How To Train Vision Transformer on Small-Scale Datasets?

Hanan Gani, Muzammal Naseer, Mohammad Yaqub

{hanan.ghani, muzammal.naseer, mohammad.yaqub}@mbzuai.ac.ae

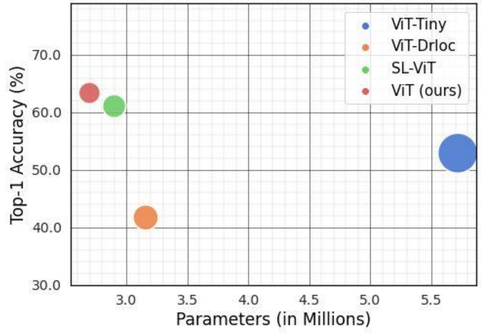
Paper-ID: 0731

https://github.com/hananshafi/vits-for-small-scale-datasets

Introduction

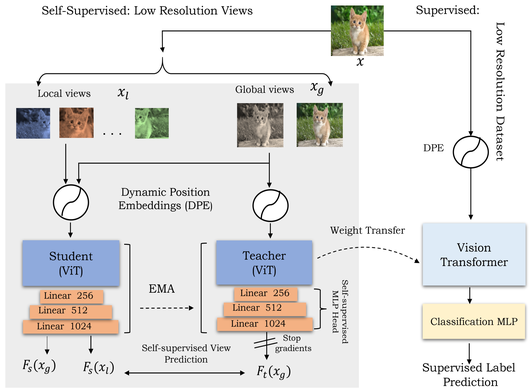
- Vision Transformers, whether monolithic or non-monolithic, both suffer when trained from scratch on small datasets.
- ViT's lack locality, inductive biases and hierarchical structure of the representations which is commonly observed in the Convolutional Neural Networks. As a result, ViTs require large-scale pre-training to learn such properties from the data for better transfer learning to downstream tasks.
- We show that inductive biases can be learned directly from the small dataset through self-supervision, thus serving as an effective weight initialization for finetuning on the same dataset
- Our proposed self-supervised inductive biases improve the performance of ViTs on small datasets without modifying the network architecture or loss functions

Highlights

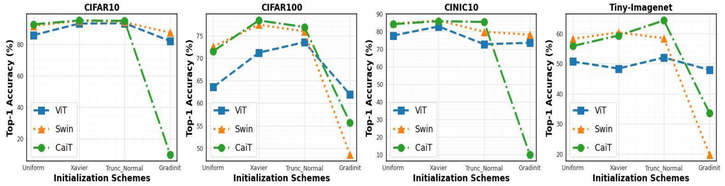


- Our approach is simple in nature and yet outperforms [1, 2, 3] by notable margins both in terms of trainable parameters and generalization (top-1 accuracy) on Tiny-ImageNet.

Methodology



- Pretraining: At the pretraining stage, we learn self-supervised weights by predicting low resolution local and global views via self- distillation through student and teacher networks [3].
- Finetuning: At the fine-tuning stage, self-supervised weights are used to initialize ViT for supervised learning. Our self-supervised inductive biases ease ViT optimization during supervised learning on small-scale datasets.



- Sensitivity towards initialization: We observe that ViT training can be unstable depending upon weight initializations e.g., CaiT performs poorly when initialized with Gradinit [4]. Similarly, the generalization of ViT and Swin varies a lot with different weight initialization methods.

