Supplementary Material: Few-Shot and Continual Learning with Attentive Independent Mechanisms

Eugene Lee Cheng-Han Huang Chen-Yi Lee
Institute of Electronics, National Chiao Tung University
Hsinchu, Taiwan
{eugenelet.ee06g, huang50213.ee04}@nctu.edu.tw cylee@si2lab.org

1. Few-Shot Learning

1.1. Observation of Attention Weight Change over Training Epochs

In this section, we provide an empirical study on the change of the attention weight corresponding the the input dimension of the attention score, $\mathrm{softmax}\left(\frac{\langle \mathbf{h}_m W_m^{\mathrm{Q}}, \hat{\mathbf{z}} W^{\mathrm{K}} \rangle}{\sqrt{d}}\right)$, over the entire training epochs. This study provides two insights: 1. the dynamics of the activation of mechanisms based on the number of training samples (shots); 2. the distribution of active and inhibited mechanisms over the training iterations. To show the dynamical change in a 2D plot, we sample classes from the validation set and observe them for the entire training process. Plots that uses Conv-4-64 and WRN-28-10 as backbone is shown in Figure 1 and Figure 2 respectively. From the plots, we can see that the activation of mechanisms are initially distributed uniformly followed by slow convergence to a sparse distribution over the training epochs, having only a few active mechanisms upon convergence. The active and inhibited attention weights are also clearly separated for all examples. Another observation is that having a larger number of training samples (shots), a smoother convergence for the activation weights across the training epochs is obtained. Smooth convergence is also obtained when a deeper backbone (WRN-28-10) is used when compared to a shallower one (Conv-4-64). This observation is intuitive as AIM is able to learn more efficiently when more samples or higher quality input features are provided, enabling the mechanisms to better model higher-order factorized information.

Competitive selection of mechanisms. From the figures shown, a distinct gap between active and inhibited mechanisms can be clearly observed. This motivates the idea of basing the activation of mechanisms on its corresponding attention value (soft decision) instead of making a hard decision that selects a total of K AIM on every inference. To demonstrate if basing the activation of AIM on the attention value would work, a simple experiment can be performed by allowing a mechanism to be active if its attention value is above 0.5 (similar to ReLU [5]), or:

$$\tilde{\mathbf{z}} = \hat{\mathbf{z}} \left(\sum_{m=1}^{M} w_m(\hat{\mathbf{z}}) W_m^{\mathrm{M}} \right), \tag{1}$$

where $w_m(\hat{\mathbf{z}})$ is given as,

$$w_m(\hat{\mathbf{z}}) = \begin{cases} \widetilde{w}_m(\hat{\mathbf{z}}), & \text{if } m \in \{m \mid \widetilde{w}_m(\hat{\mathbf{z}}) > 0.5 \text{ and} \\ & \qquad\qquad 1 \le m \le M\}, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

and $\widetilde{w}_m(\hat{\mathbf{z}})$ is defined as,

$$\widetilde{w}_m(\hat{\mathbf{z}}) = \mathrm{softmax}\left(\frac{\langle \mathbf{h}_m W_m^{\mathrm{Q}}, \hat{\mathbf{z}} W^{\mathrm{K}} \rangle}{\sqrt{d}} \right). \tag{3}$$

To keep our experiments simple, we do not induce stochasticity in the original approach and fix $K = 8$ to provide a fair comparison. We name the original method that keeps 8 mechanisms active as *hard decision* and name the method in (1) – (3) as *soft decision*. Comparison between hard decision and soft decision is shown in Table 1. From the results, we can see that when Conv-4-64 is used as backbone, higher accuracy is obtained when hard decision is used. The opposite can be observed when WRN-28-10 is used as backbone. We deduce that when extracted features are more reliable, i.e. through the use of deeper backbone or higher number of shots, the attention weights are of higher quality leading to clear distinction between relevant and less relevant mechanisms.

1.2. Observation of Attention Weight for All Classes

Different from the previous section, we show the mask instead of the attention weights here. The masks have a value of 1 for active mechanisms and 0 for inhibited mechanisms having the competitive selection based on the attention weights; for all experiments, $K = 8$ mechanisms

Table 1: Results for the comparison of using either hard decision (proposed method; $K = 8$ with $l = 0$) or soft decision (1) – (3) for the activation of mechanisms during inference. Average classification accuracies with 95% confidence intervals on the test-set are shown.

Backbone	Method	MiniImageNet, 5-Way		CIFAR-FS, 5-Way	
		1-shot	5-shot	1-shot	5-shot
Conv-4-64	Hard decision	$\mathbf{61.90 \pm 0.56\%}$	$\mathbf{74.55 \pm 0.38\%}$	$\mathbf{70.80 \pm 0.61\%}$	$\mathbf{80.50 \pm 0.40\%}$
	Soft decision	$61.74 \pm 0.57\%$	$74.41 \pm 0.38\%$	$70.18 \pm 0.61\%$	$80.38 \pm 0.39\%$
WRN-28-10	Hard decision	$\mathbf{71.03 \pm 0.57\%}$	$82.30 \pm 0.33\%$	$79.19 \pm 0.55\%$	$87.04 \pm 0.36\%$
	Soft decision	$69.96 \pm 0.56\%$	$\mathbf{82.37 \pm 0.33\%}$	$\mathbf{80.14 \pm 0.55\%}$	$\mathbf{87.41 \pm 0.35\%}$

will be active during both training and inference. We set $l = 2$ to induce stochasticity during training. Instead of sampling a single sample from each class, we take the average of the masks of each class accumulated across the entire validation set. We show heatmaps covering all classes and all 32 AIMs mechanisms on the first and final training epochs. Results that use Conv-4-64 as backbone using shots-dataset pair of 1-shot-CIFAR-FS, 5-shot-CIFAR-FS, 1-shot-MiniImageNet and 5-shot-MiniImageNet are shown in Figure 3, Figure 4, Figure 5 and Figure 6 respectively. Results that use WRN-28-10 as backbone using shots-dataset pair of 1-shot-CIFAR-FS, 5-shot-CIFAR-FS, 1-shot-MiniImageNet and 5-shot-MiniImageNet are shown in Figure 7, Figure 8, Figure 9 and Figure 10 respectively. By observing the heatmaps, we can see that the activation of mechanisms in the first epoch are uniformly distributed whereas in the final epoch, only a few set of mechanisms that are jointly used between classes accompanied by a sparse set of mechanisms that are invariant among samples from the same class. This observation meets our expectation of learning a set of experts that are each responsible for certain task. To give a better illustration on the transition of the heatmaps over the training epochs, we have attached several **.mp4** files that follows the naming convention of mask-*DATASET_MODEL_#*shot (italicized as wildcard strings) along with the supplementary materials.

