NomMer: Nominate Synergistic Context in Vision Transformer for Visual Recognition

Hao Liu† Xinghua Jiang* Xin Li Zhimin Bao Deqiang Jiang Bo Ren
Tencent YouTu Lab
{ivanhliu, clarkjiang, fujikoli, zhiminbao, dqiangjiang, timren}@tencent.com

Abstract

Recently, Vision Transformers (ViT), with the self-attention (SA) as the de facto ingredients, have demonstrated great potential in the computer vision community. For the sake of trade-off between efficiency and performance, a group of works merely perform SA operation within local patches, whereas the global contextual information is abandoned, which would be indispensable for visual recognition tasks. To solve the issue, the subsequent global-local ViTs take a stab at marrying local SA with global one in parallel or alternative way in the model. Nevertheless, the exhaustively combined local and global context may exist redundancy for various visual data, and the receptive field within each layer is fixed. Alternatively, a more graceful way is that global and local context can adaptively contribute per se to accommodate different visual data. To achieve this goal, we in this paper propose a novel ViT architecture, termed NomMer, which can dynamically Nominate the synergistic global-local context in vision transformMer. By investigating the working pattern of NomMer, we further explore what context information is focused. Beneficial from this “dynamic nomination” mechanism, without bells and whistles, the NomMer can not only achieve 84.5% Top-1 classification accuracy on ImageNet with only 73M parameters, but also show promising performance on dense prediction tasks, i.e., object detection and semantic segmentation. The code and models are publicly available at https://github.com/TencentYoutuResearch/VisualRecognition-NomMer.

1. Introduction

In computer vision, Convolutional Neural Networks (CNNs) [6, 9, 11, 21, 23, 35] have served as the de facto gold standard for many years. Recently, the Vision Transformer (ViT) [8] and its variants [25, 31] have been proposed challenging the status quo. Thanks to the global information communication and content-dependent learning nature of Self-Attention (SA), which is substantively different from local behavior in CNN, ViTs have shown superior performance on many visual recognition tasks.

However, the ViTs reasoning global dependency of split feature patch embeddings (a.k.a. tokens) is computationally expensive. To attack the issues, many recent works, such as SwinT [15], TNT-ViT [38] and HaloNet [27], introduce CNN-like inductive bias (e.g., locality, translation equivari-
nce) by building the token relations via self-attentions only within local windows and hierarchically aggregating them in a bottom-up manner. These local SA-based ViTs significantly improve the data efficiency, whereas the global contextual relation, especially in the early stage, is abandoned. To remedy this shortage, as illustrated in Fig. 1 (a) and (b), two kinds of “global-local ViTs” take a compromise, where both local and global visual dependencies are incorporated for each token. The first kind [12, 18, 37] marries the context aggregated by local SA or CNN with the one captured by global SA in parallel structure, such as multi-granular connection [24], in the basic block of hierarchical ViT. On the other hand, the local and global context are aggregated alternatively in the second kind [32, 39]. However, no matter in the parallel or successive manner, these global-local ViTs only focus on fusing global-local context while the modulation on it lacks, where the redundant information may have negative impact when recognizing various cases. For example, the “lampshade” is mis-recognized as “confectionery” by previous method, probably due to redundant contextual clues, such as colored lights, misleading the model.

According to a pioneering research [22], the human visual system can simultaneously process both peripheral and foveal vision when perceiving the real world scene, which also exhibits the intriguing property of modulation under various scenes. As a result, the redundant information can be naturally ignored. Specifically, foveal vision focuses on a local region of interest with more visual details, such as fine-grained details, colors or textures of objects in the scene, while peripheral vision refers to the one viewed at large angles but containing rough global scene information.

Inspired by the above preliminary and observations, in this paper, we regard the locally aggregated context in global-local ViT model as the foveal vision, while the globally aggregated one is treated as the peripheral-like visual information. Moreover, we propose a more effective context leverage strategy that the useful dependency information is nominated from local and global context dynamically. To achieve this nomination process, we need to solve the following two non-trivial problems: (i) How to make nominated global and local context work in harmony when processing different visual cases? (ii) How to preserve the information at most without increasing computational cost evidently when reasoning global dependency?

For the first problem, a straightforward way of nomination is to directly hard-sample from global or local context. Unfortunately, this sample process is indifferentiable, which makes the model suffer from the gradient lost problem. To overcome this issue, we coin a novel learnable Synergistic Context Nominator (SCN) in terms of aggregated context, which is in stark contrast to previous ViTs with global-local context greedily aggregated.

