PUC Chile team at TBT Task: Diagnosis of Tuberculosis Type using segmented CT scans

José Miguel Quintana[1], Daniel Florea[1], Ria Deane[1], Denis Parra[1], Pablo Pino[1], Pablo Messina[1] and Hans Löbel[1]

[1]*Department of Computer Science, School of Engineering, Pontificia Universidad Católica de Chile, Chile*

Abstract

This article describes the participation and results of the PUC Chile team in the Turberculosis task in the context of ImageCLEFmedical challenge 2021. We were ranked 7th based on the kappa metric and 4th in terms of accuracy. We describe three approaches we tried in order to address the task. Our best approach used 2D images visually encoded with a DenseNet neural network, which representations were concatenated to finally output the classification with a softmax layer. We describe in detail this and other two approaches, and we conclude by discussing some ideas for future work.

1 Introduction

ImageCLEF [1] is an initiative with the aim of advancing the field of image retrieval (IR) as well as enhancing the evaluation in various fields of IR. The initiative takes the form of several challenges, and it is specially aware of the changes in the IR field in recent years, which have brought about tasks requiring the use of different types of data such as text, images and other features moving towards multi-modality. ImageCLEF has been running annually since 2003, and since the second version (2004) there are medical images involved in some tasks, such as medical image retrieval. Since then, new tasks involving medical images have been integrated into the ImageCLEFmedical challenge group of tasks [2], and that is how the task of Tuberculosis type classification has been taking place since 2017. Although there has been changes in the data used for the newest versions of the challenge, the goal of this task is the same: automatic detection of tuberculosis (TB) types using Computer Tomography (CT) volumes as input data.

In this document we describe the participation of our team from HAIVis group [1] within the artificial intelligence laboratory [2] at PUC Chile (**PUC Chile team**) in the TB classification task at MedicalImageCLEF 2021 [2]. Our team earned the 7th place in terms of kappa metric and the fourth place in terms of accuracy in the challenge. Our best submission was a combination of deep learning techniques for two 2D views of each input CT volume, followed by a traditional multi-class classification via softmax layer.

The rest of the paper is structured as follows: Section 2 describes our data analysis, and in section 3 we provide details of our proposed approaches, including data augmentation. In

[1]http://haivis.ing.puc.cl/
[2]http://ialab.ing.puc.cl/
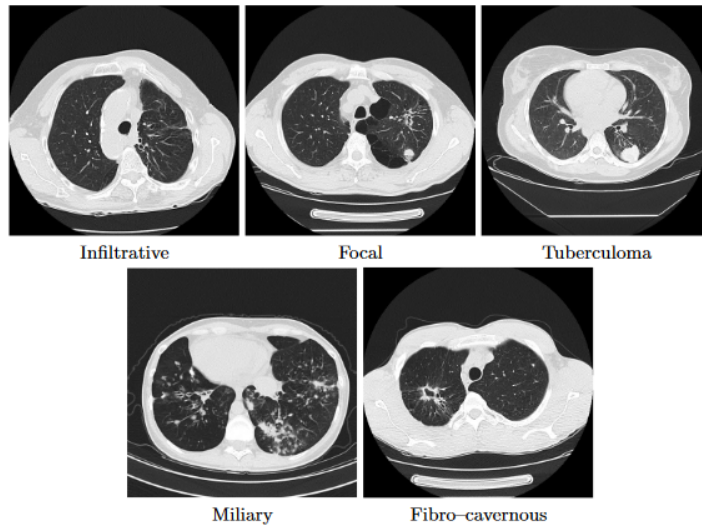
Figure 1: CT-Scans of different type of Tuberculosis

section 4 we provide details of our results, and finally in section 5 we conclude our article.

2 ImageCLEFmed Tuberculosis: tasks, data, evaluation

The challenge for the 2021 ImageCLEFmed Tuberculosis (TB) task is to automatically categorize a CT scan into one of five TB types. The generated prediction must indicate one label for each image, specifying the type of TB it contains.

