Supplementary Material – Weight Copy and Low-Rank Adaptation for Few-Shot Distillation of Vision Transformers

Diana-Nicoleta Grigore[1,◇], Mariana-Iuliana Georgescu[1,◇], Jon Alvarez Justo[2], Tor Johansen[2],
Andreea Iuliana Ionescu[3], Radu Tudor Ionescu[1,*]
[1]University of Bucharest, Romania, [2]Norwegian University of Science and Technology, Norway,
[3]University of Medicine and Pharmacy “Carol Davila”, Romania

1. Additional Results

Few-shot distillation from 1% to 10%. In Tables 1 and 2, we report results on five downstream tasks, when the students use only 2% and 5% of the ImageNet data during distillation, respectively. As noted in the main manuscript, we perform linear probing to demonstrate that our method transfers strong features. We notice that our method, WeCoLoRA, attains higher performance than the WeCo+KD method, especially when the distillation data is scarce. We observe a substantial improvement (of at least 2%) on the ImageNet downstream task, regardless of the reduction ratio or the distillation training size, when the teacher is the supervised ViT-B [4] model. We observe the same trend on the other data sets employed in the evaluation. We further note that the features learned by our distillation method also transfer to out-of-distribution data sets, such as ChestX-ray14 [13]. We consider ChestX-ray14 as out-of-distribution because it contains medical images, while the pre-training data set, ImageNet, contains natural images.

We conclude that the proposed distillation method, WeCoLoRA, is robust and obtains improved performance on multiple downstream tasks, especially when the pre-training data set is small. We also emphasize that our method does not require labeled data, and is able to compress both supervised and self-supervised models.

To better assess the performance trends on various downstream tasks when the number of samples increases from 1% to 10%, we further illustrate the performance levels obtained by WeCoLoRA vs. WeCo+KD on ImageNet-1K [3], iNaturalist [12], NWPU-RESISC45 [2], CIFAR-100 [8] and ChestX-ray14 [13] in Figures 1, 2, 3, 4, and 5, respectively. We observe that WeCoLoRA obtains significantly higher performance than WeCo+KD when there is less data involved in the knowledge distillation process (1% and 2% of the original training set [3]). Moreover, in most of the cases, WeCoLoRA also outperforms WeCo+KD when 10% of the original training set in used during knowledge distillation. We also conducted experiments on the full scale ImageNet and observed marginal differences between WeCo+KD and WeCoLoRA. We therefore conclude that WeCoLoRA is particularly useful in the few-shot KD setting.

Fewer samples, lighter student. We would like to mention that 1% of ImageNet corresponds to 12 samples per class. However, models are often evaluated on even fewer shots. To this end, we perform extra experiments for 0.25% of ImageNet, *i.e.* 3 samples per class. In this setting, we also employ a more aggressive reduction ratio ($r = 4$), which leads to a very light student model.

When the number of shots is very small, the model can collapse due to a low diversity of training samples. Although this is not the particular focus of our method, there is no obvious reason for WeCoLoRA not to be compatible with orthogonal methods that deal with the issue of collapse. To mitigate the issue of collapse, we combine WeCoLoRA with k-means++ init to select the training samples.

The results of WeCo+KD and WeCoLoRA for 0.25% of ImageNet and $r = 4$ are shown in Table 3. WeCoLoRA obtains superior results on all three data sets (ImageNet, iNaturalist and CIFAR-100). When the training samples are chosen via k-means++, we obtain slightly improved results on two datasets (see the last row in Table 3).

Layer pruning vs. WeCoLoRA. A promising approach to create lighter models without much effort is layer pruning. For transformer models, it was recently found that TopPruning, a method that drops the top-layers, obtains surprisingly good results [11]. To this end, we compare WeCoLoRA with TopPruning, using the same reduction factor of $r = 2$ for both methods. The results, which are reported in Table 4, clearly indicate that WeCoLoRA outperforms TopPruning.

Comparing teacher and student features. One question that arises when applying WeCoLoRA is if the student is indeed learning features similar to the skipped teacher layers. To address this point, we compute the mean cosine similarities between the skipped layers and the corresponding student layers (from 1 to 4), before and after applying our enhanced LoRA. As shown in Table 5, the similarities increase after distillation, indicating that the student learns

[*]Corresp. author: raducu.ionescu@gmail.com. [◇]Equal contribution.

