Constrained Few-shot Class-incremental Learning

A. Appendix

A.1. Datasets

miniImageNet. miniImageNet contains RGB images of size 84×84 from 100 different classes, where each class has 500 training images and 100 test images. It is a subset of the much larger ImageNet dataset [41] and was initially proposed for few-shot learning problems [50]. For the FSCIL evaluation, we follow the same procedure as in [48], dividing the dataset into a base session with 60 classes and eight novel sessions with a 5-way 5-shot problem each.

CIFAR100. The setup of CIFAR100 [25] is similar to miniImageNet, whereby CIFAR contains 100 different classes with 500 training images and 100 testing images per class. The resolution of the images is lower (32×32). Also here, we follow the same FSCIL procedure as in [48] with 60 base classes and eight novel sessions with 5-way 5-shot problems.

Omniglot. The Omniglot dataset [27] has a total of 1623 classes with 20 example images each. It is publicly available under the MIT license. The images are binary with a size of 105×105. We resize all the images to 32×32 floating point format in a preprocessing step. As proposed by Vinyals et al. in [50] for the few-shot learning setting, we use 1200 base classes for the meta-learning and base session while the remaining 423 classes are reserved as the novel classes for the subsequent sessions. However, as there has been no previous work targeting Omniglot in the FSCIL setting, we consider the following points for the dataset.

First, to accommodate evaluation within the base classes, we hold out the last 6 samples from the base classes, leaving the first 14 samples for the training dataset. Second, we demarcate the first 5 samples from the next-in-line 47 novel classes as the incoming support batch during sessions subsequent to the base session, so that 9 subsequent sessions can be run with 1623 classes in total in the final session. Third, we add the first 6 of the remaining 15 examples from the novel classes for the evaluation query batch, so that novel and base classes are equally weighted during the evaluation.

A.2. Ablation study

A.2.1 Reducing Dimension

We analyse the classification accuracy by reducing the dimension $d \in \{32, 64, 128, 256, 512\}$ of the output of the fully connected layer and the EM for the three datasets, whereby the maximum number of classes ($|\tilde{\mathcal{C}}^{(S)}|$) is set to 100, 100, and 1623 in CIFAR100, miniImageNet, and Omniglot, respectively. Experimental results are shown in Table A1, Table A2, and Table A3. All training hyperparamters applied in the meta-learning and retraining are kept the same, irrespective of the dimensionality. Overall, in Mode 1 and Mode 2, high dimensionality, i.e., $d \geq 256$, leads to better accuracy. This stems from the fact that these two modes mainly rely on the property of hyperdimensional vectors—where higher dimensionality is preferred—to achieve quasi-orthogonality between class vectors. However, the optimization technique employed in Mode 3 is able to find a better distribution of the prototypes with a lower dimensionality of $d = 128 > |\tilde{\mathcal{C}}^{(S)}|$ for CIFAR and miniImageNet datasets. Specifically, we see the advantage of a lower number of dimensions when increasing the number of novel classes, provided the number of dimensions is larger than the total number of classes. For example, in the last session, C-FSCIL with $d = 128$ achieves the highest accuracy on both miniImageNet (51.46%) and CIFAR100 (50.74%) in Mode 3. We could not observe this effect in Omniglot (Table A3), where the highest dimensionality is lower than the number of classes (i.e., $d = 512 < |\tilde{\mathcal{C}}^{(S)}|$). In fact, $d = 512$ results in the highest accuracy for all the modes in Omniglot.

We remark that, compared to the state-of-the-art, the superior accuracy of C-FSCIL is still maintained even with $d < |\tilde{\mathcal{C}}^{(S)}|$:

- On miniImageNet, C-FSCIL with $d = 64$ outperforms [5–7, 44, 48, 56] in Mode 1 and Mode 3, and [5–7, 44, 48] in Mode 2.
- Similarly, on CIFAR100, C-FSCIL with $d = 64$ outperforms [5–7, 44, 48] in Mode 1 and Mode 2, and [5–7, 44, 48, 56] in Mode 3.
- Likewise, on Omniglot, C-FSCIL with $d = 128$ in any mode outperforms the prototypical network [45] and CEC [56].

A.2.2 Other Attention Functions

In this section, we provide additional details on the soft absolute (softabs) attention function applied in our meta-learning, and compare it to the exponential attention commonly used in the softmax.

