Budgeted Training for Vision Transformer

Zhuofan Xia*[1], Xuran Pan*[1], Xuan Jin*[3], Yuan He[3], Hui Xue[3], Shiji Song[1], Gao Huang+[1,2]

[1] Department of Automation, BNRist, Tsinghua University
[2] Beijing Academy of Artificial Intelligence
[3] Alibaba Group

* Equal Contribution + Corresponding Author

May 01, 2023

Training Vision Transformers

Method	ImageNet acc. (top-1, %)	Distribution shifts
ViT-G (Zhai et al., 2021)	90.45	–
CoAtNet-7 (Dai et al., 2021)	90.88	–
Our models/evaluations based on ViT-G:		
ViT-G (reevaluated)	90.47	82.06
Best model in hyperparam search	90.78	84.68
Greedy soup	**90.94**	**85.02**

(Wortsman et al., 2022)

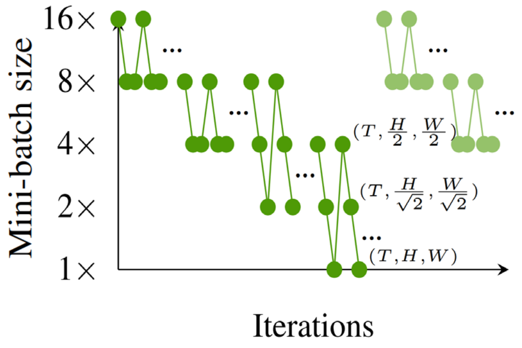
- **CoAtNet[1]**
  - **CoAtNet-7**
  - **90.88 acc1@IN-1K**
  - **2440M parameters, 2586G FLOPs**
  - **JFT-3B dataset pretrained**
  - **20.1K TPUv3-core-days**
- **Model soups[2]**
  - **ViT-G/14 with Greedy Soup**
  - **90.94 acc1@IN-1K**
  - **1843M parameters, 2860G FLOPs**
  - **JFT-3B dataset pretrained**

Training giant ViTs for superior performances often come with huge training costs.

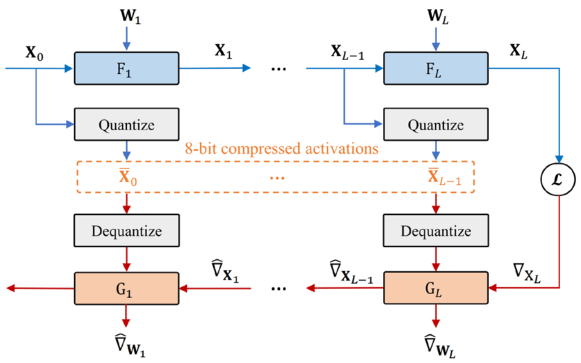
[1] Dai, Zihang, et al. "Coatnet: Marrying convolution and attention for all data sizes." Advances in Neural Information Processing Systems 34 (2021): 3965-3977.
[2] Wortsman, Mitchell, et al. "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time." International Conference on Machine Learning. PMLR, 2022.

Training Vision Transformers

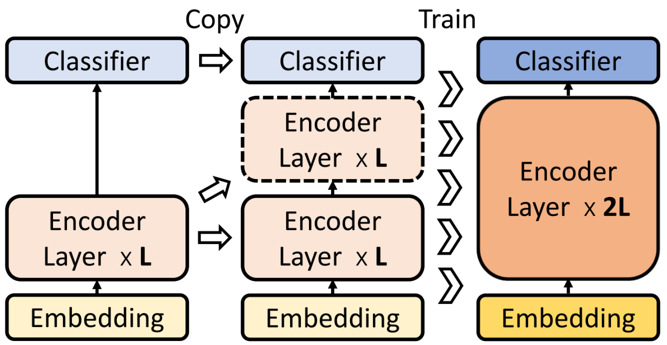
- **How to train models with less training cost?**
  - **Flexible training schedules[1]**
  - **Compressing activations to reduce memory[2]**
  - **Dynamic training-stage complexity, dropping[3] or stacking[4] layers**



