Attention Distillation for Accuracy Improvement of Vision Transformer

Taiga Tanaka[1], Ryuto Ishibashi[1], Yifan Xu[1], Lin Meng[2]

{menglin@fc.ritsumei.ac.jp}

[1] Graduate School of Science and Engineering, Ritsumeikan University,
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan

[2] College of Science and Engineering, Ritsumeikan University,
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan

Abstract. In recent years, rapid advancements in technology have propelled progress in the fields of image recognition and natural language processing, resulting in the development of numerous applications and services. In particular, the advancement of deep learning has greatly contributed to improving the accuracy of these fields, and the use of AI is becoming more widespread in everyday life[1]. Among them, Vision Transformers (ViT) has emerged as a robust architecture in image recognition. ViT requires a large dataset for training to improve accuracy. To solve this problem, DeiT using knowledge distillation has been proposed. However, this DeiT does not consider each token's contribution to [CLS] in learning. In this study, a knowledge distillation model training method called ADViT that uses the contribution of each token to Classify token ([CLS]) is proposed. Specifically, the attention map of each token for [CLS] is calculated. In each layer, learning is performed to bring the student model's attention map closer to the teacher model's attention map. Then, the accuracy is evaluated, and experimental results show the effectiveness of the proposed method, with accuracy rates of ADViT-Tiny(L1): 98.35%, ADViT-Tiny(L2): 98.06%, ADViT-Small(L1):99.00%, and ADViT-Small(L2): 98.75%, which are higher than those of DeiT, DeiT-Tiny: 96.39%, DeiT-Small: 96.83%. Future work includes managing the number of heads and reducing the computation for more optimization.

Keywords: Computer Vision, CNN, Vision Transformer, Knowledge Distillation

1 Introduction

Technological capabilities in fields such as image recognition and natural language processing have progressed incredibly in recent years. This progress has led to the rapid development of various applications and services[2], and AI is becoming increasingly prevalent in everyday life. In particular, the advancement of deep learning technology has played a pivotal role in enhancing the

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365289

accuracy of image recognition and natural language processing. Among them, Vision Transformer (ViT)[3][4] has emerged as a powerful architecture in the field of image recognition, leveraging the advantages of the Transformer[5, 6][7] model, which has been successful in natural language processing. Unlike traditional convolutional neural networks (CNNs)[8][9], ViT processes input images as a sequence of tokens, with each token corresponding to a part (patch) of the image. These tokens are aggregated through a transformer layer, and a dedicated class token [CLS] represents the final classification output.

However, to maintain the high accuracy of ViT, extensive datasets and computational resources are required, which is a significant problem in resource-limited environments. To address this problem, knowledge distillation has attracted attention as an effective method. Knowledge distillation is a method used to transfer knowledge from a highly accurate teacher model to a smaller, more efficient student model, enabling the student to retain high accuracy while lowering computational requirements. Data-efficient Image Transformer (DeiT)[10], a ViT model using knowledge distillation[11], has achieved competitive accuracy while achieving efficient training with less data and resources than ViT. DeiT aims to improve accuracy by bringing the student model's predictive distribution closer to the teacher model's predictive distribution. However, in conventional methods such as DeiT, the detailed relationship of how much each token contributes to [CLS] is unclear. In ViT, how the information of each token is integrated into [CLS] is considered a significant factor in improving the accuracy of the model.