Softabs attention. Given the cosine similarity score l_j for every class j (see (1) in the main paper), the softabs attention function is defined as

$$h(l_j) = \frac{\epsilon(l_j)}{\sum_{i=1}^{|\tilde{\mathcal{C}}^{(s)}|} \epsilon(l_i)}, \tag{11}$$

where $\epsilon(\cdot)$ is the sharpening function:

$$\epsilon(c) = \frac{1}{1 + e^{-(\beta(c-0.5))}} + \frac{1}{1 + e^{-(\beta(-c-0.5))}}, \quad \forall c \in \mathbb{R}. \tag{12}$$

The sharpening function includes a stiffness parameter β, which is set to 10 as in [21]. During meta-learning, the model is updated based on the negative log-likelihood loss

applied on $h(l_y)$. The sharpening function is maximized at $c = 1$ or $c = -1$, and minimized at $c = 0$. Hence, it promotes orthogonal prototypes. This notion of orthogonality is also reflected in the activation function applied in the prototype nudging in Mode 3 (see (8) in the main paper):

$$\sigma(c) := e^{\alpha \cdot c} + e^{-\alpha \cdot c} - 2, \forall c \in \mathbb{R}, \tag{13}$$

where $\alpha = 4$. This activation function penalizes the prototype pairs with large absolute cross-correlations.

Softmax attention. We compare the aforementioned softabs attention with the conventional exponential softmax, defined as

$$r(l_j) = \frac{e^{\tau \cdot l_j}}{\sum_{i=1}^{|\tilde{\mathcal{C}}^{(s)}|} e^{\tau \cdot l_i}}, \tag{14}$$

where $\tau = 10$ is the inverse softmax temperature. When applying the negative log-likelihood loss on $r(l_y)$, we get the commonly used categorical cross-entropy loss (CEL). The CEL aims to find anti-correlating prototypes. During prototype nudging in Mode 3, we therefore modify the activation function (13) with the objective of reaching anti-correlation:

$$\sigma'(c) := e^{\alpha \cdot c} - 1, \forall c \in \mathbb{R}, \tag{15}$$

where $\alpha = 4$ as in (13).

Comparison. We compare the classification accuracy when using either softmax or softabs attention on miniImageNet (Table A4), CIFAR100 (Table A5), and Omniglot (Table A6). In the case of the softmax attention, we also applied pretraining of the embedding and optimized the inverse softmax temperature with a grid-search. On miniImageNet, the softmax attention starts with marginally higher accuracy (0.1%) than the softabs attention in the base session, but it decays faster than the softabs when new sessions are added, independent of the mode. As a result, the softabs attention maintains higher accuracy, as high as 1.36%, during the novel sessions ($s > 1$). Similar results are observed on CIFAR100, where softabs outperforms softmax (up to 2.66%) in all sessions and all Modes 1–3. Similar results are also observed by using Mode 2 and Mode 3 on Omniglot; however, the softmax attention reaches consistently higher accuracy than softabs in Mode 1.

Fig. A1 illustrates the relations between the prototypes, either trained with the softabs or the softmax attention. When comparing the softmax with the softabs in Mode 1 (Fig. A1a vs. Fig. A1d) on the base session, the softabs attention yields cross-correlations that are close to zero (i.e., they are quasi-orthogonal), whereas the softmax promotes anti-correlating prototypes with negative cross-correlations. When new sessions are added, both attention functions yield cross-talk in Mode 1. This cross-talk is effectively reduced with the prototype nudging and retraining applied in Mode 3, where the exponential-based nudging in (15) (Fig. A1c) yields anti-correlating prototypes between the novel and the base session, whereas the double-exponential nudging in (13) (Fig. A1f) yields quasi-orthogonal prototypes. While the cross-talk is reduced on the novel classes with both activation functions, part of the class discriminability achieved during the base session is sacrificed. Fig. A2 shows similar trends on the Omniglot dataset. C-FSCIL using the softabs in Mode 1 (Fig. A2b) yields lower cross-correlations than softmax in Mode 1 (Fig. A2). The retraining in Mode 3 further reduces the cross-talk (Fig. A2c).

A.2.3 Smaller Feature Extractor

Most of the baseline methods [5–7, 44, 48] use a ResNet-18 backbone with the feature dimensionality $d_f = 512$, while we use a ResNet-12 as the feature extractor with $d_f = 640$, motivated by [11,55]. This higher feature dimensionality requires a larger number of trainable parameters[1] for ResNet-12 compared ResNet-18 (12.4 M vs. 11.2 M). Therefore, to have a fair comparison, we have also implemented a reduced ResNet-12 feature extractor with the block dimensions [64, 128, 256, $d_f = 512$] containing 8.0 M parameters. We name this smaller feature extractor as ResNet-12 (small). It results in 1.56× lower number of trainable parameters than ResNet-18.

