Full training versus fine tuning for radiology images concept detection task for the ImageCLEF 2019 challenge

Priyanshu Sinha[1], Saptarshi Purkayastha[2], and Judy Gichoya[3]

[1] Mentor Graphics India Pvt. Ltd.
priyanshu_sinha@outlook.com
[2] Indiana University Purdue University, Indianapolis, IN 46202 USA
saptpurk@iupui.edu
[3] Oregon Health Science University, Portlnd, OR 97239
gichoya@ohsu.edu

Abstract. Concept detection from medical images remains a challenging task that limits implementation of clinical ML/AI pipelines because of the scarcity of the highly trained experts to annotate images. There is a need for automated processes that can extract concrete textual information from image data. ImageCLEF 2019 provided us a set of images with labels as UMLS concepts. We participated for the first time for the concept detection task using transfer learning. Our approach involved an experiment of layerwise fine tuning (full training) versus fine tuning based on previous reported recommendations for training classification, detection and segmentation tasks for medical imaging. We ranked number 9 in this year's challenge, with an F1 result of 0.05 after three entries. We had a poor result from performing layerwise tuning (F1 score of 0.014) which is consistent with previous authors who have described the benefit of full training for transfer learning. However when looking at the results by a radiologist, the terms do not make clinical sense and we hypothesize that we can achieve better performance when using medical pretrained image models for example PathNet and utilizing a hierarchical training approach which is the basis of our future work on this dataset.

Keywords: Transfer Learning · Layer wise Fine Tuning · Deep Learning in Radiology.

1 Introduction

Concept detection from medical images remains a challenging task that limits implementation of clinical ML/AI pipelines because of the scarcity of the highly trained experts to annotate images. ImageCLEF is an annual challenge now in its

third year that seeks contributions that provide techniques to map visual information to condensed textual descriptions. The process of automatic extraction of high level concepts from low level features is difficult when the images have occlusion, background clutter, intra-class variation, pose and lighting changes[1].

Participants from past challenges in 2017 and 2018 noted a broad range of content and hence the 2019 [2] challenge was narrowed down in focus to only radiology images [3]. The focus on concept detection in the 2019 challenge is important because it is the first step of automatic image captioning, while also providing metadata to support context based image retrieval.

This was our first time participating in the ImageCLEF challenge. The challenge is a multi-label classification problem, where one radiology image can have multiple labels. Previous participants had good performance when using transfer learning, hence we focussed on optimizing the ResNet50 [5] network which had the best performance compared to VGG19 [4], Xception Net [6] and Inception-ResNetV2 [7]. We ranked number 9 in this year's challenge, with an F1 result of 0.05 after three entries. We had a poor result from performing layerwise tuning (F1 score of 0.014) which is consistent with previous authors who have described the benefit of full training during transfer learning. However when looking at the results by a radiologist, the terms do not make clinical sense and we hypothesize that we can achieve better performance when using medical pretrained image models for example PathNet which is the basis of our future work on this dataset. We describe our approach in detail in the remaining sections of this paper.