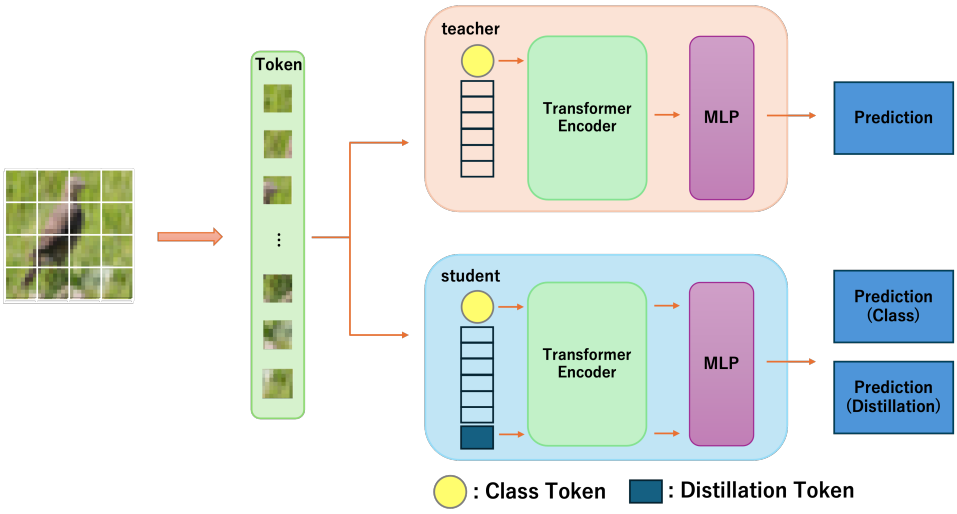

Fig. 1. The structure of the DeiT:This is the structure of DeiT, which trains the teacher model based on ViT. In the DeiT student model, a different prediction result is output using [CLS] and distillation tokens. The loss is calculated using these two prediction results and the prediction result of the teacher model.

In this paper, a new model, Attention Distillation ViT (ADViT), is proposed to improve the accuracy of distillation learning. The model takes into account the contribution rate of each token

to [CLS]. In the proposed method, the student model learns the dependency relationship with the teacher model's [CLS] and aims to improve the accuracy of the entire model by more accurately reflecting the contribution rate of each token.

This paper makes the following two contributions.

- Propose ADViT, a new method of training using knowledge distillation. The contribution rates of each token to the [CLS] in the student model are trained to approximate the contribution rates of each token to the [CLS] in the teacher model.

- Propose L1 norm, L2 norm as loss functions for ADViT.

The structure of this paper is as follows. Section. 2 introduces related works on ViT and DeiT. Section. 3 explains how the student model is trained to approximate the [CLS] and the contribution rates of each token from the teacher model. Section. 4 presents the experimental conditions, evaluation metrics, and experimental results. Section. 4.3 discusses the problems identified from the experimental results and proposes improvements for future research, and Section. 5 provides a brief conclusion.

2 Related work

2.1 Vision Transformer (ViT)

Vision Transformer (ViT) is an adaptation of the Transformer model, which originally demonstrated exceptional performance in natural language processing (NLP) tasks and has been extended to image recognition tasks. The fundamental concept of ViT is to partition the input image into fixed-size patches and process these patches as tokens, analogous to words in NLP. Each image patch is flattened into a one-dimensional vector and transformed into a feature vector using a linear layer. This process produces a fixed-size embedding representation for each patch.

In ViT, to leverage the Transformer's capability for processing sequences, a special learnable token called the Classification token [CLS] is appended at the start of the patch embeddings. This [CLS] token gathers critical information required for the final image classification. The structure of the Transformer Encoder is similar to that of a language model, consisting of Multi-Head Attention and a Feed-Forward Network (FFN), with Layer Normalization applied after each block.

The most distinctive element of ViT is that it leverages the self-attention mechanism to efficiently handle global information instead of relying on the local filtering of traditional CNNs. Self-attention uses three elements, a query ($\mathbf{Q}$), a key ($\mathbf{K}$), and a value ($\mathbf{V}$), to capture the interrelationships between each token (patch embedding). As shown in Equation (1), the similarity (attention score) is calculated based on the inner product of the $\mathbf{Q}$ and the $\mathbf{K}$, and then the $\mathbf{V}$ is weighted and summed using this to obtain a new feature representation based on the dependencies between each token. This process results in a high-dimensional representation that captures the features of the entire image and is particularly good at modeling the relationships between different regions in the image.

$$\mathrm{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \psi\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}. \tag{1}$$

In addition, in the final classification task, all patch embedding information processed by the Transformer Encoder is aggregated into the [CLS], so the final representation of this token is extracted, and the image is classified through the Multilayer Perceptron (MLP). What is unique about ViT is that this [CLS] integrates dependencies with other tokens obtained by self-attention of each patch embedding, allowing it to consider broader relationships rather than traditional local feature extraction.

