Training Vision Transformers with Only 2040 Images

Yun-Hao Cao, Hao Yu, Jianxin Wu*
State Key Laboratory for Novel Software Technology, Nanjing University
caoyh@lamda.nju.edu.cn, yuh@lamda.nju.edu.cn, wujx2001@gmail.com

TEL AVIV 2022

1. Introduction & Motivation

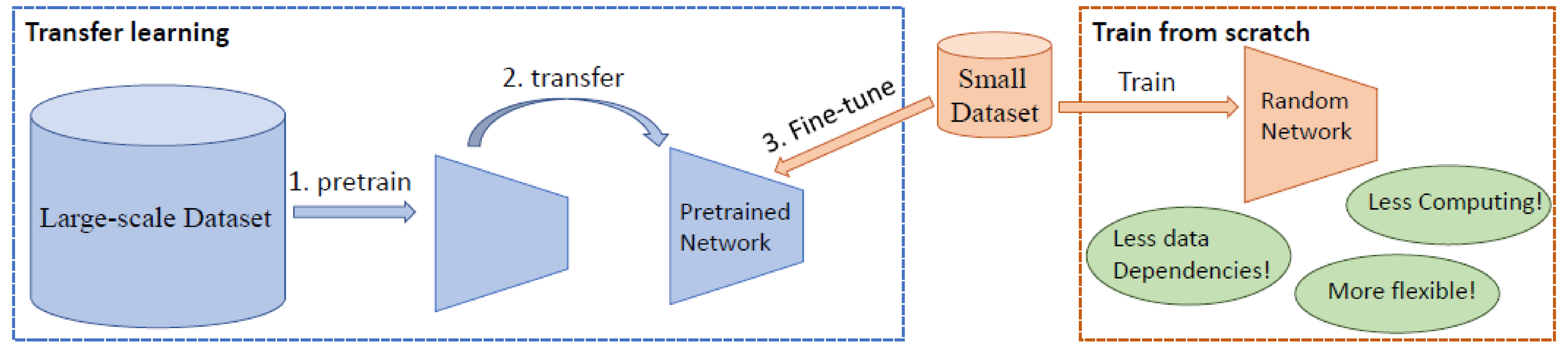
In this paper, we investigate how to train vision transformers (ViTs) with limited data **alone** (e.g., 2040 images). We propose a method called Instance Discrimination with Multi-crop and CutMix (IDMM) and achieve state-of-the-art results on 7 small datasets when training from scratch under various ViT backbones.

- ViTs are emerging as an alternative to convolutional neural networks (CNNs) for visual recognition.
- They achieve competitive results with CNNs but the lack of the typical convolutional inductive bias makes them more data-hungry than common CNNs.
- They are often pretrained on JFT-300M or at least ImageNet and few works study training ViTs with only limited data.

Key idea

- We aim to push the limit of ViTs when training from scratch on small datasets in this paper.

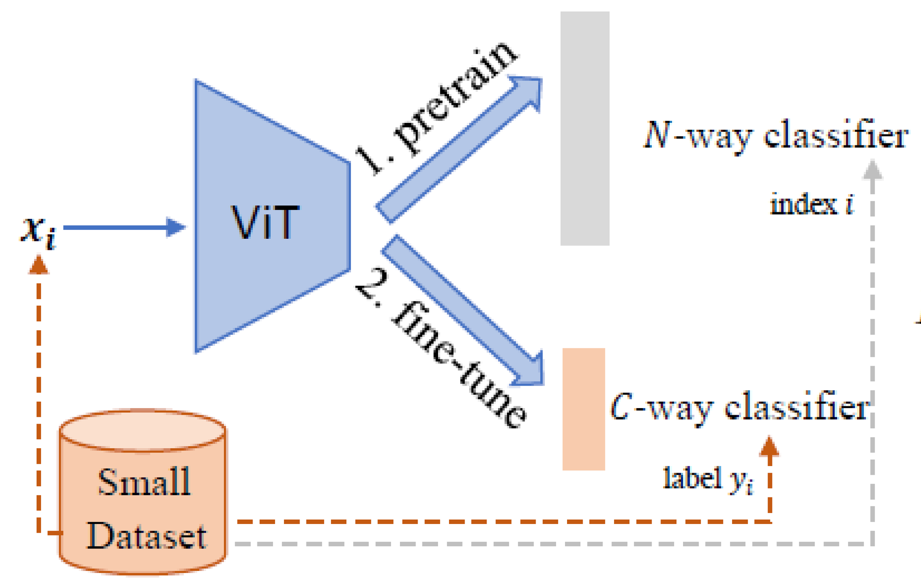
Why training from scratch?



