Norm-Regularized Token Compression in Vision Transformer Networks

Masayuki ISHIKAWA[1], Ryuto ISHIBASHI[2] and Lin MENG[1,†]

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

Abstract

Recent advancements in Vision Transformers (ViT) have seen their applications extend to object detection, image classification, and segmentation, often outperforming Convolutional Neural Networks (CNNs). However, Transformers generally impose higher computational costs compared to CNNs. Various techniques have been developed to reduce computational costs. Among the techniques for reducing the high computational cost, we have enhanced the traditional method to preserve the top K tokens and delete the rest by using four new approaches: Top k-norm, EViT-norm, TNWAF, and TAWNF. Specifically, while the Top K method decides which tokens to be detected based on the weights after attention, the Top K-norm method determines which tokens to delete after normalizing the tokens. The EViT-norm method fuses tokens into a single token by using weights. These weights are derived from contribution rates, which are determined through norm of the tokens that are to be deleted. The TAWNF method integrates the traditional approach of selecting candidate tokens for deletion based on similarity with class tokens using attention with the EViT-norm method. The TNWAF method integrates the TAWNF method, which fuses candidate tokens for deletion using weights from traditional attention following the Top K-norm method. The objective is to reduce information loss and computational costs through token fusion. Our results indicate that, for instance, using the Top K-norm method with the deletion of the lowest 10 tokens, computational costs decreased by 24.7%, with only 0.07% accuracy down. Furthermore, in the TNWAF method, when deleting the lowest 19 tokens, computational costs are reduced by 49.4%, with accuracy decreasing by 0.96%. These methods are effective in pruning models by significantly reducing computational costs without greatly affecting performance. The success achieved in norm token importance metrics in classification tasks suggests potential applicability to other model types. This is a hypothesis we plan to explore in future research. Code is available at https://github.com/maikimilk/ViT-NormReg-Compressor

1. Introduction

Computer Vision (CV) uses artificial intelligence to emulate human visual functions such as object recognition and identification. CV is increasingly utilized across various industries,

The 6th International Symposium on Advanced Technologies and Applications in the Internet of Things (ATAIT 2024) , August 19-22, 2024, Kusatsu, Japan
*These authors contributed equally.
†Corresponding author.
ri0146fe@ed.ritsumei.ac.jp (. M. ISHIKAWA); ri0097fx@ed.ritsumei.ac.jp (R. ISHIBASHI);
menglin@fc.ristumei.ac.jp (L. MENG)
http://www.ihpc.se.ritsumei.ac.jp/ (L. MENG)
0000-0003-4351-6923 (L. MENG)

including autonomous driving, facial recognition, and industrial sorting, and the application in diverse products is expected to grow. Traditionally, tasks such as object detection [1, 2], image classification [3, 4, 5, 6], and semantic segmentation [7] have been accomplished using Convolutional Neural Networks (CNNs). However, more recently, Transformer models[8, 9, 10], originally designed for Natural Language Processing (NLP), have been adapted for CV applications, leading to the development of Vision Transformers (ViT) [11] which are known to surpass the capabilities of CNNs. Despite their advantages, ViTs require significantly more computational resources than CNNs, presenting a substantial challenge for their deployment.

To address these demands, various techniques have been explored to reduce computational costs [12] [13]. Training ViT models from scratch required large datasets and extensive training to achieve convergence of the loss function, making the efficient utilization of ViTs challenging without substantial computational resources. Consequently, both companies and research institutions have been exploring methods to reduce and speed up computations. However, common strategies such as pruning and distillation can lead to information loss and reduced accuracy. In response to these challenges, this study proposes four novel methods that combine traditional approaches with a new strategy where tokens are norm, and the top K most significant tokens are retained. The remaining less significant bottom tokens are fused into a single token using weights derived from a SoftMax function. This approach aims to reduce computational costs significantly without compromising accuracy. The major contributions of this research are as follows:

- Across all pruning levels, our methods, which use normalized contribution rates for token importance, show significant improvements over traditional techniques.
- At the Pruning Level Extra-Large, our TNWAF method achieves a Top-1 Accuracy of 97.52%, slightly better by 0.3% compared to other methods but with a 0.96% decrease compared to the baseline model.
- At the Pruning Level Small, our Top K-norm method achieved a Top-1 Accuracy of 98.41%, with a minor decrease of 0.07% compared to the baseline.