Table A7 and Table A8 compare the performance on miniImageNet and CIFAR100, respectively, applying C-FSCIL with either the ResNet-12 (small) or the original ResNet-12 with $d_f = 640$. C-FSCIL using the ResNet-12 (small) maintains a high accuracy on both datasets and shows only small drops (<1%) compared to the original ResNet-12 with $d_f = 640$. Moreover, C-FSCIL with ResNet-12 (small) outperforms all the baselines [5–7, 44, 48, 56] on miniImageNet in all Modes 1–3, while requiring a lower number of trainable parameters. Similarly, on CIFAR100, C-FSCIL with ResNet-12 (small) outperforms the majority of baselines [5–7, 44, 48] in Modes 1–3 with a lower number of trainable parameters. When comparing to CEC [56], C-FSCIL with ResNet-12 (small) achieves a higher accuracy in Mode 2 and Mode 3. However, we observe that CEC uses ResNet-20, which requires a lower number of parameters.

A.3. Compression of the Explicit and GAA Memories

Here, we present a case where the memory requirements of our C-FSCIL can be further reduced by doing superposition of key-value bindings using holographic reduced representations [35]. We bind each prototype with a randomly drawn key and superimpose two key-prototype pairs, which compresses the memory by 2×. More formally, the first two

[1] Final fully connected layer excluded.

prototypes, $\mathbf{p}_1$ and $\mathbf{p}_2$, are compressed by

$$\mathbf{r} = \mathbf{p}_1 \circledast \mathbf{c}_1 + \mathbf{p}_2 \circledast \mathbf{c}_2, \tag{16}$$

where $\mathbf{c}_1$ and $\mathbf{c}_2$ are d-dimensional key-vectors randomly drawn from a normal distribution with variance $1/d$, and $\circledast$ is the circular convolution acting as binding operator. The keys are generated with a pseudorandom number generator (RG) with seed corresponding to key_i. We need to store only the seed key_i instead of the actual key vector. This key_i needs a negligible 32-bit storage per model since $\mathbf{c}_i$ can be reproduced from the key by RG. The key-value binding allows to retrieve the individual prototypes using the unbinding operation, e.g., the first prototype is retrieved by:

$$\hat{\mathbf{p}}_1 = \mathbf{r} \odot \mathbf{c}_1 \tag{17}$$

$$= \mathbf{p}_1 \circledast \mathbf{c}_1 \odot \mathbf{c}_1 + \mathbf{p}_2 \circledast \mathbf{c}_2 \odot \mathbf{c}_1 \tag{18}$$

$$\approx \mathbf{p}_1 + \mathbf{n} \tag{19}$$

where $\odot$ is the circular correlation and $\mathbf{n}$ a noise term, which decreases with increasing dimension d [35]. The presented compression can be applied in all modes. In Mode 1, the prototype vectors in the EM are compressed, whereas in Mode 3 the globally average activation vectors in the GAA memory are compressed.

Table A9 compares the accuracy of C-FSCIL with and without memory compression on miniImageNet. The compressed EM in Mode 1 remains accurate (1.7%–3.5% drop across the sessions), while the compressed GAA memory in Mode 3 yielded a larger loss (4.7%–8.5% drop). The superior accuracy of the compressed EM compared to the compressed GAA memory might stem from its quasi-orthogonal representation, which is not provided by the GAA memory.

A.4. Additional Baselines on Omniglot

For further comparison with the Omniglot dataset in the FSCIL setting, we create two new baselines based on Prototypical Networks and Continually Evolved Classifiers. For an additional comparison on Omniglot, we consider an alternative continual incremental learning setting developed by [2].

A.4.1 Prototypical Networks

The first baseline adapts the loss function and sharpening function of C-FSCIL to those used in Prototypical Networks [45]. Therefore, we call this baseline as ProtoNet*. ProtoNet* adopts the same feature extractor as C-FSCIL used for the Omniglot dataset. The averaged prototypes and query vector produced by the feature extractor are compared using the negative Euclidean distance metric, as suggested in [45]. This output attention vector goes through an exponential sharpening function, as given in (14). During the meta-learning phase, the ProtoNet* feature extractor is trained by applying the cross-entropy loss (CEL) on the sharpened attention activations. During the inference phase for the base session and the subsequent sessions, the averaged prototypes are computed using forward propagation of support examples through the meta-learned feature extractor and averaging the resulting output embeddings. For the prediction, the query vector produced by the feature extractor is compared against the averaged prototypes using the negative Euclidean distance metric. To have a fair comparison, we also varied the number of output embedding dimensions in ProtoNet*, although the original Prototypical Networks [45] used a fixed $d = 64$.

The resulting classification accuracy is presented in Table A3. C-FSCIL in any mode significantly outperforms ProtoNet* with the same d. For instance, with $d = 128$, C-FSCIL starts with 17.03% higher accuracy (80.78% vs. 63.75%) in the base session, and ends with 18.06% higher accuracy in the last session using Mode 1, which is similar to the prototype averaging applied in ProtoNet*. These accuracy gaps become larger by using either Mode 2 or Mode 3.

