Training Medical Image Analysis Systems like Radiologists

Gabriel Maicas[1]([✉]), Andrew P. Bradley[2], Jacinto C. Nascimento[3], Ian Reid[1], and Gustavo Carneiro[1]

[1] Australian Institute for Machine Learning, School of Computer Science, The University of Adelaide, Adelaide, Australia
`gabriel.maicas@adelaide.edu.au`
[2] Science and Engineering Faculty, Queensland University of Technology, Brisbane, Australia
[3] Institute for Systems and Robotics, Instituto Superior Tecnico, Lisbon, Portugal

Abstract. The training of medical image analysis systems using machine learning approaches follows a common script: collect and annotate a large dataset, train the classifier on the training set, and test it on a hold-out test set. This process bears no direct resemblance with radiologist training, which is based on solving a series of tasks of increasing difficulty, where each task involves the use of significantly smaller datasets than those used in machine learning. In this paper, we propose a novel training approach inspired by how radiologists are trained. In particular, we explore the use of meta-training that models a classifier based on a series of tasks. Tasks are selected using teacher-student curriculum learning, where each task consists of simple classification problems containing small training sets. We hypothesize that our proposed meta-training approach can be used to pre-train medical image analysis models. This hypothesis is tested on the automatic breast screening classification from DCE-MRI trained with weakly labeled datasets. The classification performance achieved by our approach is shown to be the best in the field for that application, compared to state of art baseline approaches: DenseNet, multiple instance learning and multi-task learning.

Keywords: Meta-learning · Curriculum learning · Multi-task training Breast image analysis · Breast screening · Magnetic resonance imaging

1 Introduction

Radiologists are exceptionally trained specialists who play a crucial role interpreting and assisting other doctors and specialists in diagnosing and treating diseases. Their training program typically requires the trainee to solve tasks

of increasing difficulty [1], where each task contains a relatively small number of "training images". Such a program bears little resemblance to the training of medical image analysis systems based on machine learning that are modeled to solve narrowly defined, but complex classification problems [2], requiring large training sets. Once trained, these models cannot be easily adapted to new problems – they must be re-trained with new large training sets. The use of pre-trained models [3] as a way of initializing a model is the first step towards a more similar approach to the training program of radiologists. However, pre-training does not train a model to be able to learn new tasks – instead it is a "trick" to improve convergence and generalization. Meanwhile, machine learning researchers have developed more effective *learning to learn* approaches [4] – such approaches are motivated by the ability of humans to learn new tasks quickly and with limited "training sets". The optimization in such approaches penalizes classification loss and inefficient learning on new tasks (i.e., classification problems) by using a training scheme that continuously samples new tasks, mimicking the human training process. Our hypothesis is that medical machine learning methods could benefit from such an radiologist's style training process.

In this paper, we introduce an improved model agnostic meta-learning [4] (MAML) as a way of pre-training a classifier. The training process maximizes the ability of the classifier to adapt to new tasks using relatively small training sets. We also propose a technical innovation for MAML [4], by replacing the random task selection with teacher-student curriculum learning as an improved way for selecting tasks [5]. This task selection process is based on the model's performance on the tasks, trying to mimic radiologists' training. Our improved MAML is tested on weakly-supervised breast screening from DCE-MRI, where samples are globally annotated with classes (i.e. volume-level labels): *no findings*, *benign lesions* and *malignant lesions*, but these samples do not have lesion delineations. Note that the use of weakly-labeled datasets is becoming increasingly important for medical image analysis as this is the data available in clinical practice [2].

We test our proposed approach on a dataset of dynamic contrast enhanced MRI for the breast screening classification. Results show that our proposed approach improves the area under the ROC curve (AUC), outperforming baselines such as DenseNet [6], which holds the state-of-the-art (SOTA) for many classification problems; multiple-instance learning [7], which holds SOTA for breast screening in mammography; and multi-task learning [8]. Our learning approach produces an AUC of 0.90, which is better than the best result from the baseline methods that achieves an AUC of 0.85.