1.3. Manipulating the Stochastic Sampling and Active Mechanisms Count

In this section, we show tabulated results of the manipulation of stochastic sampling and active mechanisms count as found in the main paper. The plots in the main paper show zero mean-ed results whereas the actual accuracy is reported in Table 2 and Table 3 for the maniputation of stochastic sampling count and active mechanisms count respectively. By looking at the tabulated results, we can say that the introduction of some stochasticity on the competitive selection of mechanisms during training is beneficial for the overall performance.

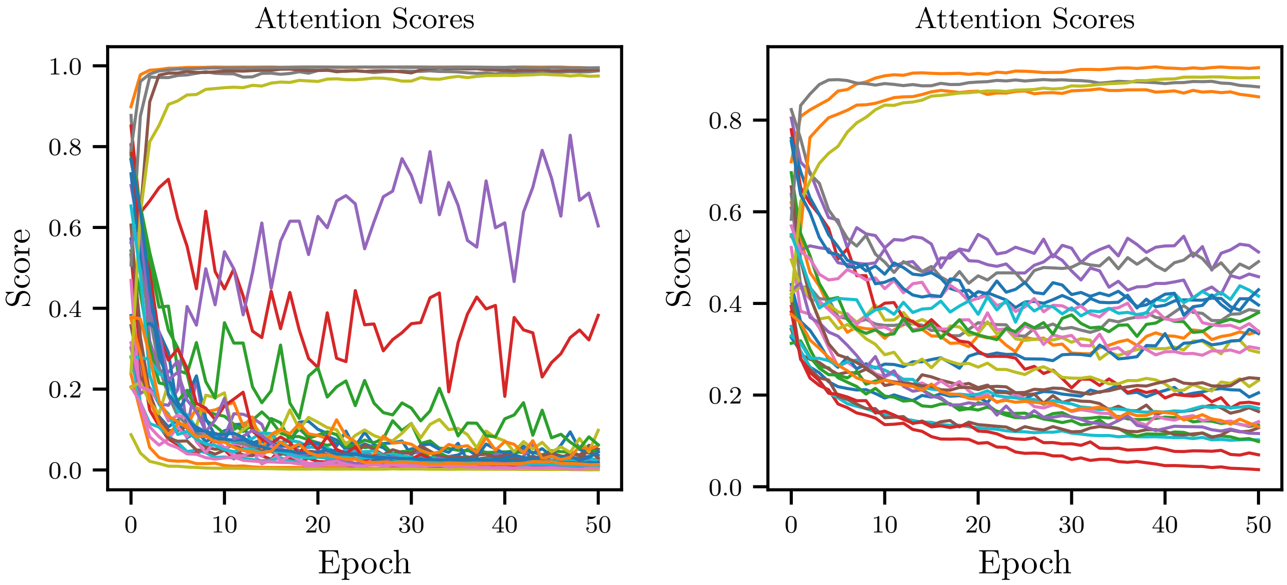
2. Continual Learning

2.1. Quantitative Analysis

We show the plots from the main paper comparing different continual learning methods in Figure 14 with the accompanied tabulated data in Table 4. Baseline shown correspond to the swapping of AIM layer with a single linear layer with number of parameters close to the originally introduced AIM layer to demonstrate that the increase in accuracy is not from over-parameterization. From the results, we can observe that with the addition of AIM as a module for continual learning, consistent improvement in accuracy can be obtained. It is also shown that the gain in accuracy does not result from the increase in parameters as shown by the accuracy attained using the baseline method.

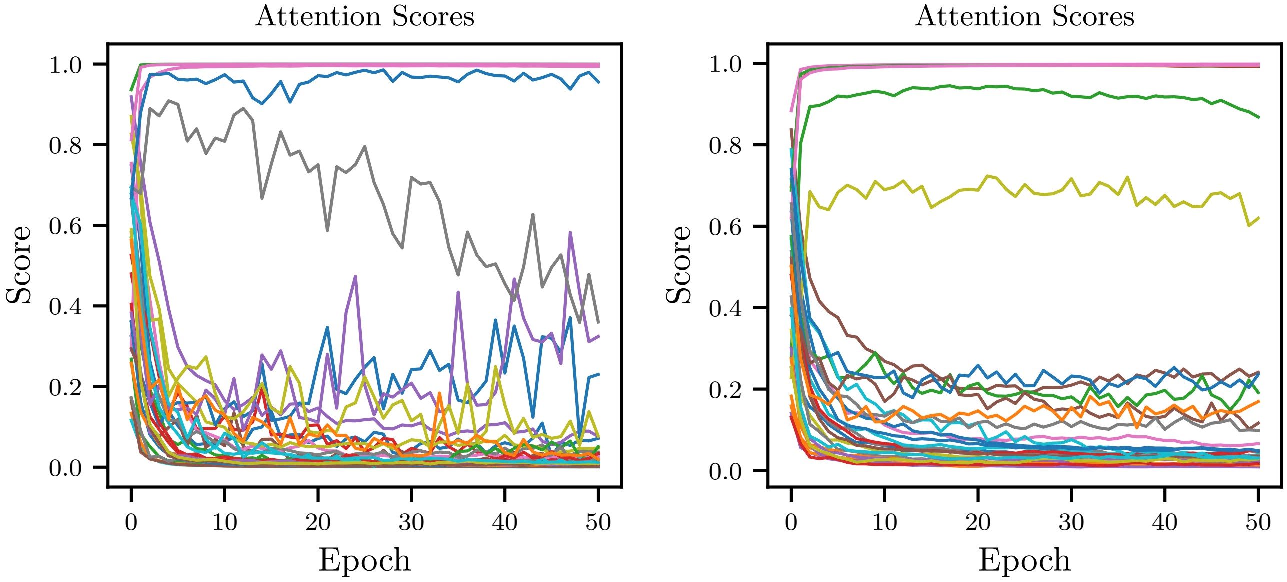
2.2. Activation of AIM

Similar to the analysis done for the activations of mechanisms for few-shot, we show the activations of mechanisms when AIM is used for continual learning. We apply AIM to both OML [2] and ANML [1] with activation heatmaps when trained on Omniglot [4], CIFAR-100 [3] and MiniImageNet [6] in Figure 14a, Figure 14b and Figure 14c respectively. We can observe that for Omniglot, the activation of mechanisms are sparsely distributed when compared to the activations obtained using CIFAR-100 and MiniImageNet. We conjecture that this is due to the simplicity of extracted representations, resulting in simpler higher-order modeling by the mechanisms. For natural images like CIFAR-100 and MiniImageNet, the features are not as distinct as the alphabets found in Omniglot, hence higher-order modeling of representations isn't as sparsely distributed. The sparsely distributed activations found in Omniglot result in distinctive increase in accuracy when compared to other datasets as shown in Table 4, e.g. at a trajectory containing 600 classes, the relative increase in accuracy when AIM is applied to OML is 20.70%. Even when MiniImageNet is used, distinctive increment in accuracy can also be observed, e.g. 19.40% relative increment in accuracy when AIM is applied to ANML, which be believe is due to the richness of information embedded in the latent representation resulting from the larger image size of 84×84. We believe that larger gain in accuracy can be attained through the introduction of a better feature extractor, i.e. an alternative to convolutional layers.



(a) CIFAR-FS, 1-shot.

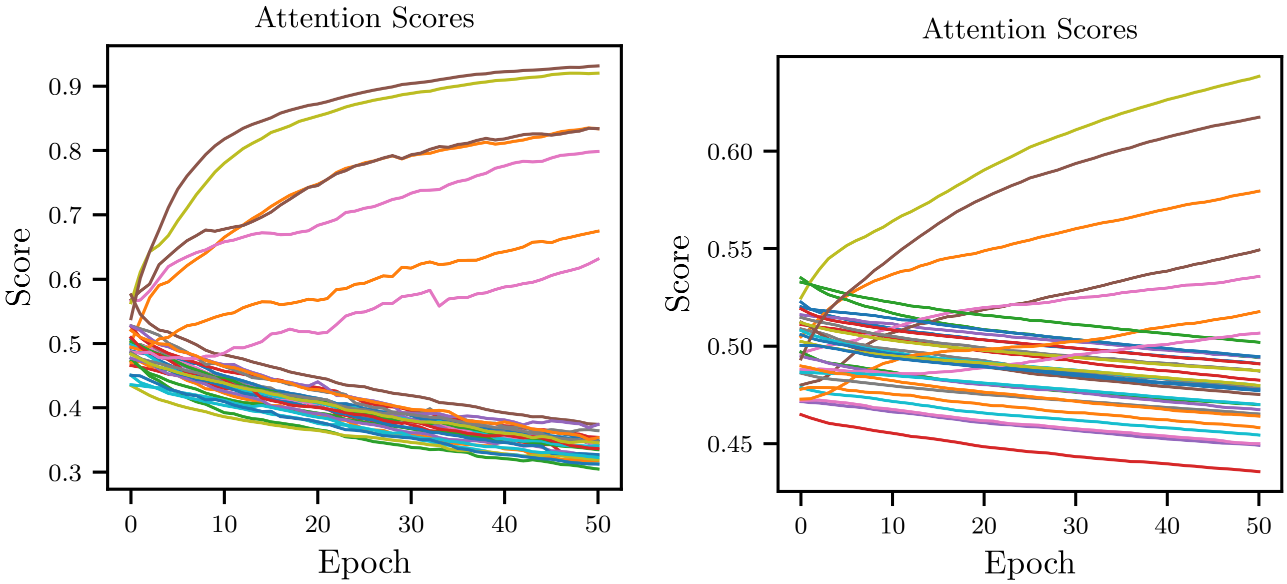
(b) CIFAR-FS, 5-shot.



(c) MiniImageNet, 1-shot.

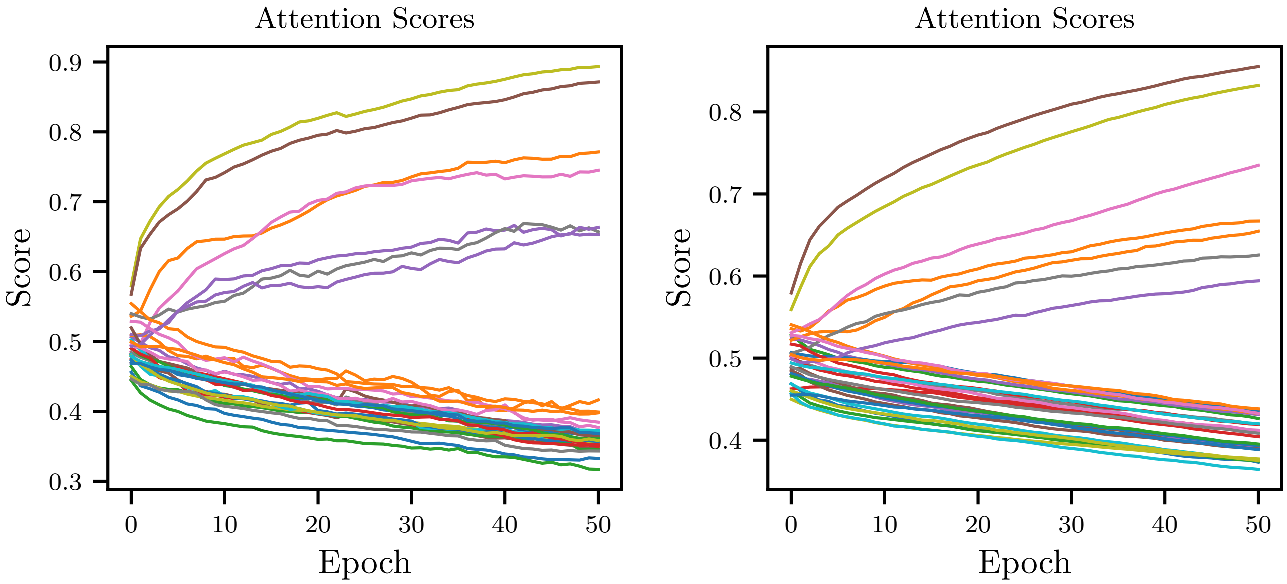
(d) MiniImageNet, 5-shot.

