Revisiting Hidden Representations in Transfer Learning for Medical Imaging

Anonymous authors
Paper under double-blind review

Abstract

While a key component to the success of deep learning is the availability of massive amounts of training data, medical image datasets are often limited in diversity and size. Transfer learning has the potential to bridge the gap between related yet different domains. For medical applications, however, it remains unclear whether it is more beneficial to pre-train on natural or medical images. We aim to shed light on this problem by comparing initialization on `ImageNet` and `RadImageNet` on seven medical classification tasks. We investigate their learned representations with Canonical Correlation Analysis (CCA) and compare the predictions of the different models. Our results show that, contrary to intuition, `ImageNet` and `RadImageNet` converge to distinct intermediate representations, and that these representations are even more dissimilar after fine-tuning. Despite these distinct representations, the predictions of the models remain similar. Our findings challenge the notion that transfer learning is effective due to the reuse of general features in the early layers of a convolutional neural network and show that weight similarity before and after fine-tuning is negatively related to performance gains.

1 Introduction

Transfer learning has become an increasingly popular approach in medical imaging, as it offers a solution to the challenge of training models with limited dataset sizes. The ability to leverage knowledge from pre-trained models has proven to be beneficial in various medical imaging applications (Raghu et al., 2019; Cheplygina et al., 2019; Shin et al., 2016). Despite its widespread use, the precise effects of transfer learning on medical image classification are still heavily understudied.

While pre-training on `ImageNet` has become a common practice in medical image classification, there have been growing concerns within the medical imaging community regarding its suitability for medical imaging tasks. Medical images differ from natural images in several ways, including local texture variations as an indication of pathology rather than a clear global subject present in natural images (Raghu et al., 2019). Additionally, medical datasets are smaller in size, have fewer classes, have higher resolution compared to `ImageNet`, and go beyond 2D. These differences between natural and medical image datasets have led to the argument that `ImageNet` may not be the optimal solution for pre-training in medical imaging due to the well-known performance degradation effect caused by domain shift. This has led to increased efforts to explore alternative solutions for pre-training, such as using existing medical datasets (El-Nouby et al., 2021), their alterations (Kataoka et al., 2020; Asano et al., 2020), and creating new medical image datasets specifically designed for pre-training, such as `RadImageNet` (Mei et al., 2022).

Recent studies have challenged the conventional wisdom that the source dataset used for pre-training must be closely related to the target task in order to achieve good performance. Evidence has emerged suggesting that the source dataset may not have a significant impact on the performance of the target task and we can pre-train on any real large-scale diverse data (Gavrikov & Keuper, 2022a;b). Further, more evidence suggests that `ImageNet` leads to the best transfer performance in terms of accuracy, as `ImageNet` not only boosts the performance (Ke et al., 2021) but also is a better source than medical image datasets (Brandt

et al., 2021). This is likely due to the focus on texture in `ImageNet` models (Geirhos et al., 2022), which has been hypothesized to be an important cue for medical image classification.

While `RadImageNet` has demonstrated `ImageNet`-level accuracy (Mei et al., 2022) on radiology image classification, it remains uncertain whether it leads to improved representations when applied to medical target datasets. Furthermore, it is essential to understand the broader implications of source datasets beyond their effect on target task performance, in order to enable practitioners to make more informed decisions when selecting a source dataset.

In light of the ongoing debate on the choice of source dataset for medical pre-training, we set out to investigate this with a series of systematic experiments on the difference of representations learned from natural (`ImageNet`) and medical (`RadImageNet`) source datasets on a range of (seven) medical targets. Our main contributions are:

- We extend the work presented in Mei et al. (2022) by doing a reproducibility study of four of their seven experiments (derived from three small medical targets: `breast`, `thyroid`, and `knee` datasets) and adding four additional medical imaging target datasets. Contrary to the findings in (Mei et al., 2022), we observe that in most cases, models pre-trained on `ImageNet` tend to perform better than those trained on `RadImageNet`. However, it is important to note that this discrepancy does not necessarily indicate the superiority of one source dataset over the other. Rather, it emphasizes the sensitivity of transfer performance to the choice of model architecture and hyperparameters.
- We investigate the learned intermediate representations of the models pre-trained on `ImageNet` and `RadImageNet` using Canonical Correlation Analysis (CCA) (Raghu et al., 2017; 2019). Our results reveal that the networks converge to distinct intermediate representations, and that these representations are even more dissimilar after fine-tuning.
- We show that although models pre-trained on `ImageNet` and `RadImageNet` learn different intermediate representations, they still produce similar predictions after fine-tuning.
- Our findings demonstrate limited feature reuse even in the first two layers of a convolutional neural network (CNN), contrary to findings by Raghu et al. (2019). We discover a negative relationship between model similarity (before and after fine-tuning) and the improvement in performance across all layers.

Our code and experiments are available on Github[1].

2 Related work

2.1 Pre-training on different data

Transfer performance degradation due to distribution shift is a known problem. This issue is particularly relevant in scenarios where the availability of large-scale in-domain supervised data for pre-training is limited. In light of this, a line of research has emerged that challenges the common practice of pre-training on `ImageNet` and instead explores pre-training on different in-domain source datasets:

Training on target data. El-Nouby et al. (2021) have shown the potential of using denoising autoencoders for pre-training on target data in a self-supervised manner. Although this approach has yielded promising results, the experiments were conducted using natural image target datasets. Even the smallest target dataset consisted of 8,000 images, making it unclear how well this approach would translate to the domain of medical imaging where the availability of large-scale target data is often limited.

Synthetic data has been shown to be a viable alternative to real-world data for pre-training, particularly in domains where labeled data is scarce. Kataoka et al. (2020) have explored pre-training on gray-scale automatically generated fractals with labels and showed that this approach can generate unlimited amounts

[1] `https://anonymous.4open.science/r/86380762B567/`

of synthetic labeled images, although it does not surpass the performance of pre-training on `ImageNet` in all cases.

Data augmentation is often used to increase dataset size, but Asano et al. (2020) showed that it can also be used to scale a single image for self-supervised pre-training, though this approach still falls short of using real diverse data. Even with millions of unlabeled images, it cannot fully bridge the gap between fully-supervised and self-supervised pre-training for deeper layers of a CNN.

Although aforementioned work in general computer vision has demonstrated the potential of synthetic and augmented data, the importance of large-scale labeled source datasets remains strong, particularly `ImageNet` in medical imaging (Brandt et al., 2021; Wen et al., 2021) despite its out-of-domain nature for medical targets. Recently, Mei et al. (2022) have demonstrated that `RadImageNet`, a large-scale dataset of radiology images similar in size to `ImageNet`, outperforms `ImageNet` on radiology target datasets. To further these findings, we investigate the potential of pre-training on the `RadImageNet` on a range of modalities that were not included in Mei et al. (2022) experiments, such as X-rays, dermoscopic images, and histopathological scans.

