Bidirectional Local-to-Global Attentions for Visual Representation

Xuan-Thuy Vo, Duy-Linh Nguyen, Adri Priadana, and Kang-Hyun Jo

Department of Electrical, Electronic and Computer Engineering,
University of Ulsan, Ulsan (4460), South Korea
Email: xthuy@islab.ulsan.ac.kr;
{ndlinh301,priadana3202}@mail.ulsan.ac.kr; acejo@ulsan.ac.kr

Abstract. Vision Transformers achieved outstanding performances across vision tasks due to their generalization ability to large models and datasets. As a main component of the Transformer, the self-attention layer has high flexibility in capturing long-range dependencies from input tokens. Recent advancements in visual backbones have jointly modeled local and global features via the integration of convolution and self-attention. However, the existing methods do not fully investigate bidirectional interactions between local and global features. This paper introduces Bidirectional Local-to-Global (BL2G) attention for capturing local and global features as well as information exchange between two features. First, BL2G utilizes window self-attention to extract local features. Modeling long-range dependencies from high-resolution image tokens can result in high memory and computation. To mitigate this problem, this work groups image tokens together by exchanging information with a fixed number of learnable group tokens. Then, grouped features are fed into the spatial MLPMixer to model the global context. Second, bidirectional interaction across local and global branches is performed to provide complementary information. Typically, in the local-to-global branch, each local query attends to global-grouped tokens, and each global query interacts with local image tokens in the global-to-local branch. These two designs are achieved by efficient cross-attention layers. To verify the effectiveness of the proposed method, experiments are conducted and evaluated on the benchmark dataset, ImageNet-1K image classification, MS-COCO object detection and instance segmentation, and ADE20K semantic segmentation. As a result, BL2G with **10.1M** and **1G** FLOPs achieves **79.1%** of Top-1 accuracy, which outperforms the recent ConvNets and Vision Transformers with similar costs. When transferring the BL2G Transformer to dense prediction tasks, the proposed method achieves comparative performances with previous methods.

Keywords: Vision Transformer · Local-to-Global Learning · Visual Representation.

1 Introduction

Transformer [27] was originally designed for machine translation and attained remarkable performance in both data modeling and efficiency. In the vision field,

DETR [1] integrates Transformer decoders into the detection head to model the relationship between image features and object queries. ViT [7] fully employs Transformer encoders for image classification and achieves promising performances compared to ConvNets. From this milestone, many methods are proposed to improve ViT in both efficiency and accuracy.

As a core component of ViT, self-attention captures long-range dependencies from input tokens. However, self-attention has quadratic complexity with token lengths. When transferring ViT models to dense prediction tasks, it creates huge computational costs. Recent methods try to mitigate this issue by introducing spatial reduction attention [29, 30, 34, 35] and window self-attention [17, 6]. In spatial reduction attention, each query attends to down-sampled key and value tokens. Although this design can reduce the computational cost to $\frac{N^2}{r^2}$ (N is a number of tokens and r is the reduction ratio), relevant regions are ignored while unimportant regions are still kept. Window self-attention [17] performs attention on non-overlapped windows and requires the cyclic shift operation to communicate information across windows. Window-based Vision Transformers [17, 6, 2, 25, 33] have improved the efficiency of ViT where self-attention results in linear complexity with token length. Spatial reduction attention can model global interactions among tokens from coarse features, while window self-attention captures local features inside windows. Both local and global information are complementary. Leveraging these two features into model blocks results in better feature representation and modeling ability [3, 9, 4, 20]. However, these methods only extract local-to-global features, and the bidirectional interaction between two features is further designed to improve performances. Based on the observation, this work promotes bidirectional local-to-global attention for visual representation.

In this paper, following methods [17, 3], window self-attention is used to model short-range dependencies between input tokens. Instead of extracting global information from down-sampled input features, this paper defines learnable group tokens that exchange information with image tokens via a cross-attention layer. Otherwise, image tokens are grouped into a fixed number of learnable tokens that are much smaller than image tokens (e.g., 8 tokens). Therefore, feeding grouped tokens into Spatial MLPMixer can result in global information at a low computational cost. To achieve bidirectional interaction across local and global features, two cross-attention layers with different $\{q, k, v\}$ pairs are adopted to efficiently exchange information between two branches. Specifically, in local-to-global modeling, local features are set as queries, and keys and values are taken from global-grouped tokens. Similarly, in the global-to-local branch, each global query interacts with local image tokens. Based on this design, information exchange between two features is fully captured.

