Curriculum Meta-Learning for Few-shot Classification

Emmanouil Stergiadis
Booking.com
emmanouil.stergiadis@booking.com

Priyanka Agrawal
Booking.com
pagrawal.ml@gmail.com

Oliver Squire
Booking.com
oliver.squire@booking.com

Abstract

We propose an adaptation of the curriculum training framework, applicable to state-of-the-art meta learning techniques for few-shot classification. Curriculum-based training popularly attempts to mimic human learning by progressively increasing the training complexity to enable incremental concept learning. As the meta-learner's goal is learning how to learn from as few samples as possible, the exact number of those samples (i.e. the size of the support set) arises as a natural proxy of a given task's difficulty. We define a simple yet novel curriculum schedule that begins with a larger support size and progressively reduces it throughout training to eventually match the desired shot-size of the test setup. This proposed method boosts the learning efficiency as well as the generalization capability. Our experiments with the MAML algorithm on two few-shot image classification tasks show significant gains with the curriculum training framework. Ablation studies corroborate the independence of our proposed method from the model architecture as well as the meta-learning hyperparameters.

1 Introduction

Meta-learning works Thrun (1998) introduced a prominent set of techniques that aim to learn transferable knowledge from seen tasks and thus can generalize well with few training examples for a range of unseen tasks and environments. These works can be broadly categorized into metric-based methods Vinyals et al. (2016); Snell et al. (2017); Sung et al. (2018), model-based methods Santoro et al. (2016); Munkhdalai and Yu (2017); Mishra et al. (2018) and gradient-based methods Ravi and Larochelle (2017); Nichol et al. (2018). Most of these works train a model with a set of past tasks drawn from the same distribution, and thus similar in nature, to the new tasks. However, training a model on extremely data-scarce tasks that mirror the unseen ones limits its ability to learn a good base learner, even when more annotated samples were in fact available for the tasks using in meta-training. Some works Nichol et al. (2018); Lee et al. (2019) alleviate this by using a larger support size for meta-training tasks, but this approach does not work consistently for all techniques (see Section 5.3) due to the discrepancy it introduces between meta-training and meta-testing. This presents a trade-off between training a better base learner versus generalization capabilities.

In this work, we propose a curriculum learning framework with a dynamic training schedule. This allows one to train a better base learner using a larger support set during initial epochs as well as adhere to the shot-size constraints towards the later epochs. While the use of curriculum learning for training machine learning models Bengio et al. (2009) is not new, the traditional curriculum learning methods use individual sample complexity and sort the samples respectively. To the best of our knowledge, this work is the first to propose a curriculum design for episodes/tasks in meta-learning approaches that uses support-size as a proxy for the difficulty of a meta-task. We conduct

5th Workshop on Meta-Learning at NeurIPS 2021, held virtually.

detailed experimental analysis (see Sections 4 - 5) with two few-shot image classification datasets to demostrate the efficacy of our proposed curriculum learning method.

2 Related Work

Meta-learning works for few-shot classification problems aim to model the knowledge of known tasks in a form that is easily transferable to new, unseen classes with very small set of examples. These works can be broadly organized into three main categories:
(i) metric-based or 'learning to compare' methods which rely on learning a model that can determine similarity between two examples. This can be used to match an unseen instance with the labeled instances Koch et al. (2015); Vinyals et al. (2016); Snell et al. (2017); Sung et al. (2018);
(ii) model-based or 'good model initialization' methods attempt to learn a model that can be fine-tuned for novel classes with a limited number of examples and few parameter updates Santoro et al. (2016); Munkhdalai and Yu (2017); Mishra et al. (2018);
(iii) gradient-based or 'learning an optimizer' methods aim at adjusting that optimization algorithm so that model can adapt with few gradient updates Finn et al. (2017); Ravi and Larochelle (2017); Nichol et al. (2018); Antoniou et al. (2019).