Figure 1: Change of attention weight or score corresponding to the input dimension over the training epochs. Different datasets pairs with different amount of training samples (shots) are shown here, using Conv-4-64 as its backbone. Each line represent an independent mechanism.



(a) CIFAR-FS, 1-shot.

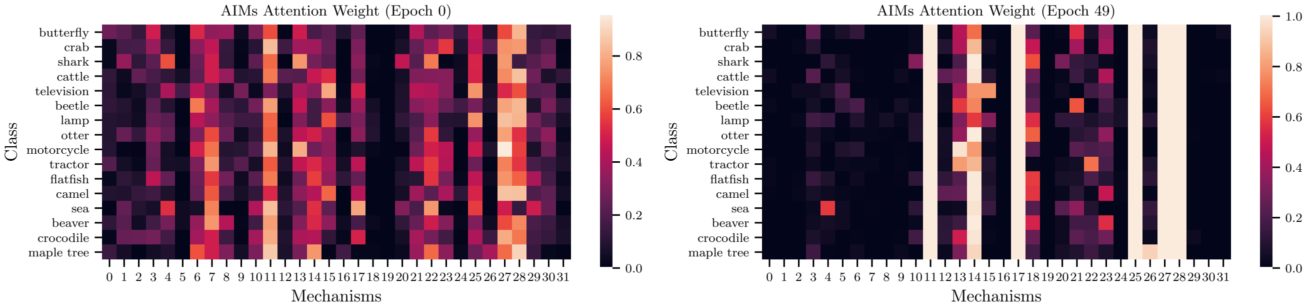
(b) CIFAR-FS, 5-shot.



(c) MiniImageNet, 1-shot.

(d) MiniImageNet, 5-shot.

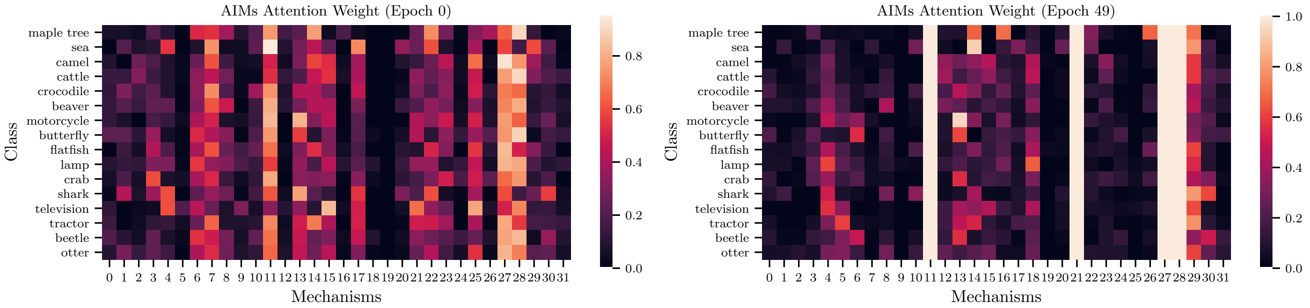
Figure 2: Change of attention weight or score corresponding to the input dimension over the training epochs. Different datasets pairs with different amount of training samples (shots) are shown here, using WRN-28-10 as its backbone. Each line represent an independent mechanism.



(a) First epoch.

(b) Last epoch.

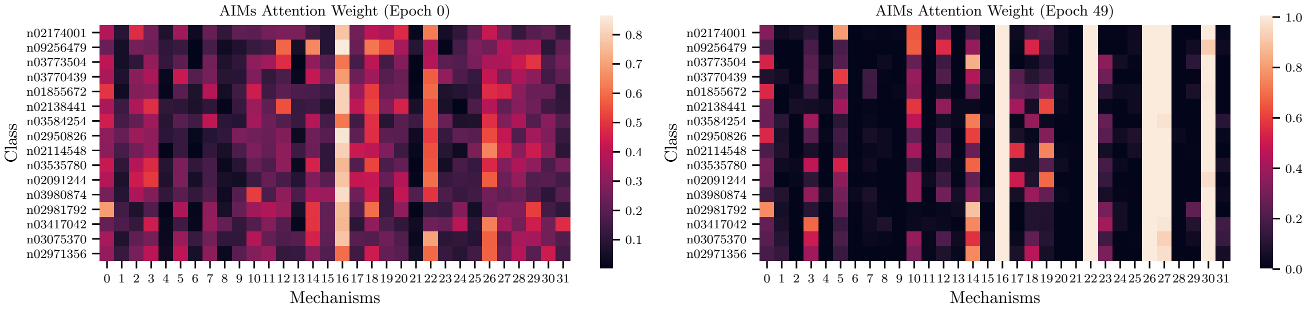
Figure 3: Activation of AIM from few-shot learning. Training on CIFAR-FS with 1-shot using Conv-4-64 as backbone.



(a) First epoch.

(b) Last epoch.

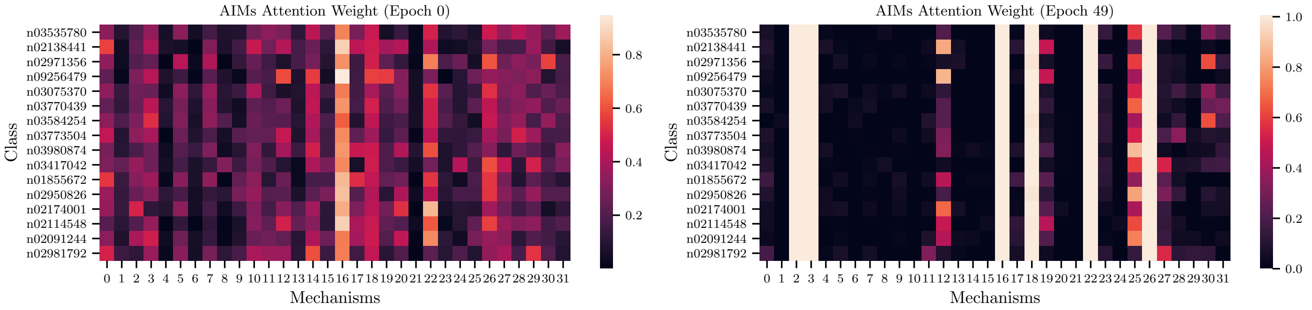
Figure 4: Activation of AIM from few-shot learning. Training on CIFAR-FS with 5-shot using Conv-4-64 as backbone.



