Finding and Classifying Tuberculosis Types for a Targeted Treatment: MedGIFT–UPB Participation in the ImageCLEF 2017 tuberculosis Task

Liviu–Daniel Ştefan[1], Yashin Dicente Cid[2,3], Oscar Jimenez–del–Toro[2,3], Bogdan Ionescu[1], and Henning Müller[2,3]

[1] University Politehnica of Bucharest, 061071 Romania
[2] University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland
[3] University of Geneva, Switzerland

Abstract. This paper describes the participation of the MedGIFT/UPB group in the ImageCLEF 2017 tuberculosis task. This task includes two subtasks: (1) multi–drug resistance detection (MDR), with the goal of determining the probability of a tuberculosis patient having a resistant form of tuberculosis and (2) tuberculosis type detection (TBT), with the goal of classifying each tuberculosis patient into one of the following five types: infiltrative, focal, tuberculoma, miliary and fibro–cavernous. Two runs were submitted for the TBT subtask and one run for the MDR subtask. Both of them use visual features learned with a deep learning approach directly from slices of patient CT (Computed Tomography) scans. For the TBT subtask the submitted runs obtained the 3rd and 8th position out of 23 runs submitted for this task, with a top Kappa value of 0.2329. In the MDR subtask, the proposed approach obtained the 7th position according to the accuracy (0.5352) out of 20 participant runs. Three main techniques were exploited during model training: pre–training the last layer of a neural network, small learning rates and data augmentation techniques. Data augmentation resulted in an effective and efficient data transformation that enhanced small lesions in the full image space.

Keywords: ImageCLEF, tuberculosis, medical image analysis, computed tomography, deep learning.

1 Introduction

According to the World Health Organization (WHO), tuberculosis remains as one of the top 10 causes of death worldwide [1]. Particularly when the disease is multi–drug–resistant (MDR), current treatment can be toxic and patient outcome is often poor [2]. An important challenge is still to correctly detect these MDR cases, as it has been estimated that only 20% of MDR tuberculosis cases are actually detected and treated accordingly [3]. Therefore, a better detection

and classification of tuberculosis types can result in a more targeted treatment for the patients [4].

ImageCLEF[4] is an image retrieval and analysis evaluation campaign (part of CLEF, the Cross Language Evaluation Forum) where many algorithms can be compared on the same basis [5]. ImageCLEF started in 2003, with a medical task included since 2004 and held every year since then [6]. In ImageCLEF 2017, the topic of one of the two medical tasks was related to tuberculosis data analysis from chest CT (Computed Tomography) images [7]. Two independent subtasks were proposed to the participants: (1) *Multi–drug resistance (MDR)*, including 230 3D CT scans from only HIV-negative patients, no relapses and classified according to their tuberculosis treatment: drug sensitive or multi–drug resistant, and (ii) *Tuberculosis type (TBT)*, containing 500 3D CT scans of TB patients, classified into five disease types: infiltrative, focal, tuberculoma, miliary and fibro–cavernous. For the MDR subtask the performance evaluation was done using ROC–curves generated from the probabilities provided by the contestants. On the other hand, for the TBT subtask the evaluation of performance was done using the unweighted Cohens Kappa. [8] gives more details on the dataset, task setup and full participant results, also for the other tasks of ImageCLEF 2017.

The MedGIFT–UBP group, submitted 3 runs in total within the ImageCLEF 2017 tuberculosis task, with participation in both subtasks. The runs were based on visual features extracted from the 2D axial slices of the patient CT scans, learned using a very deep Convolutional Neural Network (CNN). Two main research questions were addressed while training and testing algorithms with a heterogeneous and challenging dataset: 1) how to design an effective and efficient data transformation to enhance small lesions in the full image space?; 2) how to learn the ConvNet models given the limited training samples?.