Teacher	Compression factor r	Distillation method	ImageNet	ChestX-ray14	iNaturalist	RESISC45	CIFAR-100
ViT-B [4] (supervised)	2	WeCo+KD	46.9	68.3	35.1	61.6	42.0
		WeCoLoRA	**63.5**	**70.0**	**46.5**	**68.5**	**62.9**
	3	WeCo+KD	37.0	67.8	28.2	59.8	37.9
		WeCoLoRA	**39.6**	**68.1**	**29.5**	**61.5**	**38.6**
ViT-B [5] (SSL)	2	WeCo+KD	46.7	68.6	28.3	**62.9**	41.5
		WeCoLoRA	**48.2**	**68.9**	**28.5**	58.8	**44.8**
	3	WeCo+KD	33.6	66.5	20.0	53.7	35.7
		WeCoLoRA	**35.3**	**67.0**	**22.2**	**56.0**	**36.8**

Table 1. Results of WeCoLoRA and WeCo+KD in terms of accuracy (in percentages) on ImageNet-1K [3], iNaturalist [12], NWPU-RESISC45 [2] and CIFAR-100 [8], and in terms of mean AUC (in percentages) on ChestX-ray14 [13]. Results are reported for the supervised ViT-B [4] teacher and the self-supervised (SSL) ViT-B [5] teacher. During the distillation procedure, only **2**% of the ImageNet-1K training set [3] is used.

Teacher	Compression factor r	Distillation method	ImageNet	ChestX-ray14	iNaturalist	RESISC45	CIFAR-100
ViT-B [4] (supervised)	2	WeCo+KD	65.3	69.4	45.4	**73.5**	60.1
		WeCoLoRA	**67.3**	**70.0**	**49.0**	69.9	**66.6**
	3	WeCo+KD	52.2	68.4	37.0	**66.7**	46.9
		WeCoLoRA	**55.3**	**69.9**	**40.2**	66.1	**51.7**
ViT-B [5] (SSL)	2	WeCo+KD	51.0	69.3	29.9	**61.7**	47.6
		WeCoLoRA	**54.0**	**69.5**	**32.9**	61.5	**47.7**
	3	WeCo+KD	36.1	66.6	18.7	51.3	30.5
		WeCoLoRA	**37.4**	**66.5**	**22.0**	**58.6**	**38.6**

Table 2. Results of WeCoLoRA and WeCo+KD in terms of accuracy (in percentages) on ImageNet-1K [3], iNaturalist [12], NWPU-RESISC45 [2] and CIFAR-100 [8], and in terms of mean AUC (in percentages) on ChestX-ray14 [13]. Results are reported for the supervised ViT-B [4] teacher and the self-supervised (SSL) ViT-B [5] teacher. During the distillation procedure, only **5**% of the ImageNet-1K training set [3] is used.

Method	ImageNet	iNaturalist	CIFAR-100
WeCo+KD (ablated)	14.0	9.9	17.1
WeCoLoRA (ours)	21.6	14.6	**25.7**
WeCoLoRA (ours) + k-means++ init	**21.7**	**14.9**	25.5

Table 3. Accuracy rates on ImageNet-1K, iNaturalist19 and CIFAR-100, when only 0.25% ($\alpha = 0.25$) of the training set is used during distillation. K-means++ init is used to select the samples for few-shot KD (to avoid collapse). Compression factor: $r = 4$. Teacher: self-supervised ViT-B.

Method	ImageNet	iNaturalist	CIFAR-100
TopPruning [11]	52.6	41.4	56.1
WeCoLoRA (ours)	**69.2**	**49.5**	**68.3**

Table 4. Accuracy rates on ImageNet-1K, iNaturalist19 and CIFAR-100, comparing our WeCoLoRA with top-layer dropping (TopPruning). Compression factor: $r = 2$. Teacher: supervised ViT-B.

Method	Layer 1	Layer 2	Layer 3	Layer 4
WeCo (ablated)	0.693	0.458	0.387	0.268
WeCoLoRA (ours)	0.753	0.686	0.674	0.796

Table 5. Mean cosine similarities (averaged over tokens) between the student's layers and the skipped layers from the teacher. The values are calculated on features derived from 10% of ImageNet. Compression factor: $r = 3$. Teacher: self-supervised ViT-B.

