2022 26th International Conference on Pattern Recognition (ICPR)
August 21-25, 2022, Montréal, Québec, Canada

Aggregating Global Features into Local Vision Transformer

Krushi Patel†, Andrés M. Bur‡, Fengjun Li†, Guanghui Wang*
† *Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence KS, USA, 66045*
‡ Department of Otolaryngology–Head and Neck Surgery, University of Kansas, Kansas City, Kansas, USA, 66160
* *Department of Computer Science, Ryerson University, Toronto ON, Canada, M5B 2K3*
{krushi92, fli}@ku.edu, abur@kumc.edu, wangcs@ryerson.ca

2022 26th International Conference on Pattern Recognition (ICPR) | 978-1-6654-9062-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPR56361.2022.9956379

***Abstract*—Local Transformer-based classification models have recently achieved promising results with relatively low computational costs. However, the effect of aggregating spatial global information of local Transformer-based architecture is not clear. This work investigates the outcome of applying a global attention-based module named multi-resolution overlapped attention (MOA) in the local window-based transformer after each stage. The proposed MOA employs slightly larger and overlapped patches in the key to enable neighborhood pixel information transmission, which leads to significant performance gain. In addition, we thoroughly investigate the effect of the dimension of essential architecture components through extensive experiments and discover an optimum architecture design. Extensive experimental results CIFAR-10, CIFAR-100, and ImageNet-1K datasets demonstrate that the proposed approach outperforms previous vision Transformers with a comparatively fewer number of parameters. The source code and models are publicly available at: https://github.com/krushi1992/MOA-transformer**

I. Introduction

Transformer-based architecture has achieved tremendous success in the field of natural language processing (NLP) [37] [7]. Inspired by the great success of transformer in the language domain, vision transformer [8] has been proposed and achieved superior performance on the ImageNet dataset. The vision transformer splits the image into patches and feeds into the transformer, the same way as words token in NLP, and passes through several multi-head self-attention layers of the transformer to establish the long-range dependencies.

Unlike the word token, a high-resolution image contains more pixels compared to words in the passage. This leads to an increase in the computation cost as self-attention in the transformer has quadratic complexity. To alleviate this problem, various local attention-based transformers [24] [36] [48] have been proposed with a linear computation complexity. However, all the proposed approaches could not establish long-range dependencies and some of them are very complicated.

To overcome these issues in the local transformers, we develop a very simple module, named multi-resolution overlapped attention(MOA), to generate global features. The proposed module only consists of multiplication and addition operations and is embedded after each stage in the transformer before the downsampling operation. As the module is added only after each stage instead of each transformer layer, it does not add much computation cost and the number of parameters. Our experiments show that aggregating the resultant features of this module to the local transformer establish the long-range dependencies and hence significantly increases the accuracy in contrast to the total number of parameters as shown in Figure 1

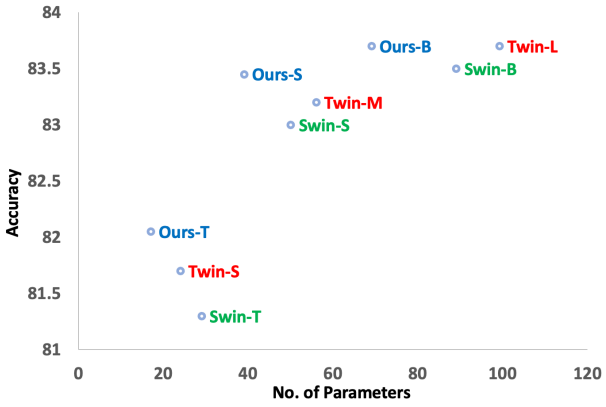

Fig. 1. Graph of accuracy vs. number of parameters for various local transformer-based models. It shows that our all versions of the model: MOA-T, MOA-S, and MOA-B have higher accuracy and comparatively fewer number of parameters.

Our proposed MOA module takes the output generated by the group of local window-based attention as an input. It first converts it to a 2D feature map, and projects it to a new low-dimension feature map. Similar to ViT [8], the projected feature map is divided into a fixed number of patches except for a few modifications. In contrast to ViT [8], the patch sizes of query and key-value are different. The resolution of the patches in the query is the same as the window size used in the local transformer layer. In contrast, the resolution of patches in key-value is slightly larger than the query patch and overlapped. The hidden dimension of the MOA global attention module is kept the same as the previous transformer layer. Therefore, the resultant features are directly aggregated to the output of the previous transformer layer.

