Vision transformers with Inductive Bias introduced through self-attention regularization

18D5251 Luiz Henrique Barbosa Mormille Prof. Masayasu Atsumi

Abstract

In recent years, the Transformer achieved remarkable results in computer vision related tasks, matching, or even surpassing those of convolutional neural networks (CNN). However, unlike CNNs, those vision transformers lack strong inductive biases and, to achieve state-of-the-art results, rely on large architectures and extensive pre-training on tens of millions of images. Introducing the appropriate inductive biases to vision transformers can lead to better convergence and generalization on settings with fewer training data. This work presents a novel way to introduce inductive biases to vision transformers: self-attention regularization. Two different methods of self-attention regularization were devised. Furthermore, this work proposes ARViT, a novel vision transformer architecture, where both self-attention regularization methods are deployed. The experimental results demonstrated that self-attention regularization leads to better convergence and generalization, especially on models pre-trained on mid-size datasets.

Keywords Inductive Bias · Vision Transformer · Self-supervised Learning

1 Introduction

1.1 Motivation

In the context learning systems, inductive biases are the set of assumptions made by a model in order to predict outputs from unseen data [1]. In other words, a potentially infinite number of hypotheses (combination of weight parameters) can exist for a finite set of training examples, and not all the hypotheses generalize well when tested on unseen data [**?**]. Learning models inherently tend to be biased toward a group of hypotheses encompassed in their architecture, and those hypotheses can be called an inductive bias.

In deep learning, the selection of models with appropriate inductive biases for the task at hand plays a crucial role in the effort to obtain better generalization [**?**, **?**]. That is particularly true for settings where a small amount of training data is available. Accordingly, models with weak inductive biases tend to converge to the local optima when trained with limited data, therefore being unstable to changes in the initial states.

The success of Convolutional Neural Networks (CNN) is often attributed to its inductive biases, such as locality and translation equivariance. In recent years, studies with the Transformer [2] have also achieved remarkable success in computer vision [3–8]. Nevertheless, state-of-the-art performance is only attained with very large architectures pre-trained on tens of millions of labeled data [9]. For example, the Vision Transformer (ViT) by Dosovitskiy *et al.* [10], when pre-trained on hundreds of millions of images, was able to achieve excellent results compared to CNNs on mid-sized or small image recognition tasks. The data hunger of vision transformers is precisely due to the fact that they have fewer inductive biases than CNNs, allowing them to search the hypothesis space more freely [11, 12]. Hence, when trained on small or medium scale datasets, such vision transformers will often converge to a local optima and generalize poorly on unseen data. Therefore, such property of vision transformers eventually hinders its training on environments with low resources, due to the heavy computational cost for training a model with hundreds of millions of parameters on tens of millions of images.

Introducing appropriate inductive biases to vision transformers can lead to better convergence and generalization. One of the most common approaches is to introduce inductive biases from CNNs to vision transformers by combining convolution layers with self-attention layers [12–17]. Distilling convolved knowledge [18] and applying self-attention to local neighbors [3] also aim to address this challenge. Another approach revolves around adaptations to the vision transformer architecture. The Pyramid Vision Transformer (PVT) [19] progressively shrinks the transformer architecture incorporating the pyramid structure of CNNs. The Swin Transformer [20] has an hierarchical architecture and uses shifted windows to compute representations from images. Nevertheless, in order to match or surpass the performance of CNNs, such vision transformers are still pre-trained on large scale datasets, like the ImageNet-21K with 14.2 million images [21], or the JFT-300M with 303 million images [22].

Therefore, the main motivation of this work is to tackle the lack of inductive biases on vision transformers and the high computational cost necessary to overcome it (associated with the training of large-capacity models on large-scale datasets).

