Improving Image Clustering with Artifacts Attenuation via Inference-Time Attention Engineering (Supplementary Material)

Kazumoto Nakamura, Yuji Nozawa, Yu-Chieh Lin,
Kengo Nakata, and Youyang Ng

Kioxia Corporation

A1 Extended Ablation Studies

A1.1 Generalization of the Proposed Method to Other Models with Different Pretraining Paradigms

To evaluate the models generalizability of our proposed method towards different pretraining paradigms, we conduct additional experiments using weakly-supervised pretrained models of CLIP [5] and supervised pretrained models of DeiT III [6]. We follow the same experimental protocols outlined in our paper. We use publicly available pretrained models in our experiments (*CLIP base*[1], *CLIP large*[2], *DeiT III base*[3], *DeiT III large*[4]). We set θ to 2.0 for all CLIP models, 2.7 for DeiT III base model and 3.25 for DeiT III large model. Note that since DeiT III was trained for ImageNet-1k, to preserve zero-shot generalization discussion, we select only CIFAR-10 and CIFAR-100 for evaluation in the case of DeiT III. The results of both clustering and k-NN classification are shown in Tab. A1. For CLIP models, the performance improves mostly with our method, similar to the results observed with DINOv2. For DeiT III, there is no obvious performance improvement, which aligns with the findings from the linear evaluation in Table 2 of Ref. [1]. Ref. [1] stated that pretraining paradigm seems to play a role in the characteristics of artifacts as CLIP and DeiT-III show artifacts at sizes smaller than DINOv2. We speculate that the supervised nature of DeiT III overfits the models to a particular dataset, decreasing the potential of performance extension during inference-time attention manipulation. We further analyzed the L_2 norms distribution in CLIP and DeiT-III models and found that ITAE successfully identified and attenuated artifacts, similar to DINOv2. We did not test *registers* and *registers + ours* settings as pretrained CLIP and DeiT III

[1] `https : / / openaipublic . azureedge . net / clip / models / 5806e77cd80f8b59890b7e101eabd078d9fb84e6937f9e85e4ecb61988df416f/ViT-B-16.pt`

[2] `https : / / openaipublic . azureedge . net / clip / models / b8cca3fd41ae0c99ba7e8951adf17d267cdb84cd88be6f7c2e0eca1737a03836/ViT-L-14.pt`

[3] `https://dl.fbaipublicfiles.com/deit/deit_3_base_224_21k.pth`

[4] `https://dl.fbaipublicfiles.com/deit/deit_3_large_224_21k.pth`

models with registers are not publicly available. However, we speculate that the complementary synergy between our method and *registers* [1] will enhance these models.

A1.2 Comparative Evaluation of Artifacts Attenuation Strategies

In our proposed method, as described in Sec. 3 of the paper, after the artifacts are identified, the corresponding attention values are attenuated to the minimum value of the patches in each head. Other strategies of attenuation can be considered. In this section, we examine three strategies: (a) Replacing artifacts with $-\infty$ (*-infinity*), (b) replacing artifacts with the average value of attention $\frac{1}{N}\sum_j \left((QK^T)_{0j}\right)$ (*average*), (c) replacing artifacts with the minimum value of attention, $\min_j \left((QK^T)_{0j}\right)$ (*minimum*), as implemented in our paper. Table A2 shows the accuracy of each of the three strategies. *average* outperforms other strategies in STL-10 for ViT-L/14 distilled and ViT-g/14. However, in the cases of CIFAR-100 and Tiny ImageNet for ViT-g/14, it falls below the accuracy of the *original*, and can be considered unstable. The results of the remaining two strategies are not so different, but *minimum* as implemented in our proposed method is slightly more accurate for more cases. There are many other possible variations of the attenuation strategies, but the fact that artifacts attenuation outperforms the original model in accuracy in almost all cases in this study indicates that attenuating attention values is effective to some extent, regardless of the specific value used for substitution.

A1.3 Comparative Evaluation with LSA

Locality Self-Attention (LSA) introduced temperature scaling and diagonal masking of the attention matrix to improve local induction bias [3]. We implemented the same diagonal masking of LSA, but only at inference time and only for the final layer of the model to conform with our framework. Table A3 shows the accuracy of LSA, our method and the original model. When comparing to the original model, LSA is effective for ViT-g/14 and ViT-L/14 distilled, but its accuracy does not clearly improve for ViT-B/14 distilled. LSA also achieves lower accuracy than our method in more cases. In this evaluation, we also investigate the combination of our method with LSA, denoted as *LSA + ours* in Tab. A3. The results show that combination with our method improved accuracy in models where LSA was effective. We speculate that while adopting the LSA alone increases the value of artifacts' attention, combining LSA with our method manages to increase the effective attention manipulated by LSA. Because of the complementary relationship between our method and LSA, we believe that the simultaneous adoption of both methods results in the greatest improvement in accuracy.

