ImageCLEF 2019: CT Image Analysis for TB Severity Scoring and CT Report Generation using Autoencoded Image Features

Siarhei Kazlouski

United Institute of Informatics Problems, Minsk, Belarus
kozlovski.serge@gmail.com

Abstract. This paper presents a possible approach for the automated analysis of 3D Computed Tomography (CT) images based on the usage of feature vectors extracted by a deep convolutional 3D autoencoder network. Conventional classification models were used on top of the "autoencoded" feature vectors as well as vectors of meta-information paired with the images. The proposed CT image analysis approach was used by participant UIIP (Siarhei Kazlouski) for accomplishing the two subtasks of the ImageCLEF Tuberculosis task of the ImageCLEF 2019 international competition. Employing the proposed approach allowed to achieve the 2nd best performance on the TB Severity Scoring subtask and the 6th best performance in the TB CT Report subtask.

Keywords: Computed Tomography, Tuberculosis, Deep Learning, Autoencoder

1 Introduction

Automated analysis of 3D CT images is an example of a task that can be solved during the development of computer assisted diagnosis systems which may be used for lung disease screening for the early detection of pathology. While promising results have been shown in automated analysis of medical images of some modalities [1, 4, 7–9], the task of CT image analysis remains challenging due to the complexity and scarcity of data. A CT image is 3D data which can often be represented as a set of 2D slices with the inter-slice distance varying between 0.5 and 5 mm. Variability in the sizes and shapes of CT image voxels implies difficulties in the application of many image analysis algorithms, while low availability of CT imaging data makes it difficult to use data-greedy approaches, for example, deep learning.

Despite the lack of data available, the approach for the analysis of 3D CT images proposed with this study employs the idea of trying to get 3D image descriptors by utilizing a 3D autoencoder network [6]. The motivation for this idea

is the potential for maximum information usage as soon as the network is able to work with the entire 3D image. Extracted features were further analyzed using conventional classification models which were ensembled with models trained on image metadata. One can note, that pure 3D classification networks could be used instead of conventional models on top of autoencoded features. The advantage of the used approach is its generality: once we got autoencoded descriptors, a conventional classification model could be easily and quickly trained for arbitraty labelling (either TB severity, or one provided in CT report findings, or some arbitrary findings classes which are not mentioned in the competition), as well as research interest.

2 Subtasks and datasets

The Tuberculosis task [2] of ImageCLEF 2019 Challenge [5] included two subtasks: TB Severety (SVR) and CT Report (CTR), both dealing with 3D CT images. The same CT imaging data was used in both subtasks and included 218 images in the training dataset and 117 in the test dataset. Along with the CT images, the lungs masks [3] and additional information about the patients was provided. The metadata included information about the presence of disability, relapse, presence of TB symptoms, co-morbidity, bacillarity, drug resistance status, patient's education, being an ex-prisoner, smoking status and alcohol addiction. The frequencies of occurrence of each metadata label are listed in Table 1.

Table 1. Frequences of positive metadata labels for SVR subtask in the datasets, %.

Label	**In Training set**	**In Test set**
Disability	16	13
Relapse	35	36
SymptomsOfTB	54	40
Comorbidity	56	48
Bacillary	85	92
HigherEducation	13	16
ExPrisoner	12	10
Alcoholic	22	25
Smoking	52	60

The subtask #1 (SVR subtask) was dedicated to the problem of categorizing TB cases into one of two classes: high severity and low severity. The task was to predict TB Severity class ("HIGH"/"LOW")

The subtask #2 (CTR subtask) was dedicated to the automated generation of CT reports which indicate the presence of several types of abnormalities in the lungs. Such automated annotation of CT scans is important for the development of the dedicated image databases. The task was to predict the presence of six

types of findings in CT scans. Information about the corresponding labels is listed in Table 2.

Table 2. Frequences of positive labels for Severity and CTR subtasks in the Training dataset.

Finding	In Training set
SeverityHigh	49
LeftLungAffected	72
RightLungAffected	81
LungCapacityDecrease	30
Calcification	13
Pleurisy	7
Caverns	40

3 Methods

This section contains a description of the methods used within the current study.

3.1 Data preprocessing

Since the key idea of the approach proposed is based on using an autoencoder network, the extremely small amount of data samples we have becomes the main challenge, especially keeping in mind the sample dimensionality. The second issue is the sample dimensionality itself, which restricts possible model architectures due to GPU memory limitations.

