Meta-learning for skin cancer detection using Deep Learning techniques

Anonymous Author(s)
Affiliation
Address
email

Abstract

This study focuses on the automatic skin cancer detection using a Meta-learning approach for dermoscopic images. The aim of this study is to explore the benefits of the generalization of the knowledge of unrelated domains in the classification performance of medical data.

In this study, a small sample of a combined dataset from 3 different sources was used to fine-tune a ResNet model pre-trained on non-medical data. The results show an increase of performance on detecting melanoma, malignant (skin cancer) and benign moles with the prior knowledge on everyday objects from ImageNet dataset by 20 points.

These findings suggest that features from unrelated domains can be used towards the classification of skin moles and that the distribution of the data affects the performance of the model.

1 Introduction

Deep learning is a technique that has been widely used in the fields of image classification and object detection due to its ability to learn highly complex patterns from vast amounts of data with reliable performance. One of the greatest challenges in the application of Deep Learning in medical domains is the limited availability of labelled data. This is mainly because it represents a considerable cost and time for a professional clinician to assess the medical condition of the patient.

One of the techniques used to deal with that limitation is Data augmentation, which consists in generating artificial examples from the target data, although this does not overcome the problem (Wong et al., 2016). Another technique which has been widely used is Transfer learning. In this approach, the network is first trained with data from a related domain, the network weights are then transferred to a new network and fine-tuned with the target data. The drawback of this method is that it presents poor performance when the amount of the target data is too small or presents a slight distribution shift in the target data (Soekhoe et al., 2016).

Despite that meta-learning is not a new concept in the deep learning community, it has not been widely explored for the application in the medical imaging domain. In this study, an exploration of a meta-learning model is reviewed, emphasizing in the dataset bias and distribution shift, factors that have been extensively addressed in image recognition problems, but have not received a significant importance for applications on medical imaging data.

1.1 Problem overview

Melanoma is the fifth most common cancer in the UK. Since 1990, its incidence has increased 134% in the UK. It is estimated that around 2,400 deaths are caused by melanoma each year in the UK.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

However, the survival rate is 90% when diagnosed on time. This suppose a relevant problem that can be solved with deep learning. The main challenge in the application of deep learning in the medical domain is the scarcity of the data. This limitation can be addressed with meta-learning approaches that generalize the knowledge of unrelated domains to help on the prediction of the task.

Since the recent advancements in the development of meta-learning frameworks, meta-learning has started to attract researchers from the medical imaging community. A first exploration of the viability of using meta-learning approaches in medical domains is explored in this study.

1.2 Meta-learning concepts and definitions

The term Meta-learning, also known as "Learning to Learn" has its origins in educational psychology, where it is described as the adaptation of the learning process according to the requirements of a specific task (BIGGS, 1985; Lemke et al., 2013). The goal of Meta-learning is to understand the learning process and exploit the acquired knowledge to improve the effectiveness of learning new tasks. In the context of artificial intelligence systems, a task can be a regression, object detection or classification problem, among all.

The knowledge derived from the learning mechanism is called Meta-knowledge, and the knowledge about which attributes are relevant to perform a task is called Metadata. A Meta-learning system uses the acquired meta-knowledge to derive the appropriate strategy to learn on a new domain of application, in other words, meta-learning is the type of learning that uses prior experience of other tasks to adapt to learn the new.

2 Related Work

The increasing popularity of deep learning in object recognition tasks is primarily due to the availability of training data. However, in medical domains, a scarce of data is faced, which have put a research direction in using data from unrelated domains to diagnose diseases from medical imaging data.

(Cheplygina et al., 2017) evaluated whether a classifier is able to predict which classification problem a dataset is sampled from, based on the performance of 6 different classifiers and 120 datasets from 6 different classification problem. Their findings shows that different datasets from the same task share similar properties, such as dataset size, type of images, number of classes and type of features. Despite the limitations of their approach, this simple method was able to extract features that are shared among datasets and find clusters. This demonstrates that there must be some intrinsic characteristics not only at the meta-level, but also between the samples of the datasets of similar tasks.