From the bidirectional local-to-global attention layer, the BL2G network is introduced by stacking the number of attention layers across four stages in a hierarchical manner. Extensive experiments are conducted to clarify the effectiveness of the proposed BL2G network. For image classification, BL2G is trained and evaluated on the benchmark dataset, ImageNet-1K. With 10.1M parame-

ters and 1 GFLOP, BL2G achieves 79.1% of Top-1 accuracy, which surpasses previous methods under similar computational costs. For object detection and instance segmentation, BL2G is fine-tuned and evaluated on the MS-COCO dataset using common detectors, RetinaNet, and Mask R-CNN. As a result, BL2G outperforms existing backbones with smaller computational costs.

2 Related Works

2.1 Vision Transformers

Transformer [27] views words as tokens and proposes the Transformer encoder and decoder to capture long-range dependencies from long sequences with weak inductive biases. However, the self-attention layer in the Transformer encoder has quadratic complexity with the token lengths. When adopting self-attention to visual tasks and treating a pixel as a token, the models suffer large memory access and computational costs due to the high input resolution. To address this issue, ViT [7] defines a 16×16 patch as a token via 16×16 convolution with stride 16, and leverages the Transformer encoder into the non-hierarchical network to extract relations across patches. With these designs, ViT establishes a new paradigm in modeling image and video inputs while achieving competitive performances with ConvNets. Remarkably, Transformer architecture can be applicable to large language models, large vision models, and multimodal models due to their strong generalization ability and high flexibility.

Self-attention captures global information from the input token, yet it has high costs and lacks inductive biases such as locality and translation invariance. PVT [29] attempts to reduce costs and proposes spatial reduction attention that sub-samples key and value tokens. PVTv2 [30] enhances inductive biases by adding convolution to the MLP layer. Swin Transformer [17] limits attention inside non-overlapped windows and requires the additional cyclic shift to exchange information across windows. Swin Transformer attains high efficiency, while the model stacks more layers to enlarge receptive fields slightly. Recent methods try to capture global receptive fields and achieve linear complexity by proposing window expanding [6, 10], window shuffling [25, 28], window shifting [17, 31], and window sliding [26, 2, 33, 21, 11].

2.2 Local-to-Global Attentions

Existing methods combine the strengths of convolution and self-attention to build hybrid networks. Strong inductive biases of convolution and high modeling capabilities of self-attention are integrated into each layer or stage of the hybrid networks. EdgeViT [20] extracts local and global features via sequential composition of depthwise convolution and spatial reduction attention. Twins [3] replace shifted window attentions in Swin Transformer with spatial reduction attention to take advantage of local-to-global features. MixFormer [2] models bidirectional interactions across spatial and channel dimensions of window self-attention and

depthwise convolution based on channel attention and spatial aggregation. Although MixFormer efficiently exchanges information across branches, the model still has limited receptive fields. Differently, this paper captures bidirectional local-to-global interactions across local self-attention and group self-attention to achieve better visual feature representation.

3 Methodology

An overview of the hierarchical BL2G architecture is shown in Figure 1. Following [29, 17, 2], the BL2G network includes four stages, and spatial dimension is progressively downsampled with a stride of $\{4, 8, 16, 32\}$. The channel dimension is doubled twice for each stage based on the base channel C. In the following, the detailed structure of BL2G attention is described in Figure 2.

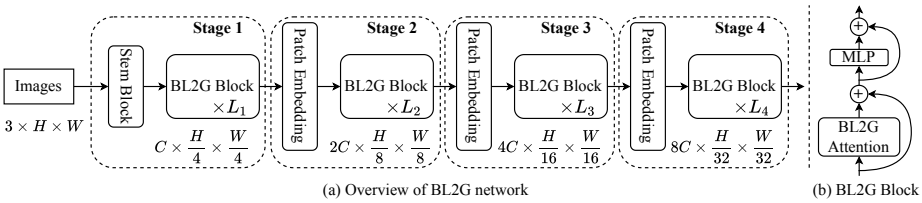


Fig. 1: (a) Overview of the BL2G network and (b) BL2G block. C, H, W are the base channel number, height, and width of the input feature. L denotes the number of stacked BL2G blocks. MLP indicates multi-layer perception.