(Wu et al., 2020)



(Pan et al., 2021)

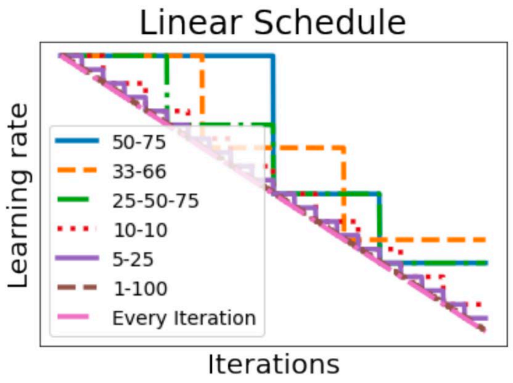


(Gong et al., 2022)

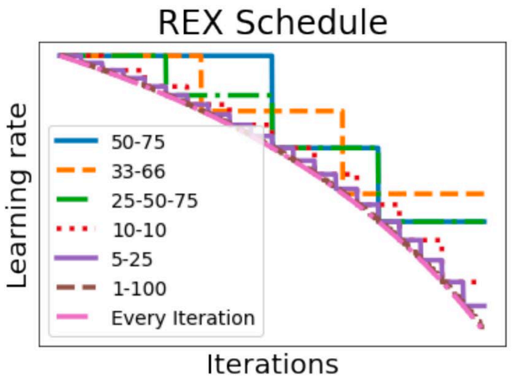
[1] Wu, Chao-Yuan, et al. "A multigrid method for efficiently training video models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[2] Pan, Zizheng, et al. "Mesa: A memory-saving training framework for transformers." arXiv preprint arXiv:2111.11124 (2021).
[3] Zhang, Minjia, and Yuxiong He. "Accelerating training of transformer-based language models with progressive layer dropping." Advances in Neural Information Processing Systems 33 (2020): 14011-14023.
[4] Gong, Linyuan, et al. "Efficient training of bert by progressively stacking." International conference on machine learning. PMLR, 2019.

Budgeted Training for ViT

- **Many labs cannot afford to train ViT under full schedule.**
  - **To train a better model under constrained training budget.**
- **Budgeted Training is proposed to fit the budget.**
  - **Learning rates schedules with fewer epochs[1,2]**
  - **Dataset pruning with fewer data[3,4]**



(Li et al., 2022)



(Chen et al., 2022)

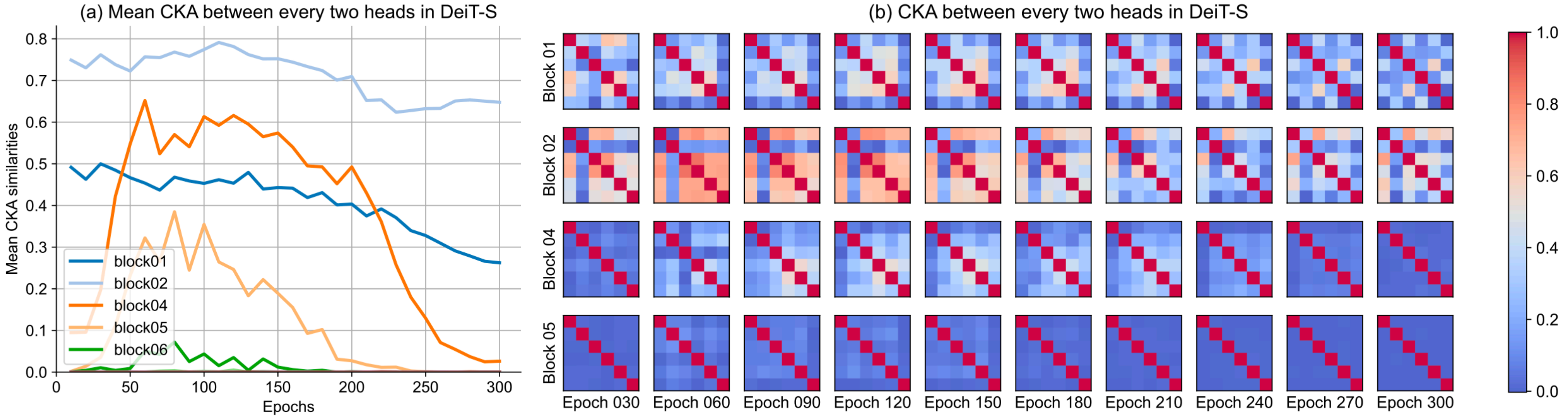
[1] Li, Mengtian, Ersin Yumer, and Deva Ramanan. "Budgeted Training: Rethinking Deep Neural Network Training Under Resource Constraints." International Conference on Learning Representations. 2020

