International Journal of AI, BigData, Computational and Management Studies
Noble Scholar Research Group | Volume 3, Issue 4, PP. 115-122, 2022
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I4P112

Vision Transformers (ViT) for Small-Scale Image Classification with Token Reduction

Sajud Hamza Elinjulliparambil
Pace University.

Abstract: Convolutional Neural Networks (CNNs) are no longer considered a superior choice in image classification over Vision Transformers (ViTs), which have shown to be highly effective on a large scale because they can exploit self-attention aspects in capturing long-range dependencies. Nevertheless, in small datasets of images like CIFAR-10, CIFAR-100 and SVHN, standard ViTs are usually inefficient in data usage, expensive to run, and redundant in token representation. This weakness is due to the fact that ViT requires large training corpora and its self-attention has a quadratic complexity with regard to the number of tokens. In order to overcome these issues, the initial studies suggested token-cutting techniques such as pruning, pooling, and a combination of CNN and ViT networks to reduce tokens on the input side and maintain the needed visual information. These ways seek to reduce overfitting and increase computation efficiency and generalization at low-data regimes. This review gives a thorough analysis of Vision Transformers in small-scale image classification with regard to developments. The paper provides an overview of ViTs development, their drawbacks when working with small datasets, an overview of initial token reduction methods, and their performance on a variety of benchmark tasks. Its discussion reflects both advantages and disadvantages of token reduction and provides the future research prospects. The article is a reference material to scholars conducting an investigation on effective transformer-based models specialized to small-scale image categorization tasks.

Keywords : Vision Transformer (Vit); Token Reduction; Small-Scale Image Classification; Self-Attention; CNN Vit Hybrid Models; Attention Pooling; Computational Efficiency; Data Efficiency; Transformer Architectures.

1. Introduction

Image classification is one of fundamental activities in computer vision that tries to give an input image a semantic label (object) belonging to a fixed set of categories [1]. During the last ten years, deep learning, especially Convolutional Neural Networks (CNNs) has made a tremendous leap, and thus, large-scale datasets like ImageNet can be handled with remarkable performance. Nonetheless, small datasets, such as CIFAR-10, CIFAR-100, and SVHN, have been permanently problematic. Small labeled samples also cause overfitting, inaccurate generalization and unstable training dynamics, particularly with large-capacity models [2]. CNNs, which are effective, can face difficulties in such situations as they depend on local receptive fields and inductive biases that are unlikely to be able to make the best use of global contextual information.

Over the past years, the transformer architecture, initially used in natural language processing has been applied to computer vision and has become known as Vision Transformers (ViTs) [3]. ViTs also have self-attention mechanisms to model long-range dependencies and thus can capture global context much better than CNNs do. It is shown that ViTs are able to reach the state-of-the-art performance on large-scale image classification datasets like ImageNet-21k, being more accurate or equal to the traditional CNNs [4]. ViTs break down images into patches, encode them in the form of tokens and run them through several transformer encoder layers, resulting in strong representations to be used in a classification task. Figure 1 has given a representation of the ViT architecture where the flow of the input image to patch embeddings, transformer processing, and final classification output is depicted.

This figure 1 illustrate the process of dividing an input image into patches, embedding these patches as tokens, passing them through transformer encoder blocks with self-attention and feed-forward layers, and finally using a classification head to predict image labels. The visualization highlights the flow from raw image to token embeddings and the transformer processing pipeline.

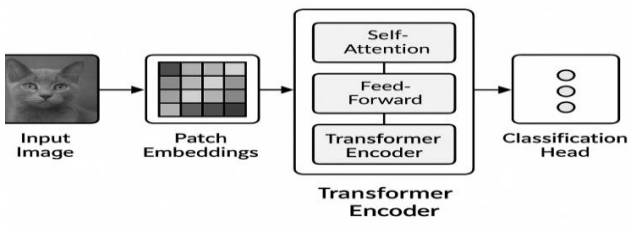

Figure 1: Overview of Vision Transformer Architecture for Image Classification

Regardless of these achievements, standard ViTs are facing limitations when used on small datasets. They are also less data-effective than CNNs because they usually need large volumes of labeled data in order to prevent overfitting [5]. Moreover, the self-attention mechanism has quadratic computational complexity with respect to tokens, which makes it very expensive to use high-resolution images. The overlapping information in patches can worsen this issue, which is why the strategies that limit the unwarranted computations without affecting the model performance are necessary[6].

