State of the Art on: Meta-learning for Few-Shot Classification

Nicola De Angeli, nicola.deangeli@mail.polimi.it

March 2020

1. Introduction to the research topic

Machine Learning is an application of Artificial Intelligence that sets as its goal to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest [35]. In recent years, a class of parametric machine learning techniques called Deep Learning has led to astonishing achievements in the field. The key aspect of Deep Learning is the ability to learn how to extract relevant features from data through the employment of many neural layers, thus not requiring field-specific expertise [31]. As a consequence, the leverage of these techniques is very approachable and results in lower costs, promoting the development of artificial intelligence applications outside the academic community.

Nonetheless, Deep Learning currently faces some obstacles that still hinder the technology to be fully exploited. Humans have a remarkable capacity to quickly learn new concepts when provided with few examples. Conversely, current popular deep learning techniques need thousand of samples to be able to generalize their knowledge and make predictions on unseen data, making them extremely data inefficient. In the context of Supervised Learning, this often translates into the need to manually label thousands of samples, which is cumbersome and time-consuming. In Reinforcement Learning (RL) this translates in having access to a large number of training trajectories and this may be unfeasible when the experience is directly observed from real-world interactions.

Meta-Learning, also known as "learning-to-learn", is a sub-field of Machine Learning that exploits previous experience to optimize learning algorithms to work well on novel tasks [16]. The discipline is currently most active in the Supervised Learning setting and is often associated with the field of Few-Shot Learning, where models are challenged to quickly learn new concepts while very few datapoints from the task at hand are available.

The problem is extremely hard, as models operating in low data regimes are especially prone to overfitting [33]. The field of Meta-Learning has been rising in the last few years, achieving super-human performances in simple few-shot classification tasks [29]. The reasons for this are twofold; on one hand, deep-learning techniques have been widely employed in Meta-Learning, on the other, the need for Deep Learning for lots of training data has motivated the development of techniques to overcome its limitations. Deep Learning and Meta-Learning thus form a symbiotic relationship that incentivizes progress in the two fields [1]

The table below lists the most prestigious conferences related to Meta-Learning, along with their respective h5-index. While they generally focus on the broader topics of Artificial Intelligence, Machine Learning, and Deep Learning, such conferences also recently held workshops [2, 3] and tutorials [1] concerning Meta-Learning in the last few years, indicating a rising interest in the subject.

Name of the conference	Google Scholar h5-index
Neural Information Processing Systems (NIPS)	169
International Conference on Learning Representations (ICLR)	150
International Conference on Machine Learning (ICML)	135
AAAI Conference on Artificial Intelligence (AAAI)	95
International Joint Conference on Artificial Intelligence (IJCAI)	67
International Conference on Artificial Intelligence and Statistics (AISTATS)	52

[1] This is also reflected in the various meta-learning libraries that have been developed on top of deep learning oriented frameworks, such as Higher [22] and Torchmeta [12].

1.1. Preliminaries

We will mainly consider the problem of Meta-Learning in the context of few-shot classification tasks, for ease of formulation and relevance to our focus.

Unlike standard classification, where samples are tuples of input and desired class label from a task, in Meta-Learning a single sample represents an entire task. A task is defined as a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, where $\mathbf{x}_i$ is an input and y_i is the desired label for $\mathbf{x}_i$ from a known set of possible labels $\mathcal{L}$. The dataset $\mathcal{D}$ is thus split into two disjoint sets, a support set S for learning the task and a prediction set B for testing the performance with respect to the task. The goal is to find a model that performs well over a distribution of tasks $p(\mathcal{D})$. The meta-learning model is formalized by resorting to the concepts of learner and meta-learner. The learner f_θ is a classifier with parameters θ and trained to perform well with respect to a particular task. The meta-learner g_ϕ is an optimizer with meta-parameters ϕ and trained to update the learner parameters optimally for a generic task from the distribution. The parameters $\theta' = g_\phi(\theta, S)$ for the new task are computed by the meta-learner starting from some initial parameters θ via the support set S and should be such that the new model $f_{\theta'}$ achieves good performance on the prediction set B. If we assume the output of the learner f_θ to be modeled as the probability $f_\theta(\mathbf{x}) = P_\theta(y|\mathbf{x})$ for all $y \in \mathcal{L}$, then the problem becomes to find the initial parameters θ and the meta-parameters ϕ maximizing

$$\mathbb{E}_{\mathcal{D}=\langle S,B\rangle \sim p(\mathcal{D})}\left[\sum_{(\mathbf{x},y)\in B} P_{g_\phi(\theta,S)}(y|\mathbf{x})\right]$$