The total dataset consists of 1,338 CT scans of TB patients, 917 assigned for training and 421 for testing. Each CT-image corresponds to only one TB category at a time. With respect to the training data set from the 917 scans, 420 have Infiltrative TB, 226 Focal, 101 Tuberculoma, 100 Miliary and 70 Fibro-cavernous. Due to segmenting each CT scan into their corresponding left and right lungs, we end up with double the amount of images.

Therefore this task corresponds to a multi-class classification problem. To rank submissions, each result is evaluated using the unweighted Cohen's Kappa as a primary metric and accuracy as a secondary metric.

2.1 Dataset Analysis

Before carrying out our different approaches to classify each CT scan into a type of tuberculosis, we studied the training dataset. Figure 1 shows the prevalence of each class.

Table 1 summarizes this information and presents the relative prevalence of each type of Tuberculosis with respect to the total amount of images in the training dataset. It is important to note the clear class imbalance present, where nearly half of all images have Infiltrative Tuberculosis.
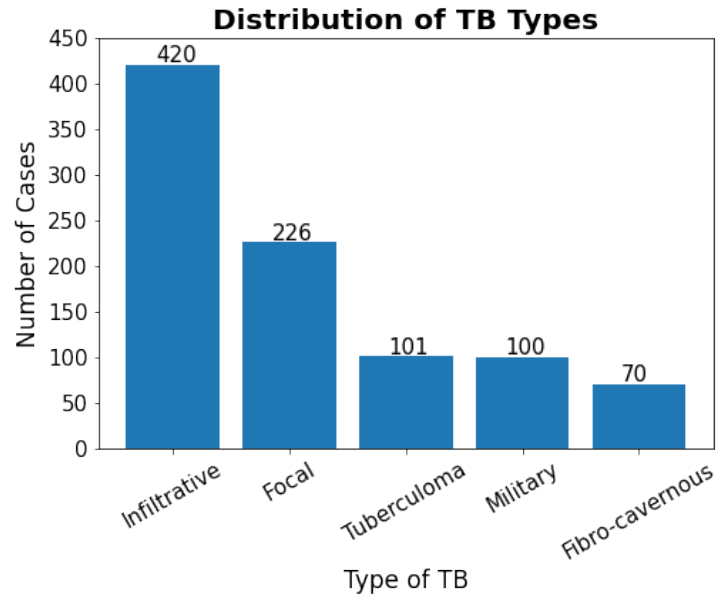
Figure 2: Number of images by class

Table 1
Number of images and percentage each type of Tuberculosis forms of the training dataset

TB Type	Number of images	Percentage of dataset (%)
Infiltrative	420	45.8
Focal	226	24.6
Tuberculoma	101	11.0
Military	100	10.9
Fibro-cavernous	70	7.6
Total	917	100.0

3 Approaches

3.1 The Human-Inspired 2D Approach

After consulting medical advice on how they would personally address the challenge, it was theorized that a model could perform better if it trained and predicted with the same data as reviewed by doctors when analyzing CT scans, this means a top view of the image from top to bottom.

The main reason for taking this approach was that it has proven to be helpful in different computer vision tasks [3] and there was no found documentation about the subject. Further investigation about the subject is aimed to be performed in future work.

3.1.1 Preprocessing

First, segmentation masks provided by ImageCLEF were applied to each 3D CT-scan, in order to separate each lung of the patient into two separate inputs. This practice was validated by medical professional, as diseases such as Tuberculosis tend to manifest on both lungs at a time. Second, each 3D matrix was then split into 2D images viewed from the z-axis, this practice is the one that gives the approach's name, as radiologists only use top-down images to review CT scans, the main reason for not using other views is that noise from other parts of the internal structure of the lungs can mislead the professional on it's diagnosis. This same thought process was used in order to train the model. It was also commented by the medical professionals that tuberculosis often concentrates on the upper part of the lungs, based around this observation 20% of images on the bottom of the scan were discarded.

After this first pre-treatment, each image was normalized and later concatenated in order to produce RGB images, task performed by the dataset, which also was in charge of performing augmentations on each item when loaded and after which, transforming them into tensors.

3.1.2 Augmentations

Each image was cropped and later re-centered and randomly scaled both on the x and y axis. After this, an angle was applied to the image, as well as shear and a random horizontal flip. This last practice was performed in order to avoid biases between left and right lungs.