3.1 BL2G Attention

The goal of BL2G attention is to extract local and global features and fully exchange information between two features. Local features are captured by performing attention inside non-overlapped windows. To mitigate the high computational cost of global self-attention, this paper introduces group attention that can extract global receptive fields and has linear complexity with image resolution. Bidirectional interaction between two branches is performed via cross-attentions with different query features.

Group Attention. Given image tokens $\mathbf{X} \in \mathbb{R}^{N \times C}$, cross-attention groups image features into a fixed number G, where $N = H * W$ is the number of image tokens. This is achieved by performing interaction between learnable group tokens $\mathbf{T}$ as query and image tokens as a pair of key and value. Each group token attends to image features globally, and performing all interactions results in the

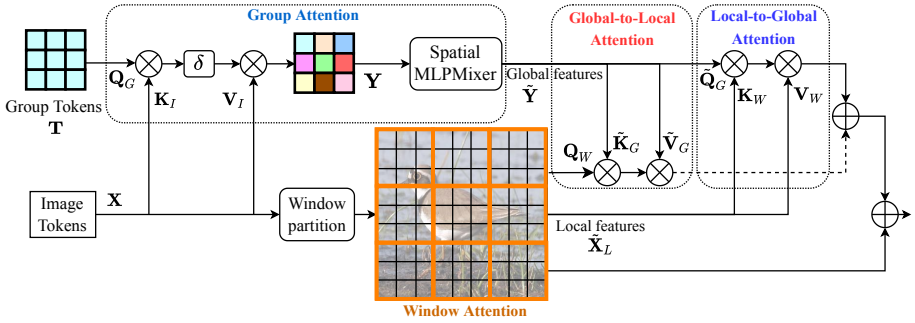


Fig. 2: Detail structure of BL2G Attention. Image tokens are grouped into a fixed number of group tokens via cross-attention. Spatial MLPMixer is used to exchange information across group features to capture global receptive fields. For image tokens, local self-attention is adopted to achieve linear complexity with image resolution. Bidirectional interaction across local and global features is captured by performing two cross-attentions with different q and k, v pairs.

attention matrix. Softmax δ is applied to each row of the attention matrix to output the attention map. Briefly, group attention is summarized as follows:

$$\mathrm{MHGA}(\mathbf{X}) = \mathrm{Concat}(\mathrm{GA}_1, \ldots, \mathrm{GA}_\mathrm{h}), \tag{1}$$

$$\mathrm{GA}_i(\mathbf{X}) = \mathrm{softmax}(\frac{\mathbf{Q}_G \mathbf{K}_I^\mathrm{T}}{\sqrt{C_h}})\mathbf{V}_I, \tag{2}$$

where GA_i is group attention for head i and h is the number of head. $\mathbf{Y} = \mathrm{MHGA}(\mathbf{X})$ indicates multi-head group attention composed of h GAs to produce output $\mathbf{Y}$. $\mathbf{Q}_G = \mathbf{T}\mathbf{W}_Q^G \in \mathbb{R}^{G \times C_h}$ is the group query updated with network parameters. $\mathbf{K}_I = \mathbf{X}\mathbf{W}_K^I, \mathbf{V}_I = \mathbf{X}\mathbf{W}_V^I \in \mathbb{R}^{N \times C_h}$ are key and value matrices. $\{\mathbf{W}_Q^G, \mathbf{W}_K^I, \mathbf{W}_V^I\} \in \mathbb{R}^{C_h \times C_h}$ are linear projections. Group attention matrix $\mathbf{A}_G = \mathbf{Q}_G \mathbf{K}_I^\mathrm{T} \in \mathbb{R}^{G \times N}$ means that N image tokens are grouped into G tokens. Obviously, the proposed GA has a linear computational cost with N while still capturing long-range dependencies from the image tokens.