1.2 Research Aims and Objectives

Based on the aforementioned motivations, the main objectives of the research presented in this dissertation are:

1. To tackle the vision transformers' lack of inductive biases in or to improve its ability to generalize well when pre-trained on smaller volumes of data.
2. To improve the ability of smaller capacity vision transformers to generalize well.
3. To reduce the computational cost associated with training high performance vision transformers.
4. To improve the ability of vision transformers when pre-trained on unlabeled data.

2 Related Work

As demonstrated in Dosovitskiy *et al.* [10], overcoming the lack of inductive biases on the ViT could be directly achieved via large architectures trained on large-scale datasets and with longer training schedules. The ViT have three variants —ViT-Base, ViT-Large and ViT-Huge— that can add to up to 632M parameters. The state-of-the-art results were attained pre-training the models on ImageNet-21k (with 21k classes and 14M images) or the JFT (with 18k classes and 303M images). Hence, overcoming the lack of inductive biases through this approach comes at a high computational cost, which can be a limiting factor on environments with fewer resources. Different approaches were also proposed in order to overcome the vision transformers' lack of inductive biases. As is well known, the most common approach is to combine self-attention layers with convolution layer [12, 13, 15], with a variety of models proposed in the literature [17, 20, 23]. Alternatively, instead of combining convolution and self-attention layers, the work of Touvron *et al.* [18] proposed to distill convolved knowledge from a CNN teacher network to a vision transformer student network. Another approach consists of applying self-attention only to local neighbors [3, 5]. Inductive biases can also be introduced in vision transformers through architecture design, such as on the PVT [19] and the Twins [24]. The Swin Transformer [20], for example, addressed this issue by constructing an hierarchical representation starting from small-sized patches that are gradually merged with their neighboring patches in deeper transformer layers. Then, self-attention is locally computed on non-overlapping windows, thus introducing inductive biases of locality, hierarchy and translation invariance [25, 26]. In this work, the lack of inductive biases on vision transformers is addressed by two novel self-attention regularization methods, as well as a novel small capacity vision transformer architecture.

3 Proposed Method

3.1 Basic concepts

3.1.1 Image Patches

An input image $X \in \mathbb{R}^{3\times H\times W}$ is divided into N patches, where (H, W) are the height and width of a 3 channels image. A patch is denoted as $x_n \in \mathbb{R}^{3\times C\times C}$, and (C, C) is the resolution of each patch. Then, a linear embedding is computed for each patch in order to obtain a 1D input for the vision transformer.

3.1.2 Self-attention

The attention map is computed on the multi-head attention (MHA) layer of each encoder block in the Transformer. We refer to the attention map produced by an encoder block as $A \in \mathbb{R}^{N\times N}$, where N is the number of patches into which an image is divided. A *row* $A_n \in \mathbb{R}^N$ contains the pairwise self-attention between patch x_n and all other patches.

3.2 Spatial distance based self-attention regularization

The first self-attention regularization method proposed in this work is based on the 2D spatial distance between image patches. The Manhattan distance is adopted to measure the 2D distance between patches. Firstly, a *distance matrix D* is computed by taking the pairwise distance between all image patches. Then, the *penalty matrix P* is computed from D. Its function is to determine a penalty over self-attention computation between each pair of patches based on their distances.

The newly proposed *distance loss* $\mathscr{L}_D$, takes both the *attention map A* and the *penalty matrix P*. The distance loss over a single image can be computed through the total sum of the pointwise product of the *attention map A* and the *penalty matrix P*. Then, the distance loss $\mathscr{L}_D$ is added to the original task loss $\mathscr{L}_T$ in the following fashion:

$$\mathscr{L}_L = \mathscr{L}_T + \lambda \mathscr{L}_D, \tag{1}$$

where $\lambda > 0$ is a hyperparameter assigned to control the balance between $\mathscr{L}_T$ and $\mathscr{L}_D$.

Therefore, without modifying the global self-attention computation, the distance loss acts as a self-attention regularizer in which the larger the distance between two patches x_n and x_m (taken from image X), the greater the penalty attributed over self-attention computation. In other words, the distance loss induces the attention map A to present low values in positions where the penalty matrix P presents high values. Hence, inductive bias is induced on self-attention maps by minimizing the distance loss.