[2] Chen, John, Cameron Wolfe, and Tasos Kyrillidis. "REX: Revisiting Budgeted Training with an Improved Schedule." Proceedings of Machine Learning and Systems 4 (2022): 64-76.

[3] Killamsetty, Krishnateja, et al. "Grad-match: Gradient matching based data subset selection for efficient deep model training." International Conference on Machine Learning. PMLR, 2021.

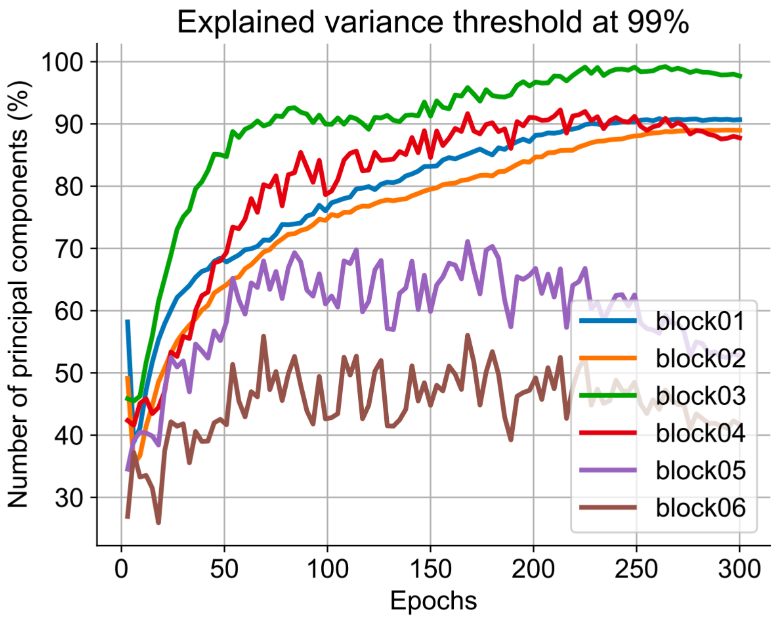
[4] Mirzasoleiman, Baharan, Jeff Bilmes, and Jure Leskovec. "Coresets for data-efficient training of machine learning models." International Conference on Machine Learning. PMLR, 2020.