(Schlegl et al., 2014) developed a convolutional neural network to classify pathologies in high-resolution Computed Tomography (CT) scans of lung tissue with partial annotations obtained from different sources. The authors report an improvement in the classification performance with the model pre-trained with natural images compared to the pre-training on medical data of a different task. Their results proves that the data from similar domains does not necessarily lead to an improvement in the learning, but instead, is the variability of the input data, such as colours, textures, shapes and angles, that the model will benefit from to generalize on new data.

(Hu et al., 2018) implemented Reptile, a state-of-the-art meta-learning model pre-trained with mini-ImageNet to detect diabetic retinopathy. Their results show a slightly increment in the performance of the meta-learning model on detecting the target class as compared to the baseline model with no pre-training. This demonstrates that the amount of training data does not correlates with the classification performance, in other words, it is possible to obtain a good generalization from a small dataset size.

(Esteva et al., 2017) implemented a GoogleNet Inception V3 model pre-trained on ImageNet. For this study, the authors designed a partitioning algorithm to balance the classes in the training set to avoid in-class bias. Their reports indicate that the transfer learning model is able to match the performance of the dermatologists on the critical diagnostic tasks. The performance of the model increases when the training data is balanced, which is an indicator of the presence of bias. This is studied in detail in (Dietterich and Kong, 1995), where the authors present a model for domain adaptation to overcome the covariate shift, by learning a representation of the data that takes into account the distribution

Table 1: Class distribution of the combined dataset

Class	Number of Samples
Benign	132 images
Malignant	15 images
Melanoma	46 images
Total	193 images

shift between the test and training data. This problem was also addressed by (Ashraf et al., 2018) who present a method to handle dataset bias for medical image data by unlearning the membership of the dataset samples using a *leave-one-dataset-out* strategy. Both results shows a considerable improvement in the performance of an unbiased classification model.

In summary, the popularity of meta-learning in the application on medical domain data still need to be explored, but represents a promising research direction for the medical community.

3 Dataset description

The images used for this study correspond to three datasets from different sources. The size of the three datasets corresponds to 27,531 dermoscopic images, from which only a sample of 193 images were used. A description of each dataset used is given below.

- ISIC 2019: This is a public dataset for skin lesion analysis towards melanoma detection on high resolution dermoscopic images. For this study, only a subset from the training set was used.
- PH2 Database: This dataset is composed by dermoscopic images acquired at the Dermatology Service of Hospital Pedro Hispano in Matosinhos, Portugal. This image database contains 200 dermoscopic images of 768x560 pixels.
- 7-Point criteria evaluation database: This database is composed of a diagnosis and seven point checklist criteria labels publicly available from the Simon Fraser University website. This dataset is composed by over 2,000 dermoscopy images that correspond to twenty classes.

The images from the three datasets were chosen randomly, which corresponds to 1.2% of the total size of the datasets. The labels of the three datasets were combined into three groups: melanoma, malignant and benign. The malignant class contains labels of skin cancer moles, such as basal cell carcinoma, squamous cell carcinoma and actinic keratosis. The rest of non-melanoma and non-cancer labels were grouped into the benign class. The class distribution of the combined dataset used for this study is given in Table 1.

4 Methodology

For this study, a ResNet50 model was used for all the experiments. This is a variation of the ResNet architecture that consist of 50 convolutional layers. This architecture was used for its high performance on ImageNet and availability in the Keras library.

All the samples of the combined dataset were normalized and reduced at 254 x 254 pixels to match the input of ResNet. No image segmentation and colour variation was performed in this study.The parameters used in the experiment are provided in table 2.

Jaccard Similarity index and F1-score were used to measure the performance of the model on each experiment. Jaccard Similarity index was chosen for its wide use in the research community and therefore to serve as a point of comparison with the related works. F-1 score was chosen as it is a useful metric in the presence of class imbalance.

Table 2: Model parameters

Learning rate	0.0001
Loss	categorical cross entropy
Momentum	0.9
Optimizer	Stochastic Gradient Descent

Table 3: Summarized results of bias detection experiments

Jaccard Similarity Index	0.0
Overall precision	0.32
Overall recall	0.32
Overall F1-score	0.32

5 Experiments and Results

A set of three different experiments were conducted for this study. These experiments are listed below.

- Bias detection
- Bench-marking experiments on ResNet model with random weight initialization.
- Pre-trained ResNet model fine-tuned with medical data.