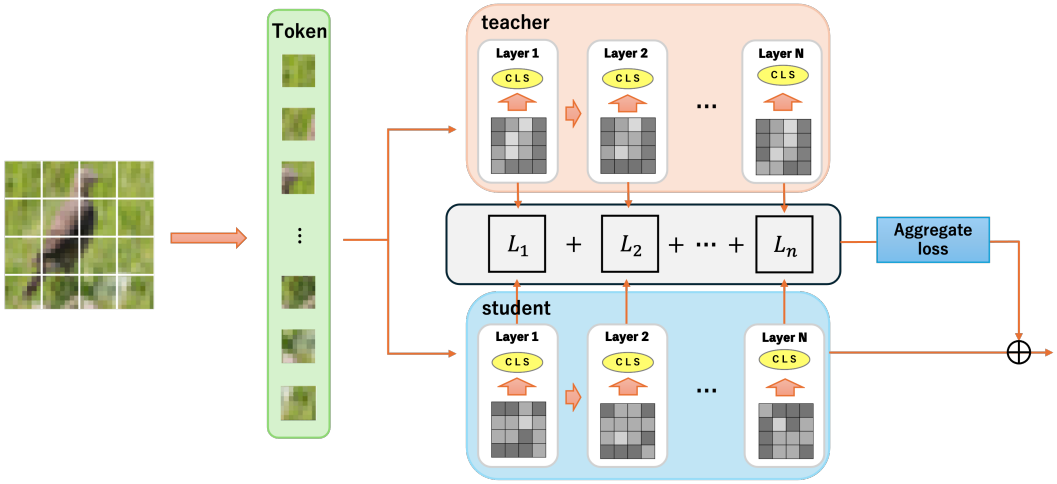

Fig. 2. Overview of Attention Distillation model: In Fig. 2, the black and white attention map shows how much each token contributes to the [CLS]. After the losses are calculated at each layer, they are summed and averaged.

2.2 Data-efficient Image Transformer (DeiT)

Data-efficient Image Transformer (DeiT) is a method for efficient learning utilizing knowledge distillation. Knowledge distillation is how a student model learns knowledge from a pre-trained teacher model.

The innovation of DeiT is the introduction of distillation tokens in addition to ViT's [CLS]. This distillation token is a means of training using the output of the teacher model and effectively performing knowledge distillation. Specifically, it is designed so that the student model can proceed with learning while referring to the output of the teacher model (Usually a large-scale CNN model (RegNetY-16GF)[12]). This token incorporates knowledge obtained from the teacher model into the model while learning dependencies with other tokens through self-attention. This makes it possible to efficiently learn with a small amount of data, which was difficult with ViT alone. Even when compared to the CNN model EfficientNet-B7[13], the accuracy is improved depending on the dataset.

In DeiT, two main methods of knowledge distillation are proposed: soft distillation and hard

distillation. These two methods learn by utilizing the difference in the output of the teacher model and the student model, and each has different characteristics.

