"Sufficient" Attention Is All You Need

Shivasankaran V P
IIT Gandhinagar
vp.shivasan@iitgn.ac.in

Abstract

Vision Transformers (ViTs) have outperformed convolutional neural networks(CNNs) and have established new benchmarks in computer vision. However, two of the main caveats of Vision Transformers (ViT) are the need for huge datasets for pre-training and huge model sizes. In this work, we explore alternative sparse self-attention patterns to the full self-attention used in ViTs that could alleviate the need for huge models and pre-training datasets. Our experiments conclusively prove that in small ViTs, full self-attention is detrimental to overall performance. Experimental results show that in small models, even random sparse self-attention performs better than full self-attention. Our best-performing sparse-self-attention ViT outperforms the full self-attention variant by 12 accuracy points.

1 Introduction

Transformers[20] originally proposed for natural language processing (NLP) have recently seen extraordinary success in computer vision (CV) by taking the form of vision transformers[7]. This wide success of transformers is usually attributed to self-attention, the main engine of transformers. Vision transformers(ViTs) have outperformed previous SOTA models in image classification[7], object detection [2] and other vision tasks. However, the boost in performance comes at a price, the models are enormous in size, and they require huge amounts of data to attain the performance boost. Self-attention acts as the main core for transformers through its capability of taking advantage of modern processing units like GPUs/TPUs and parallelizing the attention computation for each token. This ability of parallelizing computation has allowed transformers to take advantage of huge datasets and also allow computationally more complex self-attention computation, and thereby ViT's learning paradigm becomes less dependent on inductive biases and more dependent on data.

However, the SOTA ViTs often require pre-training on huge datasets like JFT-300M[18] and ImageNet[5], and when trained on smaller datasets like tiny-imagenet[11] with no pre-training tasks they fail to perform[12]. This strong dependence on data for the performance of ViTs is mainly due to the lack of any strong inductive bias for the model to take advantage of. Many works have been proposed to introduce soft inductive biases like convolutions to ViTs[8][14] and have proved to be successful, but there is a lack of work studying sparse self-attention with subtle inductive biases designed for ViTs.

The standard self-attention is quadratic with respect to the number of tokens because every query attends to every key. This dense nature of self-attention and the failure to perform with limited data[12] raises an important question *Is the dense self-attention necessary for the performance of ViTs? Are they detrimental to the performance with limited data? Could introducing sparsity to self-attention lead the model to learn the visual features with limited examples?* Previous works in NLP[25][13] and CV[24][3] have proved that dense self-attention is not necessary.

BigBird[25] is one of the prominent works studying sparse self-attention for NLP tasks. Taking inspiration from the discussed sparse self-attention patterns in BigBird, we conduct a comparative study on the performance of ViTs with different self-attention patterns under limited computation power and data. The three main contributions of this work are as follows

Preprint.

- We establish proof of concept for improvement in the performance of ViTs using sparse self-attention over full self-attention in small models.
- We prove that more attention does not necessarily translate to better performance. Specifically, we showcase random attention, which outperforms full self-attention by six accuracy points in the tiny configuration.
- Our experiments also reinforce the intuitive understanding that in bigger models, full self-attention is more beneficial than sparse self-attention.

2 Related Works

Several works have been proposed for designing efficient transformers. In the computer vision landscape, previous works have majorly focused on merging the features of self-attention using convolutions like [14][23], which resulted in models outperforming ViTs and computationally cheaper self-attention computation through the introduced inductive bias. Some prominent works like DEIT[19] uses knowledge distillation tokens, and other works like XCIT[1] use ross-covariance matrix between queries and keys for computing self-attention. Some works have taken a different path by progressively decreasing the number of tokens at each self-attention layer, like [15][16]. [14] introduces the Swin transformer, a hierarchical transformer architecture where the dense self-attention between non-overlapping patches is replaced by a self-attention which operates between overlapping patches through a shifting window. CVT[23] replaces linear projection of patches and the MLP head in vision transformers with convolutional layers. CVT removes the need for positional embeddings since the embeddings are convolutional projections instead of linear projections. [15] introduces the patchmerger module, which reduces the number of output tokens at a layer by merging the tokens with the previous layer's output tokens through learned weights. TOKENLEARNER[16] proposes a trainable tokenizer to tokenize the image. The trained tokenizer mines important tokens, and the subsequent layers calculate self-attention only on these mined tokens.

