Scaling White-Box Transformers for Vision - Appendix

Jinrui Yang[⋆1] **Xianhang Li**[⋆1] **Druv Pai**[2]

Yuyin Zhou[1] **Yi Ma**[2] **Yaodong Yu**[†2] **Cihang Xie**[†1]

[⋆]equal technique contribution, [†]equal advising

[1]UC Santa Cruz [2]UC Berkeley

A Additional Experiments and Details

A.1 Model configuration.

We provide details about CRATE-α model configurations in Table 1.

Table 1: Model configurations for different sizes of CRATE-α, parameter counts, and comparisons to CRATE models. L is depth, d is the hidden size, and K is the number of heads.

Model Size	L	d	K	CRATE-α # Params	CRATE # Params
Tiny	12	192	3	4.8M	1.7M
Small	12	576	12	41.0M	13.1M
Base	12	768	12	72.3M	22.8M
Large	24	1024	16	253.8M	77.6M
Huge	32	1280	16	526.8M	159.8M

Table 2: The comparison between CRATE-α and ViT. FLOPs and throughput are calculated based on an input size of 224x224 on an NVIDIA RTX A6000 graphics card.

Model	FLOPs (G)	#Params (M)	Throughput	Model	FLOPs (G)	#Params (M)	Throughput
CRATE-α-B/32	6.4	74.0	499	ViT-B/32	4.4	88.2	706
CRATE-α-B/16	25.8	72.3	233	ViT-B/16	17.6	86.5	375
CRATE-α-L/32	22.8	256.0	215	ViT-L/32	15.4	306.5	329
CRATE-α-L/14	119.7	253.7	56	ViT-L/14	81.1	304.1	85

A.2 Comparison of model structure with ViT.

We also compare CRATE-α to ViT in terms of computational costs, number of parameters, and inference speed. These comparisons are summarized in Table 2, where CRATE-α matches ViT's efficiency while achieving similar accuracy. With the same number of layers and embedding dimensions, CRATE-α has fewer parameters than ViT, and its FLOPs/Throughput is slightly higher.

To more accurately compare CRATE-α and ViT with larger model sizes, we conduct experiments on CRATE-α-L/16 with an image resolution of 336, nearly matching the setup of ViT-L/16. Both models use a similar amount of FLOPs: 210G for CRATE-α-L/16 compared to 191G for ViT-L/16. The throughput, or images processed per second, is also comparable at 35.53 for our model versus 35.56 for ViT-L/16. The accuracy of CRATE-α-L/16 reach 84.6%, closely approaching ViT's 85.2% under similar conditions. Meanwhile, combining the trend from Figure 1 (right) in the main paper, this narrowing performance gap from Base to Large model size suggests that CRATE-α can nearly matche ViT's performance in large-scale settings. Besides, CRATE-α inherits the mathematical interpretability

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

of the white-box models and can also achieve much better semantic interpretability evaluated by zero-shot segmentation.

A.3 Training details of CRATE-α-CLIPA models.

When employing the CRATE-α architecture to replace the vision encoder in the CLIPA [2] framework, we essentially follow the original CLIPA training recipe. The setup for the pre-training stage is presented in Table 3. During the fine-tuning stage, we made some modifications: the input image size is set to 224 × 224, the warmup steps are set to 800, and the base learning rate is set to 4e-7. When calculating the loss, we use the classification token from the vision encoder as the image feature and the last token from the text encoder as the text feature.

To explore the performance ceiling, we also train a ViT-CLIPA model from scratch. Most of the hyperparameters remain the same as those in Table 3, but there are some modifications in the pre-training stage. The batch size is set to 65,536, and the text length is set to 8 to speed up training. As with the CLIPA setup, warm-up steps are set to 3,200. Additionally, we add color jitter and grayscale augmentation, and use global average pooling instead of the classification token. These modifications help stabilize training.

Config	Value
optimizer	AdamW [5]
optimizer momentum	(0.9, 0.95)
batch size	32768
base lr	8e-6
minimal lr	0
warm-up steps	1600
schedule	cosine decay [4]
weight decay	0.2
random crop area	(40, 100)
resize method	bi-linear
temperature init	1/0.07 [1, 3]