Extensive experiments show that keeping the key-value patches slightly larger with overlap to each other leads to significant performance gain due to small information exchange between two neighborhood windows. In short, our method exploits the neighborhood information along with global information exchange between all non-local windows by embedding the proposed MOA mechanism in the local

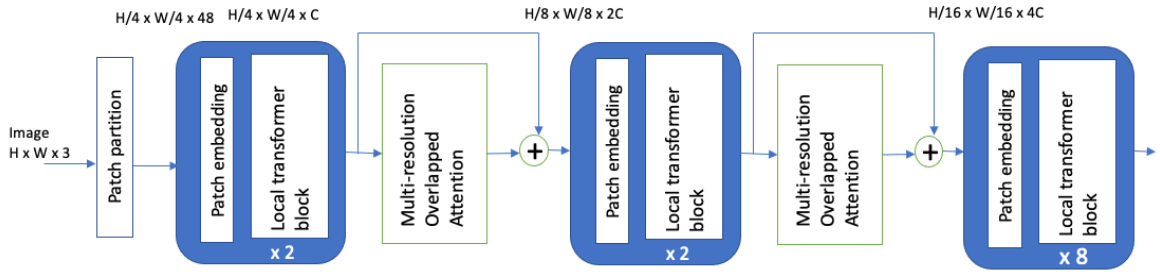

Fig. 2. The architecture of the MOA-T is composed of three stages. Each stage consists of a patch embedding/merging layer and local transformer block along with a global multi-resolution overlapped attention module after each stage except the last stage. In the beginning patch, a partition layer is included to divide the image into a fixed number of patches.

transformer.

The contributions of the proposed approach are summarized as below:

1) We propose a multi-resolution overlapped attention (MOA) module that can be plugged in after each stage in the local Transformer to promote information communication along with nearby windows and all non-local windows.
2) We thoroughly study the impact of global information in local transformer using the proposed MOA module.
3) We investigate the effect of the dimension of essential architecture components through extensive experiments and discover the optimum architecture for image classification.
4) We train the proposed model from scratch on CIFAR-10/CIFAR-100 [19] and ImageNet-1K [6] datasets and achieve state-of-the-art accuracy using a local transformer.

II. Related Work

A. Convolutional Neural Networks

After the revolutionary invention of AlexNet [20], convolutional neural network (CNN) has become a standard network for all computer vision tasks, such as image classification [25] [29], object detection [22], tracking [47], segmentation [13] [28], counting [31], and image generation [41]. Various versions of CNNs have been proposed to improve the performance by making it deeper and/or broader, such as VGG network [33], ResNet [12], Wide-ResNet [45], DenseNet [17], etc. There are also several works proposed to make it more efficient by modifying the individual convolutional layer, such as dilated convolution [42], depth-wise separable convolution [4], group convolution [20], etc. In our work, we employ the convolutional layer along with the transformer layer to reduce the overall dimension of the feature map. Our experiments show that the combination of convolutions and multi-head attention increases the performance.

B. Self Attention in CNN

Self-attention mechanisms have become ubiquitous in the field of computer vision tasks. Various works [10] [39] [2] [40] [32] [9] [49] [26] have been proposed that utilize either channel-based or position based self-attention layers to augment the convolution network. Non-local network [39] and PSANet [49] model the spatial relationship between all the pixels in the feature map and are embedded the attention module after each block in CNN, whereas SENet [16] establishes a channel relationship in the convolution network by squeezing the features using global average pooling. CBAM [40], BAM [27] and dual attention network [9] employ both channel and position based attention mechanisms separately, then combine the resultant features from both attention modules using either element-wise addition or concatenation and uses the resultant features into convolution output after each stage, whereas GCNet [2] combines SENet [16] and non-local network [39] together and propose the hybrid attention mechanism that aggregates the information of both channel and spatial relationships in the same attention module.