The remaining parts of this paper are organized as follows: Section 2 introduces the related works. In Section 3, we propose our pruning methods. Section 4 shows implementation conditions, dataset, comparison models, evaluation index, and experimental result. Section 4.3 reports discussion and future work. Finally, this paper is concluded in Section 5.

2. Related work

2.1. Vision Transformer

Vision Transformer (ViT) [11] is a model that adapts the Transformer architecture, which has become prevalent in Natural Language Processing (NLP), to handle image data by modifying the input token representation. The tokenization process involves splitting the image into patches, flattening them, and applying a linear projection to obtain D-dimensional vectors. These resulting vectors are referred to as patch embeddings, as shown in Equation (1).

Specifically, each patch x_p is projected onto a D-dimensional space using an embedding matrix **E**, as expressed in the equation:

$$\text{Patch Embedding: } x_{p_i}\mathbf{E} = \text{Linear Projection of } x_{p_i} \tag{1}$$

where x_{p_i} is the flattened representation of the i-th patch and **E** is defined in Equation (1). By treating each patch as a token, the Transformer model can process the image data. The matrix E belongs to the space of real numbers with dimensions:

$$E \in \mathbb{R}^{(P^2 \cdot C) \times D} \tag{2}$$

where P is a dimension of the patch, C is the number of channels, and D is the feature dimension.

A distinctive feature of ViT is the inclusion of a learnable class token, which serves as a special token positioned at the beginning of the input token sequence. The class token is designed to capture the positional relationships within the input. By passing this class token through a Multi-Layer Perceptron (MLP), the model can perform classification tasks. The Transformer encoder, which processes the input tokens, consists of alternating layers of multi-head self-attention and MLP blocks. Normalization layers are applied before each block, and residual connections are implemented after every block. Here, we explain the self-attention mechanism, which is a crucial component of the Transformer model and ViT. The compatibility function between queries and keys is used to compute the weights assigned to the values. The weighted sum of the values constitutes the output. The resulting output vector encodes the relevance and importance between the queries and keys. This computation is represented by Equation (3).

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, and V represent the query, key, and value matrices respectively, and d_k is the dimensionality of the keys. Since the queries and keys are derived from the same data, the self-attention mechanism can capture the relationships between all regions within the image. This provides a broader receptive field than Convolutional Neural Networks (CNNs), allowing for more flexible feature extraction. However, large-scale datasets are required to effectively extract features for training the model. Generally, larger datasets are considered better for pre-training, which can be more computationally expensive compared to CNNs.

2.2. Top K

The Top-K algorithm is a commonly used technique that involves selecting the top K elements from a set of N elements based on a specific criterion. The process begins by defining a criterion for comparison and then evaluating all elements against this criterion. The top K elements are identified and retained, while the remaining elements are discarded. The Top-K algorithm is widely employed in various applications where it is necessary to select the most relevant or significant elements from a larger set. By retaining only the top K elements, this algorithm can effectively reduce the data size, improve computational efficiency, or focus on the most important elements for further processing or analysis. It is important to note that the choice

of the selection criterion plays a crucial role in determining the effectiveness of the Top K algorithm. The criterion should be carefully designed to align with the specific requirements and objectives of the application.

2.3. EViT

EViT [14] is a method proposed to accelerate the ViT model. At the time, existing model compression techniques such as pruning and distillation, could not be directly applied to ViT due to the architectural differences between Convolutional Neural Networks (CNNs) and ViT. Therefore, there was a lack of focus on accelerating ViT models. EViT addresses this by proposing to remove and reorganize unnecessary tokens based on their contribution to the classification results, potentially reducing computational costs. The contribution of each token is determined by the difference between the class token's feature representation after attention and the feature representation of each token. Tokens with smaller feature differences are considered less important and are candidates for removal. This process involves selecting the top k elements and removing the remaining tokens, where k is a predefined value. However, removing tokens can lead to information loss and decreased classification accuracy. To mitigate this potential information loss and accuracy drop, EViT proposes fusing the candidate tokens through a weighted average and reorganizing them into a new sequence, thereby preserving the information content. By selectively pruning tokens at each layer, EViT significantly reduces computational costs.