The remainder of this paper is organized as follows: In Section 2.1, the data transformation method used to enhance small lesions in the full image space is described. Then, in Section 2.2, we investigate the network from a model-selection and optimization perspective. Section 3 reports the experimental setup. In Section 4 we report the results of our runs with respect to the top scores. Finally, in Section 5 we present conclusions and discuss current and future directions.

2 Methods

The proposed methods were built on top of the successful deep learning architecture described in [9] while tackling the problems mentioned above. In terms of volume structure modeling, a key observation is that consecutive slices are highly redundant. Therefore, a sparse structural sampling strategy was favorable in this case. To unleash the full potential of ConvNets with the provided data, several training practices were implemented to overcome the aforementioned difficulties resulting from a limited number of training samples. These practices include cross–modality pre–training and enhanced data augmentation.

[4] `http://www.imageclef.org/`

We apply for both of the subtasks the provided mask of the lungs [10] and extract the slices using a sparse structural sampling strategy described later in the paper. Then, we use a very deep CNN to perform the feature extraction from each slice of the CT volume of the patients. The parameters for each subtask are described in Section 3. Finally, we take the average of scores and use a Softmax classification to achieve the training-classification of new instances.

2.1 Data Enhancement

In this section, we first briefly describe our image representation method sampled using a sparse structural sampling strategy followed by several good practices for improving the size of the dataset for learning.

Data transformation The range of gray scale values in a natural image is of 256, but when considering color images, the possible values are 256^3, i.e. more than 16M. The range of Hounsfield Units (HU) in a CT slice could be grater than 4000 containing negative values (from -1024 to more than 3000), so a direct usage of a pre-trained GoogleNet network may not work. However, if we see each HU as a color code, then it should be possible to use a pre-trained GoogleNet network. Following this idea, we transformed each HU-based slice into a 2D RGB (Red, Green, Blue) image. Since we were interested only in the lung region containing HU in the range [-1024,300], we thresholded the image to this range, assigning the values out of the range to the respective limits. Then, each HU was mapped to a uniform distribution of the HSV space, varying from red, passing by yellow, green, cyan, blue, magenta, and red again. This mapping set to red any value outside the specified range. Furthermore, a key observation is that consecutive slices are highly redundant and therefore we used a sparse structural sampling strategy to enhance the quality of the dataset. Examples of a few 2D RGB images and their respective slices are depicted in Figure 1.

Multi–scale Cropping Augmentation To reduce the over-fitting problem due to the reduced training set we used a 10–crops data augmentation scheme to generate diverse training samples. We only crop 4 corners and 1 center of the images as well as their horizontal reflections (hence ten crops in all). We fix the input image size as 256×256 and randomly sample the cropping width and height from $\{256, 224, 192, 168\}$. After that, we resize the cropped regions to 224×224. At test time we average the predictions made by the network's Softmax layer on the 10 crops.

2.2 Network Architecture

GoogLeNet, the winner of the ILSVRC 2014 challenge, is a 22-layer deep convolutional network with layers stacked upon each other with different sizes. In order to speed up the computational efficiency it uses 1×1 convolutional operations for dimension reduction. Furthermore, it uses Average Pooling instead of

Fully Connected layers at the top of the ConvNet, eliminating a large amount of parameters that do not seem to matter much. More details can be found in its original paper [9].

We choose this architecture due to its improved utilization of the computing resources that allows to increase the depth and width of the network while keeping the computational budget constant that can therefore be trained effectively within a reasonable amount of time.

