EXPLORING THE SIMILARITY OF REPRESENTATIONS IN MODEL-AGNOSTIC META-LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In past years model-agnostic meta-learning (MAML) has been one of the most promising approaches in meta-learning. It can be applied to different kinds of problems, e.g., reinforcement learning, but also shows good results on few-shot learning tasks. Besides their tremendous success in these tasks, it has still not been fully revealed yet, why it performs work so well. Recent work proposes that MAML rather reuses features than rapidly learns. To analyze MAML, we apply representation similarity analysis (RSA), a well-established method in neuroscience, to the few-shot learning instantiation of it. Although some part of our analysis supports their general results that feature reuse is predominant, we also reveal that the representations after inner gradient steps make a broader change to the representation than the changes during meta-training. This supports the opposite proposition that MAML rapidly learns.

1 INTRODUCTION

In recent years, meta-learning, also learning to learn (Thrun & Pratt (1998)), has aroused interest in different fields of machine learning research. Meta-learning tries to use past experiences of similar problems to acquire a prior over model parameters or the learning procedure. Several researchers have used meta-learning to tackle the problem of few-shot learning (e.g., Santoro et al. (2016); Ravi & Larochelle (2017); Finn et al. (2017)) where novel tasks are trained only by seeing a few examples of each possible class. The models have to adapt quickly to the new problem task. Traditional supervised models fail at that as they are not capable of generalizing on such a few examples and either not do learn the task or quickly overfit on the training data.

Some meta-learning algorithms use recurrent (Hochreiter et al. (2001) Santoro et al. (2016) Ravi & Larochelle (2017)) or attention-based (Vinyals et al. (2016) Mishra et al. (2018)) models trained via a meta-learning objective. These techniques attempt to encapsulate the learned learning procedure in the parameters of the neural net. An alternative approach has been proposed in Maclaurin et al. (2015) and Finn et al. (2017) where a bi-level meta-learning setup exists. The *inner* optimization represents an adaptation to a given task, whereas the *outer* objective is the meta-training objective. Especially model-agnostic meta-learning (MAML) has been successful, and numerous extensions on them have been proposed to speed it up and improve its effectiveness (e.g. Finn et al. (2017); Nichol et al. (2018); Rajeswaran et al. (2019)).

Despite their success, it is still unclear why and also how MAML learns so well. In Raghu et al. the question whether MAML successfully solves few-shot learning tasks mainly because of feature reuse or rapid learning. Feature reuse means that the meta-initialization is already so good that it does not have to be changed anymore significantly for the specific tasks. In contrast, rapid learning means that efficient significant changes are still done. By freezing a lot of layers and comparing the different layers by applying canonical correlation analysis (CCA), and centered kernel alignment (CKA) (Kornblith et al. (2019)), they come to the answer that MAML rather reuses features.

We analyze the representation of few-shot learning models during training with representation similarity analysis (RSA) (Kriegeskorte et al. (2008)) and reveal further insights into their representations. On the one hand, our analysis supports their answer; on the other hand, it also reveals that inner gradient steps lead to a more significant change in the representation than the updates in meta-training. Besides, we observe that most of the inner step optimization is done in the first step.

2 RELATED WORK

The method MAML that we investigate in this paper addresses the general problem of meta-learning (Thrun & Pratt (1998)), which includes few-shot learning. Next to recurrent-based (Hochreiter et al. (2001) Santoro et al. (2016) Ravi & Larochelle (2017)), and attention-based (Vinyals et al. (2016) Mishra et al. (2018)) meta-learning models, there also exist model-agnostic approaches (Maclaurin et al. (2015); Finn et al. (2017)). Model-agnostic meta-learning (Finn et al. (2017)) is a model-agnostic approach, which means it does not require a specific model. Since MAML requires second-order derivatives, there are several approaches that approximate the second-order derivatives, e.g., FOMAML (Finn et al. (2017)), Reptile (Nichol et al. (2018)), and iMAML (Rajeswaran et al. (2019)) to overcome the need of calculating them.