2) We coin a novel Compressed Global Context Aggregator (CGCA) more effective to reduce global redundancy and to capture global correlations.

3) We propose a novel ViT framework, termed NomMer, which enables the nominated global-local context complementary with each other for various cases and tasks. We also investigate the working behavior of SCN in NomMer.

4) Thanks to the “nomination” mechanism, the NomMer can achieve 84.5% Top-1 classification accuracy on ImageNet with only 73M parameters. With fewer parameters, our small and tiny models can still achieve 83.7% and 82.6% accuracy respectively. We also witness the promising performance of NomMer on dense prediction tasks, i.e., object detection and semantic segmentation.

2. Related Work

2.1. Vision Transformer

Global Vision Transformer. The vanilla Vision Transformer (ViT) [8] greedily reasons the global dependency of visual tokens throughout the whole model. However, it may suffer from the slow model convergence and expensive computational cost. To alleviate it, DeiT [25] exploits distillation technique while PVT [31] introduces the pyramid structure [13] into ViT.
Local Vision Transformer. Due to the lacking of inductive bias and the high computational complexity of SA calculated in global range, the inherent shortages of global ViT cannot be totally eliminated. Therefore, many subsequent ViTs alternatively propose to limit the token relation building by Self-Attention (SA) only within local regions. Among, SwinT [15] coins basic block with the successive WMHSA and shifted ones to perform within- and cross-window information communication. TNT-ViT [38] proposes to represent the local structure by recursively aggregating neighboring Tokens into one Token. In the similar spirit, HaloNet [27] introduces a non-centered local attention and extends it with “haloing” operation. Although the local SA-based ViTs significantly reduce the model complexity and improve the data efficiency, whereas the global contextual relation, especially in the early stage, is abandoned.

Global-Local Vision Transformer. To overcome the “global contextual information lost” issue, many global-local ViTs [12,18,32,37,39] attempt to seek an equilibrium between the fully global and fully local contextual information leverage. Specifically, the literature [37] proposes a focal SA incorporates both fine-grained local and coarse-grained global interactions. Conformer [18] fuses local features and global representations under different resolutions based on the Feature Coupling Unit. Other than the above global-local ViTs with parallel structure, NesT [39] stacks the local SA and CNN-based global aggregation modules alternatively while the basic block consisting of successive global SA and local CNN are designed in the CVT [32].

2.2. Redundancy Reduction Methods

The design of local [27, 38] and global-local ViTs [39] can be regarded as the redundancy reduction in terms of architecture. By contrast, our proposed architecture can dynamically harmonize the local and global context through nomination mechanism. Moreover, OctConv [6] is a pioneering successful attempt on feature redundancy reduction via separating the CNN feature into low- and high-frequencies, which is realized by down-sampling and up-sampling. DRConv [5] and DynamicViT [20] are another two representative works. The DRConv proposes to dynamically select the CNN filters whereas there is still local context exploited, while the DynamicViT dynamically sparsifies tokens, which may underperform on dense prediction tasks due to the attenuation of fine-grained local interactions. The DCTransformer [17] transits the view of solving the problem into frequency domain and demonstrates the sparse representations can carry sufficient information for generating images. Similarly, the work [36] also converts the input image into frequency domain for visual understanding. The intriguing property of frequency domain motivates our method aggregating global context at feature-level in frequency domain, which is different from above both works only operating input images.

3. Methodology

3.1. Architecture Overview

The overview of the proposed architecture is shown in Fig. 2 (a), which follows the hierarchical design in other ViT models [15,18,31,37]. It mainly consists of four stages with “Reduction Modules” dispersed between adjacent stages to down-sample the token numbers and increase the channels by factor 2. Each reduction module is composed of a $3 \times 3$ convolutional layer and a max pooling layer with stride 2.