5.1 Bias detection

The *Name that dataset!* experiment was conducted to detect the presence of dataset bias. This experiment was first designed by (Khosla et al., 2012). The experiment consists in measuring the performance of a classification model on recognizing which dataset an image is from. If the model is able to recognize the dataset of origin, then we can say that the model is learning intrinsic characteristics particular to each dataset, or dataset bias.For this experiment, a simple convolutional neural network was implemented to detect the dataset bias. The results in table 3 show that the model is not able to infer the membership of the samples. This result is in alignment with our initial hypothesis, that no substantial difference between the samples of each dataset could be seen by the model, since the images of the three datasets were taken with the same device, and possibly under similar conditions.

5.2 Benchmarking experiments

For this experiment, a ResNet model was used with random weights initialization. This experiment was conducted to measure the performance of the baseline model with only the knowledge learned from the medical target data. The results of this experiment are given in figure 1. They were averaged out of 3 times using cross validation with stratified folds to ensure the class balance.

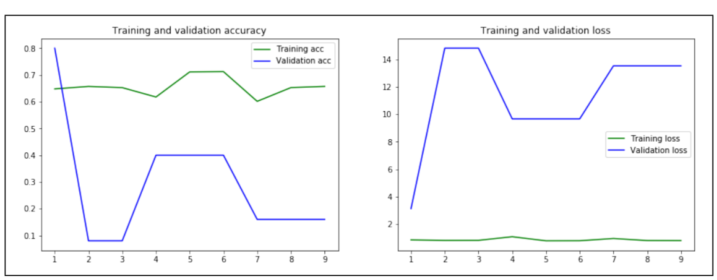

Figure 1: to the right: training and validation accuracy. To the left: training and validation loss of the benchmarking experiments

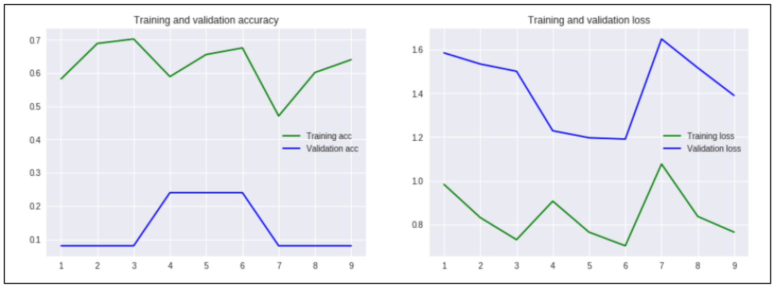

Figure 2: to the right: training and validation accuracy. To the left: training and validation loss of the meta-learning experiments without weighted classes

5.3 Meta-learning experiments

For this set of experiments, a pre-trained ResNet model was used. The weights of the network were obtained from training on the ImageNet dataset. The last softmax and dense layer was removed for the fine-tuning. The results of these experiments were averaged out of 3 times using cross validation with stratified folds. The following configuration were used for this set of experiments:

- Train and evaluation with no weighted classes.
- Train and evaluation with weighted classed.

The combined dataset is highly imbalanced, which causes overfitting due to the limited amount of samples of the minority classes. This is also related with the distribution shift between the training and test set, which does not share the same distribution of the data. To alleviate this effect, a simple class and distribution balancer algorithm is proposed. The purpose of this algorithm is to balance the classes in the training set and replicate the distribution of the test set in the training set to produce a set of weights for each class in the data that can be used during training. For its simplicity, it possess some drawbacks. One of them is that the same number of classes in the test set must exist in the training set and vice versa. See algorithm 1.

Algorithm 1: simple class and distribution balancer algorithm

Get majority class in Training set **TR->MC**
foreach *C* **do**
    Divide the number of samples of *C* in *TR* into the number of samples of *C* in Test set **TE->DTC**
    Divide *MC* into the numbers of samples of *C* in **TR->ITC**
    Set the weight for **C->DTC * ITC**
end

Figure 2 shows the results obtained by the ResNet model pre-trained with ImageNet without using the Class and distribution balancer algorithm.

The results obtained with these experiments demonstrates the benefit of using knowledge of an unrelated domain to predict the target class of the medical data as compared with the performance obtained with random initialization. This increase in the performance of classification is due to the ability of the model to extract properties that are good to generalize in the medical image dataset.

