EPiC Series in Computing

Volume 83, 2022, Pages 108–115

Proceedings of 14th International Conference
on Bioinformatics and Computational Biology

Transfer Learning Pre-training Dataset Effect Analysis for Breast Cancer Imaging

Chanaka Bulathsinghalage[1] and Lu Liu[2]

[1] North Dakota State University, Fargo, U.S.A.
chanaka.cooray@ndsu.edu

[2] North Dakota State University, Fargo, U.S.A.
lu.liu.2@ndsu.edu

Abstract

Comparing with natural imaging datasets used in transfer learning, the effects of medical pre-training datasets are underexplored. In this study, we carry out transfer learning pre-training dataset effect analysis in breast cancer imaging by evaluating three popular deep neural networks and one patch-based convolutional neural network on three target datasets under different fine-tuning configurations. Through a series of comparisons, we conclude that the pre-training dataset, DDSM, is effective on two other mammogram datasets. However, it is ineffective on an ultrasound dataset. What is more, fine-tuning may mask the inefficacy of a pre-training dataset. In addition, the efficacy/inefficacy of DDSM on the target datasets is corroborated by a representational analysis. At last, we show that hybrid transfer learning cannot mitigate the masking effect of fine-tuning.

1 Introduction

Algorithms based on Deep Learning have achieved unprecedented successes in many fields, such as natural image classification [11], natural language processing [2] and protein structure prediction [8] etc., where a large amount of labeled data are available. While for other fields, such as medical imaging analysis, such kind of labeled images are scarce. To overcome this challenge, transfer learning is resorted to pre-train the Deep Learning based algorithms with an existing source dataset (e.g. ImageNet), then fine-tune them on the target dataset.

Though the approach becomes popular in medical imaging analysis, a study [15] on understanding transfer learning for medical imaging reaches a starkly different conclusion that transfer learning pre-trained with ImageNet offers little benefit to performance. Other studies [4, 10] demonstrate that transfer learning does not necessarily result in performance improvements and pre-trained features may be less general than previously thought. However, these studies are based on the natural image datasets. If we utilize a medical imaging dataset as the pre-training dataset, can transfer learning improves the performance on the target dataset? In addition, fine-tuning is an integral part of transfer learning when the source dataset and target dataset are from different domains. If they are from similar domains, what is the effect of fine-tuning on performance? To answer these two questions, we carry out a study on understanding

H. Al-Mubaid, T. Aldwairi and O. Eulenstein (eds.), BICOB2022 (EPiC Series in Computing, vol. 83), pp. 108–115

transfer learning effect on breast cancer imaging analysis because breast cancer is the most commonly diagnosed cancer globally [22]. Early diagnosis can save million of lives.

The contributions of this paper may be summarized as follow.

First, to the best of our knowledge, the paper is the first one to study the transfer learning effect of medical pre-training datasets. Other studies focus on the natural image datasets.

Second, the paper initializes decoupling the effects of medical pre-training datasets and fine-tuning, demonstrates their effects on different target datasets and illustrates that fine-tuning may mask the inefficacy of a pre-training dataset.

2 Datasets

We focus on mammogram and ultrasound datasets in this study and it is our intention to not include MRI and histopathology imaging datasets because these two technologies are not used to screen breast cancer. Especially for histopathology imaging, which is used as the gold standard for breast cancer diagnosis. Current level of machine learning may not be ready for that yet.

Digital Database for Screening Mammography (DDSM) [12] contains 10,239 breast cancer mammography exam images from 2,620 cases, and it is the dataset used in this study to pre-train the models. The images are primarily divided into three classes: normal, benign and malignant and later the normal and benign classes are combined into one class when training the models. These images are in various dimensions, roughly around 2,500x4,500, stored in the lossless JPEG format.

Inbreast [13] is the second dataset used, and it has a total of 410 images classified as benign or malignant. The images are in either 3,328 x 4,084 or 2,560 x 3,328 resolution and saved in the DICOM format.