Before being sent to the first stage, the image in size of $H \times W \times 3$ is split into patches of $4 \times 4 \times 3$ size, which are then projected by “patch embedding” consisting of convolutional layer with $4 \times 4 \times C$ kernel and 4 stride size into $H_4 \times W_4$ visual tokens in $C$ dimension. We coin two kinds of NomMer basic blocks, i.e., Synergistic NomMer (S-NomMer) and Global NomMer (G-NomMer) blocks in our proposed architecture. Stage 1 and 2 contains $N_1$ and $N_2$ S-NomMer blocks, which are responsible for capturing the synergistic context from global-local context. As the feature maps in stage 3 and 4 become relatively small in spatial size ($H_4 \times W_4$ and $H_3 \times W_3$), the global context and local context could become homogeneous and the information is highly abstracted. Therefore, inside the Stage 3 and 4, we build the G-NomMer layers upon Global Self-Attention (SA), which pay more attention on capturing the semantic information in global range. As is shown in Fig. 2 (b), both S-NomMer and G-NomMer layers are all equipped with Feed-Forward Networks (FFN) [28], residual connection and Layer Normalization (LN) [28]. As the S-NomMer layer is the core component of our method, we will elaborate it in the next subsections.

3.2. Synergistic NomMer Layer

To implement the “dynamic nomination” of S-NomMer layer serving as the core ingredient in our proposed ViT model, our design mainly concentrates on two vital aspects: 1) generating global and local context; 2) nominating synergistic context from them for each visual tokens. Generally, as illustrated in Fig. 3, S-NomMer layer has three trainable submodules, i.e., Compressed Global Context Aggregator (CGCA), Local Context Aggregator (LCA) and Synergistic Context Nominator (SCN). The CGCA provides rough global context while LCA yields more detailed contextual information within local region. By evaluating the contributions of local or global context aggregated to each tokens, SCN flexibly modulates them and build the nomination map to mask out the features with synergistic context.
3.2.1 Compressed Global Context Aggregator

As introduced in Sec. 1, we expect the peripheral-like global context can provide the rough visual clue in terms of the scene, where more fine-grained information can be supplemented from local context. Therefore, the feature maps including visual tokens for global context aggregation should be extracted from local context. Thus, the feature maps in- 

Feed-Forward Network (FFN) with residual connections.

Figure 2. (a) Architecture of the proposed NomMer; (b) NomMer basic block consisting of NomMer layer, Layer Normalization (LN) and Feed-Forward Network (FFN) with residual connections.

Figure 3. The input feature map is applied by 2D-DCT to obtain a series of corresponding frequency blocks \( \{ f^{(m,n)} \}, m = 1, 2, \ldots, \frac{D}{N}, n = 1, 2, \ldots, \frac{C}{N} \}, \) \( f^{(m,n)} \in \mathbb{R}^{N \times N \times C} \), where each channel of \( f^{(m,n)} \) is applied by 2D-DCT to obtain a serious of corresponding frequency blocks \( \{ f^{(m,n)} \}, m = 1, 2, \ldots, \frac{D}{N}, n = 1, 2, \ldots, \frac{C}{N} \}, \) \( f^{(m,n)} \in \mathbb{R}^{N \times N \times C} \), which is denoted as:

\[
    f(i, j) = \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} c(u) c(v) F_{u,v} \cos \left( \frac{(i+0.5)\pi}{N} u \right) \cos \left( \frac{(j+0.5)\pi}{N} v \right),
\]

where \( F_{u,v} \) represent the pixel with index \((u, v)\) in \( F^{(m,n)} \) while the \( i \) and \( j \) are indexes of horizontal and vertical spatial frequencies in frequency block \( f^{(m,n)} \). \( c(\cdot) \) is the normalization scale factor ensuring the orthogonality.

Afterwards, the redundancy reduction is achieved by the low-frequency perceiver (LFP) module. In detail, LFP first drops the high-frequencies of each frequency block in proportion \( \alpha \), which is set to 0.5 in default:

\[
    \hat{f}^{(m,n)} = \{ f^{(m,n)}(i, j) \}, i, j \in \{1, 2, \ldots, l\},
\]

\[
    l = \lceil \alpha N \rceil. \tag{3}
\]

then the low frequency map is obtained by flattening each \( \hat{f}^{(m,n)} \) into a vector in \( l^2 \cdot C \) dimension. To further extract useful frequencies while reducing dimensions, a convolutional layer with \( 1 \times 1 \times C \) kernel is applied to obtain the compressed frequency map \( \tilde{f} \in \mathbb{R}^{\frac{D}{N}} \times \frac{C}{N} \times C \). As the redundancy is only reduced in the frequency domain while the spatial information still preserved, it is feasible to perform global context aggregation by using global multi-head self-attention (G-MHSA):