2.4. Token Merging

ToMe (Tokens-Merging) [15] is a method that achieves acceleration comparable to pruning while maintaining higher accuracy by merging tokens. In each Transformer block, an arbitrary number of tokens can be merged, gradually reducing the number of tokens at each layer. This process leads to improved throughput. However, it is important to note the trade-off between the number of tokens merged and the potential decrease in accuracy. The key advantage of this method is that it can obtain information from the merged tokens, allowing the attention mechanism to determine which tokens should be merged. This enables the reduction of computational costs without sacrificing accuracy. The specific process for deciding which tokens to merge involves using the self-attention mechanism's Query, Key, and Value (QKV) representations. The Key (K) contains information about each token, and the dot product similarity metric between the Keys of different tokens is used to determine the similarity between tokens. The algorithm employed to determine the token similarities and merge decisions is the Bipartite Soft Matching algorithm, which consists of the following five steps:

1. Divide the tokens into two equal-sized groups, $\mathbb{A}$ and $\mathbb{B}$.
2. Draw edges between each token in $\mathbb{A}$ and its most similar counterpart in $\mathbb{B}$.
3. Retain the edges representing the highest similarities, up to a predefined number.
4. Merge the connected tokens by averaging their features or using a similar approach.
5. Recombine the merged tokens from $\mathbb{A}$ and $\mathbb{B}$.

When merging tokens, a fine-tuning process is applied using Equation (4).

$$\text{Attention} = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + \log s\right) \tag{4}$$

where Q and K represent the query and key matrices respectively, d is the dimensionality of the keys, and s is a row vector containing the *size* of each token (number of patches the token represents). This results in the same operations as when a copy of the key exists. Furthermore, when aggregating tokens, such as during token merging, it is essential to always weight them by s.

3. Proposed Method

This paper proposes a method to reduce computational complexity by removing tokens with low contribution scores and merging them, thereby reorganizing the input sequence while minimizing information loss.

3.1. Contribution Score Metric

Unlike EViT, which used the similarity between token embeddings and the class token output from attention, we calculate the contribution score of each token to the overall image by computing the vector norm of the input token embeddings in Equation (5). A smaller vector norm suggests a smaller impact on the learning process. By avoiding the need for a class token, our approach can be applied to various image recognition tasks beyond classification.

The contribution score c_i of token i to the overall image is calculated as the L2 norm of the token embedding:

$$c_i = \|x_{p_i}\|_2 = \sqrt{\sum_{j=1}^{n} x_{p_{ij}}^2} \tag{5}$$

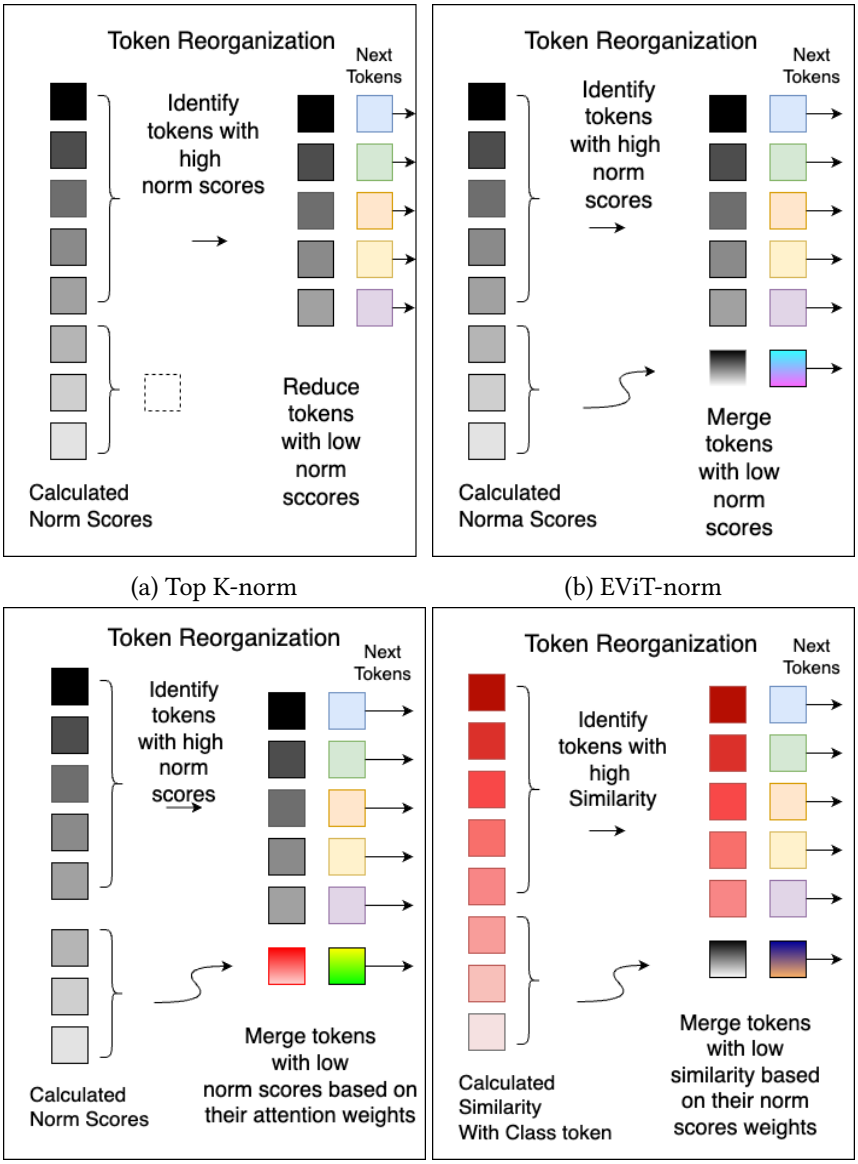
where x_{p_i} is the flattened representation of the i-th patch defined in Equation (1). $x_{p_{ij}}$ is the j-th component of the embedding vector $\mathbf{x}_{p_i}$, and n is the dimension of the token embedding vector.