3.3 Patch similarity based self-attention regularization

Our second self-attention regularization method is based on the similarity between style representations of different patches on a image. We follow the work of Gatys *et al.* [27], adpting the gram matrix as a way to obtain a style representation for each image patch. The first step of our method is then to compute a *similarity matrix S*, representing the pairwise distance between the style representation of each image patch, as shown in Fig. 1.

Then, the newly proposed *similarity loss* $\mathscr{L}_S$ takes both the *similarity matrix S* and the *attention map A*. The attention loss for a single image can be obtained from the total sum of the pointwise product between the S and A.

Similarly to the *distance loss* expressed in Section 3.2, the *similarity loss* is added to the task loss, acting as a self-attention regularizer. The intuition behind this

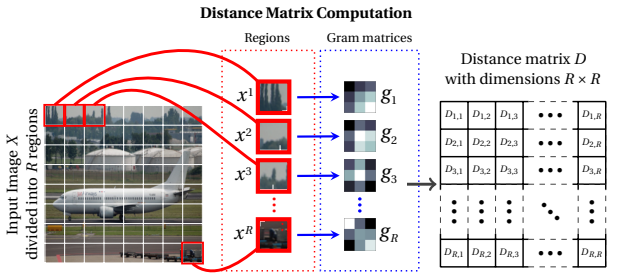

Fig. 1: Overview of the distance matrix computation. We split an image into fixed-size patches, compute the gram matrix of each individual patch, then build a distance matrix D containing the pairwise mean square errors between all gram matrices. D is a symmetric non-negative hollow matrix.

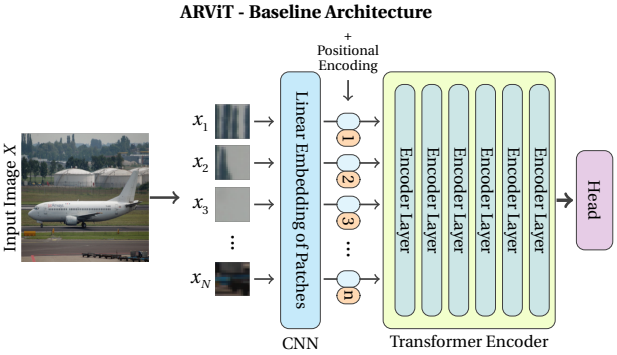

Fig. 2: ARViT architecture overview. We split an image into $C \times C$ patches, produce a linear embedding for each of them through a CNN layer with both kernel size and stride of C. We further supplement each linear embedding with a position embedding before forwarding the input sequence to the Transformer encoder. The Transformer encoder is composed of six encoder blocks, and the output of the last block is fed to a model head that addresses the self-supervised pretext-task.

method is that high self-attention values between two image patches with a similar style representation should result in a small loss; while high self-attention values between two patches whose style representation are distinct should result in a high loss. Therefore, the attention loss acts as a self-attention regularizer that induces the attention map A to present low values in positions where D presents high values, and vice-versa. Hence, without changing the global self-attention computation, inductive bias is induced through minimizing the attention loss.

3.4 ARViT Architecture

The backbone architecture of our Attention Regulated Vision Transformer (ARViT) can be interpreted as a modified and reduced version of the ViT [10]. It utilizes image patches embeddings, which are supplemented with a fixed 2D sinusoid *position embeddings* in order to retain positional information from each image patch. The ViT has 3 variants, defined by their number of encoder blocks, number of attention heads on each block and the dimension of the hidden layer. The smallest ViT variant, denoted as ViT-Base, has 12 encoder blocks with 12 heads, and the dimension of the hidden layer is 768. The first modification of the ViT is to reduce the number of encoder blocks from 12 to 6. We further reduce the dimension of the hidden layer to 516. Then, we replace the MLP with 2 layers in each encoder block with a single fully connected layer. Finally, we discard the classification token and allow the model head to address its task by using the complete output of the last encoder block. An overview of ARViT is displayed in Fig. 2.