\[
    \tilde{F}^{(G)} = \text{Conv}(G-MHSA(\tilde{f})), \tag{5}
\]

where \( \tilde{F}^{(G)} \in \mathbb{R}^{\frac{D}{N}} \times \frac{C}{N} \times N^2 \cdot C \) and “Conv(\cdot)” is a convolutional layer with \( 1 \times 1 \times N^2 \cdot C \) kernel size. Then, \( \tilde{F}^{(G)} \) is reshaped to the tensor with shape \( \frac{D}{N} \times \frac{C}{N} \times N \times N \times C \), where each block has the same shape \( (N \times N \times C) \) with \( F^{(m,n)} \). To project the frequency maps with compressed global context back to the spatial domain, the 2D-IDCT (Inverse DCT) is conducted within each channel of each block \( \tilde{F}^{(G)}_{(m,n)} \) according to:

\[
    \tilde{F}^{(G)}_{u,v} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} c(u) c(v) \cdot \hat{f}^{(G)}(i, j) \cos \left( \frac{(i+0.5)\pi}{N} u \right) \cos \left( \frac{(j+0.5)\pi}{N} v \right), \tag{6}
\]

where \( \tilde{F}^{(G)}_{u,v} \) represent the pixel with index \((u, v)\) in a restored spatial feature block while the \( \hat{f}^{(G)}(i, j) \) are spatial frequencies in one frequency block of \( \tilde{F}^{(G)} \). \( c(\cdot) \) is the normalization scale factor given in Eqn. (2). Finally, spatial feature with compressed global context \( \tilde{F}^{(G)} \) is yielded.
3.2.2 Local Context Aggregator

Local Context Aggregator (LCA) plays as the role of aggregating local context with more foveal vision-like details serving as the complementary visual information, which is expected to collaborate with compressed global context. To reach this goal, as illustrated in Fig. 3, we adopt two kinds of LCAs in this work, i.e., Local-MHSA (L-MHSA) and CNN. As suggested by many prevalent researches [18, 39], CNN introduced in the ViT model can provide more inductive bias helping model convergence. Essentially, the CNN is the content-independent aggregator while the MHSA is the content-dependent one. The combination of CNN and MHSA can be deemed as a trade-off between leveraging pre-defined inductive bias or learning it from data. However, most of previous methods adopting parallel [18] or subsequent [39] combination manner can only have single acquisition w.r.t. the inductive bias within one layer. Contrastively, a more elegant way is to determine the inductive bias usage according to the specific region, which motivates our design of LCA providing both CNN feature \( F(C) \) and L-MHSA feature \( F(L) \) for dynamic nomination. More concretely, we inherit the window-based self-attention (W-MSA) with \( M \) window size from SwinT [15] as L-MHSA. On the other hand, we adopt “Bottleneck” [9] with structure \{Conv_{1\times1}, Conv_{3\times3}, Conv_{1\times1}\} as the CNN aggregator.

3.2.3 Synergistic Context Nominator

Once the compressed global and local context features obtained, they are treated as the candidate features \( \mathbf{F} = \{F^{(L)}, F^{(C)}, F^{(G)}\} \in \mathbb{R}^{D \times D \times C} \) (highlighted by red dotted boundary in Fig. 3) for the proposed Synergistic Context Nominator (SCN). The aim of SCN is to pick up the most valuable context with synergy for each spatial location, where the detailed process is described in Fig. 3. The candidate features are first fused by element-wise addition operation and passed to a convolutional layer with \( 1 \times 1 \times 3 \) kernel to obtain a tensor \( \mathbf{T} \) of shape \( D \times D \times 3 \), where each channels in the last axis describes the probability of nominating either one type context feature from three candidates. To further acquire the nomination map, we can perform the element-wise hard sampling on the tensor \( \mathbf{T} \):

\[
\omega_{i,j} = \arg\max (\tau_{i,j,1}, \tau_{i,j,2}, \tau_{i,j,3}),
\]

where \( \tau_{i,j,c} \) is the \((i, j)\)-th element in \( c\)-th channel of \( \mathbf{T} \), \( \omega_{i,j} \) is the channel index of nominated context feature at \((i, j)\) spatial location. Correspondingly, the nomination map \( \Omega \in \mathbb{R}^{D \times D \times 3} \) can be determined. For example, if the L-MHSA context feature is nominated at position \((i, j)\), the corresponding nomination vector \( \Omega_{i,j} \) is \([1, 0, 0]\).