2.2.1 soft distillation

Soft distillation is a technique in which a student model learns by utilizing the output distribution of a teacher model. In this case, the output of the teacher model is taken as a distribution, which is compared with the prediction of the student model to calculate the loss. Specifically, it is expressed as in Equation (2).

$$\begin{aligned}\mathscr{L}_{\text{global}}^{\text{softDistill}} = (1-\lambda)\mathscr{L}_{\text{CE}}(\psi(Z_s), y) \\ + \lambda\tau^2 \text{KL}\left(\psi\left(\frac{Z_s}{\tau}\right), \psi\left(\frac{Z_t}{\tau}\right)\right)\end{aligned} \tag{2}$$

$\mathscr{L}_{\text{CE}}$ represents the cross entropy (CE) between the student model prediction Z_s and the correct label y, and KL stands for the KL divergence between the teacher model output Z_t and the student model prediction Z_s. λ is a parameter that adjusts the weighting of these two losses, and τ represents the temperature parameter for distillation. The role of the temperature parameter τ is to smooth the model output distribution, making it more flexible to learn. ψ represents the softmax function, which converts the model output into a probability distribution.

2.2.2 hard distillation

In hard distillation, the prediction results output by the teacher model are used as labels, and the labels are directly compared with the predictions of the student model using CE. For both the hard labels obtained from the teacher model and the correct labels, the CE between the output of the student model is calculated, and the average is taken. Specifically, it is expressed as in Equation (3).

$$\mathscr{L}_{\text{global}}^{\text{hardDistill}} = \frac{1}{2}\mathscr{L}_{\text{CE}}(\psi(Z_s), y) + \frac{1}{2}\mathscr{L}_{\text{CE}}(\psi(Z_s), y_t) \tag{3}$$

Each $\mathscr{L}_{\text{CE}}$ represents the CE between the student model's prediction Z_s and the correct label y, and between the student model's prediction Z_s and the teacher model's output Z_t, and the two $\mathscr{L}_{\text{CE}}$ are averaged. This approach allows DeiT to train on fewer data sets than the base ViT model and still achieve the same level of performance. This improves data efficiency and makes more efficient learning possible.

3 Methodology

In this study, to improve the performance of the ViT model, ADViT, a distillation learning method that uses a large-scale ViT model as a teacher model and adjusts the contribution of each token to [CLS] is proposed. Conventionally, DeiT is widely known as a model that uses knowledge distillation. DeiT learns using the final output obtained from the teacher model, specifically the probability distribution of the prediction result. However, in conventional methods, distillation is

performed depending on the final output, so there is a problem with information about the contribution of each internal token to the [CLS] not being reflected.

Since [CLS]s directly impact the final output in ViT, their accuracy is significantly involved in the overall prediction performance. In addition, understanding how the information of each token is aggregated and contributes to the [CLS] is thought to be the key to optimizing the internal structure of the model in a deeper sense. Therefore, in this study, instead of simply using the probability distribution of the final output for distillation, a method is proposed to realize more detailed knowledge transfer by analyzing the contribution of each token to [CLS] and incorporating this into the distillation process. This approach is expected to not only improve the performance of the entire model, but also clarify how each token affects the output.

3.1 Attention Distillation for ViT (ADViT)

Fig. 2 shows the overview of ADViT. The proposed model utilizes a ViT-based (Vision Transformer) architecture for both the teacher and student models. The training process proceeds as follows: the input image is first split into 16x16 pixel patches, which are treated as a sequence of tokens and then fed into the teacher and student models, respectively. At each layer, an attention map is computed to indicate the contribution of each patch token to the [CLS] used for class classification. Based on this attention map, the loss function is calculated between the teacher model and the student model for each layer. The loss formula calculated at each layer is shown in (4). This is the output part of the soft label formula of DeiT that is changed to an attention map. A_i^s, A_i^t denote the $i-th$ token of the attention map of the student model and teacher model, respectively.