Redundancy During Training ViT



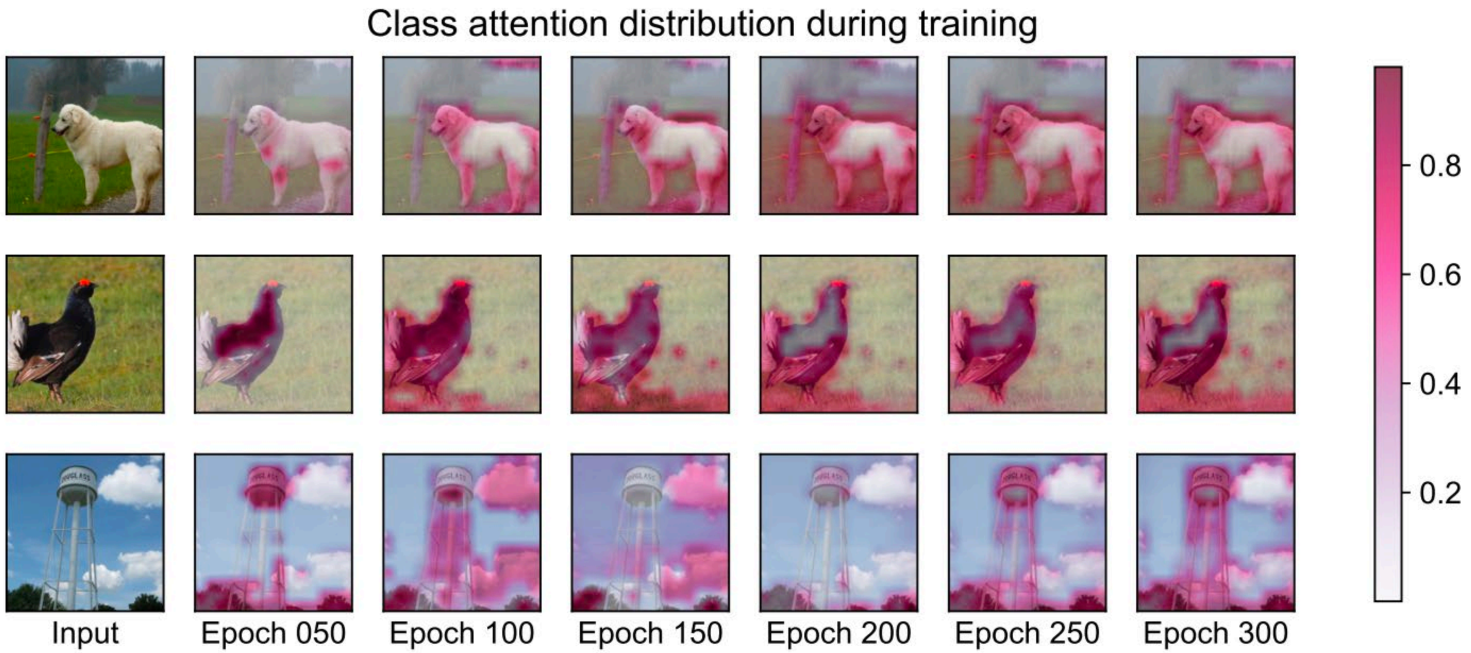
- **Redundancy in attention heads during training**
  - **CKA similarity between features on every two attention heads in DeiT-S**
  - **Features between heads are more similar at early to middle training stage**
  - **Activate fewer attention heads at early training stage!**

Redundancy During Training ViT



- **Redundancy in MLP hidden dimension during training**
  - **PCA of projected features in expanded hidden space of MLP in DeiT-S**
  - **Ratio of principal components are growing along with training**
  - **Activate fewer MLP hidden dimensions at early training stage!**

