SZTAKI @ ImageCLEFmed 2020 Tuberculosis Task

Bence Lestyan[1], András A. Benczúr[1,2], and Bálint Daróczy[1,2]

[1] Institute for Computer Science and Control (SZTAKI)
H-1111, Kende str. 13-17, Budapest, Hungary
[2] Széchenyi University
H-9026, Egyetem tér 1, Győr, Hungary
{lestyan,benczur,daroczyb}@ilab.sztaki.hu

Abstract. In this paper we describe our submission to the ImageCLEFmed 2020 Tuberculosis task and discuss additional results on the training set with various neural networks. After some centralization and normalization we independently categorized the 2D slices with convolutional neural networks (traditional and residual feed-forward networks) and we aggregated the individual predictions based on the positions of the lung and the slices. Our additional experiments with various aggregation methods indicate that individual slices do not necessary contain enough information about such complex structures.

Keywords: Computed tomography, Residual networks, Convolutional networks, Tuberculosis

1 Introduction

The goal of the ImageCLEFmed 2020 Tubercolosis task[3] [8, 6] is to detect whether the different parts of the lung are affected by Mycobacterium tuberculosis. The categories are LeftLungAffected, RightLungAffected, CavernsLeft, CavernsRight, PleurisyLeft, PleurisyRight. The data set contain 403 computed tomography scans (CT scans). Out of the 403 CT scans 283 scans are used as a training set with known labels for the participants and 120 CT scans as the test for the competition. For our experiments we split the training set into two subsets (163 as training and 120 as validation set) and evaluated our models on the smaller set. Out of the two lung masks [2, 10] we used the first segmentation method in the aggregation phase of the slice predictions.

[3] https://www.imageclef.org/2020/medical/tuberculosis

2 Models

First, we preprocessed individual CT slices. Based on the provided lung masks we centralized and rescaled the slices to lower the resolution from 512x512 to 256x256. Additionally we standard normalized the intensity values, see Fig. 1. We omitted to apply additional augmentation techniques [11] e.g. rotation, mirroring or random crop as the position of the lung is crucial. We treated the scoring procedure as a set of binary classification tasks therefore we trained separate neural networks for each category.

We chose feed-forward neural networks with a single output node to model the categories per slice. Every inner layer included Rectangular Linear Units (ReLU) as non-linear activation functions and we chose sigmoid for the output unit. We built a traditional Convolutional Neural Network (CNN [9]) with two convolutional layers with 64 5x5 sized filters and a Residual Network (ResNet [5]) with three residual blocks, for details see Table 1 and Table 2 respectively. The residual blocks contained a set of 3x3 sized convolution with 8,32,64 filters per block followed by a second convolution with the same size and a final residual connection and a downsizing unit. Between the two convolutions we used batch normalization and ReLU similarly to the original paper. Before the linear discriminative layer we downsized the tensor with average pooling. Additionally, for the CNN network we applied Dropout [11] in the second convolutional layer. We evaluated the performance of the models with the log-likelihood of the probability of the original label measured by the activation of the output unit. As an optimization method we used Adam [7] thus we omitted additional regularization in the loss function.

We measured the performance of various models on the validation set, a random subset of the training set. In the testing phase we used every training scan with the best settings. We implemented the models in PyTorch framework [4] and did all the experiments in python. Additionally, we used the provided lung masks based on the first automatic segmentation method described in [2].

2.1 Aggregation

We combined the individual predictions of the slices to compute a single score per CT scan. During our experiments we applied various methods to define a single score:

- Mean score (sc_1): mean of the individual prediction scores of the CT scan.
- Maximal score (sc_2): maximal prediction score in a CT scan.
- Minimal score (sc_3): minimal prediction score in a CT scan.
- Median score (sc_4): median prediction score in a CT scan.
- Middle score (sc_5): prediction score of the center slice.
- Majority vote (sc_6): proportion of the positive predictions.

[4] https://pytorch.org

- Mask score (sc_7): weighted prediction scores based on the proportion of the actual lung in the slices. The proportion of the lung was the proportion of the lung segment given the mask files. The masks were extracted by a fully automatic lung segmentation method described in [2]. We used the corresponding masks per lung per task.