The last two categories can also be termed as initialization-based or 'learning to fine-tune' methods. While the initialization based methods aim to enable rapid adaptation with a limited number of training examples for novel classes, these methods are shown Chen et al. (2019) to have limited generalization owing to the shifts between base and novel meta-testing classes as well as an inefficient base learner initialization owing to the difficulty of meta-training tasks. This motivates our attempt towards improving the base-learner while eliminating any discrepancy between the meta-training and meta-testing setups.

Curriculum learning Bengio et al. (2009) has been actively used for several tasks Hacohen and Weinshall (2019) to allow for ordering of training examples based on their difficulty, thus facilitating incremental learning of complex concepts. Its applications span across problems ranging from single static tasks Spitkovsky et al. (2010) to continual learning Wu et al. (2021) and transfer learning Weinshall et al. (2018); Zhang et al. (2020).
Recent work Wu et al. (2021) uses curriculum as well as meta-learning for continual relation extraction tasks to deal with catastrophic forgetting and order-sensitivity. However, unlike our method, the curriculum is limited to the traditional ordering of samples based on their complexity. To the best of our knowledge, our work is the first to propose a curriculum design for episodes/tasks in meta-learning approaches. It is important to note that while we proposed a curriculum learning schedule for gradient-based methods, our method can also be adapted for nearest-neighbor and metric-based methods.

3 Proposed Approach

Similar to the gradient-based method of model-agnostic meta-learning (MAML) Finn et al. (2017), our work employs a meta-training phase that explicitly optimises a weight initialization θ on a set of tasks, such that the base learner initialized with those weights is able to adapt to any unseen task when provided with a few samples (typically 1 to 20) of each class.

Unlike several meta learning works including MAML, we draw a clear distinction between the number of training examples at meta-test time (hereafter called $shot$) and the support size during meta-training (hereafter denoted by K). This distinction has also been made in recent works such as Nichol et al. (2018) that proposed meta-training with support size larger than the shot size i.e. $K > shot$. As the support set enables the meta-learner to adapt to a new task, we use the support size K as a proxy for a task's difficulty.

We propose that the support size K should therefore be adapted throughout meta-training. Concretely, we opt for a larger support size in the initial stages of meta-training that will be progressively decreased to eventually match that of the test distribution ($shot$) towards the end. This is equivalent to a curriculum design for the training episodes/tasks, where the initial meta-training tasks are easier due to the availability of larger support set and as the meta-training progresses, the tasks are harder due to limited support set. While several schedules can be designed to fit such as a general definition,

we limit our experiments to a rather simple yet novel schedule:
At each stage, the support set size will be equal a multiplier of the desired $shot$, where the multiplier value decreases with each stage. We first set the maximum multiplier value to a small natural number (which also corresponds to the number of curriculum stages) M as a hyper-parameter. We then reduce this multiplier by 1 in equally spaced iteration intervals until it reaches the value of 1, at which point $K = shot$ and the meta-training setup exactly matches that of meta-test. This computation corresponds to the line 7 of Algorithm 1. We denote the support size of each stage i as K_i where

$$K_i = (M - i) * shot \quad \forall i \in [0, M) \tag{1}$$

Following MAML we do not employ early stopping; instead a predetermined number of steps **N** is set as a hyper-parameter. However the checkpoint that achieved the lowest validation error is used during meta-testing to address potential overfitting. We designate half the training steps to the last stage K_M irrespective of the actual M setting in order to stabilise the meta learner and eliminate any discrepancy between meta-training and meta-testing. Therefore each stage will train for $\mathbf{n_i}$ steps where:

$$n_i = \begin{cases} \frac{N}{2M-2}, & \text{if } i < M \\ \frac{N}{2}, & \text{if } i = M \end{cases} \tag{2}$$