3.2. Token Removal

Top K-norm Method: In this method, similar to EViT, we employ a Top-K approach for token removal. As shown in Figure 1a. Top K-norm Method involves retaining the top K tokens with the highest contribution scores in Equation (5) as the next input sequence, while the remaining tokens are marked for removal and stored for token merging in other methods.

3.3. Token Merging

EViT-norm method: In this method, the norm results of the tokens designated for removal are input into the SoftMax function in Equation (6) to determine their weights. These weights are then used to compute the weighted average, resulting in the fusion of these tokens into a single token in Equation (7). As shown in Figure 1b. Let x_{new} represent the fused token, x_{pi}



(a) Top K-norm (b) EViT-norm

(c) TNWAF (d) TAWNF

Figure 1: Each Pruning Methods. In the original Vision Transformer (ViT), these techniques are employed within a single transformer block. Specifically, the methods are utilized between the multi-head self-attention and MLP components to reorganize the tokens.

denote the tokens designated for removal, c_i indicates the norm results of these tokens, and w_i denotes the weights obtained from the SoftMax function.

$$w_i = \text{SoftMax}(c_i) = \frac{e^{c_i}}{\sum_{k=1}^{n} e^{c_k}} \tag{6}$$

Where c_i represents the norm value of the i-th token designated for removal, and w_i is the weight for the i-th token calculated by the SoftMax function.

$$x_{\text{new}} = \sum_{i=1}^{n} w_i x_{p_i} \tag{7}$$

In this equation, x_{new} is the fused token calculated as the weighted average of the tokens x_{p_i}, where each token's contribution is weighted by w_i the SoftMax-derived weights.

Topk-Norm-Weighted-Attention-Fusion (TNWAF) method: This method uses the attention-derived similarity between the tokens marked for removal and the class token as weights to compute the weighted average in Equation (8). As shown in Figure 1c.

$$x_{\text{new}} = \frac{\sum_{i=1}^{n} a_i x_{p_i}}{\sum_{i=1}^{n} a_i} \tag{8}$$

Where x_{p_i} represents the i-th token designated for removal, a_i is the weight assigned to x_{p_i}, derived from its similarity to the class token. Additionally, n represents the total number of tokens designated for removal.

Topk-Attention-Weighted-Norm-Fusion (TAWNF) method: This method, similar to EViT, determines tokens for removal based on their similarity to the class token after attention, using the Top K approach. The tokens designated for removal, denoted as x_{pi}, are scored with a norm and then input into the SoftMax function to determine their weights, w_i as shown in Equation (6). A weighted average is computed to create a new token representation x_{new} that incorporates the most relevant features of the tokens, as detailed in Equation (7). As shown in Figure 1d

The resulting vector from the token merging methods is treated as a single token that encapsulates the information from the removed tokens. This method reduces computational complexity while mitigating information loss by removing tokens with minimal impact on the learning process and generating a weighted token that merges information from the removed tokens.