Figure 3 shows the results obtained by using the Class balancer algorithm with the ResNet model pre-trained with ImageNet. This results shows a boost in the performance of the model in **5 points** in Jaccard similarity index. This increase corresponds to the improvement in the accuracy of the model in the recognition of melanoma cases. An improvement in the validation accuracy and a considerable reduction in the validation loss can be seen as compared with the results obtained without using the Class and distribution balancer algorithm. The summary of the results from these experiments is given in table 4.

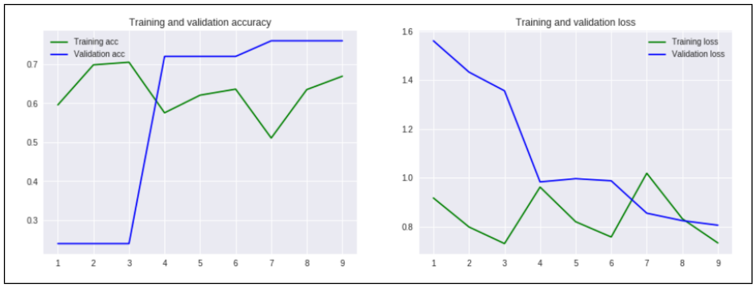

Figure 3: to the right: training and validation accuracy. To the left: training and validation loss of the meta-learning experiments using the Class balancer algorithm

Table 4: Summarized results

	Jaccard Similarity index	F-1 score
Benchmark	23.98	0.39
Fine-tunning with no weighted classes	43.03	0.54
Fine-tunning with weighting class algorithm	**47.22**	0.53

6 Conclusions

The results of the benchmarking experiments show that the model performs very well at training, but underperform in the validation set. This behaviour can be explained by the overfitting of the model towards the majority class, since the combined dataset is relatively small.

From the meta-learning experiments using ResNet for pre-training, a general improvement in the accuracy and a reduction in the loss is seen, as compared to the benchmarking experiments with random weights initialization. The results of the meta-learning experiments show that the model improved its generalization on the target dataset, and increased the recognition of the melanoma class.

The results obtained using the class weights generated by the Class and distribution balancer algorithm are interesting, as they show the same performance during training as compared with the experiments with no class weights. However, an increase in the accuracy during validation, and a significant drop in the validation loss was obtained. This boost in performance suggests that the model is sensitive to the covariate shift, therefore this should be considered in the development of meta-learning models for medical domains.

The results from this study suggest that popular deep learning models, such as ResNet, can extract knowledge of data from everyday objects and generalize for the classification of medical data, specifically the skin cancer moles, tackling one of the main challenges in the application of deep learning in medical domains, the scarcity of the data.

References

Ashraf, A., Khan, S. S., Bhagwat, N., Chakravarty, M. M., and Taati, B. (2018). Learning to unlearn: Building immunity to dataset bias in medical imaging studies. *CoRR*, abs/1812.01716.

BIGGS, J. B. (1985). THE ROLE OF METALEARNING IN STUDY PROCESSES. *British Journal of Educational Psychology*, 55(3):185–212.

Cheplygina, V., Moeskops, P., Veta, M., Bozorg, B. D., and Pluim, J. P. W. (2017). Exploring the similarity of medical imaging classification problems. *CoRR*, abs/1706.03509.

Dietterich, T. and Kong, E. B. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Hu, S., Tomczak, J., and Welling, M. (2018). Meta-learning for medical image classification. unpublished.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. (2012). Undoing the damage of dataset bias. In *Computer Vision – ECCV 2012*, pages 158–171. Springer Berlin Heidelberg.

Lemke, C., Budka, M., and Gabrys, B. (2013). Metalearning: a survey of trends and technologies. *Artificial Intelligence Review*, 44(1):117–130.

Schlegl, T., Ofner, J., and Langs, G. (2014). Unsupervised pre-training across image domains improves lung tissue classification. In *Medical Computer Vision: Algorithms for Big Data*, pages 82–93. Springer International Publishing.

Soekhoe, D., van der Putten, P., and Plaat, A. (2016). On the impact of data set size in transfer learning using deep neural networks. In *Lecture Notes in Computer Science*, pages 50–60. Springer International Publishing.

Wong, S. C., Gatt, A., Stamatescu, V., and McDonnell, M. D. (2016). Understanding data augmentation for classification: When to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE.