← Go to **CHIL 2023 Conference** homepage (/group?id=chilconference.org/CHIL/2023/Conference)

Revisiting Hidden Representations in Transfer Learning for Medical Imaging

(/pdf?id=6h2nbnvkAp)

Anonymous

15 Feb 2023 (modified: 11 Apr 2023) Submitted to CHIL2023 Readers: Conference, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers, Paper26 Authors Show Bibtex Show Revisions (/revisions?id=6h2nbnvkAp)

Keywords: Transfer learning, medical image classification, representation learning, model similarity

TL;DR: We show, on seven medical target tasks, that transfer learning is more effective when using natural images as a source, and challenge the conventional notion that transfer learning is effective due to the reuse of general features.

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.

Track: Track 2: Applications and Practice

Data Modalities: Images

Subject Areas: Representation Learning, Supervised Learning

Data Availability: Not Applicable

Code Availability: Yes

Entered Conflicts: Yes

Revealed to Dovile Juodelyte, Amelia Jiménez-Sánchez, Veronika Cheplygina

14 Feb 2023 (modified: 11 Apr 2023) Submitted to CHIL2023

Authors: *Dovile Juodelyte (/profile?id=~Dovile_Juodelyte1), Amelia Jiménez-Sánchez (/profile?id=~Amelia_Jim%C3%A9nez-S%C3%A1nchez1), Veronika Cheplygina (/profile?id=~Veronika_Cheplygina2)*

Reviewer Nomination: doju@itu.dk

Reply Type: all Author: everybody Visible To: all readers **11 Replies**

Hidden From: nobody

[−] Paper Decision

CHIL 2023 Conference Program Chairs

10 Apr 2023 CHIL 2023 Conference Paper26 Decision Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers, Paper26 Authors

Decision: Reject

[−]

Meta Review of Paper26 by Area Chair TfZy

CHIL 2023 Conference Paper26 Area Chair TfZy

07 Apr 2023 CHIL 2023 Conference Paper26 Meta Review Readers: Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Authors, Paper26 Reviewers Submitted, Program Chairs

Metareview:
The paper studies the transfer learning of DNNs from the general domain to medical domain imaging, through an empirical approach. The reviewers have agreed on the importance of the problem, but are mixed about the technical novelty. The authors have addressed most of the reviewers' concerns in the rebuttal, but the main one on technical/theoretical novelty still remains.

Confidence: 3: The area chair is somewhat confident

Recommendation: Reject

[−]

Official Review of Paper26 by Reviewer TpM4

CHIL 2023 Conference Paper26 Reviewer TpM4

18 Mar 2023 CHIL 2023 Conference Paper26 Official Review Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers Submitted, Paper26 Authors

Summary Of Paper:
This paper studies the effects of pre-training with natural vs. medical image domain sources (i.e., ImageNet vs. RadImageNet) on 7 target medical image classification tasks. The authors characterize differences between learned representations after finetuning and the predictions made, and show that pre-training with ImageNet outperforms pre-training with the medical source RadImageNet. They show that pre-training with different sources may learn distinct intermediate representations but still result in similar predictions, and offer insights on feature reuse and performance improvement with transfer learning.

Relevance To Healthcare: 4: High

Technical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work

Clarity: 4: High

Strengths:
This paper focuses on an important problem in transfer learning for medical imaging. The authors conduct extensive experiments to reveal new insights that can inform pre-training strategies for medical image classification tasks. The writing is clear and sound.

Weaknesses:
It is unclear how sensitive the results are to choice of model architecture and hyperparameters. As the differences between results in Mei 2022 and this paper may be attributed to choice of model architecture and hyperparameters, it is important that the authors provide fuller sensitivity analyses. General claims on relative performance of pre-training with ImageNet vs. RadImageNet may need to be toned down as they might be sensitive to the specific model architecture and hyperparameter choices.

Suggestions/Questions For Authors:
The influence of choice of source domain data on performance is multifactorial. ImageNet and RadImageNet are not only different in the domain of origin but also in many other characteristics. Hence, would suggest that authors take a more careful approach in attributing the results to the domain of origin of data used for pre-training.

Overall Recommendation: 4: Accept - Paper is in the top 50% of all submissions. Reviewer would recommend it to a colleague.

