Research Project Proposal: Structured Meta-Learning for Heterogeneous Tasks

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

March 2020

1. Introduction to the problem

Machine Learning is an application of Artificial Intelligence that sets as its goal to develop methods that can detect patterns in data, and then use the uncovered patterns to predict future data or other outcomes of interest [20]. In recent years, a class of parametric techniques called Deep Learning has led to astonishing achievements in the field. The key aspect of these models is the ability to learn how to extract relevant features from data through many layers of general-purpose neural networks, thus not requiring field-specific expertise [17]. Nonetheless, Deep Learning currently faces some obstacles that still hinder the technology to be exploited in many application domains. Indeed, humans have a remarkable capacity to learn new concepts when provided with few examples; conversely, current popular deep learning techniques are data-hungry, needing thousands of samples to be able to generalize their knowledge and make predictions on unseen data.

Meta-Learning [4, 29], 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 [10]. The experience is often formalized as a collection of tasks, upon which meta-learning techniques build general, task-agnostic knowledge that can be reused. The approach has been shown to address some of the challenges posed by Few-Shot Learning, where very few task-specific training datapoints are available and the problem of overfitting is particularly insidious [18]. The literature has recently provided promising results also thanks to the leverage of deep learning techniques, achieving human-like performance also in simple meta-learning tasks [16].

Beyond its recent achievements, Meta-Learning itself currently faces many challenges. Many popular approaches struggle when scaling to more powerful learners, which constrains them to poor performance when dealing with complex tasks. Another challenge is the transfer of knowledge among tasks that are particularly different. Our brain builds powerful abstractions that can be used to identify an object, no matter how it is depicted, either as a natural image, a clip-art, or another visual representation. Conversely, a problem that has been observed to occur in many state-of-the-art meta-learning approaches is the inability to generalize the knowledge when presented with a heterogeneous distribution of domains.

As also analyzed in [31], we identified three partially overlapping issues that should be addressed to scale meta-learning. Firstly, most of the current meta-learning algorithms are designed considering the simplistic assumption that the distribution of tasks is *homogeneous*, namely that tasks are coming from a single source [3], and they share the same characteristics. In contrast, real-life learning experiences are *heterogeneous*: for instance, classification tasks may vary in terms of the number of classes or examples per class and are often unbalanced. Secondly, benchmarks in Meta-Learning only measure within-dataset generalization. However, we are interested in having models that can learn from multiple sources and generalize to entirely new distributions, namely new datasets or domains. Lastly, most of the current models and benchmarks ignore the relationships between tasks and classes, disregarding structures that could be useful to share knowledge across multiple tasks.

In light of these issues, our goal is to determine whether the currently provided techniques can be extended to better scale Meta-Learning with respect to data and task heterogeneity. A model capable of operating among different data domains would be able to transfer knowledge among widely different tasks, solving the lack of training samples that is observed in certain data domains. As a practical example, the 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 shades.

2. Main related works

The Machine Learning literature provides some approaches addressing the problems of cross-domain learning and heterogeneous task distribution.

Multi-Domain Supervised Learning [7, 24, 25] tries to provide a multi-domain feature extractor by leveraging deep neural layers that are parameterized ad hoc for each domain by an adaptation procedure. Guo et al. [11] provide a useful benchmark to test many state-of-the-art techniques on cross-domain few-shot learning tasks, showing their limits when confronted with data coming from different domains.

Tseng et al. [32] propose a learned augmentation procedure based on feature-wise transformation to obtain a training set that better represents the obstacles encountered in Cross-Domain Few-Shot Learning.

In the context of heterogeneous tasks, Triantafillou et al. [31] propose Proto-MAML, a model combining the simple inductive bias of Prototypical Networks and the flexible adaptation mechanism of Model-Agnostic Meta-Learning (MAML) [9], 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.

Multimodal MAML (MMAML) [33] extends MAML by providing a parameter initialization which depends 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. Conditional Neural Adaptive Processes (CNAPs) [26] 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.

Hierarchically Structured Meta-Learning (HSML) [34] 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. In order to address task heterogeneity, Yao et al. [35] propose a relational meta-learning (ARML) framework that automatically extracts the cross-task relations and constructs a meta-knowledge graph that is queried when a new task is presented. A much simpler line of work consists in designing and leveraging a good embedding to overcome limitations related to heterogeneous data. The approach is supported by Raghu et al. [23], arguing that feature reuse is the dominant component in MAML's efficacy, with Tian et al. [30] recently confirming this hypothesis.

3. Research plan

3.1. Goal

We argue about the importance of a structured embedded representation of the task by forcing a latent space decomposition into subspaces. The same decomposition should also be reflected in the model parameters and implemented to generate domain-specific weights. Finally, we should model the relationships across tasks to efficiently reuse knowledge.

We thus propose to extend Latent Embedding Optimization (LEO) [28] to address the above remarks and obtain state-of-the-art performance on heterogeneous domains and tasks. LEO is an optimization-based meta-learning approach that embeds the model parameters to perform gradient descent in a low-dimensional space. The embedding is obtained by enriching the architecture of the learner with a simple encoder-decoder structure. The encoder takes as input the training dataset of the task and provides a task-dependent initial representation, while the decoder processes the representation and generates the corresponding model parameters. We chose LEO as our base model because of its good properties, such as the use of embedding and the leverage of gradient-based learning techniques.