2.2 Effects of transfer learning

Transfer learning is a useful method for studying representations and generalization in deep neural networks. Yosinski et al. (2014) defined the generality of features learned by a convolutional layer based on their transferability between tasks. They analyzed representations learned in `ImageNet` models and found that the early layers form general features resembling Gabor filters and color blobs, while deeper layers become more task-specific. More recently, Raghu et al. (2019) studied feature reuse in medical imaging using transfer learning and found that this reuse is limited to the lowest two convolutional layers. Besides feature reuse, they demonstrated that the scaling of pre-trained weights can result in significant improvement in convergence speed.

Instead of investigating the transferability of weights at different layers, Gavrikov & Keuper (2022b) investigated the distributions of convolution filters learned by computer vision models and found that they only exhibit minor variations across various tasks, image domains, and datasets. They noted that models based on the same architecture tend to learn similar distributions when compared to each other, but differ significantly when compared to other architectures. The authors also discovered that medical imaging models do not learn fundamentally different filter distributions compared to models for other image domains. Based on these findings, they concluded that medical imaging models can be pre-trained with diverse image data from any domain.

Our work extends previous studies on the effects of pre-training by characterizing the representations learned from `ImageNet` and `RadImageNet` and investigating the implications of the source dataset on the learned representations. Our results further challenge the notion of the generality of early convolutional layers, showing that feature reuse and target task performance are negatively related in all convolutional layers. Our results provide additional evidence that the source domain may not be of high importance for pre-training medical imaging models, as we observe that even though `ImageNet` and `RadImageNet` pre-trained models converge to distinct hidden representations, their predictions are still similar.

3 Method

We outline our overall method in Figure 1. We fine-tune publicly available pre-trained `ImageNet` and `RadImageNet` weights on medical target datasets and quantify the model similarity by comparing the network activations over a sample of images from the target datasets using two similarity measures, Canonical Correlation Analysis (CCA, Section 3.1) and prediction similarity (Section 3.2).

3.1 Canonical Correlation Analysis

CCA (Hotelling, 1936) which is a statistical method used to analyze the relationship between two sets of variables.

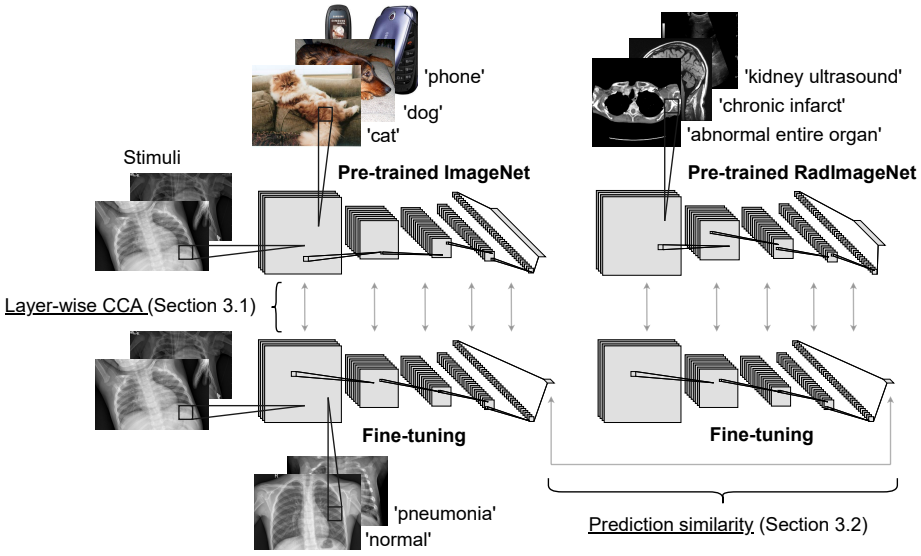


Figure 1: Overview of the experimental setup. Publicly available pre-trained `ImageNet` and `RadImageNet` weights are fine-tuned on medical targets. Model similarity is evaluated by comparing network activations over sampled stimuli images from target datasets using both CCA (described in Section 3.1) and prediction similarity (Section 3.2).

Let $\mathbf{X}$ be a dataset $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ of n data points, all consisting of p variables, and $\mathbf{Y}$ – a dataset of n data points $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ and q variables. CCA seeks to find the transformation matrices $\mathbf{A}$ and $\mathbf{B}$ that linearly combine the initial variables p and q in the datasets $\mathbf{X}$ and $\mathbf{Y}$ into $\min(p, q)$ canonical variables $\mathbf{X}\mathbf{a}^i$ and $\mathbf{Y}\mathbf{b}^i$ such that the correlation between these canonical variables is maximized:

$$\begin{aligned} \mathbf{a}^i, \mathbf{b}^i = \underset{\mathbf{a}^i, \mathbf{b}^i}{\operatorname{argmax}} \operatorname{corr}(\mathbf{X}\mathbf{a}^i, \mathbf{Y}\mathbf{b}^i) \\ \text{subject to } \forall_{j<i}\ \mathbf{X}\mathbf{a}^i \perp \mathbf{X}\mathbf{a}^j \\ \forall_{j<i}\ \mathbf{Y}\mathbf{b}^i \perp \mathbf{Y}\mathbf{b}^j \end{aligned}$$

The restrictions ensure that the canonical variables are orthogonal. This can be solved by defining substitutions $\bar{\mathbf{A}} = \mathbf{\Sigma}_X^{1/2}\mathbf{A}$ and $\bar{\mathbf{B}} = \mathbf{\Sigma}_Y^{1/2}\mathbf{B}$ obtaining:

$$\begin{aligned} \bar{\mathbf{A}}, \bar{\mathbf{B}} = \underset{\bar{\mathbf{A}}, \bar{\mathbf{B}}}{\operatorname{argmax}} \operatorname{tr}(\bar{\mathbf{A}}^\top \mathbf{\Sigma}_X^{-1/2} \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_Y^{-1/2} \bar{\mathbf{B}}) \\ \text{subject to } \bar{\mathbf{A}}^\top \bar{\mathbf{A}} = \mathbf{I} \\ \bar{\mathbf{B}}^\top \bar{\mathbf{B}} = \mathbf{I} \end{aligned}$$

Because of the orthogonality constraints, the solution is found by decomposing $\mathbf{\Sigma}_X^{-1/2} \mathbf{\Sigma}_{XY} \mathbf{\Sigma}_Y^{-1/2}$ into left and right singular vectors using singular value decomposition.

Layer-wise model similarity. Raghu et al. (2017) proposed the use of CCA for comparing representations learned by neural networks. CCA's invariance to linear combinations makes it suitable for comparing the representations learned by different models as the layer weights in neural networks are combined before being passed on (Morcos et al., 2018).

Raghu et al. (2019) used CCA to examine representations in medical imaging models. In order to maintain consistency with Raghu et al. (2019) results, we adopt the same approach of applying CCA to CNNs and use an open source CCA implementation by Raghu et al. (2017) available on GitHub[2].

