Understanding deep meta-learning

Huisman, M.

Chapter 1

Introduction

Chapter overview

The invention of neural networks marks a critical milestone in the pursuit of true artificial intelligence. However, despite their impressive performance on various tasks, deep neural networks face limitations in learning efficiently. This chapter introduces the concept of *deep meta-learning*, an approach aimed at enhancing learning efficiency in neural networks. Despite the development of numerous new deep meta-learning algorithms, we identify that there remains a research gap in understanding their performance. We motivate the importance of addressing this research gap and make steps towards filling this in the remainder of the dissertation. Moreover, this chapter outlines the specific research questions explored in this dissertation and provides an overview of its structure.

1.1 Background

Artificial deep neural networks have demonstrated remarkable capabilities, achieving human-level or even super-human performance across various tasks (Krizhevsky et al., 2012; Silver et al., 2016; Brown et al., 2020). The performance of these networks, however, hinges on the availability of large volumes of training data, often necessitating access to high-end computational resources (LeCun et al., 2015). Without sufficient data, these networks struggle to capture the intricacies of the underlying task, leading to suboptimal performance and limited generalization. This data dependency and computational demand pose significant challenges, particularly in domains where data collection is arduous, expensive, or privacy-sensitive. Overcoming these limitations and improving the data efficiency of deep neural networks is crucial to unlocking their full potential, enabling them to learn and adapt from limited data, and facilitating their deployment in resource-constrained environments, thereby democratizing the use of deep learning.

The learning efficiency of artificial deep neural networks is in stark contrast to human learning, which is highly efficient (Lake et al., 2015; Salakhutdinov et al., 2012; Schmidt, 2009). That is, humans can learn new concepts from only a handful of examples or a limited amount of experience. For instance, if shown a picture of a unique animal species that they have never encountered before, humans can often grasp its distinguishing features and correctly identify similar instances with high accuracy.

In contrast, deep neural networks typically struggle to achieve comparable performance with such limited data.

One potential explanation for this gap in learning efficiency between human learning and deep learning is that humans can rely upon a large set of prior knowledge and experience to quickly learn to perform new tasks (Jankowski et al., 2011; Lake et al., 2017). For example, humans often first learn to crawl and walk before they learn how to run. This allows us to transfer our movement patterns and knowledge from crawling and walking to more quickly learn running than would be possible if we were directly to attempt to learn how to run. Coming back to the example of learning to identify a new animal species, humans can leverage their understanding of the visual world and feature detectors from all prior experiences to quickly identify new species from only a handful of examples.

Deep neural networks, in contrast to humans, are often characterized by their high degree of specialization towards specific tasks. They typically undergo end-to-end training, meaning they learn directly from available data without any pre-existing knowledge or prior experience. This lack of prior knowledge places a considerable burden on deep neural networks, as they essentially attempt to master complex tasks without having learned foundational skills or basic concepts. In a metaphorical sense, it is as if these networks are trying to run before learning to crawl or walk. In a more practical sense, this absence of previously learned movement patterns and task-specific knowledge requires deep neural networks to rely solely on the limited data provided during training, making the learning process more challenging and time-consuming. In order to overcome this limitation, deep neural networks would greatly benefit from the ability to leverage prior experience and existing data that could be harnessed to facilitate the learning of the given task at hand. By incorporating such prior knowledge, these networks could potentially enhance their learning efficiency, accelerate the training process, and achieve higher performance with limited data.

Deep meta-learning (Schmidhuber, 1987; Thrun, 1998; Naik and Mammone, 1992) is the field of research that aims to endow deep neural networks with the ability to reuse a large set of prior experiences in order to learn to perform new tasks more efficiently. Since any prior knowledge in neural networks is encoded in the hyperparameters (such as the initialization parameters when learning a new task, the learning rate of stochastic gradient descent, or the weight update rule), the goal of deep meta-learning can be formulated as extracting a good set of hyperparameters from a set of prior tasks such that new tasks can be learned more efficiently than a neural network without such prior knowledge. We note that this is a form of transfer learning (Pan and Yang, 2009; Taylor and Stone, 2009) as we transfer knowledge obtained from previous tasks to a new task in which we are interested.