The third dataset used in this study is the MIAS dataset [21], breast cancer mammogram digital images by the Mammographic Image Analysis Society. The dataset is divided into three main classes: normal, benign and malignant and has a total of 322 images in 1,024x1,024 resolution. Similar to DDSM, normal and benign images are combined into one class when further processing.

One popular ultrasound breast cancer image dataset [17] is used as the fourth dataset in this study, and it contains 250 total images divided into 100 benign and 150 malignant images. The images are in relatively low resolution compared to the mammogram images and the image sizes are approximately 150x150 pixels.

3 Models and Results

In this study, we select three popular neural network architectures and one patch-based convolutional neural network (CNN) to evaluate their performance on three target breast cancer imaging datasets (Inbreast, MIAS and Ultrasound) when (1) training from random initialization, (2) performing transfer learning with ImageNet as their pre-training dataset and (3) performing transfer learning with DDSM as their pre-training datasets. To examine whether the pre-training dataset, DDSM, offers noticeable benefits to target datasets through parameter sharing, experiments without fine-tuning on target datasets are carried out while experiments with fine-tuning are used for comparison. The reason of excluding fine-tuning is to solely investigate the effect of the pre-training dataset. To further study the effect of fine-tuning on transfer learning, two fine-tuning configurations, FT_0.1 and FT_0.2, are assessed for each target dataset. FT_0.1 represents stratified sampling 10% of a target dataset for fine-tuning and FT_0.2 represents stratified sampling 20% of a target dataset for fine-tuning. To evaluate the classification performance of these models, we used AUC score(Area Underneath the ROC Curve) as the performance metric. It is also used in the previous study on natural imaging datasets [15].

3.1 Description of Models

VGG-16 [20], Resnet-50 [5] and Inception-v3 [23] are the selected neural network architectures used in this study because they are prevalent in existing medical imaging transfer learning studies [15, 7, 6, 1, 16, 3, 9, 24]. A patch-based CNN [18], which won the Digital Mammography DREAM Challenge in 2017, is selected as a representative of patch-based models in this study. The patch-based CNN is an end-to-end model, which includes a patch classifier based on VGG/Resnet architectures as a feature extractor. To construct this model, first the patch classifier is trained with regions of interest containing lesion annotation information. Then customized fully connection layers are appended to the trained patch classifier, which results in an image-based whole model. To accommodate inputs of different resolutions (patch VS image), all fully connection layers are implemented as convolutional operations. At last, the whole model is trained with medical images. As the patch classifier and the whole model can be trained independently on different datasets, it circumvents the requirement of lesion annotation information for all training medical images. This is a big advantage because not all medical datasets contain lesion annotation information and obtaining medical annotations is laborious and expensive.

3.2 Performance Evaluation of Transfer Learning

First, we perform experiments on the Inbreast dataset for each selected model. Each model is randomly initialized, pre-trained with ImageNet or pre-trained with DDSM and then evaluated under three fine-tuning configurations: no fine-tuning, fine-tuning with stratified 10% target data and fine-tuning with stratified 20% target data. It is impossible to directly run pre-trained models with ImageNet on medical datasets without fine-tuning. Because pre-trained models with ImageNet give predictions on 1,000 labels instead of 2 labels here. So Not Apply (NA) is used for No FT performance for these models. In addition, Path-based CNN model is not designed to train using ImageNet dataset. Table 1 shows that when these models are randomly initialized without fine-tuning, VGG-16 and Resnet-50 perform poorly and Inception-v3 and the patch-based CNN perform slightly better than random guessing. However, all models perform significantly much better (> 0.2 gains on AUC) if they are pre-trained with DDSM. When these models are randomly initialized and then fine-tuned with 10% or 20% of the Inbreast dataset, VGG-16, Resnet-50 and Inception-v3 have significant performance gains (> 0.22) and the patch-based CNN also has a performance gain of about 0.1 on AUC. ImageNet pretrained models fine-tuned show similar performance as random initialized model fine-tuned. However, if they are pre-trained with DDSM then fine-tuned with 10% or 20% of the Inbreast dataset, their performance gains are comparatively small (< 0.07 except VGG-16 fine-tuned with 20% of the Inbreast dataset). Therefore, we can **conclude that** (1) DDSM can significantly improve the classification performance on the Inbreast dataset through transfer learning, which is also demonstrated in previous research [19], (2) fine-tuning can mask the inefficacy of random initialization and (3) when the pre-training data is effective, fine-tuning plays a minor role on performance.