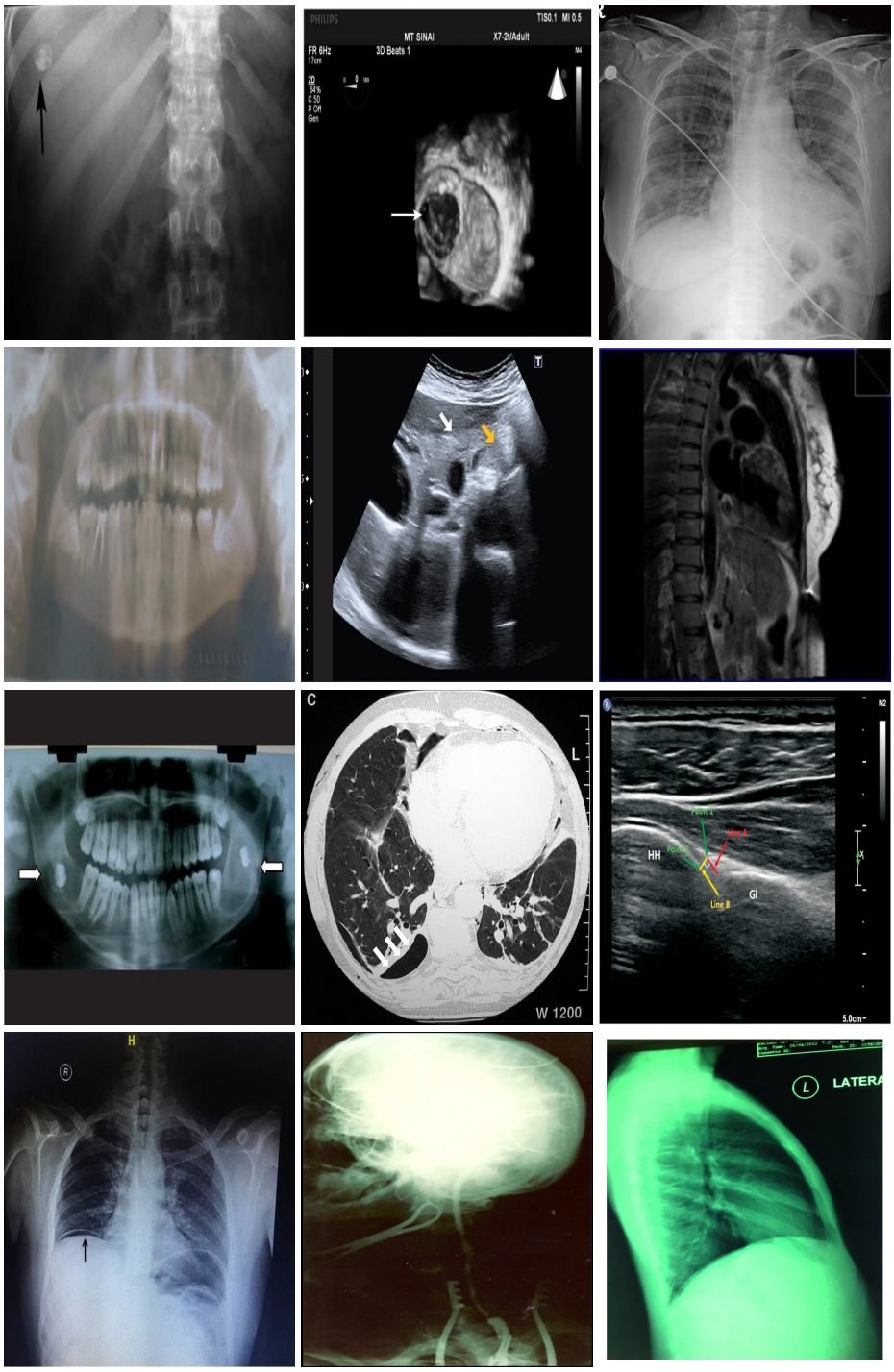
2 DATASET

A total of 6,031,814 image - caption pairs were extracted from PubMed Open Access and after processing were reduced to 72,187 radiology images from various modalities. This dataset included archived images from February 2018 to February 2019 [3]. Table 1 shows a summary of the images in the training, test and validation set. We did not use additional radiology training data for the purpose of our submission to this challenge. Each label is a UMLS concept provided as a csv file. Table 2 shows a representative sample of the data showing images in the training (First row), validation (second row) and test set (third row).

Table 1. Data and corresponding number of images.

Set	No of Images
Training Set	56629
Validation Set	14157
Test Set	10000

Table 2. Table displaying image examples from the training set (first two row), validation set (third row) and test set (fourth row).



3 STUDY EXPERIMENT

3.1 Data Analysis

The ImageCLEF images were formatted to the Imagenet directory style where the directory name is the UMLS label. This was because our approach was mainly based on transfer learning and would make repeat experiments easy to perform. Summary statistics of the dataset found *5217 unique UMLS/label concepts*. There was image imbalance with approximately 90% of the labels containing less than 100 images; and 30% labels containing a single image. Table 3 shows the top 10 concepts occurring in the highest frequency in the training set.

Table 3. Top 10 training concepts.

Concept	Frequency
C0441633	6733
C0043299	6321
C1962945	6318
C0040395	6235
C0034579	6127
C0817096	5981
C0040405	5801
C1548003	5159
C0221198	4513
C0772294	4512

Analysis of the top 25 labels (summarized in Figure 1) show that there is persistent data imbalance with one label containing more than 6500 images (C0441633 - "Scanning") and one label containing less than 2000 images (C0006104 - "Brain"). We therefore discarded labels containing less than 1000 images and used *class_weight technique from sklearn* for balancing our training data [8].