C. Vision Transformers

Similar to AlexNet, vision Transformer (ViT) [8] has changed the perspective of researchers towards solving computer vision problems. Since then, many vision transformer-based networks have been proposed to improve accuracy or efficiency. The ViT needs to be pre-trained on large datasets such as JFT300M [34] to achieve high performance. DeiT [35] solves this problem by student-teacher setup, substantial augmentation, and regularization techniques. To train the transformer on the mid-sized dataset like ImageNet-1K from scratch, the token-to-token vision transformer [43] recursively aggregate neighboring tokens (patches) into one token (patch) to reduce the number of tokens. A Cross-ViT [3] comes up with a dual branch approach with multi-scale patch size to produce robust image features and pyramid vision Transformer (PVT) [38] introduces a multi-scale-based spatial dimension design similar to FPN [23] in CNN and demonstrated good performance. Furthermore, PVT introduced a spatial reduction in key to reduce the computation cost in multi-head attention.

Various local attention-based transformers have been introduced to alleviate the quadratic complexity issues [36] [24] [48]. The HaloNet [36] introduces the idea of a slightly larger window of key than the query in a local attention mechanism and proves its effectiveness through various experiments. In our model, the key is also calculated using a slightly larger

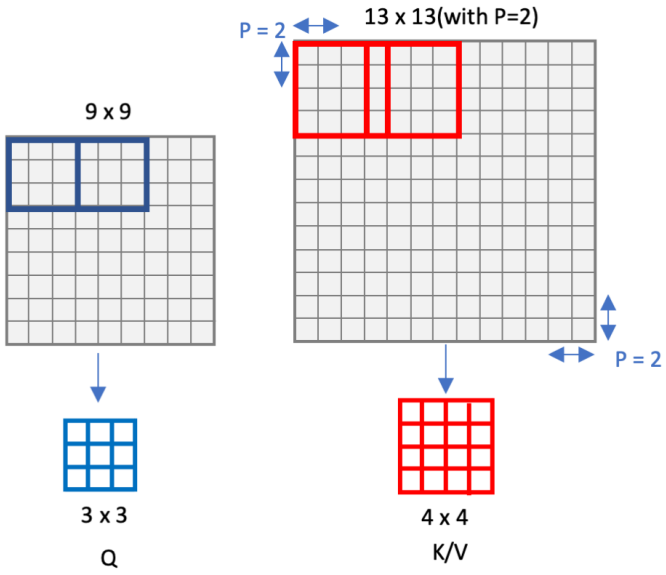

Fig. 3. Patch creation for query embedding is shown in the blue, and key/value is shown in the red for feature map size 9×9 and window size 3×3. Blue patches have the same size as the window and are non-overlapped to each other. In contrast, red patches are larger and slightly overlapped with each other. Appropriate padding is applied while creating the key-value patches.

patch, but in the context of global attention, the idea of a larger key is different from the HaloNet. A swin Transformer [24] proposes a non-overlapping window-based local self-attention mechanism to avoid quadratic complexity and achieve comparable performance and aggregated nested Transformer [48] come with the multi-scale approach with block-aggregation mechanism after each stage.

Some Transformer-based works have been proposed to utilize both local and global features [11] [5]. A Transformer in Transformer (TNT) [11] further divides the local patches (visual sentences) into smaller patches (visual words). The MHA on visual words embedding is calculated and aggregated to the sentence embedding to establish the global relationship. The twin Transformer [5] is quite the same as ours. However, global attention is applied after each local Transformer layer, increasing the computation cost significantly. In contrast, we apply it after each stage, and we have slightly larger and overlapped patches in key in multi-head attention. The proposed network efficiently utilizes global information in the local transformer and achieves higher accuracy than the above-mentioned transformer-based models.

III. Proposed Method

We aim to provide global information exchange across all windows in the local transformer by increasing the minimal computation cost and a number of parameters. An overview of our proposed model is shown in Figure 2, which shows MOA module after each stage. All stages have a similar architecture design, including patch merging layer and local transformer block except the first stage. The first stage consists of patch partition, linear embedding layer, and local transformer block. Our global MOA module is applied between each stage before the patch merging layer.

Specifically, the model takes an RGB image as an input and splits it into fix number of patches. Here each patch is treated as a token. In our experiment on the ImageNet dataset, we set the patch size to 4×4, which leads to $4 \times 4 \times 3 = 48$ feature dimensions for each patch. These row features are projected to a specific dimension C using the patch embedding layer in the first stage. The resultant features are then passed through consecutive stages consisting of patch merging layer, local transformer block, and MOA module in-between each stage. Unlike swin Transformer [24], our Transformer block employs the same self-attention mechanism as ViT [8] without any shifted window approach. Similar to swin Transformer, the number of tokens is reduced, and the output dimension is doubled in the patch merging layer after each stage. For example, the resolution after the first, second and third stage is $\frac{H}{2} \times \frac{W}{2}$, $\frac{H}{4} \times \frac{W}{4}$, and $\frac{H}{8} \times \frac{W}{8}$, respectively. The average pooling layer is inserted at the end of the last stage, followed by a linear layer to generate a classification score. The detailed explanation of each element of architecture are as follows:

A. Patch embedding layer

It is a basic linear embedding layer applied to the row features of patches to project it to a specific dimension C.

