Where are my Neighbors? Exploiting Patches Relations in Self-Supervised Vision Transformer

Elena Izzo[1], Guglielmo Camporese[1], Filippo Ziliotto[1], Luca Parolari[2], and Lamberto Ballan[3]

University of Padova - Visual Intelligence and Machine Perception (VIMP) Group
Padova, Italy

[1] [name.surname]@phd.unipd.it, [2] luca.parolari@math.unipd.it, [3] lamberto.ballan@unipd.it

Abstract

Vision Transformers (ViTs) enabled the use of transformer architecture on vision tasks showing impressive performances when trained on large datasets. However, on relatively small datasets, ViTs are less accurate given their lack of inductive bias. Relational Vision Transformer (RelViT) addresses this limitation by introducing a novel self-supervised learning (SSL) strategy which leverages patch relations within images to train ViTs more effectively. By optimizing all the output tokens of the transformer encoder related to image patches, RelViT enhances the training process by exploiting additional training signals. The initial research demonstrated that RelViT improves the classification performance on traditional supervised baselines. This research seeks to validate the scientific rationale behind RelViT's effectiveness by evaluating it on larger datasets and further investigating its impact on object detection task.

1. Introduction

Vision Transformer (ViT) [6] is a model recently developed to address computer vision tasks, such as image classification and object detection. It builds on the Transformer architecture [18], state of the art in natural language processing, considering patches as parts of the image such as words are parts of a sentence. The work [6] highlights the high potential of ViTs when trained on large datasets. However, the high performance are not maintained when dealing with small datasets due to the lack of inductive bias of the architecture. As a result, aiming at better generalization levels, some recent works modified the attention backbone of ViT introducing hierarchical feature representation [14], progressively tokenization of the image [20] or a shrinking pyramid backbone [19]. Other works, inspired by BERT [5], used self-supervised learning (SSL) paradigm firstly pre-training ViT on a massive amount of unlabelled data and, subsequently, training a linear classifier over the frozen feature of the model or fine-tuning the pre-trained model to a downstream task [2, 3, 1]. Into the mainstream of the research devoted to pushing beyond the limitations of the Vision Transformer, our RelViT model [1] introduced SSL tasks, easily integrated in the original architecture, that faces the problem of learning spatial relations among pairs of patches. It has already shown remarkable results addressing the image classification task on small datasets, however, experiments on larger datasets and other computer vision tasks are still missing. In this paper, we test RelViT's performance by further investigating it on ImageNet and object detection task.

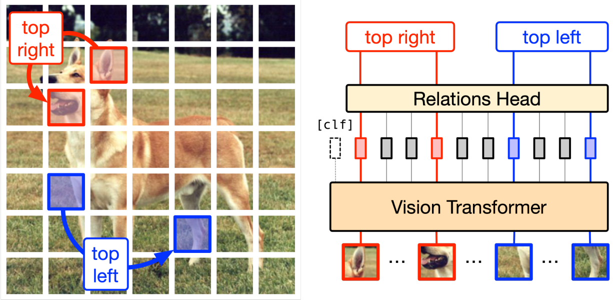

Figure 1. RelViT [1] optimizes all tokens in the image through self-supervised tasks. In the picture, tokens are optimized against *spatial relations* among patches.

2. Method

The Relational Vision Transformer model uses a standard ViT encoder, which takes image patches as input. Its output tokens are optimized through specialized heads designed for solving self-supervised learning tasks based on spatial relations among input patches. Since all heads share the same backbone, the model can solve multiple SSL tasks in parallel. The SSL signals can be leveraged by adding or removing task-related heads and their losses in summating the total loss. Thanks to this mechanism, RelViT can be trained using only self-supervision (a *pre-training* stage), full supervision, or combining the two at the same time (*downstream-only*). During the *pre-training*, it could be beneficial to use the mega-patches for increasing the contextual information in each token, and it is possible to permute the input patches without losing in generability. An overview of the RelViT idea from [1], where you can find more details, is shown in Figure 1.

3. Experiments

3.1. Datasets and Implementation Details

We report some results obtained in [1] on standard image benchmarks for classification: CIFAR-10 [10], SVHN [15], CIFAR-100 [10], Flower-102 [17], TinyImagenet [11], and ImageNet-100 (a subset of 100 labels from ImageNet) adding the experiment carried out on ImageNet-1K [4]. Moreover, we tested RelViT on various benchmarks for object detection: VOC-2007 [7], COCO [13], KITTI [8] (randomly splitted into 75% train and validation), and SVHN [16], with a number of training samples from 5k to 118k. We used the backbones ViT-S [6], Swin-T [14], and T2T-ViT-14 [20] with the same model configurations as in [1].

We developed our model in PyTorch, and the code will be available upon acceptance. For all the experiments, we investigated RelViT using the additional *spatial relations* and *absolute positions* heads for the self-supervised component (as suggested in [1]), and a single GPU for performing each experiment. For the classification benchmarks, we used the same training details reported in [1] pre-training the model from scratch on SSL tasks and subsequently fine-tuning it on classification. For the object detection task, we combined Feature Pyramid Network (FPN) [12] (to extract features from the backbone) and Mask R-CNN [9] as the detection framework.