In our case, $\mathbf{X}$ and $\mathbf{Y}$ are same-level layer activation vectors over n stimuli images sampled from a target dataset, in two models with different initializations. We extract these intermediate representations and use them as input to CCA to project the representations onto a common space, where the correlation between the projections is maximized. This common space can be thought of as a shared representation that captures the common patterns of activity across the compared networks. Then, layer-wise similarity at layer L is the average of the correlations between the canonical variables:

$$\rho_L = \sum_{i=1}^{p} \text{corr}(\mathbf{X}\mathbf{a}^i, \mathbf{Y}\mathbf{b}^i) \tag{1}$$

Intermediate representations extracted from CNNs are of shape (n, h_L, w_L, p_L), where h_L, w_L are the layer spatial dimensions and p_L is the number of channels in the layer. These representations are reshaped into $\mathbf{X}$ and $\mathbf{Y}$ matrices of shape $(n \times h_L \times w_L, p_L)$. As CCA is sensitive to the shape of the input matrices and the shapes vary across the layers within a network, we sample n, p_L and then calculate layer similarity ρ_L. This is repeated five times and the final layer similarity is obtained by averaging layer similarities ρ_L.

3.2 Prediction similarity

We calculate prediction similarity as described in Mania et al. (2019). A model mistake is defined as $q_f(x, y) = \mathbf{1}_{\text{f(x)}\neq\text{y}}$, where x is an image from the test set, y is its label and f is a network fine-tuned on the target training set. Then the prediction similarity of two networks $f_{\texttt{ImageNet}}$ and $f_{\texttt{RadImageNet}}$, fine-tuned on the same target dataset, is:

$$\mathbb{P}(q_{f_{\texttt{ImageNet}}}(x, y) = q_{f_{\texttt{RadImageNet}}}(x, y)) \tag{2}$$

Therefore, the prediction similarity is the probability that two networks will make the same errors. To gauge the prediction similarities between `ImageNet` and `RadImageNet` models, we compare them to the prediction similarity of two classifiers with the same accuracy as `ImageNet` and `RadImageNet` models but otherwise random predictions. If the mistakes made by two models with accuracy a_1 and a_2 are independent, the similarity of their predictions is equal to $a_1a_2 + (1 - a_1)(1 - a_2)$.

4 Experimental setup

4.1 Datasets

Source. We use publicly available pre-trained `ImageNet` (Deng et al., 2009) and `RadImageNet` (Mei et al., 2022) weights as source tasks in our experiments.

Target. We investigate transferability to several medical target datasets. In particular, to five radiology `RadImageNet` in-domain datasets, and two out-of-domain datasets in the fields of dermatology and microscopy. A representative image from each dataset can be seen in Figure 2.

1) `Chest`. Chest X-rays (Kermany et al., 2018) dataset contains chest X-ray images from pediatric patients aged one - five years old, labeled by expert physicians with binary labels of 'normal' or 'pneumonia'. The dataset has 5,856 images, with 1,583 labeled as 'normal' and 4,273 labeled as 'pneumonia'. The image size varies, with dimensions ranging from 72×72 to $2,916 \times 2,583$ pixels.

2) `Breast`. Breast ultrasound (Al-Dhabyani et al., 2020) dataset is collected for the detection of breast cancer. The images have a range of sizes, from 190×335 to $1,048 \times 578$ pixels. The dataset is divided into

[2] `https://github.com/google/svcca`

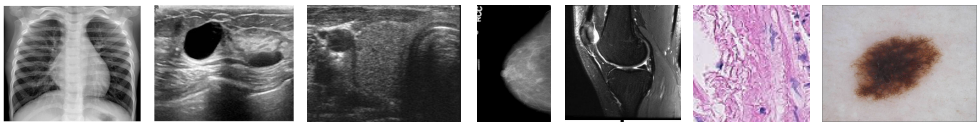


Figure 2: Example images of (a) `chest`, (b) `breast`, (c) `thyroid`, (d) `mammograms`, (e) `knee`, (f) `pcam-small`, and (g) `ISIC` datasets.

three classes: normal, benign, and malignant images. However, following Mei et al. (2022), we use a binary classification of 'benign' and 'malignant' for our analysis.

3) `Thyroid.` The Digital Database of Thyroid Ultrasound Images (DDTI) (Pedraza et al., 2015) contains 480 images of size 569×360 pixels, extracted from thyroid ultrasound videos. The images have been annotated by radiologists into five categories. Following Mei et al. (2022)'s study, these categories were transformed into binary labels: 'normal' for categories (1) normal thyroid, (2) benign and (3) no suspicious ultrasound (US) feature, and 'malignant' for categories (4a) one suspicious US feature, (4b) two suspicious US features, (4c) three or four suspicious US features and (5) five suspicious features.

4) `Mammograms.` Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) (Sawyer-Lee et al., 2016; Lee et al., 2017; Clark et al., 2013) is a dataset that targets breast cancer detection. It contains scanned film mammograms with pathologically confirmed labels: 'benign' (2,111 images) or 'malignant' (1,457 images) with image sizes ranging from $1,846 \times 4,006$ to $5,431 \times 6,871$ pixels.

5) `Knee.` MRNet (Bien et al., 2018) is a collection of 3D knee MRI scans. The labels for the dataset were obtained through manual extraction from clinical reports. Following Mei et al. (2022)'s study, we use extracted 2D sagittal views (1 to 3 samples per scan) amounting to a total of 4,235 'normal', 569 'ACL' (anterior cruciate ligament), and 418 'meniscal tear' images, all of size 256×256 pixels.

6) `PCam-small.` PatchCamelyon (Veeling et al., 2018) is a metastatic tissue classification dataset consisting of 237,680 colored patches extracted from histopathological scans of lymph node sections. The images are labeled as 'positive' or 'negative' based on the presence of metastatic tissue. To simulate a realistic target dataset size, a random subset of 10,000 images was created, with 5,026 positive and 4,974 negative samples.

7) `ISIC.` ISIC 2018 Challenge - Task 3: Lesion Diagnosis (Codella et al., 2019; Tschandl et al., 2018) - a dermoscopic lesion image dataset released for the task of skin lesion classification. The dataset comprises images of 600×450 pixels, which are split into seven disease categories. The dataset is unbalanced, with the class 'melanocytic nevus' having the most samples at 6,705, and the class 'dermatofibroma' having the least number of samples at 115.

Due to memory constraints, we reduced the original image sizes for most of the target datasets using interpolation without image cropping. Table 1 provides details of the image sizes and number of images used for fine-tuning on each target dataset. As we used publicly available pre-trained weights images were preprocessed to align with the pre-trained weights. As per the approach in Mei et al. (2022), we normalized the images with respect to the `ImageNet` dataset.

4.2 Fine-tuning

We select ResNet50 (He et al., 2016) as the standard model architecture for our experiments. This architecture is widely adopted in the field of medical imaging and has been demonstrated to be a strong performer in various image classification tasks.

We fine-tuned pre-trained networks using an average pooling layer, a dropout layer with a probability of 0.5, and the Adam optimizer with an initial learning rate of 1e-05. Due to memory limitations, the batch size was set to 32 for the datasets with the image size 224×224, and to 128 for the rest. The models were trained for a maximum of 200 epochs, with early stopping after 3 epochs of no decrease in validation loss. This was done to prevent overfitting and ensure that the models generalize well to unseen data.