Analysis and Results

Model	Params(M)	Tiny-Imagenet	CIFAR10	CIFAR100	CINIC10	SVHN
ResNet56	0.9	56.51	94.65	74.44	85.34	97.61
ResNet110	1.7	59.77	95.27	76.18	86.81	97.82
EfficientNet B0	4.0	55.48	88.38	61.64	75.64	96.06
ResNet18	11.6	53.32	90.44	64.49	77.79	96.78
ViT (scratch)	2.8	57.07	93.58	73.81	83.73	97.82
SL-ViT (Arxiv'21)	2.9	61.07	94.53	76.92	84.48	97.79
ViT-Drloc (NeurIPS'21)	3.15	42.33	81.00	58.29	71.50	94.02
ViT (Ours)	2.8	**63.36**	**96.41**	**79.15**	**86.91**	**98.03**
Swin (scratch)	7.1	60.05	93.97	77.32	83.75	97.83
SL-Swin (Arxiv'21)	10.2	64.95	94.93	79.99	87.22	97.92
Swin-Drloc (NeurIPS'21)	7.7	48.66	86.07	65.32	77.25	95.77
Swin (Ours)	7.1	**65.13**	**96.18**	**80.95**	**87.84**	**98.01**
CaiT (scratch)	7.7	64.37	94.91	76.89	85.44	98.13
SL-CaiT (Arxiv'21)	9.2	67.18	95.81	80.32	86.97	**98.28**
CaiT-DRLoc (NeurIPS'21)	8.5	45.95	82.20	56.32	73.85	19.59
CaiT (Ours)	7.7	**67.46**	**96.42**	**80.79**	**88.27**	98.18

Quantitative results: Our approach performs favorably well against different ViT baselines [1,2] as well as CNNs without adding any additional parameters or requiring changes to architecture or loss functions.

Data	ViT (scratch)	SL-ViT	ViT (Ours)	Swin (scratch)	SL-Swin	Swin (Ours)
CIFAR10	39.93	26.42	**26.01**	36.13	26.28	**25.38**
CIFAR100	65.04	48.56	**48.10**	53.83	47.29	**45.10**

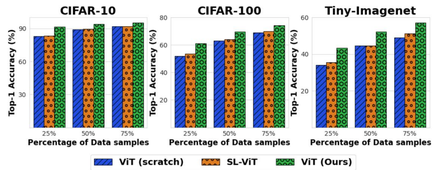
Robustness: Our training method improves model robustness against 18 natural corruptions.

Method	Aircraft	Cars
ViT-Drloc	10.40	13.82
ViT (Ours)	**66.04**	**43.89**

Finegrained datasets: Our approach outperforms the existing SOTA approach on the finegrained datasets.

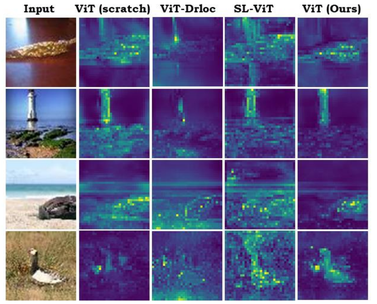
Method	Tiny-Imagenet	CIFAR10	CIFAR100
SimCLR	58.87	93.50	74.77
MOCO-V3	52.39	93.95	72.22
Ours	**63.36**	**96.41**	**79.15**

SSL comparison: Our approach performs favorably against existing SSL approaches on small-scale datasets.



Data Efficiency: Our approach consistently performs better with limited training data.

Attention to salient regions



Our proposed approach is able to capture the shape of the salient objects more efficiently with minimal or no attention to the background.

Conclusion

- In this work, we introduce an effective strategy to train Vision Transformers on small-scale low-resolution datasets without large-scale pre-training.
- We propose to learn self-supervised inductive biases directly from the small-scale datasets. We initialize the network with the weights learned through self-supervision and fine-tune it on the same dataset during the supervised training.
- We show through extensive experiments that our method can serve as a better initialization scheme and hence allows to train ViTs from scratch on small datasets.

[1] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492, 2021.
[2] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. Efficient training of visual transformers with small datasets. Advances in Neural Information Processing Systems, 34, 2021
[3] 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, pages 10347– 10357. PMLR, 2021
[4] Chen Zhu, Renkun Ni, Zheng Xu, Kezhi Kong, W. Ronny Huang, and Tom Goldstein. Gradinit: Learning to initialize neural networks for stable and efficient training. In NeurIPS, 2021