Few-Shot Unsupervised Continual Learning through Meta-Examples

Alessia Bertugli *
University of Trento
alessia.bertugli@unitn.it

Stefano Vincenzi *
University of Modena and Reggio Emilia
stefano.vincenzi@unimore.it

Simone Calderara
University of Modena and Reggio Emilia
simone.calderara@unimore.it

Andrea Passerini
University of Trento
andrea.passerini@unitn.it

Abstract

In real-world applications, data do not reflect the ones commonly used for neural networks training, since they are usually few, unlabeled and can be available as a stream. Hence many existing deep learning solutions suffer from a limited range of applications, in particular in the case of online streaming data that evolve over time. To narrow this gap, in this work we introduce a novel and complex setting involving unsupervised meta-continual learning with unbalanced tasks. These tasks are built through a clustering procedure applied to a fitted embedding space. We exploit a meta-learning scheme that simultaneously alleviates catastrophic forgetting and favors the generalization to new tasks. Moreover, to encourage feature reuse during the meta-optimization, we exploit a single inner loop taking advantage of an aggregated representation achieved through the use of a self-attention mechanism. Experimental results on few-shot learning benchmarks show competitive performance even compared to the supervised case. Additionally, we empirically observe that in an unsupervised scenario, the small tasks and the variability in the clusters pooling play a crucial role in the generalization capability of the network. Further, on complex datasets, the exploitation of more clusters than the true number of classes leads to higher results, even compared to the ones obtained with full supervision, suggesting that a predefined partitioning into classes can miss relevant structural information. The code is available at https://github.com/alessiabertugli/FUSION-ME

1 Introduction

Continual learning has been widely studied in the last few years to solve the catastrophic forgetting problem that affects neural networks. Several methods [1–6] have been proposed to solve this problem involving a replay buffer, network expansion, selectively regularizing and distillation. Some works [7–12] take advantage of the meta-learning abilities of generalization on different tasks and rapid learning on new ones to deal with continual learning problems. Few works on unsupervised meta-learning [13–15] and unsupervised continual learning [16] have been recently proposed, but the first ones deal with independent and identically distributed data, while the second one assumes the availability of a huge dataset. Moreover, the majority of continual learning and meta-learning works assume that data are perfectly balanced or equally distributed among classes. We propose a new, more realistic setting, namely FUSION (Few-shot UnSupervIsed cONtinual learning), dealing

*Equal contribution.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), 4th Workshop on Meta-Learning.

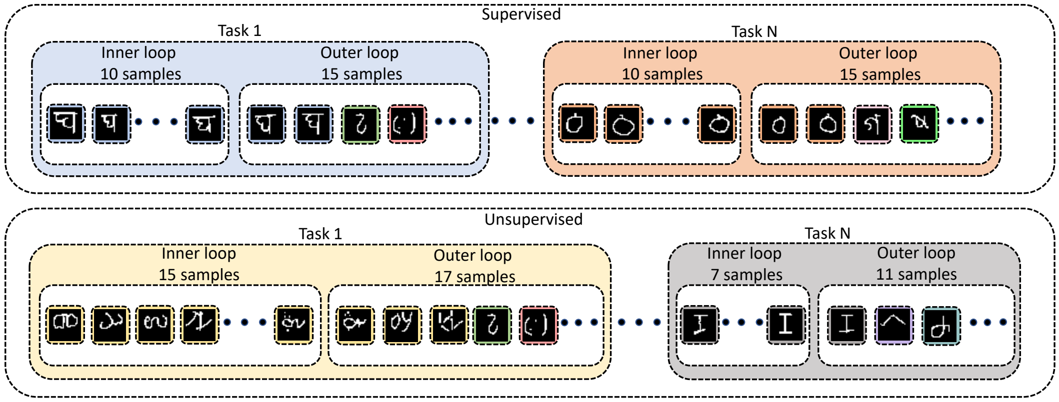

Figure 1: Supervised vs unsupervised tasks flow. In the supervised version, tasks are perfectly balanced and contain a fixed number of elements for inner loop (10 samples) and outer loop (15 samples, 5 from the current cluster and 10 randomly sampled from other clusters). In the unsupervised model, tasks are unbalanced and contain $2/3$ of cluster data for the inner loop and $1/3$ for the outer loop in addition to a fixed number of random samples.