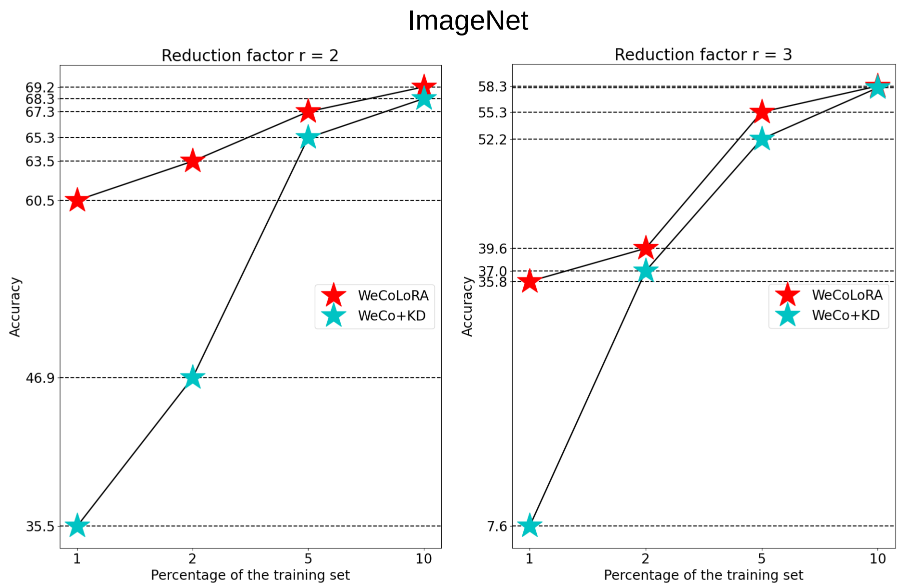

Figure 1. Accuracy rates obtained by WeCoLoRA and WeCo+KD on the ImageNet-1K [3] downstream task. Results are reported for the supervised ViT-B [4] teacher. The horizontal axis corresponds to the percentage of the original training set [3] used during knowledge distillation. Best viewed in color.

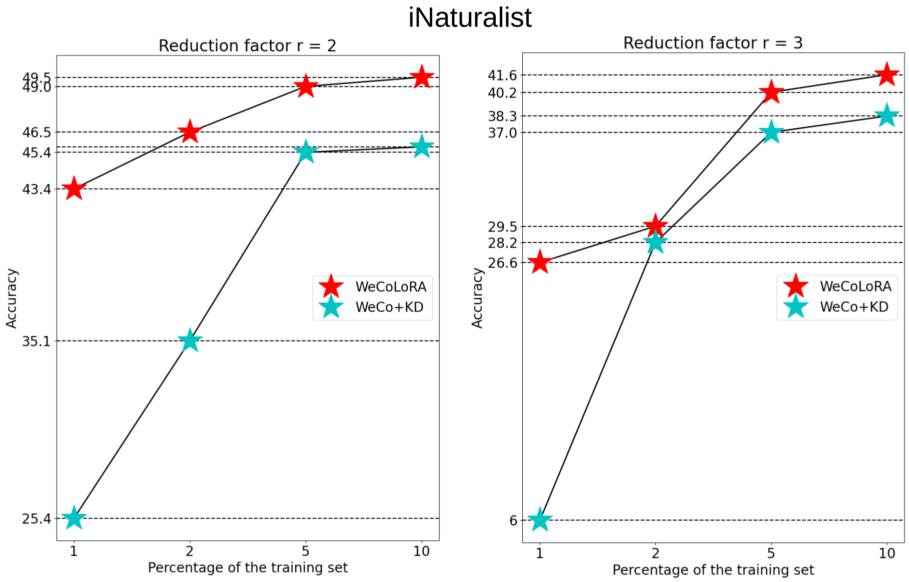

Figure 2. Accuracy rates obtained by WeCoLoRA and WeCo+KD on the iNaturalist [12] downstream task. Results are reported for the supervised ViT-B [4] teacher. The horizontal axis corresponds to the percentage of the original training set [3] used during knowledge distillation. Best viewed in color.

features similar to the skipped teacher layers. This confirms that enhanced LoRA has the intended effect.