Algorithm 1 Curriculum MAML

Require: $\mathcal{T}$: distribution over tasks
Require: α_0, β: learning rate for inner and outer loop
Require: **N**: number of iterations, **M**: multiplier
1: randomly initialize θ
2: **set** $K \leftarrow shot * M$
3: **set** $\alpha \leftarrow \alpha_0 * \sqrt{M}$
4: **set** $n \leftarrow (N/2)/(M-1)$
5: **for** i **in** $range(1, N)$ **do**
6: **if** $modulo(i, n) = 0$ **then**
7: **update** $K \leftarrow max(K - shot, shot)$
8: **update** $\alpha \leftarrow max(\sqrt{\alpha^2 - \alpha_0^2}), \alpha_0)$
9: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
10: **for** all T_i **do**
11: Sample K datapoints $D = \{x^j, y^j\}$ from T_i
12: Evaluate the cross entropy loss $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ on D
13: Compute adapted parameters with gradient descent $\theta_i \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
14: Sample L datapoints $D' = \{x^j, y^j\}$ from T_i for the meta-update
15: **end for**
16: Compute adapted parameters $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i})$ on D'
17: **end for**

We provide the pseudocode for our method in Algorithm 1 which closely follows Algorithm 2 in Finn et al. (2017). Observe that the term $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ in the inner loop is estimated with a monotonically decreasing number of examples K and thus possibly increasing variance as the curriculum progresses through its stages.

In order to capitalize on the reduced variance during the initial stages of curriculum training, we propose an annealing schedule where the learning rate is scaled proportionally to the square root of the support set's size as per equation:

$$\alpha_i = \sqrt{M - i} * \alpha_0 \quad \forall i \in [0, M) \tag{3}$$

This annealing corresponds to line 8 in Algorithm 1. Similarly to Raffel et al. (2020), we select this annealing schedule because it does not require information on the total number of steps a-priori

unlike other popular schedules like triangular learning rate (Howard and Ruder (2018)). We do not employ warm up.

4 Experimental Setup

Datasets: We conduct experiments on two few-shot image classification tasks, Omniglot Lake et al. (2015) and miniImagenet Ravi and Larochelle (2017). Omniglot consists of 1623 characters drawn from 50 different alphabets, with 20 samples provided for each class. 1100 of those character classes are used for training, 100 for validation and the remaining 423 for testing. Similar to Santoro et al. (2016) we augment omniglot with rotations by multiples of 90 degrees. MiniImagenet consists of 100 classes, of which 64 are used for training, 12 for validation and 24 for testing purposes. No data augmentation is used. We release the code to reproduce our results at *https://github.com/steremma/curr-meta-learning*.

Base models: We follow the setup from MAML[1] Finn et al. (2017). Our base learners exactly follow MAML, both in the convolutional and the feed forward variant:

- Our **convolutional Network** follows Vinyals et al. (2016) which employs 4 layers, each using 3x3 kernels and 64 filters on Omniglot or 32 filters on miniImagenet Ravi and Larochelle (2017). Convolution is followed by batch normalisation and the ReLU activation function. Lastly we address overfitting using downsampling: a 2x2 max-pooling operation on miniImagenet and strided convolution on Omniglot.

- Our **Feed Forward Network** comprises of 4 hidden layers with sizes 256, 128, 64, 64 each including batch normalization and ReLU activation, followed by a linear layer.

While first order approximation of the Hessian matrix utilised in MAML's outer loop was shown to speed-up computation with minor performance degradation, we do not apply it on our experiments. All training schemes run for a total of $N = 60,000$ steps. These steps are allocated per curriculum stage as per Eq. 2. The curriculum framework is applied over the implementation available in *learn2learn* repo[2].

Experiments: We aim to investigate the following research questions related to our proposed approach:

RQ1: *Does curriculum based training lead to an improved model for unseen tasks in comparison to baseline static meta-learner?*

We experiment with a range of curriculum stages $M = 20, 10, 5, 1$. The setting with $M = 1$ corresponds to our baseline MAML training where there is no curriculum stage and the training support set size is same to that of the test. Therefore, we treat $M = 1$ as a baseline.

RQ2: *Do hyperparameters like learning rate, query set size etc. need to be adapted for each curriculum stage?*

We observe that the support set serves as the batch for the base learner and therefore enlarging the support set should ideally stabilize the gradients. This could enable the use of larger learning rate to accelerate training. We explore this hypothesis by experimenting with adaptive learning rate per equation 3.