Representation similarity analysis (RSA) is widely used in computational neuroscience to compare a computational or behavioural model with the brain response (Kriegeskorte et al. (2008)). In Khaligh-Razavi & Kriegeskorte (2014), several (un)supervised vision models are used to show that supervised models are better for explaining IT cortical than unsupervised ones. In Cichy et al. (2017) the correlation of dynamics of the visual system is correlated with deep networks. RSA is also used in pure computational models (McClure & Kriegeskorte (2016); Dwivedi & Roig (2019); Mehrer et al.): In McClure & Kriegeskorte (2016) RSA is used as a loss function for knowledge distillation;Dwivedi & Roig (2019) uses it as a measure for the similarity between vision tasks, and Mehrer et al. analyze the consistency of neural networks with RSA. In Dwivedi et al. (2020), they propose a more general approach of comparing different activations, that RSA and CKA (Kornblith et al. (2019)) - which is used in Raghu et al. to analyze MAML - are instantiations of.

3 ANALYSIS OF REPRESENTATION IN FEW-SHOT LEARNING

MAML Finn et al. (2017) is a model-agnostic approach meaning it does not require a specific model. The idea of it is to learn an initialization of a model using exact second-order methods across tasks that are sampled from the same distribution. The optimization is done in a bi-level meta-learning setup. The meta-optimization is done in the outer loop and can be described as

$$\theta^* := \underset{\theta \in \Theta}{\operatorname{argmin}} = \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(in(\theta, \mathcal{D}_i^{tr}), \mathcal{D}_i^{test}) \tag{1}$$

The function $in(\theta, \mathcal{D}_i^{tr})$ describes the inner loop. The batch $\mathcal{D}_i^{tr}$ is initialized with θ and optimizes for one or a few training steps.

$$\phi_i = in(\theta, \mathcal{D}_i^{tr}) = \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_i^{tr}) \tag{2}$$

is an example of only using one training step.

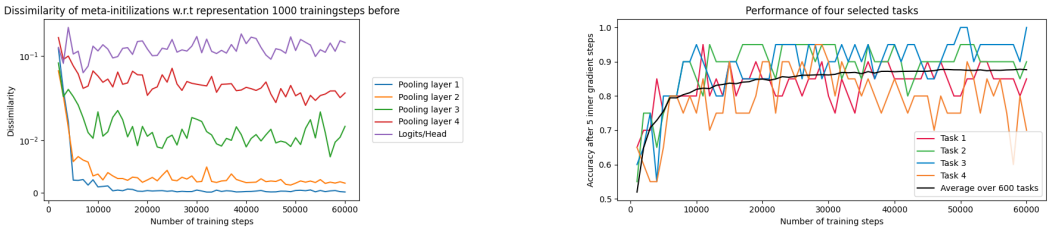
In this paper, we apply RSA to the representations which are learned by applying MAML to few-shot learning tasks. For every state of the model [1] in training, there exists an initial representation with θ as well as fine-tuned models, which are obtained after inner-loop updates ϕ_i on the support set of the meta-test tasks. We analyze different states of the meta-initialization, as well as the fine-tuned specific models for several few-shot learning tasks.

Traditionally, the meta-test task is analyzed by the performance of the query set. As this query set differs across every test task, but RSA needs the same inputs aligned, we hold out one specific evaluation task. To guarantee that a test task does not overlap partially with the evaluation task, none of the different test tasks shares input samples with the evaluation task. In addition, all of the meta-test tasks we analyze do not have any class in common with the classes from other meta-test tasks.

In our experiments, we use the same architecture which was proposed in Finn et al. (2017) which our implementation is based on[2]. In the analysis, we focus on 20-way 5-shot learning, but preliminary results of other problems showed similar behaviour.

[1] In this context, we mean by state of the model, the current initialization after training a specific amount of steps

[2] https://github.com/cbfinn/maml



(a) (b)

Figure 1: On (a), we see the dissimilarity of the meta-representations and the model's meta-representation 1000 training steps before per layer for training steps 2000 up to 60000. The dissimilarity score is computed by applying RSA on the two representations. Since RSA is returning the 1-spearman coefficient of the representations' gram matrix, 0 means dissimilar, and 1 means similar. On (b), we see the performance of MAML of four sampled tasks and optimized for five iterations in the inner-loop.

