Supplementary Materials for MagMax: Leveraging Model Merging for Seamless Continual Learning

Daniel Marczak*[1,2], Bartłomiej Twardowski[1,5,6], Tomasz Trzciński[1,2,4], and Sebastian Cygert[1,3]

[1] IDEAS NCBR
[2] Warsaw University of Technology
[3] Gdańsk University of Technology
[4] Tooploox
[5] Autonomous University of Barcelona
[6] Computer Vision Center

A More results

A.1 CIL results with different backbones

We replicate our main results from Table **??** with two different, stronger backbones: ViT-L-14 pre-trained on WebImageText [1] (different architecture, the same pre-training dataset) in Table 1 and ViT-B-16 pre-trained on LAION-400M (the same architecture, different pre-training dataset) in Table 2. MagMax still outperforms both CL and merging-based baselines. We observe smaller improvement of MagMax over the second best method than in Table **??** as the room for improvement (defined as a difference in performance between joint and zero-shot model) is smaller for these stronger backbones.

Table 1: Results with ViT-L-14 pre-trained on WebImageText [1].

Method	CIFAR100 /5	CIFAR100 /10	CIFAR100 /20	CIFAR100 /50	ImageNet-R /5	ImageNet-R /10	ImageNet-R /20	ImageNet-R /50	CUB200 /5	CUB200 /10	CUB200 /20	Cars /5	Cars /10	Cars /20	Avg
Zero-shot	75.82				87.93				62.86			77.94			76.96
Joint	93.49				93.70				87.00			92.24			91.89
LwF	86.89	83.21	82.12	82.23	90.92	**91.80**	**91.28**	90.60	**73.96**	**71.09**	66.22	80.11	79.14	75.92	81.82
EWC	88.72	85.41	85.45	83.34	91.13	91.37	91.08	**90.98**	68.66	66.45	64.65	78.97	74.05	69.74	80.71
Rand Mix	87.90	85.60	83.67	81.36	91.48	90.45	89.32	88.20	67.74	65.31	64.34	80.08	77.96	78.25	80.83
Max Abs	88.20	85.45	83.80	82.17	91.67	91.15	90.05	88.95	67.45	65.33	63.39	79.48	75.91	75.46	80.60
Avg	87.87	85.54	83.69	81.36	91.47	90.42	89.32	88.20	67.60	65.29	64.36	80.14	77.91	78.26	80.82
TIES	88.35	85.40	84.05	82.46	91.58	90.75	89.90	88.77	67.86	65.48	64.45	80.70	76.73	76.41	80.92
MagMax	**88.90**	**87.24**	**86.18**	**84.97**	**91.95**	91.63	91.22	90.72	70.30	67.57	**66.98**	**84.62**	**80.85**	**78.80**	**82.99**

* Corresponding author, email: daniel.marczak.dokt@pw.edu.pl

Table 2: Results with ViT-B-16 pre-trained on LAION-400M [2].

Method	CIFAR100 /5	CIFAR100 /10	CIFAR100 /20	CIFAR100 /50	ImageNet-R /5	ImageNet-R /10	ImageNet-R /20	ImageNet-R /50	CUB200 /5	CUB200 /10	CUB200 /20	Cars /5	Cars /10	Cars /20	Avg
Zero-shot	71.29				77.08				64.64			83.65			74.17
Joint	90.79				86.17				81.31			90.65			87.41
LwF	84.14	78.09	75.77	75.13	82.90	82.80	82.17	80.82	**71.92**	**67.93**	65.07	81.12	83.70	83.12	78.19
EWC	84.64	80.22	78.23	75.85	82.02	81.52	81.32	**80.93**	65.83	64.88	62.74	82.88	76.99	76.36	76.74
Rand Mix	82.44	79.50	76.79	74.19	82.57	80.75	78.93	77.50	66.05	65.72	65.34	84.59	83.39	**83.58**	77.24
Max Abs	82.89	79.57	77.05	75.07	82.75	81.58	79.90	78.02	65.77	64.50	64.01	84.58	81.89	80.81	77.03
Avg	82.41	79.53	76.80	74.16	82.57	80.78	78.92	77.48	66.10	65.74	65.31	84.67	83.47	83.56	77.25
TIES	82.66	79.77	77.37	75.18	82.72	81.33	79.88	78.15	66.17	65.24	64.51	84.79	82.28	82.09	77.30
MagMax	**84.85**	**81.67**	**80.31**	**78.20**	**83.47**	**83.07**	**82.23**	80.82	69.14	67.05	**65.41**	**86.86**	**84.27**	83.11	**79.32**

A.2 Sign conflicts

Fig. 1 presents the sign conflicts for class-incremental, domain-incremental and 8 datasets scenarios. We observe that sequential fine-tuning significantly reduces sign conflicts similarly to CIL results presented in Figure 3 in the main paper.

A.3 Task agnostic per-task results

Figure 2 presents more per-task task-agnostic results doe MagMax.