Furthermore, it is not obvious whether the query set should always match the support set in terms of size in each stage, or instead stay statically equal to its test-time configuration. To answer this, we experiment with both adaptive query set size $L = K$ and the static one $L = shot$.

RQ3: *Is the effect of curriculum meta learning limited to specific training configurations?*

We investigate whether improved performance is limited to specific configurations such as:

[1] Our initial experiments with REPTILE also result in gains similar to shown in Section 5.1 thus indicating applicability of proposed curriculum frameworks irrespective of the baseline meta-learner model.

[2] https://github.com/learnables/learn2learn

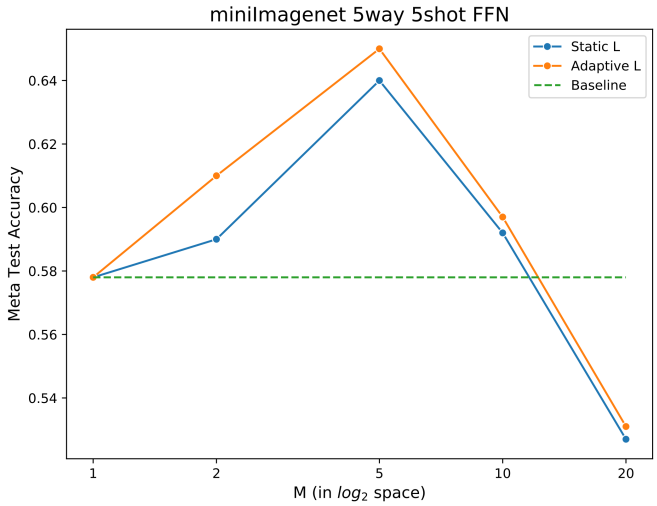


Figure 1: Meta Test performance on 5-way 5-shot miniImagenet using a Convolutional Network as the base learner. We attempt both a static ($L = shot$) and an adaptive ($L = K$) query size. For both configurations, the multiplier (M) of 5 is optimal, while adaptive L slightly outperforms the static one consistently across the range of multiplier values.

- The base learner.
- The dataset's complexity.
- The desired $shot$ configuration.

To answer this, we attempt our schedule on both base learner setups used in the original paper Finn et al. (2017), namely a fully connected and a convolution network, and both datasets (miniImagenet & Omniglot) with $shot = 1$ or $shot = 5$ examples.

RQ4: *Is gradual curriculum important?*

Several existing works (Nichol et al. (2018), Lee et al. (2019)) use a larger support set during meta-training in comparison to the $shot$ size used during meta-testing. We isolate the effect of curriculum by decoupling the effect of using bigger support size in our ablation study.

5 Results and Discussion

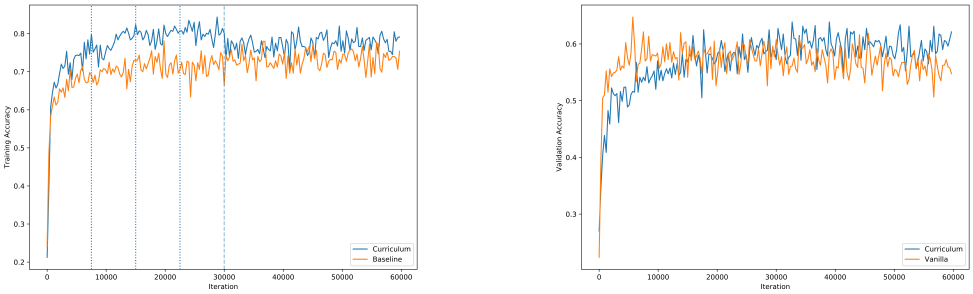
5.1 Impact of our proposed method:

As highlighted in our research question **RQ1**, we compare the performance of proposed curriculum learning framework with that of the baseline MAML model. We start by identifying a reasonable setting for M i.e. the number of curriculum stages. This analysis is done on the miniImagenet dataset with a 5-way 5-shot setup using a convolutional network as the base learner.