4. EXPERIMENTATION

4.1. Experimental Setup and Materials

The experiments in this paper are conducted on Intel(R) Core Xeon(R) CPU wih a single NVIDIA RTX A6000 GPU for training and inference. AdamW is used as an optimizer. For the learning condition, the number of epochs is set to 100. The maximum value of the learning rate (5e-4) ·B/512.0 is used in other studies [12], [16], where B indicates the batch size, which is set to 32 in this experiment.

Regarding dataset, CIFAR-10 [17] are used for classification tasks. CIFAR-10 is a dataset of object color images, including animals, vehicles, etc., as shown in Figure 2, which are

small 32 × 32 RGB images of 10 classes, with 50,000 training images and 10,000 test images. As data augmentation techniques, RandomResizedCrop at 224 × 224, RandomHorizontalFlip, RandomErasing, and RandAugment [18] are used to increase the variation of the training data.

(a) Airplane (b) Car (c) Bird (d) Cat (e) Reindeer

(f) Dog (g) Frog (h) Horse (i) Ship (j) Truck

Figure 2: CIFAR-10 class images

In terms of comparison modles, ViT-small is used, and vanilla ViT [11] is used as the baseline models. The viT-small model has 12 transformer encoder layers, 384 embedding dimensions, and 6 heads.The teacher model is pre-trained by ImageNet-21k by CIFAR-10 as downstream tasks. In addition, Top K and ToMe [15], EViT [14] are used as other comparison pruning methods. We use a reimagined version of EViT that we created, not the original EViT.

Additionally, the Top-1 Accuracy (%) and FLOPs (G) are used as the evaluation index. Each pruning method is compared based on the vanilla ViT as the base model. When comparing models, pruning amounts are adjusted to achieve comparable computational complexity.

4.2. Experimental Result

Table 1 shows the indices of each model in CIFAR-10. Among the current lightweight approaches for Vision Transformers (ViT), the ToMe method is one of the best-performing techniques. The Top k-norm method achieves competitive performance, closely approaching ToMe, with only a 0.12% down in Top-1 accuracy compared to the baseline at the Pruning Level Small. At the Pruning Level Extra-large, our proposed TNWAF method results in a Top-1 Accuracy of 97.52%, representing a 0.96% decrease compared to the baseline model. For Pruning Level Large, both our Top k-norm and TNWAF methods achieve the highest Top-1 Accuracy of 97.62% among the tested methods. At the Pruning Level Medium, the Top k-norm method achieves the best Top-1 Accuracy of 97.91%. At the Pruning Level Small, the Top k-norm method reaches a Top-1 Accuracy of 98.41%, with only a slight 0.07% decrease from the baseline model. Methods employing norm-based contributions consistently outperform those using similarity to class tokens based on traditional attention.

