Identifying tuberculosis type in CTs

Cosmin Moisii[1,2], Radu Miron[1,2] and Mihaela Elena Breaban[2]

[1] *SenticLab, Iasi, Romania*
[2] *Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iasi, Romania*

Abstract

The paper proposes and compares two distinct approaches based on deep learning for tuberculosis classification in CTs, highlighting the benefits of building the inference engine at slice-level over a volumetric approach. The methods are evaluated in the context of the ImageClef 2021 Tuberculosis task and the reported work belongs to the SenticLab.UAIC team, which ranked the first in the competition.

1. Introduction

According to the World Health Organization[1], tuberculosis (TB) is one of the top 10 causes of death worldwide and the leading cause from a single infectious agent. It is present in all countries and age groups and was the cause of a total of 1.4 million deaths in 2019, with an estimate of 10 million people infected worldwide. Generally, TB can be cured with antibiotics. An estimated 60 million lives were saved through TB diagnosis and treatment between 2000 and 2019. However, the different types of TB require different treatments, and therefore the detection of the TB type and characteristics are important real-world tasks.

In this regard, the 2021 edition of the Tuberculosis task within ImageCLEFmed [1, 2] aimed at automatically categorizing CTs of TB patients into one of the following five types: (1) Infiltrative, (2) Focal, (3) Tuberculoma, (4) Miliary, (5) Fibro-cavernous. The current paper reports the approaches developed by the SenticLab.UAIC team obtaining the best results in the competition[2].

Given the 3-dimensional nature of the CTs, several ways to tackle the classification problem in terms of input type exist. In previous work [3], we compared three different strategies: 1) compressing the 3D matrix to 2D representations by computing projections onto 3 distinct planes, 2) treating the 3D volume as a whole and thus using 3D convolutions or fusing the information at slice level and 3) bringing the inference process to the slice level. The 3rd approach proved to outperform significantly the others in terms of accuracy, at a higher cost of data preparation and much less computational burden compared to a 3D approach; training data preparation involves in this case identifying slices in the CT that present the affection specified as label for the entire CT.

pmihaela@info.uaic.ro (M. E. Breaban)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://www.who.int/news-room/fact-sheets/detail/tuberculosis
[2] https://www.imageclef.org/2021/medical/tuberculosis

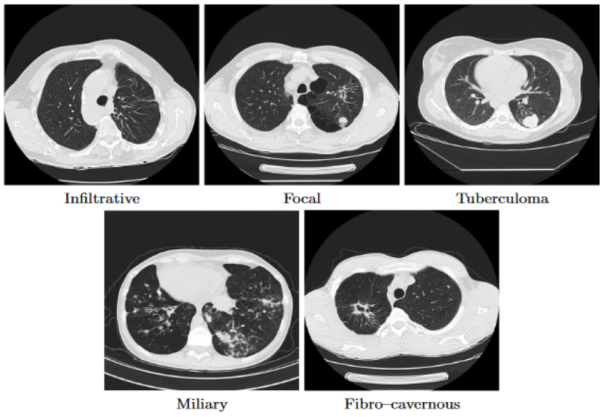

Figure 1: Tuberculosis lesion types targeted in the competition (https://www.imageclef.org/2021/medical/tuberculosis)

The current work experiments with two of the approaches above, the 3rd approach proving again to be the winner solution in the competition. Another key component of the winner solution was the aggregation step of the inference results obtained at slice level, of great importance especially for the objective of the 2021 evaluation task, where only one label had to be output per CT, although our analysis highlighted the existence of several affections for some CTs.

The paper is structured as follows. Section 2 describes the challenge and the dataset. Section 3 presents the approach we used to exploit the whole volumetric information. Section 4 describes the architectures used to process the information at slice level and the heuristics used to produce the diagnosis report at the CT level. Section 5 summarizes the results obtained on the blind test set in the competition and discusses comparatively the performance of the methods. Section 6 concludes the paper.

2. ImageCLEFmed Tuberculosis 2021: tasks, data, evaluation

The challenge in the 2021 ImageCLEFmed Tuberculosis competition is the automatic classification of CTs into one of 5 TB types - illustrated in Figure 1.