The problem of token reduction has become one of the potentially viable solutions to these issues. Certainly, token reduction strategies can enhance computational efficiency, minimise memory usage and enhance generalization on small datasets by locally pruning or aggregating less informative tokens. Methods that are investigated are attention-based pruning, pooling, and hybrid CNN-ViT models which combine the advantages of both models. The strategies allow ViTs to perform well, and overcome overfitting and resource needs.

This review aims to critically examine how Vision Transformers can be adapted to small scale image classification, in respect to token reduction mechanisms.
This paper aims to:

- Overview the history and construction of ViTs
- Name the issues related to the application of ViTs to small datasets, such as data and computational efficiency problems
- Examine token cut-down methods and their effects on performance,
- Compare findings in various small scale datasets and approaches.

Lastly, the paper also sheds light on the possible research directions in improving the efficiency and generalization of ViTs to small-data settings that can be regarded as an overall guide to the researchers and practitioners looking into the application of transformer-based solutions to the small-scale image classification task.

2. Background

To analyze Vision Transformers (ViTs) in small-scale classification and token reduction, it is also important to learn how the theory of CNN-based image classification works, and how transformers found their way into the vision field.

2.1. Convolutional Neural Networks (CNNs) for Image Classification

Historically, Convolutional neural networks (CNNs) have been the most successful architecture used to classify images, especially when small datasets are used, e.g. CIFAR-10, CIFAR-100, and SVHN [7]. They are successful because they have local receptive fields, weight sharing, and hierarchical feature extraction which makes them easily learn edges, textures, and parts of objects even with few training samples. AlexNet, VGG, ResNet, and DenseNet architectures made CNNs the new standard backbone of the majority of vision tasks [8].

CNNs however, also have structural limitations. Their use of local convolutions limits their capacity to capture global or long-range relationships in the far distant spatial areas. Whereas deeper networks or dilated convolution can improve the size of the effective receptive field, it remains the case that it fails to compete with the ability of global attention mechanisms. Such constraints stimulated the search of architectures that could describe context holistically of image that would lead to transformers in vision.The figure2 visually highlights CNN's localized feature extraction vs. ViT's global self-attention processing.

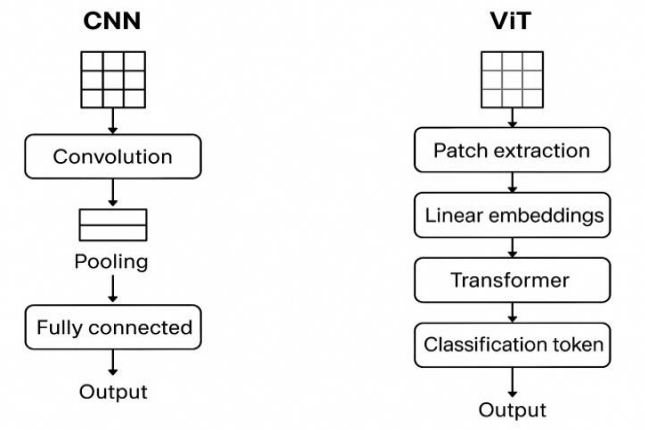

Figure 2: CNN vs ViT schematic

2.2. Transformers in Vision

The emergence of transformers in NLP prompted their extension to images, and modern attention-based vision models are based upon them.

2.2.1. Original Transformer Architecture

It is important to note that the transformers in NLP have a certain structure before they are applied to images.
Recurrence and convolution were substituted by self-attention, and the transformer architecture that researchers proposed transformed sequence modeling [9]. It has a fundamental component, multi-head self-attention which computes interactions among all pairs of tokens, which allows the model to compute long-range relationships effectively. A typical transformer encoder has:

- Multi-head self-attention layers that aggregate information globally
- Feed-forward networks (FFN) applied independently to each token
- Layer normalization and residual connections for training stability
- Positional encodings to preserve sequence order

Transformers proved highly scalable and parallelizable, leading researchers to explore their potential in vision by treating images as sequences of smaller units.