In order to overcome the mentioned issues the following concepts were used: a) image size was reasonably decreased, b) the autoencoder was trained not on per-CT, but on per-lung basis which simultaneously doubled the training dataset size and decreased the sample dimensionality two times. Data augmentation was also used. Detailed steps of data preprocessing during the training stage were as follows:

1) Each CT image was split into two parts, each containing one lung. The split was performed roughly, into equal parts by splitting on the middle Y coordinate. The part containing the left lung was used as it is, and the part containing the right lung was reflected along Y axis in order to make the right lung oriented similar to left one. Lung images were normalized to 0-1 scale. The provided lungs masks were treated the same way (except for normalization).

2) For each lung a random transformation was generated and applied to the lung and its mask. The mask was binarized after the transformation and applied to the lung image (all voxels outside of the masked lung area were set to zero). The resulting image was resized to 128 x 128 x 128 pixels and normalized once again. Transformations included 3D shift, rotation, scale, crop and shear, and were applied sequentially with a probability of 50% for each transform. The ranges of parameters used for the transformation are presented in Table 3.

Table 3. Parameters of augmentation applied to the Training dataset.

Transform	Parameter value
Shift, pixels	up to 5% of each side (X, Y, Z) size
Rotation center	Shifted from image center up to 3% of each side (X, Y, Z) size
Rotation angle	up to 5 degrees
Scaling	up to 5%
Shear	up to 0.01 absolute value, each of six components

3.2 Training and validation subsets

The training dataset provided by the organizers was split into training and validation subsets for models development.

Two splits were used, one for autoencoder model training, and another one for classification models training. For autoencoder training, a randomly sampled 90% of training data was used for training and the remaining 10% was used for validation. For classification models the fixed random 5-fold cross-validation split was used. Smaller validation size in the autoencoder case is motivated by maximizing training set size, while validation loss for the autoencoder model is less important. For classification tasks scoring is crucial, so cross-validation was used.

3.3 Autoencoder Training

A custom convolutional architecture was used for the autoencoder model. Several trials with different hyperparmeters (number of layers, kernel size, filter number) were exectuted and the model with the best achieved validation loss was selected. The selected architecture is presented in Table 4. Each convolutional layer had a kernel size of (3,3,3) and was followed by a 3D max pooling layer with parameters (2,2,2) in the encoder part and a 3D upsample layer with parameter (2,2,2) in the decoder part. This setup resulted in an encoded feature vector of size 256.

The autoencoder model was trained using Adam optimizer and was performed in three stages. On the first stage the model was trained on the mixture of left and right lung images using data augmentation as described before. The training was performed until any significant improvements in validation loss were observed. Specifically, model trained for 72 epochs was selected on this stage. On the second stage the retrieved pretrained model was finetuned with a 10 times smaller learning rate separately for the left and right lungs using data augmentation, resulting in two different models, one for the left and one for the right lungs. Training stop criteria was the same, resulting in 20 epochs of training. So, on the first two stages 92 randomly augmented versions of each original CT were used for model training. Finally, learning rate was decreased 10 times again and each of the two models were finetuned for 2 epochs on competition data without data augmentation.

On the inference stage both autoencoder models were used to generate feature vectors for the left and right lung of each CT image, so the resulting feature

Table 4. Autoencoder architecture.

Layer	Number of filters
Convolution3D	128
Convolution3D	64
Convolution3D	64
Convolution3D	32
Convolution3D	32
Convolution3D	32
Flatten	
Convolution3D	32
Convolution3D	32
Convolution3D	32
Convolution3D	64
Convolution3D	64
Convolution3D	128

vectors of CT were a concatenation of the left and right lung encoded descriptors, with 512 components in total.

Analysis of encoded vectors showed that around half of the components of the vector are very close to zero for all images, which probably reflects the non-optimal architecture and/or weights of model. As soon as no better model could be retrieved in experiments, it was decided just to drop the "zero" components of the encoded vector, which resulted in the final version of the encoded CT images descriptors containing 220 components each.

3.4 Prediction of CT Report labels and TB Severity.

Since all of the predictions in both subtasks are binary classification problems, they were treated and solved in the same way. The idea may be formulated as follows: for each of the requested binary class labels (which are: Severity-High, LeftLungAffected, RightLungAffected, LungCapacityDecrease, Calcification, Pleurisy, Caverns) build a binary classification model which will take encoded image descriptors and available meta information about patients as input features.

Because of the different nature of encoded image descriptors and image meta information it was decided to train separate classification models for each feature type and then ensemble the models' output rather than concatenating feature vectors.

