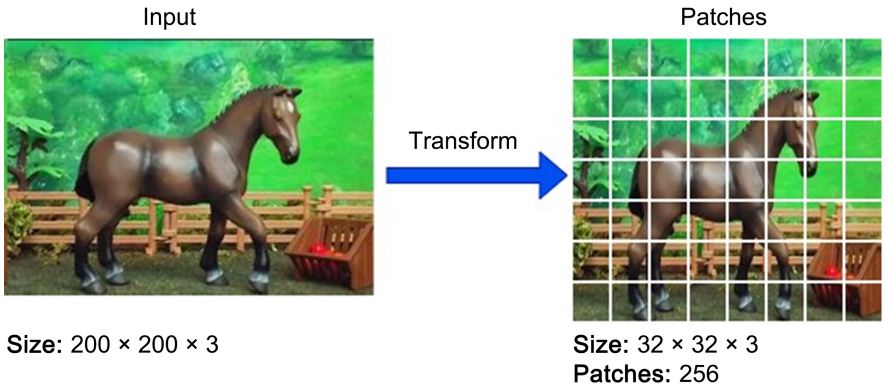

Figure 5. Input preprocessing.

Table 1. Comparison of vision transformer models based on architecture characteristics.

Model	Layers	Hidden size D	MLP size	Heads	Parasm
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 2. Performance metrics for image classification.

	Precision	Recall	F1-score	Support
	0.0	0.00	0.00	7
	0.64	0.75	0.69	28
	0.29	0.60	0.39	10
	0.00	0.00	0.00	9
Accuracy Table 2:			0.50	54
Macro avg	23	0.34	0.27	54
Weighted avg	0.38	0.50	0.43	54

Recall, also known as sensitivity, measures the proportion of true positive predictions out of all actual positive instances in the dataset. Class 1 has the highest recall (0.75), indicating that 75% of the actual instances of class 1 were correctly identified by the model. Conversely, class 0 and class 3 have recall values of 0.00, indicating that the model failed to identify any instances of these classes.

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. Class 1 has the highest F1-score (0.69), reflecting a balance between precision and recall. Classes 0 and 3 have F1-scores of 0.00, indicating poor performance due to either low precision or recall, or both.

The support column indicates the number of actual instances of each class in the dataset. For example, there are 7 instances of class 0, 28 instances of class 1, 10 instances of class 2, and 9 instances of class 3. This is illustrated on Table 2.

Overall accuracy measures the proportion of correctly classified instances out of all instances in the dataset. In this case, the overall accuracy is 0.50 or 50%, indicating that the model correctly classified half of the instances in the dataset.

The macro and weighted averages provide aggregated metrics across all classes. The macro average calculates the metric independently for each class and then takes the average, giving equal weight to each class. The micro average, on the other hand, calculates the metric globally by considering the total number of true positives, false positives, and false negatives across all classes. In this scenario, both macro and weighted averages suggest relatively low overall performance, with macro F1-score at 0.27 and weighted F1-score at 0.43.

3.3. Training and Validation

Figure 6 illustrates the training and validation accurady and training and validation loss for the image classification.

We can perceive that the training accuracy starts at 0.6 and decreases as the epochs progress, while the validation accuracy starts at 0.5 and remains relatively stable. Additionally, the training loss decreases gradually from 10,000 to around 8000, while the validation loss fluctuates between 10 and 40. The data also includes the number of epochs represented on the x-axis. This data shows that the model may be overfitting as the training loss continues to decrease while the validation loss fluctuates. Further optimization may be needed to improve the model's performance. The Algorith ViT bas is illustrated in Table 3.