2.2.2. Vision Transformer (ViT)

The vision transformer is the adaptation of transformers to images which represents a trend of not using convolution-based architectures [10].That was made possible by the Vision Transformer (ViT), which proposes to process an image as a sequence of fixed (size) patches similar to word tokens in NLP. ViT consists of the major elements:

- Patch Embedding: The picture is segmented into patches (e.g. 16×16) which do not overlap, and flattened, and then projected in an embedding space.
- Positional Encoding: Positional vectors are learnable, which is utilized to preserve spatial relationships.
- Transformer Encoder: Stacked blocks of self-attention and feed-forward blocks predict global dependencies.
- Classification Token: This is a special token that sums up the whole image representation.

ViT showed an advantage or competitive results on large-scale data sets like ImageNet-21k and JFT-300M. Nevertheless, when working on small data sets (e.g. CIFAR-10/100) it became evident that ViT is quite data-intensive and does not perform as well without large-scale pretraining or some other regularizations.

Table 1: Comparison of CNN and ViT Models on Large-Scale Image Classification

Model (Year)	Dataset	Parameters (M)	FLOPs (B)	Notes
ResNet-152	ImageNet-1k	60.2	11.3	Strong CNN baseline
EfficientNet-B7	ImageNet-1k	66	37	Highly optimized CNN
ViT-B/16	ImageNet-21k → 1k	86	17.6	Requires large pretraining
ViT-L/16	ImageNet-21k → 1k	307	63.6	Very high capacity
DeiT-S	ImageNet-1k	22	4.6	ViT trained without large-scale pretraining

This table indicates that there is a difference in performance when using traditional CNNs and different ViT models. When having large datasets, ViTs usually need many parameters and require a lot of computation resources but can be more effective than CNNs. CNNs are more efficient and data-friendly particularly in small scale applications.

3. Challenges of ViT on Small-Scale Datasets

It is important to determine the major issues that restrict the direct use of Vision Transformers to small-scale data before examining the token reduction strategies. ViTs are effective on large data, but have various real-world limitations in the face of scarce data or computing power.

3.1. Data Efficiency

In order to grasp ViT limitations we start by noting that they rely on large-scale datasets. Great reliance on large labeled datasets is a characteristic of Vision Transformers to be trained successfully. ViTs learn more directly than CNNs because ViTs do not apply inductive biases like locality, translation equivariance, etc [11]. ViTs trained directly achieve lower performance on small datasets such as CIFAR-10 and CIFAR-100, except with significant data augmentation, knowledge distillation, or massive pretraining (e.g. ImageNet-21k), compared to CNNs. This is highly inefficient in terms of data, which leads to serious overfitting, especially when training ViTs with millions of parameters on datasets with only tens of thousands of images. Smaller models like DeiT-S mitigate this problem to a certain extent with a more potent regularization mechanism

but still, the main problem is that ViTs require large datasets to be effective at generalization and thus, unless further optimization techniques are implemented, can hardly be used in small-scale classification [12].

3.2. Computational Complexity

The second impediment to small-scale adoption is the cost of self-attention is high. ViTs are a self-attention mechanism, which means that its computational complexity is quadratic in terms of the number of tokens [13]. In the case of an image divided into N patches, the N-head attention operation can be scaled to $O(N^2)$ and thus training and inference are computationally expensive. This is compounded by the fact that the resolutions of the image are higher in which case the number of patches increases exponentially. In small datasets that tend to have low-resolution images, such a computational cost is not as dramatic but nonetheless a problem with the deeper or larger ViT models. As an example, ViT-B/16 and ViT-L/16 have huge memory footprint and long training times, at moderate input dimensions. These costs often constrained ViT experiments on smaller datasets in research settings unless smaller models or training processes were used. Computational overhead is therefore an important issue particularly in resource constrained settings and small input modalities.

3.3. Redundancy in Token Representations

Finally, redundant patch information contributes both to computational inefficiency and overfitting.In case images are partitioned into homogenous patches (e.g. 16×16), a large portion of the patches can be redundant or low-information (e.g. the background or a repeated texture). Although these patches contribute to the final classification result only a little bit, they are still processed by the transformer as complexly as possible. This causes redundant computation and can increase noise in small-scale environments which can further contribute to overfitting [15].The identification of this redundancy encourages the study of methods of reducing tokens (pruning, merging, pooling or learned selection mechanisms). The goals of these methods are dynamic or statical reduction of the number of tokens to allow the model to concentrate on informative regions as well as enhance computational efficiency and generalization.