Conventional classification models were used for working with both types of features and included scikit-learn package implementation of SVM, K-neighbors classifier (kNN), random forest classifier (RF) and AdaBoost classifier. Hyper parameters for each of the models were tuned by cross-validation using random parameter search.

In case of meta information features were used "as is", while for autoencoded image features their PCA-transformed presentation with 3,5,10,50 components were used as well.

The resulting algorithm can be described as follows.

1) For each of 5 folds, take the autoencoded features and fit PCA using the training set and transform validation set using 3, 5, 10, 50 components. Thus six alternative feature vectors were presented for each image in each cross-validation split: meta information (META), encoded features (AEC), and PCA-transformed encoded features with 3, 5, 10, 50 components (PCA3, PCA5, PCA10, PCA50).

2) Sample random hyper parameters for each of 4 classifier types. Parameter ranges are presented in Table 5 (only valid parameters combinations were used).

For each of the target labels:

3) Get the mean AUC-ROC score at cross-validation for all models and sampled hyper parameters.

4) Select the best models according to the achieved scores. Models selection was performed not just by the top score value, but using the following heurisctic: 1) classifier-feature type combination were used only once, 2) in the case that scores are reasonably close, more simple models have priority (examples: a) if the kNN model with 11 neighbors scores 0.80 and with 2 neighbors scores 0.78, the second one is used; b) if the same model scores 0.8 on the 50 component PCA features and 0.77 on the 3-component, the second one is used).

Table 5. Classifier parameters search ranges.

Classifier	**Parameters range (format: min .. max .. step)**
KNeighborsClassifier	neighbors number: 1 .. 20 .. 1)
RandomForestClassifier	estimators number: 1 .. 100 .. 5
	max tree depth number: 1 .. 6 .. 1
AdaBoostClassifier	estimators number: 1 .. 200 .. 5
	learning rate : 0.1 .. 1 .. 0.2
SVM	C: 10e-5 .. 10e5 .. by power of 10
	degree : 1 .. 7 .. 1
	kernel: 'linear', 'poly', 'rbf'

The application of the described algorithm resulted in the selection of from 1 to 4 best models for each target class prediction, which were ensembled during final prediction by the simple averaging of class probabilities. A summary on the selected models and their validation performance is presented in Table 6.

4 Submissions and results

As the result of this study, the described method was applied to generate predictions for the competition test dataset. Predictions were submitted by participant UIIP for the CTR and SVR subtasks. The full list of the submitted results for both subtasks is available at the task web page[1].

[1] https://www.imageclef.org/2019/medical/tuberculosis/

Table 6. Selected classification models.

Label	Classifier	Features	AUC-ROC
LeftLungAffected	RF(E*=20, D**=1)	PCA5	0.77
	SVM(linear, C=10e5)	PCA5	0.81
RightLungAffected	RF(E=20, D=1)	PCA5	0.79
	SVM(rbf, C=10e5)	PCA5	0.77
Calcification	SVM(poly, d=2, C=0.01)	META	0.84
	RF(E=16, D=1)	PCA10	0.80
	kNN(neighbors=11)	PCA5	0.82
	SVM(linear, C=10e3)	AEC	0.90
Caverns	AdaBoost(E=15, lr=0.2)	AEC	0.89
Pleurisy	RF(E=6, D=2)	META	0.82
	RF(E=20, D=1)	AEC	0.89
	RF(E=10, D=1)	PCA10	0.90
LungCapacityDecr.	RF(E=6, D=2)	META	0.78
	RF(E=20, D=1)	AEC	0.83
	SVM(linear, C=10e3)	PCA5	0.72
Severity	kNN(neighbors=11)	PCA5	0.71
	RF(E=8, D=2)	PCA10	0.82
	RF(E=30, D=2)	AEC	0.87

*estimators number
**max depth

Before generating the final submission, the autoencoder model and selected classifiers used for predicting the test data were trained on the whole available training dataset. Final probabilities of each target class were calculated as the average probability of the selected classification models.

Table 7 shows the best results achieved by the participants in the CTR subtask. The run submitted by UIIP achieved the 6th best mean AUC, while also demostrating the worst minimum AUC for the prediction of the presence of lung abnormalities. The average mean result on the contrary to the worst minimum AUC demonstrate, that selected models might work pretty well for some of the target labels in the report, while failing for other labels. Since all targets demonstrate similar performances on the validation set, test results may be caused by overfitting to validation and the different distribution of train and test sets in the competition, or by some errors in validation set generation (in particular, the uneven distribution of meta information and targets classes itself was not carefully treated in experiments).