As we can see in Fig. 1, curriculum-based training significantly outperforms the baseline setup for a large range of settings M. For this dataset, we find that a reasonable choice is $M = 5$ and we continue to use this setting for our following experiments. Fig. 1 additionally illustrates a comparison between static and adaptive query set size. More details on this can be seen in Section 5.3

Interestingly, even a smaller value of $M = 2$ shows improved performance in comparison to the baseline[3]. This is important in practice, because in several applications there is a practical limitation for the maximum M one can use: The minimum amount of labeled samples required for each class

[3]In the 5-shot setting our reproduction of baseline diverges from the one reported in Finn et al. (2017). However values of M around 5 result in test accuracy even higher than the one reported in the original paper.



(a) Training Accuracy

(b) Validation Accuracy

Figure 2: Training and Validation accuracy for miniImagenet 5-way 5-shot with $M = 5$. Vertical lines denote stage switching. Prior to the last stage the base learner is trained on $K > shot$ examples which is reflected in the much higher training accuracy.

participating in meta-training is increased from $shot$ to $M * shot$. This is for example the reason we do not report 5-shot performance on Omniglot, as this would require 25 samples per class in the first stage when only 20 are available[4]. In many applications this limitation does not pose an issue. For example, consider a setup in which candidate labels are dynamically added to a classification system over time. In such cases, a practitioner might have plenty of labelled samples for historical classes at their expense but only **shot** samples for the newly added classes. In such cases the increased requirements for historical labelled data do not limit the selection of M.

On the other hand, it is important to note that for very high values of $M = \{10, 20\}$ the performance starts dropping. This possibly happens because it becomes difficult for the model to adapt to the meta-testing environment due to its exposure to high amounts of training data during meta-training.

This trade-off between the base learner optimization and generalization capacity is illustrated by juxtaposing the training accuracy curves in Fig. 2. During the first curriculum stages, the support set is the largest, resulting in very high training accuracy, which does not translate to validation accuracy because of the huge discrepancy between training and validation. In contrast, when entering the last stage where $K = shot$ the training accuracy unsurprisingly takes a hit as the number of training examples halves, however it stays above the baseline that was training on this size all along. At the same time the validation accuracy increases as the gap between the training and validation setups has vanished.

5.2 Generality of the proposed method

As per our **RQ3**, we investigate whether the results seen in Section 5.1 continue to hold with different base learners. For this purpose we design 2 setups on Omniglot, only differing in their base learner configuration:

- OmniGlot 5-way 1-shot with a Feed Forward Network as the base learner.
- OmniGlot 5-way 1-shot with a Convolutional Neural Network as the base learner.

The results are shown in the first three rows of Table 1. We first report the baseline version we reproduced which closely mirrors the results reported in the original MAML paper. The following two rows correspond to different curriculum variants, both of which outperform the baseline irrespective of the base learner. This result confirms that our method is largely agnostic to the choice of base learner.

Fortunately, our baseline experiments exactly match the numbers reported in Finn et al. (2017) for all other experimental setups

[4]Ideally, one can conduct curriculum training for 5-shot setting on Omniglot by either allowing for oversampling or reducing the number of curriculum stages

Model	Omniglot		miniImagenet	
	FFN 5-way 1-shot	CNN 5-way 1-shot	CNN 5-way 5-shot	CNN 5-way 1-shot
Baseline MAML	0.90	0.99	0.58	0.48
Curriculum learning				
$L = shot$ (static)	0.92 (+.02)	**1.00** (**+.01**)	0.64 (+.06)	0.51 (+.03)
$L = K$ (adaptive)	**0.94** (**+.04**)	**1.00** (**+.01**)	**0.66** (**+.08**)	**0.54** (**+.06**)
Static Support Size				
$K = M * shot$, $L = shot$	0.87 (-.03)	0.98 (-.01)	0.51 (-.07)	0.45 (-.03)
$K = M * shot$, $L = K$	0.91 (+.01)	0.99 (+.01)	0.55 (-.03)	0.46 (-.02)

Table 1: Results on the OmniGlot and miniImagenet datasets. All variants use a multiplier $M = 5$. Our static variants use $K = M * shot$ throughout training. For the miniImagenet results, we report baseline results as reproduced in our environment which in the 5-shot case diverge from the ones reported in Finn et al. (2017). The original paper mentions a 5-shot test accuracy of **0.63** instead of **0.58** as reported here. However, the deltas would likely lie in the same range as our variants are based on the same implementation as the reported baseline.