A common testbed for deep meta-learning algorithms that we also employ throughout this dissertation is few-shot learning (Wang et al., 2020b; Ravi and Larochelle, 2017; Vinyals et al., 2016), where the neural network has to learn different tasks only from a handful of examples, or *shots* per class. In few-shot learning scenarios, deep neural networks are presented with a set of classes or categories, each accompanied by a limited number of labeled instances (shots). The objective is to enable the network to generalize from this sparse data and quickly adapt to new classes with only a few examples. This setting closely emulates the remarkable ability of humans to rapidly acquire new knowledge and skills from minimal exposure and allows us to assess the ability of the networks to reuse prior knowledge in order to quickly learn to perform new tasks.

1.2 Motivation

The concept of meta-learning with neural networks was pioneered in the 1980s - 2000s (Schmidhuber, 1987; Bengio et al., 1991; Hochreiter et al., 2001; Thrun, 1998; Naik and Mammone, 1992), but it is only in recent years that the field has experienced a surge in popularity, driven by promising results presented by Vinyals et al. (2016), Finn et al. (2017), and other researchers. This increased attention has led to the introduction of numerous novel techniques, resulting in significant advances in the few-shot learning performance of these methods. However, despite these achievements, the reasons behind why certain meta-learning algorithms outperform others often remain elusive. The underlying principles that determine the success of different meta-learning approaches have received limited attention and are, therefore, not well understood. This knowledge gap hampers our ability to make informed design choices and improve deep meta-learning algorithms.

It is crucial to address this research gap to gain knowledge about meta-learning algorithms beyond simple performance comparisons, such as method A outperforming method B. By delving deeper into the underlying principles, we can gain a comprehensive understanding of why certain meta-learning algorithms succeed while others fall short. This deeper insight, in turn, could allow us to design enhanced deep meta-learning algorithms and reason about the expected impact of specific design decisions on the performance of different algorithms. In essence, it empowers us to move beyond empirical observations and establish a theoretical foundation for deep meta-learning, enabling more robust and effective algorithm development.

Furthermore, an equally important aspect that has received limited attention is investigating how theoretical principles can be leveraged to improve the performance of meta-learning algorithms. By bridging the gap between theory and practice, we can uncover valuable insights into how theoretical foundations can be practically applied to enhance the capabilities of deep meta-learning algorithms. Understanding the interplay between theoretical principles and algorithmic design choices can guide the development of more efficient and effective meta-learning techniques, fostering advancements in few-shot learning and other related fields.

In summary, advancing the understanding of deep meta-learning algorithms is a critical research direction. By uncovering the underlying principles that govern their success, we can gain valuable insights beyond simple performance comparisons. This knowledge not only facilitates the development of improved deep meta-learning algorithms but also provides a theoretical framework to reason about the influence of different design decisions on their performance. Additionally, exploring the integration of theoretical principles into practical algorithm development opens up exciting avenues to enhance the capabilities of deep meta-learning approaches. In this dissertation, we aim to address these research gaps, so that we can move the field forward even further, paving the way for more robust and principled meta-learning techniques with broader applicability and superior performance.

1.3 Research questions

This dissertation is a collection of research articles that we have produced following this line of research. In this section, we give an overview of the research questions that we address as well as their contextual framework and relationships.

1.3.1 Stateless neural meta-learning

MAML (Finn et al., 2017) and the meta-learner LSTM (Ravi and Larochelle, 2017) are two popular deep meta-learning techniques. Interestingly, in writing the overview of deep meta-learning (see Chapter 2), we intuitively realized that the meta-learner LSTM is a slightly more expressive algorithm than MAML. In this dissertation, we mathematically prove this intuition: the meta-learner LSTM encompasses and can learn everything that MAML is capable of learning, and even more. In other words, the meta-learner LSTM subsumes MAML. Despite this fact, MAML consistently outperforms the meta-learner LSTM in various tasks in terms of the few-shot learning ability. The reasons behind this performance gap remain unknown and require further investigation.

Understanding the factors contributing to the superior performance of MAML compared to the meta-learner LSTM is of utmost importance. By unraveling the mechanisms responsible for this discrepancy, we can gain valuable insights into the inner workings of these meta-learning techniques. This research question prompts us to empirically explore potential factors such as architectural differences, optimization strategies, or inherent biases that might favor MAML's performance. Investigating these factors could shed light on the subtle nuances that influence the success or failure of deep meta-learning algorithms, leading to more informed design choices and ultimately driving improvements in the field. In short, we aim to answer the following research question.

RQ1: What causes the performance gap between MAML and the meta-learner LSTM?