After obtaining the grouped tokens, SpatialMLPMixer [23] is adopted to revise and exchange global features between grouped tokens. SpatialMLPMixer attains balances between parameter numbers, GFLOPs, and accuracy. Moreover, SpatialMLPMixer efficiently works with a fixed number of input tokens, which is well-appropriate for the grouped tokens. Typically, SpatialMLPMixer consists of two fully connected (FC) layers, and the GELU() activation function is inserted between them to learn the non-linear function in high spatial dimension:

$$\tilde{\mathbf{Y}} = \mathbf{W}_Y \otimes \mathrm{FC}_2(\mathrm{GELU}(\mathrm{FC}_1(\mathbf{Y}^\mathrm{T}))), \tag{3}$$

where $\mathbf{Y} \in \mathbb{R}^{G \times C}$ is the output of group attention. $\mathrm{FC}_1, \mathrm{FC}_2$ are fully connected layers applied across the spatial dimension. $\mathbf{W}_Y \in \mathbb{R}^{C \times C}$ denotes linear projection that mixes grouped tokens across the channel dimension (Channel MLP).

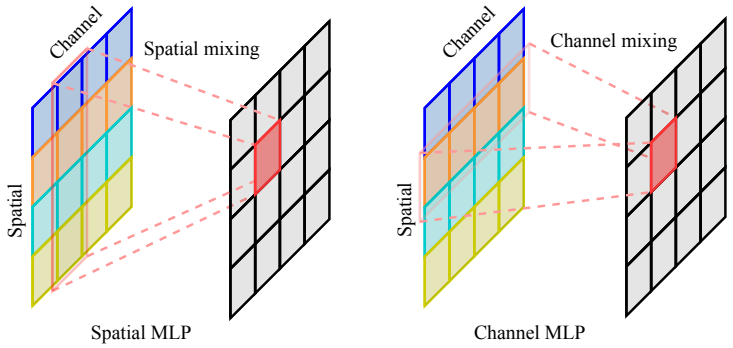


Fig. 3: Illustration of SpatialMLPMixer across spatial and channel dimension.

Figure 3 illustrates information mixing across spatial and channel dimensions. Each grouped token is fully connected to all other tokens. Therefore, global information across grouped tokens is updated.

Window Attention. The goal of window attention [17] is to capture local features from the image tokens. Window attention achieves high efficiency compared to global self-attention, which is compatible with high input resolution. Specifically, the image tokens are partitioned into non-overlapped windows, and self-attention performs interactions inside each window. The output of window attention is the local features $\tilde{\mathbf{X}}_L$:

$$\tilde{\mathbf{X}}_L = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^{\text{T}}}{\sqrt{C}})\mathbf{V}, \tag{4}$$

where $\mathbf{Q} = \mathbf{X}_W\mathbf{W}_Q, \mathbf{K} = \mathbf{X}_W\mathbf{W}_K, \mathbf{V} = \mathbf{X}_W\mathbf{W}_V \in \mathbb{R}^{N_W \times w^2 \times C}$ are query, key, value matrices projected from no-overlapped windows $\mathbf{X}_W \in \mathbb{R}^{N_W \times w^2 \times C}$, where N_W is number of windows and w^2 is window area.

Bidirectional Interaction. After revising the grouped tokens $\tilde{\mathbf{Y}}$ and acquiring local features $\tilde{\mathbf{X}}_L$, global information to the local image features and vice versa are propagated through *Global-to-Local Attention* (G2L) and *Local-to-Global Attention* (L2G) processes.

For (G2L) attention, the global features $\tilde{\mathbf{Y}}$ produced by SpatialMLPMixer are ungrouped via cross-attention operation. Where local features, as queries, attend to global features as a pair of keys and values, global information is returned to the image features. Formally, global-to-local attention is computed as follows:

$$\text{G2L}(\tilde{\mathbf{X}}_L, \tilde{\mathbf{Y}}) = \text{softmax}(\frac{\mathbf{Q}_W\tilde{\mathbf{K}}_G^{\text{T}}}{\sqrt{C}})\tilde{\mathbf{V}}_G, \tag{5}$$