(a) First epoch.

(b) Last epoch.

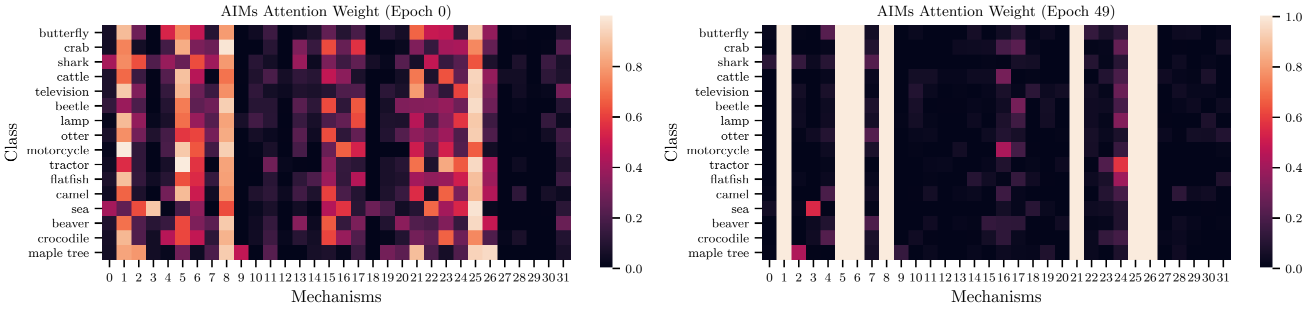
Figure 5: Activation of AIM from few-shot learning. Training on MiniImageNet with 1-shot using Conv-4-64 as backbone.



(a) First epoch.

(b) Last epoch.

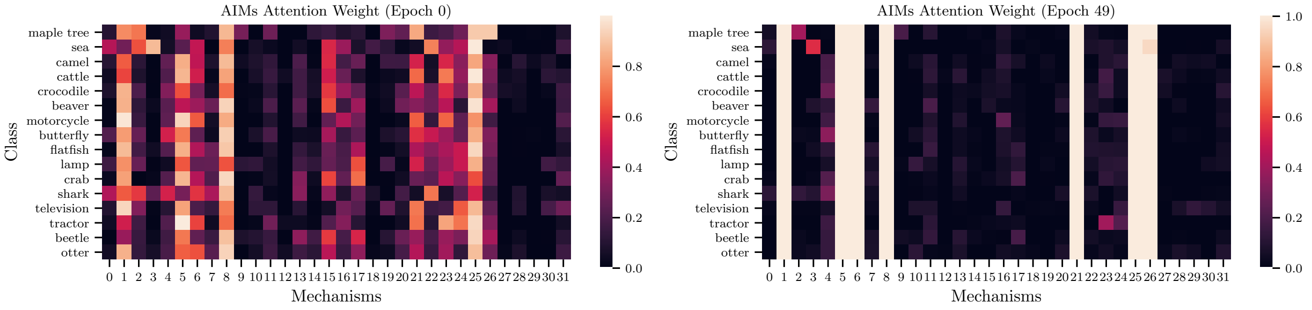
Figure 6: Activation of AIM from few-shot learning. Training on MiniImageNet with 5-shot using Conv-4-64 as backbone.



(a) First epoch.

(b) Last epoch.

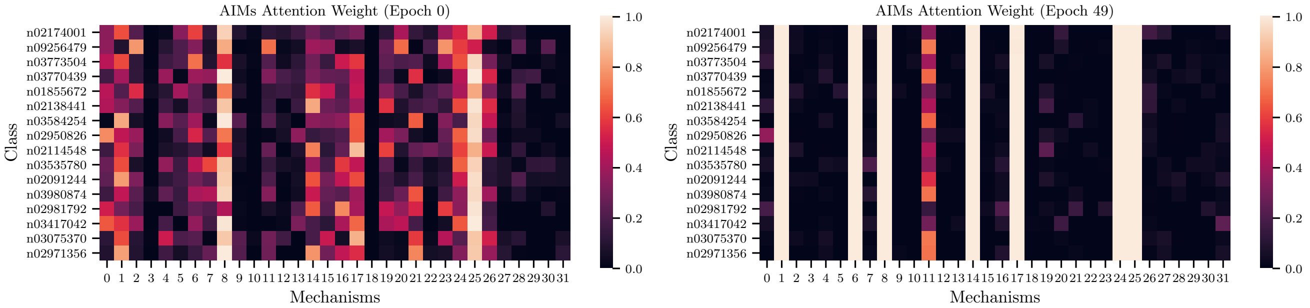
Figure 7: Activation of AIM from few-shot learning. Training on CIFAR-FS with 1-shot using WRN-28-10 as backbone.



(a) First epoch.

(b) Last epoch.

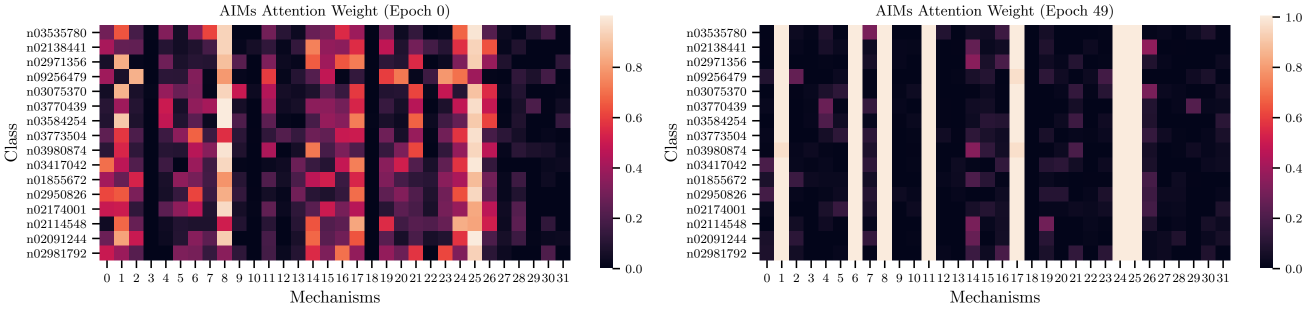
Figure 8: Activation of AIM from few-shot learning. Training on CIFAR-FS with 5-shot using WRN-28-10 as backbone.