Table 1: Target datasets with number of images, number of classes and image size used to fine-tune the pre-trained `ImageNet` and `RadImageNet` weights.

Dataset	Size	Classes	Image size
`Chest`	5,856	2	112×112
`Breast`	780	2	224×224
`Thyroid`	480	2	224×224
`Mammograms`	3,568	2	224×224
`PCam-small`	10,000	2	96×96
`ISIC`	10,015	7	112×112

Table 2: Mean AUC ± std (both ×100) after fine-tuning on target datasets. Underlined is the highest mean AUC per dataset.

	`ImageNet`		`RadIamgeNet`		**Random init**
Target dataset	**No Freeze**	**Freeze**	**No Freeze**	**Freeze**	**No Freeze**
`Thyroid`	52.3 ± 8.9	<u>58.5 ± 5.1</u>	50.2 ± 4.1	49.4 ± 5.9	52.9 ± 5.1
`Breast`	<u>90.5 ± 3.8</u>	89.8 ± 3.0	88.2 ± 4.2	72.4 ± 22.1	51.2 ± 10.0
`Chest`	98.7 ± 0.4	98.6 ± 0.6	<u>98.8 ± 0.2</u>	98.7 ± 0.3	82.5 ± 1.0
`Mammograms`	63.4 ± 4.3	<u>68.8 ± 2.0</u>	62.0 ± 12.2	66.2 ± 6.3	49.6 ± 3.7
`Knee`	<u>91.5 ± 1.1</u>	89.0 ± 3.1	89.3 ± 6.3	63.8 ± 5.7	68.3 ± 11.0
`ISIC`	<u>94.2 ± 1.3</u>	93.1 ± 2.3	90.8 ± 0.8	90.6 ± 0.7	84.0 ± 2.3
`Pcam-small`	<u>93.8 ± 1.1</u>	93.2 ± 2.5	87.1 ± 2.3	86.0 ± 1.9	73.4 ± 8.2

In addition to full fine-tuning, we used a freezing strategy where we froze all the pre-trained weights to train the classification layer first and then fine-tuned the whole network with the same hyperparameters as above.

Models were implemented using Keras (Chollet et al., 2015) library and fine-tuned on 3 NVIDIA GeForce RTX 2070 GPU cards.

4.3 Evaluation

We fine-tune the pre-trained networks on each target dataset using five-fold cross-validation approach. To ensure patient-independent validation where patient information is available (`chest`, `thyroid`, `mammograms`, `knee`), the target data is split such that the same patient is only present in either the training or validation split. We evaluate fine-tuned network performance using AUC (area under the receiver operating characteristic curve), while a one-against-all AUC, weighted by class priors, is used for multi-class problems. Model similarity is evaluated using CCA and prediction similarity as described in Section 3.

5 Results

We carried out a series of experiments to evaluate the effect of pre-training on `ImageNet` and `RadImageNet` on model accuracy and learned representations after fine-tuning on medical targets. In the following section, we present the results of our experiments and provide a thorough analysis of the findings. The results offer new insights into the effects of transfer learning and provide a foundation for future research in this field.

5.1 `ImageNet` tends to outperform `RadImageNet`

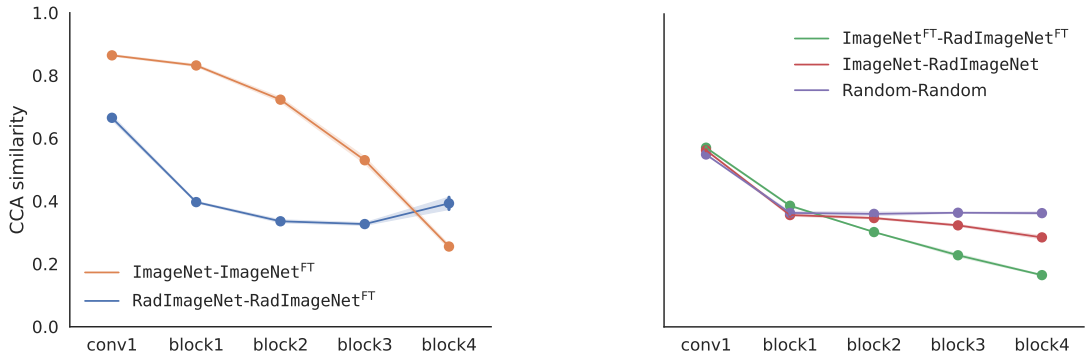
We show the AUC performances in Table 2. Overall, the unfrozen `ImageNet` leads to the highest AUCs in four out of the seven datasets, followed by the frozen `ImageNet` in two datasets. Only `Chest` reaches the highest performance with frozen `RadImageNet` weight, though we note that all performances were close for

this dataset. In general, we found that the performance of `ImageNet` and `RadImageNet` was comparable for radiology targets. However, when evaluating on the `ISIC` and `pcam-small` datasets, `ImageNet` produced considerably better results. This discrepancy is likely due to the fact that `RadImageNet` consists of grayscale images, which might limit its performance in dermatology and microscopy images despite its medical nature.

Compared to Mei et al. (2022), we get similar `ImageNet` AUC for `knee` and `breast`, but much lower `RadImageNet` AUC. We note that Mei et al. (2022) evaluated `RadImageNet` performance using a subset of 349 images from the `thyroid` dataset, compared to 480 images available and split the classification of ACL and meniscal tear into two separate tasks for the `knee` dataset. Furthermore, the results reported by Mei et al. (2022) were aggregated over four different model architectures and multiple hyperparameter (freezing and learning rate) combinations, making it difficult to directly compare our ResNet50 results. This highlights the sensitivity of transfer performance to the choice of model architecture and hyperparameters.

5.2 Layer-wise representations become more different after fine-tuning

In Figure 3 we show layer-wise `ImageNet` and `RadImageNet` CCA similarity to themselves after fine-tuning, `ImageNet`$^{\text{FT}}$ and `RadImageNet`$^{\text{FT}}$ respectively (Figure 3a), as well as layer-wise `ImageNet` and `RadImageNet` CCA similarity before and after fine-tuning (Figure 3b). `ImageNet` weights change less during fine-tuning (orange line on the left). The two networks converge to distinct solutions after fine-tuning (green line on the right), even more distinct than before fine-tuning, and their similarity is significantly lower when compared to the similarity of two random initializations. Here we only provide results on `knee`, however we observed similar patterns for the other target datasets (Appendix A).