In practice, the distribution $p(\mathcal{D})$ is often unknown and we resort to learning from a fixed collection of tasks, a meta-dataset, which can be modeled as samples from such distribution. This dataset can be split into a training set and a validation set to estimate the meta-learning model performance, analogously to traditional learning tasks. It is important to understand the difference between the training set and the support set. The former is a collection of tasks as datasets and thus contains data from multiple tasks, while the latter contains datapoints from a specific task. The same argument holds for the validation set and the prediction set.

The paradigm of episodic training proposed by Vinyals et al. [54] is the standard in the area of few-shot classification, as it provides a simple way to sample, starting from a dataset featuring a number C of classes, many few-shot classification tasks that are compatible with the learner, which in general can differentiate among a number of classes $N < C$. A K-shot classification task is thus obtained by first sampling N different classes from the C available in the dataset, each sampled class is then randomly assigned a specific label and a number K of sampled datapoints belonging to the class, the obtained dataset is finally partitioned in support and prediction set.

Many meta-learning techniques have been implemented using TensorFlow [4] or PyTorch [41], two popular machine learning frameworks for auto-differentiation.

1.2. Research Topic

Meta-Learning is a family of techniques that aim to generalize the knowledge derived from previously encountered tasks to perform better on new, unseen tasks. The approach has been shown to successfully address some of the challenges posed by Few-Shot Learning, where very few task-specific training datapoints are available. The literature has recently provided promising results also thanks to the rapid developments and latest breakthroughs in Deep Learning, achieving human-like performance in simple meta-learning tasks [29].

Meta-learning techniques may also shed light on the inner mechanisms of the human brain, as humans also base their behavior and learning on previous experience. Though ambitious, Meta-Learning might play in the future an important role in the discovery of Artificial General Intelligence.

Finally, by focusing on this topic, we aspire to broaden the availability of deep learning techniques. A model that is capable to learn new, complex tasks and generalize knowledge with few training samples would prove beneficial to experts confronting niche machine learning tasks with little training data available online, which is, for instance, the case when working with clinical data.

2. Main related works

2.1. Classification of the main related works

The meta-learning literature provides a useful taxonomy used to classify the various methods.

Metric-Based approaches predict the class probability y related to a new input $\mathbf{x}$, given the support set S, by leveraging a kernel function that measures the similarity between $\mathbf{x}$ and each datapoint in S. The performance of a metric-based model thus entirely depends on the quality of its kernel function. *Model-Based approaches* make no assumption on the shape of the output, focusing on models that can learn quickly with few training steps thanks to their inner structure or the help of a powerful meta-learner. Their performance often relies on manual design choices, a possible future improvement thus would be to provide a way to learn optimal configurations. *Optimization-Based approaches*, such as MAML [14], find a shared initialization of the model parameters θ_0, across all tasks of the distribution that can be quickly adapted to task-specific parameters in few steps of gradient descent. The main problem of this class of algorithms is the heavy load of computation they require when meta-training complex learners because of second-order derivatives. Simplified models have been proposed to scale optimization-based approaches and overcome their limitations, performing on par or better than the original MAML model [15, 36, 44, 47, 59]. *Conditional neural processes (CNPs)* [17] combine the benefits of Gaussian processes, able to incorporate prior knowledge in function approximation but computationally expensive, and Deep Neural Networks. The result is a model that provides the flexibility of Stochastic Processes while featuring a neural network structure that can be trained via gradient descent. Examples of few-shot learning models based on CNPs [17] are VERSA [19] and CNAPs [46]. While CNPs perform very well in Few-Shot regression, classification, and image completion tasks, they struggle to obtain similar performances in the many-shot case.

In the last few years, the field of Meta-Learning has thrived and managed to provide methods capable of obtaining human and superhuman-level performance in simple tasks such as one-shot classification. Nonetheless, it is evident that recent approaches still struggle when confronted with more complex tasks, such as in the case of the Omniglot challenge [29].