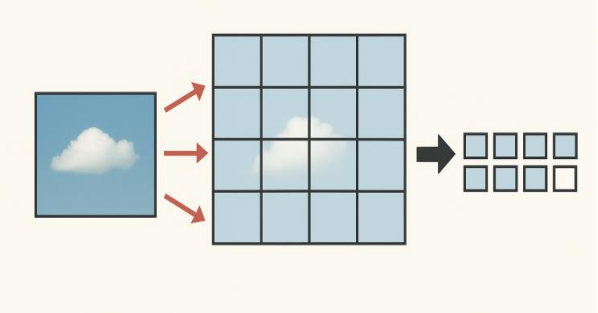

Figure 3: Redundant Tokens in ViT for Small Images

As figure 3 demonstrating how an input image is divided into patches, highlighting multiple patches with nearly identical or non-informative features (e.g., uniform background). Arrows or markers should indicate how redundant tokens contribute to unnecessary self-attention computation. The figure should visually motivate token reduction by showing computational waste and redundancy.

4. Token Reduction Techniques in ViTs

The high token count of image patching was realized before researchers realized the need to contemplate using Vision Transformers for small classification tasks. The need to make things efficient and strong led to the appearance of the strategy of token reduction [16].

4.1. Motivation

In order to explain token reduction, it is first important to explain the necessity of the same.Standard ViTs use a fixed set of image patches, and consider each patch as a token. Nevertheless, these tokens in high density particularly in small images are redundant or low-information areas. Any minimization of tokens, which does not reduce crucial visual material, has two key benefits:

- Improved computational efficiency: Fewer tokens directly reduce the quadratic cost of self-attention.
- Better generalization on small-scale datasets: Removing redundant or noisy patches decreases overfitting and helps the model focus on salient features.

Thus, token reduction provides a principled way to adapt ViTs to small-scale and low-resource settings, motivating several early techniques explored

4.2. Early Approaches

Research focused on pruning, pooling, and hybrid feature extraction approaches.

4.2.1. Dynamic Token Pruning

Pruning methods reduce token count by eliminating less informative patches. The initial NLP transformer pruning methods served as the source of inspiration for dynamic token pruning approaches. Pruning in vision entails the rejection of tokens by using attention scores, gradient-based importance or feature saliency[17]. Adapted studies have proposed a set of explainability-based metrics that determine the tokens that have little influence on the final prediction.

The pruning process typically includes:

- Calculation of importance scores on a token-to-token basis.
- Reduction in tokens to a certain level.
- Transferring the remaining tokens across transformer blocks.

These techniques are less expensive to compute, and may allow reducing overfitting to smaller data sets, but initial algorithms tend to over-prune useful tokens in case the importance estimation is volatile.

4.2.2. Pooling- Based Token Reduction.

The pooling techniques combine several patches into less informative but more valuable tokens.The pooling-based strategies decrease the number of tokens through the merging or aggregation of patches. Early methods included:

- Mean smoothing of neighboring patches.
- The pooling is based on attention, in which the learned importance is used to weigh the tokens.

PatchMerging (applied in early hierarchical transformers) which merges adjacent patches to downsample space [18] . Pooling is computationally easy and it tries to maintain key global data and is therefore appealing when dealing with tiny datasets. Nevertheless, excessive aggressive pooling can obscure fine grains.

4.2.3. Hybrid CNN–ViT Approaches

The information is first extracted with CNNs to get the local features and then tokens are reduced before being introduced to the transformer. Multiple initial works suggested the integration of CNNs and ViTs to enhance the efficiency of data [19]. The CNN layers record local textures and spatial structure, thereby giving a reduced number of the feature maps. These low-resolution feature tokens undergo long-range modeling in a ViT. Approaches often used:

- A CNN stem (e.g., patch embedding resembling ResNet)
- Fewer spatial dimensions which lead to fewer tokens.
- Layers of transformer encoders that are put on compressed tokens.

These hybrid approaches provide a trade-off between CNN inductive bias and transformer flexibility, which is especially needed with small-scale data where ViTs alone are not able to extrapolate.

Table 2: Comparison of Token Reduction Strategies in Early ViT Research

Method Type	Dataset	Tokens Reduced (%)	Accuracy Impact	FLOPs Reduction
Dynamic Token Pruning	CIFAR-100	20–40%	Slight drop or neutral	Moderate
Attention-Based Pooling	CIFAR-10	25–50%	Neutral or slight gain	High
Average Spatial Pooling	CIFAR-100	30–50%	Slight drop	High
Hybrid CNN–ViT Stem	CIFAR-10/100	40–60% (implicit)	Noticeable gain	Moderate

This table summarizes early token reduction efforts, comparing pruning, pooling, and hybrid architectures across small-scale datasets. It includes approximate reductions in tokens, accuracy changes, and computational savings (FLOPs).