In the NLP field, sparse self-attention has been more closely studied in terms of which factions in dense self-attention attributes to the success of transformers like BIGBIRD[25] and FNET[13]. FNET replaces the self-attention layer in transformers by unparameterized standard 2D Fourier Transform and achieved 92-97% accuracy of BERT[6] on GLUE[21] benchmark. BIGBIRD proposes a Turing complete linear sparse self-attention mechanism, which preserves the properties of full self-attention.

3 Big Bird

BIGBIRD formulates self-attention in a more generalized form as shown in Eq 1[25] by treating tokens as nodes in a graph. An edge in the graph represents two attending tokens. The redefined self-attention formulation gives more flexibility in defining sparse self-attention clearly and concisely.

$$ATTN_D(X)_i = x_i + \sum_{h=1}^{H} \sigma(Q_h(x_i)K_h(X_{N(i)})^T) \cdot V_h(X_{N(i)}) \tag{1}$$

where $X = (x_1, x_2, ..., x_n)$ is the input sequence and D is directed graph with vertices as tokens. If x_i attends to x_j, then the corresponding nodes in D will have an edge accordingly. Note that $X_{N(i)}$ are the tokens to which x_i attends. In the extreme case where D is a complete graph, the equation calculates the full self-attention proposed in the original transformer paper[20]. The main advantage of rewriting the self-attention equation in this form is that now the sparsification of self-attention can be approached as a graph-sparsification problem. The authors of BIGBIRD use this self-attention formulation to prove their sparse proposed linear sparse-self attention is Turing complete and universal approximators. Refer to the original paper[25] for detailed proofs.

The authors of BIGBIRD hypothesise that a small average path between nodes and the notion of locality between nodes is desirable for sparse self-attention. The sparse self-attention of BIGBIRD consists of three attention patterns, namely random, window, and global attention. Refer to Fig 1. Random and window attention patterns are designed from their hypothesis, while global is a theoretical ramification.

Random Attention Random graphs have the power to approximate the spectral properties of complete graphs[17][9]. In order to facilitate fast information flow between tokens, they use the Erdos-Rényi model, where the shortest path between nodes is logarithmic with respect to the total number of nodes[10][4]. Each query attends to a random number of keys.

Window Attention Graphs with high clustering coefficients are beneficial because they bring a notion of a locality to self-attention. In order to balance with the Random attention pattern designed above and maintain the fast flow of information between tokens, they use the Watts and Strogatz model[22] to construct small world graphs. Each query at location i attends to the keys in its window of size w centred at i.

Global Attention Each query attends to added global tokens, and every global token attends to every token.

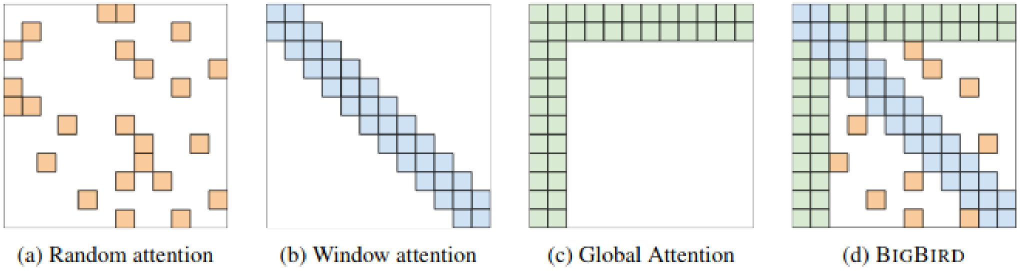
(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