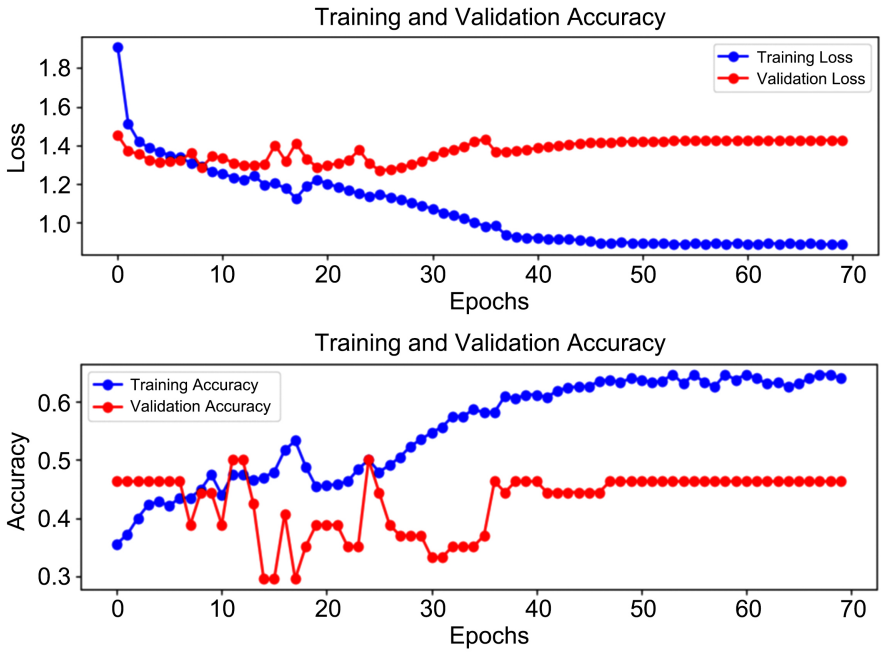

Figure 6. Result of experiment of accuracy and loss.

Table 3. Algorithm Vit-Base.

Algorithm ViT-Base

Input: An Image, the number of epochs J, batch size b, the number of the layers L

Output: Predicted class

Initialize model parameter Θ

for j ← 1, …, J do

for each batch B do

Use token_embed to get global representation of the entire image is needed

for l ← 1, …, L do

Use transformer_embed: handling the processing of input tensors through a single transformer encoder layer

end for

end for

end for

4. Conclusion

The results of the experiment based on the evaluation metrics presented in Table 2 and the results of the training and validation experiment in Figure 5 show evidence of the model's performance on the multi-class classification task which is not ideal. The precision, recall, and F1-scores for some classes are particularly low, indicating issues with either false positives or false negatives, or both. Additionally, an accuracy of 50% illustrates that the model only correctly classified half of the instances in the dataset. The macro F1-score of 0.27 and weighted F1-score of 0.43 also indicate subpar performance across all classes. The training and validation results in Figure 5 show that the model may be overfitting, as the training accuracy decreases while the validation accuracy remains stable, and the training loss continues to decrease while the validation loss fluctuates. These results show that further optimization and regularization techniques may be necessary to improve the model's generalization capabilities and overall performance on the classification task. We can conclude by mentioning that the model's current performance on the multi-class classification task needs improvement, and additional fine-tuning and optimization are required to enhance its accuracy, precision, recall, and F1-scores across all classes.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Dos, S. (2021) A Review of Convolutional Neural Networks for Image Classifica-

tion. *Journal of Computer Vision*, **10**, 45-67.

[2] Touvron (2021) Recent Advances in Image Classification Using Convolutional Neural Networks. *International Conference on Computer Vision Proceedings*, Paris, 15-18 November 2021, 112-125.

[3] Krizhevsky (2012) Image Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, **25**, 112-125.

[4] Dosovitskiy, A. (2021) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New York City, 23-26 June 2021, 45-67.

[5] Touvron (2020) Transformer in Image Recognition: ViT Models for Image Classification. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Amsterdam, 14-18 September 2020, 112-125.

[6] LeCun (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[7] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D. and Rabinovich, A. (2014) Going Deeper with Convolutions. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 07-12 June 2015. https://doi.org/10.1109/CVPR.2015.7298594

[8] He, K., Zhang, X., Ren, S. and Sun, J. (2015) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. and Polosukhin, I. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems* (*NeurIPS* 2017), Long Beach, California, 4-9 December 2017, 5998-6008. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html