1.3.2 Pre-training vs gradient-based meta-learning

A related method for improving the learning efficiency of deep neural networks compared with the deep meta-learning algorithms (MAML, meta-learner LSTM) investigated in the previous research question, is transfer learning by means of pre-training and finetuning. The idea of this method is to train a neural network to perform a given source task for which abundant data is available. This is in contrast to how deep meta-learning techniques are often trained for few-shot learning settings as they are trained on various tasks for which only a handful of examples are available to learn from. When presented with a new task, the finetuning approach transfers the parameters obtained by training on the source task to the target task and consequently fine-tunes these parameters on this target task. During finetuning, only the head of the network is adjusted whilst the body of the network is kept frozen. In the case of image classification, on which we focus in this dissertation, this corresponds to reusing the same feature detectors and only adapting how the different features are linearly combined to produce class scores.

Whilst the previous research question allowed us to gain a deep understanding of the differences between two deep meta-learning algorithms (MAML and the meta-learner LSTM), which have been observed to be successful at few-shot learning in various scenarios, recent results (Chen et al., 2019; Tian et al., 2020; Mangla et al., 2020) suggest that when evaluated on tasks from a different data distribution than the one used for training, the simple pre-training and finetuning baseline may be more effective than more complicated popular meta-learning techniques such as MAML and Reptile (Nichol et al., 2018). This is surprising as the learning behavior of MAML was shown to mimic that of finetuning: both rely on reusing learned features (Raghu et al., 2020). This begs the question of what causes the observed performance differences between these approaches, giving rise to the following research question.

RQ2: How do the learning behaviors of finetuning, MAML, and Reptile differ from each other and how does this influence their ability to quickly learn to perform new tasks?

1.3.3 LSTMs for few-shot learning

The study of research question 1 (RQ1) prompted us to look deeper into using an LSTM for meta-learning. More specifically, the meta-learner LSTM uses an LSTM at the *meta-level*: to perform weight updates for a base-learner network. In 2001, however, Hochreiter et al. (2001) showed that an LSTM that operates at the *data level* trained with backpropagation across different tasks is capable of meta-learning. The LSTM operating at the data level is fed an entire training set, and predictions for query inputs are then conditioned on the resulting hidden state. Despite the promising results of this approach on small problems, and more recently, also on reinforcement learning problems (Duan et al., 2016; Wang et al., 2016), the approach has received little attention in the supervised few-shot learning setting. In an earlier work that has tested the LSTM (Santoro et al., 2016), the performance was observed to be inferior compared with other meta-learning algorithms such as MAML and the meta-learner LSTM. This is surprising, as an LSTM at the data level is a maximally expressive meta-learning algorithm (Finn and Levine, 2018). We revisit this approach and test it on modern few-shot learning benchmarks. The research question we investigate is the following.

RQ3: Are LSTMs good few-shot learners when evaluated on modern benchmarks?

1.3.4 Subspace adaptation prior

The previous research questions have all focused on attempting to distill knowledge from empirical results. This begs the question of whether the opposite direction can also yield fruitful results, that is, whether the integration of machine learning knowledge can be used to enhance the few-shot learning capabilities of neural networks. Inspired by classical machine learning knowledge that a more expressive model (with more parameters) is more prone to overfitting than one with fewer parameters, we hypothesize that it may be beneficial for deep meta-learning methods to adjust only a subset of parameters rather than adapting *all* parameters of trainable layers when learning new tasks. More specifically, we hypothesize that it is beneficial to learn which subsets of parameters to adjust in every layer, rather than only the final layer as is common in the pre-training and finetuning strategy. The idea behind this hypothesis that simply adapting all paramaters neglects potentially more efficient learning strategies for a given task distribution and may be susceptible to overfitting, especially in few-shot learning where tasks must be learned from a limited number of examples. In short, we aim to answer the following research question.

RQ4: Can the few-shot learning ability of deep neural networks be improved by meta-learning which subsets of parameters to adjust?

1.4 Outline of the dissertation

This dissertation constitutes a collection of research papers and every chapter corresponds to one paper.

In Chapter 2, we survey the field of deep meta-learning. This serves as a detailed introduction and overview of the field. In this chapter, we provide the reader with the theoretical foundation for describing deep meta-learning algorithms, and we investigate and summarize key methods, which are categorized into i) metric-, ii) model-, and iii) optimization-based techniques. In addition, we identify the main open challenges, such as performance evaluations on heterogeneous benchmarks, and reduction of the computational costs of meta-learning. This chapter corresponds to the following published research article.