with unlabeled and unbalanced tasks in a meta-continual learning fashion and a novel method MEML (Meta-Example Meta-Learning), that is able to face this complex scenario. In the task construction phase, rather than directly exploiting high dimensional raw data, an embedding learning network is used to learn a fitted embedding space to facilitate clustering. Precisely, the k-means algorithm is applied to build tasks composed of unbalanced data, each one with the assigned pseudo-label. Our meta-learning model relies on a double-loop procedure that receives data in an online incremental learning fashion. The classification layers are learned through a single inner loop update, adopting an attentive mechanism that extracts the most relevant features -meta example- of the current unbalanced task; this considerably reduces the training time and memory usage. In the outer loop, to avoid forgetting and improve generalization, we train all model layers exploiting, as input, an ensemble between data of the same class of the stream and data randomly sampled from the overall trajectory (see Figure 1). We test our model and setup on Omniglot [17] and Mini-ImageNet [18], achieving favorable results compared to baseline approaches. We show the importance of performing the single inner loop update on the meta-example with respect to both updating over a random sample and updating over multiple samples of the same task. We empirically verify that with tasks generated in an unsupervised manner, the need for balanced data is not crucial compared to the variability in the data and the exploitation of small clusters.

2 Few-Shot Unsupervised Continual Learning

We propose a novel setting that deals with unsupervised meta-continual learning and study the effect of the unbalanced tasks derived by an unconstrained clustering approach.

As done in [13], the task construction phase exploits the k-means algorithm over suitable embeddings obtained through an unsupervised pre-training. This simple but effective method assigns the same pseudo-label to all data points belonging to the same cluster. The first step employs two different models: Deep Cluster [19] for Mini-ImageNet, and ACAI [20] for Omniglot. Both these methods consist of unsupervised training and produce an embedding vector set $Z = Z_0, Z_1, ..., Z_N$, where N is the number of data points in the training set. ACAI is based on an autoencoder while Deep Cluster on a deep feature extraction phase followed by k-means clustering. They outline some of the most promising approaches to deal with unlabeled, high dimensional data to obtain and discover meaningful latent features. Applying k-means over these embeddings leads to unbalanced clusters, which determine unbalanced tasks. This is in contrast with typical meta-learning and continual learning problems, where data are perfectly balanced. To recover a balanced setting, in [13], the authors set a threshold on the cluster dimension, discarding extra samples and smaller clusters. A recent alternative [21] forces the network to balance clusters, but this imposes a partitioning of the embedding space that contrasts with the extracted features. We believe that these approaches are sub-optimal as they alter the data distribution. In an unsupervised setting, where data points are grouped based on the similarity of their features, variability is an essential factor. By keeping also the small tasks, our model generalizes better and reaches higher accuracy at meta-test time. In a data imbalanced setting, the obtained meta-representation is more influenced by large clusters. Since the latter may contain more generic features than the smaller ones, the model is able to generalize better by mostly learning from them. Despite this, the small clusters may contain important information for different classes presented during evaluation. To corroborate this claim, we investigate balancing techniques, both at data-level, such as data augmentation and at model-level, such as balancing parameters into the loss term. Once the tasks are built, they are sampled one at a time for the meta-continual train. The training process happens in a class incremental way, where a task correspond to one cluster. During this phase, the network have to learn a good representation that will be able to generalize to unseen tasks, while avoiding forgetting. The meta-train phase relies completely on pseudo-labels. At meta-continual test time, novel and unseen tasks are presented to the network. The representation learned during meta-train remains fixed, while only the prediction layers are fine-tuned, testing on few data of novel classes.