2.2. Brief description of the main related works

2.2.1 Metric-Based Meta Learning

These approaches focus on learning generalizable embeddings, based on on the assumption that the embeddings capture all necessarily discriminative representations of data and that simple non-parametric classifiers are sufficed. This is the case of Matching Networks, where Vinyals et al. [54] propose to learn the embedding with a differentiable nearest neighbor objective. Relation Networks [51] learn a deep distance metric for images to predict labels of new samples given other labeled datapoints. Prototypical Networks [50] learn a metric space where classification can be performed by computing the distance of a point to prototype representations of each class, leading to good few-shot learning performance but quickly saturating when the number of shots is large [52]. Few-shot Embedding Adaptation with Transformer (FEAT) [57] leverages the Transformer self-attention mechanism to identify relationships among new and previously seen instances. Task dependent adaptive metric for improved few-shot learning (TADAM) [38] learns a feature extractor and metric scaling for the task at hand to provide better results. Triantafillou et al. [52] propose Proto-MAML, a model combining the simple inductive bias of Prototypical Networks and the flexible adaptation mechanism of MAML, as well as Meta-Dataset, a novel meta-learning benchmark that pays attention to the relationship within classes when generating new episodes to obtain more realistic tasks.

2.2.2 Model-Based Meta-Learning and Fast weights

Ravi and Larochelle [45] employ Long Short-term Memory cells [24] to learn a few-shot meta-learner that is capable of training a model in few update steps. Santoro et al. [48] employ Memory-Augmented Neural Networks such as Neural Turing Machines [20, 21] to combine the benefits of gradient-based learning and memory methods.

As a downside, the considered memory-addressing procedure is manually selected, whereas an optimal one could be learned automatically.

Fast weights architectures divide weights in slow weights learned on all the tasks and fast weights generated for the task at hand. Schmidhuber [49] showed that a slow feed-forward neural network can be used to generate context-dependent weight changes for a second network, but only small scale experiments were conducted. Subsequent work demonstrated practical applications of fast weights [18], where a generator network is learned through evolution to solve an artificial control problem. Successively Ha et al. [23] explore fast-weights for recurrent neural networks under the name of *hyper-networks*. Different variations of fast weights have been proposed to generate weights of convolutional neural networks [37], including conditional modulation via affine transformations of features [38, 42] and policies [9, 10, 26]. Meta Networks [34] features a learner that provides the meta-learner with metadata on the task dataset to generate better-performing fast weights. In general, model-based approaches resort in very powerful models, but they require careful hyper-parameter tuning, normalization because of training instability [7, 23].

2.2.3 Optimization-Based Meta Learning

Model-Agnostic Meta-Learning (MAML) [14] is a simple, model-agnostic procedure that learns an optimal parameter initialization for a given learner by gradient descent. The need for multiple backward passes during meta-training makes MAML prohibitively expensive to run on very complex learners. Some first-order approximation may alleviate the computation load at the expense of performance [36]. The proposed training paradigm, characterized by an inner-outer loop on the tasks, has been widely reused in subsequent approaches from the literature. Latent Embedding Optimization (LEO) [47] embeds the model parameters to perform gradient descent in a low-dimensional space. Because the training procedure is based on MAML, LEO shares with the former many issues, such as heavy computational cost. The problem is however partially addressed thanks to the low dimensionality of the embedding. Fast Context Adaptation via Meta-Learning (CAVIA) [59] extends MAML by dividing the model parameters in context and shared parameters, leading to less propensity to meta-overfitting, easier parallelization, and better interpretability. MAML has also been shown to not perform well when tasks differ greatly [28]. On this matter, Multimodal MAML [55] extends MAML by providing a parameter initialization dependent on the mode of the task in case of multimodal task distributions. The approach does not, however, contemplate the possibility for tasks from different modes to share some relevant knowledge, which may act as a beneficial regularization. Hierarchically Structured Meta-Learning (HSML) [56] organizes tasks in a hierarchical clustering structure to provide both knowledge customization on the task and knowledge sharing among tasks. However, the ways in which the model can dynamically extend the task hierarchy are limited, which may be especially suboptimal in a continual learning setting where the structure of tasks may change over time.

Transformation Networks (T-nets) [32] learn a distance metric that warps the activation space such that a single gradient descent step yields parameters that are well suited for the task at hand. The distance metric is efficiently learned by interleaving linear projections that are meta-optimized only in the outer loop and are shared across tasks of the distribution. This approach is equivalent to learn a preconditioning matrix of the gradient. Similarly, Warped Gradient Descent (WarpGrad) [15] extends T-nets with non-linear warp projections, and can be applied to non-feed-forward neural networks. Moreover, the authors define a meta-objective in a joint search space across tasks, resorting to a novel meta-optimization procedure that is agnostic on the number of steps of the inner loop. Other forms of preconditioning have been proposed, by parameterizing the task gradient in different ways [33, 39].

