Train ViT on Small Dataset With Translation Perceptibility

BMVC 2023

Huan Chen, Wentao Wei, Ping Yao

Institute of Computing Technology, Chinese Academy of Sciences;

University of Chinese Academy of Sciences ; Chien-Shiung Wu College, Southeast University

中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

INTRODUCTION

- Vision Transformers (ViTs) exhibit deficiencies when trained on smaller datasets, specifically lacking locality, inductive biases, and hierarchical structure, which are inherent in convolutional approaches.
- Inspired by the translation equivariance of CNNs, we propose a novel self-supervised auxiliary task that enables ViTs to acquire translation perceptibility.
- Our method delivers competitive performance on small datasets across various resolutions without necessitating architectural modifications, and it can be seamlessly integrated with previous methods for enhanced utility.

Translation Perceptibility

Consider "x" and "y" as the input and output respectively, let "TS" denote the translation-set, and let "F" and "trans" stand for the model and translation function respectively.

- Translation Invariance means that the system produces exactly the same response (output) regardless of how its input is translated.

$$y = F(trans(x, TS)) = F(x)$$

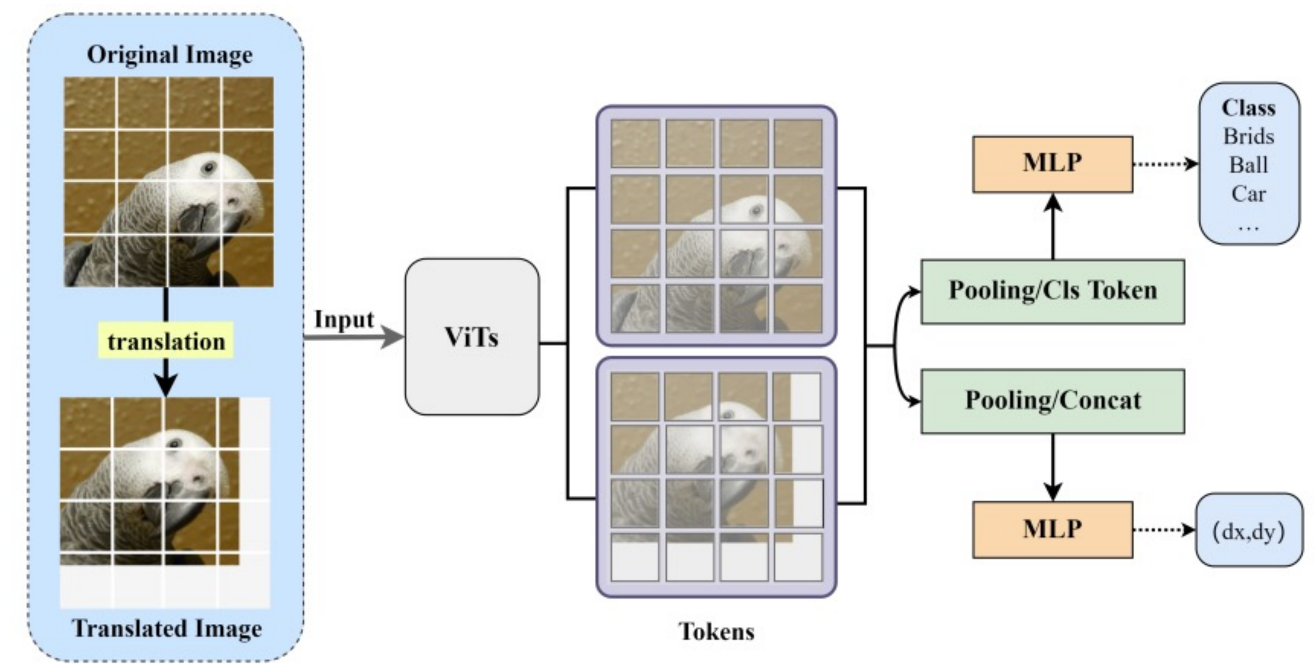
- Translation Invariance means that the system produces exactly the same response (output) regardless of how its input is translated.

$$y = F(trans(x, TS)) = trans(F(x), TS)$$

- Translation Perceptibility means that the system can work differently in different locations, but its output changes regularly with the input.

$$y = F(x)$$
$$y_{trans} = F(trans(x, TS))$$
$$TS = MLP(y, y_{trans})$$

METHODS



Pipe: In order to guide the model in learning translation perceptibility, we first apply an arbitrary translation to the input image along any direction and generate the corresponding translation labels. Subsequently, both the original and translated images are fed into the network for processing. The output tokens are utilized for classification tasks as well as translation perception prediction tasks.