2   Literature Review

Breast screening from DCE-MRI aims at early detection of breast cancer in women at high-risk [9]. Currently, this screening process is mostly done manually, where its success depends on the radiologist's abilities [10]. An automated breast screening system working as a second reader can help radiologists reduce

variability and increase the sensitivity and specificity of their readings. Traditionally, such systems rely on classifiers trained with large-scale strongly labeled datasets (i.e., containing lesion delineation and global classification) [11–15]. The non-scalability of this process (due to costs related to the annotation process) motivated the development of learning methods that can use weakly-labeled training sets [7] (i.e., samples contain only global classification). However, these methods still follow traditional machine learning approaches, which means that they still need large-scale training sets, even when the model has been pre-trained from other classification problems [3].

Contrasting with traditional machine learning algorithms, humans excel at learning new skills and new "classification" problems, where new learning tasks often require fewer training samples than the ones before. This *learning to learn* ability has inspired the development of a new generation of machine learning algorithms. For example, multi-task learning uses an optimization function that is trained to simultaneously minimize the loss of several different, but related classification problems [8], helping the regularization of the training procedure. Nevertheless, multi-task learning does not address the issue of making a model effective at learning new classification problems with small datasets. This issue is addressed by *meta-learning* [4], which has been designed to solve the *few-shot* learning problem, where the classifier is trained to train for new classification problems with previously unseen classes containing a small number of images. In meta-learning for few-shot classification, the model is *meta-trained* to solve classification problems for many randomly sampled tasks (i.e., the tasks are not fixed as in multi-task learning). Then the model is *meta-tested* by classifying unseen classes after being able to adapt using *few* training images of such unseen classes.

We explore the potential to improve the meta-learning process using a more useful (i.e., non random) task sampling procedure. For example, formulating the task sampling as a multi-armed bandit problem has been shown to produce faster convergence and better generalization [16]. Similarly, Matiisen *et al.* [5] proposed a new form of curriculum learning [17] that selects new tasks based not on their performance but on their performance improvement. However, these task sampling approaches have been applied in traditional machine learning problems, such as supervised and reinforcement learning problems, which means that our proposed application of curriculum learning for task selection in meta-learning is novel, to the best of our knowledge[1].

3   Methodology

Our methodology consists of three stages (see Fig. 1). We first **meta-train** the model using different tasks (each containing relatively small training sets) to find

---

[1] While writing the final draft of this paper, we noticed a recent approach by Sharma et al. [18]. However, they sample tasks for the problem of multi-task learning. In addition, sampling tasks is not based on the improvement of performance, but tasks where the performance is worse.
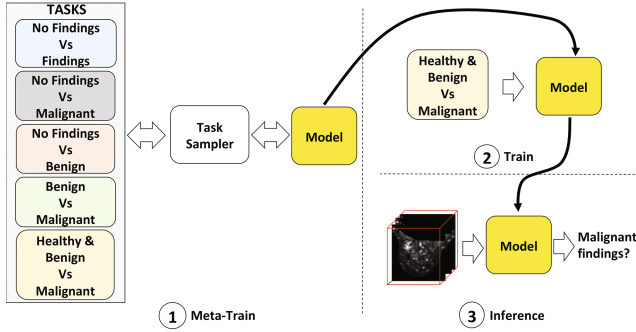
Fig. 1. The model is first meta-trained using several tasks containing relatively small training sets. The meta-trained model is then used to initialize the usual training process for breast screening (i.e., healthy and benign versus malignant). The probability of malignancy is estimated from a forward pass during the inference process.

a good initialization that is then used to **train** the model for the breast screening task (i.e., the healthy and benign versus malignant task). The **inference** is performed using previously unseen test data. Below, we define the dataset and describe each stage.