The training dataset consists of chest CT scans of 917 TB patients, each CT scan being categorized in only one TB class. The test set consists of 421 CT scans. Part of the training data also has some additional metadata, but because such information was not available for the test data, we did not include it in the analysis.

The resolution is 512x512 with a variable number of slices - 580 at maximum (illustrated in Figure 2) and various spacing, the slice thickness varying from 0.6 to 5 mm with a median at 2.5 mm.

The distribution of classes is imbalanced, as illustrated in Figure 3.

The metrics used to measure the performance of the algorithms are Cohen's Kappa and accuracy, the former being used to rank the entries in the competition.

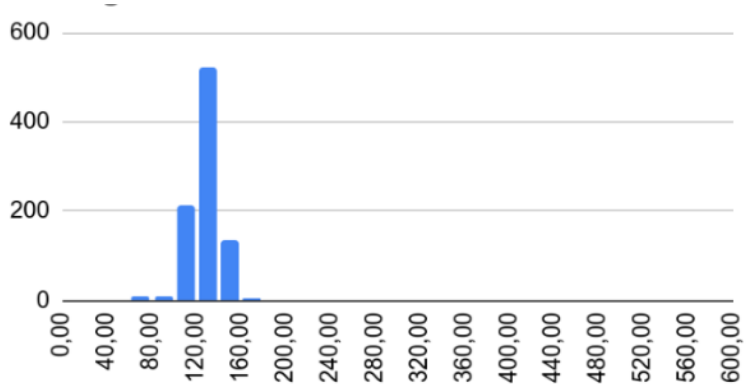

Figure 2: Distribution of the number of slices per CT

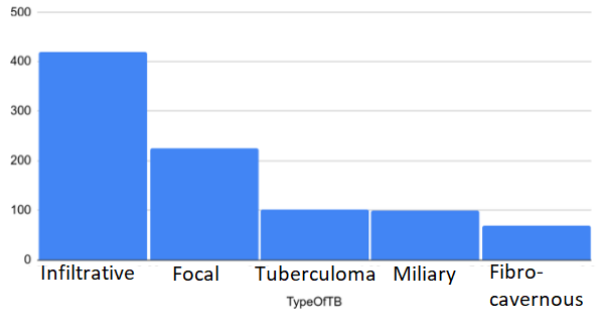

Figure 3: Distribution of classes in the training set

3. Learning from volumetric data

2D convolutional neural networks have been very successful in a wide range of 2D image vision tasks from classification to object detection and segmentation. Ever since the appearance of Alex-Net [4], state of the art results have been obtained on several benchmarks.

For these reasons we chose to work with convolutions for this competition. Although impressive results have been obtained on 2D image tasks using 2D convolutions, 3D convolutions still have to emerge as de facto architecture for 3D image tasks. Since the convolution operation is a local one, searching for features in the neighbourhood of a pixel and tuberculosis type might be influenced by several pathologies found in different and distant slices of the same patient, we choose as our main model 3D ResNet with Non Local Features [5]. ResNets [6] have become popular due to their residual connections which prevent failure when training very deep neural networks.

Pretraining has had a significant role in increasing the performances of convolutional neural networks. We chose to use pretrained 3D ResNet50 with Non Local features on Kinetics. Due to good results mentioned in [7] we chose Inflated 3D architecture with weights pretrained on

ImageNet [1].

We converted each volume to a sequence of images, each image representing the RGB representation of each slice. We chose window width equal to 1500 and window level equal to -600 for the whole volume and stored only 8bit of information for each slice, thus obtaining an $[0, 255]$ range 1 channel image. In order to take advantage of the pretrained models which require 3 channel images as input, we simply duplicated the first channel over the second and the third channel. We chose to resize each image slice to 128×128 and 256×256 pixels. Each input image is a $N \times N \times 120$ part of the whole volume, where N is the pixel size of an image. Padding with 0 filled slices is done if necessary.

We used 2 training phases for the final model. The first phase uses as input slices with 128×128 dimensions and is trained with no augmentations techniques. The total number of epochs is 100. As loss function we use Cross Entropy. Initial learning rate is $1e-3$. Learning rate scheduler is Linear Scheduler with a decreasing factor of 0.5 each 20 epochs. We use Stochastic Gradient Descent as optimizer with a weight decay of $1e-6$. We call this the $pretraining$ phase.