3.2. Experimental Results

We investigated RelViT on larger datasets testing its generalization abilities on ImageNet-1K. Table 1 reports the results on various datasets for classification. The RelViT approach improves the supervised baselines on small and huge datasets, gaining +2.94% on ImageNet-1K and proving its effectiveness in this complex scenario. It is worth noticing that the 68.28% of accuracy has been obtained using just 1 GPU and a small batch size of 256. For the object detection task, we carried out experiments under both the *downstream-only* scenario and pre-training on SSL tasks and subsequently fine-tuning for object detection using both the patches' permutations and the mega-patches as hyper-parameters. Table 2 reports RelViT results on VOC-2007 using multiple backbones and hyper-parameters, whereas Table 3 shows the results on various benchmarks using ViT-S as the backbone. Independently by the dataset, hyper-parameters, and backbone, the RelViT approach always outperforms the supervised baselines up to +8.64% in mAP_{50} on VOC-2007 using Swin as backbone and +2.32% in mAP_{box75} on COCO.

	Backbone	Supervised	RelViT	Improv.
CIFAR-10	ViT-S/4	86.09 ±0.46	**90.23** ±0.09	**+4.14 ↑**
SVHN	ViT-S/4	96.01 ±0.07	**97.14** ±0.03	**+1.13 ↑**
CIFAR-100	ViT-S/4	59.19 ±0.84	**64.99** ±0.46	**+5.85 ↑**
Flower-102	ViT-S/32	42.08 ±0.29	**45.78** ±0.75	**+3.70 ↑**
TinyImagenet	ViT-S/8	43.19 ±0.78	**51.98** ±0.20	**+8.79 ↑**
Imagenet100	ViT-S/32	58.04 ±0.91	**66.46** ±0.45	**+8.42 ↑**
Imagenet	ViT-S/32	65.34	**68.28**	**+2.94 ↑**

Table 1. Comparison between the RelViT model and the supervised ViT baselines on several datasets. RelViT is pre-trained on SSL tasks and subsequently finetuned for classification, whereas the baselines are trained on classification.

Backbone	Method	Upstream	Perm.	Mega-patch	mAP	mAP_{50}	mAP_{75}
ViT [6]	Supervised	-	-	-	14.31	30.04	11.59
	RelViT	✗	✗	✗	14.59	30.86	15.16
	RelViT	✓	✗	✗	15.16	31.51	12.78
	RelViT	✓	✗	✓	**16.87**	**34.06**	**14.73**
	(**Improv.**)				(**+2.56 ↑**)	(**+4.02 ↑**)	(**+3.14 ↑**)
Swin [14]	Supervised	-	-	-	15.43	32.40	12.73
	RelSwin	✗	✗	✗	12.98	28.37	10.21
	RelSwin	✓	✓	✗	20.98	40.02	19.60
	RelSwin	✓	✓	✓	**21.65**	**41.04**	**20.65**
	(**Improv.**)				(**+6.22 ↑**)	(**+8.64 ↑**)	(**+7.92 ↑**)
T2T-ViT [20]	Supervised	-	-	-	14.20	30.09	11.32
	RelT2T-ViT	✗	✗	✗	14.74	30.38	12.54
	RelT2T-ViT	✓	✓	✗	15.85	32.17	13.45
	RelT2T-ViT	✓	✓	✓	**16.61**	**33.60**	**14.53**
	(**Improv.**)				(**+2.41 ↑**)	(**+3.51 ↑**)	(**+3.21 ↑**)

Table 2. RelViT results on VOC-2007 using different backbones. *Upstream* defines the usage of SSL tasks during pre-training instead of jointly learning them along the downstream task. *Perm.* and *Mega-patch* set the use of permutation and mega-patches.

Dataset	Method	Upstream	Perm.	Mega-patch	mAP_{box}	mAP_{box50}	mAP_{box75}
COCO mini-val	Supervised	-	-	-	20.43	35.36	20.27
	RelViT	✓	✗	✗	**22.44**	**37.35**	**22.59**
	RelViT	✓	✗	✓	22.03	36.79	22.37
	(**Improv.**)				(**+2.01 ↑**)	(**+1.99 ↑**)	(**+2.32 ↑**)
SVHN	Supervised	-	-	-	32.55	74.01	21.73
	RelViT	✗	✗	✗	34.42	76.52	24.51
	RelViT	✓	✗	✗	32.30	74.84	20.68
	RelViT	✓	✗	✓	**34.36**	**76.71**	**24.10**
	(**Improv.**)				(**+1.81 ↑**)	(**+2.70 ↑**)	(**+2.37 ↑**)
KITTI	Supervised	-	-	-	32.17	55.73	33.26
	RelViT	✗	✗	✗	29.82	52.81	30.50
	RelViT	✓	✗	✗	**37.02**	**63.52**	**35.79**
	RelViT	✓	✗	✓	33.86	58.39	34.40
	(**Improv.**)				(**+4.85 ↑**)	(**+7.99 ↑**)	(**+2.53 ↑**)

Table 3. RelViT vs. supervised ViT baselines on various datasets for object detection. Columns details in the caption of Table 2.

4. Conclusion

In this work, we have investigated the effectiveness of RelViT on ImageNet-1K and several datasets for object detection task. The results show the generalizability and outperforming properties of RelViT, expanding its scope beyond image classification on small datasets.

References

[1] Guglielmo Camporese, Elena Izzo, and Lamberto Ballan. Where are my neighbors? exploiting patches relations in self-supervised vision transformer. *British Machine Vision Conference (BMVC)*, 2022. 1, 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the North American Chapter of the Association for Comput. Linguistics (NAACL)*, 2019. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021. 1, 2

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 2

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[10] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 2

[11] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. 2015. 2

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[15] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 2

[16] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 2

[17] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *Proc. of Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP)*, 2008. 2

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1

[19] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[20] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2