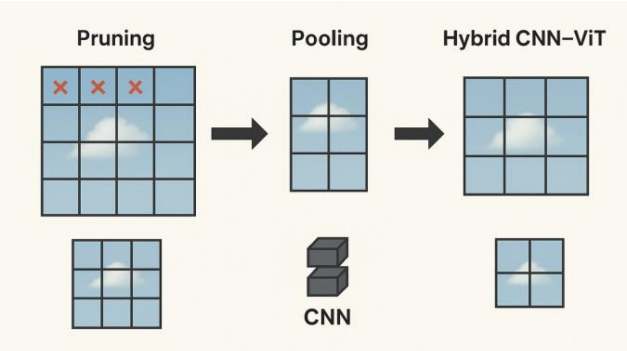

Figure 4: Visualization of Token Reduction Strategies

The figure 4 includes arrows showing reduction steps and side-by-side examples to highlight the differences between methods.

5. Applications on Small-Scale Image Classification

Studies also did a lot of research on the adaptation of Vision Transformers to small-scale datasets. As ViTs were initially used with large-scale datasets like ImageNet-21k, their results on lower-resolution benchmarks had to be changed to enhance the efficiency of the data and regulate model complexity. In this part, the effect of token reduction on performance is evaluated on popular small-scale data sets.

5.1. CIFAR-10 and CIFAR-100

Some of the most popular models of benchmarking lightweight image classification architectures are CIFAR-10 and CIFAR-100. Initial experiments determined that typical ViTs, trained in a fully randomised manner, tend to perform poorly on them due to the fact that the original ViT architecture heavily depends on large scale pretraining [20]. The limited number of images (60,000) is mismatched with the high ability transformer architecture model that frequently overfits or underfits depending on the size and training schedule of the model.

A number of studies were trying to reduce this problem. The use of data augmentation and regularization strategies was also one of the effective approaches. The use of Mixup, CutMix, RandAugment, and stochastic depth techniques proved very useful in ViT training on CIFAR datasets. Knowledge distillation was another powerful approach, with the DeiT model being the first to show that ViTs can competitively be trained without ImageNet-21k pretraining, after being trained on a CNN teacher model. Nonetheless, CIFAR images (32×32 resolution) after augmentation and distillation intrinsically generate a vast quantity of small and highly redundant tokens. A large portion of patches have identical edges, colors and textures, which makes them of little use in discriminative learning. The result of this redundancy is to increase the unnecessary computational overhead and increase the chance of overfitting.

Randomization methods such as: token reduction by removing tokens of low importance, or aggregating patches into a large aggregate representation became a viable way out. These methods decrease the effective number of tokens and at the same time enable ViTs to consider only the most informative parts, furthermore reducing the level of computational complexity. In practice, the loss of tokens in generalization on CIFAR-10 and CIFAR-100 was enhanced by simplifying the input representation and eliminating noise generated by overlapping patches.