2.2.4 Conditional Neural Processes

Versatile Amortized Inference (VERSA) [19] employs an amortization network taking as input a few-shot learning dataset with an arbitrary number of shots and classes to provide a distribution over task-specific parameters. Despite performing extremely well in one-shot classification tasks, the model is not able to preserve state-of-the-art results when the number of shots increases. Conditional Neural Adaptive Processes (CNAPs) [46] comprises a classifier learner whose parameters are adapted by a Conditional Neural Process taking as input the task dataset.

Proposed future lines of work concerning CNAPs include the use of gradients and function approximations in the adaptation mechanism, as well as considering distributional extensions.

2.2.5 Multitask Learning

Multitask and Meta-Learning are very related fields of Machine Learning. Both exploit the presence of shared structures across multiple tasks to speed up the training, with a different goal in mind. Informally, Multitask Learning aims to efficiently learn several tasks that are presented together, rather than training separate models for each one. The hope is that, if tasks are related, sharing the knowledge across them will allow training a single, compact model quicker and more efficiently. Meta-Learning instead, has the goal of extracting the knowledge from a distribution of tasks, such as learning a prior, to generalize and learn quicker on new unseen tasks.

Several approaches exist to model shared information across tasks. They can be roughly separated into two different categories i.e., methods where parameters of the model are close to each other in a geometric sense [13, 53] and approaches where the parameters of the model share a common structure [11, 30, 40, 43, 58]. This structure can be a clustering assumption [58], a (Gaussian) prior for the parameters of all tasks [30] or some advanced structure like the Kingman's coalescent [11] which is a continuous-time partitioned prior. Argyriou et al. [8] propose an inductive bias on task parameters assuming them to lie in a low dimensional linear subspace. Successively, Agarwal et al. [5] consider all task parameters to lie on a manifold. Based on the subspace assumption, Kumar and Daume III [27] propose a framework to selectively share the information across tasks, assuming that each task parameter vector is a linear combination of a finite number of underlying basis tasks. Other works differentiate between tasks and address the fact that some of them might be unrelated, by assuming the existence of *disjoint groups of tasks* [25], or allowing two tasks from different groups to overlap by having one or more bases in common [27].

2.3. Discussion

The meta-learning literature has produced a large variety of interesting and performing approaches in the last few years, mainly thanks to the recent breakthroughs in Deep Learning. However, despite the latest efforts and achievements in the field, the literature still does not offer a method that can generalize its knowledge on tasks with new, unseen data domains, which hinders the ability to apply deep learning solutions when the amount of available training data is low. Many popular approaches also struggle when scaling to the complexity of the learner, which ultimately constrains them to poor performance when confronting complex tasks. In some cases, heterogeneity of the various tasks may also be the cause of counter-productive transfer learning, where irrelevant knowledge from a task is erroneously reused when dealing with another task. The problem may be addressed by reinforcing the task-specific inductive bias, though it is important to introduce regularization to avoid unwanted overfitting.

In light of these issues, one of our objectives is to determine whether the currently provided techniques can be further extended to generalize previous knowledge on new tasks presenting unseen data domains. Current techniques were designed and tested on simple benchmarks featuring samples from a single domain, such as Omniglot [6]. A model capable of operating among different data domains would be able to transfer knowledge among widely different tasks, potentially solving the lack of training samples that are observed in certain data domains. As a practical example, our desired model would be able to generalize the recognition of malignant tumors in x-ray images to images obtained through other less popular techniques or instruments which may feature different colors and patterns.

References

[1] ICML 2019 Tutorial: Meta-Learning. https://sites.google.com/view/icml19metalearning, 2019. [Online; accessed 30-March-2020].

[2] NeurIPS 2019 Workshop: Meta-Learning (MetaLearn 2019). http://metalearning.ml/2019/, 2019. [Online; accessed 30-March-2020].

[3] NeurIPS 2019 Workshop: Learning Transferable Skills. https://www.skillsworkshop.ai/, 2019. [Online; accessed 30-March-2020].

[4] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[5] Arvind Agarwal, Samuel Gerber, and Hal Daume. Learning multiple tasks using manifold regularization. In *Advances in neural information processing systems*, pages 46–54, 2010.

[6] Simon Ager. Omniglot. https://www.omniglot.com/, 1998. [Online; accessed 30-March-2020].