3.1 Dataset

Let the dataset be represented by $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{t}_i, b_i, d_i, y_i)\}_{i=1}^{|\mathcal{D}|}$ where $\mathbf{v} : \Omega \to \mathbb{R}$ is the first subtraction DCE-MRI volume (Ω denotes the volume lattice), $\mathbf{t} : \Omega \to \mathbb{R}$ is the T1-weighted volume, $b \in \{\text{left}, \text{right}\}$ indicates if this is the left or right breast of the patient, $d_i \in \mathbb{N}$ denotes patient identification, and $y \in \mathcal{Y} = \{0, 1, 2\}$ is the volume label ($y_i = 2$: breast contains a malignant lesion, $y_i = 1$: breast contains at least one benign and no malignant findings, and $y_i = 0$: no findings). We divide $\mathcal{D}$ using the patient identification into the training set $\mathcal{T}$, validation set $\mathcal{V}$ and testing set $\mathcal{S}$, with no overlap between these sets.

For the meta-training phase, we use the meta-training set defined by $\{\mathcal{D}_j\}_{j=1}^5$ where each meta-set $\mathcal{D}_j \subseteq \mathcal{T}$ contains the relevant volumes for the classification task K_j, defined as follows: (1) K_1 classifies volumes that contain any findings (benign or malignant); (2) K_2 discriminates between volumes with no findings and malignant findings; (3) K_3 discriminates between volumes with no findings and benign findings; (4) K_4 discriminates volumes with benign findings against malignant findings; and (5) K_5 addresses breast screening, i.e. finding volumes that contain malignant findings.

3.2 Model

We **meta-train** a model across a number of tasks so that it can be quickly trained to new unseen tasks from few images, or fine-tuned to become more effective at one of the tasks used in the meta-training phase. See algorithm 1 for an overview of the methodology.

Algorithm 1 Overview of the meta-training procedure

1: **procedure** META-TRAIN($\{K_1 \ldots K_5\}$, $\{\mathcal{D}_1 \ldots \mathcal{D}_5\}$, model f_θ)
2:     Initialise model parameters θ
3:     **for** $m = 1$ *to* M **do**   ▷ Meta-update Loop
4:         **Create** meta-batch $\mathcal{K}_m$ by sampling $|\mathcal{K}_m|$ tasks from $\{K_1 \ldots K_5\}$
5:         **for** each task $K_j \in \mathcal{K}_m$ **do**
6:             **Adapt** model with (1) using samples from $\mathcal{D}_j$ ▷ Adaptation
7:             **Update** model parameters with (2)   ▷ Meta-update

---

Let f_θ be the model parameterized by θ. For each meta update, the model adapts to the multiple tasks using the meta-batch set $\mathcal{K}_m$. The tasks included in $\mathcal{K}_m$ are sampled according to one of the methods described below in Sect. 3.3. For each task $K_j \in \mathcal{K}_m$, we sample from $\mathcal{D}_j$ a training set $\mathcal{D}_j^{tr}$ and a validation set $\mathcal{D}_j^{val}$ with N^{tr} and N^{val} volumes, respectively. The model parameter θ **adaptation** is performed with the following gradient descent at time step t:

$$\theta_j'^{(t)} = \theta^{(t)} - \alpha \frac{\partial \mathcal{L}_{K_j}\left(f_{\theta^{(t)}}\left(\mathcal{D}_j^{tr}\right)\right)}{\partial \theta}, \tag{1}$$

where α denotes the adaptation learning rate, and $\mathcal{L}_{K_j}\left(f_\theta\left(\mathcal{D}_j^{tr}\right)\right)$ is the cross-entropy loss to train for the classification task K_j. Finally, given the adapted models $f_{\theta_j'^{(t)}}$ for each task $K_j \in \mathcal{K}_m$, the model parameter θ is **meta-updated** from the error on the validation volumes $\mathcal{D}_j^{val}$ of the task w.r.t. the initial parameters $\theta^{(t)}$:

$$\theta^{(t+1)} = \theta^{(t)} - \beta \sum_{K_j \in \mathcal{K}_m} \frac{\partial \mathcal{L}_{K_j}\left(f_{\theta_j'^{(t)}}\left(\mathcal{D}_j^{val}\right)\right)}{\partial \theta}, \tag{2}$$

where β denotes the meta-learning rate. In summary, the **meta-training** phase consists of updating the parameters of the model based on the error in validation images after being *adapted* to a task using few images. This is equivalent to the following optimization:

$$\min_\theta \sum_{K_j \in \mathcal{K}_m} \mathcal{L}_{K_j} f_{\theta_j'^{(t)}}(\mathcal{D}_j^{val}) = \min_\theta \sum_{K_j \in \mathcal{K}_m} \mathcal{L}_{K_j}\left(f_{\theta^{(t)} - \alpha \frac{\partial \mathcal{L}_{K_j}\left(f_{\theta^{(t)}}\left(\mathcal{D}_j^{tr}\right)\right)}{\partial \theta}}(\mathcal{D}_j^{val})\right) \tag{3}$$

The resulting model f_θ obtained after the completion of the meta-training process is then fine-tuned using the cross entropy loss for the breast screening binary classification problem. This process consists of the **training phase**, where we use the training set $\mathcal{T}$ for training and validation set $\mathcal{V}$ for model selection. The final model is tested during the **inference phase** by feeding testing volumes from $\mathcal{S}$ through the network to estimate their probability of malignancy.

3.3   Task Sampling

The sampling process to select $|\mathcal{K}|$ tasks from $\bigcup_{j=1}^{5} K_j$ (step 4 of Algoritjm 1) is currently based on random sampling [4]. However, we consider this to be a crucial step in that algorithm, and therefore propose four sampling methods for step 4 of Algorithm 1. In particular, we study the following sampling methods: (1) **Random:** randomly sample all tasks with replacement [4]; (2) **All-task:** sample all $|\mathcal{K}| = 5$ tasks exactly once; (3) **Teacher-Student Curriculum Learning (CL)** [5]: sample tasks that can achieve a higher improvement on their performance. This is formalized by a partially observable Markov decision process (POMDP) parametrized by the *state*, which is the current parameter vector $\theta^{(t)}$; the next *action* to perform, which is the task K_j to train on; the *observation* O_{K_j}, consisting of the AUC improvement after adapting the parameters from $\theta^{(t)}$ to $\theta'^{(t)}$ for task K_j; and the *reward* R_{K_j}, which is computed from the AUC improvement of the current observation O_{K_j} minus the AUC improvement obtained from the last time the task K_j was sampled. The goal of the sampling algorithm is to maximize the score of all tasks, which is solved based on reinforcement learning using Thompson sampling. More specifically, a buffer $\mathcal{B}_j$ stores the last B rewards for task K_j, and at sampling time, a *recent reward* is randomly chosen from each of the buffers $\mathcal{B}_j$. The next task for the meta-training is the one associated with the buffer that produced the highest absolute valued *recent reward*. This procedure chooses to lean a task until its improvement stabilizes, and then different tasks will be sampled and so on. Note that by sampling according to the absolute value, tasks where the performance is decreasing will tend to be sampled again; and (4) **Multi-armed bandit (MAB)** [16]: sample in the same way as the CL approach above, but the observation O_{K_j} is stored in the buffer instead of the reward R_{K_j}. Also, the next task is selected based on the highest valued *recent observation* (not its absolute value).

4   Experiments and Results

We assess our methodology on a breast DCE-MRI dataset containing 117 patients, divided into a training set with 45 patients, a validation with 13 and a test set with 59 patients [15,19]. Each sample for each patient in this dataset contains T1-weighted and dynamically-contrast enhanced MRI volumes. Given the current interest in decreasing the number of scans [12,15], only the first subtraction volume is used. Although all patients contain at least one lesion (benign or malignant, confirmed by biopsy), not all breasts contain lesions. The T1-weighted volume is only used to automatically segment and extract the left and right breasts into volumes of size $100 \times 100 \times 50$ [15] and assign separate labels to them, where the label of a breast can be "no-finding","malignant" (if it contains at least one malignant lesion), or "benign" (if all lesions are benign). All evaluations below are based on the area under the ROC curve (AUC).

The model f_θ, implemented in 3D, is based on the DenseNet [6], which currently holds the best classification performance in several computer vision applications. The model architecture and hyper-parameters are selected based on the

**Table 1.** Baseline AUC for classifiers trained on breast screening.

Baseline	AUC
DenseNet [6]	0.83
MIL [7]	0.85
Multi-task [8]	0.85

Table 2. AUC for our proposed models depending on the meta-batch size and task sampling methods and trained for breast screening.

Model	$|\mathcal{K}|$	AUC per sampling method			
		Random	All-task	MAB	CL
BSML	3	0.86	N/A	0.88	**0.90**
BSML	5	0.85	0.89	0.89	**0.90**
BSML-NS	4	0.85	0.88	0.87	0.89

highest AUC for the breast screening problem in the validation set. The architecture is composed of five dense blocks of two dense layers each and is trained with a learning rate of 0.01 and a batch size of 2 volumes. For our proposed methodology (labeled as BSML), the number of meta-updates is $M = 3000$, the meta-learning rate $\beta = 0.001$, the number of training and validation volumes selected for task K_j from the meta-set $\mathcal{D}_j$ is $N^{tr} = N^{val} = 4$, the number of gradient descent updates is 5, and the adaptation learning rate $\alpha = 0.1$. We check the influence of the meta-batch size $|\mathcal{K}| \in \{3, 5\}$. Also, we evaluate the influence of all task sampling approaches listed in Sect. 3.3. Finally, we also run experiments to check the performance of our model when the task of breast screening is not used for meta-training (BSML-NS). This means that the training process has to learn an unseen task starting from the initialization achieved in the meta-training step. In this case, we use $|\mathcal{K}| = 4$ and test the influence of the different task sampling approaches.

Our proposed model is compared against the following baselines: (1) a *DenseNet* trained for the breast screening binary task; (2) the pre-trained DenseNet (1) fine-tuned using a multiple-instance learning framework *(MIL)* [7] – this approach holds the SOTA for the breast screening problem in mammography; and (3) a DenseNet trained with a *multi-task* loss [8] using the 5 tasks defined in Sect 3.1.

| (a) | (b) | (c) | (d) |
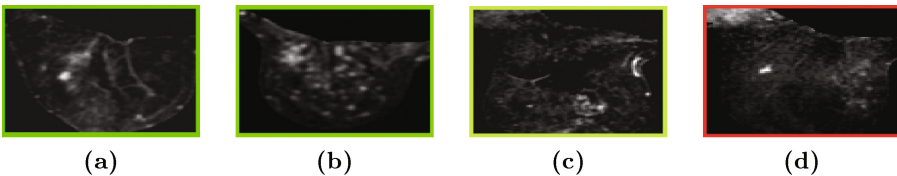
Fig. 2. Classification examples. Image (2a) shows a correct negative classification of a volume containing a benign lesion, images (2b) and (2c) show a correct positive classification of a volume containing a malignant lesion, and image (2d) shows an false negative classification of a volume containing a small malignant lesion.

Tables 1 and 2 contain the AUC for baselines and experiments detailed above. Figure 2 shows examples of the classification produced by our methodology.