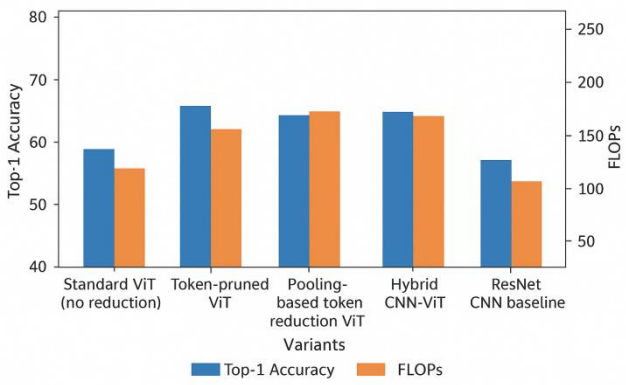

Figure 5: Performance Comparison of ViT Variants on CIFAR-10/100

5.2. Other Benchmark Datasets

In addition to CIFAR datasets, there were a number of small-scale benchmarks that were popularly used to test the lightweight variants of ViT [21]. These are SVHN, TinyImageNet, Flowers-102 and others MNIST-family data. The datasets are different in terms of complexity as well as color distribution and this makes them helpful in studying the interaction between token reduction and dataset characteristics. On SVHN, say, there frequently are large homogeneous backdrops of numbers. This causes token redundancy to be a more visible problem since a large portion of patches is a blank area or a solid color gradient. The token reduction was especially useful in this case, as patients pruned of low information patches prior to transformer application, leading to sample efficiency and less computation. In TinyImageNet (made up of 64x64 images), the higher the resolution, the higher the number of tokens and the higher the pressure on sequence length minimization is. Initial versions of transformers combining CNN feature extractors with ViT encoder versions like CvT and ConViT achieved relatively high results due to the downsampling of the CNN stages and the semantically rich tokens that the CNN stages generated prior to the transformer encoder. This was an effective hybrid form of strategy of token reduction.

In Flowers-102 and MNIST-like images, the accuracy was always boosted by token reduction to reduce noise and limit the model to how many model capacity structures are needed to capture the images. Hybrid or token-reduced ViTs in most cases compared to or outperformed CNNs with a lower number of FLOPs, showing that ViTs could be competitive in small-data settings with enablements.

6. Discussion

The paper summarizes the the review and assesses further implications of applying the token reduction techniques to Vision Transformers on the small-scale image classification. Although token reduction enhances the applicability of ViTs to low-data settings, a number of challenges still exist. The subsections below highlight the advantages, constraints, and future trends as defined by the study trends.

6.1. Benefits of Token Reduction

To Vision Transformer small-scale data, token reduction presents a number of important benefits. The first significant advantage is the remarkable decrease of the cost of computation. As far as the complexity of self-attention is quadratic in the number of tokens, simple reduction in the number of tokens can result in significant increases in efficiency. This can be reduced especially on small images, such as those in CIFAR-10 or SVHN, where the traditional ViT patching procedure yields a large number of small redundant tokens[22]. The second important advantage is that there is enhanced generalization. Redundant or low information patches have a tendency to add noise to the learning process which exacerbates the overfitting probability in small data regimes. This is done by pruning or pooling such tokens to make the model produce visual features that are most meaningful. This performs the effectual regularization of the model and the lack of extraneous complexity, which results in improved results on test data. The literature also continuously found that token-reduced ViTs were more accurate than their full-token counterparts when trained on smaller datasets, and had lower FLOPs and parameters.

6.2. Limitations

In spite of these benefits, there are also a number of restrictions in token reduction. Among the concerns that should be pointed out is the possibility that essential visual information can be erased accidentally. In case reduction strategy is too aggressive or not adaptive enough, then discriminative patches particularly tiny patches with fine details may be cut. This is especially an issue when dealing with CIFAR-100, which has various object types in the images with minor variations. The other constraint is that it requires hyperparameter tuning. The number of tokens to prune, the timing of pruning them and the reduction strategy can greatly affect the performance of the model. Some of the introduced token reduction methods did not have standardized heuristics or theoretical guarantees, necessitating a large amount of manual experimentation. Besides, certain reduction strategies also add an additional complexity to the design of the models, which could neutralize the simplicity that initially made ViTs appealing in the first place. Another difficulty with positional or spatial coherence is that when tokens are removed or merged, there is a challenge to maintain coherence. ViTs are very sensitive to positional embeddings because manipulation of tokens can cause spatial consistency to be broken unless it is treated using special methods, like attention-weighted pooling.

6.3. Future Directions

On the basis of current trends in research, various potential directions were expected in the development. Developing more efficient training strategies on small datasets by ViTs was one of the expected directions. These are better data augmentation pipelines and better regularization schemes, and distillation-based training to bridge the gap between large-data models and small-data tasks further. Reduced adaptive tokens were another area of progress that was predicted. Initial approaches used both fixed pooling or pruning heuristics, and subsequent studies were likely to focus on the attention-based and dynamically learned reduction techniques which would help in retaining valuable information and also reject the redundant ones. Moreover, the field of hybrid architectures was supposed to continue dominating research. The models that interpolate both the inductive bias of CNNs that are proficient in local feature representation and the global contextual modelling of transformers were simulated to offer the best of the two worlds. Initial experiments such as CvT and ConViT had already shown that hybrid models were capable of performing better than plain ViTs on small-scale benchmarks, which implied that additional research on hybrid encoders, multi-stage feature hierarchies, and structured token reduction would become significant directions to achieve the optimal performance.