3 Meta-Example Meta-Learning

Our network is composed of a Feature Extraction Network (FEN) and a CLassification Network (CLN), both updated during the meta-training phase through a meta-learning procedure based on the construction of a meta-example. MAML and all its variants rely on a two-loop mechanism that allows learning new tasks from a few steps of gradient descent. Recent investigations on this algorithm explain that the real reason for MAML's success resides in feature reuse instead of rapid learning [22], proving that learning meaningful representations is a crucial factor. Based on this assumption, we focus on the generalization ability of the feature extraction layers. We remove the need for several inner loops, maintaining a single inner loop update through an attentive procedure that considerably reduces the training time and computational resources needed for training the model and increases the global performance. At each time-step, as pointed out in Figure 1, a task $\mathcal{T}_i = (\mathcal{S}_{cluster}, \mathcal{S}_{query})$ is randomly sampled from tasks distribution $p(\mathcal{T})$. $\mathcal{S}_{cluster}$ contains elements of the same cluster and is defined as $\mathcal{S}_{cluster} = \{(X_k, Y_k)\}_{k=0}^{K}$, with $Y_0 = ... = Y_K$, where $Y_0 = ... = Y_k$ is the cluster pseudo-label. Instead, $\mathcal{S}_{query}$ contains a variable number of elements belonging to the current cluster and a fixed number of elements randomly sampled from all other clusters, and is defined as $\mathcal{S}_{query} = \{(X_q, Y_q)\}_{q=0}^{Q}$. All the elements belonging to $\mathcal{S}_{cluster}$ are processed by the frozen FEN, parameterized by θ, computing the feature vectors $R_0, R_1, ..., R_K$ in parallel for all task elements as $R_{0:K} = f_\theta(X_{0:K})$. The obtained embeddings are refined with an attention function parameterized by ρ computes the attention coefficients α from the features vectors:

$$\alpha_{0:K} = \textit{Softmax}[f_\rho(R_{0:K})]. \tag{1}$$

Then, the final aggregated representation learning vector ME, called *meta-example*, captures the most salient features, and is computed as follows:

$$ME = \sum_{k=0}^{K}[R_k \cdot \alpha_k]. \tag{2}$$

The single inner loop is performed on this meta-example, which adds up the weighted-features contribution of each element of the current cluster. Then, the cross-entropy loss ℓ between the predicted label and the pseudo-label is computed and both the classification network parameters W and the attention parameters ρ ($\psi = \{W_i, \rho\}$) are updated as follows:

$$\psi \leftarrow \psi - \alpha \nabla_\psi \ell_i(f_\psi(\textit{ME}), Y_0), \tag{3}$$

where α is the inner loop learning rate. Finally, to update the whole network parameters $\phi = \{\theta, W_i, \rho\}$, and to ensure generalization across tasks, the outer loop loss is computed S_{query}. The outer loop parameters are thus updated as follows:

$$\phi \leftarrow \phi - \beta \nabla_\phi \ell_i(f_\phi(X_{0:Q}), Y_{0:Q}), \tag{4}$$

where β is the outer loop learning rate.

Table 1: Meta-test results on Omniglot.

Algorithm/Classes	10	50	75	100	150	200
Oracle OML [24]	88.4	74.0	69.8	57.4	51.6	47.9
Oracle MEML	**92.3**	**78.2**	**72.7**	**60.9**	**51.8**	**51.4**
OML balanced 500	67.8	27.6	29.4	24.5	18.7	15.8
OML balancing param	59.4	27.2	24.3	18.4	15.5	11.8
OML augmentation	72.2	35.1	32.5	27.5	21.8	17.3
OML	74.6	32.5	30.6	25.8	19.9	16.1
OML single update	67.5	32.0	30.2	24.3	18.4	15.3
MEML mean	60.6	31.2	25.8	21.3	17.0	13.7
MEML (Ours)	**84.6**	37.3	37.5	30.9	25.4	20.7
MEML RS (Ours)	81.6	**56.4**	**54.0**	**44.6**	**34.1**	**27.4**

Table 2: Meta-test results on Mini-ImageNet.