4 RESULTS

In this section, we show the results of multiple experiments. Figure 1a shows the dissimilarity developing during training for the four pooling layers and the final dense layer before applying the softmax function. The dissimilarity is denoted in log-scale. We compare the similarity of the representation to the representation obtained by the network 1000 training steps before. The application of RSA on the evaluation task to compare the representation returns a dissimilarity score. After several thousand training steps, we see that the dissimilarity has converged to a value and then only fluctuates around it for the rest of the training time. We can also observe that the earlier a layer in the network is, the more similar its representations are. The dissimilarity of representations from the first pooling layers is close to zero, whereas the last pooling layers' dissimilarity is fluctuating around about 0.1.

For further analysis, we sampled four tasks and evaluated the representation on a fifth one. In Figure 1b their performance after several training steps is shown. We observe they start fluctuating around a value after some thousand training steps increasing slightly. Also, the average performance across 600 tasks is rapidly increasing during the first 10000 iterations and then only slowly.

In Figure 2 we see the results of applying RSA on the representation before and after (1,5,10) inner loop gradient steps of the four selected tasks for every pooling layer and also for the final layers after 20000, 30000, 40000, 50000, and 60000 training steps. As seen in Figure 1b, the different models' performance is comparatively similar, although it is fluctuating quite a bit. First, we see that the dissimilarity scores in earlier pooling layers are smaller than in the later layers. This underlines the theory that fine-tuning benefits rather from reusing features than from quickly learning. However, we see that in every layer, all fine-tuned representations, even only after one inner step, are quite far away from their meta initialization by θ. In general, the learned representation in meta-optimization is closer to the representation 10000 training steps before than to the fine-tuned query tasks. This means that every fine-tuned optimization is rapidly adapting to the novel situation not only in the later layers but also in early pooling layers.

Although the representations after fine-tuning are similar to their starting representation, compared to the changes of updating during meta optimization (see 1a), the change is large. Interestingly the first inner step is responsible for most of the change in the representations. The representations after 5 or even 10 inner steps are not large compared to the one after 1 step[3]. This can be seen by looking at the dotted line, which connects the representation after 1, 5 and 10 inner steps.

Another interesting observation is that, especially in later layers, the same tasks staring from different meta-initializations are heading to a similar representation region that is more similar to themselves than to initialization. For the last layer, this makes sense as the loss function directly optimizes for it. For the intermediate layers before, there could be many other possible representations as we