Huisman, M., van Rijn, J. N., & Plaat, A. (2021). A survey of deep meta-learning. Artificial Intelligence Review, 54(6), 4483-4541. Springer.

In Chapter 3, we investigate our first research question based on the observed performance gap between the meta-learner LSTM and MAML. We show that the reason for this surprising performance gap is related to second-order gradients. We construct a new algorithm (named TURTLE) to gain more insight into the importance of second-order gradients. TURTLE is simpler than the meta-learner LSTM yet more expressive than MAML and outperforms both techniques at few-shot sine wave regression and 50% of the tested image classification settings (without any additional hyperparameter tuning) and is competitive otherwise, at a computational cost that is comparable to second-order MAML. We find that second-order gradients also significantly increase the accuracy of the meta-learner LSTM. This chapter is based on the following published research article.

Huisman, M., Plaat, A., & van Rijn, J. N. (2022). Stateless neural meta-learning using second-order gradients. Machine Learning, 111(9), 3227-3244. Springer.

In Chapter 4, we investigate our second research question. More specifically, we investigate the observed performance differences between finetuning, MAML, and another meta-learning technique called Reptile, and show that MAML and Reptile specialize for fast adaptation in low-data regimes of similar data distribution as the one used for training. Our findings show that both the output layer and the noisy training conditions induced by data scarcity in the few-shot learning setting play important roles in facilitating this specialization for MAML. Lastly, we show that the pre-trained features as obtained by the finetuning baseline are more diverse and discriminative than those learned by MAML and Reptile. Due to this lack of diversity and distribution specialization, MAML and Reptile may fail to generalize to target tasks that are more distant to the observed training tasks whereas finetuning can fall back on the diversity of the learned features. This chapter is based on the following research articles.

Huisman, M., Plaat, A. & van Rijn, J. N. (2021). A preliminary study on the feature representations of transfer learning and gradient-based meta-learning techniques. In Fifth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems.

Huisman, M., Plaat, A. & van Rijn, J. N. (2023). Understanding Transfer Learning and Gradient-Based Meta-Learning Techniques. Accepted for publication in Machine Learning. Springer.

In Chapter 5, we investigate our third research question. That is, we revisit the classical LSTM approach to deep meta-learning and show that surprisingly, LSTM outperforms the popular meta-learning technique MAML on a simple few-shot sine wave regression benchmark, but that LSTM, expectedly, falls short on more complex few-shot image classification benchmarks. We identify two potential causes and propose a new method called *Outer Product LSTM (OP-LSTM)* that resolves these issues and displays substantial performance gains over the plain LSTM. Compared to popular meta-learning baselines, OP-LSTM yields competitive performance on within-domain few-shot image classification, and performs better in cross-domain settings by 0.5% to 1.9% in accuracy score. While these results alone do not set a new state-of-the-art, the advances of OP-LSTM are orthogonal to other advances in the field of meta-learning, yielding new insights in how LSTM work in image classification, allowing for a whole range of new research directions. This chapter is based on the following research article.

Huisman, M., Moerland, T. M., Plaat, A., & van Rijn, J. N. (2023). Are LSTMs good few-shot learners? Machine Learning, 112, 4635–4662. Springer.

In Chapter 6, we investigate our fourth and final research question. That is, we investigate whether the few-shot learning performance of neural networks can be improved by meta-learning which parameters to adjust. To investigate this, we propose *Subspace Adaptation Prior* (SAP), a novel gradient-based meta-learning algorithm that jointly learns good initialization parameters (prior knowledge) and layer-wise *parameter subspaces* in the form of operation subsets that should be adaptable. In this way, SAP can learn which operation subsets to adjust with gradient descent based on the underlying task distribution, simultaneously decreasing the risk of overfitting when learning new tasks. We demonstrate that this ability is helpful as SAP yields superior or competitive performance in few-shot image classification settings (gains between 0.1% and 3.9% in accuracy). Analysis of the learned subspaces demonstrates that low-dimensional operations often yield high activation strengths, indicating that they may be important for achieving good few-shot learning performance. This chapter is based on the following research article.

Huisman, M., Plaat, A., & van Rijn, J. N. (2023). Subspace Adaptation Prior for Few-Shot Learning. Machine Learning. Springer.

In the following chapter, we give a detailed overview of the field deep meta-learning, serving as a basis for the rest of this dissertation.