A1.4 Comparative Evaluation of Artifacts Identification Strategies

In Sec. 4 of the paper, we employ the L_2 norms of the query as QKV patches to identify the artifact. However, it is also possible to utilize the L_2 norms of

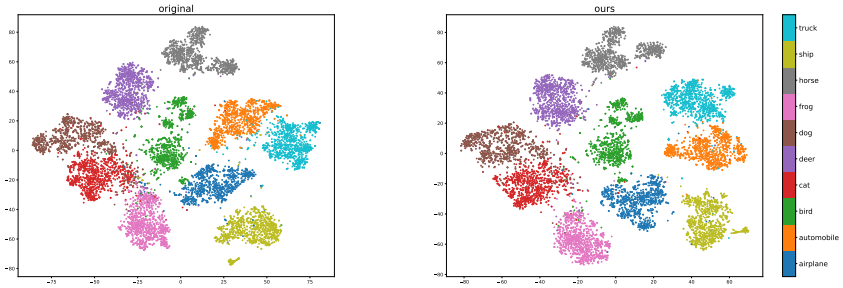

Fig. A1: t-SNE visualization: The left map is the output feature map of original model and the right map is the feature map of the model incorporating our proposed method. For our method in this visualization, we set $\theta = 3.5$ as obtained from Sec. 5.2 of the paper for better representing the potential of our method. Colors indicate true labels of image data points. Feature map of our proposed method shows fewer breakaway data points. Best viewed in color (model: ViT-B/14 distilled, dataset: CIFAR-10).

the key or value patches, as described in Sec. 3. Moreover, in prior work, artifacts were identified by the L_2 norms of the final output patch tokens of the model [1]. In this section, we discuss these other strategies of artifacts identification. The results of clustering are shown in Tab. A4 after identifying the artifacts by a threshold value θ for each model from the histogram of L_2 norms of each strategy. From the result, we observe that the method using *output* has a high accuracy in ViT-g/14. However, the clustering accuracy is not stable in ViT-B/14 distilled. As described in Sec. 3, the histogram obtained from the L_2 norms of *output* exhibits unclear bimodality. Hence, there is difficulty in determining an appropriate threshold for identifying artifacts. Also, there is a disadvantage in terms of computational cost when computing artifacts from the *output* as a back pass loop from the output of the model is needed. For comparison between utilizing *query*, *key*, and *value*, we observe that the accuracy of *key*, like the *output*, is not stable for ViT-B/14 distilled, although to a lesser extent. *value* and *query* have almost the same level of accuracy. Therefore, for the sake of clarity, we mainly utilized *query* in our experiments.

A1.5 Output Feature Representation Visualization

Figure A1 shows the t-SNE [4] visualization of the output features for the original model and model incorporating our proposed method by using the dataset of CIFAR-10. For the image clustering of CIFAR-10 with ViT-B/14 model, experiment using the original model has a clustering accuracy of 83.63, which improves to 84.49 for $\theta = 3.0$ and 84.86 for $\theta = 3.5$ by applying our proposed method. It is observed that feature map of our proposed method shows fewer breakaway data points. To quantify this, in the discussion here, we define breakaway points as data points with a silhouette score smaller than 0, calculated using the true

labels. For the original model, there are 635 breakaway points out of 10,000 data points. The number of breakaway points decreases to 569 for $\theta = 3.0$ and 535 for $\theta = 3.5$. These improvements in output features' quality enhance the subsequent image clustering accuracy.

A2 Limitation

A2.1 Limitations on Data

Because this method utilizes a pretrained model and does not involve any re-training, there is a possibility that clustering may not work well on some highly specific datasets due to the bias of the pretrained model. However, for the datasets that we have evaluated, our method proves effective.

A2.2 Limitations on Methodology

Methods without re-training such as our proposed method maybe difficult to extract performance beyond the potential of the original model. However, algorithms of tuning-free merging of weights from other external models [2, 7] have been proposed. These methods may provide a complimentary solution to our method for better performance.