Segmentation results using convolutional networks. To showcase the versatility of our approach, we test WeCoLoRA on a medical image segmentation task, by integrating it into the U-Net architecture [10]. The segmentation model employs ResNet-18 [6] as backbone. Since the segmentation model is based on convolutional layers, we replace LoRA [7] with ConvLoRA [1]. The experiments are performed on the Multiple Sclerosis Lesion Segmentation benchmark (MSLesSeg 2024) [9]. We report results in terms of the Dice coefficient in Table 6, where we compare our WeCoLoRA with the strongest baseline, namely WeCo+KD. The results demonstrate that WeCoLoRA ob-

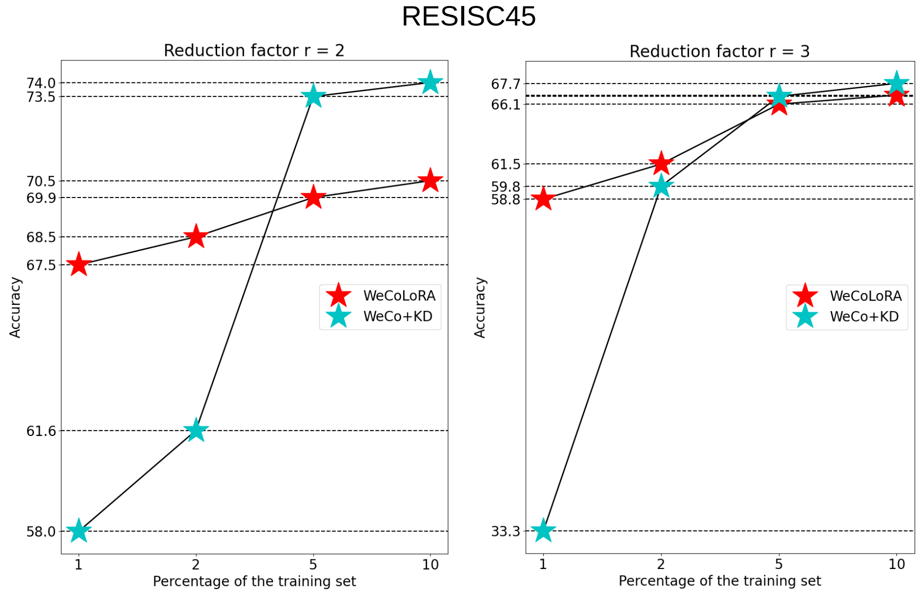

Figure 3. Accuracy rates obtained by WeCoLoRA and WeCo+KD on the NWPU-RESISC45 [2] downstream task. Results are reported for the supervised ViT-B [4] teacher. The horizontal axis corresponds to the percentage of the original training set [3] used during knowledge distillation. Best viewed in color.