7. Conclusion

Vision Transformers is an important contribution to the field of image classification, with good modeling ability based on self-attention all over the world. Nonetheless, ViTs have shown to be difficult to apply to small-scale datasets, as they have large data usage, a high cost to compute, and are additionally sensitive to overlapping or noisy patch presentations. ViTs generally are poorer than CNNs at CIFAR-10 and CIFAR-100, which requires architecture adjustments to address the issue. One such direction that has become promising to help in addressing these challenges is token reduction. Such methods as dynamic token pruning and pooling-based reduction and hybrid CNN ViT architectures show a significant promise of enhancing the efficiency and generalization of ViTs in small data environments. The strategies ease the input, reduce the computation costs and counteract overfitting through highlighting of information areas within the image. Initial findings

always indicate that ViTs that are token-reduced can be trained to perform considerably better than standard transformer models and even match CNN baselines on various small image benchmarks.

Although these are the benefits, there are also limitations of token reduction strategies. Fast optimization can eliminate important visual information, positional features can be broken, and most algorithms need large amounts of hyperparameter optimization. Nevertheless research indicates a number of promising directions to pursue in the future such as making selection more adaptively, making regularization methods more powerful, and coming up with more successful hybrid models who can exploit the local inductive biases as well as global attention. On the whole, the review points out that, with a well-planned token reduction approach, Vision Transformers can be successfully transferred to the small-scale image classification task. Further development of effective ViT architectures will allow the wider adoption of their use in low-resource settings, with more research opportunities on small, data-efficient transformer architectures.

References

1. Elngar, A. A., et al. (2021). Image classification based on CNN: A survey. Journal of Cybersecurity and Information Management, 6(1), 18–50.
2. Li, Z., et al. (2021). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems, 33*(12), 6999–7019.
3. Ekman, M. (2021). *Learning deep learning: Theory and practice of neural networks, computer vision, natural language processing, and transformers using TensorFlow.* Addison-Wesley Professional.
4. Yuan, L., et al. (2021). Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
5. Bazi, Y., et al. (2021). Vision transformers for remote sensing image classification. *Remote Sensing, 13*(3), 516.
6. Goel, K., et al. (2020). Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*.
7. Hossain, M. A., & Sajib, M. S. A. (2019). Classification of image using convolutional neural network (CNN). *Global Journal of Computer Science and Technology, 19*(2), 13–14.
8. Swapna, M., Sharma, Y. K., & Prasadh, B. M. G. (2020). CNN architectures: AlexNet, LeNet, VGG, GoogLeNet, ResNet. *International Journal of Recent Technology and Engineering, 8*(6), 953–960.
9. Pazouki, E. (2021). A transformer self-attention model for time series forecasting (pp. 1–10).
10. Dong, H., Zhang, L., & Zou, B. (2021). Exploring vision transformers for polarimetric SAR image classification. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–15.
11. Xu, Y., et al. (2021). VITAE: Vision transformer advanced by exploring intrinsic inductive bias. In *Advances in Neural Information Processing Systems, 34*, 28522–28535.
12. Touvron, H., et al. (2021). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR.
13. Hou, Y., & Zheng, L. (2021). Multiview detection with shadow transformer (and view-coherent data augmentation). In *Proceedings of the 29th ACM International Conference on Multimedia*.
14. Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*
15. Marin, D., et al. (2021). Token pooling in vision transformers. *arXiv preprint arXiv:2110.03860*.
16. Jiang, Z.-H., et al. (2021). All tokens matter: Token labeling for training better vision transformers. In *Advances in Neural Information Processing Systems, 34*, 18590–18602.
17. Wu, Z. (2021). *Video Understanding: Data Privacy, Pipeline Simplicity, and Implementation Efficiency* (Doctoral dissertation).
18. Liu, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
19. Jiang, Z., et al. (2021). Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model. *Computational Intelligence and Neuroscience, 2021*, 7529893.
20. Singla, S., Singla, S., & Feizi, S. (2021). Improved deterministic L2 robustness on CIFAR-10 and CIFAR-100. *arXiv preprint arXiv:2108.04062*.
21. Hassani, A., et al. (2021). Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*.
22. Basart, S. (2021). *Towards Robustness of Neural Networks* (Doctoral dissertation). The University of Chicago.