One challenge lying in the proposed SCN module is that the hard sampling process is not differentiable whereas the weights in SCN need to be updated during training. To tackle this issue, we introduce a reparameterization method, termed Gumbel Softmax trick [10], which allows the gradients to back-propagate through the discrete sampling process. In the end, the feature map \( \mathbf{F}^{(S)} \in \mathbb{R}^{D \times D \times C} \) with nominated synergistic context is output by masking the nomination map on the candidate feature set:

\[
f^{(S)}_{i,j} = \sum_{p=1}^{3} \Omega_{i,j} f^{(S)}_{p,i,j},
\]

where \( f^{(S)}_{i,j} \) is the \((i, j)\)-th feature vector of \( \mathbf{F}^{(S)} \) and \( p \) is the type index of context feature at location \((i, j)\).

3.3. NomMer Variants

To fully explore the potential of NomMer with different configurations, we build several variants of it, i.e., NomMer-T, NomMer-S and NomMer-B, which refer to tiny, small and base model separately. The detailed configurations are given in supplementary material.

4. Experiments

4.1. Image Classification on ImageNet-1K

Experimental setting. We compare our proposed NomMer against several baselines on ImageNet-1K [7]. For
a fair comparison, we follow the experimental settings in [25]. Concretely, all our models are pre-trained for 300 epochs with the input size of $224^2$. The initial learning rate and batch size are set to $10^{-3}$ and 1,024, respectively. For optimization, AdamW [16] optimizer with a cosine learning rate scheduler is used. The weight decay is set to 0.05 and the maximal gradient norm is clipped to 5.0. We also inherit the data augmentation and regularization techniques from [25]. The stochastic depth drop rates are set to 0.1, 0.3 and 0.5 for tiny, small and base models individually. When reporting the results of $384^2$ input, we fine-tune the models with a total batch size of 512 for 30 epochs. The learning rate and weight decay are $10^{-5}$ and $10^{-8}$.

**Performance.** In Tab. 1, we compare our NomMer with state-of-the-art CNN and Transformer architectures on image classification task. The results show that the proposed NomMer consistently outperforms other approaches with similar model size and computational budgets. Compared with CNN-based RegNetY [19], our models achieve improvements ranging from 1.6% to 2.6% under three configurations with input image size of $224^2$. Compared with the ViT models, the proposed NomMer also witnesses the superior performance. More specifically, our NomMer-B with 73M parameters achieves a 84.5% ImageNet Top-1 accuracy, surpassing DeiT-B [25], Swin-B [15] and Conformer-B [18] by 2.7%, 1.2% and 0.4%, respectively. Besides, the lightweight version (NomMer-T) also achieves the best performance. When finetuned on the $384^2$ images, the similar trend is also observed. In addition, we further verify the effectiveness of our proposed method pre-trained on the larger ImageNet-21K dataset (see supplementary material).

### 4.2. Object Detection on COCO

**Experimental setting.** To verify NomMer’s versatility, we benchmark it on object detection with COCO 2017 [14]. The models pretrained on ImageNet-1K [7] are used for initializing the backbone of Cascade Mask R-CNN [3] framework. Similar to SwinT [15], we follow $3 \times$ schedule training with 36 epochs for a fair comparison. During training, multi-scale training strategy is employed to randomly resize image’s shorter side to the range of [480, 800]. And we use AdamW [16] for optimization with initial learning rate $10^{-4}$ and weight decay 0.05. In the similar spirit, 0.1, 0.3 and 0.5 stochastic depth drop rates are set to regularize the training for tiny, small and base models.

**Performance.** The box and mask mAPs on COCO validation set are summarized in Tab. 2, from which we can see that NomMer significantly boosts the $AP^b$ and $AP^m$. In details, the box’s mAP and mask’s mAP of NomMer-B are 0.8% and 0.6% higher than that of the strong baseline Swin-B [15], which demonstrates the importance of global representations for high level tasks in our method. When evaluating our tiny model, it surpasses the second best method Focal-T [37] by 0.3%, which also suggests that NomMer can also perform well on the prediction tasks with fewer parameters. To further investigate the versatility of the proposed model when it works in different detection frame-
works, we also conduct a series of experiments to compare NomMer with other SOTAs. A more detailed description is given in supplementary material.