B Additional analyses

B.1 Layer-wise weight changes

To better undestand the process of fine-tuning and merging with MagMax, we analyze the magnitudes of τ_{MagMax_t} parameters. We group these parameters either by their type (e.g. layer normalization, attention or MLP) or by the block index to which they belong. We present the analysis in Figure 3 and observe that the magnitudes of layer normalization are much higher that the magnitudes of other layers. Moreover, magnitude seems not to depend on the depth. Note, that we only analyze weight matrices and disregard the biases.

B.2 Distribution of parameters in task vectors

Figure 4 presents the distribution of parameters in the task vectors.

References

1. 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: ICML (2021)
2. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv: 2111.02114 (2021)

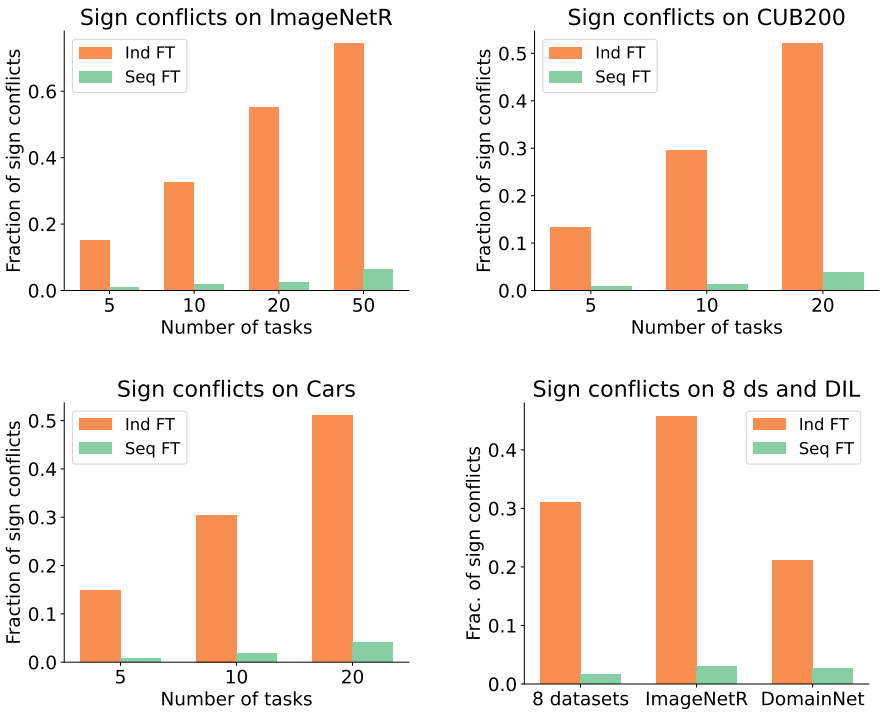

Figure 1: Sign conflicts for CIL, DIL and 8 datasets settings.