Table 1: Transfer learning classification performance (AUC) on the Inbreast dataset

Algorithm	Pre-training	No FT	FT_0.1	FT_0.2
VGG-16	Random Ini	**0.340**	0.763	0.784
	ImageNet	NA	0.760	0.770
	DDSM	0.778	0.835	0.872
Resnet-50	Random Ini	**0.244**	0.776	0.802
	ImageNet	NA	0.790	0.809
	DDSM	0.833	0.870	0.882
Inception-v3	Random Ini	**0.583**	0.812	0.816
	ImageNet	NA	0.806	0.824
	DDSM	0.794	0.825	0.839
Patch-based CNN [18]	Random Ini	**0.512**	0.605	0.636
	DDSM	0.804	0.835	0.871

Second, we perform the same experiments on the MIAS dataset. In Table 2, similar patterns can be observed and we can reach the **same conclusions**. DDSM is an effective transfer learning pre-training dataset on the MIAS dataset. Fine-tuning masks the poor performance of random initialization, while plays a trivial role on classification performance when the pre-training dataset is effective.

Table 2: Transfer learning classification performance (AUC) on the MIAS dataset

Algorithm	Pre-training	No FT	FT_0.1	FT_0.2
VGG-16	Random Ini	**0.513**	0.850	0.856
	ImageNet	NA	0.850	0.846
	DDSM	0.863	0.862	0.867
Resnet-50	Random Ini	**0.401**	0.852	0.851
	ImageNet	NA	0.846	0.850
	DDSM	0.872	0.881	0.880
Inception-v3	Random Ini	**0.181**	0.861	0.860
	ImageNet	NA	0.845	0.853
	DDSM	0.858	0.861	0.866
Patch-based CNN [18]	Random Ini	**0.473**	0.558	0.568
	DDSM	0.771	0.776	0.760

At last, we perform the same experiments on the Ultrasound dataset. In Table 3, without fine-tuning, no matter these models are randomly initialized or pre-trained with DDSM, their performance is not impressive. However, when they are fine-tuned, their performance is improved tremendously except when Inception-v3 is randomly initialized or pre-trained with ImageNet and fine-tuned with 10% of data. Therefore, we can **conclude that** (1) DDSM is not effective on improving classification performance on the Ultrasound data partly due to different data modalities or resolutions and (2) fine-tuning can mask the inefficacy of pre-training datasets.

Table 3: Transfer learning classification performance (AUC) on the ultrasound dataset

Algorithm	Pre-training	No FT	FT_0.1	FT_0.2
VGG-16	Random Ini	**0.671**	0.932	0.977
	ImageNet	NA	0.951	0.957
	DDSM	**0.405**	0.963	0.984
Resnet-50	Random Ini	**0.516**	0.844	0.954
	ImageNet	NA	0.803	0.933
	DDSM	**0.388**	0.904	0.965
Inception-v3	Random Ini	**0.632**	0.672	0.842
	ImageNet	NA	0.687	0.817
	DDSM	**0.495**	0.870	0.978
Patch-based CNN [18]	Random Ini	**0.526**	0.951	0.982
	DDSM	**0.769**	0.954	0.983

3.3 Representational Analysis of DDSM

To further investigate the effect of the pre-training dataset, DDSM, on the target datasets, we compare the hidden representations learned by Resnet-50 on different target datasets using (SV)CCA [14], which is short for (Singular Vector) Canonical Correlation Analysis. The tool collects the ordered collection of outputs of neurons on a sequence of inputs and obtains neuron activation vectors. Given the activation vectors for two sets of neurons (for example, the same neurons under two conditions), CCA seeks linear combinations of each that are as correlated as possible.

First, we divide each target dataset into two stratified groups, one with 20% of data and the other containing 80% of data. We use SV(CCA) to compare the hidden representations of the following layers, conv1, block1, block2, block3 and block4 under two conditions, before and after training on target datasets. In the first condition, Resnet-50 is pre-trained with DDSM and then fed with the 80% of a target dataset. In the second condition, Resnet-50 is first pre-trained with DDSM, then fine-tuned with the 20% of a target dataset, and at last fed with the left 80% of the target dataset.

Figure 1 shows that the CCA similarities on the Inbreast and MIAS datasets are on decreasing trends, while the CCA similarities on the ultrasound dataset are on an increasing trend. For the Inbreast and MIAS datasets, CCA similarities are high, which means the hidden representations learned from DDSM resemble the hidden representations learned from the Inbreast/MIAS dataset. What is more, the low layers (close to the model input) have higher CCA similarities than the top layers (close to the model output), which corresponds to the assumption of transfer learning. Low level features are similar between tasks therefore can be transferred and high level features are more task-specific. For the Ultrasound dataset, CCA similarities are very low, which means the hidden representations learned from DDSM are quite different from the hidden representations learned from the Ultrasound dataset. Therefore, DDSM is an effective pre-training dataset for the Inbreast and MIAS datasets. However, it is not an effective one for the Ultrasound dataset.

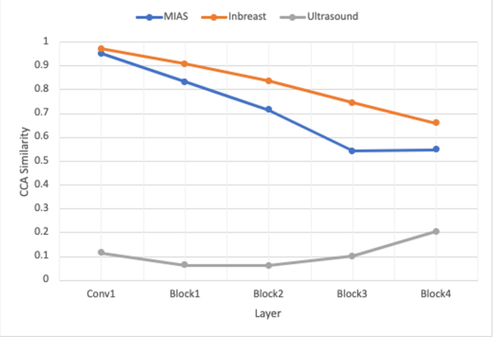

Figure 1: Resnet-50 per-layer CCA similarities before and after training on target datasets.

3.4 Hybrid Transfer Learning Performance Evaluation

Previous research [15] recommends a hybrid transfer learning approach, reusing pre-trained weights up to e.g. block2, redesigning the top of the network (which has the bulk of the parameters) to be more lightweight, initializing these layers randomly. We perform the hybrid approach by reusing pre-trained weights up to block2, evaluate its performance under two fine-tuning configurations, FT_0.1 and FT_0.2, and compare with the results of transfer learning that reuses all pre-trained weights. Table 4 shows that hybrid transfer learning and regular transfer learning obtain similar performance. Therefore, the conclusions drew in paragraphs of 3.2 are also applied to hybrid transfer learning. DDSM is an effective pre-training dataset for the Inbreast and MIAS datasets and it is not adequate for the Ultrasound dataset. Fine-tuning can mask the inefficacy of random initialization when hybrid transfer learning is used.

Table 4: Comparing transfer learning reusing all weights and hybrid transfer learning (All VS Hybrid)

Dataset	Pre-training	FT_0.1	FT_0.2
MIAS	Random Ini	0.852 VS 0.849	0.851 VS 0.845
	ImageNet	0.846 VS 0.853	0.850 VS 0.847
	DDSM	0.881 VS 0.883	0.880 VS 0.892
Inbreast	Random Ini	0.776 VS 0.785	0.802 VS 0.803
	ImageNet	0.790 VS 0.772	0.809 VS 0.817
	DDSM	0.870 VS 0.871	0.882 VS 0.877
Ultrasound	Random Ini	0.844 VS 0.930	0.954 VS 0.982
	ImageNet	0.803 VS 0.911	0.933 VS 0.986
	DDSM	0.904 VS 0.936	0.965 VS 0.970

4 Conclusion

In this study, we investigate the effects of a transfer leaning pre-training dataset, DDSM, on two mammogram image datasets and one ultrasound image dataset with three popular neural network architectures and one patch-based CNN. For each model, experiments of random initialization, pre-trained with ImageNet and pre-trained with DDSM are carried out under three configurations of fine-tuning to decouple the compounding effect of the pre-training dataset and fine-tuning in transfer learning. By comparing these experiment results, we draw the conclusions that (1) DDSM is an effective pre-training dataset for the Inbreast and MIAS datasets, (2) DDSM is not a good pre-training dataset for the ultrasound dataset, (3) fine-tuning can mask the inefficacy of a pre-training dataset and give false impression of high classification performance. The efficacy/inefficacy of DDSM is also demonstrated on these target datasets from the perspective of representational analysis, in which increasing and decreasing trends of per-layer CCA similarites are indicators. At last, we display that hybrid transfer learning cannot mitigate the masking effect of fine-tuning.

The study inspects the effects of a transfer learning pre-training dataset in an empirical approach and a principled method to systematically study the problem will be appreciated and give insights on the generalization of transfer learning performance. In addition, excluding fine-tuning from pre-training datasets to study their effects separately is not possible for natural image datasets in medical imaging applications.

5 Acknowledgments

The work is supported by the National Science Foundation under NSF EPSCoR Track-1 Cooperative Agreement OIA #1946202 and used advanced cyberinfrastructure resources provided by the University of North Dakota Computational Research Center and the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University.

References

[1] Quan Chen, Xiang Xu, Shiliang Hu, Xiao Li, Qing Zou, and Yunpeng Li. A transfer learning approach for classification of clinical significant prostate cancers from mpmri scans. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 101344F. International Society for Optics and Photonics, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Carlos A Ferreira, Tânia Melo, Patrick Sousa, Maria Inês Meyer, Elham Shakibapour, Pedro Costa, and Aurélio Campilho. Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2. In *International Conference Image Analysis and Recognition*, pages 763–770. Springer, 2018.

[4] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Ahmed Hijab, Muhammad A Rushdi, Mohammed M Gomaa, and Ayman Eldeib. Breast cancer classification in ultrasound images using transfer learning. In *2019 Fifth International Conference on Advances in Biomedical Engineering (ICABME)*, pages 1–4. IEEE, 2019.

[7] Marcia Hon and Naimul Mefraz Khan. Towards alzheimer's disease classification through transfer learning. In *2017 IEEE International conference on bioinformatics and biomedicine (BIBM)*, pages 1166–1169. IEEE, 2017.

[8] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[9] SanaUllah Khan, Naveed Islam, Zahoor Jan, Ikram Ud Din, and Joel JP C Rodrigues. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, 125:1–6, 2019.

[10] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[12] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017.

[13] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

[14] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *arXiv preprint arXiv:1706.05806*, 2017.

[15] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

[16] A Sai Bharadwaj Reddy and D Sujitha Juliet. Transfer learning with resnet-50 for malaria cell-image classification. In *2019 International Conference on Communication and Signal Processing (ICCSP)*, pages 0945–0949. IEEE, 2019.

[17] Paulo Sergio Rodrigues. Breast ultrasound image. *Mendeley Data*, 1, 2017.

[18] Thomas Schaffter, Diana SM Buist, Christoph I Lee, Yaroslav Nikulin, Dezső Ribli, Yuanfang Guan, William Lotter, Zequn Jie, Hao Du, Sijia Wang, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA network open*, 3(3):e200265–e200265, 2020.

[19] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12, 2019.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] P SUCKLING J. The mammographic image analysis society digital mammogram database. *Digital Mammo*, pages 375–386, 1994.

[22] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[24] Sulaiman Vesal, Nishant Ravikumar, AmirAbbas Davari, Stephan Ellmann, and Andreas Maier. Classification of breast cancer histology images using transfer learning. In *International conference image analysis and recognition*, pages 812–819. Springer, 2018.