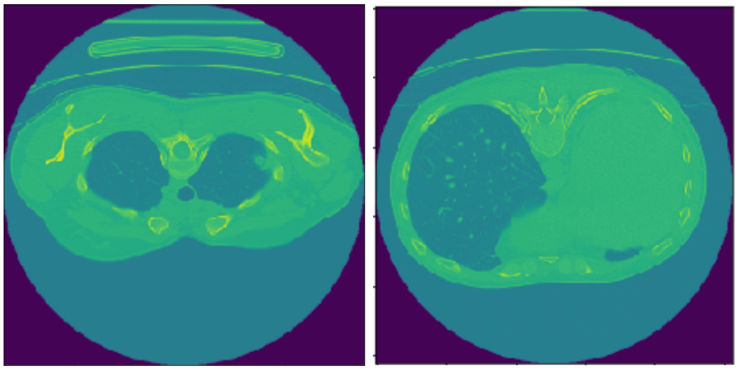

Fig. 1: Examples of modified CT slices. Note, standard normalization is a linear transformation.

Table 1: Convolutional network layout. We denote 2D convolution, maximal pooling [9] and Dropout [11] with C, M and DO respectively.

Layer	**#nodes**	**#parameters**	**output**
C5x5 + M2x2	64	1.6k	64x126x126
C5x5 + M2x2 + DO	64	102k	64x61x61
Output layer	1	238k	1

3 Results

All of our submitted runs included mean score over the CNN results. Our main submission (#68061) achieved mean AUC of 0.595. The remaining runs contained single category scores with random scores for the rest of the categories (we estimated the AUC as $AUC_{mean} * 6 - 0.5 * 5$)). The estimated per category AUC of our submission can be seen in Table 3. We noticed that two of the categories achieved an AUC under 0.5 thus if we negate the scores the AUC values will flip to the upper half and the adjusted mean AUC will be 0.6548. Important to mention, that these adjustments only provide us information about

Table 2: Residual network layout.

Layer	#nodes	#parameters	output
Input layer	16	0.2k	16x256x256
Residual layer 1		3k	8x256x256
Residual layer 2		14k	32x128x128
Residual layer 3		73k	64x64x64
Average pooling 8x8		0	64x8x8
Output layer	1	4k	1

the distinguishing capability of the models (how the model differentiate negative and positive examples), in a realistic scenario the final decisions would be still wrong as without any test data we would not know that we need to flip the scores. During the challenge and afterwards we experimented over the small training (163 CT scans) and validation (120 CT scans) sets with several models and aggregation methods. Table 4 show the mean AUC results on the validation set and the detailed AUC scores can be seen for the left and right lung in Table 5 and in Table 6 respectively. The method (mean score of CNN) in our main submission achieved a mean AUC 0.595 on the validation set however the best method (median score of CNN) performed significantly better with AUC of 0.642. If we select the best model (ResNet or CNN) with the median score per category the mean AUC will be similar to the median CNN with 0.659. In comparison, if we select properly both the model and the aggregation method for each category the mean AUC increases to 0.686, a significant gain on the validation set in comparison to the submitted run.

Table 3: Estimated individual AUC values.

category	run	AUC
Estimated LeftLungAffected	#68052	0.734
Estimated CavernsLeft	#68055	0.452
Estimated PleurisyLeft	#68050	0.728
Estimated RightLungAffected	#68059	0.74
Estimated CavernsRight	#68049	0.41
Estimated PleurisyRight	#68058	0.674
mean	#68061	0.595
mean adjusted		0.6548

4 Conclusions and Future Work

In this paper we described our submission and some additional experiments over the data set of the ImageCLEFmed 2020 Tuberculosis task. We trained traditional feed-forward convolutional and residual neural networks over the in-

Table 4: Mean AUC results on the validation set.

model	**aggregation**	**mean AUC**
CNN (submitted)	sc_1	0.595
CNN	sc_2	0.594
CNN	sc_3	0.614
CNN	sc_4	0.642
CNN	sc_5	0.626
CNN	sc_6	0.558
CNN	sc_7	0.591
ResNet	sc_1	0.620
ResNet	sc_2	0.614
ResNet	sc_3	0.6
ResNet	sc_4	0.584
ResNet	sc_5	0.616
ResNet	sc_6	0.577
ResNet	sc_7	0.614
Best model	sc_4	0.659
Best model & aggr.		0.686