RESULTS

Model	Imagenet-200	CIFAR10	CIFAR100	CINIC10	SVHN
ResNet18	53.32	90.44	64.49	77.79	96.78
ResNet56	56.51	94.65	74.44	85.34	97.61
ResNet101	59.77	95.27	76.18	86.81	97.82
EfficientNet B0	55.48	88.38	61.64	75.64	96.06
ViT(scrach)	54.07	93.58	73.81	83.73	97.82
SL-ViT	58.75	94.53	76.92	84.48	97.79
ViT-Drloc	54.44	81.00	58.29	71.50	94.02
ViT-vfsd(reproduce)	58.56	96.06	76.41	86.90	98.02
ViT-Trans(ours)	59.47	96.26	77.16	86.45	98.09
ViT-vfsd-Trans(ours)	**59.48**	**96.74**	**78.01**	**87.64**	**98.20**
Swin(scrach)	60.05	93.97	77.32	83.75	97.83
SL-Swin	64.95	94.93	79.99	87.22	97.92
Swin-Drloc	48.66	86.07	65.32	77.25	95.77
Swin-vfsd(reproduce)	64.28	96.52	80.67	87.96	98.02
Swin-Trans(ours)	62.27	96.87	80.28	88.26	98.15
Swin-vfsd-Trans(ours)	**65.05**	**97.08**	**81.25**	**88.63**	**98.17**

Quantitative results: Performance-wise, our method excels across datasets without extra inference parameters.

Model	WHU-RS19	UCMerced_LandUse	flowers102
ViT(scrach)	82.69	83.57	68.67
ViT-vfsd	89.76	91.66	69.01
ViT-Trans(ours)	91.83	94.52	73.65
ViT-vfsd-Trans(ours)	**93.27**	**95.24**	**74.72**
Swin(scrach)	85.10	88.81	79.13
Swin-vfsd	87.02	94.76	80.62
Swin-Trans(ours)	**94.71**	**97.62**	**85.37**
Swin-vfsd-Trans(ours)	**94.71**	96.43	84.66

Larger dimensions: Our method excels with larger input dimensions.