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

[−] Authors' Response

CHIL 2023 Conference Paper26 Authors Dovile Juodelyte (/profile?id=~Dovile_Juodelyte1) (privately revealed to you)

27 Mar 2023 (modified: 29 Mar 2023) CHIL 2023 Conference Paper26 Official Comment Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers Submitted, Paper26 Authors

Comment:

We would like to sincerely thank the reviewer for their time and effort in reviewing our paper, as well as the valuable comments and suggestions that have been very helpful in improving the quality of our work. We have carefully considered your feedback and made the necessary revisions to better reflect our findings on the comparative performance of ImageNet and RadImageNet models. Additionally, we have expanded our discussion on the dataset differences beyond the difference in domain. We have provided comments alongside the changes made below:

Model architecture and hyperparameters

- General claims on relative performance of pre-training with ImageNet vs. RadImageNet may need to be toned down as they might be sensitive to the specific model architecture and hyperparameter choices.
    - We recognize that general claims regarding the relative performance of pre-training with ImageNet versus RadImageNet are outside the scope of this study. Our goal in including these comparisons was to provide additional empirical data on the performance of RadImageNet, which is a very new dataset. We want to highlight performance sensitivity to hyperparameters, which is mentioned in the results section and for that reason we further elaborate on it in the updated discussion section.
    - Removed from the abstract:

      “We find that overall the models pre-trained on ImageNet outperform those trained on RadImageNet.”
    - Updated in the introduction:

      “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.”
- As the differences between results in Mei 2022 and this paper may be attributed to choose of model architecture and hyperparameters, it is important that the authors provide fuller sensitivity analyses.
    - Our primary focus is on representations rather than performance. Specifically, we are interested in examining the differences between source datasets, which is why we intentionally used a single model architecture and did not conduct a sensitivity analysis. There are other studies that have looked into hyperparameter effects on the transfer performance, we have added references to these studied in the discussion section.
    - Here is the updated section of the 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.”

Source domain

- Would suggest that authors take a more careful approach in attributing the results to the domain of origin of data used for pre-training
    - Removed from the discussion:

      “In our results ImageNet generally outperformed RadImageNet on seven medical target datasets. This is counterintuitive, as RadImageNet is designed to be a medical-specific source dataset and is comparable in size to ImageNet, yet our results indicate that ImageNet still offers better transfer performance. "

- Added to the discussion:

  "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."

We hope that these changes address your concerns and contribute to the overall quality of the paper.

[−]

Official Review of Paper26 by Reviewer LGwn

CHIL 2023 Conference Paper26 Reviewer LGwn

15 Mar 2023 CHIL 2023 Conference Paper26 Official Review Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers Submitted, Paper26 Authors

Summary Of Paper:
This paper looked at transfer learning using natural versus medical images for pretraining. In particular, the authors compared ImageNet versus RadImageNet. It is an interesting question to explore, whether pretraining on natural images is equivalent, superior, or worse than pretraining on a similar dataset. The authors are to be commended for their experiments exploring this interesting question. A strength of this paper is the exploration across multiple datasets.

Relevance To Healthcare: 4: High

Technical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work

Clarity: 4: High

Strengths:
Strengths include testing across multiple datasets with multiple experiments assessing their findings.

Weaknesses:
The authors state in the discussion, "In our results ImageNet generally outperformed RadImageNet on seven medical target datasets. This is counterintuitive, as RadImageNet is designed to be a medical-specific source dataset and is comparable in size to ImageNet, yet our results indicate that ImageNet still offers better transfer performance." Looking at Figure 6, it seems like there is overlap between STDs/means, and no statistical test to support the claim that ImageNet outperformed RadImageNet. The authors should include some basis to strengthen their claims regarding the superior performance of ImageNet over RadImageNet.

Additionally the performances of ImageNet did seem higher without STD overlap on ISIC and Pcam-small. This is actually not counterintuitive because radiology images are vastly different than skin lesion or pathology images, which likely are more similar to natural images. RadImageNet is a medical-specific source for radiology, and images of pathology and skin arer vastly different.

Suggestions/Questions For Authors:
Could the authors comment on the poor performance on the Thyroid and mammogram datsets? A sample size issue?

The authors state, "early layers of ImageNet and RadImageNet are as similar as two randomly initialized layers". This is a somewhat surprising result – could it have anything to do with the lack of color and texture in radiology imaging?