To achieve our goal, we add structure to the latent embedded representation in LEO by substituting the encoder-decoder architecture and feature extractor with the more interpretable component known as Domain Invariant Variational Autoencoders (DIVA) [14]. DIVA extends the standard variational autoencoder model by partitioning the latent space in three subspaces representing domain, class, and residual information of the input, with the final objective of learning a disentangled representation. An important advantage provided by introducing disentanglement is the high interpretability of the results, which may help us when drawing conclusions from our experiments. The division in subspaces of the embedding will also let us reuse and transfer the learned representation referring to a part of it. The advantage of disentangled representation in supervised and reinforcement learning has been theorized before [1, 5, 6, 12, 27], speculating on the suitability of a factorized representation for transfer learning and generalization. Recently the suitability of factorized representations for transfer learning has been empirically validated in few works [2, 13, 15]. We argue that disentanglement is a desirable property for meta-learning applications and this work aims to test this hypothesis. With the same goal, other adversarial architectures for Domain Adaptation and Generalization could be adopted as well [8, 22].

Considering the difficulty of the problem we face, we limit ourselves to work on classification tasks. The nature of our research will be mainly experimental, in line with most of the recent works produced in Machine Learning and Meta-Learning. We hope that the results of our approach will be later formalized in the literature to be better understood and built upon.

3.2. Research Activities

We provide below a list of the tasks in which the research is decomposed, along with a brief description for each of them.

1. *Literature review*: we reviewed relevant books and papers provided by the meta-learning literature to better understand the focus of the field, the latest advancements, and the main open issues, producing documents outlining State of the Art and our Project Proposal.

2. *Implementation*: we implement a framework for running experiments on our machinery and analyzing results, as well as several baseline architectures such as MAML, LEO, and DIVA using PyTorch [21].

3. *Observation*: we demonstrate and confirm the flaws of the simplest models we implemented when dealing with heterogeneous domains and tasks. To do so, we create a new dataset based on Omniglot [3] called Corrupted-Omniglot (C-Omniglot), adding 16 different perturbations as seen in C-MNIST [19].

4. *Experimentation*: we design and implement our proposed model in PyTorch combining LEO and DIVA. The model will then be tested on C-Omniglot and Meta-Dataset [31] and its classification accuracy compared to competitors in the field, such as MultiMAML [33] and HSML [34]. We would also like to understand whether disentanglement arises in the embedded representation we designed. Finally, we analyze the results to produce a detailed report.

5. *Paper*: we write a paper presenting our approach and the results we obtained. We submit the paper to a renowned Machine Learning conference such as the International Conference on Learning Representations (ICLR) or the AAAI Conference on Artificial Intelligence (AAAI) in September 2020.

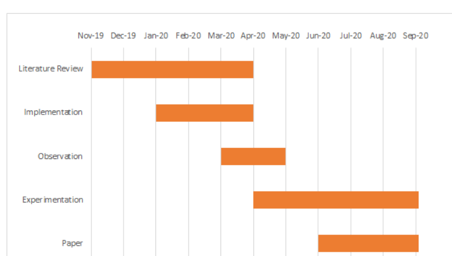

Figure 1: Gantt chart for the project

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[2] Alessandro Achille, Tom Eccles, Loic Matthey, Chris Burgess, Nicholas Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. In *Advances in Neural Information Processing Systems*, pages 9873–9883, 2018.

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

[4] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[5] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012.

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[7] Hakan Bilen and Andrea Vedaldi. Universal representations: The missing link between faces, text, planktons, and cat breeds. *arXiv preprint arXiv:1701.07275*, 2017.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *arXiv:1606.03657 [cs, stat]*, June 2016.

[9] 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.

[10] 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.

[11] Yunhui Guo, Noel CF Codella, Leonid Karlinsky, John R Smith, Tajana Rosing, and Rogerio Feris. A new benchmark for evaluation of cross-domain few-shot learning. *arXiv preprint arXiv:1912.07200*, 2019.

[12] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

[13] Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1480–1490. JMLR. org, 2017.

[14] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. *arXiv preprint arXiv:1905.10427*, 2019.

[15] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019.

[16] 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.

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

[18] 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.

[19] Norman Mu and Justin Gilmer. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*, 2019.

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

[21] 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`.

[22] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. 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 5102–5112, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/peng19b.html`.

[23] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

[24] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, pages 506–516, 2017.

[25] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018.

[26] 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.

[27] Karl Ridgeway. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*, 2016.

[28] 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`.

[29] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: The meta-meta-... hook*. Diplomarbeit, Technische Universität München, München, 1987.

[30] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.

[31] 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`.

[32] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv preprint arXiv:2001.08735*, 2020.

[33] 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.

[34] 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`.

[35] Huaxiu Yao, Xian Wu, Zhiqiang Tao, Yaliang Li, Bolin Ding, Ruirui Li, and Zhenhui Li. Automated relational meta-learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rklp93EtwH`.