A3 Licence info

Table A5 shows the license info of images used in Fig. 4 of the paper. The images are overlaid with attention map in the figure.

Table A1: Clustering & k-NN classification results across various pretraining paradigms and model sizes (*DINOv2 base*: ViT-B/14 distilled, *DINOv2 large*: ViT-L/14 distilled, *CLIP base*: ViT-B/16, *CLIP large*: ViT-L/14, *DeiT III base*: ViT-B/16, *DeiT III large*: ViT-L/16). Clustering results are reported in ACC while k-NN classification results are reported in standard k-NN classification accuracy.

Dataset	Experiment	Model	Model Size	original	ours
CIFAR-10	Clustering	DINOv2	base	83.63 ± 1.13	$\mathbf{84.49 \pm 1.19}$
			large	82.16 ± 1.48	$\mathbf{82.49 \pm 1.55}$
		CLIP	base	72.34 ± 0.80	$\mathbf{77.47 \pm 1.34}$
			large	$\mathbf{79.45 \pm 1.47}$	79.25 ± 1.49
		DeiT III	base	82.54 ± 3.48	$\mathbf{82.89 \pm 1.26}$
			large	84.18 ± 3.11	$\mathbf{84.43 \pm 2.76}$
CIFAR-100	Clustering	DINOv2	base	64.26 ± 0.30	$\mathbf{65.02 \pm 0.14}$
			large	68.69 ± 0.34	$\mathbf{69.04 \pm 0.22}$
		CLIP	base	42.92 ± 0.22	$\mathbf{49.63 \pm 0.25}$
			large	47.88 ± 0.31	$\mathbf{56.94 \pm 0.21}$
		DeiT III	base	$\mathbf{60.64 \pm 0.57}$	60.41 ± 0.22
			large	$\mathbf{67.19 \pm 0.57}$	67.16 ± 0.59
STL-10	Clustering	DINOv2	base	75.65 ± 1.04	$\mathbf{82.76 \pm 1.27}$
			large	65.78 ± 1.22	$\mathbf{70.51 \pm 1.42}$
		CLIP	base	85.61 ± 1.74	$\mathbf{86.57 \pm 1.38}$
			large	83.67 ± 1.34	$\mathbf{84.65 \pm 1.39}$
Tiny ImageNet	Clustering	DINOv2	base	67.81 ± 0.24	$\mathbf{68.23 \pm 0.25}$
			large	71.98 ± 0.15	$\mathbf{73.19 \pm 0.21}$
		CLIP	base	35.43 ± 0.16	$\mathbf{39.53 \pm 0.15}$
			large	52.54 ± 0.16	$\mathbf{55.45 \pm 0.14}$
CIFAR-100	k-NN	DINOv2	base	87.31	**87.58**
			large	91.12	**91.39**
		CLIP	base	71.72	**73.56**
			large	78.81	**80.90**
		DeiT III	base	**82.22**	81.97
			large	86.13	**86.22**
ImageNet-1k	k-NN	DINOv2	base	82.04	**82.07**
			large	83.50	**83.62**
		CLIP	base	73.12	**74.26**
			large	79.25	**80.35**

Table A2: Image clustering result across various attenuation strategies and model sizes (*small*: ViT-S/14 distilled, *base*: ViT-B/14 distilled, *large*: ViT-L/14 distilled, *giant*: ViT-g/14) reported in ACC.

Model Size	Method	CIFAR-10	CIFAR-100	STL-10	Tiny ImageNet
base	original	83.63 ± 1.13	64.26 ± 0.30	75.65 ± 1.04	67.81 ± 0.24
	-infinity	84.47 ± 1.33	64.92 ± 0.23	82.68 ± 1.25	$\mathbf{68.27 \pm 0.23}$
	average	84.27 ± 1.30	64.88 ± 0.33	$\mathbf{82.87 \pm 1.28}$	68.34 ± 0.22
	minimum	$\mathbf{84.49 \pm 1.19}$	$\mathbf{65.02 \pm 0.14}$	82.76 ± 1.27	68.23 ± 0.25
large	original	82.16 ± 1.48	68.69 ± 0.34	65.78 ± 1.22	71.98 ± 0.15
	-infinity	$\mathbf{82.49 \pm 1.56}$	69.03 ± 0.21	70.53 ± 1.22	73.18 ± 0.19
	average	82.41 ± 1.27	68.75 ± 0.37	$\mathbf{72.13 \pm 1.39}$	72.67 ± 0.18
	minimum	$\mathbf{82.49 \pm 1.55}$	$\mathbf{69.04 \pm 0.22}$	70.51 ± 1.42	$\mathbf{73.19 \pm 0.21}$
giant	original	78.09 ± 1.25	68.99 ± 0.39	55.91 ± 1.14	73.25 ± 0.16
	-infinity	78.64 ± 1.87	$\mathbf{69.56 \pm 0.33}$	56.00 ± 0.84	73.52 ± 0.16
	average	$\mathbf{79.19 \pm 1.87}$	68.58 ± 0.25	$\mathbf{56.25 \pm 0.88}$	72.97 ± 0.19
	minimum	78.59 ± 1.91	69.50 ± 0.28	56.01 ± 0.93	$\mathbf{73.54 \pm 0.17}$