Table 1: Results on ImageNet-1K Image Classification

Method	#params(M)	GFLOPs	Top-1 Acc.(%)
DeiT-T [24]	6.0	1.3	72.2
MobileViTv1-XS [18]	2.3	1.0	74.8
LVT [22]	3.4	0.9	74.8
PVT-T [29]	13.2	1.6	75.1
MobileViTv2-0.75 [19]	2.9	1.0	75.6
ResT-Lite [34]	10.5	1.4	77.2
PoolFormer-S12 [32]	11.9	1.8	77.2
Swin-1G [17]	7.3	1.0	77.3
EdgeViT-XS [20]	6.7	1.1	77.5
DFvT-S [8]	11.2	0.8	78.3
PVTv2-B1 [30]	13.1	2.1	78.7
BL2G (Ours)	**10.1**	**1.0**	**79.1**

where $\mathbf{Q}_W = \tilde{\mathbf{X}_L}\mathbf{W}_Q^W, \tilde{\mathbf{K}}_G = \tilde{\mathbf{Y}}\mathbf{W}_K^G, \tilde{\mathbf{V}}_G = \tilde{\mathbf{Y}}\mathbf{W}_V^G$ are query, key, and value features, and $\mathbf{W}_Q^W, \mathbf{W}_K^G, \mathbf{W}_V^G$ are linear projections.

For L2G attention, the local features acquired by Window Attention interact with global features, which enable local-to-global relations to flow from the window attention branch to the other. This is achieved by the cross-attention layer, where the local image features are queried by global group features:

$$\mathrm{L2G}(\tilde{\mathbf{X}}_L, \tilde{\mathbf{Y}}) = \mathrm{softmax}(\frac{\tilde{\mathbf{Q}}_G\mathbf{K}_W^{\mathrm{T}}}{\sqrt{C}})\mathbf{V}_W, \tag{6}$$

where $\tilde{\mathbf{Q}}_G = \tilde{\mathbf{Y}}\mathbf{W}_Q^G, \mathbf{K}_W = \tilde{\mathbf{X}}_L\mathbf{W}_K^W, \mathbf{V}_W = \tilde{\mathbf{X}}_L\mathbf{W}_V^W$ are query, key, and value features, and $\mathbf{W}_Q^W, \mathbf{W}_K^G, \mathbf{W}_V^G$ are linear projections.

Finally, ungrouped features and local-to-global features are fused via summation and the shortcut connection.

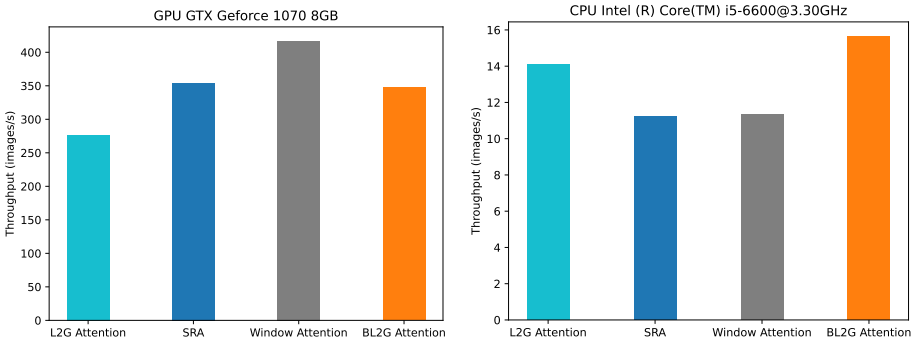
3.2 Model Configuration

Based on the proposed attention, the BL2G block is obtained, which includes two layer normalizations, BL2G attention (spatial mixing), MLP (channel mixing), and two residual connections inserted between two mixings. Similar to hierarchical backbones [29, 17, 6, 9], the BL2G network is introduced by stacking BL2G blocks across four stages, and spatial dimension is downsampled through Patch Embedding. Specifically, the number of BL2G blocks is set to {2, 2, 6, 6} and the base channel is 32. The number of heads in BL2G attention is configured with {2, 4, 8, 16}, and an MLP ratio of 4 is kept unchanged across 4 stages.

4 Experiments

4.1 Image Classification

Settings. The BL2G is trained and evaluated on the ImageNet-1K [5] which includes 1.2M training images and 50K validation images. Following common



(a) GPU throughput comparison (b) CPU throughput comparison

Fig. 4: Throughput comparison between baseline attention: L2G Attention in EdgeViT [20], Spatial Reduction Attention in PVTv2 [30], Window Attention in Swin Transformer [17], and our BL2G Attention. Throughput is measured with full precision and the same batch size (256 on GPU and 16 on CPU).

settings [29, 17, 2, 24], the model is trained for 300 epochs with a batch size of 1280. The optimizer is AdamW with a learning rate of $3e^{-3}$, a warmup epoch of 5, a weight decay of 0.05, and a momentum of 0.9. The input images are resized to 224×224. Standard data augmentations [24, 17, 29, 30] are adopted to improve performance, such as drop path, cut mix, mixup, rand augment, and label smoothing.