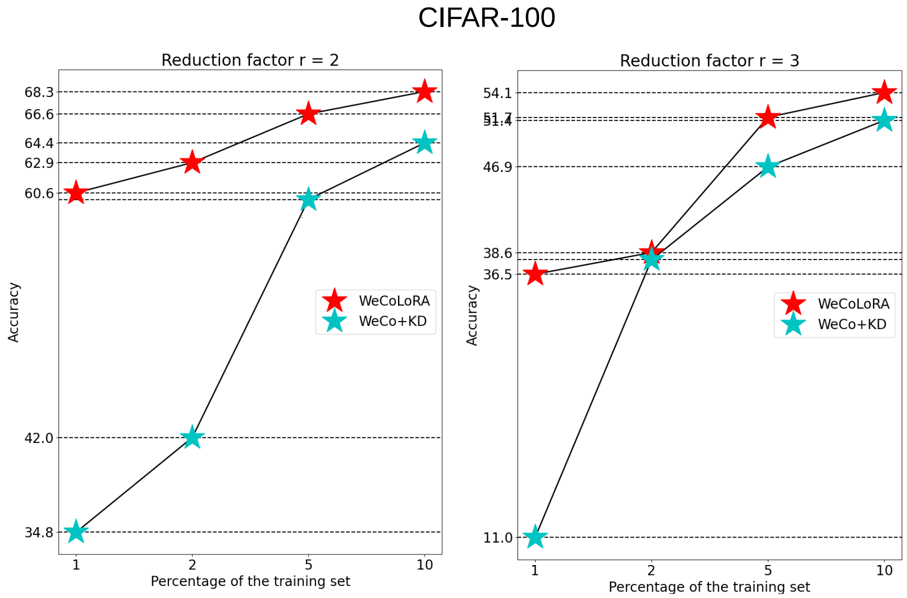

Figure 4. Accuracy rates obtained by WeCoLoRA and WeCo+KD on the CIFAR-100 [8] downstream task. Results are reported for the supervised ViT-B [4] teacher. The horizontal axis corresponds to the percentage of the original training set [3] used during knowledge distillation. Best viewed in color.

tains higher performance than WeCo+KD, when a convolutional backbone is employed on a segmentation task. This demonstrates the compatibility of WeCoLoRA with both transformer and convolutional architectures, as well as its applicability to diverse tasks, namely classification and segmentation.

Deeper teacher, higher compression factors. To demon-

Method	Dice Coefficient
WeCo+KD (ablated)	0.7665
WeCoLoRA (ours)	**0.7708**

Table 6. Segmentation results in terms of Dice coefficient obtained with WeCo+KD and WeCoLoRA on the MSLesSeg benchmark. Compression factor: $r = 2$.

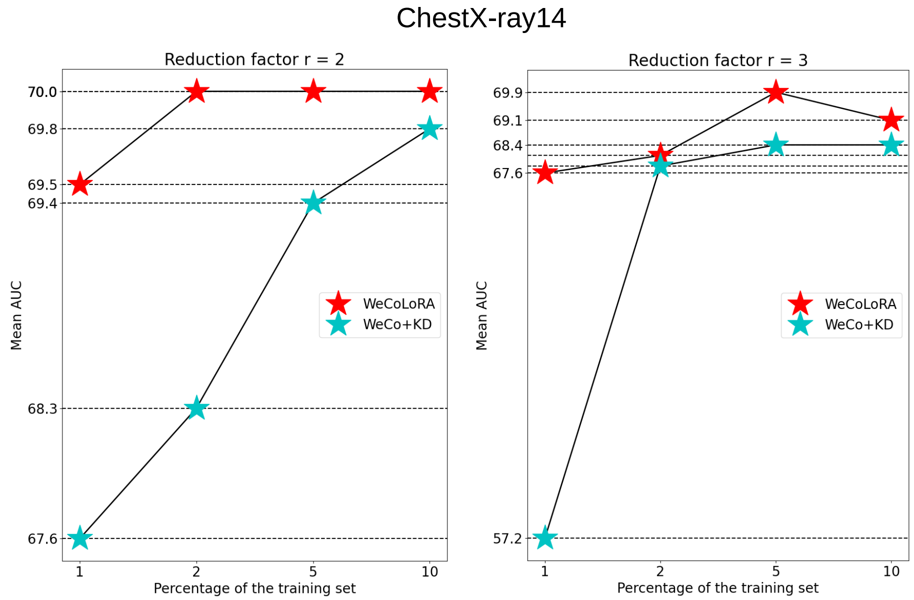

Figure 5. Mean AUC scores (in percentages) obtained by WeCoLoRA and WeCo+KD on the ChestX-ray14 [13] downstream task. Results are reported for the supervised ViT-B [4] teacher. The horizontal axis corresponds to the percentage of the original training set [3] used during knowledge distillation. Best viewed in color.