The authors state, "This is expected, as natural images often contain straight lines and edges that are absent in medical images, resulting in less prominent edges in the filters learned from medical images." It's not clear to me whether this statement is broadly applicable to all medical images – light photography of skin lesions and pathology images would appear to have edges. But also, X-rays definitely have edges (bones). The authors should either temper this claim or provide additional evidence for it.

Overall Recommendation: 4: Accept - Paper is in the top 50% of all submissions. Reviewer would recommend it to a colleague.

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

[−] Authors' Response

CHIL 2023 Conference Paper26 Authors Dovile Juodelyte (/profile?id=~Dovile_Juodelyte1) (privately revealed to you)

27 Mar 2023 (modified: 29 Mar 2023) CHIL 2023 Conference Paper26 Official Comment Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers Submitted, Paper26 Authors

Comment:

We would like to sincerely thank the reviewer for their time and effort in reviewing our paper, as well as the valuable comments and suggestions that have been very helpful in improving the quality of our work. Below we address the points listed in the weaknesses and suggestions/questions:

ImageNet outperforming RadImageNet

- We recognize that general claims regarding the relative performance of pre-training with ImageNet versus RadImageNet are outside the scope of this study. Our goal in including these comparisons was to provide additional empirical data on the performance of RadImageNet, which is a very new dataset. We have rewritten the part of the discussion “In our results ImageNet generally outperformed RadImageNet on seven medical target datasets…”, to better reflect this:

  “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.”
- RadImageNet's lower performance on ISIC and Pcam-small compared to ImageNet can indeed be attributed to the grey-scale/radiology images used in RadImageNet. This is acknowledged in the results section:

  “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.”

[−] Authors' Response cont.

CHIL 2023 Conference Paper26 Authors Dovile Juodelyte (/profile?id=~Dovile_Juodelyte1) (privately revealed to you)

27 Mar 2023 CHIL 2023 Conference Paper26 Official Comment Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers Submitted, Paper26 Authors

Comment:

Questions

- Could the authors comment on the poor performance on the Thyroid and mammogram datsets? A sample size issue?
  - The poor performance of Thyroid dataset may partially be attributed to the difficulty of the ultrasound nodule detection. The dataset comprises various diagnoses and cases including thyroiditis, goiter, nodules, and cancer. The limited dataset size may not fully capture the complexity of these diagnoses. Mei et al. (2022) also report lower AUC for this dataset compared to other datasets. It is worth mentioning that we did not perform any hyperparameter tuning except for increasing image size in cases where the models performed exceptionally poorly. Increasing image size from 112x112 to 224x224 improved performance on this dataset. We were not able to increase the image size more due to memory constraints.

  - Low image resolution is likely the primary reason for the poor performance on mammograms. We significantly reduced the image sizes ranging from 1,846 × 4,006 to 5,431 × 6,871 pixels to 224x224. As in the case of thyroid dataset we increased image size from the initial 112X112 to 224x224, which greatly improved performance. Furthermore, low quality of the images in the dataset, which are scanned mammograms, may have also contributed to poor performance. Upon inspection, we found that some of the images were only partially scanned, with the lower section of the image being black.
- The authors state, "early layers of ImageNet and RadImageNet are as similar as two randomly initialized layers". This is a somewhat surprising result – could it have anything to do with the lack of color and texture in radiology imaging?
  - We would argue that color does not have significant impact on layer similarity because convolutional layers learned from grayscale inputs produce kernel channels that are nearly identical, with only slight variations in weights due to random initialization prior to pre-training. However, RGB channels are highly correlated too, resulting in convolutional layers with correlated kernel channels that may be shifted or inverted, these are the types of transformations that CCA should be invariant to. We have attempted to empirically verify this by simulating random filters that mimic the behavior of correlated channels in RadImageNet (by adding noise) and ImageNet (by adding noise and transformations), which had no impact on CCA similarity. Texture differences on the other hand, we would argue, are likely the primary driving factor behind the difference in CCA similarity.
- The authors state, "This is expected, as natural images often contain straight lines and edges that are absent in medical images, resulting in less prominent edges in the filters learned from medical images." It's not clear to me whether this statement is broadly applicable to all medical images – light photography of skin lesions and pathology images would appear to have edges. But also, X-rays definitely have edges (bones). The authors should either temper this claim or provide additional evidence for it.
  - This has been rephrased to:

    "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."

