CheXtraining: A Study on Training Robust Models for Interpretation of Photos of Chest X-rays

Anfal Siddiqui
Stanford University
450 Serra Mall, Stanford, CA 94305
anfal@stanford.edu

Abstract

Recent studies have shown that current chest x-ray (CXR) interpretation models have a drop in performance when applied to photos of chest x-rays, which can significantly hinder the ability to leverage smartphone cameras and apps to get fast and accurate pathology detection from the images in areas where radiologists are not readily available. In this work, we explore a variety of different pretraining and training strategies, including fine-tuning using photos of CXRs, mixed training, and MoCo Contrastive Learning, to produce models that can accurately interpret both digital CXRs and photos of them. We found that many of these techniques, while successful at improving performance on photos of CXRs, also result in a drop in performance on digital CXRs relative to our baseline, indicating there is a large distributional shift between the two types of images and that a single model for both will not achieve peak performance. Additionally, fine-tuning using photos of CXRs and MoCo were found to provide inconsistent benefits. Our best performance came by training a model using a mixture of both types of CXRs to produce a robust encoder and then separately fine-tuning two separate classifier layers for it using training data of each type, with performance of the two resulting models consistently surpassing baseline performance on both types of CXRs. We also performed a fixed-budget training study using different proportions of each type of CXR that reaffirmed the domain mismatch between the two types, but also illustrated that a model must be trained on some amount of digital CXRs in order to achieve its best performance on photos of them.

1. Introduction

Chest x-rays (CXR) are the world's most common imaging examination and are used for screening numerous life-threatening pathologies. Recent deep learning models for chest x-ray interpretation that detect the presence of certain pathologies have achieved or even exceeded radiologist performance [3]. However, real-life deployments of these models require proper integration into all healthcare settings. Due to a lack of local radiologists, many healthcare setups provide photos of x-rays (rather than digital ones) to remote practitioners [8]. Moreover, with the ubiquity of smartphones, being able to develop apps that can consume photographs of CXRs and provide immediate diagnostic support from a model is very tantalizing and can have a major impact on quickly getting patients proper care.

However, photographs of CXRs present new challenges to models as they introduce visual artifacts such as glare, mixed lightning conditions, blur, and more that are not seen in the digital CXR datasets said models are traditionally trained with [11]. Recent work has shown that the best CXR interpretation models have a statistically significant drop in performance when applied to such photos [10], demonstrating the need for further study on how to build robust models that generalize well to this new domain.

To that end, our work focuses on investigating which pretraining and training strategies are most beneficial at building a robust chest x-ray interpretation model that can accurately interpret both standard/pure digital CXRs and photographs of them. Additionally, in order to provide insight into which type of CXR yields the most benefit when training models and to inform further data collection efforts, we also perform a fixed-budget training study to determine how different proportions of images of the two types affect the quality and performance of the models produced.

Explicitly, all of the models produced in our work take in as input either a chest x-ray or a photo of one and use a convolutional neural network to output the probabilities of it showing each of five particular pathologies: Pleural Effusion, Edema, Atelectasis, Consolidation, and Cardiomegaly. We chose these pathologies because they are well represented in the datasets we utilized and have served as the primary focus in CXR deep learning literature due to their prevalence and clinical significance [3].

2. Related Work

2.1. Digital CXR Interpretation Models

CXR interpretation has been studied extensively using deep learning techniques in recent years. One prominent study is that of Irvin et al., in which they produced a model using a DenseNet-121 architecture that exceeded radiologist performance on Cardiomegaly, Edema, and Pleural Effusion [3]. Moreover, the authors also released a large dataset of labeled CXRs, called CheXpert, and have hosted an on-going competition centered around it that has spurred development in this space. As measured by this competition's leaderboard, the current state-of-the-art model was produced by Yuan et al. They forwent traditional cross entropy loss during training and instead used a novel loss function to maximize AUC, the default metric used in medical classification tasks [15]. Although this clever approach produced impressive results (AUC of 0.9305), their work is deeply oriented around maximizing this singular metric and does not report how their model fares on other metrics or in real-world settings with distributional shift (such as those presented by photos of CXRs), which arguably limits its impact.