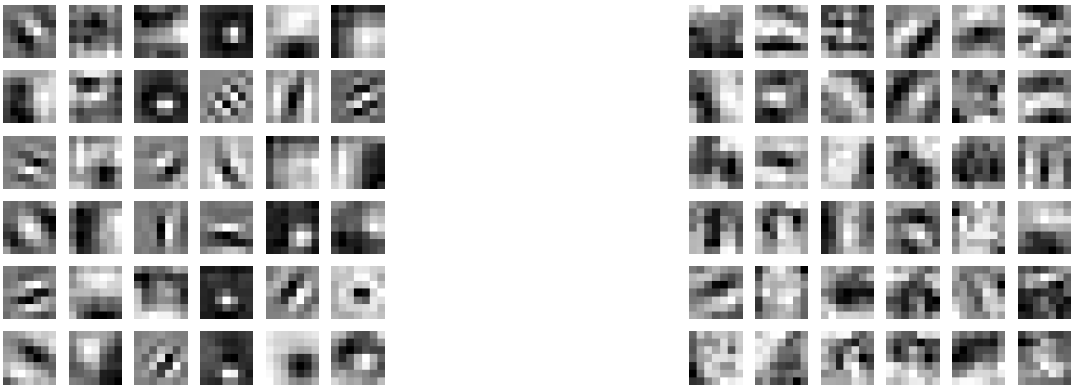
(a) Before versus after fine-tuning

(b) `ImageNet` versus `RadImageNet`

Figure 3: Layer-wise CCA similarity (Equation 1) of (a) a network to itself before and after fine-tuning on `knee` and (b) `ImageNet` to `RadImageNet`. `ImageNet-ImageNet`$^{\text{FT}}$ similarity (orange line) is higher (the weights change less after fine-tuning) than `RadImageNet-RadImageNet`$^{\text{FT}}$ similarity (blue line). `ImageNet`$^{\text{FT}}$ and `RadImageNet`$^{\text{FT}}$ are highly dissimilar (green line), even more dissimilar than before fine-tuning (red line). They are also more dissimilar than two randomly initialized networks (purple line). Error bars present mean $\pm$ std over five-fold cross-validation.

For experiments where pre-trained weights were initially frozen and fine-tuned after training the classification layer, we observed essentially similar trends of weight similarity as shown in Figure 3, but with higher variability between the folds of the cross-validation.

The results of the layer-wise CCA similarity analysis reveal that `ImageNet` and `RadImageNet` converge to distinct solutions after fine-tuning on the same target dataset, to the extent that they become even more dissimilar than before fine-tuning. This outcome contradicts our expectation that the representations of the two networks would become more similar after training on the same target dataset. This discrepancy may be due to memorization of the target dataset by one or both of the networks, as suggested by the findings of Morcos et al. (2018). They found that networks trained to classify randomized labels, hence memorizing the data, tend to converge to more distinct solutions compared to networks that generalize to unseen data.



(a) ImageNet (b) RadImageNet

Figure 4: First 36 conv1 filters of ResNet50 pre-trained on (a) ImageNet and (b) RadImageNet. Observe that the filters in ImageNet have a more pronounced resemblance to Gabor filters.

The stability of early layer weights during fine-tuning is often attributed to their capture of general features, such as edge detectors (Raghu et al., 2019; Yosinski et al., 2014), which are necessary regardless of the target domain and task. Our results are consistent with this hypothesis, in that we observe that early layers change less compared to later layers (Figure 3a). However, we also found an interesting and contradictory result, in that the early layers of ImageNet and RadImageNet are as similar as two randomly initialized layers. This observation is inconsistent with the intuition that early layers capture general features, as we would expect them to be more similar than randomly initialized layers if they indeed capture general features.

When we examine the first convolutional layer filters pre-trained on ImageNet and RadImageNet (Figure 4), we observe that the filters in ImageNet more closely resemble Gabor filters, while those in RadImageNet are more fuzzy. This is expected, as natural images often contain regular structures, such as 90 degree angles and edges, that are typically less prominent in some of radiology images, resulting in less distinct edges in the filters learned from radiology images. Interestingly, these different first layer features in both ImageNet and RadImageNet, without changing significantly during fine-tuning (Figure 3), lead to comparable performance in most cases (Table 2).

5.3 Networks make similar mistakes after fine-tuning

In order to further understand the similarity between ImageNet and RadImageNet pre-trained networks, we compared their predictions after fine-tuning on medical targets, as shown in Figure 5. Despite the networks converging to different hidden representations, as evidenced in Figure 3, their predictions were found to be more similar than expected for independent predictions. The fact that both ImageNet and RadImageNet were fine-tuned on the same target dataset may explain this observation. However, it also suggests the possibility that the networks are learning similar misleading cues in the data, resulting in similar misclassifications.

5.4 Higher weight similarity associated with less AUC improvement

Our results challenge the commonly held belief that the benefits of transfer learning in deep neural networks come from feature reuse in the early layers. Figure 6 shows that the gain in AUC due to pre-training is negatively related to the weight CCA similarity between the pre-trained and fine-tuned networks. This means that models that reused the pre-trained weights without making changes during fine-tuning performed worse, as opposed to those that adapted the weights. This negative relationship was observed across all layers and was particularly pronounced in the RadImageNet pre-trained networks.

Our findings align with recent results in the Natural Language Processing field which demonstrate that the benefits of pre-training are not related to knowledge transfer (Krishna et al., 2021). Additionally, our results complement the findings of Raghu et al. (2019) who showed that there are feature-independent benefits of

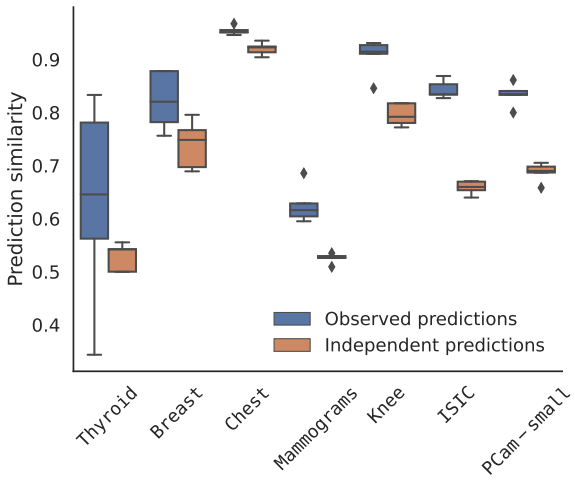
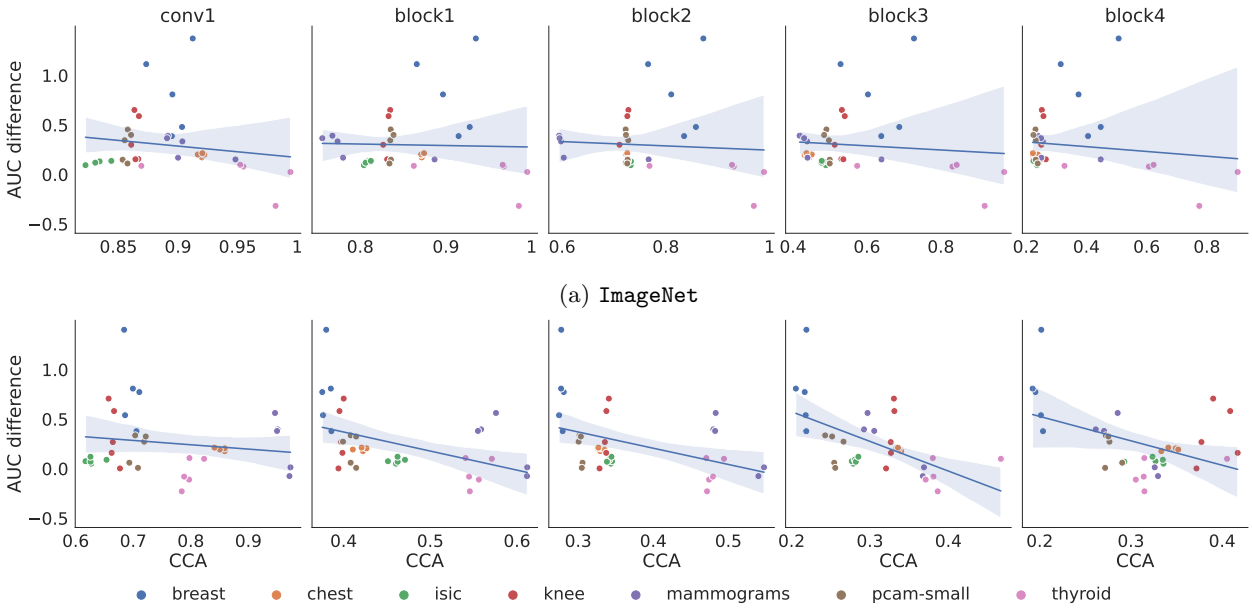