[7] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, pages 3981–3989, 2016.

[8] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[9] Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Arian Hosseini, Pushmeet Kohli, and Edward Grefenstette. Learning to understand goal specifications by modelling reward. 2018.

[10] Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. Gated-attention architectures for task-oriented language grounding. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[11] Hal Daumé III. Bayesian multitask learning with latent hierarchies. *arXiv preprint arXiv:0907.0783*, 2009.

[12] Tristan Deleu, Tobias Würfl, Mandana Samiei, Joseph Paul Cohen, and Yoshua Bengio. Torchmeta: A Meta-Learning library for PyTorch, 2019. URL https://arxiv.org/abs/1909.06576. Available at: https://github.com/tristandeleu/pytorch-meta.

[13] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[15] Sebastian Flennerhag, Andrei A Rusu, Razvan Pascanu, Hujun Yin, and Raia Hadsell. Meta-learning with warped gradient descent. *arXiv preprint arXiv:1909.00025*, 2019.

[16] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimilano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. *arXiv preprint arXiv:1806.04910*, 2018.

[17] Marta Garnelo, Dan Rosenbaum, Chris J Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J Rezende, and SM Eslami. Conditional neural processes. *arXiv preprint arXiv:1807.01613*, 2018.

[18] Faustino Gomez and Jürgen Schmidhuber. Evolving modular fast-weight networks for control. In *International Conference on Artificial Neural Networks*, pages 383–389. Springer, 2005.

[19] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018.

[20] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[21] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

[22] Edward Grefenstette, Brandon Amos, Denis Yarats, Phu Mon Htut, Artem Molchanov, Franziska Meier, Douwe Kiela, Kyunghyun Cho, and Soumith Chintala. Generalized inner loop meta-learning. *arXiv preprint arXiv:1910.01727*, 2019.

[23] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[25] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *ICML*, volume 2, page 4, 2011.

[26] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[27] Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.

[28] Alexandre Lacoste, Boris Oreshkin, Wonchang Chung, Thomas Boquet, Negar Rostamzadeh, and David Krueger. Uncertainty in multitask transfer learning. *arXiv preprint arXiv:1806.07528*, 2018.

[29] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104, 2019.

[30] Alessandro Lazaric and Mohammad Ghavamzadeh. Bayesian multi-task reinforcement learning. In *ICML-27th International Conference on Machine Learning*, pages 599–606. Omnipress, 2010.

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[32] Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2927–2936, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/lee18a.html`.

[33] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[34] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563. JMLR. org, 2017.

[35] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[36] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[37] Junhyuk Oh, Satinder Singh, Honglak Lee, and Pushmeet Kohli. Zero-shot task generalization with multi-task deep reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2661–2670. JMLR. org, 2017.

[38] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018.

[39] Eunbyung Park and Junier B Oliva. Meta-curvature. In *Advances in Neural Information Processing Systems*, pages 3309–3319, 2019.

[40] Alexandre Passos, Piyush Rai, Jacques Wainer, and Hal Daume III. Flexible modeling of latent task structures in multitask learning. *arXiv preprint arXiv:1206.6486*, 2012.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[42] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[43] Piyush Rai and Hal Daumé III. Infinite predictor subspace models for multitask learning. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 613–620, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[44] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, pages 113–124, 2019.

[45] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2016. URL `https://openreview.net/forum?id=rJY0-Kcll`.

[46] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. In *Advances in Neural Information Processing Systems*, pages 7957–7968, 2019.

[47] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=BJgklhAcK7`.

[48] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016.

[49] Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

[50] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[51] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[52] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rkgAGAVKPr`.

[53] Aleš Ude, Andrej Gams, Tamim Asfour, and Jun Morimoto. Task-specific generalization of discrete and periodic dynamic movement primitives. *IEEE Transactions on Robotics*, 26(5):800–815, 2010.

[54] 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.

[55] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. Multimodal model-agnostic meta-learning via task-aware modulation. In *Advances in Neural Information Processing Systems*, pages 1–12, 2019.

[56] Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. Hierarchically structured meta-learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7045–7054, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/yao19b.html`.

[57] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Learning embedding adaptation for few-shot learning. *arXiv preprint arXiv:1812.03664*, 2018.

[58] Kai Yu, Volker Tresp, and Anton Schwaighofer. Learning gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine Learning (ICML-05)*, pages 1012–1019, 2005.

[59] Luisa M Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.