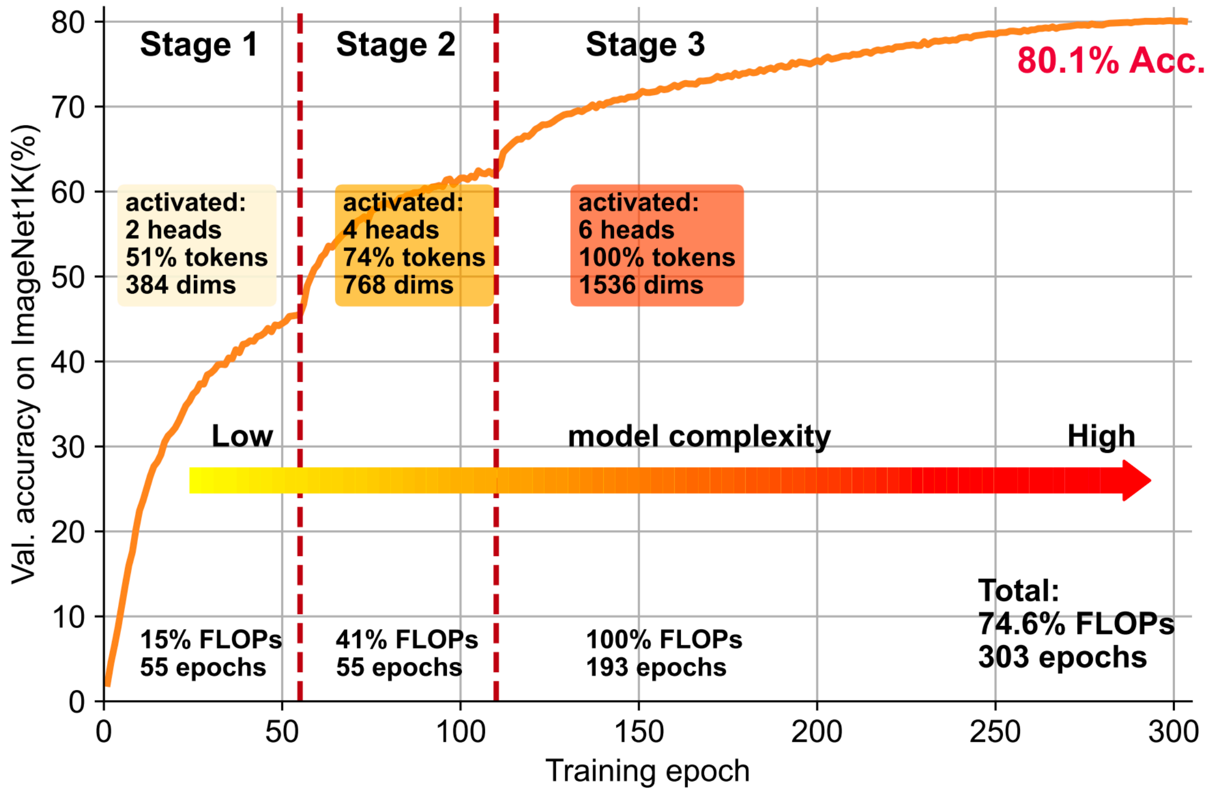
Redundancy During Training ViT



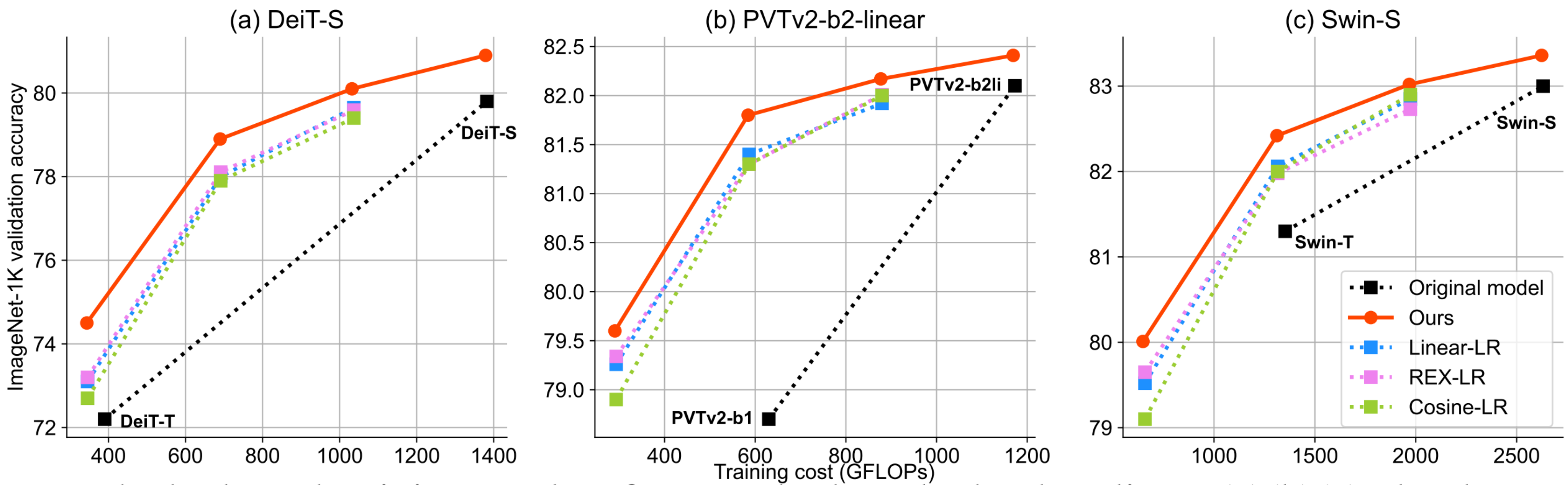
- **Redundancy in visual tokens during training**
  - **Class attention distribution visualization of DeiT-S**
  - **Patches with higher scores concentrate along with training**
  - **Activate smaller number of tokens at early training stage!**