Figure 5: Prediction similarity (Equation 2) between $\texttt{ImageNet}^{\text{FT}}$ and $\texttt{RadImageNet}^{\text{FT}}$ (blue box plot) compared to prediction similarity of two networks that would make independent mistakes (orange box plot). $\texttt{ImageNet}^{\text{FT}}$ and $\texttt{RadImageNet}^{\text{FT}}$ predictions are more correlated than expected for independent predictions on average across all target datasets, with notable variation observed for thyroid dataset.



(a) `ImageNet`

(b) `RadImageNet`

Figure 6: AUC pre-training gains over random initialization for seven target datasets vs CCA similarity before and after fine-tuning on those targets, for (a) `ImageNet` and (b) `RadImageNet`, with 95% confidence interval. Higher CCA similarity after fine-tuning is associated with lower AUC gains, observed across all layers with a more pronounced effect on `RadImageNet`. Note that the scaling on the x-axes are different in each plot for visibility, and for `RadImageNet` the CCA similarity is lower overall.

using pre-trained weights, such as better scaling compared to random initialization. These results highlight the complex nature of transfer learning and the need for further investigation into the underlying mechanisms that drive its performance benefits.

6 Discussion

In our experiments `ImageNet` generally outperformed `RadImageNet` on medical target datasets, in contrast to the earlier results reported in Mei et al. (2022). However, this highlights the sensitivity of source dataset transfer performance to the model architecture and hyperparameters, rather than the inherent superiority of one source dataset over the other. While our study did not comprehensively analyze this sensitivity, interested readers can refer to related studies, such as Raghu et al. (2019) and Wen et al. (2021), which offer valuable insights into the impact of model architecture and hyperparameters on transfer performance.

Contrary to our preconceptions about transfer learning, our analysis with CCA found that the models converged to distinct intermediate representations and that these representations are even more dissimilar after fine-tuning on the same target dataset. Despite distinct intermediate representations, model predictions on an instance level exhibit a significant degree of similarity. This extends Mania et al. (2019) findings which showed that `ImageNet` models are similar in their predictions even with different architectures. We extend upon the findings of Raghu et al. (2019), who demonstrated limited feature reuse in the lowest two layers of a CNN. We discover that even the features in the first two layers are not reused to the same extent as previously believed.

While our results are similar across seven different target datasets, these tasks are not fully representative of the medical imaging field as a whole. Researchers have hypothesized that for 3D target tasks such as CT or MRI, 3D pre-training might be a better alternative to 2D pre-training (Wong et al., 2018; Brandt et al., 2021), and studies have shown that incorporating information from the third dimension might be beneficial for performance (Yang et al., 2021; Taleb et al., 2020). However, `RadImageNet` only has 2D images, even though some of the original images are 3D. From our results, we are unable to conclude that `ImageNet` weights are generally more suitable for medical target tasks.

Another limitation is in only using AUC as the metric for evaluating performance. Although this is a commonly used metric for classification tasks in medical imaging, there might be nuances across applications where it could be important to consider alternative metrics, such as calibration (Reinke et al., 2021).

Our results suggest that the implications of using `ImageNet` in medical image classification go beyond performance alone. In particular, there might be an issue with the memorization of spurious patterns in the data. This can potentially have consequences with respect to algorithmic bias and fairness. For example, see Gichoya et al. (2022) where `ImageNet` pre-trained networks memorize patient race. Memorization also makes a network more vulnerable to adversarial attacks (Bortsova et al., 2021). It would be interesting to further explore these properties in future research.

While `RadImageNet`'s comparable size to `ImageNet` allowed for a unique opportunity to compare natural and medical source datasets, it is important to note that the two datasets have significant differences beyond their domains. For instance, `RadImageNet` has differences in color, number of classes, and diversity in data due to its limited number of patients. To further explore the impact of these differences in greater detail, future research could consider including Ecoset (Mehrer et al., 2021), a natural image dataset with 565 basic-level categories selected to better reflect the human perceptual and cognitive experience.

7 Conclusions

Transfer learning is a key strategy to leverage knowledge from the models pre-trained on large-scale datasets to deal with the challenge of small medical datasets. In this study, we investigated the transferability of two different domain sources (natural: `ImageNet` and medical: `RadImageNet`) to several target medical image classification tasks. Our results show that pre-training on `ImageNet` outperformed `RadImageNet` in most cases. Furthermore, we delved deeper into the learned representations after fine-tuning by using CCA and comparing the similarity of predictions. Although the models converged to distinct representations, we found they made similar predictions. Finally, we observed limited reuse of features, even in the early convolutional layers, after fine-tuning.

References

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, February 2020. ISSN 2352-3409. doi: 10.1016/j.dib.2019.104863. URL https://www.sciencedirect.com/science/article/pii/S2352340919312181.

Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image, February 2020. URL http://arxiv.org/abs/1904.13132. arXiv:1904.13132 [cs].

Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Russell J. Stewart, Francis G. Blankenberg, David B. Larson, Ricky H. Jones, Curtis P. Langlotz, Andrew Y. Ng, and Matthew P. Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS medicine*, 15(11):e1002699, November 2018. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002699.

Gerda Bortsova, Cristina González-Gonzalo, Suzanne C. Wetstein, Florian Dubost, Ioannis Katramados, Laurens Hogeweg, Bart Liefers, Bram van Ginneken, Josien P. W. Pluim, Mitko Veta, Clara I. Sánchez, and Marleen de Bruijne. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Medical Image Analysis*, 73:102141, October 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102141. URL https://www.sciencedirect.com/science/article/pii/S1361841521001870.

Irma van den Brandt, Floris Fok, Bas Mulders, Joaquin Vanschoren, and Veronika Cheplygina. Cats, not CAT scans: a study of dataset similarity in transfer learning for 2D medical image classification, July 2021. URL http://arxiv.org/abs/2107.05940. arXiv:2107.05940 [cs].