It was considered to flip all images only to one side, this idea was later rejected due to the fact that human professionals are skilled enough to detect the presence of tuberculosis in a lung regardless of his orientation or which one of the two is the scan of.

As for the parameters used, the cropping center for each axis was decided by selecting the original center of the image and displacing it by a random amount of pixels between the values of 0 and 32. As for the angle of the rotation and the shear, a random float between 0 and 6.0 and another one between 0 and 4.0 were used respectively. Finally, for the scaling, each axis was multiplied by 2^x, with x being a random float between the values of 0 and 0.15.
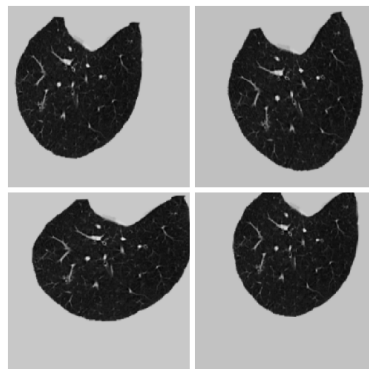
Figure 3: Multiple augmentations on the same crop of a CT Scan

3.1.3 Model

The model used was lightly tweaked DenseNet121 pretrained on ImageNet, using an Adam optimizer in order for training to be faster, this due to the high amount of images used during the training process. As for the loss, a weighted cross entropy was used in order to decrease class imbalance. Finally, learning rate was reduced on plateau of the validation set's loss.

The model trained on each of the images, receiving both an image and a label for the input and failed to converge on a good solution for the test set even after many epochs of training. It was suspected that this was due to the grand amount of images not containing any information which were given as inputs. This was later proven right when inspecting the output predictions for images belonging to a same CT scan; clusters of correct predictions where found in the middle of many different predictions, later revealed by a medical experts to be the exact layers to have the disease present in them.

Due to a lack of time, no solution to the issue was available to be implemented and is left to be further developed in future works, due to the uncertainty in the effectiveness of this approach.

3.2 A Simple 2D Approach

This approach was a modified version of one proposed in [4], which tried to represent 3D images by squeezing the volumetric data to 2D projections. Unfortunately, it did not work as expected due to an exploding gradient problem in its execution which, because of lack of time, was not able to be resolved.

We believe it is still useful to explain this approach and study the reasons it did not work.

3.2.1 Preprocessing

As mentioned in 3.1, this approach was a modified version of one proposed in [4]. In this case, instead of calculating the mean, maximum and standard deviation in each dimension, we only calculated the maximum for each axis and created a single 2D, three channel image from the three matrices that appeared, which can be interpreted as an RGB image.

First of all, we applied the segmentation mask provided by ImageCLEF to each 3D CT-scan, dropped the first 10% and last 20% of the image to eliminate unnecessary information, and then calculated the maximum values across each dimension. This produced three 2D matrices, which were subsequently concatenated to produce one single RGB image. The reason we only calculated the maximum was because we believe that, due to the nature of tuberculosis being various small nodules present across the lungs, calculating the average and the standard deviation across dimensions were not accurate measures to determine if there is tuberculosis in the lungs. This is mainly because most of the lung is empty space, so the average and standard deviations would be close to zero. On the other hand, the maximum would show if there were higher values present in the lung, an indication of tuberculosis, which we hypothesize could also help determine the type of tuberculosis.

3.2.2 Augmentations

The augmentations applied in this approach were resizing, to have three equally sized channels, random horizontal flips, and normalization of the final image.

3.2.3 Model

The model used was one network, composed of a fine-tuned ResNet50 pretrained on ImageNet, available from the Torchvision library in PyTorch[3], using Cross Entropy as the loss function, due to the nature of the challenge being a multi-class classification problem, and Adam as an optimization algorithm. To add regularization to the model, we implemented a drop rate of 0.3. Additionally, we set class weights in our loss function to resolve the imbalanced dataset problem.