Proposed Approach

Dynamically adjust activation rates of ViT components



Experiment Results



- **Main results on ImageNet-1K**
  - **Various model architectures, including DeiT, PVTv2, Swin Transformer**
  - **Outperforms various budgeted training baselines**
  - **Consistent improvements on different training budgets**

Experiment Results

Pretrain method	CIFAR-10 Top-1 Acc.	CIFAR-100 Top-1 Acc.
Original	98.78%	89.44%
Ours (75% budget)	98.91%	89.65%

Pretrain method	Schedule	Segmentor	mIoU
PVTv2-b2linear	40K	S-FPN	45.10
Ours (75% budget)	40K	S-FPN	45.29
Swin-S	160K	UperNet	47.64
Ours (75% budget)	160K	UperNet	47.46

Pretrain method	Schedule	AP^b	AP^b_{50}	AP^b_{75}	AP^b_s	AP^b_m	AP^b_l	AP^m	AP^m_{50}	AP^m_{75}	AP^m_s	AP^m_m	AP^m_l
PVTv2-b2linear	1x	44.1	66.3	48.4	28.0	47.4	58.0	40.5	63.2	43.6	21.5	43.0	58.2
Ours (75% budget)	1x	44.1	66.1	48.2	28.3	47.4	57.1	40.3	63.3	43.0	24.7	43.5	54.2
Swin-S	3x	48.5	70.2	53.5	33.4	52.1	63.3	43.3	67.3	46.6	28.1	46.7	58.6
Ours (75% budget)	3x	48.2	70.2	53.1	32.1	51.7	62.6	43.2	67.0	46.6	27.3	46.8	58.3

- **Competitive transfer learning results on downstream tasks**
  - **DeiT-S on CIFAR-10/100 finetuned classification**
  - **Swin-S and PVTv2-b2li on MSCOCO object detection with Mask-RCNN, 1x & 3x schedule**
  - **Swin-S and PVTv2-b2li on ADE20K semantic segmentation with various models**

Experiment Results

Method	Model	Schedule / Fraction	Training cost	Top-1 Acc.
GradMatch-PB	ResNet-18	5% of ImageNet	31.9G	45.15
Ours	DeiT-T	[11,15,17]	31.4G	**57.88**
GradMatch-PB	ResNet-18	10% of ImageNet	63.70G	59.04
Ours	DeiT-T	[8,24,39]	63.23G	**60.20**
GradMatch-PB	ResNet-18	30% of ImageNet	191.10G	68.12
Ours	DeiT-T	[22,51,127]	190.85G	**69.49**

- **Comparison over dataset pruning method**
  - **Better performances on models with similar FLOPs**
  - **Consistent improvements on various training budget**
  - **Significant margin in low budget scheme**

Thank you!

Paper

Contact: xzf20@mails.tsinghua.edu.cn

May 01, 2023