Table A3: Image clustering result when adopting LSA in various model sizes (*small*: ViT-S/14 distilled, *base*: ViT-B/14 distilled, *large*: ViT-L/14 distilled, *giant*: ViT-g/14) reported in ACC.

Model Size	Method	CIFAR-10	CIFAR-100	STL-10	Tiny ImageNet
base	original	83.63 ± 1.13	64.26 ± 0.30	75.65 ± 1.04	67.81 ± 0.24
	LSA	83.25 ± 1.41	64.35 ± 0.28	75.77 ± 1.11	67.83 ± 0.15
	ours	$\mathbf{84.49 \pm 1.19}$	$\mathbf{65.02 \pm 0.14}$	$\mathbf{82.76 \pm 1.27}$	68.23 ± 0.25
	LSA + ours	83.97 ± 1.55	64.86 ± 0.37	82.45 ± 1.45	$\mathbf{68.33 \pm 0.14}$
large	original	82.16 ± 1.48	68.69 ± 0.34	65.78 ± 1.22	71.98 ± 0.15
	LSA	82.84 ± 1.68	69.33 ± 0.34	69.39 ± 1.22	72.22 ± 0.20
	ours	82.49 ± 1.55	69.04 ± 0.22	70.51 ± 1.42	73.19 ± 0.21
	LSA + ours	$\mathbf{83.14 \pm 1.47}$	$\mathbf{69.75 \pm 0.32}$	$\mathbf{76.58 \pm 1.46}$	$\mathbf{73.60 \pm 0.17}$
giant	original	78.09 ± 1.25	68.99 ± 0.39	55.91 ± 1.14	73.25 ± 0.16
	LSA	78.82 ± 2.05	69.30 ± 0.33	56.91 ± 0.71	73.36 ± 0.16
	ours	78.59 ± 1.91	69.50 ± 0.28	56.01 ± 0.93	73.54 ± 0.17
	LSA + ours	$\mathbf{79.70 \pm 1.56}$	$\mathbf{69.84 \pm 0.35}$	$\mathbf{58.95 \pm 0.92}$	$\mathbf{73.90 \pm 0.18}$

Table A4: Image clustering result across various artifacts identification strategies and model sizes (*small*: ViT-S/14 distilled, *base*: ViT-B/14 distilled, *large*: ViT-L/14 distilled, *giant*: ViT-g/14) reported in ACC.

Model Size	Method	CIFAR-10	CIFAR-100	STL-10	Tiny ImageNet
base	original	83.63 ± 1.13	64.26 ± 0.30	75.65 ± 1.04	67.81 ± 0.24
	query	$\mathbf{84.49 \pm 1.19}$	$\mathbf{65.02 \pm 0.14}$	82.76 ± 1.27	68.23 ± 0.25
	key	83.13 ± 1.06	64.78 ± 0.31	79.03 ± 1.22	68.14 ± 0.11
	value	84.29 ± 0.98	64.98 ± 0.22	$\mathbf{82.77 \pm 1.44}$	$\mathbf{68.30 \pm 0.18}$
	output	$\mathbf{84.49 \pm 1.67}$	64.53 ± 0.29	75.82 ± 1.06	67.96 ± 0.18
large	original	82.16 ± 1.48	68.69 ± 0.34	65.78 ± 1.22	71.98 ± 0.15
	query	82.49 ± 1.55	69.04 ± 0.22	70.51 ± 1.42	73.19 ± 0.21
	key	82.92 ± 1.73	69.09 ± 0.32	$\mathbf{70.66 \pm 1.31}$	73.17 ± 0.17
	value	82.49 ± 1.56	69.05 ± 0.18	70.56 ± 1.41	$\mathbf{73.21 \pm 0.16}$
	output	$\mathbf{82.97 \pm 1.82}$	$\mathbf{69.15 \pm 0.30}$	69.58 ± 1.14	73.03 ± 0.18
giant	original	78.09 ± 1.25	68.99 ± 0.39	55.91 ± 1.14	73.25 ± 0.16
	query	78.59 ± 1.91	69.50 ± 0.28	56.01 ± 0.93	$\mathbf{73.54 \pm 0.17}$
	key	78.63 ± 1.88	69.39 ± 0.27	56.01 ± 0.94	73.52 ± 0.20
	value	78.59 ± 1.91	69.51 ± 0.30	56.01 ± 0.93	73.53 ± 0.19
	output	$\mathbf{78.65 \pm 1.86}$	$\mathbf{69.54 \pm 0.29}$	$\mathbf{56.02 \pm 0.92}$	73.52 ± 0.16