As mentioned earlier, this approach presented exploding gradient problems. When these started to appear, we implemented gradient clipping and went varying the learning rate. The learning rate that performed best, before presenting exploding gradients, had a value of 0.0001. Unfortunately, this was not enough to resolve the problem. We believe this is due to using only the maximum that, although normalized, still caused all images to present high values and prevented the network to learn correctly.

In the following subsection, we present a different approach which did not present exploding gradient issues but also implemented a modification of the volumetric squeezing approach presented in [4].

3.3 The 2D Approach Scoring the Best

In this particular approach, the first step was to apply the segmentation masks provided by ImageCLEF, particularly using the method proposed in [5], to separate the lungs of each image and delete the non-important parts of the CT scan. After that, we divided the segmented image in half, obtaining two 3D matrices, one for each lung.

Each 3D CT scan can be reduced to a 2D representation, by computing different statistics across each dimension of the image. We used the procedure suggested in [4], that consisted in computing the mean, maximum and standard deviation over the 3D images, but applying it only over the 1st and 2nd axis (lateral and superior). Using this, we obtained 2 images per matrix, which can be interpreted as an RGB image. After that we applied the preprocessing steps implemented in [4], which consisted in increasing the voxels intensity in the CT by 1024HU, dividing the maximum values by 1500, and dividing the mean values and standard deviation values by their maximum. Additionally, we resized the images to 256 x 256 pixels. In the end, we end up with 2 RGB images for each lung, which represent the lateral and superior views of them.

---

[3]https://pytorch.org/vision/stable/models.html
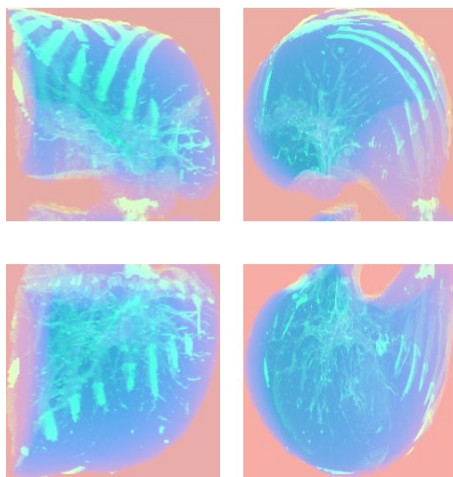
Figure 4: Images after preprocessing

3.3.1 Augmentations

The images pass through multiple augmentations, each one of them with a varying probability of being applied. In summary, the augmentations implemented were image rotation, with a rotation range of 25 degrees, width and height shift, with a shift range of 15% of the image, zoom, with a range of 20% of the image, and horizontal and vertical flipping.

3.3.2 Model

With the 2 RGB images for each lung obtained from 3.2, we further applied the augmentations in 3.2.1 to obtain the images to feed the model. We trained one network for each axis, using a fine-tuned DenseNet121 pre-trained on ImageNet, with average pooling in the output layer, and added some layers at the end to reduce dimensions. Subsequently, we concatenated the output of the two networks and added a softmax layer on top to get the final prediction. Using no regularization, we obtained a Cohen Kappa value on training and validation of approximately 0.11, meanwhile with L2 regularization we got 0.236 with 0.511 accuracy on the training set, and 0.186 with 0.467 accuracy on our validation set. On the test set we got 0.120 Cohen's Kappa with 0.401 of accuracy.

In an effort to reduce the impact of the most represented classes in the dataset, we tried weighting the loss of each label, according to it´s representation in the dataset. Regarding this, we tried with multiple configurations of fine tuning of the network, achieving the best results with the last twenty layers unfrozen. Using that configuration, we scored a Cohen's Kappa of 0.206 on training set and 0.105 on validation set.