- We cannot always rely on such large-scale datasets from the perspective of data, computing and flexibility.

2. Framework of proposed IDMM

The full learning process contains two stages and we first perform self-supervised pretraining and then supervised finetuning **on the same target dataset**.
We focus on the self-supervised pretraining stage and our method is based on parametric instance discrimination.



$$L_{\text{InsDis}}^{(i)} = -\mathbf{w}_i^T\mathbf{z}_i + \log\sum_{j=1}^{N} e^{\mathbf{w}_j^T\mathbf{z}_i}$$

Gradient analysis

✓ There exists an extremely infrequent update problem for instance discrimination when the number of instances N becomes large.

$$\frac{\partial L}{\partial \mathbf{w}_k} = -\delta_{\{k=i\}}\mathbf{z}_i + \frac{e^{\mathbf{w}_k^T\mathbf{z}_i}}{\sum_{j=1}^{N} e^{\mathbf{w}_j^T\mathbf{z}_i}}\mathbf{z}_i = (P_k^{(i)} - \delta_{\{k=i\}})\mathbf{z}_i$$

✓ CutMix and label smoothing can help update the weight matrix more frequently

$$L_{\text{InsDis}}^{(i)} = -C_i\mathbf{w}_i^T\tilde{\mathbf{z}}_{ii'} - C_{i'}\mathbf{w}_{i'}^T\tilde{\mathbf{z}}_{ii'} - C\sum_{j\neq i,i'}\mathbf{w}_j^T\tilde{\mathbf{z}}_{ii'} + \log\sum_{j=1}^{N} e^{\mathbf{w}_j^T\tilde{\mathbf{z}}_{ii'}}$$

$$\frac{\partial L}{\partial \mathbf{w}_k} = \left(P_k^{(ii')} - C_i\delta_{\{k=i\}} - C_{i'}\delta_{\{k=i'\}} - C(1-\delta_{\{k=i\}}-\delta_{\{k=i'\}})\right)\tilde{\mathbf{z}}_{ii'}$$

Why do we choose instance discrimination?

- The final learnable fc layer make it more flexible when compared to other methods.
- N is small in this paper since we focus on small datasets.
- Stability. Instability is a major issue that impacts self-supervised ViT training and the form of instance discrimination (cross entropy) is more stable and easier to optimize.

3. Experiments

Comparison of different pretraining methods

Backbone	pretraining method	pretraining epochs	Flowers	Pets	Dtd	Indoor67	CUB	Aircraft	Cars
DeiT-Tiny [31]	random init.	0	58.1	31.8	49.4	31.0	23.8	14.6	12.3
	SimCLR [8]	800	71.1	52.1	55.9	50.7	36.2	43.2	64.3
	SupCon [17]		72.3	50.3	55.6	49.3	37.8	29.4	66.2
	MoCov2 [9]		61.8	41.5	50.6	41.1	31.6	37.7	44.0
	MoCov3 [10]		67.0	52.9	52.9	49.4	20.5	32.0	53.7
	DINO [7]		64.1	51.3	51.7	46.9	41.8	**45.7**	65.3
	IDMM (ours)		**79.9**	**56.7**	**61.2**	**53.9**	**43.1**	43.2	**66.4**

- Our method performs best on all these datasets, except for aircraft.
- The advantage of our method is more obvious when the number of training images is small.