<table>
<thead>
<tr>
<th>Method</th>
<th>#param. (M)</th>
<th>FLOPs (G)</th>
<th>AP50 (%)</th>
<th>AP75 (%)</th>
<th>APm (%)</th>
<th>AP50m (%)</th>
<th>AP75m (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res50 [9]</td>
<td>82</td>
<td>739</td>
<td>46.3</td>
<td>64.3</td>
<td>50.5</td>
<td>40.1</td>
<td>61.7</td>
</tr>
<tr>
<td>X101-32 [35]</td>
<td>101</td>
<td>819</td>
<td>48.1</td>
<td>66.5</td>
<td>52.4</td>
<td>41.6</td>
<td>63.9</td>
</tr>
<tr>
<td>X101-64 [35]</td>
<td>140</td>
<td>972</td>
<td>48.3</td>
<td>66.4</td>
<td>52.3</td>
<td>41.7</td>
<td>64.0</td>
</tr>
<tr>
<td>Swin-T [15]</td>
<td>86</td>
<td>745</td>
<td>50.5</td>
<td>69.3</td>
<td>54.9</td>
<td>43.7</td>
<td>66.6</td>
</tr>
<tr>
<td>Swin-S [15]</td>
<td>107</td>
<td>838</td>
<td>51.8</td>
<td>70.4</td>
<td>56.3</td>
<td>44.7</td>
<td>67.9</td>
</tr>
<tr>
<td>Swin-B [15]</td>
<td>145</td>
<td>982</td>
<td>51.9</td>
<td>70.9</td>
<td>56.5</td>
<td>45.0</td>
<td>68.4</td>
</tr>
<tr>
<td>Focal-T [37]</td>
<td>87</td>
<td>770</td>
<td>51.5</td>
<td>70.6</td>
<td>55.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NomMer-T</td>
<td>80</td>
<td>755</td>
<td>51.8</td>
<td>70.8</td>
<td>56.0</td>
<td>44.7</td>
<td>67.6</td>
</tr>
<tr>
<td>NomMer-S</td>
<td>99</td>
<td>851</td>
<td>52.4</td>
<td>71.5</td>
<td>56.8</td>
<td>45.1</td>
<td>68.8</td>
</tr>
<tr>
<td>NomMer-B</td>
<td>130</td>
<td>1006</td>
<td>52.7</td>
<td>71.6</td>
<td>57.2</td>
<td>45.6</td>
<td>68.9</td>
</tr>
</tbody>
</table>

Table 2. Results on COCO object detection and instance segmentation with Cascade Mask R-CNN.

4.3. Semantic Segmentation on ADE20K

Experimental setting. For another dense prediction task, Semantic Segmentation, we further evaluate our model on the ADE20K [40] dataset. Specifically, our NomMer serves as the backbone of UperNet [33] which is a prevalent segmentation method. Unless explicitly specified, we use a standard recipe by setting the image size to 512^2 and train the models for 160k iterations with batch size 16.

Performance. The results of Upernet with various backbones on the ADE20K [40] dataset are reported in the Tab. 3, where both single-scale and multi-scale evaluation results are included. Obviously, our method significantly outperforms previous state-of-the-arts under different configurations. Under single-scale setting, our NomMer separately achieves 50.0%, 48.7%, and 46.1% mIoU with base, small and tiny model configurations, which are 1.0%, 0.7%, 0.3% higher than the strong baseline Focal [37] counterparts. And we can also observe the consistent performance improvement under the multi-scale evaluation. Conclusively, our NomMer can steadily improve the performance of various visual recognition tasks owing to the synergistic context nomination.

4.4. Ablation Study

To better investigate the effectiveness of different aspects in our proposed S-NomMer layer, we conduct extensive ablation studies on both classification and downstream tasks of which the results are summarized in Tab. 4.

Effect of Local Context Aggregator. We validate the effectiveness of the LCA combing both L-MHSA and CNN. Compared with the performance of “-w/o Global&CNN” only equipped with L-MHSA, the additional CNN (“-w/o Global”) consistently increases 0.3% accuracy on ImageNet [7], 0.2% box’s mAP and 0.3% mask’s mAP on COCO [14], and 0.1% mIoU on ADE20K [40]. This confirms the benefit of inductive bias provided by CNN and also shows the combination can capture the salient fine-grained features through synergy.