Table 3: **Pre-training hyper-parameters for CLIPA.**

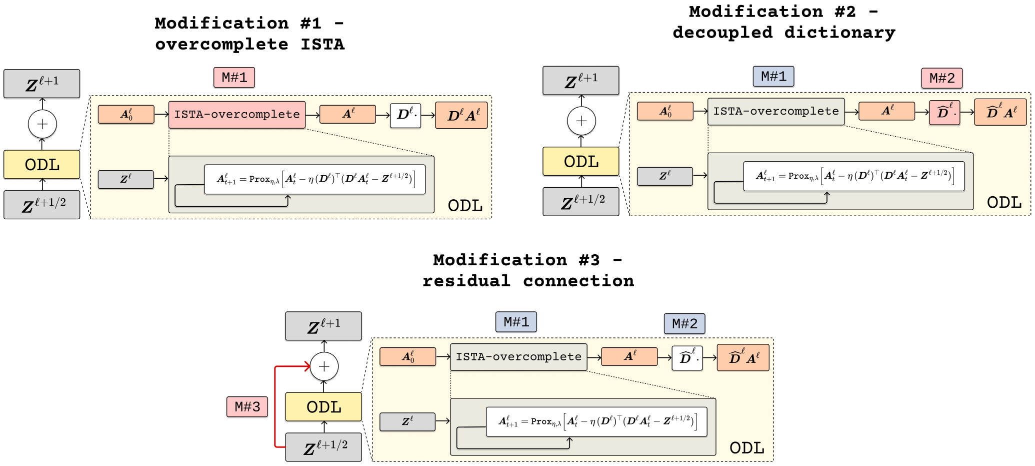
Modification #1 - overcomplete ISTA
M#1
$Z^{\ell+1}$
+
ODL
$Z^{\ell+1/2}$
A_0^ℓ
ISTA-overcomplete
A^ℓ
$D^\ell \cdot$
$D^\ell A^\ell$
Z^ℓ
$A_{t+1}^\ell = \text{Prox}_{\eta,\lambda}\left[A_t^\ell - \eta(D^\ell)^\top(D^\ell A_t^\ell - Z^{\ell+1/2})\right]$
ODL
Modification #2 - decoupled dictionary
M#1
M#2
$\widehat{D}^\ell \cdot$
$\widehat{D}^\ell A^\ell$
Modification #3 - residual connection
M#3

Figure 1: One layer of the CRATE-α model architecture (with more details for the three modifications described in Section 3.

Visualization of self-attention maps of CRATE-α. We provide visualization of attention maps of CRATE-α in Fig. 2.

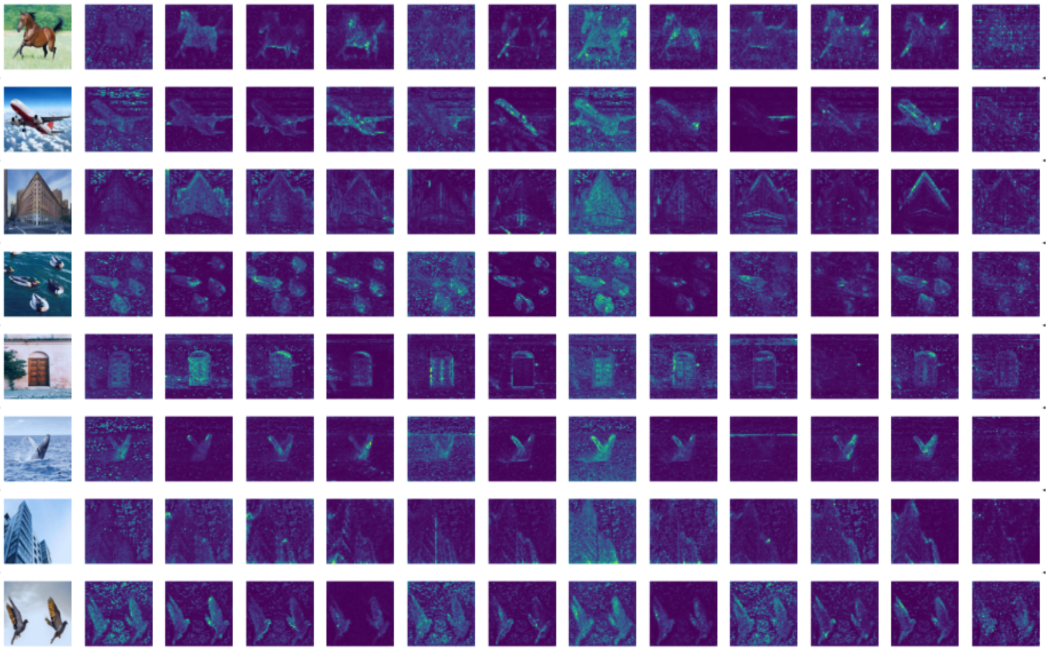

Figure 2: We visualize the self-attention maps of the CRATE-α Base model using 8×8 patches trained using classification. Similar to the original CRATE [6], our model also demonstrates the capability to automatically capture the structural information of objects. For each row, the original image is displayed on the left, while the corresponding self-attention maps are shown on the right. The number of self-attention maps corresponds to the number of heads in the CRATE-α model.