State-of-the-art results when training from scratch

Backbone	Method	Fine-tuning resolution	Fine-tuning epochs	Flowers	Pets	Dtd	Indoor67	CUB	Aircraft	Cars
DeiT-Tiny [31]	IN super.	224	200	97.3	88.6	73.2	75.6	76.8	78.7	90.3
	random init.	224	800	67.8	44.5	54.5	40.6	24.3	33.2	38.8
	IDMM (ours)	224	800	83.4	59.0	61.8	56.1	45.0	46.0	73.7
		224→448	800→100	**85.6**	**64.2**	**64.9**	**59.9**	**50.9**	**48.6**	**77.8**
DeiT-Base [31]	IN super.	224	200	97.7	91.4	74.9	78.1	81.9	82.8	92.6
	random init.	224	800	67.3	48.4	46.0	44.0	27.7	30.1	33.3
	IDMM (ours)	224	800	88.1	63.2	62.3	57.4	47.8	43.1	64.5
		224→448	800→100	**90.6**	**67.2**	**67.3**	**61.7**	**54.3**	**46.6**	**70.7**
PVTv2-B0 [35]	IN super.	224	200	98.0	90.5	75.9	76.7	81.4	88.3	92.6
	random init.	224	800	90.3	80.5	57.7	66.3	66.6	74.8	87.9
	IDMM (ours)	224	800	94.6	84.7	69.3	69.6	73.8	79.8	90.9
		224→448	800→100	**95.9**	**88.0**	**73.2**	**73.7**	**77.6**	**83.3**	**92.0**
PVTv2-B3 [35]	IN super.	224	200	98.7	93.6	78.1	80.8	85.5	91.7	94.4
	random init.	224	800	90.5	83.4	64.5	67.5	66.2	85.0	89.9
	Ours	224	800	95.9	89.8	68.9	73.2	79.0	90.5	94.0
		224→448	800→100	**96.7**	**91.9**	**71.8**	**76.3**	**82.8**	**91.8**	**94.3**
T2T-ViT-7 [41]	IN super.	224	200	97.7	90.5	75.2	76.6	79.0	83.8	92.8
	random init.	224	800	82.1	66.2	58.5	57.7	35.7	57.2	60.3
	IDMM (ours)	224	800	90.8	75.0	64.7	66.0	59.0	71.4	89.9
		224→448	800→100	**91.7**	**76.9**	**65.7**	**68.9**	**63.2**	**72.9**	**91.2**

Transferring ability on small datasets

Backbone	Pretraining Datasets	Pretraining Method	Flowers	Pets	Dtd	Indoor67	CUB	aircraft	Cars
PVTv2-B0	SIN-10k	IDMM	**93.8**	**83.6**	**66.8**	**69.4**	**70.7**	**81.3**	**87.5**
		MoCov3	91.0	81.4	62.3	66.3	63.7	74.5	86.2
		DINO	92.3	82.3	65.9	68.5	65.8	76.9	86.4
		supervised	92.9	81.7	66.1	65.9	66.6	78.7	86.0
PVTv2-B3	SIN-10k	IDMM	**95.9**	**88.4**	**70.1**	**73.6**	**76.8**	**87.5**	**92.9**
		MoCov3	93.7	87.1	66.0	70.5	63.7	82.2	92.3
		DINO	95.0	87.8	68.3	73.4	72.4	86.1	92.5
		supervised	90.9	80.9	62.9	63.3	65.6	83.8	89.7
T2T-ViT-7	SIN-10k	IDMM	**89.8**	**74.1**	**63.5**	**62.6**	**55.2**	**72.7**	**82.4**
		supervised	80.8	57.8	57.5	50.7	35.6	56.8	59.9

- ViTs have good transferring ability even when pretrained on small datasets.
- Our IDMM achieves the best results.

Application on ImageNet

Backbone	Method	Epochs	Acc. (%)
PVTv2-B0	random init.	100	68.6
	MoCov3 (SIN-10k)		68.8
	IDMM (SIN-10k)		**69.5**
	IDMM (SIN-total 10k)		**69.5**
	random init.	300	70.0
	IDMM (SIN-10k)		**70.9**
DeiT-Tiny	random init.	100	66.8
	IDMM (SIN-10k)		**67.8**
	random init.	300	72.2
	IDMM (SIN-10k)		**72.9**

- Representations learned on small datasets can serve as a good initialization even for ImageNet training.

4. Contributions & Conclusions

✓ We propose IDMM for self-supervised ViT training and achieve state-of-the-art results when training from scratch for various ViTs on 7 small datasets.

✓ We give theoretical analyses on why we should prefer parametric instance discrimination when dealing with small data from the loss perspective.

✓ we show how strategies like label smoothing and CutMix alleviate the infrequent updating problem from the gradient perspective.

✓ We analyze the transferring ability of small datasets and find that ViTs also have good transferring ability even when pretrained on small datasets.