We hope that these changes and comments address your concerns and contribute to the overall quality of the paper.

[−] **Authors' Response**

[Deleted] Dovile Juodelyte (/profile?id=~Dovile_Juodelyte1) (privately revealed to you)

CHIL 2023 Conference Paper26 Official Comment Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers Submitted, Paper26 Authors

[−] Official Review of Paper26 by Reviewer BrG6

CHIL 2023 Conference Paper26 Reviewer BrG6

14 Mar 2023 (modified: 15 Mar 2023) CHIL 2023 Conference Paper26 Official Review Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers Submitted, Paper26 Authors

Summary Of Paper:

The paper observes that ImageNet pre-training outperforms RadImageNet in the majority of medical datasets. It further investigates the learned intermediate representations of the models pre-trained on ImageNet and RadImageNet using Canonical Correlation Analysis. The result shows that although models pre-trained on ImageNet and RadImageNet learn different intermediate representations, they still produce similar predictions after fine-tuning.

Relevance To Healthcare: 4: High

Technical Novelty And Significance: 2: The contributions are only marginally significant or novel

Clarity: 3: Medium

Strengths:

1. The paper conducts the experiment on seven different datasets to prove their findings.

2. The methods of CCA and prediction similarity are clear and should be reproducible.

Weaknesses:

1. The data preprocessing is not clear. How do you normalize and preprocess the seven different datasets?
2. Some experiments need to be done. In Fig.3, how about the result of before versus after fine-tuning and ImageNet versus RadImageNet on other datasets in addition to Knee?
3. It would be better to report the classification score like accuracy and F1 via ImageNet and RadImageNet on seven different datasets.

Suggestions/Questions For Authors:
Clearly describe the preprocessing method on seven different datasets. Report the classification score like accuracy and F1 via ImageNet and RadImageNet.

Overall Recommendation: 3: Marginal Accept - I tend to vote for accepting this submission, but rejecting it would not be that bad.

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

[−]
Authors' Response

CHIL 2023 Conference Paper26 Authors Dovile Juodelyte (/profile?id=~Dovile_Juodelyte1) (privately revealed to you)

27 Mar 2023 CHIL 2023 Conference Paper26 Official Comment Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers Submitted, Paper26 Authors

Comment:
We would like to sincerely thank the reviewer for their time and effort in reviewing our paper, as well as the valuable comments and suggestions that have been very helpful in improving the quality of our work. Below we address the points listed in the weaknesses and suggestions/questions:

- Some experiments need to be done. In Fig.3, how about the result of before versus after fine-tuning and ImageNet versus RadImageNet on other datasets in addition to Knee?
    - Due to space constraints we did not include results on other targets in the main text. Nevertheless, we observed similar patterns for the other target datasets. We have now added an appendix with results on other targets for reference. These results are also available on the Github repository.
- Clearly describe the preprocessing method on seven different datasets.
    - This has been included in the 4.1 section Datasets (in red in the updated manuscript):

      "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"
- Report the classification score like accuracy and F1 via ImageNet and RadImageNet.
    - We would like to kindly mention that it is a common practice in medical imaging to report AUC, which is also a recommended metric to use (Reinke et al., 2021). While we acknowledge that the classification performance of pre-training with ImageNet versus RadImageNet is outside the scope of this study, we included these comparisons to provide additional empirical data on the performance of RadImageNet, which is a very new dataset. Our decision to report in this way was based on previous reporting standards. We would like to emphasize that the choice of classification metric we used does not have any impact on the results of the representations learned by the neural networks.

We hope that these changes address your concerns and contribute to the overall quality of the paper.

[−]
Official Review of Paper26 by Reviewer xCCU

CHIL 2023 Conference Paper26 Reviewer xCCU

07 Mar 2023 CHIL 2023 Conference Paper26 Official Review Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers Submitted, Paper26 Authors

Summary Of Paper:
In this paper, the authors explore the differences in the representations of two pre-trained models on medical images. Specifically, they compare the models pre-trained on ImageNet and RadImageNet for medical image classification tasks. They analyzed the variability of the two pre-trained models in different aspects (e.g., performance, intermediate representations, model similarity, etc.).

Relevance To Healthcare: 5: Very High

Technical Novelty And Significance: 2: The contributions are only marginally significant or novel

Clarity: 3: Medium

Strengths:

1. The paper is well-written and easy to understand. The motivation of the study is clear, and the problem is interesting for the medical imaging community.
2. Empirical findings are insightful. The authors demonstrate their findings from different perspectives: performance in the downstream medical imaging classification tasks, comparing the intermediate representations and model similarity.
3. The authors provide implementation details to facilitate the reproducibility of their study and have released their code.

Weaknesses:

1. The novelty of this work is limited. This paper delivers a good analysis (case study), but brings less innovation.
2. The main contribution of this paper is to compare the methods pre-trained on ImageNet and RadImageNet for the medical image classification tasks, which lacks the methodology contribution.
3. The findings in this paper are based on empirical findings and are missing theoretical support.

Suggestions/Questions For Authors:
The authors only compare the performance of methods pre-trained on ImageNet and RadImageNet, which limits the findings to be valuable only for these two datasets. Does the model pre-trained on ImageNet (natural images) also outperform other models pre-trained on medical images? The authors are suggested to pre-train models on different natural image datasets and medical image datasets and compare their initialization and downstream task performance.

Overall Recommendation: 2: Marginal Reject - I tend to vote for rejecting this submission, but accepting it would not be that bad.

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

[−]
Authors' Response

CHIL 2023 Conference Paper26 Authors Dovile Juodelyte (/profile?id=~Dovile_Juodelyte1) (privately revealed to you)

27 Mar 2023 (modified: 28 Mar 2023) CHIL 2023 Conference Paper26 Official Comment Readers: Program Chairs, Paper26 Senior Area Chairs, Paper26 Area Chairs, Paper26 Reviewers Submitted, Paper26 Authors

Comment:
We would like to sincerely thank the reviewer for their time and effort in reviewing our paper, as well as the valuable comments and suggestion. Below we address the points listed in the weaknesses related to novelty/methodological contribution/theoretical support and provide clarifications of our approach.

Novelty

1. The novelty of this work is limited. This paper delivers a good analysis (case study), but brings less innovation.
2. The main contribution of this paper is to compare the methods pre-trained on ImageNet and RadImageNet for the medical image classification tasks, which lacks the methodology contribution.
3. The findings in this paper are based on empirical findings and are missing theoretical support.
    - Although our study may not have prioritized novel or methodological contributions, we would argue that our work aligns well with the Applications and Practice track. As outlined in the Call for Papers, this track seeks to showcase works that apply robust methods, models, or practices to identify, characterize, audit, evaluate, or benchmark systems. Our study utilized a robust method, CCA, to audit and characterize intermediate representations acquired by convolutional layers, in order to shed light on transfer learning, a crucial, yet poorly understood, technique in medical imaging. While

there is still much work to be done, we have managed to uncover some new insights that we believe would be beneficial to the community.

Limitation of the findings to two source datasets

- Does the model pre-trained on ImageNet (natural images) also outperform other models pre-trained on medical images? The authors are suggested to pre-train models on different natural image datasets and medical image datasets and compare their initialization and downstream task performance.
    - Our study focused on ImageNet and RadImageNet because ImageNet is widely used for pre-training in medical imaging, as it generally offers the best transfer performance. Previous studies have demonstrated that other medical or natural source datasets do not match ImageNet performance (Brandt et al., 2021, Wen et al., 2021). However, medical source datasets in previous studies were considerably smaller in size compared to ImageNet, the largest one being 100,000 images. RadImageNet was released as an alternative to ImageNet with a comparable size, offering a unique opportunity for comparison eliminating implications of differences in dataset size. It is worth noting that beyond the domain, there are significant differences between RadImageNet and ImageNet, such as color, number of classes, and diversity in data due to the limited number of patients in RadImageNet. It would be indeed interesting to compare our results to pre-training on Ecoset (Mehrer et al., 2021), a natural image dataset with 565 basic-level categories selected to better reflect the human perceptual and cognitive experience, we have updated discussion section to include these considerations:

      “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.”

We hope that these comments address your concerns, provide clarity regarding our approach and contribute to the overall quality of the paper