Algorithm/Classes	2	4	6	8	10
Oracle OML [24]	50.0	31.9	27.0	16.7	13.9
Oracle MEML	**66.0**	**33.0**	**28.0**	**29.1**	**21.1**
OML	49.3	41.0	19.2	18.2	12.0
MEML 64	58.0	41.2	**40.0**	27.3	18.8
MEML 128	56.0	41.7	21.6	16.2	11.4
MEML 256	**70.0**	**48.4**	36.0	**34.0**	**21.6**
MEML 512	54.7	36.4	26.2	14.1	21.4
MEML 64 RS	54.0	39.0	31.2	27.3	16.4

4 Experiments

4.1 Balanced *vs.* Unbalanced Tasks

To justify the use of unbalanced tasks and show that allowing unbalanced clusters is more beneficial than enforcing fewer balanced ones, we present in Table 1 some comparisons achieved on the Omniglot dataset. First of all, we introduce a baseline in which the number of clusters is set to the true number of classes, removing from the task distribution the ones containing less than N elements and sampling N elements from the bigger ones (OML). We thus obtain a perfectly balanced training set at the cost of less variety within the clusters; however, this leads to poor performance as small clusters are never represented. Setting a smaller number of clusters than the number of true classes gives the same results (OML balanced 500). This test shows that cluster variety is more important than balancing for generalization. To verify if maintaining variety and balancing data can lead to better performance, we try two balancing strategies: augmentation, at data-level, and balancing parameter, at model-level. For the first one, we keep all clusters, sampling N elements from the bigger and using data augmentation for the smaller to reach N elements (OML augmentation). At model-level, we multiply the loss term by a balancing parameter, to weight the update for each task based on cluster length (OML balancing param). These two tests, especially the latter one, result in lower performance with respect to the unbalanced setting, suggesting that the only thing that matters is cluster variety. We can also presume that bigger clusters may contain the most meaningful and general features, so unbalancing does not negatively affect the training of our unsupervised meta-continual learning model. Finally, as we want to confirm that this intuition is valid in a more general unsupervised meta-learning model, we perform the balanced/unbalanced experiments also on CACTUs [13]. The results are shown in Table 3 (Top) and attest that the model trained on unbalance data outperforms the balanced one, further proving the importance of task variance to better generalize to new classes at meta-test time. We report the results training the algorithms on 20 ways for generality purposes and 5 shots and 15 shots, in order to have enough data points per class to create the imbalance.

4.2 Meta-example Single Update *vs.* Multiple Updates

In Table 1, we show that the model trained with the attention-based method consistently outperforms all the other baselines. The single update gives the worst performance, but not really far from the multiple updates one, confirming the idea that the strength of generalization relies on the feature reuse. Also, the mean test has performance comparable with the multiple and single update ones, proving the effectiveness of the attention mechanism to determine a suitable and general embedding vector for the CLN. Training time and resources consumption is considerably reduced with our model based on a single update on the generated meta-example (see Supplementary Material). We also test our technique in a standard meta-learning setting. We compare our meta-example based algorithm MEML to MAML [23] on Omniglot dataset in Table 3 (bottom), consistently outperforming it. We report the results training on 20 ways and 1 and 5 shots. In particular, the 5 shots test highlights the effectiveness of our aggregation method.

4.3 MEML *vs.* Oracles

To see how the performance of MEML is far from those achievable with the real labels, we also report for all datasets the accuracy reached in a supervised setting (*oracles*) on both Omniglot (see Table 1) and Mini-ImageNet (see Table 2). We define Oracle OML the supervised model present in [24],

Table 3: Balanced *vs.* unbalanced CACTUs-MAML (top) and MEML, with our meta-example update, compared to basic MAML (bottom) on Omniglot dataset.

Algorithm/Ways, Shots	5,1	5,5	20,1	20,5
CACTUs-MAML Balanced 20,5	60.50	84.00	40.50	67.62
CACTUs-MAML Unbalanced 20,5	**62.50**	**85.50**	**42.62**	**71.87**
CACTUs-MAML Balanced 20,15	67.00	86.00	32.50	64.62
CACTUs-MAML Unbalanced 20,15	**72.00**	**89.00**	**40.00**	**66.25**
MAML 20,1	78.00	97.50	77.62	92.87
MEML 20,1 (Ours)	**97.50**	**99.97**	**88.13**	**99.37**
MAML 20,5	88.00	99.50	74.62	92.75
MEML 20,5 (Ours)	**95.00**	**99.95**	**85.63**	**96.25**

and Oracle MEML the supervised model updated with our meta-example strategy. Oracle MEML outperforms Oracle OML on Omniglot and Mini-ImageNet, suggesting that the meta-examples strategy is beneficial even in a fully supervised case. MEML reaches higher performance compared to the other OML baselines but lower on Omniglot compared to the Oracle OML. On Mini-ImageNet, our model trained with 256 clusters outperforms both oracles. To further improve the performance avoiding forgetting at meta-test time, we add a rehearsal strategy based on reservoir sampling on the CLN (MEML RS). This generally results in superior performance on Omniglot. On Mini-ImageNet the performance with and without rehearsal are similar, due to the low number of test classes in the dataset that alleviates catastrophic forgetting.