[3]Note, that training was optimizing for 5 steps

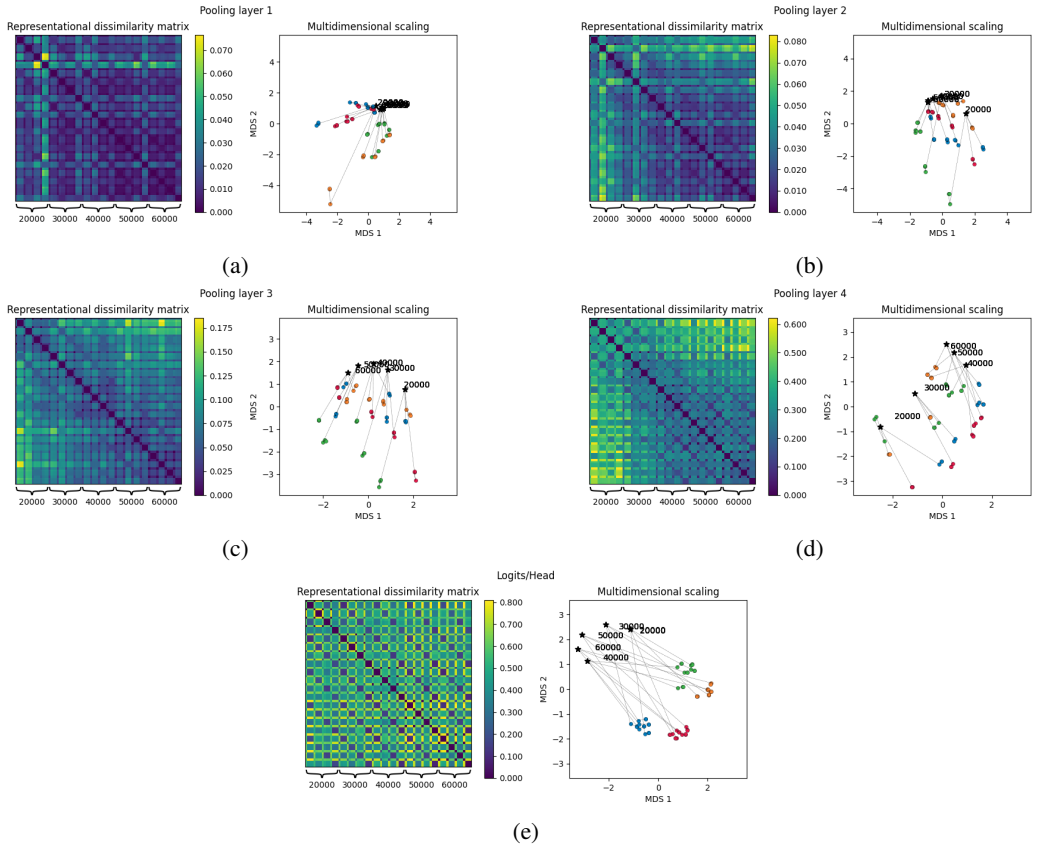


Figure 2: In this figure, the similarity of the representations obtained by applying an evaluation task on the MAML model after 20000, 30000, 40000, 50000, and 60000 training steps are shown. On the left, the representation similarity matrix (RDM) is visualized, whereas, on the right side, one can observe the multidimensional scaling (MDS). Every subplot shows the similarity of one specific layer. For every initialization θ, we also show the representation after updating for 1, 5, and 10 inner steps. A dotted line connects them.

still have to consider that we fine-tuned only on 20 data samples per task. This also leads to the result that inner optimization is valuable and leads to better representation than not training them.

5 Conclusion

In this paper, we applied RSA on both the initial but also the fine-tuned representation of MAML. On the one hand, our analysis confirmed that the representations in early layers do not change a lot compared to the ones in later layers, which underlines the theory that MAML rather reuses features instead of learning rapidly. However, we also showed that in early layers, the change of the representations in the inner-loop of meta-testing are significantly larger than in training the meta-initialization. This reveals that MAML partially still quickly learns also in early layers. Furthermore, most of the change in the representation seems to be done by only the first inner-step. In general, we think that the question of whether MAML reuses features or rapidly learns also depends on its definition. As there is no clear definition and measure for that, it is hard to agree finally.

Nevertheless, our analysis shows the potential of using RSA to analyze the representations of MAML. We think that more questions can be discussed in the future by looking at the representations of fine-tuned MAML models.

REFERENCES

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, and Aude Oliva. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153:346–358, 2017.

Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 12387–12396. Computer Vision Foundation / IEEE, 2019.

Kshitij Dwivedi, Jiahui Huang, Radoslaw Martin Cichy, and Gemma Roig. Duality diagram similarity: A generic framework for initialization selection in task transfer learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVI*, volume 12371 of *Lecture Notes in Computer Science*, pp. 497–513. Springer, 2020.

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, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1126–1135, 2017.

Sepp Hochreiter, A. Steven Younger, and Peter R. Conwell. Learning to learn using gradient descent. In *Artificial Neural Networks - ICANN 2001, International Conference Vienna, Austria, August 21-25, 2001 Proceedings*, pp. 87–94, 2001.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, 10, 11 2014.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Dougal Maclaurin, David Duvenaud, and Ryan P. Adams. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 2113–2122, 2015.

Patrick McClure and Nikolaus Kriegeskorte. Representational distance learning for deep neural networks. *Frontiers Comput. Neurosci.*, 10:131, 2016.

Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. Individual differences among deep neural network models. *Nature Communications*, (1):5725. doi: 10.1038/s41467-020-19632-w.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999, 2018.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. *CoRR*, abs/1909.04630, 2019.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy P. Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 1842–1850, 2016.

Sebastian Thrun and Lorien Y. Pratt. Learning to learn: Introduction and overview. In Sebastian Thrun and Lorien Y. Pratt (eds.), *Learning to Learn*, pp. 3–17. Springer, 1998.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3630–3638, 2016.