3.2 Training

Each input image was resized into 224x224 pixels without cropping. We used a batch size of 32 with learning rate 0.0001. The batches were formed by randomly shuffling the dataset. Optimization was performed using Adam optimizer with default beta_1 (0.9) and beta_2 (0.999). Image augmentation during training was performed using the Keras ImageDataGenerator. Augmentations performed include rescaling, rotation, zooming, shearing and horizontal flipping. A total of 100 epochs were executed. We split the data to 85% training set and 15% validation set. The network was trained using the Keras framework with tensorflow as the backend, running on a NVIDIA Quadro P6000 GPU.

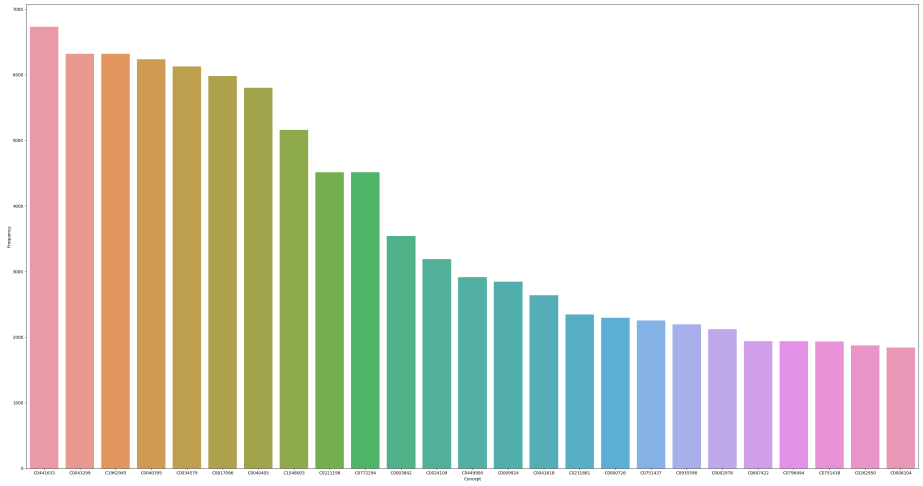


Fig. 1. Frequency plot of top 25 concepts.

We treated this as multi label classification problem and limited our training to the top 25 labels. Our base model was ResNet50, from which we removed the fully connected top layers and added our own auxiliary convolutional layer along with dense layers. To prevent overfitting, we used dropout between dense layers. After evaluating our performance with fine tuning the last layers and reviewing the literature on fine tuning versus full training [9], we embarked on layerwise fine tuning using Resnet50(run 2). In the second run we sequentially trained each layer while freezing others. For this approach we decreased the learning rate for higher layers and fine tuned it layer wise by unfreezing layers below a particular layer.

4 Evaluation and Analysis

Tajbakhsh et al [9] performed the most comprehensive experiment evaluating the approach of fine tuning a network versus training a network from scratch. In their review of classification, detection and segmentation tasks using multiple imaging modalities including radiology, colonoscopy and ultrasound, they demonstrate better performance with layerwise fine tuning. Our attempt to replicate their superior performance when approaching concept detection task on the ImageCLEF 2019 dataset led to lower performance when layerwise fine tuning (F1 score of 0.014) versus whole fine tuning the network as a whole (F1 score 0.05) summarized in table 4. Our poor comparative performance may be due to poor selection of hyperparameters for fine tuning the network.

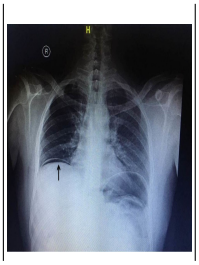
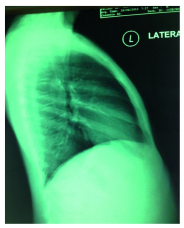
Our approach included a clinical review of some of the sample output by a radiologist who is one of the authors of this paper, and we notice a large discrepancy in the utility of the generated concepts (Table 5). For example the first

Table 4. F1 Score of Different Runs.