B. Patch merging layer

Patch merging layer reduces the number of tokens by concatenating the features of 2×2 neighboring patches and doubles the number of hidden dimensions by applying a linear layer on the concatenated 4C - dimensional features.

C. Local Transformer Block

The local transformer block consists of a local window-based standard multi-head attention module, followed by a two-layer MLP with GELU non-linearity. A layer norm is used before each multi-head attention module and each MLP with residual connection after each module.

D. Multi-resolution Overlapped Attention Block

To utilize the advantage of global information in local transformer, we apply a global attention module named multi-resolution overlapped attention (MOA) in-between each stage. The architecture of the MOA mechanism is the same as the standard multi-head attention except for a few modifications. Similar to standard MHA, it first divides the feature map into the fixed size of patches. However, unlike the standard MHA, patches for generating key and value embeddings are slightly larger and overlapped, while the patches for query embedding are non-overlapped as shown in Figure 3.

As shown in the Figure 3, the input to MOA block is of size $W \times H \times$ hidden dim, Where $W = \frac{W}{2}$, $\frac{W}{4}$ or $\frac{W}{8}$, $H = \frac{H}{2}$, $\frac{H}{4}$ or $\frac{H}{8}$, and hidden dim = 96, 192, or 384. Calculating query, key, and value embeddings directly from the input is quite expensive in computation. For example, in context to the ImageNet dataset, the feature map size of the

input to MOA block after the first stage is $56 \times 56 \times 96$. Deriving query embedding directly from the input feature with a patch size 14 will lead to the resultant feature of dimension $14 \times 14 \times 96 = 18816$. Therefore, we first reduce the hidden dimension with factor R by applying 1×1 convolution, which reduces the computation cost. The resultant feature dimension after applying the convolution is $H \times W \times \frac{hiddendim}{R}$. This leads to feature size in one query patch is $14 \times 14 \times \frac{hiddendim}{R}$, which is projected to the one-dimensional vector of size: $1 \times 1 \times hiddendim$. The total number of the query is $\frac{H}{14} \times \frac{W}{14}$. Similarly, the key and value vector are projected, but the patch size is slightly larger than the query as shown in Figure 3. In our model, we set the key-value patch size to 16. Therefore, the number of key-value will be according to the equation: $(\frac{H-16+(2\times padding)}{stride} + 1, \frac{W-16+(2\times padding)}{stride} + 1)$. Multi-head attention is applied to this query, key, and value embedding, followed by two-layer MLP with GELU non-linearity in between. Similar to the Transformer block, layer norm is applied along with residual connection after each MOA module. At last, on the resultant features, 1×1 convolution is applied, followed by broadcast addition of resultant features with the output of the previous transformer block, which contains the local information.

E. Relative Position Index

We use relative position bias $B \in R^{M^2 \times N^2}$, as used by [1] [15] [14] [30] , in the heads of both local and global attentions during similarity computation:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + B)V \quad (1)$$

where $Q \in R^{M^2 \times d}$ is a query matrix, $K, V \in R^{N^2 \times d}$ are the key and value matrices; d is the hidden dimension, M^2 is the total number of patches in the queue and N^2 is total the number of patches in the key.

F. Architecture Detail

By following the previous works [24] [5], we build three versions of the model: MOA-T, MOA-S, and MOA-B for the ImageNet dataset and only two versions of the model: MOA-T and MOA-B for the CIFAR -10/100 dataset as it is quite smaller. Table I shows the architecture configurations for the CIFAR and the ImageNet datasets. In the CIFAR based models, both MOA-T and MOA-B contain the same number of Transformer layers: 12, but have a different number of hidden dimensions. In context to the ImageNet based models, the total number of layers for MOA-T and MOA-S is 12 and 24 respectively, but the hidden dimension is kept the same, whereas MOA-S and MOA-B have the same number of Transformer layers: 24, with contrast hidden dimensions 96 and 124 respectively.