The losses calculated at each layer are then summed and averaged. The calculated loss is combined with the classification loss of the student model and used as the training objective for the final student model. This loss design enables the effective transfer of knowledge from the teacher model to the student model, aiming to enhance training accuracy.

$$\begin{aligned}\mathscr{L}_{\mathrm{KL}}^{\mathrm{Distill}} = (1-\lambda)\mathscr{L}_{\mathrm{CE}}(\psi(A_s), y) \\ + \lambda\tau^2 \mathrm{KL}\left(\psi\left(\frac{A_s}{\tau}\right), \psi\left(\frac{A_t}{\tau}\right)\right)\end{aligned} \tag{4}$$

3.2 Types of Losses

In this study, in addition to the KL loss and CE commonly used in DeiT, training is performed using L1 and L2 norms. Conventional methods use a probability distribution based on the number of classes as a loss function. Still, in this study, the contribution of a class to a token is calculated using an Attention Map of 196 tokens. In this case, it is thought that applying softmax makes it difficult to see the difference in loss for each element. To prevent this problem, training using L1 and L2 norms is adopted, as shown in Equations (5) and (6).

$$\mathscr{L}_{\mathrm{l1}}^{\mathrm{Distill}} = \|A\|_1 = \sum_{i=1}^{n} |A_i^s - A_i^t| \tag{5}$$

$$\mathscr{L}_{l2}^{\text{Distill}} = \|A\|_2 = \sum_{i=1}^{n} (A_i^s - A_i^t)^2 \tag{6}$$

$\mathscr{L}_{l1}^{\text{Distill}}$ represents the L1 norm and calculates the difference between the contributions (A_i^s, A_i^t) of the $i-th$ token of the student model and the teacher model to the [CLS]. $\mathscr{L}_{l2}^{\text{Distill}}$ represents the L2 norm and calculates the difference between the squared contributions (A_i^s, A_i^t) of the $i-th$ token from the student model and the teacher model to the [CLS].

Table 1: Results of learning each model

Model	**Method**	**Type of Loss**	**Top1 Accuracy** (%)	**Params** (M)
CNN	ResNet10t[14]	CE	94.40	4.93
	ResNet34	CE	97.34	21.29
	EfficientNet-B1[15]	CE	95.42	6.53
	EfficientNet-B4	CE	98.12	17.57
Tiny@224	ViT[3]	CE	95.59	5.53
	DeiT[10]	hard	96.39	5.53
	ADViT	KL	96.17	5.53
	ADViT	**L1**	**98.35**	**5.53**
	ADViT	**L2**	**98.06**	**5.53**
Small@224	ViT	CE	96.34	21.67
	DeiT	hard	96.83	21.67
	ADViT	KL	96.94	21.67
	ADViT	**L1**	**99.00**	**21.67**
	ADViT	**L2**	**98.75**	**21.67**
Teacher	ViT-Base	-	99.34	85.81

4 Evaluation

4.1 Experimental Configurations

4.1.1 Dataset

This experiment uses CIFAR-10[16] as the classification task. CIFAR-10 is a dataset of color images containing objects such as animals and vehicles, and consists of small color images of 32×32 pixels with ten classes. The dataset consists of 50,000 training images and 10,000 test images. RandomResizedCrop at 224×224, RandomHorizontalFlip, and RandomErasing are used for data augmentation.

Table 2: Results of classification accuracy for test data

Test Class	ADViT-Small(L1)	DeiT-Small
airplane	**99.50%**	98.50%
automobile	**99.30%**	98.40%
bird	**98.80%**	95.80%
cat	**97.70%**	92.80%
deer	**99.40%**	96.60%
dog	**97.20%**	94.50%
frog	**99.50%**	98.50%
horse	**99.50%**	98.40%
ship	**99.80%**	97.30%
truck	**99.00%**	97.50%

4.1.2 Implementation Conditions

The experiments in this paper are conducted on an AMD Ryzen 9 7900X3d 12-core Processor CPU with NVIDIA GeForce RTX 4090 GPU for training and inference. AdamW is used to optimize the model parameters. In the training condition, the number of epochs for the classification task is set to 300, respectively.