Run	ID	F1 Score
Run 1	26815	0.05
Run 2	27011	0.014

row demonstrated a chest xray with a pneumoperitoneum, and our model does not generate terms closely related to the actual radiograph interpretation. We hypothesize that a stepwise approach to training where ontology hierarchies for example laterality and anatomy are maintained may generate a superior performance that is clinically meaningful.

Table 5. Sampled images from the test dataset with the generated concepts and radiologist generated terms.

Sample Image	Concept Detected	Radiologist generated terms
	– C0751437(adenohypophyseal dis) – C0079595 (diagnostic imaging technique) – C0023884 (gastrointestinal tract) – C0003842 (arterial)	– Chest Xray – pneumoperitoneum – frontal radiograph
	– C0751438 (posterior pituitary dis) – C0013516 (dx ultrasound-heart) – C0221874 (spacers) – C0042449 (venous subtree)	– Skull radiograph – cervical hardware – mandibular fusion – hardware – lateral view
	– C0023884(gastrointestinal tract)	– Lateral radiograph – Chest Xray

5 CONCLUSION

Despite previous documentations of superior performance with layer wise fine tuning of medical image tasks, we had a poor performance with this approach for concept detection. There is an opportunity to improve on layer wise fine tuning for such tasks. We advance the challenge by reviewing clinical relevance of output, for which despite our performance at number 9 in the challenge we found that the clinical utility of the concepts detected was low and hypothesize that we can achieve better performance and improved clinical utility using a hierarchical approach to training.

References

1. D. Katsios and E. Kavallieratou, "Concept Detection on Medical Images using Deep Residual Learning Network," CLEF, 2017.
2. Ionescu, Bogdan and Müller, Henning and Péteri, Renaud and Dang-Nguyen, Duc-Tien and Piras, Luca and Riegler, Michael and Tran, Minh-Triet and Lux, Mathias and Gurrin, Cathal and Cid, Yashin Dicente and Liauchuk, Vitali and Kovalev, Vassili and Ben Abacha, Asma and Hasan, Sadid A. and Datla, Vivek and Liu, Joey and Demner-Fushman, Dina and Pelka, Obioma and Friedrich, Christoph M. and Chamberlain, Jon and Clark, Adrian and de Herrera, Alba García Seco and Garcia, Narciso and Kavallieratou, Ergina and del Blanco, Carlos Roberto and Rodríguez, Carlos Cuevas and Vasillopoulos, Nikos and Karampidis, Konstantinos, "Overview of ImageCLEF 2019 : Challenges, Datasets and Evaluation", Experimental IR Meets Multilinguality, Multimodality, and Interaction, in Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019)
3. O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology objects in context (ROCO): A multimodal image dataset," in Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings, vol. 11043, D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, and P. Jannin, Eds. Cham: Springer International Publishing, 2018, pp. 180–189.
4. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
5. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv, Sep. 2014.
6. F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800–1807.
7. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," arXiv, Feb. 2016.
8. s̈klearn.utils.class_weight.compute_class_weight — scikit-learn 0.21.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html. [Accessed: 27-May-2019].

9. N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and Jianming Liang, "Convolutional neural networks for medical image analysis: full training or fine tuning?," IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1299–1312, Mar. 2016.
10. Pelka, Obioma and Friedrich, Christoph M and García Seco de Herrera, Alba and Müller, Henning, "Overview of the ImageCLEFmed 2019 Concept Prediction Task", CLEF2019 Working Notes, CEUR Workshop Proceedings (CEUR- WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-2380/, 2019
11. Bogdan Ionescu and Henning Müller and Renaud Péteri and Yashin Dicente Cid and Vitali Liauchuk and Vassili Kovalev and Dzmitri Klimuk and Aleh Tarasau and Asma Ben Abacha and Sadid A. Hasan and Vivek Datla and Joey Liu and Dina Demner-Fushman and Duc-Tien Dang-Nguyen and Luca Piras and Michael Riegler and Minh-Triet Tran and Mathias Lux and Cathal Gurrin and Obioma Pelka and Christoph M. Friedrich and Alba García Seco de Herrera and Narciso Garcia and Ergina Kavallieratou and Carlos Roberto del Blanco and Carlos Cuevas Rodríguez and Nikos Vasillopoulos and Konstantinos Karampidis and Jon Chamberlain and Adrian Clark and Antonio Campello, "ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature", Experimental IR Meets Multilinguality, Multimodality, and Interaction, in Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019)