Table 8 shows the best results achieved by the participants in the SVR subtask. The run submitted by UIIP achieved the 2nd highest value for AUC, and shared the 1st best accuracy with UIIP_BioMed participant. Achieved AUC correlates with validation scores and demonstrates the efficiency of the used approach.

Table 7. The best participants' runs submitted for the CTR subtask.

Group Name	Mean AUC	Min AUC	Rank
UIIP_BioMed	0.7968	0.6860	1
CompElecEngCU	0.7066	0.5739	2
MedGIFT	0.6795	0.5626	3
San Diego VA HCS/UCSD	0.6631	0.5541	4
HHU	0.6591	0.5159	5
UIIP	0.6464	0.4099	6
MostaganemFSEI	0.6273	0.4877	7
UniversityAlicante	0.6190	0.5366	8
PwC	0.6002	0.4724	9
LIST	0.5523	0.4317	10

Table 8. The best participants' runs submitted for the SVR subtask.

Group Name	AUC	Accuracy	Rank
UIIP_BioMed	0.7877	0.7179	1
UIIP	0.7754	0.7179	2
HHU	0.7695	0.6923	3
CompElecEngCU	0.7629	0.6581	4
San Diego VA HCS/UCSD	0.7214	0.6838	5
MedGIFT	0.7196	0.6410	6
UniversityAlicante	0.7013	0.7009	7
MostaganemFSEI	0.6510	0.6154	8
SSN College of Engineering	0.6264	0.6068	9
University of Asia Pacific	0.6111	0.6154	10
FIIAugt	0.5692	0.5556	11

5 Conclusions

The results of this study allow us to draw the following conclusions:

- Despite data scarcity, deep autoencoder networks may be used for extracting reasonable descriptors of 3D CT data if some tricks for training set extension are used.
- Meta information about patients are helpful for the more accurate predictions of TB characteristics.
- Although the used approach demonstrated good performance in the SVR subtask, it was not very reliable for the generation of the CT report, which means the suggested method is not very stable or at least needs more careful validation.

Acknowledgements

This study was partly supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, U.S. Department of Health and

Human Services, USA through the CRDF project DAA3-18-64818-1 "Year 7: Belarus TB Database and TB Portals".

References

1. Al-Kofahi, Y., Zaltsman, A., Graves, R., Marshall, W., Rusu, M.: A deep learning-based algorithm for 2-D cell segmentation in microscopy images. BMC Bioinformatics **19**(1), 365 (Oct 2018). https://doi.org/10.1186/s12859-018-2375-z, https://doi.org/10.1186/s12859-018-2375-z
2. Dicente Cid, Y., Liauchuk, V., Klimuk, D., Tarasau, A., Kovalev, V., Müller, H.: Overview of ImageCLEFtuberculosis 2019 - automatic ct-based report generation and tuberculosis severity assessment. In: CLEF2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 9-12 2019)
3. 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)
4. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., , the CAMELYON16 Consortium: Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA **318**(22), 2199–2210 (12 2017). https://doi.org/10.1001/jama.2017.14585, https://doi.org/10.1001/jama.2017.14585
5. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
6. Krizhevsky, A., E. Hinton, G.: Using very deep autoencoders for content-based image retrieval. 19th European Symposium on Artificial Neural Networks, Bruges, Belgium (January 2011)
7. Liauchuk, V., Kovalev, V.: Detection of lung pathologies using deep convolutional networks trained on large X-ray chest screening database. In: Proceedings of the 14th international conference on Pattern Recognition and Information Processing (PRIP'2019). Minsk, Belarus (May 21-23 2019)
8. Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., Hedlund, M., Tellez, D., Ciompi, F., Zerhouni, E., Lanyi, D., Viana, M., Kovalev, V., Liauchuk, V., Phoulady, H.A., Qaiser, T., Graham, S., Rajpoot, N., Sjblom, E., Molin, J., Paeng, K., Hwang, S., Park, S., Jia, Z., Chang, E.I.C., Xu, Y., Beck, A.H., van Diest, P.J., Pluim, J.P.: Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. Medical Image Analysis **54**,

111 – 121 (2019). https://doi.org/https://doi.org/10.1016/j.media.2019.02.012, http://www.sciencedirect.com/science/article/pii/S1361841518305231
9. Zaidi, S.M.A., Habib, S.S., Van Ginneken, B., Ferrand, R.A., Creswell, J., Khowaja, S., Khan, A.: Evaluation of the diagnostic accuracy of computer-aided detection of tuberculosis on chest radiography among private sector patients in Pakistan. Scientific reports **8**(1), 12339 (2018). https://doi.org/10.1038/s41598-018-30810-1