Subsequently, we explore whether the observed performance boost is related to the complexity of the dataset, or the desired $shot$ setting. Also in Table 1 we show results on miniImagenet using the convolutional base learner for $shot = 1, 5$. In both setups, we observe exactly the same behaviour as in Omniglot, i.e. the curriculum schedule results in improved performance. These findings corroborate that our method is robust to the base learner, the complexity of the dataset and the desired $shot$ configuration.

5.3 Performance with adaptation of hyperparameters:

We analyze if setting hyperparameters in accordance to the support set's size K at a given curriculum stage yields further improvement. For this study, we select two prominent hyperparameters: **learning rate** α and **query set size** L.

Adapting learning rate (α): As discussed in section 3, the larger support set used in early stages of the curriculum should yield a less noisy gradient to the base learner in the inner loop. In addition, we hypothesize that increasing the learning rate of the base learner could lead to faster and potentially better optimization. Specifically we attempt to scale the learning rate as per equation 3 using the base $\alpha_0 = 0.5$ as in the original implementation. While curriculum learning with static learning rate already performs better than the baseline, we found that annealing provides further consistent gains in terms of test accuracy as shown in Fig. 3. Therefore, throughout all experiments illustrated in our paper we use an adaptive learning rate.

Adaptive Query set size (L): For each of the setups listed in Section 5.2, we first report the baseline version we reproduced, which closely matches the results reported in the original MAML paper. We then report performance for two variations of our proposed curriculum schedule: keeping the query size $L = shot$ throughout training, or adapting it to $L = K$ for each stage of training. Comparing the rows under the **Curriculum learning** header in Table 1 shows that adapting the query set size consistently increases gains across experimental setups.

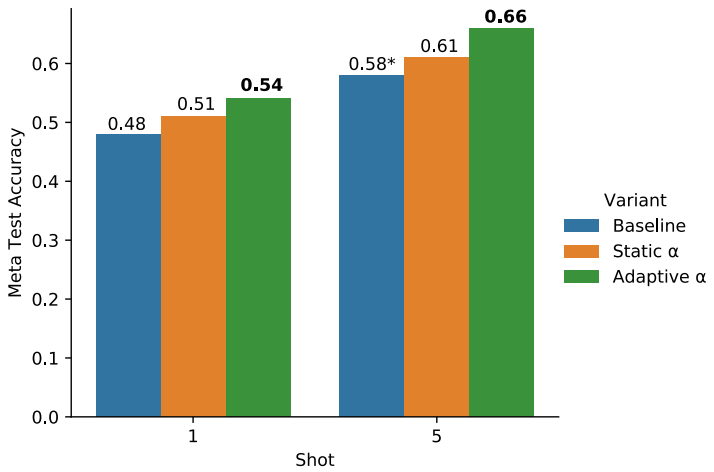


Figure 3: Meta-Test performance on 5-way 5-shot miniImagenet using a Convolutional Network as the base learner. We explore the effect of annealing the learning rate proportionally to the square root of the support set's size per Equation (3).

5.4 Effect of statically larger support size

Lastly, for all setups, we report ablation numbers where we train without curriculum, but instead with a larger support size equal to $K = M * shot$ (the maximum value used in the curriculum variations) kept constant throughout training. This is to explore whether our improvement comes from the schedule, or simply by using a support size higher than shot as reported in Nichol et al. (2018); Lee et al. (2019). There are again two variations on this ablation, where the query size is statically equal to $shot$, or to K corresponding to the last two rows in Table 1. Our ablation reveals that neither version achieves the improved performance of curriculum meta-learning. The observed gains should therefore be attributed to the proposed schedule on the support set size and not its absolute size at a given stage.