The second phase is the proper training. This time we use 256×256 images as input. We use as augmentations Horizontal and Vertical Flips, Contrast and Color distortions and Gaussian Blur as well, each with a 0.5 chance. With 0.5 chance we also invert the volume. The total number of epochs is 100. Initial learning rate is $1e-3$ with a decrease factor of 0.5 each 15 epochs. The loss used is Cross Entropy with label smoothing in order to avoid overfitting. Each volume was normalized using ImageNet mean and standard deviation. We use Stochastic Gradient Descent as optimizer with a weight decay of $1e-6$. We use as initial weights the final weights obtained by previous $pretraining$ phase in order to not start from scratch. For this approach we did not use the masks for lung segmentation provided by the competition, nor any other method to segment the lungs. The entire CT was used as is.

We use test time augmentations. We perform for each image 6 inference steps. One step with the original image and another step with the reversed image. For the other 4 steps we use random augmentations that we used during the training phase. We used the last model saved during training for inference and also the last 10 models saved during training as an ensemble, leading to a total of 66 predictions per CT. We use different techniques for aggregating the results of the ensemble and different test time augmentations. The first method is to pick as final label the most frequent label predicted (**FS 3DNLR50**). In case of frequency equality the prediction with the highest score is chosen. The second method is based on the mean of the scores for each label (**MS 3DNLR50**). The label with the highest mean is picked. The third method uses as final label the one with the highest score predicted among all predictions (**HS 3DNLR50**). The second method obtains the best performance. The single model inference consisting of applying the model in the last epoch, with no test-time augmentation, gives the poorest results (**S 3DNLR50**). We believe this is due to the fact that a single volume can't always present pathologies for a single type of Tuberculosis. This is hinted by the instability of the performance metrics (loss and accuracy) computed on the training set, even on the final epochs with a small learning rate. Label smoothing also prevents overfitting acting as a strong regularizer. Training on the 128×128 images without label smoothing reaches almost perfect accuracy after 150

[1] https://github.com/facebookresearch/video-nonlocal-netmain-results

epochs, whereas training with label smoothing reaches less than 80% accurcy.

4. Learning at slice level

Having a closer look at the training set, one can observe that usually the lesions on the lungs are located only on a small number of slices from the whole volume. A natural idea is to try a 2D model that could differentiate between healthy lung slices and lung slices with lesions and construct the CT report based on the findings at slice level. For this purpose we need training data labeled at slice level and not CT level.

We manually selected slices, at lung level, that we thought were relevant for the respective label. This means we carefully selected only the slices that contained the representative label even though, to our opinion confirmed later by a radiologist, that CT contained pathologies corresponding to other labels as well. The selection was made by us, briefly trained by small descriptions we found on the internet. We strived to make a balanced dataset, but due to the nature of the pathology some were easier to gather than others. Also due to the easiness of selecting healthy slices we took the opportunity to make a large healthy slices set.

Using the same architecture as last year [3], Efficient-net B4 [8], we trained a classifier with 6 labels (the initial 5 tuberculosis types + healthy class). We then aggregated the predictions probabilities for each volume and used it as a data point for a different simpler classifier (we tried Multi Layer Perceptron and Logistic Regression) using as output the final volume label. This way we aggregated all slices into a single label. We further increased the scores by using test-time augmentations (only horizontal flip) and averaging several second-stage classifier results that were trained with different parameters.

4.1. Training

We used approximately the same approach we used last year at ImageCLEF CTReport challenge which we will briefly describe here.

Given a selected sliced we grouped it together with the previous and the next slice in the volume. We changed its window and level values to highlight the lung features. The selected slices were split into half, corresponding to each lung, and we kept only the side with the affection. We cropped the images, using a simple threshold method to remove the padding and kept only the body. The resulting images were resized to 256×256 pixels. These were then normalized with values in the range $[0,255]$ corresponding to 3 black and white images which were concatenated at channel level. These mini volumes of 3 consecutive slices, we thought, could better highlight the difference between an infiltration and an artery, or a cavern and a lumen as these are can be very similar at a certain point in space but continue in a different manners.