4 Experiments

4.1 Datasets and evaluation

All models were pre-trained on a self-supervised rotation estimation task on the ILSVRC-2012 ImageNet dataset [28]. Then, all models were finetuned on 5 different downstream classification tasks: CIFAR-10, CIFAR-100, Oxford Flowers-102, Imagenette and Imagewoof, where the Top-1 accuracy was adopted as an evaluation metric.

4.2 Model performance

4.2.1 ARViT's performance

The first experiment aimed to compare the performance of ARVIT against a similar capacity vision transformer — the ViT-Tiny — with substantial improvements observed. On CIFAR-10, ARViT outperforms ViT-Tiny by roughly 10%, while on CIFAR-100 the improvement is greater than 23%. On the Flowers-102 dataset, ARViT surpasses ViT-Tiny by approximately 8%, and on Imagenette and Imagewoof the improvement is roughly 6% and 11% respectively (Table 1).

4.2.2 Spatial distance based self-attention regularization

When regularizing self-attention on ARViT using our 2D spatial distance method, significant gains in performance were observed. Namely, ARViT's accuracy was improved by further 1.63% on CIFAR-10, 0.36% on CIFAR-100, 1.47% on Flowers-102, 0.87% on Imagenette and 2.89% on Imagewoof (Table 1).

4.2.3 Patch similarity based self-attention regularization

Deploying the patch similarity based self-attention regularization method on ARViT also resulted on marginal improvements in the Top-1 accuracy on all 5 downstream classification tasks. Namely, a 5% gain on the CIFAR-10 dataset, a 13% on the CIFAR-100, a 4% gain on Flowers-102, a 5% gain on Imagenette and a 10% gain on Imagewoof (Table 1).

4.3 Computational cost

All models were trained with batch size of 80, distributed on four NVIDIA GEFORCE GTX 2080 Ti GPUs, with a capacity of 11 Gb each. Regarding the training time for ARViT, without any regularization method, each epoch was trained on an average of 16m44s. When deploying

the 2D spatial distance based self-attention regularization, each epoch was trained on an average of 18m20s. When deploying the patch similarity based self-attention regularization method to ARViT, the observed training time was 18m44s per epoch. As a comparison, the average training time for each epoch on ViT-Tiny and ViT-B [10] were 22m45s and 1h49m12s, respectively.

5 Conclusion

In this work, we tackled the lack of inductive biases in vision transformers, which is usually overcome by pre-training large architectures on large-scale datesets. We proposed two self-attention regularization methods based on the two-dimensional distance between image patches, and on the similarity between different patches of an image obtained from the distance between their gram matrices. We focused on experimenting with a low-resources set-up, and deployed our method on our proposed architecture, denoted ARViT [29]. Furthermore, all our models were pre-trained on a self-supervised task using the ILSVRC-2012 ImageNet dataset, with approximately 1.3 million images. Our experiments showed that our proposed self-regularization methods improved the performance of ARViT as far as up to 13% once finetuned on benchmark classification tasks.

