RevColV2: Exploring Disentangled Representations in Masked Image Modeling

Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

1 More Training Details

This section gives more training details on MIM pre-training and fine-tuning on downstream tasks, such as ImageNet classification, COCO detection, and ADE20K segmentation. For ImageNet experiments, the base learning rate is based on batch size 256.

1.1 Training Details on MIM pre-training.

We use the same setting for different sizes RevCol models on MIM pre-training. The detail hyper-parameters are shown in Table 1. Following exists works [1, 2], we do not use stochastic depth [3] and other regularization strategies in MIM pre-training.

config	value
optimizer	AdamW
base learning rate	1.5e-4
weight decay	0.05
optimizer momentum	β_1, β_2=$0.9, 0.95$
batch size	4096
learning rate schedule	cosine decay
warmup epochs	40
training epochs	1600
augmentation	RandomResizedCrop

Table 1: MIM Pre-training settings.

1.2 Details on Image22K intermediate fine-tuning.

We further intermediately fine-tune RevColV2 models on ImageNet-22K dataset. The fine-tuning details is shown in Table 2. The hyper-parameters generally follow [4, 2].

1.3 End-to-end fine-tuning details on ImageNet-1K.

We end-to-end fine-tune RevCol variants on ImageNet-1K after MIM pre-training and intermediately fine-tuning on ImageNet-22K. Table 3 shows the detail training settings after MIM pre-training.

We also show training settings on ImageNet-1K after ImageNet-22K fine-tuning. Table 4 gives the detailed hyper-parameters.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

config	value
optimizer	AdamW
base learning rate	2.5e-4
weight decay	0.05
optimizer momentum	β_1, β_2=0.9, 0.999
layer-wise lr decay	0.8
batch size	4096
learning rate schedule	cosine decay
warmup epochs	5
training epochs	90
augmentation	RandAug (9, 0.5)
label smoothing	0.1
mixup	0.8
cutmix	1.0
drop path	0.1 (B), 0.2 (L)
head init	0.001
ema	None

Table 2: End-to-end IN-22K intermediate fine-tuning settings.

config	value
optimizer	AdamW
base learning rate	5e-4
weight decay	0.05
optimizer momentum	β_1, β_2=0.9, 0.999
layer-wise lr decay	0.75
batch size	1024
learning rate schedule	cosine decay
warmup epochs	5
training epochs	100 (B), 50 (L)
augmentation	RandAug (9, 0.5)
label smoothing	0.1
mixup	0.8
cutmix	1.0
drop path	0.1
head init	0.001
ema	0.9999

Table 3: End-to-end ImageNet-1K fine-tuning settings

1.4 Details on ADE20K semantic segmentation

For semantic segmentation, we evaluate different backbones on ADE20K dataset. We fine-tune the pre-trained networks on ADE20K with 160,000 iterations. For UperNet framework [5], the learning rate is 4e-5 with batch size 16, using AdamW optimizer. The layer-wise learning rate decay rate is set as 0.65 for both base and large size models. The drop path rate is 0.1. For Mask2Former framework [6], the learning rate is 2e-5 with batch size 16. The drop path rate is set as 0.3 and the layer-wise learning decay rate is 0.9.

1.5 Details on COCO object detection and instance segmentation

For object detection and instance segmentation, we evaluate RevColV2 backbones with Mask R-CNN [7] and Cascade Mask R-CNN [8] detectors. We use ImageNet-1K MIM pre-trained weights as initialization and fine-tune the models with 50 epochs and a batch size of 32, learning rate 1e-4 for Mask R-CNN framework. The large scale jittering data augmentation strategy is used with scale range [0.1, 2.0]. The drop path rates for RevCOlV2 are set as 0.2 (base) and 0.3 (large) and the layer-wise learning rate decay rates are set as 0.9. For Cascade Mask R-CNN framework, we train

config	value
optimizer	AdamW
base learning rate	2.5e-5
weight decay	0.01
optimizer momentum	β_1, β_2=0.9, 0.999
layer-wise lr decay	0.9
batch size	512
learning rate schedule	cosine decay
warmup epochs	None
training epochs	30
augmentation	RandAug (9, 0.5)
label smoothing	0.1
mixup	None
cutmix	None
drop path	0.1(B), 0.2 (L)
head init	0.001
ema	0.9999

Table 4: End-to-end ImageNet-1K fine-tuning settings (after IN-22K intermediate fine-tuning).

models with 100 epochs following [9] with large scale jittering augmentation strategy. The learning rate is 1e-4 with batch size 64. We do not use soft-NMS in our experiments.

2 More Results

2.1 Compared with supervised baseline

To verify the effectiveness of RevColV2 architecture with MIM pre-train, we compare the performance on ImageNet-1K fine-tune using MIM pre-trained model weights and random initialization. We use the same setting with [1] in this supervised baseline, except additional 0.999 EMA strategy. The base/large models achieve 83.1% and 82.6% top-1 accuracy on ImageNet-1K. The MIM pre-trained RevColV2 models outperform supervised baseline by a large margin (**+1.6%** and **+3.7%**).

2.2 Linear probing results

We report the linear probing results on ImageNet-1K after pre-training for RevColV2 models and other counterparts on Table 5. Following [1], we fix the pre-trained backbone models and train a classification head for 90 epochs with LARS optimizer. We append this classification head on the last level of bottom-up columns in RevColV2. The linear probing performance of RevColV2 models surpasses other encoder only models such as SimMIM [10] and autoencoder models such as MAE [1].

Table 5: Linear probing results on ImageNet-1K dataset.

Model	Size	Target	Params	FLOPs	LIN
ImageNet-1K pre-train:					
BEIT-B [11]	224^2	DALL-E	87M	18G	56.7
SimMIM-B [10]	224^2	Pixel	88M	16G	56.7
RevColV2-B	224^2	Pixel	88M	19G	**67.7**
MAE-L [1]	224^2	Pixel	307M	62G	75.8
RevColV2-L	224^2	Pixel	327M	67G	**79.3**

References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[2] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023.

[3] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016.

[4] Yuxuan Cai, Yizhuang Zhou, Qi Han, Jianjian Sun, Xiangwen Kong, Jun Li, and Xiangyu Zhang. Reversible column networks. *arXiv preprint arXiv:2212.11696*, 2022.

[5] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[6] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021.

[7] 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.

[8] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[9] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 280–296. Springer, 2022.

[10] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[11] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.