IV. Experimental Evaluations

We verify our model through extensive experiments on CIFAR-10/CIFAR-100 and ImageNet-1K for image classification. We design three architecture versions: MOA-T, MOA-S, and MOA-B, for the classification tasks.

A. CIFAR-10/100 Results

CIFAR-10 and CIFAR-100 datasets consist of 50,000 training and 10,000 test images of resolution 32×32 with the total number of classes 10 and 100, respectively. We train the network for 300 epochs using AdamW [18] optimizer with an initial learning rate of 0.009 and weight decay of 0.05. We utilize a cosine decay learning rate schedular along with 20 warm-up epochs. We implemented two models: MOA-T and MOA-B for the CIFAR dataset with total batch-size 128 and stochastic drop-rate 0.2 [21].

Table II shows the performance of our model on the CIFAR-10 and CIFAR-100 datasets. We presented only two models with the same number of layers but with different hidden dimensions for this dataset. As shown in the table, it can be seen that both models outperform all the previous Transformer-based models by a significant amount. It improves the performance by 0.59% and 0.98% on CIFAR-10 and 0.56% and 0.23% on CIFAR-100 for the Tiny and Base models, respectively, compared to Swin Transformer. For the Base model, our model achieves state-of-the-art accuracy on local vision Transformer with a comparatively fewer number of parameters and GFLOPs. The accuracy of other models is reported by training the models from scratch with the same training setting reported in the papers [24] [35] [38].

B. ImageNet Results

ImageNet-1K dataset consists of around 1.28M training images and 50K validation images with 1000 classes. We resize all the images to the resolution 224×224 during training. We follow the same training technique, like Swin and Twin, and train the network for 300 epochs using AdamW [18] optimizer with a cosine learning rate schedular and 20 warmup epochs. We keep the batch-size 128 for MOA-T and 64 for MOA-S and MOA-B models per GPU. We employ a total of four GPUs together during training leads to a total batch-size 512 for MOA-T and 256 for MOA-S and MOA-B models. We utilize the same augmentation technique used by [24] such as a mixture of cutmix [44] and mixup [46] and regularization technique stochastic drop rate. We set the drop rate [21] of 0.2, 0.3, and 0.5 respectively for MOA-T, MOA-S, and MOA-B.

Table III shows our model's result and a similar Transformer-based model on the ImageNet-1K classification task. Our proposed models: MOA-T, MOA-S, and MOA-B, achieve higher accuracy than most of the Transformer-based models with significant parameter reduction. MOA-T outperforms Twin-S and Swin-T by 0.34% with around 22% fewer parameters. Our MOA-S improves the performance by 0.5% and 0.3% compared to Swin-S and Twin-M respectively, even with the lower batch size during training. Our MOA-B achieves the state-of-the-art accuracy of 83.7% on ImageNet-1K with comparatively fewer parameters with a smaller batch size than the remaining vision transformers. Our model increases the computation cost by a negligible amount, but the

TABLE I
MODEL CONFIGURATION FOR CIFAR/IMAGENET DATASET

Model	Dataset	Input-Size	Window-Size	No. of Layers	No. of Heads	Hidden Dim	Patch -Size
MOA-T	CIFAR	32×32	4×4	[2, 2, 6, 2]	[3, 6, 12, 24]	[96, 192, 384, 768]	1
MOA-B	CIFAR	32×32	4×4	[2, 2, 6, 2]	[4, 8, 16, 32]	[128, 256, 512, 1024]	1
MOA-T	ImageNet	224×224	14×14	[2, 2, 8]	[3, 6, 12]	[96, 192, 384]	4
MOA-S	ImageNet	224×224	14×14	[2, 2, 20]	[3, 6, 12]	[96, 192, 384]	4
MOA-B	ImageNet	224×224	14×14	[2, 2, 20]	[4, 8, 16]	[128, 256, 512]	4

TABLE II
RESULTS ON CIFAR - 10/100

Model	CIFAR-100(%)	CIFAR-10(%)	Parameters
Deit-T	70.33	89.2	5M
PVT-T	72.80	91	13M
Swin-T	78.07	94.41	27.5M
MOA-T	**78.63**	**95**	30M
DeiT-B	71.54	93	85M
PVT-B	70.1	89.87	61M
Swin-B	78.45	94.47	86.7M
MOA-B	**78.68**	**95.05**	53M