5   Discussion and Conclusion

We presented a methodology to train medical image analysis systems that tries to mimic the process of training a radiologist. This is achieved by meta-training the model with several tasks containing small meta-training sets, followed by a subsequent training to solve the particular problem of interest. We established a new SOTA for the weakly supervised breast screening problem when compared to several baselines such as DenseNet [6], a multi-task trained DenseNet [8] and a DenseNet fine-tunned in a MIL framework [7]. Note that the MIL setup does not achieve a large improvement as reported in the original paper [7]. We believe that this is due to the use of DenseNet, which tends to show better classification results than Alexnet [7]. Also, it is worth mentioning that our proposed method has not shown any false positive classification in the test set.

As reflected in the experiments, the sampling of the tasks to meta-train is an important step of our proposed methodology. In particular, the CL sampling showed more accurate classification than random sampling, which yields similar results to the baselines. The MAB sampling improved over random selection, but it is still not as competitive as curriculum learning. We conjecture that sampling according to the best performance (i.e., MAB) keeps selecting more often the tasks that produce the highest reward, while CL samples tasks with a larger margin for improvement because they can achieve a larger slope in the learning curve. Consequently, CL aims at improving the reward for ALL tasks. Also, the meta-batch size does not appear to have much influence in the results. Furthermore, the BLML-NS results in Table 2 show that our proposed methodology can be successfully trained for breast screening even when this task is not included in the meta-training phase. In particular, notice that the AUC is competitive, being 1 point smaller than our best result (that includes breast screening in meta-training), but between 4 and 6 points better than the baselines.

References

1. The Royal Australian and New Zealand College of Radiologists: Training in Clinical Radiology (2009)
2. Wang, X., Peng, Y., et al.: Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR (2017)
3. Bar, Y., Diamant, I., et al.: Deep learning with non-medical training used for chest pathology identification. In: Medical Imaging: Computer-Aided Diagnosis (2015)
4. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
5. Matiisen, T., Oliver, A., Cohen, T., Schulman, J.: Teacher-student curriculum learning. arXiv preprint arXiv:1707.00183 (2017)

6. Huang, G., Liu, Z.: Densely connected convolutional networks. In: CVPR (2017)
7. Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 603–611. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_69
8. Xue, W., Brahm, G.: Full left ventricle quantification via deep multitask relationships learning. Med. Image Anal. **43**, 54–65 (2018)
9. Smith, R.A., Andrews, K.S.: Cancer screening in the United States, 2017: a review of current American cancer society guidelines and current issues in cancer screening. CA Cancer J. Clin. **67**, 100–121 (2017)
10. Vreemann, S., Gubern-Merida, A.: The frequency of missed breast cancers in women participating in a high-risk MRI screening program. Breast Cancer Res. Treat. **169**, 323–331 (2018)
11. Gubern-Mérida, A., Martí, R.: Automated localization of breast cancer in DCE-MRI. Med. Image Anal. **20**, 265–274 (2015)
12. Dalmış, M.U., Vreemann, S.: Fully automated detection of breast cancer in screening MRI using convolutional neural networks. J. Med. Imaging **5**, 014502 (2018)
13. Amit, G., et al.: Hybrid mass detection in breast MRI combining unsupervised saliency analysis and deep learning. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 594–602. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_68
14. Jäger, P.F., et al.: Revealing hidden potentials of the q-Space signal in breast cancer. In: Descoteaux, M., et al. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 664–671. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_76
15. Maicas, G., Carneiro, G., Bradley, A.P., Nascimento, J.C., Reid, I.: Deep reinforcement learning for active breast lesion detection from DCE-MRI. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 665–673. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_76
16. Gutiérrez, B., Peter, L., Klein, T., Wachinger, C.: A multi-armed bandit to smartly select a training set from big medical data. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 38–45. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_5
17. Bengio, Y., Louradour, J., et al.: Curriculum learning. In: ICML (2009)
18. Sharma, S., Jha, A.K., Hedge, P., Ravindran, B.: Learning to multi-task by active sampling. In: ICLR (2018)
19. McClymont, D., Mehnert, A., et al.: Fully automatic lesion segmentation in breast MRI using mean-shift and graph-cuts on a region adjacency graph. In: JMRI (2014)