4.4 Number of Clusters

In an unsupervised setting, the number of original classes could be unknown. Consequently, it is important to assess the performance of our model by varying the number of clusters at meta-train time. With a coarse-grain clustering, a low number of clusters are formed and distant embeddings can be assigned to the same pseudo-label, grouping classes that can be rather different. On the other hand, with a fine-grain clustering, a high number of clusters with low variance are generated. Both cases lead to poor performance at meta-test time. We test on Omniglot (see Table 1), setting the number of clusters to: the true number of classes (OML); a lower number of clusters (OML balanced 500), resulting in more than 20 samples each. Since the Omniglot dataset comprehends 20 samples per class, in the first case it results in unbalanced tasks, while in the second we sample 20 elements from the bigger clusters. The performance of the 1100 clusters test is consistently higher than that obtained with the 500 clusters test, confirming that variability is more important than balancing. On Mini-ImageNet, we test our method in Table 2 with 64, 128, 256, and 512 clusters (MEML number of clusters). Since Mini-ImageNet contains 600 examples per class, after clustering we sample examples between 10 and 30, proportionally to the cluster dimension. We obtain the best results with 256 clusters and the meta-example approach, outperforming not only the other unsupervised experiment but also the supervised oracle. We observe that using 512 clusters degrades performance with respect to the 256 case, suggesting that tasks constructed over an embedding space with too specific features fail to generalize. Using a lower number of clusters, such as 64 or 128, also achieves worse performance. This time, the embedding space is likely aggregating distant features, leading to a complex meta-continual training, whose pseudo-classes are not clearly separated.

5 Related Work

5.1 Supervised and Unsupervised Continual Learning

Continual learning is one of the most challenging problems arising from neural networks that are heavily affected by catastrophic forgetting. The proposed methods can be divided into three main categories. *Architectural strategies*, are based on specific architectures designed to mitigate catastrophic forgetting [25, 26]. *Regularization strategies* are based on putting regularization terms into the loss function, promoting selective consolidation of important past weights [1, 5]. Finally *rehearsal strategies* focus on retaining part of past information and periodically replaying it to the model to strengthen connections for memories, involving meta-learning [4, 27], combination of rehearsal and regularization strategies [2, 3], knowledge distillation [28–31], generative replay [32–34] and channel gating [35]. Only a few recent works have studied the problem of unlabeled data, which mainly involves representation learning. CURL [16] proposes an unsupervised model built on a representation learning network. This latter learn a mixture of Gaussian encoding task variations, then integrates a generative memory replay buffer as a strategy to overcome forgetting.

5.2 Supervised and Unsupervised Meta-Learning

Meta-learning, or learning to learn, aims to improve the neural networks ability to rapidly learn new tasks with few training samples. The majority of meta-learning approaches proposed in literature are based on Model-Agnostic Meta-Learning (MAML) [23, 36–38]. Through the learning of a profitable parameter initialization with a double loop procedure, MAML limits the number of stochastic gradient descent steps required to learn new tasks, speeding up the adaptation process performed at meta-test time. Although MAML is suitable for many learning settings, few works investigate the unsupervised meta-learning problem. CACTUs [13] proposes a new unsupervised meta-learning method relying on clustering feature embeddings through the k-means algorithm and then builds tasks upon the predicted classes. UMTRA [14] is a further method of unsupervised meta-learning based on a random sampling and data augmentation strategy to build meta-learning tasks, achieving comparable results with respect to CACTUs. UFLST [15] proposes an unsupervised few-shot learning method based on self-supervised training, alternating between progressive clustering and update of the representations.