A.4.2 Continually Evolved Classifiers

The second baseline is the Continually Evolved Classifiers (CEC) [56], which achieved state-of-the-art accuracy on miniImageNet and CIFAR100 in FSCIL. After evaluating the performance of CEC with different feature extractors, including a ResNet-18, a ResNet-20, and the feature extractor from our C-FSCIL, we found the ResNet-20 to achieve the highest accuracy. Table A3 shows the accuracy when varying the embedding dimension. CEC achieved highest accuracy when the dimension is set to $d = 64$, which is indeed the default dimension of CEC. Overall, our C-FSCIL in Mode 3 outperforms the CEC baseline by a large margin of 8.30% and 9.59% in session 1 and 10, respectively.

A.4.3 Alternative FSCIL setting on Omniglot

We also consider an alternative continual incremental learning setting developed by [2], that arranges a larger number of classes in the novel sessions. In this setting, the model is meta-learned over the entire 964 base classes defined in the original Omniglot dataset, and tested on the 659 classes in the test dataset, while incrementally exposing 10 classes per session starting with 10 classes and finishing with 600 classes.

We compare our work with ANML [2] as the best performing model. The results are shown in Fig. A3: C-FSCIL consistently performs better than ANML and minimizes the accuracy degradation, as more novel classes are incrementally added from 10 to 600. ANML incurs a drop of 31.1% compared to 10.1% in our Mode 3. This indicates the higher scalability of C-FSCIL to cover a large number of classes in its lifespan.

Table A1. Dimension ablation on miniImageNet. Classification accuracy (%) of C-FSCIL in the 5-way 5-shot FSCIL setting.

Session (s)		1	2	3	4	5	6	7	8	9
No. of classes $\lvert\tilde{\mathcal{C}}^{(s)}\rvert$		60	65	70	75	80	85	90	95	100
Mode	d									
Mode 1	512	76.37	70.94	66.36	62.64	59.31	56.02	53.14	51.04	48.87
	256	**76.78**	**71.08**	**66.37**	**62.75**	**59.33**	**56.39**	**53.34**	**51.11**	**48.94**
	128	76.32	70.72	65.93	62.16	58.63	55.74	52.83	50.73	48.48
	64	76.30	70.74	65.91	62.32	59.00	55.76	52.76	50.48	48.30
	32	74.10	68.58	63.86	60.23	56.95	53.67	50.80	48.44	46.33
Mode 2	512	76.45	**71.23**	**66.71**	**63.01**	**60.09**	56.73	53.94	52.01	50.08
	256	**76.70**	70.95	66.19	62.80	59.65	**56.80**	**54.29**	**52.08**	**50.58**
	128	76.23	70.25	65.33	62.24	59.49	56.95	53.91	51.67	49.65
	64	75.83	68.54	62.57	58.59	56.09	53.49	50.62	48.66	46.95
	32	73.38	66.68	61.90	58.37	54.72	51.06	48.69	46.71	44.73
Mode 3	512	76.40	71.14	66.46	63.29	60.42	57.46	54.78	53.11	51.41
	256	**76.75**	**71.17**	**66.50**	**63.39**	**60.86**	58.05	55.30	53.08	51.41
	128	76.25	70.51	65.80	63.29	60.72	**58.18**	**55.63**	**53.44**	**51.46**
	64	76.30	70.55	65.36	62.49	59.76	57.01	54.00	51.51	49.41
	32	73.97	67.82	62.61	59.20	56.25	52.52	49.54	47.41	45.99

Table A2. Dimension ablation on CIFAR100. Classification accuracy (%) of C-FSCIL in the 5-way 5-shot FSCIL setting.

Session (s)		1	2	3	4	5	6	7	8	9
No. of classes $\lvert\tilde{\mathcal{C}}^{(s)}\rvert$		60	65	70	75	80	85	90	95	100
Mode	d									
Mode 1	512	**77.47**	**72.20**	**67.53**	**63.23**	**59.58**	**56.67**	**53.94**	**51.55**	**49.36**
	256	77.10	72.07	67.48	63.22	59.56	56.52	53.87	51.31	49.10
	128	76.95	71.78	67.09	62.75	59.08	56.35	53.63	51.24	48.72
	64	76.92	71.50	66.72	62.46	58.74	55.65	52.86	50.38	48.23
	32	74.85	69.82	64.91	60.33	56.99	53.92	51.32	48.87	46.75
Mode 2	512	**77.50**	**72.45**	**67.94**	**63.80**	**60.24**	**57.34**	54.61	**52.41**	**50.23**
	256	77.37	72.29	67.96	63.57	60.04	57.02	**54.63**	52.15	50.13
	128	76.93	72.28	67.44	63.69	59.83	57.55	55.01	52.24	49.33
	64	76.60	70.60	66.23	62.08	58.56	55.95	53.00	50.33	48.06
	32	74.52	69.09	64.06	59.56	56.00	52.95	50.26	47.85	45.22
Mode 3	512	**77.47**	**72.40**	67.47	63.25	59.84	56.95	54.42	52.47	50.47
	256	77.13	72.05	**67.66**	**63.65**	**60.10**	57.27	55.07	52.73	50.70
	128	77.00	72.28	67.40	63.45	59.72	**57.59**	**55.33**	**53.01**	**50.74**
	64	77.00	71.45	67.26	63.00	59.75	56.94	54.41	51.92	49.20
	32	74.92	68.98	64.20	58.81	55.41	52.56	50.04	47.41	45.33

Table A3. Dimension ablation on Omniglot. Classification accuracy (%) on Omniglot in the 47-way 5-shot FSCIL setting.