We further tried using shared weights along the networks of the axes, in order to reduce the quantity of parameters and over fitting. This didn't improve the results, and got a best Cohen Kappa score of 0.141 on training and 0.098 on the validation set.
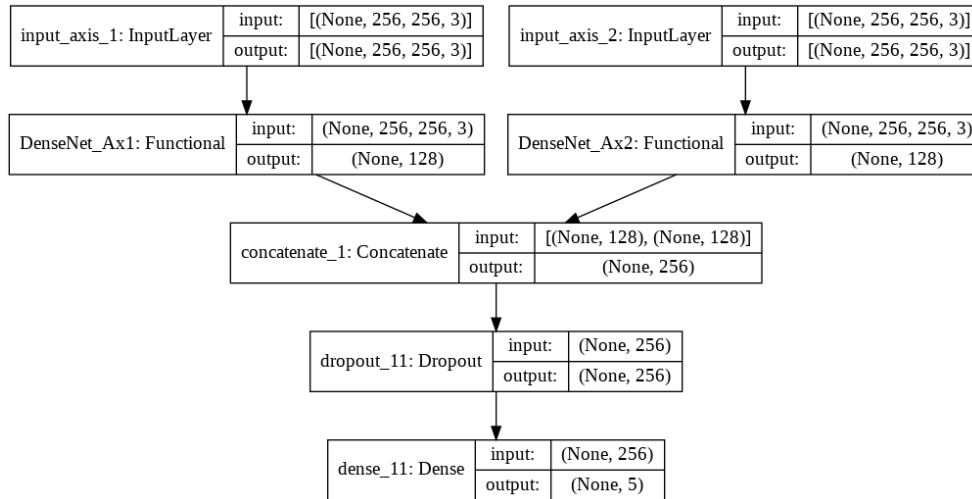
Figure 5: Visual representation of the model

3.3.3 Evaluation of the Model

In order to evaluate the model, we preprocessed the data as explained in 3.3, resulting in 2 images per lung. Then, we got the predictions for each side of the lung and sum the softmax results, keeping the highest value as the final prediction.

4 Results

In this section we present the results of our approaches at our own developing servers and also at the crowdai.org platform.

Table 2 shows the results obtained while training, including a fourth approach that was not submitted.

Table 2
Train/validation results for each approach in our development servers

Rank	Approach	Train CK	Train Acc	Val. CK	Val. Acc
1	Best performing 2D Approach	0.236	0.511	0.186	0.467
2	2D Approach with shared weights	0.141	0.316	0.098	0.194
3	Human inspired approach	-0.040	0.337	-	-
4	Simple 2D Approach	0.000	0.472	-	-

Table 3 presents the evaluation metrics obtained for each approach when evaluated with the test dataset.

*Weighted based on class prevalence

Table 3
Final results for each approach at crowdai.org

Rank	Approach	Kappa (K)	Accuracy (Acc)
1	Best performing 2D Approach	0.120	0.401
2	Human inspired approach	-0.040	0.337
3	Weighted random choice*	-0.048	0.245

5 Conclusions

In this article we have provided details of the participation of the PUC Chile team, for the Tuberculosis type (TBT) classification task within the ImageCLEFmedical challenge 2021. In the process of building our final submission, we tested several approaches. Our final submission was based on a DenseNet architecture for visually encoding the input medical volumes represented as 2D images, followed by a softmax classification layer. In future work, we plan to address the task based on the process that actual radiologists follow when classifying CT scans for Tuberculosis, which we described in section 3.1. Another idea we plan to further investigate is using perceptual image similarity [6] to leverage approaches based on K-NN, which have had interesting results in previous challenges. Finally, we plan at using methods which can directly deal with volumes (3D images) rather than 2D images.

Acknowledgements

References

[1] B. Ionescu, H. Müller, R. Péteri, 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 Multi-linguality, 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 ImageCLEFtu-berculosis 2021 - CT-based tuberculosis type classification, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.

[3] O. Mendez, S. Hadfield, N. Pugeault, R. Bowden, Sedar: Reading floorplans like a human—using deep learning to enable human-inspired localisation, International Journal of Computer Vision 128 (2020) 1286–1310.

[4] R. Miron, C. Moisii, M. Breaban, Revealing lung affections from cts. a comparative analysis of various deep learning approaches for dealing with volumetric data, arXiv preprint arXiv:2009.04160 (2020).

[5] V. Liauchuk, V. Kovalev, Imageclef 2017: Supervoxels and co-occurrence for tuberculosis ct image classification, in: CLEF2017 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Dublin, Ireland, 2017.

[6] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, 2018. `arXiv:1801.03924`.