Biomedical Signal Processing and Control 69 (2021) 102828

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Breast mass classification with transfer learning based on scaling of deep representations

Michal Byra

Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland

ARTICLE INFO

Keywords:
Breast mass classification
Convolutional neural networks
Deep learning
Transfer learning
Ultrasound imaging

ABSTRACT

Ultrasound (US) imaging is widely used to help radiologists in diagnosing breast cancer. In this work, we propose a deep learning based approach to breast mass classification in US. Transfer learning with convolutional neural networks (CNNs) is commonly used to develop object recognition models in medical image analysis. The most widely used fine-tuning techniques aim to modify weights of pre-trained networks to address target medical problems. However, fine-tuning can be difficult when the number of trainable parameters of the pre-trained network is large and the available medical data are scarce. To address this issue, we propose a novel transfer learning technique based on deep representation scaling (DRS) layers, which are inserted between the blocks of a pre-trained CNN to enable better flow of information in the network. During network training, we only update the parameters of the DRS layers in order to adjust the pre-trained CNN to process breast mass US images. We present that the DRS based approach greatly reduces the number of trainable parameters, and achieves better or comparable performance to the standard transfer learning techniques. The proposed DRS layer method combined with the standard fine-tuning techniques achieved excellent breast mass classification performance, with area under the receiver operating characteristic curve of 0.955 and accuracy of 0.915.

1. Introduction

Ultrasound (US) imaging is widely used to help radiologists in assessing and diagnosing breast cancer in women [1]. US imaging is portable, noninvasive and less expensive than other medical imaging modalities, like mammography or magnetic resonance imaging. However, analysis of US images is difficult and associated with high inter-rater reliability. To accurately differentiate malignant and benign breast masses radiologists have to possess a deep knowledge about characteristic US image features related to malignancy (e.g. mass shape and echogenicity). Various machine learning methods have been proposed to aid the radiologists with breast mass differentiation [2,3]. Nowadays, deep learning methods are gaining interest in breast mass classification.

Deep convolutional neural networks (CNNs) can automatically process input images to determine important image features and provide the desired output [4–6]. However, performance of deep learning methods is usually related to the volume of available training data. When the available training data are scarce, it is usually infeasible to train a well performing model from scratch. To address this problem, transfer learning techniques have been investigated and became the methodology of choice for the development of deep learning methods in medical image analysis [7]. The goal of transfer learning is to utilize pre-trained models to address new recognition tasks. This way a model developed on a large dataset can be adjusted to efficiently process data from a different domain and compensate the lack of training data for the target task.

CNNs pre-trained on the ImageNet dataset are among the most popular models used for transfer learning and breast mass differentiation in US [8,9]. The ImageNet dataset includes over 1 000 000 RGB images corresponding to 1 000 objects and has been used to develop various classification CNNs. In the case of the transfer learning, the most basic approach is to utilize an ImageNet model as a fixed feature extractor combined with a standard classifier, such as the support vector machine method [10,11]. In this case, medical images are used as an input to the pre-trained model and deep features are usually extracted from the last convolutional layer of the network. This approach has been successfully used for breast US mass differentiation in several papers [12–15]. Another transfer learning approach is to fine-tune a pre-trained model using target medical data [16,17]. This approach usually provides better results than the feature extraction technique. However, if the target dataset is small and the number of trainable parameters of the

E-mail address: mbyra@ippt.pan.pl.

https://doi.org/10.1016/j.bspc.2021.102828
Received 23 March 2021; Received in revised form 15 May 2021; Accepted 30 May 2021
Available online 15 June 2021
1746-8094/

model is large, fine-tuning of the entire network may lead to convergence problems and over-fitting [18]. To reduce the number of trainable parameters, investigators usually decide which network layers to fine-tune and which layers to leave frozen. Commonly, only the last layers of the pre-trained model are fine-tuned, which is motivated by the observation that the first layers of the network include feature extractors that are general and shareable between different tasks. In comparison, last layers code features related to recognition of objects from the source dataset, and modification of these layers may be therefore more beneficial for the extraction of features for the new task [19]. Nevertheless, modification of the first layers, including edge and color blob detectors, may also play an important role. To fine-tune a deep model pre-trained on RGB images with grayscale US images, the grayscale pixel intensities are commonly duplicated into all color channels. This raises a question whether the detectors in the first layers of the model pre-trained on RGB images process the target data efficiently. This issue may result in the extraction of worse performing features in the first layers, affecting the processing in deeper layers [20]. In the case of the breast mass differentiation, fine-tuning has been applied in several papers [14,21–25]. Authors investigated the usefulness of various pre-trained deep models and fine-tuning strategies.

In this work, we propose a novel transfer learning approach to breast mass classification. Our method can be used to effectively adjust the entire pre-trained network to the target task, and it corresponds to a lower number of trainable parameters than the standard fine-tuning techniques. The proposed approach is based on the deep representation scaling (DRS) layers, which we insert between pre-trained blocks of the model. The aim of these layers is to transform deep representations to enable better flow of information in the pre-trained network. There are several motivations for our approach:

- In the case of the regular fine-tuning, each trainable parameter of the pre-trained model is updated separately with the back-propagation algorithm. In our case, we equip the network with the DRS layers, initialized as identity mappings, which govern the updates of the block parameters. With the DRS layers, we can directly and coherently scale entire deep representations. In this sense, our approach can be perceived as a parameterized fine-tuning, where we update specific groups of elements in the same way, reducing the number of trainable parameters and addressing potential over-fitting issues.
- DRS layers can enhance (or attenuate) the propagation of certain features though the network, and effectively address the saturation problem caused by the presence of the activation functions. With the DRS layers, we can transform deep representations to make them pass the activation functions. For a large visual mismatch between the source and target data, pre-trained convolutional filters may output feature maps that are noisy. With the DRS layers, we can suppress such noisy maps and exclude them from the processing.

This paper is organized in the following way. First, we present how to equip a pre-trained CNN with the DRS layers. Next, we perform several experiments to illustrate the usefulness of our method in the case of the breast mass classification. Most importantly, we show how the inclusion of the DRS layers impacts the processing of deep representations by the activation function layers of the pre-trained model.

2. Methods

2.1. Residual networks

In this work, we utilized the ResNet101 residual network pre-trained on the ImageNet dataset to demonstrate the usefulness of our approach [8,26]. This model is widely used as a backbone network for various transfer learning tasks and has been utilized for breast mass classification in previous papers [22]. Architecture of the ResNet101 CNN is presented in Fig. 1. Standard ResNet is a CNN that includes stacked residual blocks. Each block can be expressed in the following way [26]:

$$\begin{aligned} \mathbf{y}_l &= \mathbf{x}_l + \mathscr{F}(\mathbf{x}_l; W_l), \\ \mathbf{x}_{l+1} &= f(\mathbf{y}_l), \end{aligned} \tag{1}$$

where $\mathbf{x}_l$ and $\mathbf{x}_{l+1}$ stand for the input and the output of the l-th residual block, $\mathscr{F}$ is the residual function with the associated weights W_l, and f is the rectified linear unit (ReLu) activation function. The function $\mathscr{F}$ includes multiple convolutional, batch normalization and activation functions layers. The number of the residual blocks for the ResNet101 is equal to 33.

2.2. Deep representation scaling

Residual block equation shows that the deep representations from the previous layers are propagated through the network, which generally improves the training and enables development of deeper networks [26]. However, such propagation of representations though a pre-trained ResNet model may not lead to good transfer learning performance, especially when the visual mismatch between the source and target datasets is large. For example, this may result in activation function saturation and extraction of worse performing features in deeper layers of the network. To address the problem, we propose the following modification of the residual block:

$$\begin{aligned} \mathbf{y}_l &= \mathbf{x}_l + \mathscr{F}(\mathbf{x}_l; W_l), \\ \mathbf{z}_l &= \mathscr{G}(\mathbf{y}_l; S_l, D_l), \\ \mathbf{x}_{l+1} &= f(\mathbf{z}_l), \end{aligned} \tag{2}$$

where function $\mathscr{G}$, associated with weights S_l and D_l, stands for an intermediate operation used for scaling of deep representations (DRS layer). This operation consists of two linear functions and has the following form:

$$y_l^{'}(i,j,k) = S_l^m(i,j)y_l(i,j,k) + S_l^b(i,j), \tag{3}$$

$$z_l(i,j,k) = D_l^m(k)y_l^{'}(i,j,k) + D_l^b(k), \tag{4}$$

where $y_l(i,j,k)$ are the elements of representation $\mathbf{y}_l \in \mathscr{R}^{H\times W\times C}$, with H, W and C equal to the height, width and the number of channels of $\mathbf{y}_l$. The first linear scaling, denoted by S, transforms the deep representation in respect to the spatial dimensions. The second linear scaling, denoted by D, is used to separately scale each channel of the input tensor. With the DRS layers specified by parameters S_l and D_l, we can transform deep

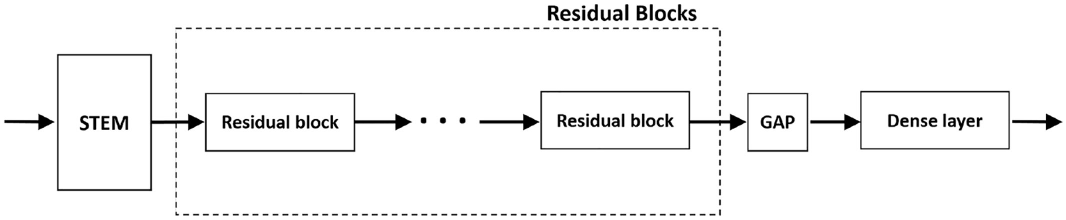

Fig. 1. Simplified architecture of the ResNet101 network including 33 residual blocks See Fig. 2a) for an illustration of a residual block. GAP stands for the global average pooling layer.

representations before they are processed by the ReLU activation function layer. Moreover, to adjust the pre-trained model to a new task, we can fine-tune only the DRS layers, initialized as an identity mapping, while keeping the other elements of the residual block frozen. Our modification of the residual block is presented in Fig. 2.

There are many candidates for the transformation function $\mathscr{G}$. We believe that the proper transformation should posses the following three properties:

- We should be able to initialize $\mathscr{G}$ as an identity mapping. This way, the layers associated with the function $\mathscr{G}$ initially don't change the flow of the information in the pre-trained network during the training. Initialization of the DRS layers with random parameters could easily result in activation function saturation and undermine the entire processing of deep representations in the pre-trained model.
- The number of the trainable parameters of the DRS layer should be lower than for the corresponding residual function. This way, with the $\mathscr{G}$ function we can both modify the input $\mathbf{x}_l$ and the output of the residual function $\mathscr{F}$ based on a smaller number of trainable parameters than it would be required for the function $\mathscr{F}$ alone.
- DRS layers should process the deep representations in a more global way than the regular convolutional operators included in $\mathscr{F}$ to effectively and coherently scale entire input tensors.

In our work, the choice of the transformations in $\mathscr{G}$ was inspired by the literature on the usefulness of attention gates and squeeze and excitation (SE) blocks in CNNs [27,28]. However, we used the ideas behind these methods in a slightly different way. The attention gates and SE blocks were originally utilized to process deep representations individually. The outputs of these operations were conditioned on the input data. In medical image analysis, the attention gates were used to spatially filter deep representations and attenuate regions that are not important for object recognition [28]. The SE blocks were developed to scale deep representations in a channel-wise manner to enhance the information present in particular channels. In our case, the $\mathscr{G}$ function was used to match the pre-trained network with the target dataset. The operations in $\mathscr{G}$ were determined for the entire target dataset. The aim of the scaling in Eq. (3) was to spatially perturb input tensors, potentially reducing (or enhancing) specific spatial patterns present in deep representations. Linear transformation described in Eq. (4) was applied to enhance (or attenuate) certain representations in a channel-wise manner, similarly to the SE blocks. Moreover, the bias terms in Eqs. (3) and (4) could be used to mitigate the problems related to the activation function saturation. For example, a large change of the bias term in Eq. (4) could render the particular tensor channel further propagated or blocked by the ReLu activation function. Moreover, the operations specified by Eqs. (3) and (4) are associated with a small number of trainable parameters, much smaller than for the residual function $\mathscr{F}$.

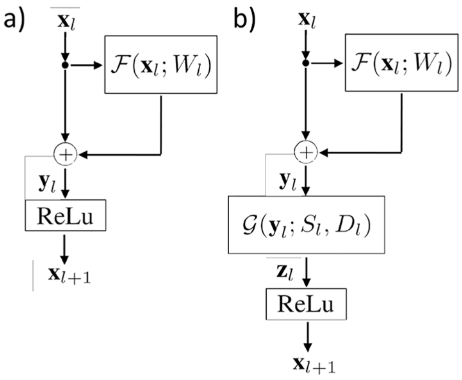

Fig. 2. Residual block a) without and b) with the deep representation scaling layers designed to improve the flow of information in the pre-trained ResNet101 model.

3. Performance evaluation

3.1. Dataset

To assess the proposed deep learning method, we used the BUSI dataset consisting of 647 grayscale breast US images. 437 images (67%) corresponded to benign masses and 210 images (33%) presented malignant breast masses [29]. The dataset was collected during regular scanning performed at the Baheya Hospital, Cairo, Egypt using a LOGIQ E9 and LOGIQ E9 Agile US scanners. Sample US images from the dataset are presented in Fig. 3.

We duplicated grayscale image pixel intensities to all color channels to enable transfer learning with the ResNet101 pre-trained on RGB images [12]. Next, US images were resized to dimensions of 224x224 and normalized in the same way as the ImageNet data originally used for the pre-training [26]. To perform experiments, data were randomly divided into train/validation/test sets with a 452/65/130 split. Ratio of malignant breast masses was the same for each split and equal to approximately 33%.

3.2. Transfer learning techniques

The proposed method based on the DRS layers was compared with several standard transfer learning techniques. For each approach, the last dense layer (classification layer) of the pre-trained ResNet101 was replaced with a dense layer suitable for the binary classification of breast masses, initialized with random weights. In this work, the following transfer learning techniques were implemented:

- **Feature extraction**: we froze all layers of the model and trained only the dense layer. In this case, the pre-trained model was only used to extract features from the global average pooling layer and to train a linear classifier.
- **Fine-tuning, last block**: in comparison to the feature extraction technique, in this case we also fine-tuned the last residual block of the network.
- **Full fine-tuning**: we fine-tuned the entire pre-trained network with the target US data, including all residual blocks and the dense layer.
- **DRS** (proposed): we equipped each of the 33 residual blocks of the ResNet101 with a DRS layer initialized as an identity mapping, see Fig. 2. During the training, only the DRS layers and the dense classification layer were trainable.

The number of trainable parameters (excluding the task specific dense layer) for each transfer learning method is presented in Table 1. In comparison to the full fine-tuning, the transfer learning technique based on the DRS layers has around 426 times less trainable parameters.

To further assess our approach, we also investigated whether the proposed DRS method can be combined with the two standard fine-tuning techniques. In this case, we also fine-tuned either the last or all blocks of the network equipped with the DRS layers. This way, the model could simultaneously scale deep representations as well as adjust convolutional filters during the training.

3.3. Classification metrics

To assess the breast mass classifiers we calculated the receiver operating characteristic curve (ROC) and the area under the ROC curve (AUC). Accuracies, sensitivies and specificietes were determined based on the point on the ROC curve closest to curve upper left corner [30]. Bootstraping was applied to calculated the standard deviations of the classification scores. Additionally, the bootstrapped AUC values obtained for different methods were compared using Wilcoxon rank sum

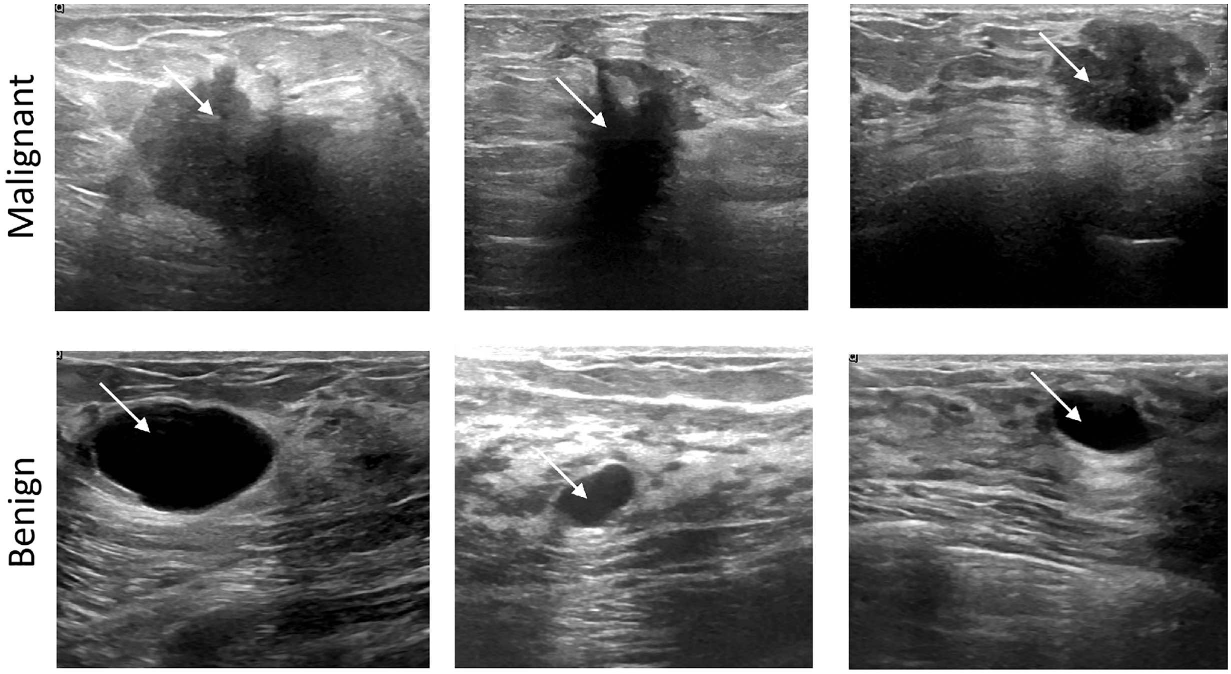

Fig. 3. Sample images from the BUSI dataset we used to evaluate transfer learning methods.

Table 1
Number of trainable parameters (e.g. filter weights) for each implemented transfer learning method (excluding the last task specific dense layer) for the ResNet101 pre-trained on the ImageNet dataset.

Method	Trainable parameters
Feature extraction	0
Fine-tuning, last block	4 471 394
Fine-tuning, all blocks	42 653 790
Deep representation scaling (DRS)	101 471

test at the significance level of 0.05.

3.4. Training

All models were trained using Adam optimizer to minimize the standard binary cross-entropy loss function. To address the problem of class imbalance, we weighted the loss function with class weights inversely proportional to class frequencies in the training set. Grid search was applied to select better performing hyper-parameters. The grid had the learning rate in range of [0.001, 0.01, 0.1] and decay rate of the 1st moment estimates (β_1) in range of [0.85, 0.9, 0.95]. The decay rate for the 2nd moment estimates (β_2) and the batch size were set to 0.999 and 24, respectively. Image augmentation was applied to generate more data for the training. During the training, we monitored the accuracy on the validation set and terminated the training if no improvement was observed after 20 epochs. Learning rate was exponentially decreased every 4 epochs by a factor of 0.9 if no improvement was observed on the validation set. For each set of the hyper-parameters, the training was repeated three times and the better performing model on the validation set was selected for test evaluations. Calculations were performed in Python using TensorFlow on a computer equipped with a RTX 2080 Ti graphics card [31].

3.5. Scaling effects

To better understand the impact of the DRS layers on the information flow in the network, we investigated whether the DRS layers actually change what is propagated through the ReLu activation functions, specified by f in the residual block equation (Eq. (1)). As presented in several studies on adversarial learning, even simple manipulations of input image color distribution (e.g. color inversion) may have a negative impact on network's performance [32,33]. In the case of the transfer learning, large visual mismatch between the source and target data may result in the dying ReLu problem when the negative inputs are not transmitted through the activation functions, perturbing processing in deeper layers. To assess this problem, we compared the outputs of the activation function layers of the models trained with and without the DRS layer.

Let $\mathbf{t}_l^n$ stand for the element-wise signum function of $\mathbf{x}_l^n$ for the n-th input image. We can define the following activation rate function for the l-th residual block of the network:

$$A_l = \frac{1}{NHWC}\sum_n\sum_i\sum_j\sum_k t_{l+1}^n(i,j,k), \tag{5}$$

where N stands for the number of test images, H, W and C are the dimensions of $\mathbf{t}_l^n$, similarly as in Eq. (3) and (4). The activation rate function A_l is equal to 1 if the output of the Relu activation function in the l-th block is strictly positive for all test images, or 0 if it is non-positive (≤ 0). Therefore, the activation rate function measures how many inputs go through the activation function (at average).

To better understand the differences in processing of deep representations between the two ResNet101 models, we can additionally define the following counting function:

$$C_l(i,j,k) = \begin{cases} 1, & \text{if } t_{l+1}^{DRS}(i,j,k) \neq t_{l+1}^{FE}(i,j,k) \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $\mathbf{t}_{l+1}^{DRS}$ and $\mathbf{t}_{l+1}^{FE}$ stand for the element-wise signum functions of $\mathbf{x}_{l+1}^{DRS}$ and $\mathbf{x}_{l+1}^{FE}$, corresponding to the l-th residual block of the models with and without the DRS layers, respectively. The second model (feature extraction technique) processed the target US images in the same way as in the case of the source ImageNet dataset. We can use the counting function to define the activation rate change function in the following way:

$$\Delta A_l = \frac{1}{NHWC}\sum_n\sum_i\sum_j\sum_k C_l^n(i,j,k), \tag{7}$$

where C_l^n stands for the counting function of the l-th block obtained for the n-th test image. The rate has simple interpretation. ΔA_l is equal to 0 if the ReLu activation function both propagates and blocks exactly the same elements of $\mathbf{x}_{l+1}^{DRS}$ and $\mathbf{x}_{l+1}^{FE}$, and 1 for the opposite case. With the activation rate change function we can assess whether the incorporation of the DRS layers results in propagation of different features through the network. Additionally, we calculated the mean and max absolute errors between the weights and bias terms of the DRS layers and parameters corresponding to the identity mapping. This was performed to assess the level of parameter perturbation in the DRS layers and to assess if both scaling functions were utilized to transform deep representations. Calculations were performed for all transfer learning techniques utilizing DRS layers.

4. Results

4.1. Classification

Classification results obtained for each transfer learning technique are presented in Table 2. Here, the technique based on the feature extraction achieved the worst performance, with AUC value of 0.903. The proposed approach based on DRS layers achieved AUC value of 0.935, which was significantly higher than for the feature extraction method (p-value$<$0.05) and comparable to other fine-tuning techniques. Moreover, by combining the DRS method with one of the standard fine-tuning techniques we achieved AUC value of 0.955, which was significantly higher than for all other methods. This result shows that the DRS method can be used jointly with other transfer learning techniques to improve their performance. ROC curves for the better performing approach and the feature extraction baseline method are presented in Fig. 4.

4.2. Visualisations

Activation functions: Fig. 5a) presents the activation rate functions, Eq. (5), calculated for each residual block using ImageNet validation set (dataset used for the pre-training of the ResNet101) and for the feature extraction technique on the BUSI test set. In both cases the activation rate function was equal to around 60% for the majority of residual blocks, but significantly decreased for the last blocks, resulting in much more sparse deep representations. Similar results were obtained for the transfer learning techniques utilizing the DRS layers, Fig. 5b), with slight differences visible in the case of the last residual blocks. Results presented in Fig. 5 shows that while the pre-trained network was not developed to process breast US images, the activation rate functions closely imitated those obtained for the ImageNet dataset.

Table 2
Test set scores calculated for each implemented transfer learning technique. The better classification performance was obtained for the approaches utilizing both the DRS layers and the fine-tuning techniques.

Method	AUC	Accuracy	Sensitivity	Specificity
Feature extraction	0.903±0.022	0.823±0.029	0.904±0.042	0.784±0.045
Fine-tuning, last block	0.934±0.018	0.884±0.020	0.904±0.031	0.875±0.026
Fine-tuning, all blocks	0.916±0.020	0.884±0.028	0.833±0.036	0.909±0.036
DRS	0.935±0.018	0.869±0.030	0.833±0.038	0.886±0.047
DRS + fine-tuning, last block	0.955±0.011	0.915±0.020	0.904±0.904	0.920±0.024
DRS + fine-tuning, all blocks	0.954±0.016	0.923±0.019	0.976±0.033	0.897±0.033

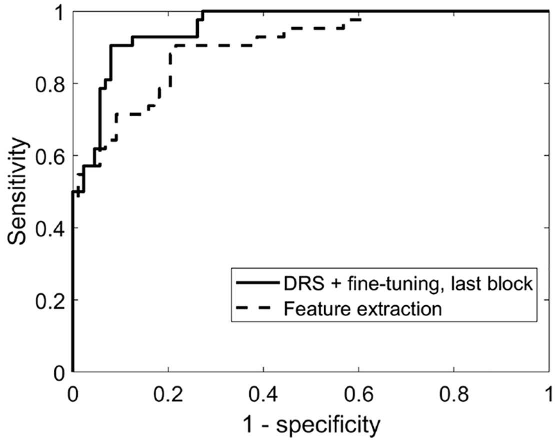

Fig. 4. ROC curves obtained for the feature extraction method (AUC of 0.903) and the fine-tuning of last network block combined with the DRS layers (AUC of 0.955).

Although the activation rate functions in Fig. 5 were similar for all approaches, Fig. 6 shows that the presence of the DRS layers resulted in propagation of different features through the network. For the DRS method combined with the fine-tuning of the last block, the activation rate change functions increased approximately linearly with the depth of the residual blocks to decrease in the last layers, presumably due to the drop of the activation rate function observed in Fig. 5. In the case of the combination of the DRS layers and the full fine-tuning, the activation rate change function quickly increased to around 40% to similarly decrease to around 18% for the last blocks, as in the case of the other two methods. The larger increase of the activation rate change function for the full fine-tuning was probably due to the fact that the training resulted in modification of the convolutional filters corresponding to different blocks of the network.

DRS layer parameters: mean and max absolute errors between the layer parameters after the training and the weights and bias terms corresponding to the identity mapping are presented in Fig. 7. The mean and max absolute errors were approximately equal to around 1% and 5%. Presumably, even small changes of the parameters could impact the information flow presented in Fig. 6. Both DRS layer transformations, the depth-wise D and the spatial S, were utilized for all three investigated transfer learning techniques. In the case of the spatial transformation, the largest max absolute errors were obtained for the middle and last residual blocks. In comparison, max errors for the depth-wise scaling were less variable across the blocks. However, mean errors for the spatial transformation calculated for the weights and bias terms significantly decreased for several residual blocks, suggesting that the spatial transformation was not always fully utilized.

5. Discussion

In this work, we proposed a novel deep learning based approach to breast mass classification in US. In comparison to the previous methods, based either on fine-tuning or feature extraction, we introduced and applied a transfer learning technique based on the scaling of deep representations. The proposed approach is associated with a much smaller number of trainable parameters compared to the full fine-tuning technique and can be used to adjust the entire network to the new classification task. Moreover, our results indicated that the proposed method can be combined with standard fine-tuning strategies to further improve the performance. By combining the DRS layers with the fine-tuning technique we could simultaneously scale deep representations and modify convolutional filters of the pre-trained model.

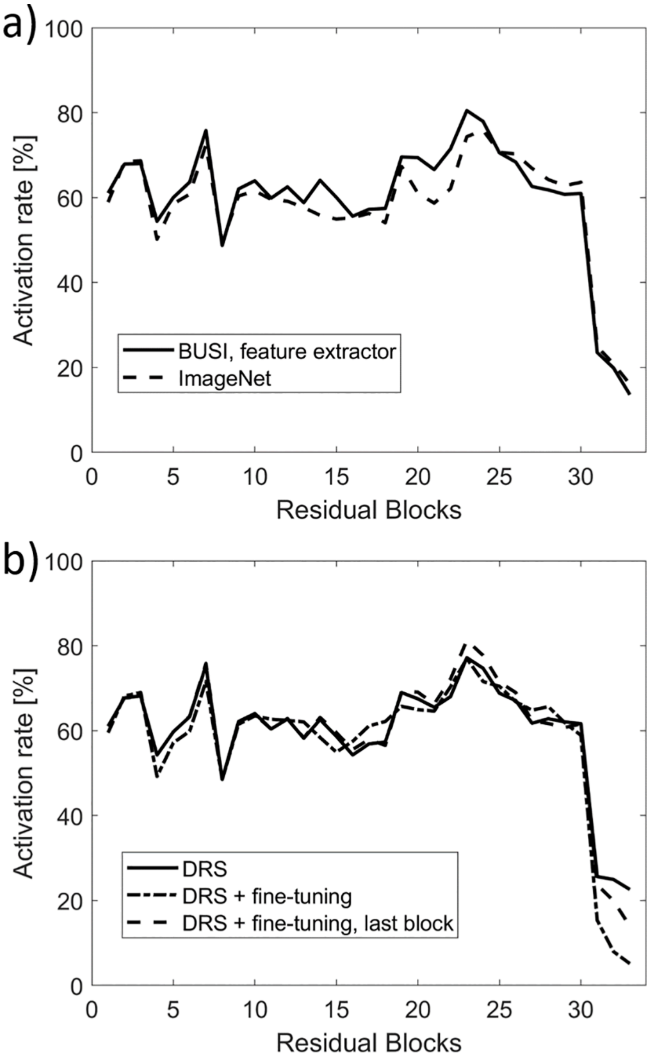

Fig. 5. Activation rate functions (Eq. (5)) calculated for a) the ImageNet validation set and feature extraction method on the BUSI test set, and b) the results for the transfer learning techniques utilizing DRS layers. For all cases the activation functions were equal to around 60% for the majority of the residual blocks, but decreased for the last blocks, showing that the networks promote sparse representations for the classification. Calculations were performed each of the 33 residual blocks of the ResNet101.

It is difficult to directly compare our results with the results reported in the previous studies due to different methodologies and employed datasets. Direct comparisons are especially difficult in the case of the fine-tuning techniques, which depend on training data volume. In our study, the approach based on DRS layers and fine-tuning achieved high AUC value of 0.955. For the feature extraction technique, we obtained AUC value of 0.9. Generally, our results are in an agreement with the previous studies on the usefulness of transfer learning techniques for breast mass classification in US. Antropova et al. used features extracted from pre-trained VGG19 network to train support vector machine classifiers and achieved AUC value of around 0.9 [12]. Similarly, Byra et al. utilized features extracted from the VGG19 network and achieved AUC value of 0.881 [14]. However, in the case of the fine-tuning, other authors utilized larger datasets. For example, Han et al. used fine-tuning to develop a breast mass classification models based on a set of 7408 US images and achieved high AUC value of 0.96 [21]. Qi et al. utilized fine-tuning and a set of 8145 US images and achieved high AUC value of 0.98 [23]. Tanaka et al. applied full fine-tuning with the ResNet152 CNN based on 1536 US images and achieved AUC value of 0.935 [24]. Al-Dhabyani et al. utilized a pre-trained ResNet CNN and the BUSI dataset to develop a model for breast mass classification [22]. However, direct comparison between our work and the results of Al-Dhabyani et al. is difficult, because the authors did not provide detailed descriptions of the applied transfer learning methods. The authors achieved accuracy of 82% (traditional augmentation technique), which was at the level of our feature extraction method.

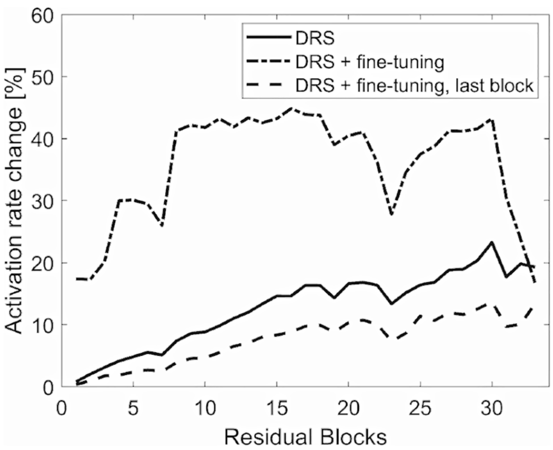

Fig. 6. Activation rate change functions (Eq. (7)) calculated on the test set for the transfer learning techniques utilizing DRS layers. Plots illustrate the differences in information propagation through the ReLu activation functions. Incorporation of the DRS layers promoted propagation of different features through the pre-trained ResNet101.

The experiments presented that DRS layers had impact on the processing of information within the network. As presented in Figs. 5 and 6, the network equipped with the DRS layers extracted different features to perform breast mass classification. The activation rate change function was equal to around 15% for the last residual blocks. This shows that the DRS layers caused saturation of deep representations and simultaneously enhanced propagation of representations that otherwise would be blocked by the activation functions. Moreover, Fig. 7 depicted that the DRS layers were used to transform deep representations in the case of all three investigated transfer learning techniques utilizing DRS layers.

There are several issues with our study. First, there were many potential candidates for the scaling functions. While in our work we utilized linear spatial and depth-wise operations, we could also used nonlinear scaling functions. Second, we did not examine other CNN architectures, such as the VGG19 or the InceptionV2. However, our transfer learning technique is general and can be applied with any pre-trained network.

6. Conclusions

We presented a novel transfer learning technique, aiming at transforming deep representations rather than direct modification of network's pre-trained layers. We successfully applied the technique to differentiate malignant and benign breast masses. The experiments showed that our approach achieved similar or better results than the commonly used techniques based on fine-tuning. Moreover, our approach, when combined with the standard techniques, achieved excellent performance. Additionally, we performed experiments to better understand how the presence of the DRS layers impacts the processing of information in the activation function layers. The results showed that the DRS layers enabled extraction of different features for the classification than in the case of the unmodified pre-trained model. In the future, we plan to investigate other potential forms for the DRS layers, and also examine the usefulness of the method using different network architectures.

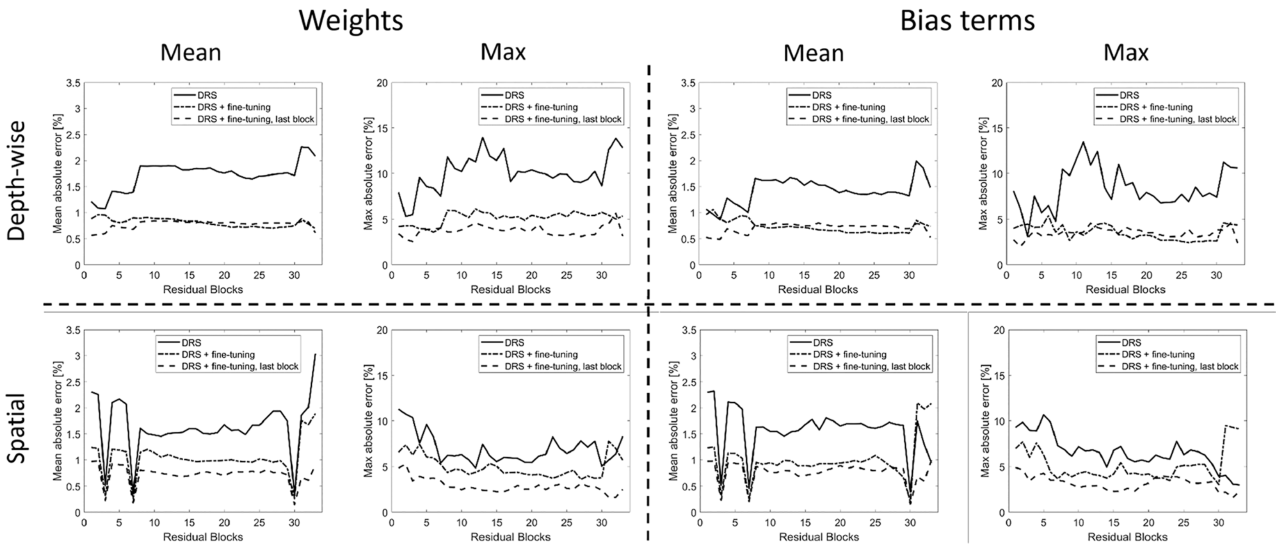

Fig. 7. Mean and max absolute errors calculated between the DRS layer parameters after the training and the parameters corresponding to the identity mapping. Results show that both the spatial and depth-wise transformations were utilized to adjust the pre-trained ResNet101 to differentiate breast masses for all three investigated transfer learning techniques.

CRediT authorship contribution statement

Michal Byra: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians 68 (6) (2018) 394–424.

[2] W.G. Flores, W.C. de Albuquerque Pereira, A.F.C. Infantosi, Improving classification performance of breast lesions on ultrasonography, Pattern Recognition 48 (4) (2015) 1125–1136.

[3] G.-G. Wu, L.-Q. Zhou, J.-W. Xu, J.-Y. Wang, Q. Wei, Y.-B. Deng, X.-W. Cui, C. F. Dietrich, Artificial intelligence in breast ultrasound, World Journal of Radiology 11 (2) (2019) 19.

[4] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: A survey, arXiv preprint arXiv:2001.05566.

[5] Z. Wang, J. Chen, S.C. Hoi, Deep learning for image super-resolution: A survey, IEEE transactions on pattern analysis and machine intelligence.

[6] Z.-Q. Zhao, P. Zheng, S.-T. Xu, X. Wu, Object detection with deep learning: A review, IEEE Transactions on Neural Networks and Learning Systems 30 (11) (2019) 3212–3232.

[7] M.A. Morid, A. Borjali, G. Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, Computers in Biology and Medicine 104115 (2020).

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 248–255, 2009.

[9] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE 109 (1) (2020) 43–76.

[10] A. Sharif Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 806–813.

[11] S. Dara, P. Tumma, Feature extraction by using deep learning: A survey, in: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1795–1801.

[12] N. Antropova, B.Q. Huynh, M.L. Giger, A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets, Medical Physics 44 (10) (2017) 5162–5171.

[13] M. Byra, Discriminant analysis of neural style representations for breast lesion classification in ultrasound, Biocybernetics and Biomedical Engineering 38 (3) (2018) 684–690.

[14] M. Byra, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, M. Andre, Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion, Medical Physics 46 (2) (2019) 746–755.

[15] J. Virmani, R. Agarwal, et al., Deep feature extraction and classification of breast ultrasound images, Multimedia Tools and Applications 79 (37) (2020) 27257–27292.

[16] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[17] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, PMLR, 97–105, 2015.

[18] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2, 2014, pp. 3320–3328.

[19] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging 35 (5) (2016) 1299–1312.

[20] Y. Xie, D. Richmond, Pre-training on grayscale imagenet improves medical image classification, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, 0–0.

[21] S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W.-C. Bang, Y.-K. Seong, A deep learning framework for supporting the classification of breast lesions in ultrasound images, Physics in Medicine & Biology 62 (19) (2017) 7714.

[22] W. Al-Dhabyani, Deep learning approaches for data augmentation and classification of breast masses using ultrasound images, International Journal of Advanced Computer Science and Applications 10 (5).

[23] X. Qi, L. Zhang, Y. Chen, Y. Pi, Y. Chen, Q. Lv, Z. Yi, Automated diagnosis of breast ultrasonography images using deep neural networks, Medical Image Analysis 52 (2019) 185–198.

[24] H. Tanaka, S.-W. Chiu, T. Watanabe, S. Kaoku, T. Yamaguchi, Computer-aided diagnosis system for breast ultrasound images using deep learning, Physics in Medicine & Biology 64 (23) (2019), 235013.

[25] E. Zhang, S. Seiler, M. Chen, W. Lu, X. Gu, BIRADS features-oriented semi-supervised deep learning for breast ultrasound computer-aided diagnosis, Physics in Medicine & Biology 65 (12) (2020), 125005.

[26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[27] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[28] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, Attention gated networks: Learning to leverage salient regions in medical images, Medical Image Analysis 53 (2019) 197–207.

[29] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data in Brief 28 (2020), 104863.

[30] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.

[31] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, in: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, 265–283.

[32] H. Hosseini, B. Xiao, M. Jaiswal, R. Poovendran, On the limitation of convolutional neural networks in recognizing negative images, in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2017, 352–358.

[33] B. Kim, H. Kim, K. Kim, S. Kim, J. Kim, Learning not to learn: Training deep neural networks with biased data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9012–9020.