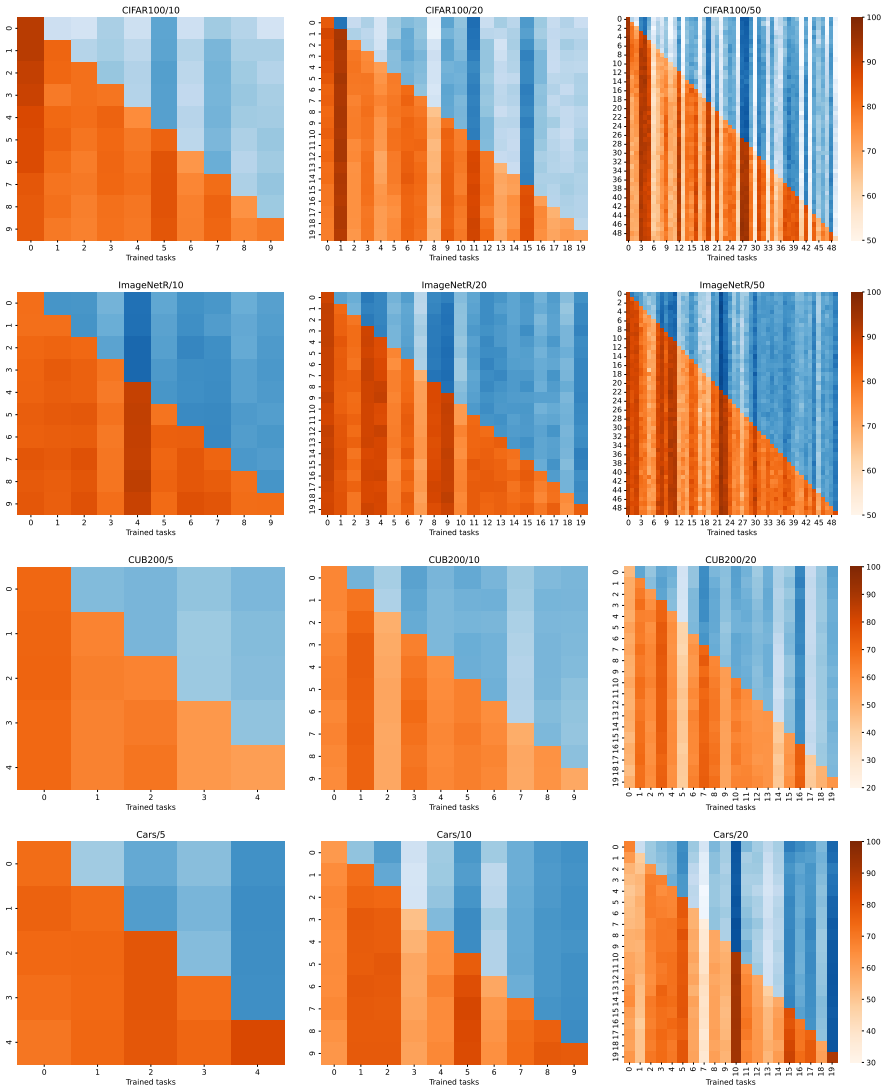

Figure 2: Task-agnostic results of MagMax in different settings.

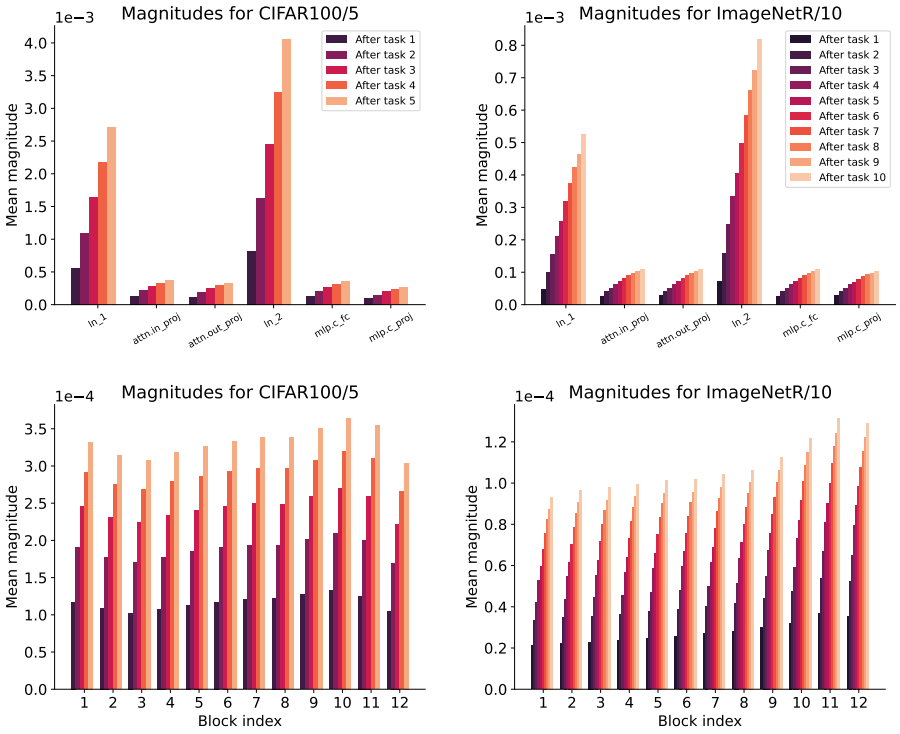

Figure 3: Mean magnitudes of τ_{MAGMAX_t} parameters grouped by layer type (top) or block index (bottom) for CIFAR100/5 (left) and ImageNetR/10 (right). **Top:** parameters of LayerNorm layers change the most. **Bottom:** the magnitude of parameter change does not depend much on a block index (depth).

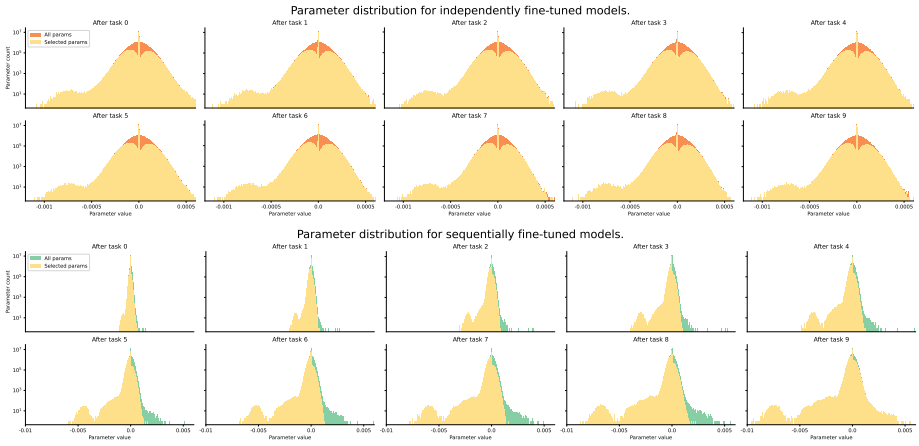

Figure 4: When fine-tuned independently (top), task vectors have similar distributions of parameters. Moreover, similar distribution contributes to the task vector merged by maximum magnitude selection. However, when fine-tuned sequentially (bottom), the distribution of parameters in task vectors differs – later task vectors have larger parameters ans, as a results, they contribute more to the final task vector. Note that the vertical axis is logarithmic and that the scale of the independent and sequential distributions differ.