Method	Compression factor r	CIFAR-100	RESISC-45	ChestX-ray14
WeCo+KD (ablated)	6	22.3	52.8	62.8
WeCoLoRA (ours)	6	24.1	53.6	**63.7**
WeCoLoRA+supervised fine-tuning	6	18.4	**54.5**	63.3
WeCoLoRA+classification head transfer	6	**27.1**	54.2	63.6
WeCo+KD (ablated)	8	20.6	49.2	62.5
WeCoLoRA (ours)	8	25.0	**53.8**	63.1
WeCoLoRA+supervised fine-tuning	8	21.8	53.3	62.4
WeCoLoRA+classification head transfer	8	**25.9**	53.6	**63.3**

Table 7. Results of various training paradigms in terms of accuracy (in percentages) on CIFAR-100 [8], NWPU-RESISC45 [2] and in terms of mean AUC (in percentages) on ChestX-ray14 [13]. Results are reported for the unsupervised ViT-L teacher or backbone [4]. During the fine-tuning or distillation procedure, only 1% of ImageNet-1K [3] training set is used.

strate the applicability of WeCoLoRA to deeper teachers, and its robustness to higher compression factors, we perform additional experiments with the ViT-L teacher based on supervised pre-training, considering compression factors of $r = 6$ and $r = 8$. In Table 7, we report the results of WeCo+KD and WeCoLoRA on CIFAR-100, RESISC-45 and ChestX-ray14. WeCoLoRA outperforms WeCo+KD for all compression factors, thus showcasing consistent performance gains across various compression factors and teacher models.

WeCoLoRA based on fine-tuning instead of distillation. The feature distillation performed by WeCoLoRA is unsupervised, *i.e.* our framework does not require classification labels during distillation. An alternative approach is to employ supervised fine-tuning instead of unsupervised feature distillation. As shown in Table 7, the fine-tuning combined with WeCoLoRA produces worse results on CIFAR-100, while leading to similar results on RESISC45 and ChestX-ray14. We thus conclude that the supervised fine-tuning is not always beneficial.

WeCoLoRA with classification head transfer. One way to potentially boost the performance of WeCoLoRA is to transfer the classification head from the corresponding teacher model, instead of initializing the classification head of the student model from scratch. This idea is explored in Table 7 (last row), where it exhibits performance boosts on CIFAR-100. The results on RESISC45 and ChestX-ray14 do not clearly show the benefit of transferring the classification head.

2. Limitations

The main limitation of our method is its applicability to architectures that use multiple consecutive blocks with the same configuration, *e.g.* vision transformers and ResNets [6]. This restriction is imposed by our weight copying mechanism. Our ablation results indicate that the weight copying step is very useful in the few-shot distillation scenario, as it significantly boosts performance (see Table 1 from the main article). Simply removing the weight copying step is not a viable option, since the performance would drastically degrade. To make our framework applicable to any architecture, the weight copying mechanism could be enhanced with adaptor blocks, which would be able to reshape the copied weights to the appropriate size. However, the adaptor blocks need to be tailored for each specific pair of teacher and student models. This will increase the complexity of the hyperparameter tuning stage, which, in the current form, is quite straightforward, *i.e.* aside from typical hyperparameters, such as the learning rate and the mini-batch size, WeCoLoRA only adds the compression ratio r and the rank of the low-rank matrices k as extra hyperparameters.

References

[1] Sidra Aleem, Julia Dietlmeier, Eric Arazo, and Suzanne Little. ConvLoRA and AdaBN Based Domain Adaptation via Self-Training. *Proceedings of ISBI*, pages 1–5, 2024. 3

[2] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 1, 2, 4, 5

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255, 2009. 1, 2, 3, 4, 5

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*, 2021. 1, 2, 3, 4, 5

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of CVPR*, pages 16000–16009, 2022. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of CVPR*, pages 770–778, 2016. 3, 6

[7] Edward J. Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of ICLR*, 2022. 3

[8] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 1, 2, 4, 5

[9] Alessia Rondinella, Elena Crispino, Francesco Guarnera, Oliver Giudice, Alessandro Ortis, Giulia Russo, Clara Di Lorenzo, Davide Maimone, Francesco Pappalardo, and Sebastiano Battiato. Boosting multiple sclerosis lesion segmentation through attention mechanism. *Computers in Biology and Medicine*, 161:107021, 2023. 3

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of MICCAI*, pages 234–241, 2015. 3

[11] Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. On the effect of dropping layers of pre-trained transformer models. *Computer Speech & Language*, 77:101429, 2023. 1, 2

[12] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist Species Classification and Detection Dataset. In *Proceedings of CVPR*, pages 8769–8778, 2018. 1, 2, 3

[13] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *Proceedings of CVPR*, pages 3462–3471, 2017. 1, 2, 5