TABLE III
RESULTS ON IMAGENET-1K

Model	Accuracy(%)	Parameters	GFLOPs
Deit-Small/16	79.9	22.1M	4.6
CrossViT-S	81.0	26.7M	5.6
T2T-ViT-14	81.5	22M	5.2
PVT-Small	79.8	24.5M	3.8
TNT-T	73.9	6.1M	1.4
Twins-PCPVT-S	81.2	24.1M	3.8
Swin-T	81.3	29M	4.5
Twins-SVT-S	81.7	24M	2.9
MOA-T	**82.05**	17M	4.8
T2T-ViT-19	81.9	39.2M	8.9
PVT-Medium	81.2	44.2M	6.7
TNT-S	81.5	23.8M	5.2
Twins-PCPVT-B	82.7	43.8	6.7
Swin-S	83.0	50M	8.7
Twins-SVT-B	83.2	56M	8.6
MOA-S	**83.5**	39M	9.4
ViT-Base/16	77.9	86.6M	17.6
Deit-Base/16	81.8	86.6M	17.6
T2T-ViT-24	82.3	64.1M	14.1
CrossViT-B	82.2	104.7M	21.2
PVT-Large	81.7	61.4M	9.8
TNT-B	82.9	65.6M	14.1
Swin-B	83.3	15.4M	15.4
Twins-SVT-L	**83.7**	99.2M	15.1
MOA-B	**83.7**	68M	16.2

performance improvement and parameter reduction are highly rewardable.

V. ABLATION STUDY

In this section, we conduct ablation experiments to understand the effect of the dimension of each component, such as window size, the overlapped area between the key-value patches, and the reduction factor in global attention, in our model. We employ the Tiny model to perform all ablation experiments, and all the experiments are performed either on CIFAR-100 or ImageNet dataset. The training configurations remain the same as reported in the experiment section.

TABLE IV
RESULTS WITH DIFFERENT WINDOW-SIZE ON IMAGENET

Window-Size	Dataset	No. of Stage	Accuracy	Parameters
2×2	CIFAR -100	4	76.04	29.7M
4×4	CIFAR-100	4	78.61	30M
8×8	CIFAR-100	3	76.02	16M
7×7	ImageNet	4	81.4	31M
14×14	ImageNet	3	82.07	17M
28×28	ImageNet	2	78.2	6M

A. Window-size

The sequence length of the local-Transformer is one of the essential factors on which computation cost relies. As the sequence length increases, computation cost in the self-attention mechanism increases as well. In a local vision Transformer, sequence length depends on the window size. There is always a trade-off between the accuracy and computation cost based on the sequence length. We perform experiments with various window sizes in our model and find that 4×4 and 14×14 window size works well on CIFAR-100 and ImageNet datasets, respectively, as shown in Table IV. Furthermore, we remove the stages where the window size is greater than the feature map size to significantly reduce the number of parameters.

B. Overlapped Portion

To initiate the neighborhood information transmission, we propose to use slightly larger and overlapped keys. To investigate the effect of the portion of the overlapped area, we perform experiments with different percentages of overlapped portions in keys as shown in Table V. It can be seen from the results that the performance is increased in terms of accuracy as the percentage decreases, which means only slight information exchange between the neighborhood windows are required to improve the performance. Furthermore, fewer overlapped portions decrease the sequence length, which reduces the number of parameters and GFLOPs.

C. Reduction

Before the MOA global attention, the hidden dimension is reduced to decrease the number of parameters and computation

TABLE V
RESULTS ON CIFAR-100 WITH DIFFERENT PERCENTAGE OF THE OVERLAPPED PORTION

% Overlap	Accuracy	Parameters
17%	78.63	30.05M
33%	78.52	30.06M
50%	78.38	30.08M
66%	78.38	30.59M

TABLE VI
RESULTS WITH DIFFERENT WINDOW-SIZE ON CIFAR-100

Reduction	Accuracy	Parameters
8	78.38	31.67M
16	78.34	30.59M
32	78.63	30.06M
64	78.51	29.78M
num-heads	78.41	31.43M

TABLE VII
SIGNIFICANCE OF GLOBAL ATTENTION AND OVERLAPPED PATCHES

Model	Accuracy	Parameters
Without Global	75.56	27M
With Global (ViT)	78.34	30.59M
With Global (Ours)	78.63	30.06M

cost. Table VI shows the performance of our model with various values of R. From the result, it is evident that R = 32 achieves the best result with comparatively less number of parameters and computation cost than a smaller value of R.