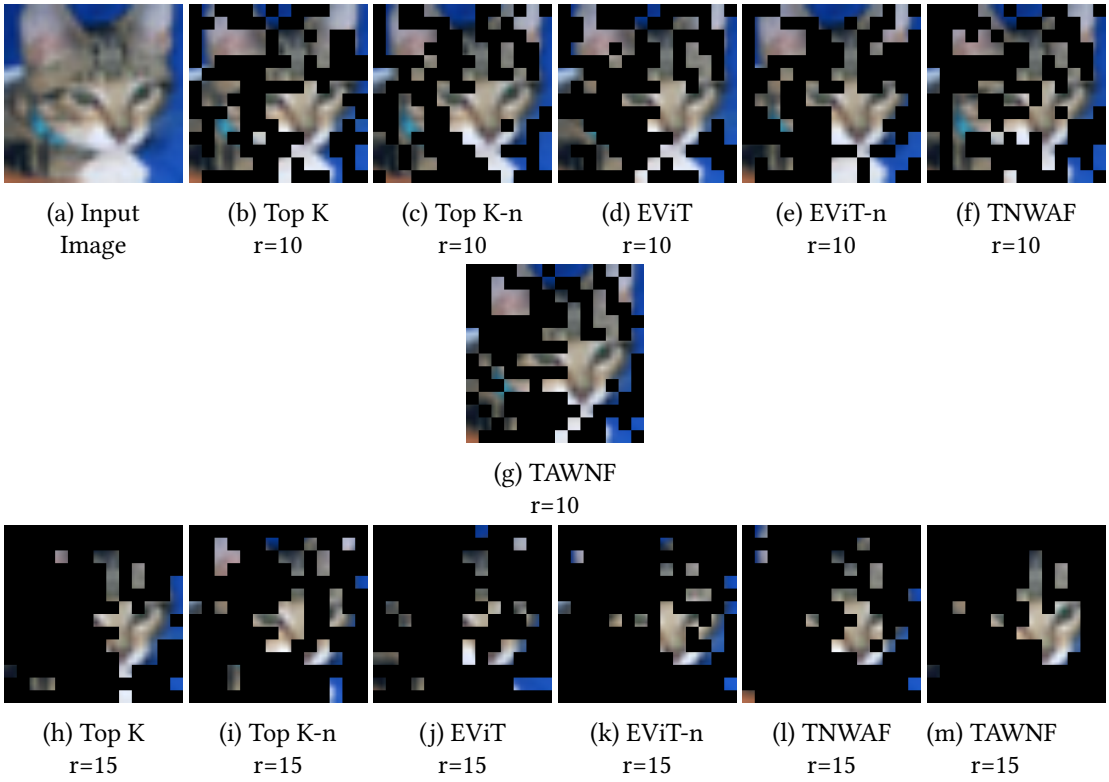
Table 1
Experimental Result

Model	Method	r*	Top-1 Acc. (%)	FLOPs (G)	FLOPs ↓(%)	Params (M)
Pruning Level Extra-large						
ViT-Small	-	-	98.48	4.25	-	21.9
	ToMe	16	98.19	2.16	49.2	21.9
	Top K	18	97.12	2.15	49.4	21.9
	EViT	19	97.09	2.15	49.4	21.9
Ours	**Top K-norm**	18	97.22	2.15	49.4	21.9
	EViT-norm	19	97.22	2.15	49.4	21.9
	TNWAF	19	**97.52**	2.15	49.4	21.9
	TAWNF	19	97.07	2.15	49.4	21.9
Pruning Level large						
	ToMe	14	98.31	2.41	43.3	21.9
	Top K	16	97.44	2.38	44.0	21.9
	EViT	17	97.38	2.38	44.0	21.9
Ours	**Top K-norm**	16	**97.62**	2.38	44.0	21.9
	EViT-norm	17	97.51	2.38	44.0	21.9
	TNWAF	17	**97.62**	2.38	44.0	21.9
	TAWNF	17	97.38	2.38	44.0	21.9
Pruning Level Medium						
	ToMe	12	98.26	2.67	37.2	21.9
	Top K	14	97.65	2.61	38.6	21.9
	EViT	15	97.66	2.61	38.6	21.9
Ours	**Top K-norm**	14	**97.91**	2.61	38.6	21.9
	EViT-norm	15	97.87	2.61	38.6	21.9
	TNWAF	15	97.81	2.61	38.6	21.9
	TAWNF	15	97.70	2.61	38.6	21.9
Pruning Level Small						
	ToMe	8	98.53	3.20	24.7	21.9
	Top K	9	97.98	3.20	24.7	21.9
	EViT	10	98.14	3.20	24.7	21.9
Ours	**Top K-norm**	9	**98.41**	3.20	24.7	21.9
	EViT-norm	10	98.31	3.20	24.7	21.9
	TNWAF	10	98.28	3.20	24.7	21.9
	TAWNF	10	98.07	3.20	24.7	21.9

* The number of reduced tokens.

4.3. Discussion and Future Work

The experimental result demonstrates the effectiveness of using normalization in the calculation of token importance, a key component of our proposed method. In cases of substantial token



(a) Input Image | (b) Top K r=10 | (c) Top K-n r=10 | (d) EViT r=10 | (e) EViT-n r=10 | (f) TNWAF r=10

(g) TAWNF r=10

(h) Top K r=15 | (i) Top K-n r=15 | (j) EViT r=15 | (k) EViT-n r=15 | (l) TNWAF r=15 | (m) TAWNF r=15

Figure 3: Pruned images

reduction (Extra Large), our TNWAF method does not exhibit a decrease in accuracy compared to other methods. Conversely, with minimal token reduction (Small), our Top k-norm method achieves a reduction of FLOPs by 24.7% while experiencing only 0.07% accuracy down compared to the baseline model. These results clearly show the superiority of our methods over traditional techniques that rely on similarity to class tokens based on attention across all pruning levels. This advantage underscores the potential of norm-based approaches in improving model efficiency and performance.