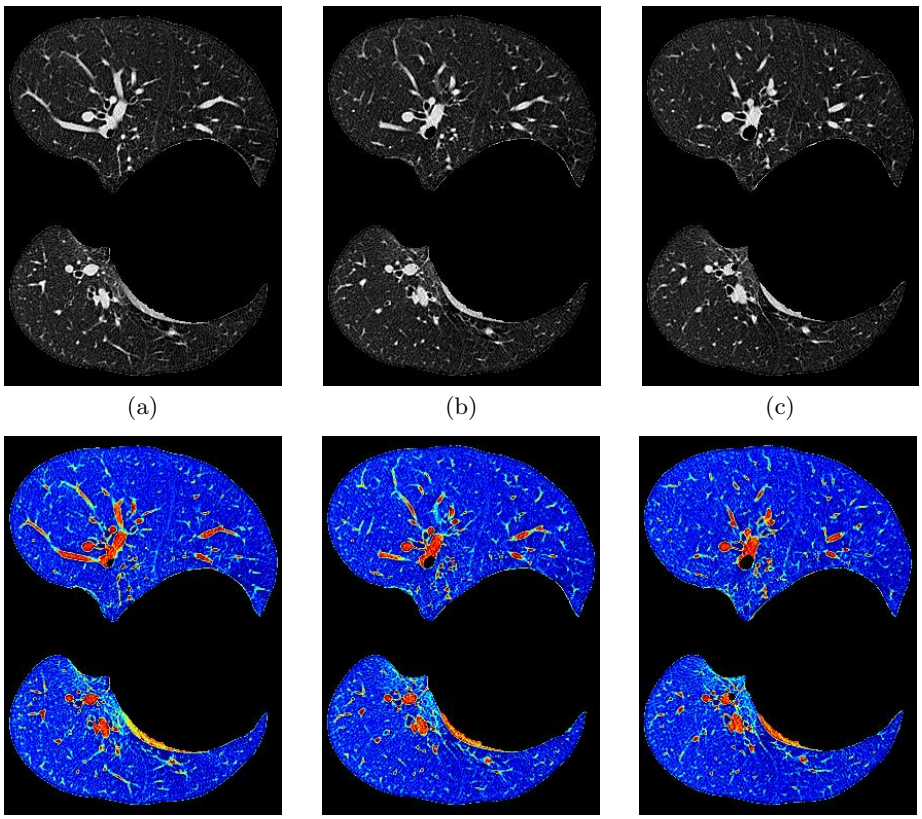

Fig. 1. The figure shows examples of the gray scale (a–c) and RGB modalities (d–f).

3 Experimental Setup

In this section, we give a detailed description of our proposed methods. We first present the network input and the training details. After that, we describe our testing strategies for both of the subtasks.

3.1 ImageCLEFtuberculosis: Task Setup for the Tuberculosis Type

Training *Run 1: TBT_T_GNet*: In this run, the network weights are learned using the mini-batch stochastic gradient descent with momentum set to 0.9. At each iteration, a mini-batch of 32 samples is constructed by sampling 32 training slices from the dataset using a batch accumulation of 2. We did not use any data augmentation techniques in this run. The learning rate is initially set to 10^{-5}, and then the rate is changed to 10^{-6} after 1/3 of the iterations, then to 10^{-7} after 2/3 of the iterations, and finally 10^{-8} after which the training is stopped.

As the network in this run takes RGB images as input, we pre–trained it using the ImageNet [11] model as initialization.

Run 2: TBT_TEST_RUN_2_GoogleNet_10crops_at_(different_scales): In this run, the network weights are learned using the mini-batch stochastic gradient descent with momentum set to 0.9. At each iteration, a mini-batch of 32 samples is constructed by sampling 32 training slices from the dataset. In net training, a sub–image is randomly cropped from the selected slice using the techniques described in the previous section. The learning rate is initially set to 10^{-5} scheduled with a polynomial decreasing policy at a power of 5. As the network in this run takes gray level images as input, we pretrained it using the ImageNet model as initialization. First, we discretize the slices into the interval from 0 to 255 by a linear transformation. This step makes the range to be the same with RGB images. Then, we modify the weights of first convolution layer of the ImageNet RGB model to handle the input of grayscale images. Specifically, we average the model filters of first layer across the channel. This initialization method works pretty well and reduce the effect of overfitting in experiments.

Testing *Run 1: TBT_T_GNet*: At test time, given a volume, we extract all the slices (samples). The class scores for the whole volume are then obtained by averaging the scores across the slices. Finally, we use a softmax classification to achieve the training–classification of new instances. This yields an accuracy of 46% on the validation set.

