Transfer Learning with a Layer Dependent Regularization for Medical Image Segmentation

Nimrod Sagie[1], Hayit Greenspan[1], and Jacob Goldberger[2(✉)]

[1] Tel-Aviv University, Tel-Aviv, Israel
[2] Bar-Ilan University, Ramat-Gan, Israel
Jacob.Goldberger@biu.ac.il

Abstract. Transfer learning is a machine learning technique where a model trained on one task is used to initialize the learning procedure of a second related task which has only a small amount of training data. Transfer learning can also be used as a regularization procedure by penalizing the learned parameters if they deviate too much from their initial values. In this study we show that the learned parameters move apart from the source task as the image processing progresses along the network layers. To cope with this behaviour we propose a transfer regularization method based on monotonically decreasing regularization coefficients. We demonstrate the power of the proposed regularized transfer learning scheme on COVID-19 opacity task. Specifically, we show that it can improve the segmentation of coronavirus lesions in chest CT scans.

Keywords: Transfer learning · Regularization · COVID-19 opacity

1 Introduction

Collecting annotated medical data is usually an expensive procedure that requires the collaboration of radiologists and researchers. One of the main differences between the medical imaging domain and computer vision is the need to cope with a limited amount of annotated samples [2,5,11,21]. Transfer learning is a popular strategy to overcome the difficulties posed by limited annotated training data. The goal of transfer learning is to transfer knowledge from a source task to a target task by using the parameter set of the source task in the process of learning the target task. Transfer learning utilizes models that are pre-trained on large datasets, that can either be scenery datasets such as ImageNet or medical datasets. There is a plethora of work on using transfer learning in different medical imaging applications (e.g. [3,22]). Due to the popularity of transfer learning in medical imaging, there has been also work analyzing its precise effects (see e.g. [13,15,19]).

This research was supported by the Ministry of Science & Technology, Israel.

C. Lian et al. (Eds.): MLMI 2021, LNCS 12966, pp. 161–170, 2021.
https://doi.org/10.1007/978-3-030-87589-3_17

A common procedure when using transfer learning is to start with a pre-trained model on the source task and to fine-tune the model, i.e. train it further, using a small set of data from the target task. Variants of transfer learning include fine-tuning of all network parameters, only the parameters of the last few layers, or simply just use the pre-trained model as a fixed feature extractor which is followed by a trained classifier. Injecting information into a network via parameter initialization is problematic since this information can be lost during the optimization procedure. Li et al. [9] recently proposed that, in addition to initialization, the pre-trained parameters can be also used as a regularization term. They implemented an L_2 penalty term to allow the fine-tuned network to have an explicit inductive bias towards the original pre-trained model.

In this study we show that the learned parameters move apart from the source task as the image processing progresses along the network layers, and that this occurs even if we regularize the learned parameters. To cope with this we propose a regularization method based on monotonically decreasing regularization coefficients that allows a gradually increasing distance of the learned parameters from the pre-trained model along the network layers. We applied this transfer learning regularization strategy to the task of COVID-19 opacity segmentation and show that it improves the segmentation of Coronavirus lesions in chest CT scans.

2 Transfer Learning via Gradual Regularization

Parameter regularization is a common technique for preventing overfitting to the training data. Let θ be the parameter set of a given neural network. The L_2 regularization modifies the loss function $\text{Loss}(\theta)$ which we minimize by adding a regularization term that penalizes large weights:

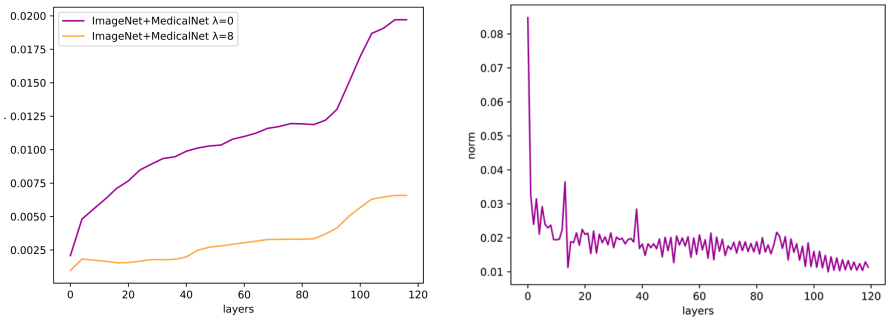
$$\text{Loss}(\theta) + \lambda \|\theta\|^2, \tag{1}$$

where λ is the regularization coefficient. Adding the L_2 term results in much smaller weights across the entire model, and for this reason it is known as weight decay. Network parameters are usually initialized by zero (with a small random perturbation to avoid trivial solutions) and the regularization term prevents the parameters from deviating too much from the initial zero values.

Transfer learning is a network training method where a model trained on a task with a large available annotated data, is reused as the starting point for a model on a second task. Several recent studies have suggested exploiting the full potential of the knowledge already acquired by the model on the source task, by penalizing the difference between the parameters of the source task and the parameters of the target task we aim to learn [9,10]. In transfer learning the target network is initialized by the source network parameters. Hence, a suitable L_2 regularized loss for transfer learning is:

$$\text{Loss}(\theta) + \lambda \|\theta - \bar{\theta}\|^2, \tag{2}$$

where $\bar{\theta}$ is the parameter set of the source task model. The value for λ in the range of $(0, \infty)$ controls the amount of knowledge we want to transfer from the source task to the target task. In practice, λ is a hyper-parameter that can be tuned using cross-validation.



(a) (b)

Fig. 1. (a) Average L_2 distance between the parameters of source and target networks at each layer, with ($\lambda = 8$) and without ($\lambda = 0$) regularization. (b) Average L_2 norm of the parameters of source network at each layer.

We next illustrate the tendency of the parameters of the target model to more deviate from the pre-trained values in deeper network layers. We used an image segmentation task implemented by a U-net architecture. The details of the source and target models are given below. We calculated the average L_2 distance between the original and the tuned parameters at each network layer. The distance between the target and the source values of each parameter is normalized by the norm of the source value. We examined two transfer learning cases, fine-tuning without regularization ($\lambda = 0$) and fine-tuning with a fixed regularization ($\lambda = 8$) (Eq. 2) that was found to be the optimal value for that setup. Figure 1a shows that at $\lambda = 0$, the distance of the tuned parameters from their original values increases along the network layers. For the case of $\lambda = 8$, as expected, the regularization reduces the distance between the pre-trained and the tuned model. However, the trend toward increased deviation along the network layers remains. Figure 1b shows the average parameter norms at each layer of the source network. We can see that, in contrast to transfer learning, in training from scratch there is no increased deviation from the near zero random starting point along the network layers.

Based on the analysis described above, in this study we propose to apply the transfer regularization gradually such that the transfer regularization coefficient λ decreased monotonically along the network layers. A larger value of λ results in a more aggressive knowledge transfer from the source to the target. The first network layers perform low-level processing that do not vary much between tasks applied to similar data types. As the data processing progresses along

the network layers, the network is more focused on the target task which is different from the source task. Changing the parameters of a layer also modifies the input to the next layer, which causes the difference between the source and target tasks to accumulate along the network layers. Hence, it makes sense to gradually decrease the penalty of moving away from the pre-trained model as the data processing progresses along the network layers.

Denote the parameters of a target domain network by $\theta = (\theta_1, \theta_2, ..., \theta_k)$ such that θ_i are the parameter set of the i-th layer of the network and k is the number of layers in the network. In a similar way denote the parameters of the source network layers by $\bar{\theta} = (\bar{\theta}_1, \bar{\theta}_2, ..., \bar{\theta}_k)$.

The proposed regularized cost function for transfer learning is:

$$\text{Loss}(\theta) + \sum_{i=1}^{k} \lambda_i \|\theta_i - \bar{\theta}_i\|^2, \tag{3}$$

such that

$$\infty \geq \lambda_1 \geq \lambda_2 \geq \ \cdots \ \geq \lambda_k \geq 0.$$

Setting the transfer regularization hyper-parameter λ to ∞ results in freezing the regularized parameters. By setting the hyper-parameter λ to zero, we obtain standard transfer learning where the only way knowledge is transferred to the target task is via parameter initialization. In the case of the final layers that are learned from scratch, we can still initialize them with small random numbers and use standard L_2 regularization during training.

In this study we focus on U-net networks for image segmentation tasks. The U-net architecture [16] has become the state-of-the-art for medical image semantic segmentation. It is composed of two main pathways: a contraction path (the encoder) that captures the context by processing low-level information, and the expanding path (the decoder), which enables precise localization. The U-net encoder performs low and mid-level processing of the pixel map leading to a latent image representation. In contrast, the U-net decoder, generates the network's decisions based on the computed representation and is focused on a specific task accomplished by the network. The most common way of utilizing transfer learning with U-net is by initializing the encoder with pre-trained weights and then either freezing it, or allowing re-training, depending on the target's data size and computational power limitations. The decoder, which is task-dependent, is trained from scratch. We propose to exploit the full potential of the knowledge already acquired by the model on the source task, by enabling changes in weights, but under a certain constraint. The proposed cost function is:

$$\text{Loss}(\theta) + \sum_{i=1}^{k} \lambda_i \|\theta_{\text{encoder},i} - \bar{\theta}_{\text{encoder},i}\|^2 + \lambda' \|\theta_{\text{decoder}}\|^2 \tag{4}$$

s.t. $\bar{\theta} = (\bar{\theta}_{\text{encoder}}, \bar{\theta}_{\text{decoder}})$ and $\theta = (\theta_{\text{encoder}}, \theta_{\text{decoder}})$ are the parameters of the source and target networks, respectively and i goes over the encoder layers. In this scheme we refine the encoder regularization by setting a gradually decreasing regularization coefficients along the encoder layers as described in Eq. (3).

There are many ways to define a decreasing coefficient sequence. In this study we used slowly decreasing functions in the form of:

$$\lambda_i = \max(0, \lambda_0 - \alpha \cdot \log(i)) \qquad i = 1, ..., k \tag{5}$$

such that λ_0 and α are hyper-parameters that can be tuned on a validation set using a grid search.

3 Network Implementation Details

We next describe the network architecture and pre-training used. We focused here on the task of COVID-19 opacity segmentation. We used a 2-D U-net [16] with a DenseNet121 [6] backbone. In our implementation, the decoder was composed of decoder blocks and a final segmentation head, which consists of a convolutional layer and softmax activation. Each decoder block consists of a transpose convolution layer, followed by two blocks of convolutional layers, batch normalization, and ReLU activation. For the cost function, we used weighted cross-entropy, where the weights were calculated using the class ratio in the dataset.

We investigated regularization in several different pre-training scenarios. We implemented three source tasks and used them to pre-train the encoder on the target task (the decoder was trained from scratch). The three source tasks were as follows:

- **Natural image pre-training network:** U-net with an encoder that was trained on ImageNet.
- **Medical image pre-training network:** U-net with encoder that was trained from scratch on several publicly available medical imaging segmentation tasks [20]. The network has a shared encoder for global feature extraction followed by several medical task-specific decoders [17]. We term this network "MedicalNet".
- **Combined natural and medical image pre-training network:** The U-net encoder was initialized with ImageNet weights and then trained on the medical datasets as above. We term this network "ImageNet+MedicalNet".

The overall system consisted of the trained model and a series of image processing techniques for both the pre, and the post-processing stages. For pre-processing, all the input slices were clipped and normalized to $[0, 1]$ using a window of $[-1000, 0]$ HU and then resized to a fixed spatial input size of 384×384. The trained network was applied to each slice separately. To construct the 3-D segmentation, we first concatenated the slice-level probabilities generated by the model, and then applied a post-processing pipeline that included morphological operations and removal of opacities outside the lungs.

4 Experiments and Results

We evaluated the system on the task of COVID-19 opacity segmentation using a small COVID-19 dataset [7] containing 29 non-contrast CT scans from three different distributions, from which 3,801 slices were extracted. Lungs and areas of infection were labeled by two radiologists and verified by an experienced radiologist. The given labels were of the lungs and infection. The train-validation-test split was: 21 cases (2446 slices) for training, 3 cases (442 slices) for validation, and 5 cases (913 slices) for testing, chosen at random. We compared two transfer regularization methods:

Table 1. Segmentation results for various source networks and transfer regularization schemes.

Pre-trained network	Regularization coefficient	Dice	Sensitivity	Precision
No pre-training	0	0.650	0.701	0.727
ImageNet	0	0.677	0.701	0.730
	3	0.638	0.706	0.632
	5	0.633	0.699	0.631
	8	0.627	0.674	0.638
MedicalNet	0	0.687	0.724	0.782
	3	0.711	0.835	0.734
	5	0.718	0.759	0.807
	8	0.710	0.832	0.704
	$15 - 1.5 \cdot \log(\text{layer-index})$	0.722	0.775	0.801
	$15 - 2.0 \cdot \log(\text{layer-index})$	0.776	0.772	0.828
	$20 - 1.5 \cdot \log(\text{layer-index})$	0.749	**0.847**	0.762
	$20 - 2.0 \cdot \log(\text{layer-index})$	0.767	0.766	0.824
ImageNet+MedicalNet	0	0.724	0.752	0.803
	3	0.743	0.836	0.778
	5	0.754	0.843	0.783
	8	0.764	0.789	0.825
	$15 - 1.5 \cdot \log(\text{layer-index})$	0.794	0.785	0.860
	$15 - 2.0 \cdot \log(\text{layer-index})$	0.784	0.781	0.851
	$20 - 1.5 \cdot \log(\text{layer-index})$	**0.799**	0.782	0.850
	$20 - 2.0 \cdot \log(\text{layer-index})$	0.792	0.765	**0.868**

– **Fixed regularization** [18]**:** Experiments performed with constant values of λ, starting with $\lambda = 0$; i.e., standard transfer learning via parameter initialization, up to $\lambda = 50$. A high penalty for deviation from the learned weights, which can be considered as basically freezing the encoder.

– **Layer-wize based regularization:** Experiments performed with a gradually decreased λ as a function of the U-net encoder layer's depth.

Given a 3-D chest CT scan, the system produced the correlated 3-D prediction mask for the lungs, as well as the COVID-19 related infections. Once the 3-D segmentation mask for the test set had been extracted, we compared it to the ground truth reference mask for the opacity class.

Table 1 summarizes the segmentation results for the three source tasks. The best segmentation results were attained with ImageNet+MedicalNet, for both the fixed and the monotonically decreasing regularizations. For the fixed regularization, $\lambda = 8$ was obtained as the optimal value, as the Dice score improved by 5.5% from 0.724 ($\lambda = 0$) to 0.764, with a p-value of 0.006. For the monotonically decreasing regularization, $\lambda = 20 - 1.5 \cdot \log(i)$ was found to be the optimal formula on a validation set. In this case the Dice score improved from the case of no regularization ($\lambda = 0$) by 10.3% with p-value < 0.0001.

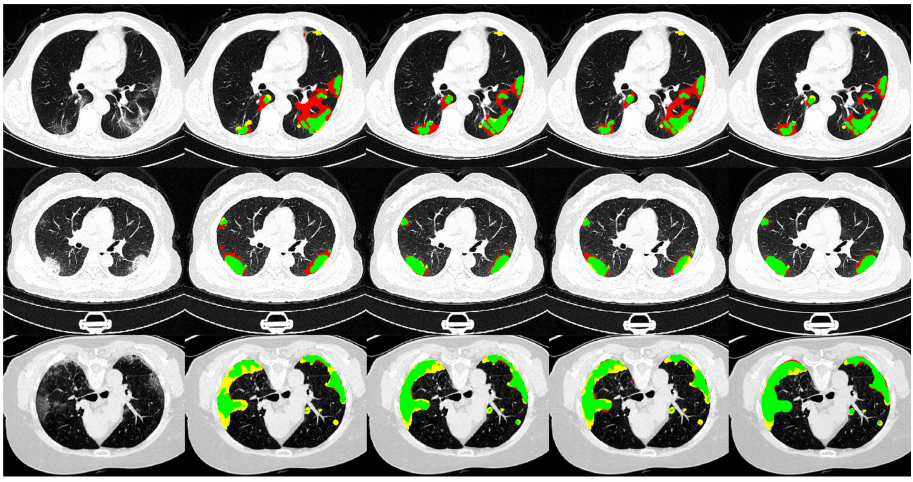


Fig. 2. A qualitative comparison of COVID-19 opacity segmentation with different transfer learning regularizations. Three examples are shown. Green, red, and yellow represent TP, FP, and FN prediction, respectively. (Color figure online)

These results demonstrate that using an inductive bias towards the source parameters for transfer learning, overpowers initialization on its own, since the distributions of the source task and the target task are more similar. Thus, by using the regularization term, either as a function of the layer number or as a constant number, the segmentation results can be improved in cases where the transfer learning is from a source domain close to the target domain. In cases where the transfer learning comes from a source domain with a very different

distribution than the target domain, as in the case from natural images to non-contrast chest CT images, it is better to allow deviation from the learned weights.

Qualitative results are shown in Fig. 2. For each input slice, the CT slice and the segmentation results are given for several values of λ, fixed or monotonically decreasing function, obtained by using the ImageNet+MedicalNet as source task. The given examples show the system prediction for slices from three different test cases with different disease demonstrates generalization capabilities of the proposed method in capturing ground-glass and consolidative opacities. It can be also seen that at $\lambda = 8$ and at $\lambda = 20 - 1.5 \cdot \log(l)$, the red and the yellow regions are demonstrably lessened compared to at $\lambda = 0$ and at $\lambda = 50$, which is indicative of improved results of the optimal regularization term.

Table 2. Classification results of the RSNA 2019 Brain CT Hemorrhage Challenge for various transfer regularization schemes.

Methods	Regularization coefficient	Accuracy
No regularization	0	0.583
Fixed regularization	4	0.699
Monotonically decreasing	$5 - 0.5 \cdot \log(\text{layer-index})$	**0.727**

There are several published results on the same COVID-19 dataset [7]. Wang et al. [23] suggested a Hybrid-encoder transfer learning approach. Laradji et al. [8] used a weakly supervised consistency-based strategy with point-level annotations. Muller et al. [12] implemented a 3-D U-Net and using a patch-based scheme. Paluru et al. [14] recently suggested an anamorphic depth embedding-based lightweight model. The reported Dice scores were 0.704 [23], 0.750 [8], 0.761 [12], 0.798 [14] and 0.698 [1]. Comparison here, however, is problematic due to different data-splits and different source tasks used for transfer learning. We note, however, that our transfer regularization approach is complementary to previous works and can be easily integrated into their training procedure.

To show that layerwize transfer learning regularization is a general concept we demonstrate it on another target task: The RSNA 2019 Brain CT Hemorrhage Challenge [4]. Detecting the hemorrhage, if present, is a critical step in treating the patient and there is a demand for computer-aided tools. The goal is to classify each single slice, to one of the following categories: normal, subarachnoid, intraventricular, subdural, epidural, and intraparenchymal hemorrhage. There is a large variability among images within the same class, making the classification task very challenging. We used the encoder described above, and we initialized it with the parameters of MedicalNet. On top of the encoder, we added two fully-connected layers for the classification task. By concatenating three instances of the same slice with different HU windowing (brain window, subdural window, and bone window) and a [0, 1] normalization, we formed a three channeled input. Since the dataset is highly imbalanced, we excluded most of the normal slices and slices with noisy labels, so eventually we were left with 23,031 images,

that were split randomly into train (n = 13,819), validation (n = 4,606) and test (n = 4,606) sets. The parameters of the regularization term were tuned on the validation set using a grid search. Table 2 shows the classification results on the test set in terms of accuracy. The results demonstrate the added value of adding such regularization term, fixed or monotonically decreasing, to the standard classification loss.

To conclude, this study described a transfer learning regularization scheme based on using the parameters of the source task as a regularization term where the regularization coefficients decrease monotonically as a function of the layer depth. We concentrated on image segmentation problems handled by the U-net architecture where the encoder and the decoder need to be treated differently. We addressed the specific task of segmenting COVID-19 lesions in chest CT images and showed that adding a decreased regularization along the layer axis to the cost function, leads to improved segmentation results. The proposed transfer regularization method is general and can be incorporated in any situation where transfer learning from a source task to a target task is implemented.

References

1. Bressem, K.K., Niehues, S.M., Hamm, B., Makowski, M.R., Vahldiek, J.L., Adams, L.C.: 3D U-net for segmentation of COVID-19 associated pulmonary infiltrates using transfer learning: state-of-the-art results on affordable hardware. CoRR abs/2101.09976 (2021)
2. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. **54**, 280–296 (2019)
3. De Fauw, J., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. **24**(9), 1342–1350 (2018)
4. Flanders, A.E., et al.: Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol. Artif. Intell. **2**(3), e190211 (2020)
5. Greenspan, H., van Ginneken, B., Summers, R.M.: Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans. Med. Imaging **35**(5), 1153–1159 (2016)
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
7. Jun, M., et al.: COVID-19 CT lung and infection segmentation dataset. Zenodo, 20 April 2020
8. Laradji, I., et al.: A weakly supervised consistency-based learning method for COVID-19 segmentation in CT images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2453–2462 (2021)
9. Li, X., Grandvalet, Y., Davoine, F.: Explicit inductive bias for transfer learning with convolutional networks. In: International Conference on Machine Learning, pp. 2825–2834 (2018)
10. Li, X., Grandvalet, Y., Davoine, F.: A baseline regularization scheme for transfer learning with convolutional neural networks. Pattern Recogn. **98**, 107049 (2020)

11. Litjens, G., et al.: A survey on deep learning in medical image analysis. arXiv preprint arXiv:1702.05747 (2017)
12. Müller, D., Rey, I.S., Kramer, F.: Automated chest CT image segmentation of Covid-19 lung infection based on 3D U-net. arXiv preprint arXiv:2007.04774 (2020)
13. Neyshabur, B., Sedghi, H., Zhang, C.: What is being transferred in transfer learning? arXiv preprint arXiv:2008.11687 (2020)
14. Paluru, N., et al.: Anam-Net: anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images. IEEE Trans. Neural Netw. Learn. Syst. **32**(3), 932–946 (2021)
15. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: understanding transfer learning for medical imaging. In: Advances in Neural Information Processing Systems (NIPS) (2019)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Sagie, N., Almog, S., Talby, A., Greenspan, H.: COVID-19 opacity segmentation in chest CT via HydraNet: a joint learning multi-decoder network. In: Medical Imaging 2021: Computer-Aided Diagnosis, vol. 11597. SPIE (2021)
18. Sagie, N., Greenspan, H., Goldberger, J.: Transfer learning via parameter regularization for medical image segmentation. In: The European Signal Processing Conference (EUSIPCO) (2021)
19. Shirokikh, B., Zakazov, I., Chernyavskiy, A., Fedulova, I., Belyaev, M.: First U-Net layers contain more domain specific information than the last ones. In: Albarqouni, S., Bakas, S., Kamnitsas, K., Cardoso, M.J., Landman, B., Li, W., Milletari, F., Rieke, N., Roth, H., Xu, D., Xu, Z. (eds.) DART/DCL -2020. LNCS, vol. 12444, pp. 117–126. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60548-3_12
20. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
21. Tajbakhsh, N., et al.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging **35**(5), 1299–1312 (2016)
22. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
23. Wang, Y., et al.: Does non-COVID19 lung lesion help? Investigating transferability in COVID-19 CT image segmentation. Comput. Methods Programs Biomed. **202**, 106004 (2021)