In the discussions presented by Visiaraize, it is apparent that different methodologies lead to varied remaining tokens, as shown in Figure 3. The Top K-norm method, which exhibits the best performance at r=10, predominantly retains tokens associated with salient image features such as ears and eyes while eliminating background tokens (see Figure 3c). This suggests that focusing on key features while disregarding extraneous background elements enhances performance. Conversely, the Top K method results in the removal of significant tokens, such as those of the ears and eyes, compared to the Top K-norm method, which likely contributes to a reduction in accuracy, as illustrated in Figure 3b. Furthermore, at r=15, the TNWAF method, which is the most effective, predominantly omits background elements while preserving tokens representing key image features like the nose, eyes, and ears, as depicted in Figure 3l. This contrasts with other images where tokens representing the eyes, ears, and nose are removed,

or background tokens are retained, underscoring the variability in token preservation across different methods.

In the future, while we eliminate a fixed amount of tokens, future efforts will focus on optimizing the token pruning process by identifying the optimal token removal rate for each layer. Furthermore, given the success achieved by applying norm-based measures to the importance of tokens in classification tasks, we intend to explore whether similar benefits can be obtained in models beyond classification tasks.

5. Conclusion

In this paper, we propose several methods for pruning in Vision Transformers (ViT). These include using normalized contribution rates as metrics of importance for token deletion and applying these rates to compute weights for token fusion. Additionally, we integrate these methods with traditional techniques that are sinilar to attention-based class tokens. The experimental results demonstrate the significance of using normalized contribution rates over the widely used attention-based class token similarity for assessing token importance during deletion. The findings also indicate that the optimal pruning method varies depending on the amount of token deletion. Future experiments will explore whether adaptive token deletion, varying by layer, might be more effective than a fixed quantitative approach.

References

[1] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149. doi:10.1109/TPAMI.2016.2577031.

[2] X. YUE, H. LI, L. MENG, Yolo-sm: a lightweight single-class multi-deformation object detection network, IEEE Transactions on Emerging Topics in Computational Intelligence (2024).

[3] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2015. arXiv:1512.03385.

[4] H. Li, L. Meng, Hardware-aware approach to deep neural network optimization, Neurocomputing 559 (2023) 126808.

[5] H. LI, Z. WANG, X. YUE, W. WANG, H. TOMIYAMA, M. Lin, An architecture-level analysis on deep learning models for low-impact computations, Artificial Intelligence Review 56 (2023) 1971–2020.

[6] X. YUE, H. LI, F. Yoshiyuki, L. MENG, Dynamic dataset augmentation for deep learning-based oracle bone inscriptions recognition, 76, ACM, 2022.4.

[7] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015. arXiv:1505.04597.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, 2023. arXiv:1706.03762.

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

[10] R. Ishibashi, H. Kaneko, N. Nojiri, K. Saho, L. Meng, Optimized vision transformer for dementia diagnosis using micro-doppler radar, in: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2023.
[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. arXiv:2010.11929.
[12] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, 2021. arXiv:2012.12877.
[13] Z. Kong, P. Dong, X. Ma, X. Meng, M. Sun, W. Niu, X. Shen, G. Yuan, B. Ren, M. Qin, H. Tang, Y. Wang, SPViT: Enabling Faster Vision Transformers via Soft Token Pruning, 2022. arXiv:2112.13890.
[14] Y. Liang, C. Ge, Z. Tong, Y. Song, J. Wang, P. Xie, Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations, 2022. arXiv:2202.07800.
[15] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, J. Hoffman, Token Merging: Your ViT But Faster, 2023. arXiv:2210.09461.
[16] H. Yin, A. Vahdat, J. Alvarez, A. Mallya, J. Kautz, P. Molchanov, AdaViT: Adaptive Tokens for Efficient Vision Transformer, 2022. arXiv:2112.07658.
[17] A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images (2009).
[18] E. D. Cubuk, B. Zoph, J. Shlens, Q. V. Le, RandAugment: Practical automated data augmentation with a reduced search space, 2019. arXiv:1909.13719.