Results. Table 1 reports the results on ImageNet-1K image classification. With 10.1M parameters and 1 GFLOPs, BL2G achieves 79.1% of Top-1 accuracy which outperforms baseline Swin Transformer [17] by 1.8%, PVT-T [29] by 4.0%, competitive method EdgeViT-XS [20] by 1.6%, and recent method DFvT-S [8] by 0.8%. Figure 4 shows throughput comparisons between efficient methods. As a result, BL2G attention achieves similar speeds while attaining better accuracy than baseline attention.

4.2 Object Detection and Instance Segmentation

Settings. The proposed method is conducted on dense prediction tasks to validate the effectiveness of the BL2G backbone. MS-COCO [16] is used to train and evaluate the models. This dataset includes 118K training images and 50K validation images with 80 categories. The original backbone ResNet [13] in the detection model [15] and instance segmentation model [12] is replaced with the BL2G backbone. 1× training schedule (12 epochs) is employed to compare performances with common methods [15, 12, 32, 29]. The AdamW is adopted as an optimizer with a learning rate of $1e^{-4}$, a weight decay of 0.05, and a momentum of 0.9. Input images are resized to 1333×800 and a batch size of 16 is configured.

Table 2: Results on MS-COCO object detection using RetinaNet [15]

Backbone	#params(M)	GFLOPs	AP^{box}
ResNet-18 [13]	21	189	31.8
EMO-2M [33]	12	167	36.2
PoolFormer-S12 [32]	22	207	36.2
ResNet-50 [13]	38	250	36.3
PVT-T [29]	23	183	36.7
BL2G (Ours)	**18**	**167**	**37.5**

Fig. 5: Qualitative results of the BL2G Transformer with detection head Mask R-CNN [12].

Results. Table 2 compares our BL2G with other backbones [13, 33, 32, 29] using detector RetinaNet [15]. BL2G achieves 37.5% AP^{box} that surpasses recent method EMO [33] by 1.3% with similar GFLOPs, baseline ResNet-50 [13] by 1.2% while saving 33% GFLOPs, and PVT-T [29] by 0.8% with smaller computational costs.

Table 3 shows that the BL2G achieves consistent improvements compared to other competitors [13, 29]. The BL2G outperforms the baseline ResNet-50 by 2.0% AP^{mask} while saving 28% GLFOPs, and the PVT-T by 1.3% AP^{mask} while saving 11% GLFOPs. Qualitative results of the proposed method are illustrated in Figure 5.

Table 3: Results on MS-COCO instance segmentation using Mask R-CNN [12]

Backbone	#params(M)	GFLOPs	AP^{box}	AP^{mask}
ResNet-18 [13]	31	207	34.0	31.2
ResNet-50 [13]	44	260	38.0	34.4
ResNet-101 [13]	63	336	40.4	36.4
PVT-T [29]	33	208	36.7	35.1
BL2G (Ours)	**28**	**185**	**38.9**	**36.4**

4.3 Semantic Segmentation

Settings. The backbone BL2G is trained and evaluated on ADE20K dataset [36] using the semantic segmentation method [14]. Experimental configuration [14, 29] is adopted to train and evaluate the BL2G backbone. The model is trained for 80K interactions with a batch size of 16. The optimizer is AdamW with a learning rate of $2e^{-4}$ and a weight decay of $1e^{-4}$. Input images are resized to 512×512. The metric mIoU is used to evaluate the models, and only single-scale testing is adopted.