Effect of Compressed Global Context Aggregator. As is shown in Tab. 4, we find the performance of all tasks can be consistently improved by integrating the global context into local ones, even if the global context is aggregated from simple max-pooling features. When the aggregation is performed in the frequency domain through our DCT-based CGCA, the average accuracy of each task is further boosted, which indicates our learnable CGCA can well strike the trade-off between redundancy reduction and useful information preservation for the visual recognition tasks.

Effect of Synergistic Context Nominator. To verify the capability of another core component SCN in our architecture, we compare the performance of our NomMer with and without the SCN equipped on each task. As demonstrated in Tab. 4, our model can obtain the best results by adopting nominator, with at least 0.4% improvement on different tasks than the non-nominator version in which various types of contexts are directly fused through element-wise addition. These results further prove the effectiveness of the synergistic context learned by SCN.

4.5. Qualitative Analysis

To further investigate the working pattern of our proposed NomMer, in Fig. 4, we visualize the synergistic nomination maps from intermediate layers of S-NomMer block from base model on classification task. We surprisingly observe that nomination maps exhibit several intriguing properties. In low-level nomination maps exhibit (“Layer 1,”) the CNN context features are always predominant, which is
Table 4. Ablation study of NomMer on three benchmarks based on the NomMer-T architecture.

<table>
<thead>
<tr>
<th>Method</th>
<th>Local</th>
<th>Global</th>
<th>Nominator</th>
<th>ImageNet Top-1 (%)</th>
<th>COCO APb</th>
<th>APm</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NomMer-T w/o Global &amp; CNN</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>81.4</td>
<td>50.4</td>
<td>43.5</td>
<td>44.7</td>
</tr>
<tr>
<td>NomMer-T w/o Global</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>81.7</td>
<td>50.6</td>
<td>43.8</td>
<td>44.8</td>
</tr>
<tr>
<td>NomMer-T w/o Gumbel &amp; DCT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>82.0</td>
<td>50.9</td>
<td>44.0</td>
<td>45.1</td>
</tr>
<tr>
<td>NomMer-T w/o Gumbel</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>82.2</td>
<td>51.2</td>
<td>44.3</td>
<td>45.4</td>
</tr>
</tbody>
</table>

Figure 4. Visualization of nomination maps, attention maps and CNN CAM [30] maps of nominated features from NomMer-B on classification task. Red: CNN context $F^C$. Green: Local context $F^L$. Blue: Compressed global context $F^G$. “Layer B.L” stands for that map is from the $L$-th NomMer layer of the $B$-th NomMer block. The pink hollow boxes in attention maps represent the locations of nominated local or global context features. Best viewed in color and zoom in.

Figure 5. Class activation attention maps on classification task. Best viewed in color.

consistent with the conclusion given in research [34], that early convolutions help transformers see better. Along with the model going deeper, nomination maps present various synergistic patterns of context in different layers. Specifically, the local context features aggregated by CNN and L-MHSA contribute most in “Layer 1_4”, while the adjacent “Layer 1_5” focuses on the global context accompanied with local CNN context. These phenomena demonstrate that our model can successfully obtain synergistic context with redundancy reduced, where the synergy not only happens between local and global context, but also cross layers.

Moreover, one can also observe that the global context features are inclined to be nominated in the smooth regions, e.g., “Layer 1_5” with less color and texture variations. We attribute this behavior to the missing of salient details in smooth regions where model needs to refer more information from larger scope, which is illustrated by the global attention maps in Fig. 4 from “Layer 2_5”. Comparatively, there are more fine-grained texture or patterns involved in the local context aggregated by CNN and L-MHSA, which can be demonstrated by the local attention maps and CNN CAM maps (from “Layer 2_5”) obtained by algorithm [30].

Thanks to the nomination mechanism, NomMer-B can capture more discriminative features from synergistic context than other ViTs, such as DeiT [25] only exploiting global context. By visualizing attention maps of class activation based on method [4], in Fig. 5, we find that the attentions of NomMer-B often concentrate on the exclusive features of object, such as the keyboard of “combination lock” and the dot pattern of “slug” while DeiT-B presents more unstable activation maps.

5. Conclusion

In this work, we introduced a novel vision transformer architecture solving visual recognition tasks by nominating synergistic context. Extensive experiments on image classification and dense prediction tasks demonstrated its superiority over state-of-the-arts. The visualization of the nomination maps learned by our method can also provide an empirical guidance for the design of architecture, which will be further explored in our future work.
References