Table A5: License info of images in Fig. 4 of the paper (※urls of number 2 and 9 are currently invalid).

Number	Image id	URL	License
1	526751	http://farm4.staticflickr.com/3288/2933360267_ae24740821_z.jpg	Attribution-NonCommercial-NoDerivs License
2	574315	http://farm3.staticflickr.com/2010/2247055627_5269f84985_z.jpg	Attribution-NonCommercial-NoDerivs License
3	5037	http://farm8.staticflickr.com/7379/9599671465_8a2f486da1_z.jpg	Attribution-NoDerivs License
4	246883	http://farm4.staticflickr.com/3067/2869541146_a627d12677_z.jpg	Attribution-NonCommercial-ShareAlike License
5	253433	http://farm1.staticflickr.com/162/361912851_59c9993d91_z.jpg	Attribution-NonCommercial-NoDerivs License
6	231237	http://farm4.staticflickr.com/3607/3342869781_cd4a4b1154_z.jpg	Attribution-NonCommercial-ShareAlike License
7	289659	http://farm8.staticflickr.com/7031/6786602747_b7b811b0d5_z.jpg	Attribution-NonCommercial-NoDerivs License
8	163290	http://farm7.staticflickr.com/6166/6190069448_f9da6727e6_z.jpg	Attribution-NonCommercial License
9	66817	http://farm8.staticflickr.com/7279/7864913910_9e85e0a82a_z.jpg	Attribution License
10	424545	http://farm4.staticflickr.com/3193/3054220374_d2a3456295_z.jpg	Attribution-NonCommercial-NoDerivs License
11	378673	http://farm8.staticflickr.com/7124/7810951178_b622f1466c_z.jpg	Attribution-NonCommercial-NoDerivs License
12	405279	http://farm1.staticflickr.com/236/443807489_3d7fba2557_z.jpg	Attribution License
13	332570	http://farm2.staticflickr.com/1051/1392968224_0f863f4054_z.jpg	Attribution-NonCommercial License
14	131386	http://farm6.staticflickr.com/5074/5860045248_f99b35c5c8_z.jpg	Attribution-ShareAlike License
15	338560	http://farm5.staticflickr.com/4044/4583091116_28eaab2a2b_z.jpg	Attribution License

Attribution-NonCommercial-ShareAlike License : http://creativecommons.org/licenses/by-nc-sa/2.0/
Attribution-NonCommercial License : http://creativecommons.org/licenses/by-nc/2.0/
Attribution-NonCommercial-NoDerivs License : http://creativecommons.org/licenses/by-nc-nd/2.0/
Attribution License : http://creativecommons.org/licenses/by/2.0/
Attribution-ShareAlike License : http://creativecommons.org/licenses/by-sa/2.0/
Attribution-NoDerivs License : http://creativecommons.org/licenses/by-nd/2.0/

References

1. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024)
2. Huang, C., Ye, P., Chen, T., He, T., Yue, X., Ouyang, W.: Emr-merging: Tuning-free high-performance model merging. arXiv preprint arXiv:2405.17461 (2024)
3. Lee, S.H., Lee, S., Song, B.C.: Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492 (2021)
4. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)
5. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International conference on machine learning (2021)
6. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
7. Xu, Z., Yuan, K., Wang, H., Wang, Y., Song, M., Song, J.: Training-free pretrained model merging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5915–5925 (2024)