Run 2: TBT_TEST_RUN_2_GoogleNet_10crops_at_(different_scales): At test time, given a volume, we extract all the slices. The class scores for the whole volumes are then obtained by averaging the scores across the slices and crops therein. Finally, we use a Softmax classification to achieve the training–classification of a new volume. This yields an accuracy of 41% on the validation set.

In this run we simultaneously enriched the dataset both in quality and quantity, but this resulted in performance deterioration due to the noise introduced. We also tried the data enrichment approach using the RGB modality, but we did not notice any improvement.

3.2 ImageCLEFtuberculosis: Setup and Results for the Multi-drug Resistant Task

Training *MDR_TST_RUN_1*: The network weights are learned using the mini-batch stochastic gradient descent with momentum set to 0.9. At each iteration,

a mini-batch of 32 samples is constructed by sampling 32 training slices from the dataset using a batch accumulation of 2. In net training, a 224×224 sub–image is randomly cropped from the selected frame. The learning rate is initially set to 10^{-5} scheduled with a polynomial decreasing policy at a power of 5.

Testing *MDR_TST_RUN_1*: In the test phase, given a volume, we extract all the slices. The class scores for the whole volume is then obtained by averaging the scores across the slices.

4 Results

For the TBT subtask the submitted runs that obtained the 3rd and 8th position out of 23 runs submitted for this task, with a top Kappa value of 0.2329.

In Table 1, the performance of the proposed and the top techniques for the TBT subtask can be seen. The Table reports scores that correspond to the test set, from the metric proposed by the organizers. Our best score is reported in run1, which uses the RGB modality without data augmentation techniques. On the other hand, the run2 reports scores obtained by using the grayscale modality with data augmentation techniques. In the latter result, we consider that the representation is affected by noise when increasing the number of samples.

Table 1. Selected subset of results from ImageCLEFtuberculosis (2017) – TBT subtask. Comparison of our submitted runs to top scores.

Group name	Run name	Run type	Kappa	ACC
SGEast	TBT_resnet_full	Not applicable	0.24	0.4
SGEast	TBT_LSTM_17_wcrop	Not applicable	0.23	0.39
MedGIFT-UPB	**Run 1**	**Automatic**	**0.23**	**0.38**
MedGIFT-UPB	**Run 2**	**Automatic**	**0.19**	**0.37**

In the MDR subtask, the proposed approach obtained the 7th position according to the accuracy (0.5352) out of 28 participant runs. In Table 2, the performance of the proposed and the top techniques for the MDR subtask can be seen. The table reports scores that correspond to the test set using the metrics proposed by the organizers.

Overall, our group obtained good positions both in the TBT and MDR tasks.

5 Conclusions

In this work we propose two fully automatic tuberculosis classification methods and one automatic predictor for assessing the probability of TB patients having a multi-drug resistant TB. These approaches were evaluated in the TBT and

Table 2. Selected subset of results of the ImageCLEF 2017 tuberculosis task – MDR subtask. Comparison of our submitted runs to the best results.

Group Name	Run name	Run type	AUC	ACC
MedGIFT	MDR_Top1_correct	Automatic	0.58	0.51
MedGIFT	MDR_submitted_topBest3_correct	Automatic	0.57	0.46
MedGIFT	MDR_submitted_topBest5_correct	Automatic	0.56	0.48
MedGIFT-UPB	**Run 1 (baseline)**	Automatic	**0.51**	**0.53**

MDR subtasks of the ImageCLEF 2017 tuberculosis task. Due to the fact that the datasets are relatively small, we tested several good practices for training the ConvNets. Relying on the proposed training strategies, the approaches achieved an accuracy of 38.7% on the TB type dataset and 51% on the MDR dataset. A research direction is to introduce more aggressive data augmentation techniques designed to improve the network generalization capabilities aligned with new techniques to generate medical hypotheses.