Figure 1: Self-attention patterns in BIGBIRD

4 Experiments

We adapt the proposed sparse self-attention implementation of BIGBIRD[1] for ViT, and the adapted code can be found here `https://github.com/Shiva-sankaran/SparseAttentionViT`. We will use BBVIT to denote ViT trained using one or more of the attention patterns discussed in section 3 and ViT to denote the original vision transformer proposed in [7]. Since this work is directed towards establishing proof of concept for the performance of sparse self-attention in models with a small number of parameters, we train BBVIT and ViT with the following two configurations, as listed below. Note that the only difference between BBVIT and ViT is the type of self-attention used. In ViT, we use full self-attention, and in BBVIT, we a sparse self-attention. We use Imagenette dataset[2] which is a subset of Imagenet[5] for evaluation. Imagenette consists of only ten classes, and each class has roughly 1000 images for training. We use only top-1 accuracy to evaluate the models, as top-5 accuracy could be misleading due to only ten classes being present.

Tiny configuration
Model configuration: $d_{model} = 512$, $N = 3$ and $h = 8$.
Image configuration: Image size = $(184, 184)$. Patch size = $(8, 8)$.

Moderate configuration
Model configuration: $d_{model} = 1024$, $N = 6$ and $h = 16$.
Image configuration: Image size = $(368, 368)$. Patch size = $(16, 16)$.

We train each model for 200 epochs with a learning rate of 1e-3 and a plateau scheduler to dynamically change the learning rate. Adam optimizer and cross-entropy loss function were used for training. The trained checkpoints can be found here [3].

[1] `https://github.com/google-research/bigbird`

[2] `https://github.com/fastai/imagenette`

[3] `https://drive.google.com/drive/folders/1tJhMrptkGyYOKNCNLOpR4gHJa5HNv4py?usp=sharing`

Model	Parameters(in millions)
ViT-Tiny	7.5
BBViT-Tiny	6.8
ViT-Moderate	51.7
BBViT-Moderate	45.5

Table 1: Total number of trainable parameters comparison

5 Results

To better understand the effect of each attention, we train a total of 6 additional variants. Each of these variants has the same transformer parameters, so the only difference between them is the self-attention used in the models. The compiled test results on the Imagenette dataset can be seen in table 2 and table 3. For the moderate configuration, we do not test on combinations of sparse attention patterns due to time and resource constraints.

Model	Top-1 Acc
ViT	53.33
BBViT	64.58
BBViT-R	59.37
BBViT-W	64.25
BBViT-G	62.29
BBViT-R+W	63.39
BBViT-W+G	**65.66**
BBViT-R+G	62.39

Table 2: Top-1 accuracy for models trained with tiny configuration. R-Random attention. W-Window attention. G-Global attention

Model	Top-1 Acc
ViT	**75.89**
BBViT	73.85

Table 3: Top-1 accuracy for models trained with moderate configuration. R-Random attention. W-Window attention. G-Global attention

6 Discussions

Observing table 2, we see that BBViT variants outperform ViT by a big margin(~6-12 points). BBViT variant trained with window and global sparse attention patterns performs the best. BBViT trained with pure random attention pattern has the least performance among BBViT variants.

Surprisingly, full self-attention performs worse by 6 points compared to random attention, which leads to the question of whether full self-attention is detrimental to the performance of small ViTs with limited data. From table 2, we also note that adding additional sparse attention patterns is not conducive to efficient learning, or in other words, more attention does not always translate to better performance. The better performance of BBViT variants could be attributed to the subtly introduced inductive bias via the attention patterns, especially the window attention, which introduces the notion of locality, one of the primary inductive biases of CNNs.

On the contrary, ViT outperforms BBViT in the moderate configuration by 2 points. From table 3 and table 2, we can say that full self-attention needs a bigger host ViT for it to facilitate learning, and sparse self-attention learns visual features better than full self-attention in smaller ViT variants.

7 Future Work

Due to time and memory constraints, we could not use Tiny-Imagenet. Verifying whether the reported behaviour of sparse self-attention can be seen in Tiny-Imagenet would strengthen the argument for sparse self-attention patterns. Since the best-performing variant has a strong sense of locality, it is important to experiment with convolution/window-based ViTs like CvT[23], SWIN[14]. The authors of BIGBIRD[25] proposed sparse self-attention patterns for NLP tasks, which might not translate well to visual features. So, experimenting with sparse self-attention patterns for computer vision tasks could lead to an ideal sparse self-attention for small ViT models.

8 Conclusion

In this paper, we establish proof of concept for the edge in the performance of sparse self-attention patterns in small ViT variants compared to full self-attention patterns. Of all the experimented sparse attention patterns, the best performing variant is the window+global sparse attention pattern, which outshines full-self attention by 12 points. Further, we also note that random sparse attention performs better than full-self attention for small ViTs. Our experiments also point out that more attention does not necessarily lead to better performance. As per the general expectation, full self-attention eventually beats sparse self-attention models as the model grows.

References

[1] Alaaeldin Ali et al. "Xcit: Cross-covariance image transformers". In: *Advances in neural information processing systems* 34 (2021), pp. 20014–20027.

[2] Nicolas Carion et al. *End-to-End Object Detection with Transformers*. 2020. DOI: 10.48550/ARXIV.2005.12872. URL: https://arxiv.org/abs/2005.12872.

[3] Tianlong Chen et al. "Chasing sparsity in vision transformers: An end-to-end exploration". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 19974–19988.

[4] Fan Chung and Linyuan Lu. "The average distances in random graphs with given expected degrees". In: *Proceedings of the National Academy of Sciences* 99.25 (2002), pp. 15879–15882.

[5] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[6] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[7] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2020. DOI: 10.48550/ARXIV.2010.11929. URL: https://arxiv.org/abs/2010.11929.

[8] Benjamin Graham et al. "Levit: a vision transformer in convnet's clothing for faster inference". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 12259–12269.

[9] Shlomo Hoory, Nathan Linial, and Avi Wigderson. "Expander graphs and their applications". In: *Bulletin of the American Mathematical Society* 43.4 (2006), pp. 439–561.

[10] Eytan Katzav, Ofer Biham, and Alexander K Hartmann. "Distribution of shortest path lengths in subcritical Erdős-Rényi networks". In: *Physical Review E* 98.1 (2018), p. 012301.

[11] Ya Le and Xuan Yang. "Tiny imagenet visual recognition challenge". In: *CS 231N* 7.7 (2015), p. 3.

[12] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. "Vision transformer for small-size datasets". In: *arXiv preprint arXiv:2112.13492* (2021).

[13] James Lee-Thorp et al. "Fnet: Mixing tokens with fourier transforms". In: *arXiv preprint arXiv:2105.03824* (2021).

[14] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.

[15] Cedric Renggli et al. "Learning to Merge Tokens in Vision Transformers". In: *arXiv preprint arXiv:2202.12015* (2022).

[16] Michael S Ryoo et al. "TokenLearner: What Can 8 Learned Tokens Do for Images and Videos?" In: *arXiv preprint arXiv:2106.11297* (2021).

[17] Daniel A Spielman and Shang-Hua Teng. "Spectral sparsification of graphs". In: *SIAM Journal on Computing* 40.4 (2011), pp. 981–1025.

[18] Chen Sun et al. "Revisiting unreasonable effectiveness of data in deep learning era". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 843–852.

[19] H Touvron et al. "HJ egou,"Training data-efficient image transformers & distillation through attention,"". In: *arXiv preprint arXiv:2012.12877* (2020).

[20] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[21] Alex Wang et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding". In: *arXiv preprint arXiv:1804.07461* (2018).

[22] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks". In: *nature* 393.6684 (1998), pp. 440–442.

[23] Haiping Wu et al. "Cvt: Introducing convolutions to vision transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 22–31.

[24] Zhuofan Xia et al. "Vision transformer with deformable attention". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 4794–4803.

[25] Manzil Zaheer et al. "Big bird: Transformers for longer sequences". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17283–17297.