(a) First epoch.

(b) Last epoch.

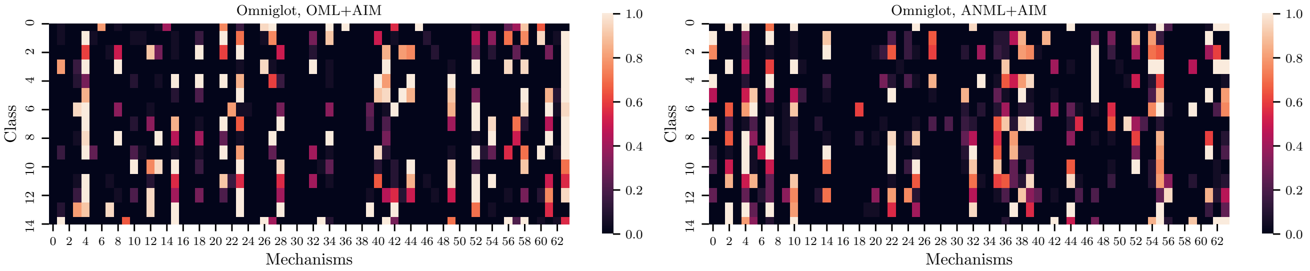
Figure 9: Activation of AIM from few-shot learning. Training on MiniImageNet with 1-shot using WRN-28-10 as backbone.



(a) First epoch.

(b) Last epoch.

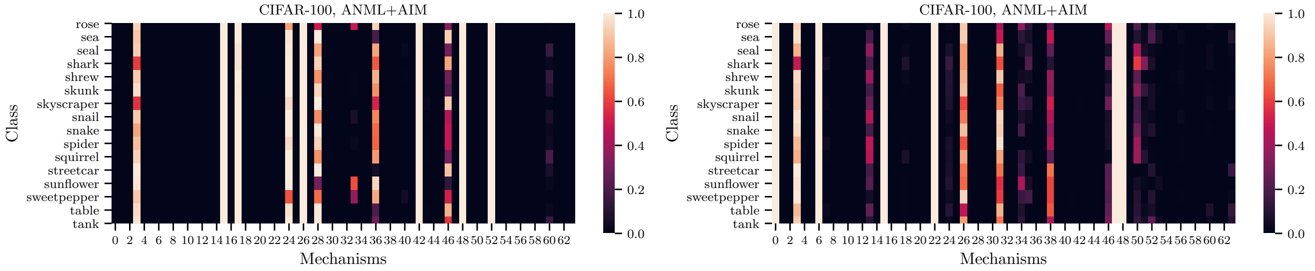
Figure 10: Activation of AIM from few-shot learning. Training on MiniImageNet with 5-shot using WRN-28-10 as backbone.



(a) OML+AIM

(b) ANML+AIM

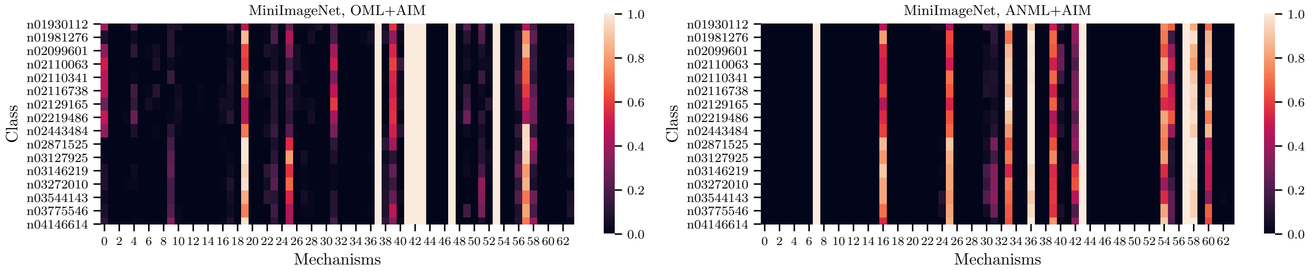
Figure 11: Activation of AIM from continual learning. Subset of classes from Omniglot are shown.



(a) OML+AIM

(b) ANML+AIM

Figure 12: Activation of AIM from continual learning. Subset of classes from CIFAR-100 are shown.



(a) OML+AIM

(b) ANML+AIM

Figure 13: Activation of AIM from continual learning. Subset of classes from MiniImageNet are shown.

Table 2: Results for varying the stochastic sampling count $K + l$. Zero mean-ed plot is found in the main paper. Throughout the experiment, $K = 8$ and l is varied. Average classification accuracies with 95% confidence intervals on the test-set are shown.