Veronika Cheplygina, Marleen de Bruijne, and Josien P. W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54: 280–296, May 2019. ISSN 1361-8423. doi: 10.1016/j.media.2019.03.009.

Francois Chollet et al. Keras, 2015. URL https://keras.io.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, December 2013. ISSN 1618-727X. doi: 10.1007/s10278-013-9622-7. URL https://doi.org/10.1007/s10278-013-9622-7.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC), March 2019. URL http://arxiv.org/abs/1902.03368. arXiv:1902.03368 [cs].

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are Large-scale Datasets Necessary for Self-Supervised Pre-training?, December 2021. URL http://arxiv.org/abs/2112.10740. arXiv:2112.10740 [cs].

Paul Gavrikov and Janis Keuper. CNN Filter DB: An Empirical Investigation of Trained Convolutional Filters. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19044–19054, New Orleans, LA, USA, June 2022a. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.01848. URL https://ieeexplore.ieee.org/document/9880244/.

Paul Gavrikov and Janis Keuper. Does Medical Imaging learn different Convolution Filters?, October 2022b. URL http://arxiv.org/abs/2210.13799. arXiv:2210.13799 [cs, eess].

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, November 2022. URL http://arxiv.org/abs/1811.12231. arXiv:1811.12231 [cs, q-bio, stat].

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, June 2022. ISSN 2589-7500. doi: 10.1016/S2589-7500(22)00063-2. URL https://www.sciencedirect.com/science/article/pii/S2589750022000632.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

Harold Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321–377, 1936. ISSN 0006-3444. doi: 10.2307/2333955. URL https://www.jstor.org/stable/2333955. Publisher: [Oxford University Press, Biometrika Trust].

Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without Natural Images. *Asian Conference on Computer Vision (ACCV)*, 2020.

Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y. Ng, and Pranav Rajpurkar. CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 116–124, Virtual Event USA, April 2021. ACM. ISBN 978-1-4503-8359-2. doi: 10.1145/3450439.3451867. URL https://dl.acm.org/doi/10.1145/3450439.3451867.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y. L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A. N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131.e9, February 2018. ISSN 1097-4172. doi: 10.1016/j.cell.2018.02.010.

Kundan Krishna, Jeffrey Bigham, and Zachary C. Lipton. Does Pretraining for Summarization Require Knowledge Transfer? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3178–3189, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.273. URL https://aclanthology.org/2021.findings-emnlp.273.

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):170177, December 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.177. URL https://www.nature.com/articles/sdata2017177. Number: 1 Publisher: Nature Publishing Group.

Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model Similarity Mitigates Test Set Overuse, May 2019. URL http://arxiv.org/abs/1905.12580. arXiv:1905.12580 [cs, stat].

Johannes Mehrer, Courtney J. Spoerer, Emer C. Jones, Nikolaus Kriegeskorte, and Tim C. Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Science*, 118:e2011417118, February 2021. ISSN 0027-8424. doi: 10.1073/pnas.2011417118. URL https://ui.adsabs.harvard.edu/abs/2021PNAS..11811417M. ADS Bibcode: 2021PNAS..11811417M.

Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiology: Artificial Intelligence*, 4(5):e210315, September 2022. doi: 10.1148/ryai.210315. URL https://pubs.rsna.org/doi/10.1148/ryai.210315. Publisher: Radiological Society of North America.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a7a3d70c6d17a73140918996d03c014f-Abstract.html.

Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, volume 9287, pp. 188–193. SPIE, January 2015. doi: 10.1117/12.2073532. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9287/92870W/An-open-access-thyroid-ultrasound-image-database/10.1117/12.2073532.full.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability, November 2017. URL http://arxiv.org/abs/1706.05806. arXiv:1706.05806 [cs, stat].

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging, October 2019. URL http://arxiv.org/abs/1902.07208. Number: arXiv:1902.07208 arXiv:1902.07208 [cs, stat].

Annika Reinke, Minu D Tizabi, Carole H Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, et al. Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642*, 2021.

Rebecca Sawyer-Lee, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin. Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM), 2016. URL https://wiki.cancerimagingarchive.net/x/lZNXAQ. Version Number: 1 Type: dataset.

Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, May 2016. ISSN 1558-254X. doi: 10.1109/TMI.2016.2528162.

Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3D Self-Supervised Methods for Medical Imaging. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18158–18172. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d2dc6368837861b42020ee72b0896182-Abstract.html.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161, August 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.161. URL https://doi.org/10.1038/sdata.2018.161.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation Equivariant CNNs for Digital Pathology, June 2018. URL http://arxiv.org/abs/1806.03962. arXiv:1806.03962 [cs, stat].

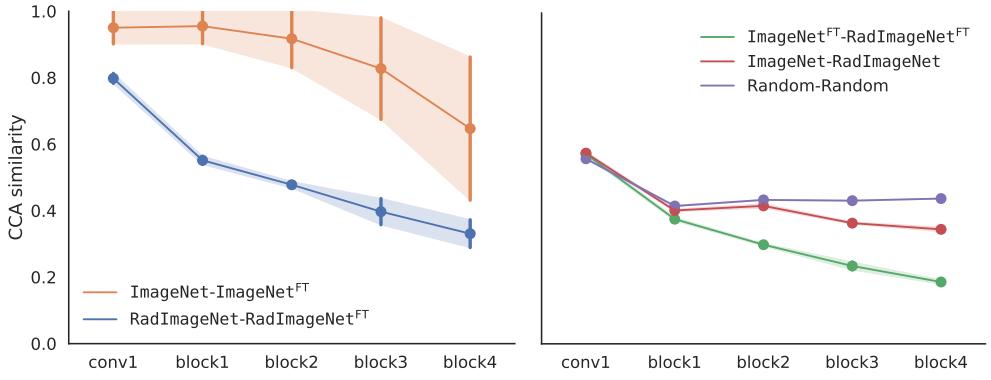
Yang Wen, Leiting Chen, Yu Deng, and Chuan Zhou. Rethinking pre-training on medical imaging. *Journal of Visual Communication and Image Representation*, 78:103145, July 2021. ISSN 1047-3203. doi: 10.1016/j.jvcir.2021.103145. URL https://www.sciencedirect.com/science/article/pii/S1047320321000894.

Ken CL Wong, Tanveer Syeda-Mahmood, and Mehdi Moradi. Building medical image classifiers with very limited data using segmentation networks. *Medical image analysis*, 49:105–116, 2018.

Jiancheng Yang, Xiaoyang Huang, Yi He, Jingwei Xu, Canqian Yang, Guozheng Xu, and Bingbing Ni. Reinventing 2D Convolutions for 3D Images. *IEEE Journal of Biomedical and Health Informatics*, 25 (8):3009–3018, August 2021. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2021.3049452. URL http://arxiv.org/abs/1911.10477. arXiv:1911.10477 [cs, eess].

