UDC 004.89:004.422.6 DOI: https://doi.org/10.30837/IYF.CVSAMM.2024.157

# RELAXATION OF CONTIGUITY CONSTRAINT IN VARIED-SIZE WINDOW ATTENTION IN VISION TRANSFORMERS

Lukashov D.S.
Science Supervisor – PhD in Technical Sciences, Ac. Prof. Naumeyko I.V.
Kharkiv National University of Radio Electronics,
Department of Applied Mathematics,
Kharkiv, Ukraine
e-mail: dmytro.lukashov@nure.ua

Метою цієї роботи є дослідження можливості подальшої релаксації обмежень у структурі вікон Swin трансформерів із введенням поняття змінного розміру вікна. Ця робота пропонує повністю відійти від вікон та зняти обмеження на те, що механізм самоуваги має бути застосовано між елементами, що знаходяться у неперервному вікні. Вибір елементів між якими має застосовуватись механізм самоуваги пропонується віддати на навчання основному трансформеру, а не окремій мережі.

A crucial part of basically any state-of-the-art deep computer vision network is a special subnetwork called backbone. It is responsible for extracting features from images which are then used by other subnetworks to solve a task in question (classification, detection, segmentation etc.). Currently, one of the best, if not the best, networks to be used as a backbone is swin transformer [1] and its modifications, such as varied-size vision transformer [2].

Swin transformer is a vision transformer that uses so called "shifting windows" to avoid performing MSA (multi-head self-attention) across all patches of an image [3].

Varied-size vision transformers paper relaxes fixed window sizes and non-overlapping constraint, and it also proposes a method to efficiently choose window sizes using a small convolutional network. It also explores a possibility to not use shifting by arguing that is redundant since the overlapping is allowed. In the result, authors have been able to improve the performance of the original swin transformer architecture by a few percents which is a significant amount considering the performance of the modern computer vision algorithms.

The success of the relaxation of the window constraints in swin transformers suggests that it is worth exploring further in this direction. These proceedings propose a method to learn "window" configurations without including any prior knowledge into the initial configuration. However, it is necessary to point out that "window" is no longer a correct word because there is no contiguity constraint, so it is more appropriate to call learned patterns "focus groups".

The obvious downside of the proposed approach is that more data, time and processing resources required to learn focus groups. However, there are two very significant benefits: flexibility and transferability. The first one is obvious, the network is free to learn any configuration that it thinks is appropriate which

can lead to performance improvements. The second one means that it is possible to transfer the approach to other data modalities where it is not obvious or not possible to choose good windows.

To formally define an approach, consider the following two matrices $X \in \mathbb{R}^{s \times d}$ and $W \in \mathbb{R}^{g \times s}$ where $s$ is the sequence length, $d$ is the dimension of a transformer and $g$ is the number of focus groups. Each focus group is defined by the following equation:

$$F^{(i)} = X_{\text{index\_of}(\text{topk}(W_i, k), W_i)}. \quad (1)$$

There $k$ is the length of a focus group, it is chosen so that $gk = s$ to preserve the sequence length by concatenating $F^{(i)}$ across the first dimension so that it is possible to do the residual connection with the input $X$. The topk function selects the top $k$ biggest elements of a vector and the index_of function selects the indices in the second argument which correspond to the values in the first. It is obvious that $F^{(i)} \in \mathbb{R}^{k \times d}$. The standard MSA is then applied to each $F^{(i)}$ and the results of that operation are then concatenated as has been mentioned above.

One big problem of this approach is that the functions topk and index_of are non-differentiable. However, this can be easily fixed by defining the operation $G$ in the following way:

$$G(x, w) = x,\ \partial G / \partial x = 1,\ \partial G / \partial w = x.$$

This operation acts as an element-wise multiplication where $w$ is assumed to always be 1 (or vector/matrix/tensor of ones). It does not change gradient flow to $x$ but allows $w$ to receive gradients from the network. So, to circumvent the problem in question it is just necessary to extend the equation (1):

$$F^{(i)} = G(X_{\text{index\_of}(\text{topk}(W_i, k), W_i)}, \text{topk}(W_i, k)).$$

Essentially what this does is imagining as if the weights $W_i$ were ones and zeros broadcasted to the shape $s \times d$ and were element-wise multiplied with $X$.

In summary, the proceedings propose to relax the constraints of swin transformers even more than the varied-size window attention does by introducing the notion of learnable focus groups. The potential drawbacks and benefits of such a relaxation have been discussed and the method for learning focus groups have been introduced. The next logical step is to experiment with the approach on different data sets and evaluate its performance.

References:

1. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows / [Z. Liu, Y. Lin, Y. Cao та ін.]. // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2021. – P. 10012–10022.

2. VSA: Learning Varied-Size Window Attention in Vision Transformers / Q. Zhang, Y. Xu, J. Zhang, D. Tao. – 2022.

3. Theoretical feasibility of fully linear multiple full-size attentions / D. Lukashov, I. Naumeyko, N. Lukashova. // Information Technology and Implementation (IT&Is-2023) Kyiv. – 2023. – P. 276–278.