Backbone	Stochastic sampling count $K + l$	MiniImageNet, 5-Way 1-shot	MiniImageNet, 5-Way 5-shot	CIFAR-FS, 5-Way 1-shot	CIFAR-FS, 5-Way 5-shot
Conv-4-64	8	$61.90 \pm 0.56\%$	$74.55 \pm 0.38\%$	$70.80 \pm 0.61\%$	$80.50 \pm 0.40\%$
	10	$61.90 \pm 0.57\%$	$74.55 \pm 0.38\%$	$71.09 \pm 0.62\%$	$80.48 \pm 0.40\%$
	12	$62.13 \pm 0.58\%$	$74.66 \pm 0.38\%$	$70.68 \pm 0.62\%$	$80.36 \pm 0.39\%$
	16	$61.74 \pm 0.57\%$	$74.24 \pm 0.38\%$	$69.95 \pm 0.63\%$	$79.96 \pm 0.40\%$
	20	$61.70 \pm 0.57\%$	$74.12 \pm 0.38\%$	$70.01 \pm 0.62\%$	$79.60 \pm 0.41\%$
	24	$60.73 \pm 0.57\%$	$73.58 \pm 0.39\%$	$68.82 \pm 0.63\%$	$79.21 \pm 0.40\%$
	28	$60.34 \pm 0.56\%$	$73.19 \pm 0.38\%$	$67.79 \pm 0.64\%$	$78.90 \pm 0.41\%$
	32	$59.43 \pm 0.56\%$	$72.89 \pm 0.38\%$	$66.90 \pm 0.63\%$	$78.35 \pm 0.41\%$
WRN-28-10	8	$71.03 \pm 0.57\%$	$82.30 \pm 0.33\%$	$79.19 \pm 0.55\%$	$87.04 \pm 0.36\%$
	10	$71.22 \pm 0.57\%$	$82.25 \pm 0.34\%$	$80.20 \pm 0.55\%$	$87.34 \pm 0.36\%$
	12	$71.08 \pm 0.57\%$	$82.25 \pm 0.34\%$	$80.20 \pm 0.55\%$	$87.19 \pm 0.36\%$
	16	$70.57 \pm 0.57\%$	$82.25 \pm 0.34\%$	$79.95 \pm 0.56\%$	$87.10 \pm 0.36\%$
	20	$70.38 \pm 0.57\%$	$81.86 \pm 0.34\%$	$80.17 \pm 0.56\%$	$87.07 \pm 0.36\%$
	24	$70.20 \pm 0.57\%$	$81.65 \pm 0.34\%$	$80.18 \pm 0.56\%$	$86.68 \pm 0.36\%$
	28	$70.40 \pm 0.57\%$	$81.60 \pm 0.35\%$	$80.33 \pm 0.56\%$	$86.63 \pm 0.37\%$
	32	$69.34 \pm 0.55\%$	$81.19 \pm 0.34\%$	$80.13 \pm 0.56\%$	$86.22 \pm 0.38\%$

Table 3: Results for varying the active mechanism count K. Zero mean-ed plot is found in the main paper. Throughout the experiment, K is varied and $l = 0$. Average classification accuracies with 95% confidence intervals on the test-set are shown.

Backbone	Active Mechanism Count	MiniImageNet, 5-Way		CIFAR-FS, 5-Way	
	K	1-shot	5-shot	1-shot	5-shot
Conv-4-64	1	$25.54 \pm 0.26\%$	$62.63 \pm 0.40\%$	$36.38 \pm 0.40\%$	$68.72 \pm 0.44\%$
	2	$62.00 \pm 0.57\%$	$74.62 \pm 0.38\%$	$69.95 \pm 0.61\%$	$80.45 \pm 0.40\%$
	4	$61.93 \pm 0.56\%$	$74.54 \pm 0.38\%$	$70.30 \pm 0.60\%$	$80.37 \pm 0.39\%$
	8	$61.59 \pm 0.56\%$	$74.54 \pm 0.38\%$	$70.15 \pm 0.62\%$	$80.46 \pm 0.39\%$
	12	$61.39 \pm 0.56\%$	$74.62 \pm 0.38\%$	$70.65 \pm 0.61\%$	$80.77 \pm 0.38\%$
	16	$61.60 \pm 0.55\%$	$74.67 \pm 0.38\%$	$70.66 \pm 0.60\%$	$80.68 \pm 0.39\%$
	20	$61.68 \pm 0.56\%$	$74.67 \pm 0.39\%$	$70.00 \pm 0.61\%$	$80.52 \pm 0.39\%$
	24	$61.81 \pm 0.56\%$	$74.67 \pm 0.38\%$	$70.01 \pm 0.61\%$	$80.55 \pm 0.39\%$
	28	$61.81 \pm 0.56\%$	$74.63 \pm 0.39\%$	$70.74 \pm 0.60\%$	$80.46 \pm 0.39\%$
	32	$61.89 \pm 0.56\%$	$74.79 \pm 0.39\%$	$70.56 \pm 0.61\%$	$80.39 \pm 0.39\%$
WRN-28-10	1	$61.01 \pm 0.55\%$	$73.84 \pm 0.37\%$	$72.21 \pm 0.57\%$	$81.03 \pm 0.41\%$
	2	$69.98 \pm 0.56\%$	$81.89 \pm 0.34\%$	$79.88 \pm 0.55\%$	$86.90 \pm 0.36\%$
	4	$70.03 \pm 0.56\%$	$82.10 \pm 0.33\%$	$79.53 \pm 0.55\%$	$86.25 \pm 0.37\%$
	8	$69.76 \pm 0.57\%$	$82.30 \pm 0.33\%$	$79.19 \pm 0.55\%$	$86.26 \pm 0.37\%$
	12	$69.93 \pm 0.56\%$	$82.26 \pm 0.33\%$	$79.53 \pm 0.55\%$	$86.68 \pm 0.37\%$
	16	$70.24 \pm 0.55\%$	$82.15 \pm 0.33\%$	$79.19 \pm 0.55\%$	$86.74 \pm 0.37\%$
	20	$69.87 \pm 0.55\%$	$82.08 \pm 0.34\%$	$79.42 \pm 0.54\%$	$86.83 \pm 0.37\%$
	24	$69.80 \pm 0.56\%$	$82.52 \pm 0.33\%$	$79.89 \pm 0.54\%$	$87.08 \pm 0.37\%$
	28	$69.58 \pm 0.55\%$	$82.30 \pm 0.33\%$	$79.66 \pm 0.54\%$	$86.88 \pm 0.37\%$
	32	$70.09 \pm 0.55\%$	$82.29 \pm 0.33\%$	$80.01 \pm 0.53\%$	$87.02 \pm 0.37\%$

Table 4: Average meta-testing test accuracy of continual learning on various datasets. During training, trajectory of samples are introduced, i.e. meta-test train images are fed to the model sequentially without the usage of rehearsal memory and evaluation using meta-testing test set is performed at the end. Relative increment (decrease) in accuracy through the introduction of AIM is shown in green (red), e.g. when 10 classes in a trajectory is introduced for Omniglot, a relative increment in accuracy of +3.45 over OML is attained when AIM is inserted, shown as $\text{OML} + \text{AIM}$.