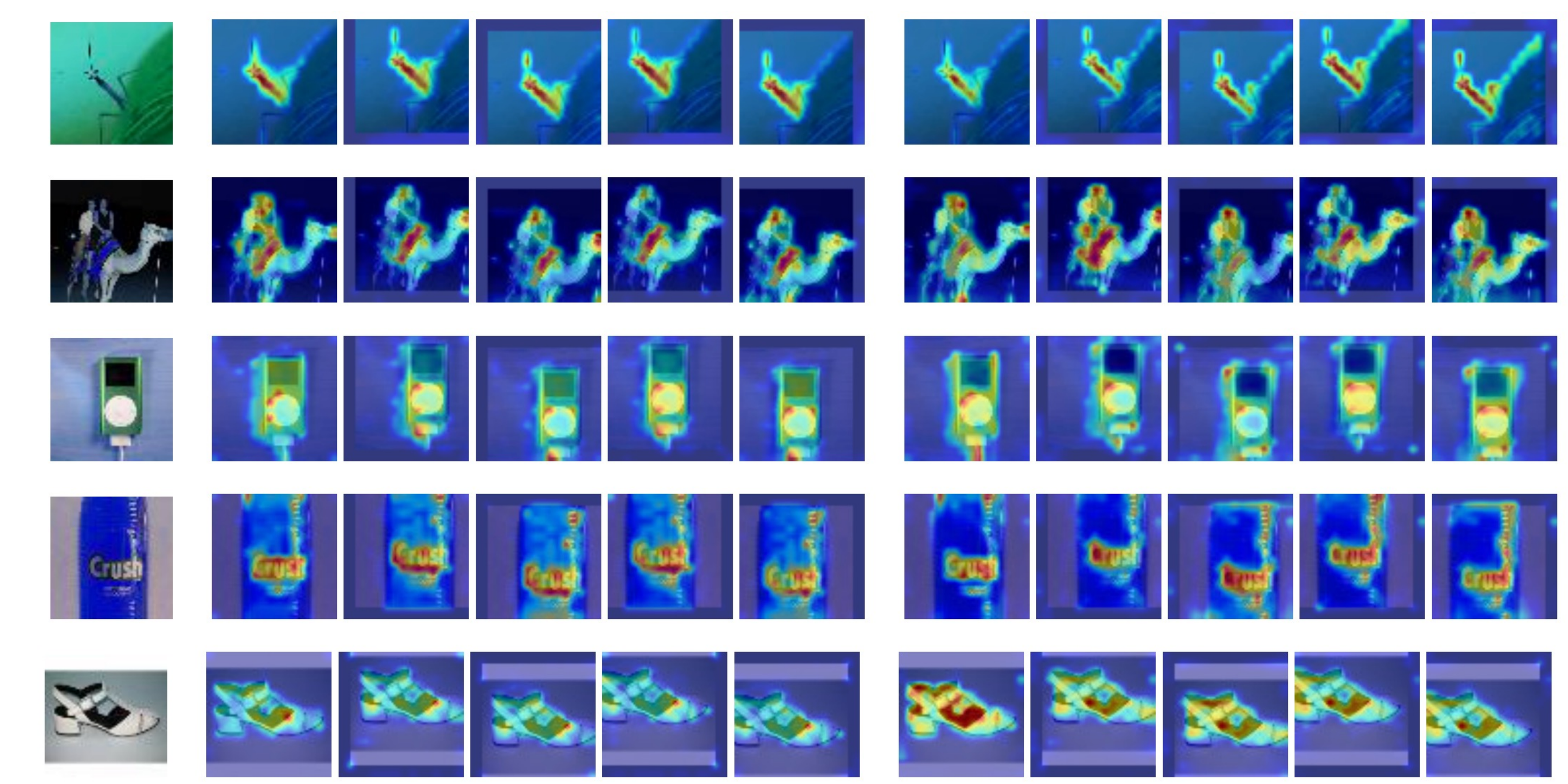
Model	Imagenet-200	CIFAR10	CIFAR100
CaiT(scrach)	58.87	94.91	76.89
CaiT-vfsd	62.18	96.50	79.64
CaiT-Trans(ours)	62.00	96.73	80.66
CaiT-vfsd-Trans(ours)	**62.84**	**97.32**	**80.90**

Models: Aside from ViT/Swin, our method remains effective on CaiT as well.

Model	Imagenet-200	Imagenet-100	Model	Imagenet-200	Imagenet-100
ViT(scrach)	54.07	62.56	Swin(scrach)	60.05	66.36
ViT-Trans(ours)	**59.47**	**65.50**	Swin-Trans(ours)	**62.27**	**69.00**
SL-ViT	58.75	66.96	SL-Swin	64.95	71.88
SL-ViT-Trans(ours)	**61.49**	**69.64**	SL-Swin-Trans(ours)	**66.80**	**74.81**
ViT-Drloc	54.44	64.52	Swin-Drloc	-	67.08
ViT-Drloc-Trans(ours)	**57.30**	**65.36**	Swin-Drloc-Trans(ours)	-	**69.96**
ViT-vfsd	58.56	65.38	Swin-vfsd	64.28	69.38
ViT-vfsd-Trans(ours)	**59.48**	**65.66**	Swin-vfsd-Trans(ours)	**65.05**	**71.30**

Extensibility: Our method can be integrated with previous state-of-the-art methods to achieve even better performance.

RESULTS



Attention to salient regions. Comparing our method (left) and vfsd[1] (right) using attention rollout on low-res Imagenet-100 samples. Our approach shows greater resilience to image translation.

CONCLUSION

- We propose a self-supervised training method for Vision Transformers (ViTs) on small datasets, guiding ViTs to learn translation perceptibility
- Our approach outperforms state-of-the-art methods on small datasets with varying resolutions, and its benefits amplify as the input size increases.
- Our approach can integrate previously advanced methods, demonstrating its extensive extensibility.

Reference

1. Gani H, Naseer M, Yaqub M. How to train vision transformer on small-scale datasets?[J]. arXiv preprint arXiv:2210.07240, 2022.