5.3 Meta-Learning for Continual Learning

Meta-learning has extensively been merged with continual learning for different purposes. We can highlight the existence of two strands of literature [39]: *meta-continual learning* with the aim of incremental task learning and *continual-meta learning* with the aim of fast remembering. Continual-meta learning approaches mainly focus on making meta-learning algorithms online, with the aim to rapidly remember meta-test tasks [40, 7, 11]. More relevant to our work are meta-continual learning algorithms [8, 24, 41, 12, 10, 9, 42], which use meta-learning rules to “learn how not to forget". OML [24] and its variant ANML [41] favor sparse representations by employing a trajectory-input update in the inner loop and a random-input update in the outer one. The algorithm jointly trains a representation learning network (RLN) and a prediction learning network (PLN) during the meta-training phase. Then, at meta-test time, the RLN layers are frozen and only the PLN is updated. ANML replaces the RLN network with a neuro-modulatory network that acts as a gating mechanism on the PLN activations following the idea of conditional computation.

6 Discussion

In this work, we tackle a novel problem concerning few-shot unsupervised continual learning. We propose a simple but effective model based on the construction of unbalanced tasks and meta-examples. Our model is motivated by the power of *representation learning*, which relies on few and raw data with no need for human supervision. With an unconstrained clustering approach, we find that no balancing technique is necessary for an unsupervised scenario that needs to generalize to new tasks. In fact, the most robust and general features are gained though task variety; even if favoring larger clusters leads to more general features, smaller ones should not be discarded as they can be representative of less common tasks. This means that there is no need for complex representation learning algorithm that try to balance clusters elements. A future achievement is to deeply investigate this insight by observing the variability of the embeddings in the feature space. A further improvement consists in the introduction of FiLM layers [43] into the FEN to change data representation at meta-test time and the introduction of an OoD detector to face with Out-of-Distribution tasks. The performances of our model with meta-examples suggest that a single inner update can increase performances if the most relevant features for the task are selected. To this end, a more refined technique, relying on hierarchical aggregation techniques, can be considered.

References

[1] Kirkpatrick, J.N., Pascanu, R., Rabinowitz, N.C., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences of the United States of America **114 13** (2016) 3521–3526

[2] Lopez-Paz, D., Ranzato, M.A.: Gradient episodic memory for continual learning. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., eds.: Advances in Neural Information Processing Systems 30. Curran Associates, Inc. (2017) 6467–6476

[3] Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with A-GEM. In: ICLR. (2019)

[4] Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., , Tesauro, G.: Learning to learn without forgetting by maximizing transfer and minimizing interference. In: International Conference on Learning Representations. (2019)

[5] Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17, JMLR.org (2017) 3987–3995

[6] Rebuffi, S.A., Kolesnikov, A.I., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 5533–5542

[7] Jerfel, G., Grant, E., Griffiths, T., Heller, K.A.: Reconciling meta-learning and continual learning with online mixtures of tasks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., eds.: Advances in Neural Information Processing Systems 32. Curran Associates, Inc. (2019) 9119–9130

[8] Vuorio, R., Cho, D.Y., Kim, D., Kim, J.: Meta continual learning. ArXiv **abs/1806.06928** (2018)

[9] Rajasegaran, J., Khan, S., Hayat, M., Khan, F.S., Shah, M.: itaml : An incremental task-agnostic meta-learning approach. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

[10] Liu, Q., Majumder, O., Ravichandran, A., Bhotika, R., Soatto, S.: Incremental learning for metric-based meta-learners. ArXiv **abs/2002.04162** (2020)

[11] Harrison, J., Sharma, A., Finn, C., Pavone, M.: Continuous meta-learning without tasks (2020)

[12] Yao, H., Wei, Y., Huang, J., Li, Z.: Hierarchically structured meta-learning. In: Proceedings of the 36rd International Conference on International Conference on Machine Learning. ICML'19 (2019)

[13] Hsu, K., Levine, S., Finn, C.: Unsupervised learning via meta-learning. In: International Conference on Learning Representations. (2019)

[14] Khodadadeh, S., Bölöni, L., Shah, M.: Unsupervised meta-learning for few-shot image and video classification. ArXiv **abs/1811.11819** (2018)