dividual slices of the CT scans and combined the predictions based on the importance of the slices according to their position and how well they represent both of the lungs. We found that median score performed best on average although in some categories the middle slice score or the mask score outperformed other aggregation methods. Both ResNet and traditional CNN performed similarly in our experiments on the validation set while the residual network needed significantly higher computational power. Our simplest run which was submitted to the challenge had very low mean AUC score 0.595 meanwhile with additional aggregations we improved the same method on the validation set to achieve a mean AUC 0.684. We plan to replace 2D convolutions with 3D convolutions to take advantage of the complex structure of CT scans. Additionally, we intend to further expand our experiments with bi-directional Recurrent Neural Networks (RNN [4]) to read through the CT scans from both ends and classify the sequence as a whole, utilize Markov Random Fields [1] over the prior predictions and generate additional samples with slice transition refinement with inter-slice reconstruction and with category-wise Generative Adversarial Networks [3] to boost the training procedure. Based on the submissions of other participants (SenticLab.UAIC mean AUC 0.924 or SDVA-UCSD mean AUC 0.875) we believe individual slice predictions may not be representative enough to describe CT scans as a whole to detect Mycobacterium tuberculosis.

5 Acknowledgement

The publication was supported by the Hungarian Government project GINOP-2.2.1-18-2018-00004: AI based lung cancer diagnosis by chest CT, 2018-1.2.1-NKP-00008: Exploring the Mathematical Foundations of Artificial Intelligence,

Table 5: Per category AUC results on the validation set for the left lung. The best results are highlighted in red.

model	**LeftLungAffected**	**CavernsLeft**	**PleurisyLeft**
CNN sc_1	0.675	0.506	0.575
CNN sc_2	0.593	0.556	0.512
CNN sc_3	0.762	0.525	0.625
CNN sc_4	0.612	0.7	0.631
CNN sc_5	0.717	0.525	0.575
CNN sc_6	0.725	0.503	0.5
CNN sc_7	0.706	0.506	0.643
ResNet sc_1	0.687	0.681	0.618
ResNet sc_2	0.593	0.7	0.575
ResNet sc_3	0.706	0.581	0.637
ResNet sc_4	0.65	0.587	0.515
ResNet sc_5	0.668	0.681	0.562
ResNet sc_6	0.662	0.628	0.5
ResNet sc_7	0.743	0.637	0.612

by the Higher Education Institutional Excellence Program, and by the Momentum Grant of the Hungarian Academy of Sciences. B.D. was supported by an MTA Premium Postdoctoral Grant 2018.

Table 6: Per category AUC results on the validation set for the right lung. The best results are highlighted in red.

model	**RightLungAffected**	**CavernsRight**	**PleurisyRight**
CNN sc_1	0.712	0.543	0.543
CNN sc_2	0.7	0.625	0.581
CNN sc_3	0.693	0.55	0.531
CNN sc_4	0.712	0.525	0.675
CNN sc_5	0.712	0.593	0.631
CNN sc_6	0.5	0.575	0.546
CNN sc_7	0.543	0.581	0.568
ResNet sc_1	0.562	0.6	0.575
ResNet sc_2	0.537	0.612	0.668
ResNet sc_3	0.587	0.543	0.543
ResNet sc_4	0.55	0.587	0.618
ResNet sc_5	0.575	0.593	0.618
ResNet sc_6	0.562	0.578	0.531
ResNet sc_7	0.5	0.581	0.612

References

1. Daróczy, B., Vaderna, P., Benczúr, A.: Machine learning based session drop prediction in lte networks and its son aspects. In: 2015 IEEE 81st Vehicular Technology Conference (VTC Spring). pp. 1–5. IEEE (2015)
2. 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. pp. 31–35. CEUR Workshop Proceedings, CEUR-WS (May 2015)
3. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
4. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: Lstm: A search space odyssey. IEEE transactions on neural networks and learning systems **28**(10), 2222–2232 (2016)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
6. Ionescu, B., Müller, H., Péteri, R., Abacha, A.B., Datla, V., Hasan, S.A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G.: Overview of the ImageCLEF 2020: Multimedia Retrieval in Medical, Lifelogging, Nature, and Internet Applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), vol. 12260. LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece (September 22-25 2020)
7. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Kozlovski, S., Liauchuk, V., Dicente Cid, Y., Tarasau, A., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2020 - automatic CT-based report generation. In: CLEF2020 Working Notes. CEUR Workshop Proceedings
9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
10. Liauchuk, V., Kovalev, V.: Imageclef 2017: Supervoxels and co-occurrence for tuberculosis ct image classification. In: CLEF2017 Working Notes. CEUR Workshop Proceedings, CEUR-WS, Dublin, Ireland (September 11-14 2017)
11. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)