Additionally, since statically using a higher K and/or L does not yield this improvement we conclude that it is indeed the adaption of those parameters and not their maximum size that results in the performance boost.

6 Conclusions

We present a simple, yet effective curriculum method to boost few-shot classification performance within the meta-learning framework. We apply the proposed method on top of MAML, a well known meta learning algorithm, and show consistent gains across base learners, datasets and $shot$ configurations. Furthermore, our method is conceptually applicable to other particular meta learning frameworks and can be easily tested in combination with modern, state-of-the-art meta learning algorithms. To facilitate future work we make our source code publicly available.

References

Thrun, S. In *Learning to Learn*; Thrun, S., Pratt, L., Eds.; Springer US: Boston, MA, 1998; pp 181–209.

Vinyals, O.; Blundell, C.; Lillicrap, T.; kavukcuoglu, k.; Wierstra, D. Matching Networks for One Shot Learning. Advances in Neural Information Processing Systems. 2016.

Snell, J.; Swersky, K.; Zemel, R. S. Prototypical Networks for Few-shot Learning. *CoRR* **2017**, *abs/1703.05175*.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H. S.; Hospedales, T. M. Learning to Compare: Relation Network for Few-Shot Learning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018; pp 1199–1208.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; Lillicrap, T. Meta-Learning with Memory-Augmented Neural Networks. Proceedings of The 33rd International Conference on Machine Learning. 2016; pp 1842–1850.

Munkhdalai, T.; Yu, H. Meta Networks. Proceedings of the 34th International Conference on Machine Learning. International Convention Centre, Sydney, Australia, 2017; pp 2554–2563.

Mishra, N.; Rohaninejad, M.; Chen, X.; Abbeel, P. A Simple Neural Attentive Meta-Learner. International Conference on Learning Representations. 2018.

Ravi, S.; Larochelle, H. Optimization as a Model for Few-Shot Learning. ICLR. 2017.

Nichol, A.; Achiam, J.; Schulman, J. On First-Order Meta-Learning Algorithms. *CoRR* **2018**, *abs/1803.02999*.

Lee, K.; Maji, S.; Ravichandran, A.; Soatto, S. Meta-Learning with Differentiable Convex Optimization. *CoRR* **2019**, *abs/1904.03758*.

Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. ICML '09. 2009.

Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese Neural Networks for One-shot Image Recognition. 2015.

Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Proceedings of the 34th International Conference on Machine Learning. International Convention Centre, Sydney, Australia, 2017; pp 1126–1135.

Antoniou, A.; Edwards, H.; Storkey, A. How to train your MAML. International Conference on Learning Representations. 2019.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; Huang, J.-B. A Closer Look at Few-shot Classification. International Conference on Learning Representations. 2019.

Hacohen, G.; Weinshall, D. On The Power of Curriculum Learning in Training Deep Networks. ICML. 2019.

Spitkovsky, V. I.; Alshawi, H.; Jurafsky, D. From Baby Steps to Leapfrog: How "Less is More" in Unsupervised Dependency Parsing. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California, 2010; pp 751–759.

Wu, T.; Li, X.; Li, Y.; Haffari, R.; Qi, G.; Zhu, Y.; Xu, G. Curriculum-Meta Learning for Order-Robust Continual Relation Extraction. *CoRR* **2021**, *abs/2101.01926*.

Weinshall, D.; Cohen, G.; Amir, D. Curriculum Learning by Transfer Learning: Theory and Experiments with Deep Networks. ICML. 2018; pp 5235–5243.

Zhang, S.; Zhang, X.; Zhang, W.; Søgaard, A. Worst-Case-Aware Curriculum Learning for Zero and Few Shot Transfer. *arXiv e-prints* **2020**, arXiv:2009.11138.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* **2020**, *21*, 1–67.

Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. 2018; pp 328–339.

Lake, B. M.; Salakhutdinov, R.; Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **2015**, *350*, 1332–1338.