4.1.3 Comparison Models

This experiment uses ViT, DeiT, the proposed method ADViT, the CNN ResNet[14], and EfficientNet[15]. The sizes of the ViT-based models are tiny and small, respectively. Each ViT model has 12 transformer encoder layers, but the embedding dimension and the number of heads vary depending on the model size. The tiny model has an embedding dimension of 192 and 3 heads, and the small model has an embedding dimension of 384 and 6 heads. CNNs with parameters close to those of the ViT model are used as comparison targets. The teacher model for DeiT and ADViT is ViT-Base, and hard labels are used for the loss of DeiT, while KL, L1 norm, and L2 norm are used for the loss of ADViT. The KL temperature function $\tau = 1.0$ is set. The teacher models are pre-trained on ImageNet-12k and ImageNet-1k[17] and fine-tuned on CIFAR-10. The CNN used in this experiment has parameters close to ViT. For ResNet, ResNet-10t and ResNet-34 are used, and for EfficientNet, EfficientNet-B1 and EfficientNet-B4 are used.

4.1.4 Evaluation Index

In this experiment, Top-1 Accuracy(%) and Parameter(M) are used as the evaluation index. The proposed method is compared with each of the conventional methods.

Table 3: Comparison of prediction results on test images.

Model	Result	Class	Test data images
ADViT-Small	✓	bird	  
DeiT-Small	×		
ADViT-Small	✓	car	  
DeiT-Small	×		

4.2 Experimental Results

4.2.1 Comparison of learning results for each model

Table 1 shows the classification accuracy and parameter results of the CNN-based model, ViT-based models, and the proposed method. Among the CNN-based models, EfficientNet-B4:98.12% achieved the highest accuracy. The conventional ViT-Tiny-based model, DeiT-Tiny, had an accuracy of 96.39%, while the ViT-Small-based model, DeiT-Small, had an accuracy of 96.83%, both of which were lower than EfficientNet-B4. Among the ViT-Tiny-based models, the proposed ADViT model using the L1 norm for loss achieved the highest accuracy of 98.35%. Among the ViT-Small-based models, the proposed ADViT model using the L1 norm for loss achieved the highest accuracy of 99.00%. Among the ADViT models, models using the L2 norm also achieved accuracies of 98.06% on Tiny and 98.75% on Small, demonstrating higher accuracy than DeiT. However, the model using KL loss showed an accuracy of 96.17% on Tiny and 96.94% on Samll, which was less accurate than DeiT on Tiny. Also, when comparing the number of parameters, ADViT achieved the highest accuracy among similar parameters.

4.2.2 Comparison of classification accuracy using test data

Table 2 shows the classification accuracy results for 10 classes using the CIFAR-10 test data with ADViT-Small and DeiT-Small, which had the highest accuracy among the proposed methods. ADViT shows higher accuracy for all ten classes, from airplanes to trucks. In particular, for cats, DeiT showed a success rate of 92.80% while ADViT showed a success rate of 97.70%, which is an improvement of approximately 5%. Table 3 also shows images in the test data for Bird and Car that were correctly predicted by ADViT-Small but not by DeiT-Small.

4.3 Discussion

ADViT has improved accuracy compared to other ViT and CNN models, and its effectiveness has been confirmed. In addition, the contribution of each token to the [CLS] in ViT is considered an important factor in classification. In addition, when comparing the accuracy of each class in the test data, The improvements in accuracy were observed for all classes, suggesting that ADViT can be used for a variety of image types. Therefore, it is believed that ADViT can achieve highly accurate image recognition in many situations. In this experiment, the KL loss used as the loss function did not improve accuracy compared to the L1 and L2 norm. This may be because the softmax was applied to the contribution of 196 tokens, which did not result in a sufficient difference in loss. This problem is believed to be improved by adjusting the temperature parameter τ. In addition, compared to EfficientNet-B4, the accuracy of DeiT is about 1.5% lower, which is probably due to the fact that CNN is not used as the teacher model. In this experiment, ViT-Base was used as the teacher model to match the conditions of the proposed method but to further improve the accuracy of the comparison target; it is necessary to use a CNN model such as RegNet[12] as the teacher model.