Session (s)		1	2	3	4	5	6	7	8	9	10
No. of classes $\lvert\tilde{\mathcal{C}}^{(s)}\rvert$		1200	1247	1294	1341	1388	1435	1482	1529	1576	1623
Mode/ Work	d										
Mode 1	512	**84.16**	**83.82**	**83.69**	**83.32**	**83.22**	**82.78**	**82.70**	**82.32**	**81.77**	**81.56**
	256	82.26	81.99	81.95	81.73	81.74	81.42	81.42	81.18	80.70	80.39
	128	80.78	80.97	80.50	80.24	79.80	79.45	79.01	78.41	78.11	78.19
	64	72.89	72.87	72.49	72.18	71.52	70.85	70.73	69.87	69.59	69.30
	32	57.07	56.70	56.14	55.72	54.97	54.36	53.85	53.19	52.45	52.52
Mode 2	512	**86.87**	**86.77**	**86.57**	**86.44**	**86.40**	**86.20**	**86.25**	**85.96**	**85.63**	**85.49**
	256	84.32	84.35	84.23	83.94	84.02	83.93	83.86	83.78	83.44	83.19
	128	82.85	83.29	82.70	82.65	82.14	82.03	81.91	81.31	80.68	81.15
	64	76.07	76.73	76.46	75.98	75.86	75.17	75.20	74.31	74.43	74.28
	32	64.31	63.71	63.94	63.72	63.04	62.66	62.04	61.58	61.35	61.15
Mode 3	512	**87.21**	**87.03**	**86.89**	**86.60**	**86.43**	**86.32**	**86.13**	**85.98**	**85.59**	**85.70**
	256	84.59	84.57	84.39	84.11	84.25	83.89	83.95	83.94	83.62	83.35
	128	83.51	83.29	83.14	82.87	82.54	81.90	82.03	81.50	81.29	81.31
	64	76.67	76.30	76.46	76.22	75.66	75.51	75.53	74.19	74.23	74.25
	32	64.99	64.66	64.61	63.72	63.34	63.14	62.66	62.23	61.69	62.17
ProtoNet [45]	256	**70.61**	**70.20**	**70.01**	**69.68**	**69.48**	**68.99**	**68.74**	**68.07**	**67.60**	-
	128	63.75	63.34	63.15	62.44	62.38	62.00	61.61	61.29	60.68	60.13
	64	49.69	49.10	48.63	48.13	47.67	46.97	46.73	46.11	45.47	45.08
	32	36.53	36.11	35.74	35.45	34.87	34.37	34.08	33.42	32.99	32.71
CEC [56]	512	76.44	76.62	76.21	76.10	75.37	74.92	74.60	74.04	73.43	73.19
	256	76.94	76.94	76.56	76.35	75.62	75.20	74.84	74.45	73.94	73.59
	128	77.10	77.15	76.94	76.89	76.26	75.79	75.46	75.05	74.58	74.28
	64	**78.91**	**79.07**	**78.74**	**78.60**	**77.94**	**77.55**	**77.18**	**76.77**	**76.39**	**76.11**
	32	74.51	74.59	74.32	73.97	73.31	72.76	72.28	71.84	71.55	71.41

Table A4. Attention ablation on miniImageNet. Classification accuracy (%) of C-FSCIL with $d = 512$ in the 5-way 5-shot FSCIL setting.

Session (s)		1	2	3	4	5	6	7	8	9
No. of classes $\lvert\tilde{\mathcal{C}}^{(s)}\rvert$		60	65	70	75	80	85	90	95	100
Mode	Attention									
Mode 1	softabs	76.37	**70.94**	**66.36**	**62.64**	**59.31**	**56.02**	**53.14**	**51.04**	**48.87**
Mode 1	softmax	**76.48**	**70.94**	**66.36**	62.43	58.94	55.67	52.56	50.21	47.91
Mode 2	softabs	76.45	**71.23**	**66.71**	**63.01**	**60.09**	**56.73**	**53.94**	**52.01**	**50.08**
Mode 2	softmax	**76.48**	71.18	66.67	62.80	59.54	56.33	53.04	51.04	49.09
Mode 3	softabs	76.40	**71.14**	**66.46**	**63.29**	**60.42**	**57.46**	**54.78**	**53.11**	**51.41**
Mode 3	softmax	**76.47**	70.86	65.90	62.53	59.72	56.88	54.47	51.83	50.05

Table A5. Attention ablation on CIFAR100. Classification accuracy (%) of C-FSCIL with $d = 512$ in the 5-way 5-shot FSCIL setting.