D. Effect of Overlapped Key-Value

To verify the effect of overlapped and larger key-value patches, we train the model without overlapping patches and compare the results. Furthermore, we also conduct an experiment without applying global attention in-between each stage to verify the significance of global information exchange. From the result in Table VII, it can be seen that including global attention and overlapped key-value patches achieve the best performance.

VI. CONCLUSION

The paper has investigated the effect of aggregating global information in local Transformer after each stage and neighborhood pixel information transmission. We have also proposed a multi-resolution overlapped attention (MOA) module that can be plugged in after each stage in the local transformer to promote information communication along with nearby windows. Our results show that both types of features: global and local, are crucial for image classification. As a result, exploiting both features leads to significant performance gain on the standard classification datasets such as CIFAR10/100 and the ImageNet with comparatively fewer parameters.

REFERENCES

[1] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, J. Gao, S. Piao, M. Zhou *et al.*, "Unilmv2: Pseudo-masked language models for unified language model pre-training," in *International Conference on Machine Learning*. PMLR, 2020, pp. 642–652.

[2] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[3] C.-F. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," *arXiv preprint arXiv:2103.14899*, 2021.

[4] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[5] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *arXiv preprint arXiv:2104.13840*, 2021.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[10] K. Gajurel, C. Zhong, and G. Wang, "A fine-grained visual attention approach for fingerspelling recognition in the wild," *arXiv preprint arXiv:2105.07625*, 2021.

[11] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *arXiv preprint arXiv:2103.00112*, 2021.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] L. He, J. Lu, G. Wang, S. Song, and J. Zhou, "Sosd-net: Joint semantic object segmentation and depth estimation from monocular images," *Neurocomputing*, vol. 440, pp. 251–263, 2021.

[14] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3588–3597.

[15] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3464–3473.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-100 (canadian institute for advanced research)." [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[21] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," *arXiv preprint arXiv:1605.07648*, 2016.

[22] K. Li, M. I. Fathan, K. Patel, T. Zhang, C. Zhong, A. Bansal, A. Rastogi, J. S. Wang, and G. Wang, "Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations," *arXiv preprint arXiv:2104.10824*, 2021.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.
[25] W. Ma, X. Tu, B. Luo, and G. Wang, "Semantic clustering based deduction learning for image recognition and classification," *Pattern Recognition*, vol. 124, p. 108440, 2022.
[26] W. Ma, T. Zhang, and G. Wang, "Miti-detr: Object detection based on transformers with mitigatory self-attention convergence," *arXiv preprint arXiv:2112.13310*, 2021.
[27] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.
[28] K. Patel, A. M. Bur, and G. Wang, "Enhanced u-net: A feature enhancement network for polyp segmentation," *arXiv preprint arXiv:2105.00999*, 2021.
[29] K. Patel and G. Wang, "A discriminative channel diversification network for image classification," *Pattern Recognition Letters*, vol. 153, pp. 176–182, 2022.
[30] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.
[31] U. Sajid, X. Chen, H. Sajid, T. Kim, and G. Wang, "Audio-visual transformer based crowd counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2249–2259.
[32] U. Sajid and G. Wang, "Towards more effective prm-based crowd counting via a multi-resolution fusion and attention network," *Neurocomputing*, 2021.
[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[34] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.
[35] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.
[36] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 894–12 904.
[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[38] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.
[39] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
[40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
[41] W. Xu and G. Wang, "A domain gap aware generative adversarial network for multi-domain image translation," *IEEE Transactions on Image Processing*, vol. 31, pp. 72–84, 2021.
[42] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
[43] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv preprint arXiv:2101.11986*, 2021.
[44] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.
[45] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.
[46] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
[47] T. Zhang, X. Zhang, Y. Yang, Z. Wang, and G. Wang, "Efficient golf ball detection and tracking based on convolutional neural networks and kalman filter," *arXiv preprint arXiv:2012.09393*, 2020.
[48] Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. O. Arik, and T. Pfister, "Nested hierarchical transformer: Towards accurate, data-efficient and-interpretable visual understanding," *arXiv preprint arXiv:2105.12723*, 2021.
[49] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 267–283.