Meta-Learning for Medical Image Classification

Shi Hu
Informatics Institute
University of Amsterdam
s.hu@uva.nl

Jakub Tomczak
Informatics Institute
University of Amsterdam
jakubmkt@gmail.com

Max Welling
Informatics Institute
University of Amsterdam
m.welling@uva.nl

Abstract

Most of the current state-of-the-art methods to classify medical images is to first train a deep model on ImageNet, then transfer all network weights to a new network except for the last softmax layer, and then fine-tune on the target dataset. When the amount of training data in the target dataset is sufficient, this method is able to surpass the level of a trained doctor on several datasets; however, when it is insufficient, which is common in a lot of real medical applications, this method may lead to mediocre results. To address the small dataset problem, we apply a meta-learning method to train, and then fine-tune on the target dataset. We show our results surpass the state-of-the-art method on a popular medical image dataset.

1 Introduction

Meta-learning (or learning to learn) has been an active research area in recent years due to its promise to be able to generalize well given limited amount of training data [1]. Meta-learning generally involves two components, i.e., the learner, which learns a new task, and the meta-learner, which trains the learner. There are many ways to make use of meta-learning in deep learning, and in this work, we are interested in using it to find an attractive initialization point, such that the subsequent fine-tuning step on the target dataset can achieve more accurate results than the current state-of-the-art method [2, 3].

Although meta-learning is an active research area in the machine learning community, it has not been widely applied in the medical imaging domain. One reason is that many current meta-learning approaches do not work well if there is a significant domain shift between meta training and meta test set. For example, Vinyals et al. show their meta-learning model is unable to work well if there is little concept overlap between the two sets [4].

2 Methodology

In recent years, deep learning [5] has been rapidly adopted in various real-world tasks, such as image recognition, speech recognition and autonomous driving, etc. One prerequisite for deep learning to perform well is to have a sufficient amount of training data. This amount should typically be on the order of hundreds of thousands. However, in many real-world domains, it is very expensive to annotate the object of interest, such as a medical image. For example, it can take up to minutes for a certified doctor to label the medical condition of one medical image. To address the small dataset problem, we apply a meta-learning method to obtain attractive initialization weights of a network. Then we transfer these weights except for the last softmax layer to a new network of the same architecture, and fine-tune on the target dataset. We achieve better classification results than the current state-of-the-art method.

The meta-learning model used in this work is called Reptile [6], which is closely related to but different from MAML [7]. The latter is one of the first meta-learning methods that aims to find

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

attractive initialization weights of a network for further fine-tuning. The former then simplifies this model and makes it more scalable. We use the same network architecture as in these works in pre-training (for baseline) and meta-learning (for the proposed model). In both cases, during fine-tuning we remove the batch normalization layer [8]. The reason is we found the batch normalization weights are less transferable when we tried to reproduce the top results in this Kaggle competition using the current state-of-the-art method with the Inception-v3 model [9] and ImageNet [10] for pre-training.

We train Reptile under the "5-shot 5-way" setting on mini-ImageNet [11] (the hyper-parameters for this setting are shown in Table 4 of their paper [6]). The mini-ImageNet dataset consists of 100 classes randomly selected from ImageNet, where each class has 600 images. The reason we use this dataset instead of the original ImageNet is because it is convenient for rapid prototyping and experiments.

For our proposed model, we transfer all weights trained by Reptile except for the last softmax layer to a new network, then fine-tune on the target dataset. For fine-tuning, we use the Adam optimizer [12] with its default parameter values: learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size of 32.

There are two baseline models. For baseline 1, we randomly initialize the network weights and then fine-tune on the target dataset. For baseline 2, we first train the network on mini-ImageNet dataset, and then fine-tune on the target dataset via transfer learning excluding the last softmax layer. The hyper-parameters used in both baselines are the same as the ones used in the fine-tuning stage of our proposed method.

All models are selected via early stopping. This includes the pre-training model for the baseline, the meta-learning model for the proposed method, and the fine-tuning models for both. We run every experiment for 200,000 iterations, and the models are evaluated every 1,000 iterations. Each experiment was run four times with four different random seeds.

3 Experiments

3.1 Data

We use the Kaggle's Diabetic Retinopathy Detection dataset [13]. According to its website's description, "[d]iabetic retinopathy is the leading cause of blindness in the working-age population of the developed world. It is estimated to affect over 93 million people." Hence, it is important to accurately classify the condition of a diabetic retinopathy (DR) screening, such that the patient can receive the appropriate treatment. The images in this dataset have 5 condition levels, namely, healthy, mild, moderate, severe or proliferative. The task is that given a DR screening, use the model to predict its condition. This dataset has 35,126 training images, and we randomly split these images into 80% training and 20% validation, and make sure the screenings of both eyes from the same patient falls in the same set.

3.2 Results

We use the quadratic weighted kappa (QWK) to evaluate the models, which is the same evaluation metric used in the Kaggle competition for this dataset. The reason for using this evaluation metric rather than the classification accuracy is because the dataset is highly unbalanced: there is one class which takes up more than 50% of the data. Thus, one can achieve high classification accuracy by predicting all labels to be the majority class. In contrast, QWK will assign the lowest score 0 to this case. Table 1 shows the fine-tuning results comparison on the Kaggle test set. We vary the amount of training and validation data by keeping 50% and 100% for each condition level. In other words, if we only use 50% of the training data, then we will also only use 50% of the validation data to simulate the "small dataset" scenario. We find our method performs the best in both cases, and the performance gain widens when we reduce the amount of training data. If we further reduce the amount of training data, the amount of validation data will then be very small, so early stopping will not be very effective for model selection.

Table 1: Comparison on fine-tuning test results

Percentage kept	Baseline 1	Baseline 2	Our method
50%	0.145 ± 0.021	0.101 ± 0.027	**0.161 ± 0.028**
100%	0.165 ± 0.040	0.171 ± 0.028	**0.175 ± 0.028**

4 Conclusion

In this short paper we present a simple experiment to demonstrate how to use meta-learning to learn network initialization weights for fine-tuning on a medical image dataset. We achieve better classification performance than the current state-of-the-art method. The results are interesting because the medical image dataset is very different than the natural images presented in the ImageNet, and the between-class differences are much smaller for medical images than natural images, which increases the difficulty for correct classification. For future work, we would like to scale up the meta-learning algorithm such that it can use advanced models, such that Inception-v4 [14], as well as a bigger training set, such as ImageNet.

Acknowledgments

We thank Chongxuan Li for useful discussions and Prof. Bram van Ginneken for introducing the Kaggle dataset. SH was funded by the NWO (Grant DLMedIa). JMT was funded by the European Commission within the MSC-IF (Grant No. 702666).

References

[1] Sebastian Thrun and Lorien Pratt. Learning to learn. *Springer Science & Business Media*, 1998.

[2] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *The Journal of the American Medical Association*, 2016.

[3] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017.

[4] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. *NIPS*, 2016.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

[6] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv*, 2018.

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 2017.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. *ICML*, 2015.

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *CVPR*, 2016.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009.

[11] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *ICLR*, 2017.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[13] Diabetic retinopathy detection. `https://www.kaggle.com/c/diabetic-retinopathy-detection`. Accessed: 2018-04-03.

[14] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *AAAI*, 2017.