References

1. Organization, W.H., et al.: Global tuberculosis report 2016. (2016)
2. Ahuja, S.D., Ashkin, D., Avendano, M., Banerjee, R., Bauer, M., Bayona, J.N., Becerra, M.C., Benedetti, A., Burgos, M., Centis, R., Chan, E.D., Chiang, C.Y., Cox, H., D'Ambrosio, L., DeRiemer, K., Dung, N.H., Enarson, D., Falzon, D., Flanagan, K., Flood, J., Garcia-Garcia, M.L., Gandhi, N., Granich, R.M., Hollm-Delgado, M.G., Holtz, T.H., Iseman, M.D., Jarlsberg, L.G., Keshavjee, S., Kim, H.R., Koh, W.J., Lancaster, J., Lange, C., de Lange, W.C.M., Leimane, V., Leung, C.C., Li, J., Menzies, D., Migliori, G.B., Mishustin, S.P., Mitnick, C.D., Narita, M., O'Riordan, P., Pai, M., Palmero, D., Park, S.k., Pasvol, G., Pea, J., Prez-Guzmn, C., Quelapio, M.I.D., Ponce-de Leon, A., Riekstina, V., Robert, J., Royce, S., Schaaf, H.S., Seung, K.J., Shah, L., Shim, T.S., Shin, S.S., Shiraishi, Y., Sifuentes-Osornio, J., Sotgiu, G., Strand, M.J., Tabarsi, P., Tupasi, T.E., van Altena, R., Van der Walt, M., Van der Werf, T.S., Vargas, M.H., Viiklepp, P., Westenhouse, J., Yew, W.W., Yim, J.J.: Multidrug resistant pulmonary tuberculosis treatment regimens and patient outcomes: an individual patient data meta–analysis of 9,153 patients. PLoS med **9**(8) (2012) e1001300
3. Rendon, A., Tiberi, S., Scardigli, A., DAmbrosio, L., Centis, R., Caminero, J.A., Migliori, G.B.: Classification of drugs to treat multidrug–resistant tuberculosis (mdr–tb): evidence and perspectives. Journal of Thoracic Disease **8**(10) (2016) 2666
4. Horsburgh, C.R.J., Barry, C.E.I., Lange, C.: Treatment of tuberculosis. New England Journal of Medicine **373**(22) (2015) 2149–2160
5. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., García Seco de Herrera, A., Bromuri, S., Amin, M.A., Kazi Mohammed, M., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., Roldán García, M.d.M.: General overview of ImageCLEF at the CLEF 2015 labs. In: Working Notes of CLEF 2015. Lecture Notes in Computer Science. Springer International Publishing (2015)

6. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics **39**(0) (2015) 55 – 61
7. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, Dublin, Ireland, CEUR-WS.org ¡http://ceur-ws.org¿ (September 11-14 2017)
8. Ionescu, B., Müller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.T., Dicente Cid, Y., Eickhoff, C., Garcia Seco de Herrera, A., Gurrin, C., Islam, Bayzidul and, K.V., Liauchuk, V., Mothe, J., Piras, L., Riegler, M., Schwall, I.: Overview of ImageCLEF 2017: Information extraction from images. In: CLEF 2017 Proceedings. Lecture Notes in Computer Science, Dublin, Ireland, Springer (September 11-14 2017)
9. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. CoRR **abs/1409.4842** (2014)
10. Dicente Cid, Y., Jiménez del Toro, O.A., Depeursinge, A., Müller, H.: Efficient and fully automatic segmentation of the lungs in ct volumes. In Goksel, O., Jiménez del Toro, O.A., Foncubierta-Rodríguez, A., Müller, H., eds.: Proceedings of the VISCERAL Anatomy Grand Challenge at the 2015 IEEE ISBI. CEUR Workshop Proceedings, CEUR-WS (May 2015) 31–35
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3) (2015) 211–252