Session (s)		1	2	3	4	5	6	7	8	9
No. of classes $\lvert\tilde{\mathcal{C}}^{(s)}\rvert$		60	65	70	75	80	85	90	95	100
Mode	Attention									
Mode 1	softabs	**77.47**	**72.20**	**67.53**	**63.23**	**59.58**	**56.67**	**53.94**	**51.55**	**49.36**
Mode 1	softmax	76.35	71.03	66.31	62.13	58.45	55.40	52.69	50.20	47.99
Mode 2	softabs	**77.50**	**72.45**	**67.94**	**63.80**	**60.24**	**57.34**	**54.61**	**52.41**	**50.23**
Mode 2	softmax	76.33	71.09	66.49	62.16	58.8	55.78	53.17	50.75	48.39
Mode 3	softabs	**77.47**	**72.40**	**67.47**	**63.25**	**59.84**	**56.95**	**54.42**	**52.47**	**50.47**
Mode 3	softmax	76.35	70.88	65.96	61.99	58.17	55.00	52.38	49.92	47.81

Table A6. Attention ablation on Omniglot. Classification accuracy (%) of C-FSCIL with $d = 512$ in the 47-way 5-shot FSCIL setting.

Session (s)		1	2	3	4	5	6	7	8	9	10
No. of classes $\lvert\tilde{\mathcal{C}}^{(s)}\rvert$		1200	1247	1294	1341	1388	1435	1482	1529	1576	1623
Mode	Attention										
Mode 1	softabs	84.16	83.82	83.69	83.32	83.22	82.78	82.70	82.32	81.77	81.56
Mode 1	softmax	**85.42**	**85.15**	**85.05**	**84.69**	**84.57**	**84.22**	**84.11**	**83.94**	**83.57**	**83.54**
Mode 2	softabs	86.87	**86.77**	86.57	**86.44**	**86.40**	**86.20**	**86.25**	**85.96**	**85.63**	**85.49**
Mode 2	softmax	**86.93**	86.69	**86.60**	86.42	86.17	86.00	86.03	85.78	85.40	85.36
Mode 3	softabs	**87.21**	**87.03**	**86.89**	**86.60**	**86.43**	**86.32**	**86.13**	**85.98**	**85.59**	**85.70**
Mode 3	softmax	87.01	86.89	86.80	86.57	86.29	86.08	86.02	85.78	85.36	85.33

Table A7. Feature extractor ablation on miniImageNet. Classification accuracy (%) in the 5-way 5-shot FSCIL setting.

Session (s)		1	2	3	4	5	6	7	8	9
No. of classes $\lvert\tilde{\mathcal{C}}^{(s)}\rvert$		60	65	70	75	80	85	90	95	100
Mode/ Work	Feature extractor									
AL-MML [48]	ResNet-18	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42
IDLVQ-C [5]	ResNet-18	64.77	59.87	55.93	52.62	49.88	47.55	44.83	43.14	41.84
Semantic KD [6]	ResNet-18	<62	<59	<54	<50	<49	<45	<42	<40	<39
VAE [7]	ResNet-18	<62	<60	<54	<52	<50	<49	<46	<44	<43
F2M [44]	ResNet-18	67.28	63.80	60.38	57.06	54.08	51.39	48.82	46.58	44.65
CEC [56]	ResNet-18	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63
C-FSCIL Mode 1	ResNet-12	76.37	70.94	66.36	62.64	59.31	56.02	53.14	51.04	48.87
C-FSCIL Mode 2	ResNet-12	76.45	71.23	66.71	63.01	60.09	56.73	53.94	52.01	50.08
C-FSCIL Mode 3	ResNet-12	76.40	71.14	66.46	63.29	60.42	57.46	54.78	53.11	51.41
C-FSCIL Mode 1	ResNet-12 (small)	76.08	70.63	66.11	62.23	58.91	56.12	53.11	51.02	48.93
C-FSCIL Mode 2	ResNet-12 (small)	75.90	70.52	66.01	62.11	58.86	56.19	53.23	51.31	49.53
C-FSCIL Mode 3	ResNet-12 (small)	76.12	70.20	65.29	62.25	59.35	56.76	54.18	52.15	50.47

Table A8. Feature extractor ablation on CIFAR100. Classification accuracy (%) in the 5-way 5-shot FSCIL setting.

Session (s)		1	2	3	4	5	6	7	8	9
No. of classes $\lvert\tilde{\mathcal{C}}^{(s)}\rvert$		60	65	70	75	80	85	90	95	100
Mode/ Work	Feature extractor									
AL-MML [48]	ResNet-18	64.10	55.88	47.07	45.16	40.11	36.38	33.96	31.55	29.37
Semantic KD* [6]	ResNet-18	<64	<57	<51	<46	<43	<41	<39	<37	<35
VAE* [7]	ResNet-18	<62	<58	<57	<52	<51	<49	<46	<45	<42
F2M [44]	ResNet-18	64.71	62.05	59.01	55.58	52.55	49.96	48.08	46.67	44.67
CEC [56]	ResNet-20	73.07	68.88	65.26	61.19	58.09	55.57	53.22	51.34	49.14
C-FSCIL Mode 1	ResNet-12	77.47	72.20	67.53	63.23	59.58	56.67	53.94	51.55	49.36
C-FSCIL Mode 2	ResNet-12	77.50	72.45	67.94	63.80	60.24	57.34	54.61	52.41	50.23
C-FSCIL Mode 3	ResNet-12	77.47	72.40	67.47	63.25	59.84	56.95	54.42	52.47	50.47
C-FSCIL Mode 1	ResNet-12 (small)	76.58	71.51	66.79	62.49	58.8	55.72	52.91	50.56	48.39
C-FSCIL Mode 2	ResNet-12 (small)	76.57	71.86	67.34	63.05	59.46	56.42	53.80	51.37	49.26
C-FSCIL Mode 3	ResNet-12 (small)	76.58	71.74	66.71	62.20	58.94	56.21	53.63	51.41	49.50

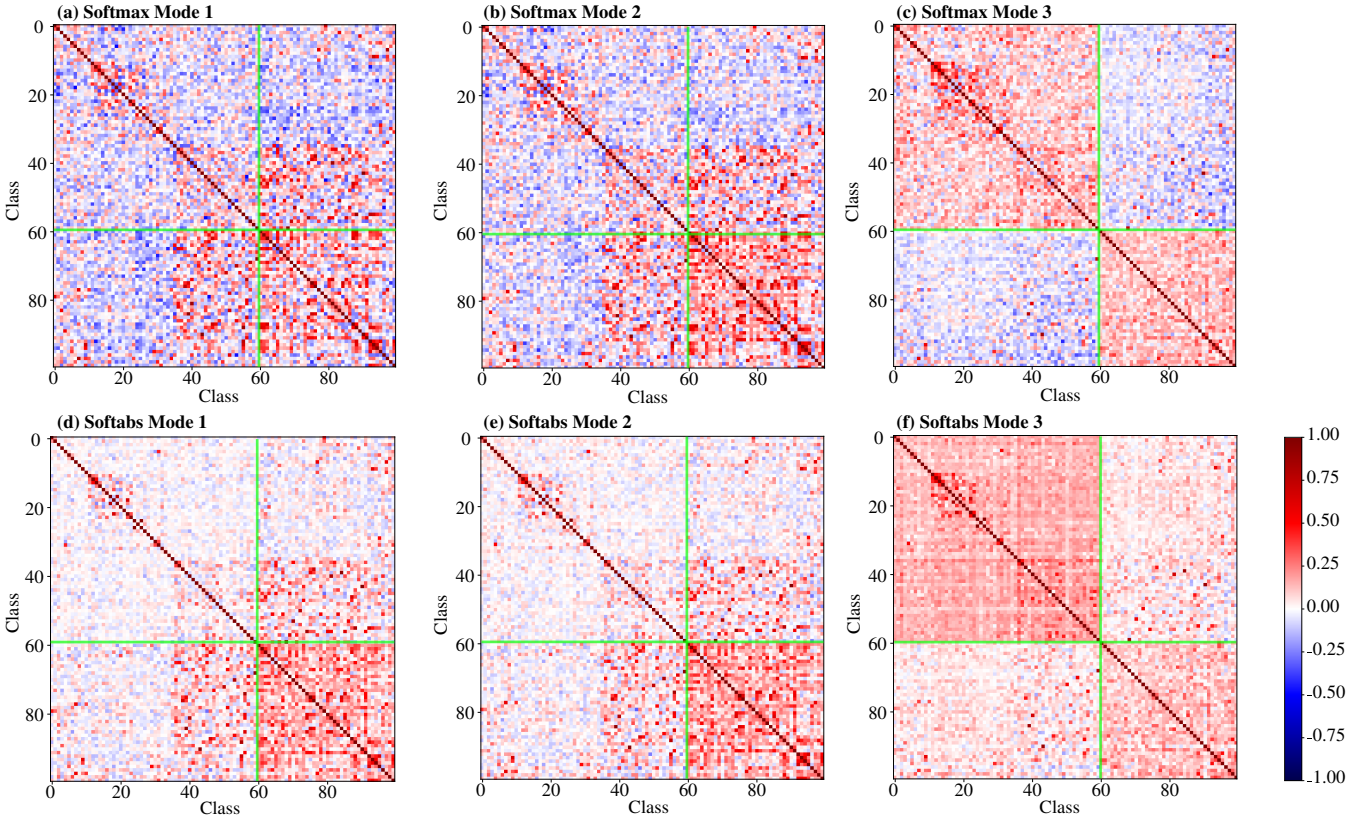

Figure A1. Cosine similarities between the prototypes on miniImageNet using different attention functions (softmax vs softabs) across the three modes. The green cross splits the base session (60 classes) and the novel sessions (40 classes in total).

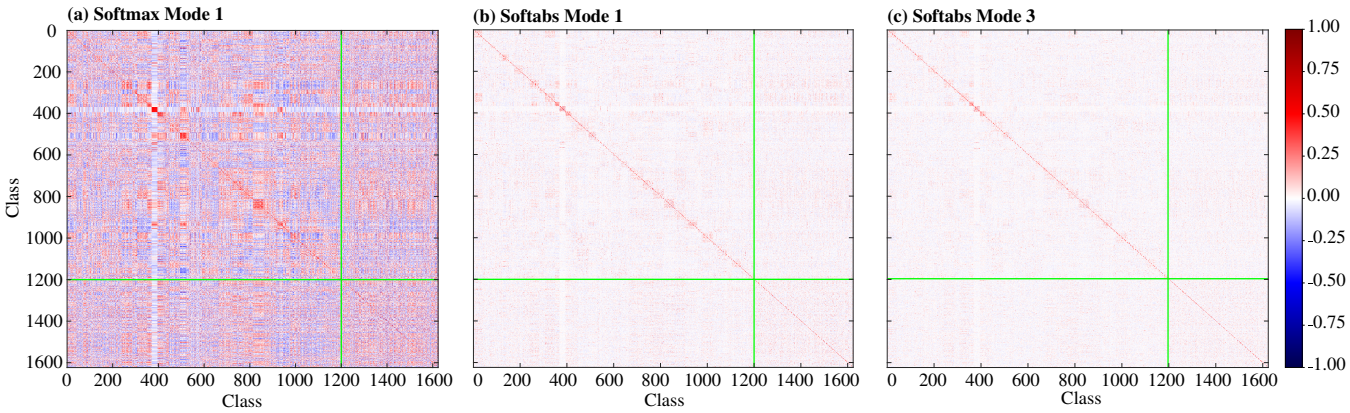

Figure A2. Cosine similarities between the prototypes on Omniglot in Mode 1 using different attention functions (softmax vs softabs). It also compares Mode 1 and Mode 3 when using softabs attention function. The green cross splits the base session (1200 classes) and the novel sessions (423 classes in total).

Table A9. Memory compression on miniImageNet. Classification accuracy (%) of C-FSCIL in the 5-way 5-shot FSCIL setting.

Session (s)		1	2	3	4	5	6	7	8	9
No. of classes $\lvert\tilde{\mathcal{C}}^{(s)}\rvert$		60	65	70	75	80	85	90	95	100
Mode	Compression									
Mode 1	No compression	76.37	70.94	66.36	62.64	59.31	56.02	53.14	51.04	48.87
Mode 1	2× compressing EM	74.65	69.31	64.10	60.43	56.84	53.51	49.94	48.05	45.34
Mode 3	No compression	76.40	71.14	66.46	63.29	60.42	57.46	54.78	53.11	51.41
Mode 3	2× compressing GAAM	71.72	66.40	61.41	57.13	53.56	50.38	47.74	45.28	42.91

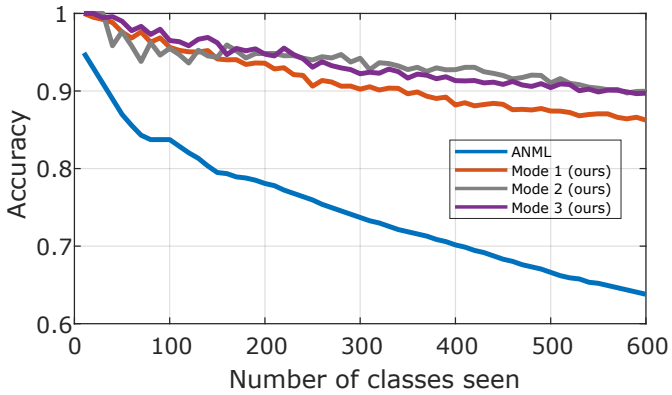

Figure A3. Classification accuracy (%) on Omniglot in the alternative FSCIL setting with c-way 5-shot where c is the number of seen classes; ANML refers to the best performing model in [2].