Visualization of loss curves. We visualize the training loss curves of the four models, including CRATE and its three variants, in Fig. 3. We visualize the training loss curves of CRATE-α-Base with different patch sizes in Fig. 4. In Fig. 5, we also visualize the training loss curves of models trained with efficient scaling strategy described in Section 4.4 in the main paper.

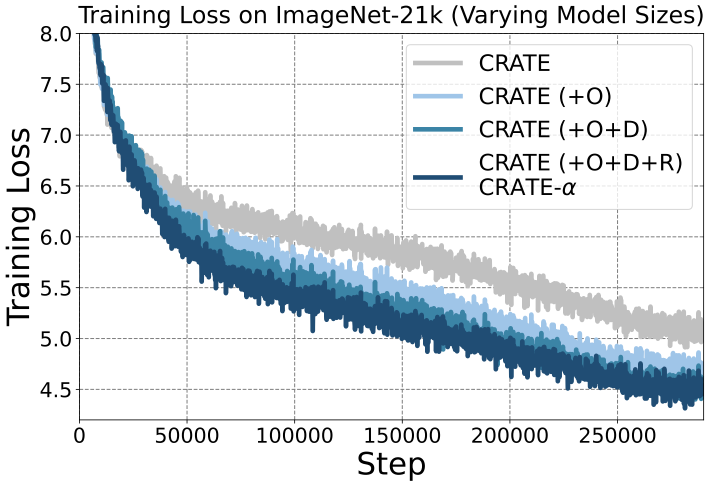

Figure 3: Training loss curves of different model architectures (mentioned in Fig. 1 in the main paper) on ImageNet-21K. The patch size is 32 for all four models shown in this figure. (+O: +overcomplete dictionary, +D: +decoupled dictionary, +R: +residual connection.)

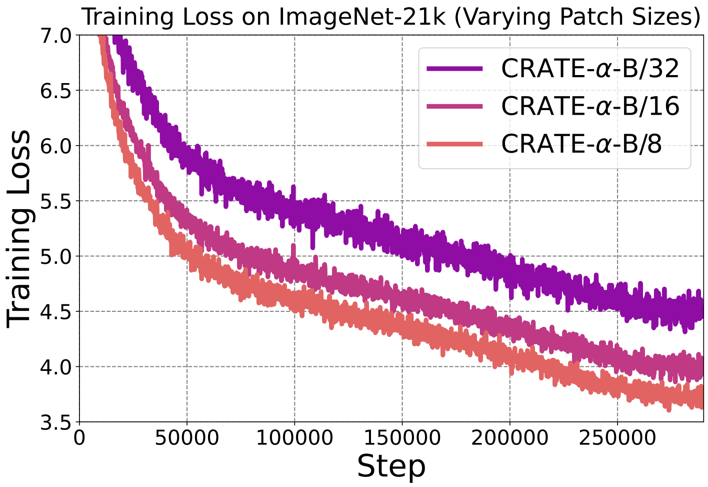

Figure 4: Comparing training loss curves across CRATE-α-Base with different patch sizes.

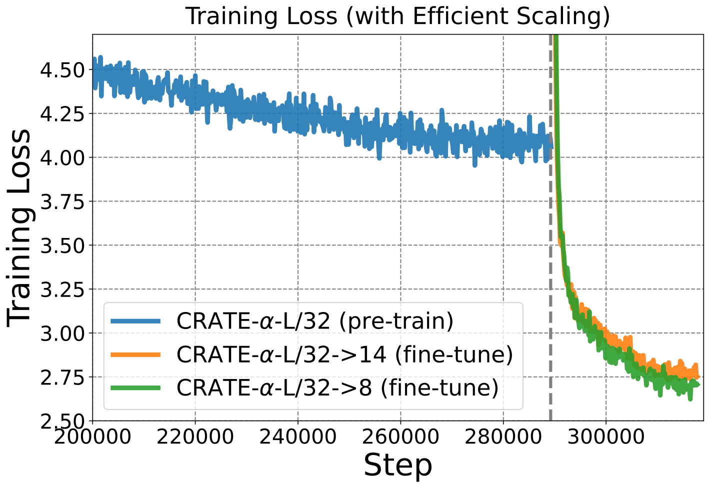

Figure 5: Comparing training loss curves when using the efficient scaling strategy. The blue curve corresponds to the CRATE-α-Large/32 model (in the pre-training stage). After pre-training the CRATE-α-Lage/32, we further fine-tune it with smaller patch sizes (longer token length), including patch size 14 (orange curve) and patch 8 (green curve).

References

[1] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. July 2021.

[2] Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, 2023.

[4] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

[6] Yaodong Yu, Tianzhe Chu, Shengbang Tong, Ziyang Wu, Druv Pai, Sam Buchanan, and Yi Ma. Emergence of segmentation with minimalistic white-box transformers. In *Conference on Parsimony and Learning*, pages 72–93. PMLR, 2024.