Method	Number of classes								
	10	50	75	100	200	300	400	500	600
	Dataset: Omniglot								
Baseline	10.00	2.00	1.33	1.00	0.50	0.33	0.25	0.20	0.17
OML	94.34	80.69	76.30	73.61	62.96	56.61	51.27	47.34	44.56
ANML	82.60	84.92	83.60	81.92	76.85	72.83	69.12	65.74	63.41
OML+AIM	97.70 (+3.45)	98.60 (+1.28)	98.55 (+5.50)	97.75 (+8.41)	96.28 (+11.76)	94.08 (+17.58)	91.09 (+24.26)	85.51 (+22.93)	80.37 (+20.70)
ANML+AIM	94.25 (+11.65)	97.32 (+12.40)	93.05 (+9.45)	89.34 (+7.42)	84.52 (+7.67)	76.50 (+3.67)	66.83 (-2.29)	62.58 (-3.17)	59.68 (-3.74)
	2	4	6	8	10	15	20	25	30
	Dataset: CIFAR-100								
Baseline	50.00	25.00	16.67	12.50	10.00	6.67	5.00	4.00	3.33
OML	81.57	61.46	58.24	52.39	53.65	39.42	33.58	28.53	26.78
ANML	86.86	73.03	63.66	56.60	50.01	45.02	40.06	36.29	34.15
OML+AIM	85.65 (+4.08)	79.46 (+18.00)	68.03 (+9.79)	60.44 (+8.04)	53.39 (-0.26)	46.16 (+6.74)	42.02 (+8.43)	36.17 (+7.64)	33.59 (+6.81)
ANML+AIM	84.10 (-2.76)	70.10 (-2.93)	60.90 (-2.76)	58.35 (+1.76)	55.88 (+5.87)	46.72 (+1.70)	39.84 (-0.23)	36.47 (+0.18)	28.81 (-5.34)
	2	4	6	8	10	12	15	18	20
	Dataset: MiniImageNet								
Baseline	50.00	25.00	16.67	12.50	10.00	8.33	6.67	5.56	5.00
OML	63.00	42.17	30.80	28.40	23.31	19.26	16.82	13.97	11.54
ANML	73.25	48.42	33.42	28.44	23.43	19.49	16.66	13.81	12.73
OML+AIM	75.75 (+12.75)	56.13 (+13.96)	42.92 (+12.12)	38.94 (+10.53)	33.52 (+10.20)	28.57 (+9.30)	27.81 (+10.99)	23.85 (+9.89)	23.03 (+11.48)
ANML+AIM	85.25 (+12.00)	64.13 (+15.71)	52.97 (+19.56)	52.79 (+24.35)	44.90 (+21.47)	39.28 (+19.79)	36.67 (+20.01)	33.01 (+19.20)	32.13 (+19.40)

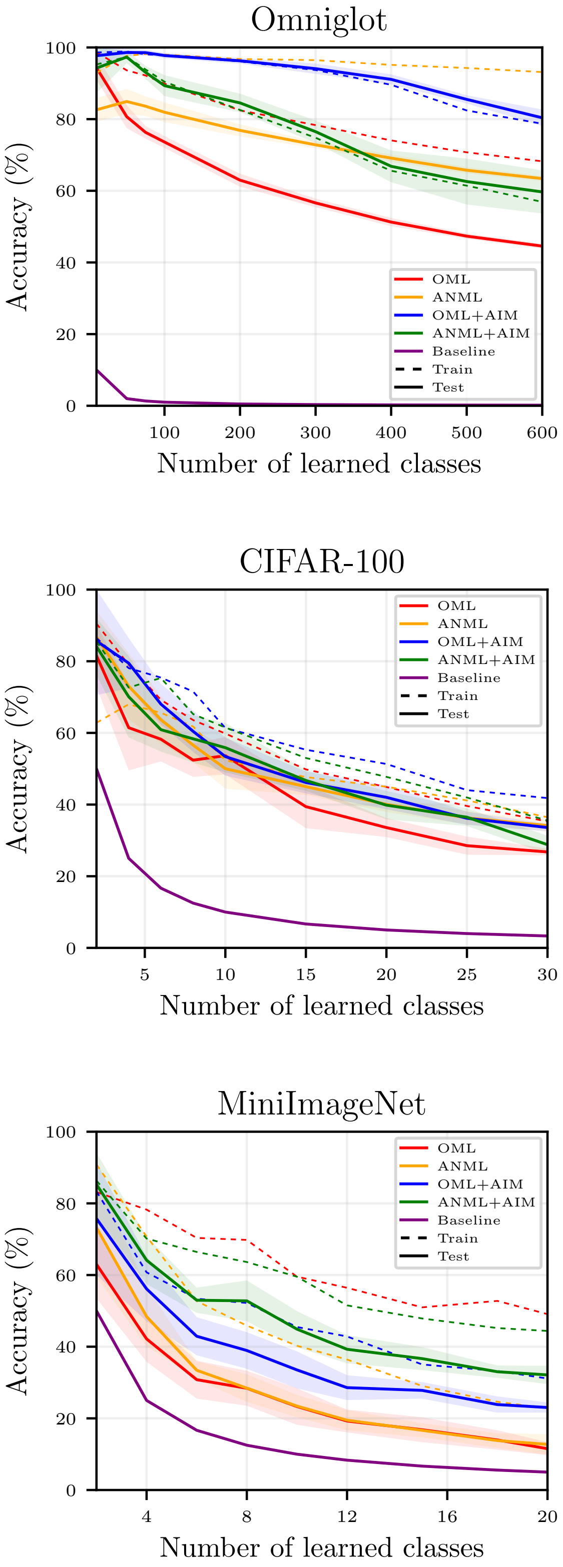


Figure 14: Evaluation of continual learning methods using dataset of various scales. Meta-test testing (training) trajectories are shown in solid (dashed) lines. All curves are averaged over 10 runs with standard deviation shown.

References

[1] Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O Stanley, Jeff Clune, and Nick Cheney. Learning to continually learn. *arXiv preprint arXiv:2002.09571*, 2020.

[2] Khurram Javed and Martha White. Meta-learning representations for continual learning. In *Advances in Neural Information Processing Systems*, pages 1820–1830, 2019.

[3] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[4] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[5] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[6] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.