[15] Ji, Z., Zou, X., Huang, T., Wu, S.: Unsupervised few-shot learning via self-supervised training. ArXiv **abs/1912.12178** (2019)

[16] Rao, D., Visin, F., Rusu, A.A., Teh, Y.W., Pascanu, R., Hadsell, R.: Continual unsupervised representation learning (2019)

[17] Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266) (2015) 1332–1338

[18] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09. (2009)

[19] Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: European Conference on Computer Vision. (2018)

[20] Berthelot*, D., Raffel*, C., Roy, A., Goodfellow, I.: Understanding and improving interpolation in autoencoders via an adversarial regularizer. In: International Conference on Learning Representations. (2019)

[21] Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. In: International Conference on Learning Representations. (2020)

[22] Raghu, A., Raghu, M., Bengio, S., Vinyals, O.: Rapid learning or feature reuse? towards understanding the effectiveness of maml. In: ICLR. (2020)

[23] Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In Precup, D., Teh, Y.W., eds.: Proceedings of the 34th International Conference on Machine Learning. Volume 70 of Proceedings of Machine Learning Research., International Convention Centre, Sydney, Australia, PMLR (06–11 Aug 2017) 1126–1135

[24] Javed, K., White, M.: Meta-learning representations for continual learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., eds.: Advances in Neural Information Processing Systems 32. Curran Associates, Inc. (2019) 1818–1828

[25] Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. ArXiv **abs/1606.04671** (2016)

[26] Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y.W., Pascanu, R., Hadsell, R.: Progress & compress: A scalable framework for continual learning. In: ICML. (2018)

[27] Spigler, G.: Meta-learnt priors slow down catastrophic forgetting in neural networks. ArXiv **abs/1909.04170** (2019)

[28] Furlanello, T., Lipton, Z.C., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: ICML. (2018)

[29] Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2018) 2935–2947

[30] Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Lifelong learning via progressive distillation and retrospection. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018)

[31] Lee, K., Lee, K., Shin, J., Lee, H.: Overcoming catastrophic forgetting with unlabeled data in the wild. In: ICCV. (2019)

[32] Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, Red Hook, NY, USA, Curran Associates Inc. (2017) 2994–3003

[33] Silver, D.L., Mahfuz, S.: Generating accurate pseudo examples for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (June 2020)

[34] Liu, X., Wu, C., Menta, M., Herranz, L., Raducanu, B., Bagdanov, A.D., Jui, S., van de Weijer, J.: Generative feature replay for class-incremental learning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020) 915–924

[35] Abati, D., Tomczak, J., Blankevoort, T., Calderara, S., Cucchiara, R., Bejnordi, B.E.: Conditional channel gated networks for task-aware continual learning. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

[36] Rajeswaran, A., Finn, C., Kakade, S.M., Levine, S.: Meta-learning with implicit gradients. In: NeurIPS. (2019)

[37] Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. ArXiv **abs/1803.02999** (2018)

[38] Rusu, A.A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., Hadsell, R.: Meta-learning with latent embedding optimization. In: International Conference on Learning Representations. (2019)

[39] Caccia, M., Rodriguez, P., Ostapenko, O., Normandin, F., Lin, M., Caccia, L., Laradji, I., Rish, I., Lacoste, A., Vazquez, D., et al.: Online fast adaptation and knowledge accumulation: a new approach to continual learning. arXiv preprint arXiv:2003.05856 (2020)

[40] Finn, C., Rajeswaran, A., Kakade, S., Levine, S.: Online meta-learning. In Chaudhuri, K., Salakhutdinov, R., eds.: Proceedings of the 36th International Conference on Machine Learning. Volume 97 of Proceedings of Machine Learning Research., Long Beach, California, USA, PMLR (09–15 Jun 2019) 1920–1930

[41] Beaulieu, S., Frati, L., Miconi, T., Lehman, J., Stanley, K.O., Clune, J., Cheney, N.: Learning to continually learn. 24th European Conference on Artificial Intelligence (ECAI) (2020)

[42] Tao, X., Hong, X., Chang, X., Dong, S., Wei, X., Gong, Y.: Few-shot class-incremental learning. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

[43] Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. arXiv preprint arXiv:1709.07871 (2017)