Table 4: Results on ADE20K semantic segmentation using Semantic FPN [14]

Backbone	#params(M)	GFLOPs	mIoU
ResNet-18 [13]	15.5	32.2	32.9
PVT-T [29]	17.0	33.2	35.7
ResNet-50 [13]	28.5	45.6	36.7
PoolFormer-S12 [32]	16.2	31.0	37.2
BL2G (Ours)	**12.6**	**25.7**	**38.8**

Results. Table 4 shows that the BL2G Transformer obtains better mIoU performances with less computational cost and parameters. In particular, BL2G achieves 38.8 mIoU which outperforms the PVT-T [29] by 3.1% while saving 22% GFLOPs, the baseline ResNet-50 [13] by 2.1% while saving 43% GFLOPs. It verifies that the proposed BL2G Transformer enables better visual representation learning.

5 Conclusion

This paper proposes the BL2G Vision Transformer as an efficient and versatile backbone. BL2G is designed to capture local and global features, and introduce a new bidirectional interaction between two features. Addressing issues in the Transformer encoder, group attention is proposed to alleviate the high computational cost of global self-attention while effectively modeling global information.

Bidirectional interaction promotes modeling ability in local and global features for window attention and group attention. Extensive experiments are conducted to validate the effectiveness of the proposed method on ImageNet-1K image classification, MS-COCO object detection, instance segmentation, and ADE20K semantic segmentation. As a result, BL2G Transformer achieves competitive performances compared to recent methods across visual tasks. In the future, the proposed network will be scaled to include more variants and deployed for real-world applications.

Acknowledgement

This result was supported by "Region Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE)(2021RIS-003).

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
2. Chen, Q., Wu, Q., Wang, J., Hu, Q., Hu, T., Ding, E., Cheng, J., Wang, J.: Mixformer: Mixing features across windows and dimensions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5249–5259 (2022)
3. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. Advances in Neural Information Processing Systems **34**, 9355–9366 (2021)
4. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. Advances in neural information processing systems **34**, 3965–3977 (2021)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134 (2022)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
8. Gao, L., Nie, D., Li, B., Ren, X.: Doubly-fused vit: Fuse information from vision transformer doubly with local representation. In: European Conference on Computer Vision. pp. 744–761. Springer (2022)
9. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: Convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12175–12185 (2022)

10. Hassani, A., Shi, H.: Dilated neighborhood attention transformer. arXiv preprint arXiv:2209.15001 (2022)
11. Hassani, A., Walton, S., Li, J., Li, S., Shi, H.: Neighborhood attention transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6185–6194 (2023)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6399–6408 (2019)
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
18. Mehta, S., Rastegari, M.: Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=vh-0sUt8HlG
19. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=tBl4yBEjKi
20. Pan, J., Bulat, A., Tan, F., Zhu, X., Dudziak, L., Li, H., Tzimiropoulos, G., Martinez, B.: Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In: European Conference on Computer Vision. pp. 294–311. Springer (2022)
21. Pan, X., Ye, T., Xia, Z., Song, S., Huang, G.: Slide-transformer: Hierarchical vision transformer with local self-attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2082–2091 (2023)
22. Pan, Z., Zhuang, B., He, H., Liu, J., Cai, J.: Less is more: Pay less attention in vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2035–2043 (2022)
23. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in neural information processing systems **34**, 24261–24272 (2021)
24. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
25. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer. In: European conference on computer vision. pp. 459–479. Springer (2022)

26. Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B., Shlens, J.: Scaling local self-attention for parameter efficient visual backbones. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12894–12904 (2021)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
28. Vo, X.T., Nguyen, D.L., Priadana, A., Jo, K.H.: Hierarchical vision transformers with shuffled local self-attentions. In: 2023 International Workshop on Intelligent Systems (IWIS). pp. 1–6. IEEE (2023)
29. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
30. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
31. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal attention for long-range interactions in vision transformers. Advances in Neural Information Processing Systems **34**, 30008–30022 (2021)
32. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10819–10829 (2022)
33. Zhang, J., Li, X., Li, J., Liu, L., Xue, Z., Zhang, B., Jiang, Z., Huang, T., Wang, Y., Wang, C.: Rethinking mobile block for efficient attention-based models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1389–1400 (2023)
34. Zhang, Q., Yang, Y.B.: Rest: An efficient transformer for visual recognition. Advances in neural information processing systems **34**, 15475–15485 (2021)
35. Zhang, Q., Yang, Y.B.: Rest v2: simpler, faster and stronger. Advances in Neural Information Processing Systems **35**, 36440–36452 (2022)
36. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019)