In future research, the number of heads used in the student and teacher models should be unified to improve accuracy. In addition, while this research mainly focuses on improving accuracy, it is necessary to consider thoroughly reducing the computational cost in the future. Specifically, pruning techniques [18] and token merging (ToMe) are being considered as methods to reduce the computational load . We also try to apply the model for the real application

5 Conclusion

In this study, we proposed ADViT, a novel knowledge distillation framework for Vision Transformers that leverages the contribution of individual tokens to the [CLS] token to enhance image recognition accuracy. By distilling the attention-based importance maps from a larger teacher model to a compact student model, we aimed to preserve critical structural information during the training process.

Experimental results on the CIFAR-10 dataset demonstrate the superior performance of ADViT over conventional methods. Specifically, ADViT-Tiny and ADViT-Small achieved peak accuracies of 98.35% and 99.00%, respectively, when utilizing the L1 norm for loss calculation. This represents a significant improvement of approximately 2% over the baseline ViT and DeiT models of similar sizes, and even outperformed high-performance CNNs like EfficientNet-B4. Furthermore, our class-specific analysis revealed that ADViT provides more robust feature representations, as evidenced by a substantial 5% accuracy boost in challenging categories such as ”cat.” These findings confirm that attention-based distillation is a highly effective strategy for training lightweight yet accurate ViT architectures.

Despite these achievements, several challenges remain. The discrepancy in the number of attention heads between student and teacher models and the choice of temperature parameters in KL divergence require further optimization. Future research will focus on unifying architectural configurations and integrating advanced efficiency techniques, such as pruning and ToMe, to further reduce computational overhead. Ultimately, we aim to deploy this high-performance, low-latency

framework in real-world applications and resource-constrained environments.

References

[1] Li H, Wang Z, Yue X, Wang W, Tomiyama H, Meng L. An architecture-level analysis on deep learning models for low-impact computations. Artificial Intelligence Review. 2023;56(3):1971-2010.

[2] Yue X, Meng L. YOLO-MSA: A Multi-scale Stereoscopic Attention Network for Empty-Dish Recycling Robots,. IEEE Transactions on Instrumentation and Measurement. 2023;72:1-14.

[3] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.

[4] Ren J, Li H, Wang A, Saho K, Meng L. Radar-based gait analysis by Transformer-liked network for dementia diagnosis,. Biomedical Signal Processing and Control. 2024.

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[6] Ishibashi R, Kaneko H, Meng L. Enhancing DETR with Attention-Based Thresholding for Efficient Early Japanese Book Reorganization; 2023. .

[7] Ishibashi R, Nojiri N, Saho K, Meng L. Optimized Vision Transformer for Dementia Diagnosis using Micro-Doppler Radar (Accepted). In: 2023 IEEE Intl. Conf. on Systems, Man, and Cybernetics (SMC); 2023. .

[8] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural computation. 1989;1(4):541-51.

[9] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.

[10] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. PMLR; 2021. p. 10347-57.

[11] Gou J, Yu B, Maybank SJ, Tao D. Knowledge distillation: A survey. International Journal of Computer Vision. 2021;129(6):1789-819.

[12] Radosavovic I, Kosaraju RP, Girshick R, He K, Dollár P. Designing network design spaces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 10428-36.

[13] Tan M. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:190511946. 2019.

[14] He K, Zhang X, Ren Sea. Deep residual learning for image recognition. In: Proc. of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.

[15] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR; 2019. p. 6105-14.

[16] Krizhevsky A, Hinton G, et al.. Learning multiple layers of features from tiny images. Toronto, ON, Canada; 2009.

[17] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015;115:211-52.

[18] Zhu M, Tang Y, Han K. Vision transformer pruning. arXiv preprint arXiv:210408500. 2021.