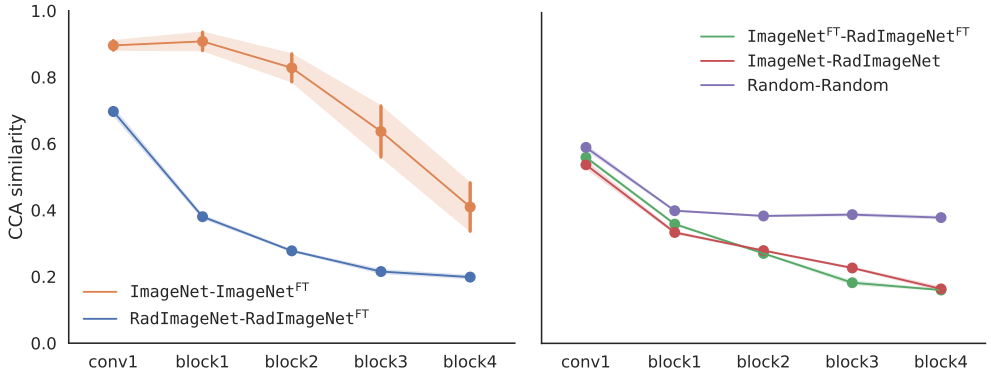
Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://papers.nips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html.

A Layer-wise CCA similarity



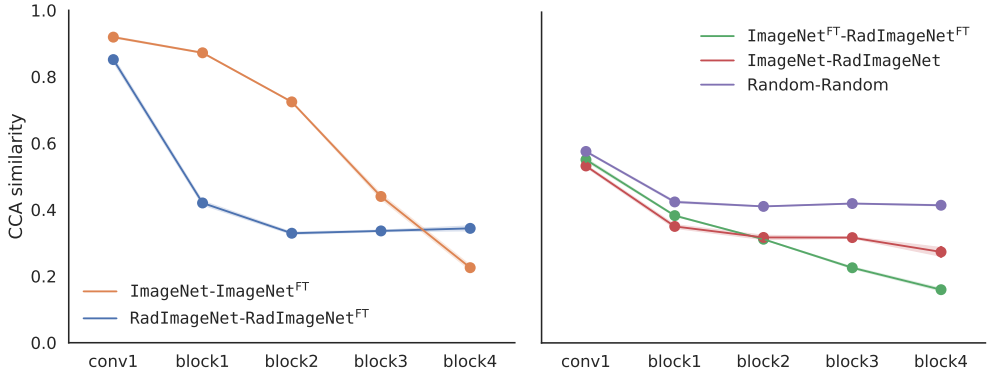
(a) Before versus after fine-tuning (b) `ImageNet` versus `RadImageNet`

Figure 7: Layer-wise CCA similarity of networks fine-tuned on `thyroid`.



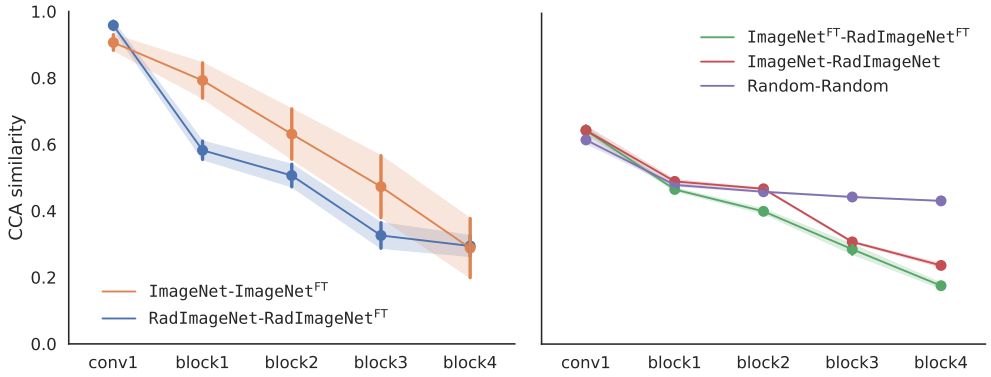
(a) Before versus after fine-tuning (b) `ImageNet` versus `RadImageNet`

Figure 8: Layer-wise CCA similarity of networks fine-tuned on `breast`.



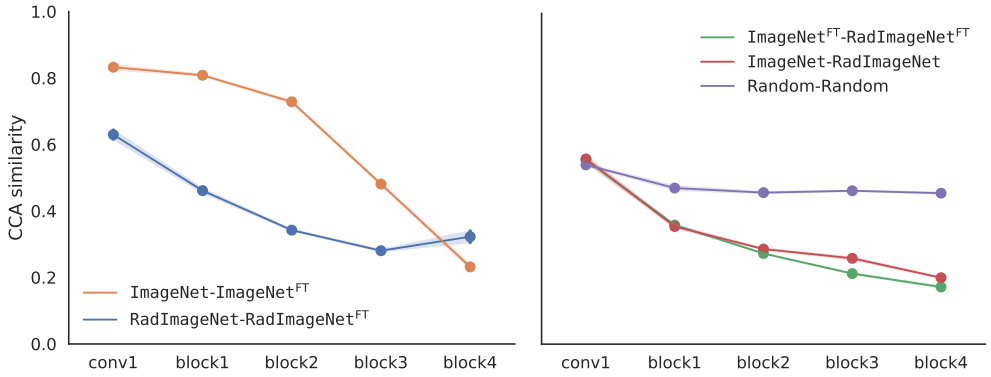
(a) Before versus after fine-tuning (b) `ImageNet` versus `RadImageNet`

Figure 9: Layer-wise CCA similarity of networks fine-tuned on `chest`.



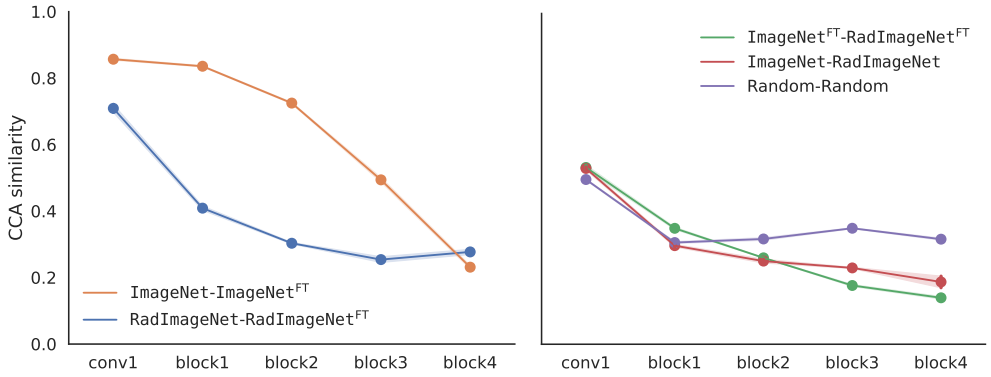
(a) Before versus after fine-tuning (b) `ImageNet` versus `RadImageNet`

Figure 10: Layer-wise CCA similarity of networks fine-tuned on `mammograms`.



(a) Before versus after fine-tuning (b) `ImageNet` versus `RadImageNet`

Figure 11: Layer-wise CCA similarity of networks fine-tuned on `isic`.



(a) Before versus after fine-tuning (b) `ImageNet` versus `RadImageNet`

Figure 12: Layer-wise CCA similarity of networks fine-tuned on `pcam-small`.