As augmentations we used a random crop of size 224×224, a random horizontal flip with a probability of 0.5 and normalized the image. We trained an EfficientNet-B4 , with 90 epochs and batches of 32. We used this network to predict on each slice of a volume in the training set (processed in the fashion we explained above after the volume was resized to a fixed value of depth 128) and obtain the probabilities of each affection type. These probabilities were concatenated into an array of size $[128, 6]$ corresponding to [no of slices, no of labels]. These

arrays were used as input to train a simpler classifier (for example a logistic regression classifier), each array corresponding to a label. We did not use any masks nor segmentation algorithms to extract the lungs.

The first submission (**Ef MLP**) using this approach resulted in a kappa score of 0.203 and used 9/10 of the data, the rest 1/10 being used for internal validation. The second stage classifier was an MLP classifier with 2 hidden layers of size 100 and 30 respectively. No test-time augmentation was performed. The second submission (**Ef MLP LogReg**) with a kappa score of 0.221 was a mean of 4 predictions: one MLP classifier and one LogReg classifier tested on the original and flipped images. This submission scored the highest. The rest of the submissions (**Ef comb**) correspond to different training parameters and means of scores (second stage classifiers training on flipped images, means of first stage classifier probabilities on original and flipped images, etc).

5. Comparative results

Table 1 summarizes the results obtained in the competition on the test set.

Table 1
Results reported on the test set. The first four entries use a 3d approach, while all the others make predictions at slice level.

Method	kappa	accuracy
S 3DNLR50	0.169	0.366
HS 3DNLR50	0.174	0.397
FS 3DNLR50	0.183	0.401
MS 3DNLR50	0.187	0.404
Ef comb1	0.192	0.444
Ef comb2	0.194	0.444
Ef MLP	0.203	0.458
Ef comb3	0.205	0.449
Ef MLP LogReg	0.221	0.466

The winning submission corresponds to a kappa score of 0.221. The low scores obtained generally in the competition are not a surprise for us, since, during the phase of manual slice labeling, we identified CTs in the training set presenting more affection types and not only the labeled one. In our opinion, the task should be framed as a multi-label classification problem, giving the possibility to report all the affections present. We could not find the rational behind CT labeling for the cases that present more than one lesion type, and neither could the AI, as the results indicate.

6. Conclusions

In the light of the results we obtained in both the 2020 and the 2021 ImageClef TB evaluation tasks, we may conclude that an approach based on inference at slice level is superior to those using the entire volume in such classification tasks. The effort of including in the training

set the needed information in the form of labeled slices, basically reducing to identifying the sequence of slices presenting a certain affection, is definitely rewarding, not only in terms of accuracy gain, but also in terms of the inference type, models built in this way being able to provide more valuable information like localization and size of the affection.

7. Acknowledgements

This research was partially supported by the Competitiveness Operational Programme Romania under project number SMIS 124759 - RaaS-IS (Research as a Service Iasi).

References

[1] B. Ionescu, H. Müller, R. Peteri, A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, S. Kozlovski, V. Liauchuk, Y. Dicente, V. Kovalev, O. Pelka, A. G. S. de Herrera, J. Jacutprakart, C. M. Friedrich, R. Berari, A. Tauteanu, D. Fichou, P. Brie, M. Dogariu, L. D. Ştefan, M. G. Constantin, J. Chamberlain, A. Campello, A. Clark, T. A. Oliver, H. Moustahfid, A. Popescu, J. Deshayes-Chossart, Overview of the ImageCLEF 2021: Multimedia retrieval in medical, nature, internet and social media applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), LNCS Lecture Notes in Computer Science, Springer, Bucharest, Romania, 2021.

[2] S. Kozlovski, V. Liauchuk, Y. Dicente Cid, V. Kovalev, H. Müller, Overview of ImageCLEFtuberculosis 2021 - CT-based tuberculosis type classification, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Bucharest, Romania, 2021.

[3] R. Miron, C. Moisii, M. Breaban, Revealing lung affections from cts. A comparative analysis of various deep learning approaches for dealing with volumetric data, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_105.pdf.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[5] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, 2018. arXiv:1711.07971.

[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015). URL: http://arxiv.org/abs/1512.03385. arXiv:1512.03385.

[7] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, CoRR abs/1705.07750 (2017). URL: http://arxiv.org/abs/1705.07750. arXiv:1705.07750.

[8] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, arXiv preprint arXiv:1905.11946 (2019).