References

1. T. M. Mitchell, *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research . . . , 1980.
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
3. N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*, pp. 4055–4064, PMLR, 2018.
4. I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention Augmented Convolutional Networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, Korea (South)), pp. 3285–3294, IEEE, Oct. 2019.
5. H. Zhao, J. Jia, and V. Koltun, "Exploring Self-Attention for Image Recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Seattle, WA, USA), pp. 10073–10082, IEEE, June 2020.
6. P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
7. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *arXiv:2005.12872 [cs]*, May 2020. arXiv: 2005.12872.
8. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," p. 13.
9. Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, "Visformer: The vision-friendly transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 589–598, 2021.
10. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv:2010.11929 [cs]*, Oct. 2020. arXiv: 2010.11929.
11. K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, and Y. Xu, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, 2022. ISBN: 0162-8828 Publisher: IEEE.
12. Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "Vitae: Vision transformer advanced by exploring intrinsic inductive bias," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28522–28535, 2021.
13. C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 357–366, 2021.
14. X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen, "Conditional Positional Encodings for Vision Transformers," *arXiv:2102.10882 [cs]*, Mar. 2021. arXiv: 2102.10882.
15. K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15908–15919, 2021.
16. Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "LocalViT: Bringing Locality to Vision Transformers," *arXiv:2104.05707 [cs]*, Apr. 2021. arXiv: 2104.05707.
17. Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 367–376, 2021.
18. 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*, pp. 10347–10357, PMLR, 2021.
19. 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," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.
20. 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," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.
21. 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*, pp. 248–255, Ieee, 2009.
22. C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice), pp. 843–852, IEEE, Oct. 2017.
23. B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "Levit: a vision transformer in convnet's clothing for faster inference," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12259–12269, 2021.
24. 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," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9355–9366, 2021.
25. Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, "Self-supervised learning with swin transformers," *arXiv preprint arXiv:2105.04553*, 2021.
26. Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202–3211, 2022.
27. L. A. Gatys, A. S. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," *arXiv:1508.06576 [cs, q-bio]*, Sept. 2015. arXiv: 1508.06576.
28. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, May 2017.
29. L. H. Mormille, C. Broni-Bediako, and M. Atsumi, "Regularizing self-attention on vision transformers with 2D spatial distance loss," *Artificial Life and Robotics*, vol. 27, pp. 586–593, Aug. 2022.

Table 1: Comparison between the performance of ARViT on different downstream tasks when regularized with different methods and set-ups. Furthermore, ViT-Tiny, a variant of the ViT [10], is included in the table for comparison. It has the same number of encoder blocks and attention heads as ARViT. Values in bold indicate the models with best performance on each downstream task.

Model	Regularization Type	Regularized Layer	Patch resolution	CIFAR-10	CIFAR-100	Flowers	Imagenette	Imagewoof
ViT-Tiny	-	-	-	73.30%	53.55%	73.91%	84.72%	61.77%
ARViT-Base (ours)	-	-	-	83.13%	76.27%	81.61%	91.18%	73.01%
ARViT-L1 (ours)	2D distance	1st layer	16	82.02%	75.89%	80.12%	90.58%	72.92%
ARViT-L2 (ours)	2D distance	2nd layer	16	83.63%	76.20%	81.27%	91.56%	74.81%
ARViT-L3 (ours)	2D distance	3rd layer	16	84.76%	76.63%	83.08%	92.05%	75.90%
ARViT-L4 (ours)	2D distance	4th layer	16	83.31%	76.52%	81.98%	91.88%	75.24%
ARViT-L5 (ours)	2D distance	5th layer	16	83.70%	76.01%	81.21%	91.30%	73.46%
ARViT-L6 (ours)	2D distance	6th layer	16	79.81%	74.84%	79.14%	88.77%	70.44%
ARViT-R1-1 (ours)	Region similarity	1st layer	16	88.29%	88.63%	85.58%	**95.48%**	82.31%
ARViT-R1-2 (ours)	Region similarity	2nd layer	16	86.99%	86.40%	85.15%	94.28%	79.61%
ARViT-R1-3 (ours)	Region similarity	3rd layer	16	87.17%	84.91%	85.33%	95.36%	82.20%
ARViT-R1-4 (ours)	Region similarity	4th layer	16	87.54%	86.13%	85.31%	**95.48%**	**82.93%**
ARViT-R1-5 (ours)	Region similarity	5th layer	16	**88.45%**	**89.65%**	**85.82**%	95.40%	82.70%
ARViT-R1-6 (ours)	Region similarity	6th layer	16	87.55%	87.31%	85.15%	95.78%	81.15%