Even more recently, the work of Wang et al. developed a novel model architecture for the detection of Covid-19 from CXRs, achieving a 93.3% accuracy on their test set [14]. While this work demonstrated the benefits of strong CXR interpretation models towards global health outcomes and presented a unique architecture, it did not utilize any transfer learning methods to leverage existing advances made using CheXpert and other large datasets, which was a missed opportunity at testing the potential versatility of existing models towards a completely novel disease.

2.2. Generalizability of CXR Interpretation Models

Several recent works have examined how well current CXR interpretation models perform on new domains, including photographs of CXRs. In CheXpedition, Rajpurkar et al. evaluated some of the best performing models on CheXpert on a dataset of CXR photos, finding their performance was roughly equivalent on the two domains [9]. However, their evaluation dataset was fairly small (using samples from only 500 patients), leaving doubt on how robust these results were. Perhaps recognizing this shortcoming, the work of Phillips et al. collected a substantially larger dataset of natural and synthetic CXR photographs for deeper study, called CheXphoto [8]. Rajpurakar et al. then utilized CheXphoto to follow up on their original findings in [10] and [11], this time finding most of the same models previously studied suffered a statistically significant drop in performance (although some models still exceeded radiologist performance in this domain for certain pathologies). Despite these improved findings, none of the follow-up papers addressed how models can be made more robust to the distributional shift.

2.3. Building Models for CXR Photographs

The work of Le et al. was among one of the first to utilize CheXphoto. They used a two-step approach: (1) using one model to localize the actual radiograph in the photograph of a CXR (2) extracting the localized radiograph and feeding it into another model for interpretation [5]. Their approach yielded strong results with a mean F1 score of 0.6988. However, it has the significant downside of having to train two completely separate models for two separate tasks, which can be impractical.

Similar to our work, Kuo et al. investigated various training approaches for making models that perform well on photos of CXRs. They found success with their "recalibration" approach, where a large dataset of digital CXR images is augmented using transformations to appear as if it contains photographs of CXRs and then is used to train a new model [4]. The resultant model achieved performance on photographs of CXRs that was on par with models for pure digital CXRs. However, a major shortcoming of Kuo et al.'s work - and one that we seek to address in our own - is that it did not explore whether models can be made that perform well on both digital CXRs and photographs of them. Indeed, their "recalibrated" model does worse on every pathology for digital CXRs as compared to photographs of them.

2.4. Contrastive Learning for CXR

Prior works have explored applying the increasingly popular Momentum Contrast (MoCo) Contrastive Learning approach using CheXpert. Sowrirajan et al. found that using MoCo for pre-training provided a boost in disease diagnosis performance when compared to ImageNet initialization (although the boost diminished as the number of labeled examples increased during fine-tuning) [12], while Sriram et al. found the technique provided a boost in performance for different tasks, such as adverse event prediction [13]. The benefits of Contrastive Learning on these very distinct tasks speaks to the underlying robustness of the models it produces. Although neither of the above papers tested the method under domain shift, this robustness does indicate potential viability of the technique when applied to photographs of CXRs.

3. Datasets

We have made use of two datasets: CheXpert and CheXphoto.

3.1. CheXpert

CheXpert[1] is a dataset of pure, digital chest x-rays labeled for the presence of 14 different pathologies/observations. For each observation, a chest x-ray in the dataset is labeled as positive, negative, or uncertain, where the labels were extracted from the corresponding radiologist report in an automated fashion [3]. The public release of CheXpert is composed of a training set of 223,414 images and a validation set of 234 images 1. We have held out the official validation set as our own test set and randomly sampled 10% of the training set to be our validation set (a 201,073/22341 split).

Figure 1. A sample from CheXpert

3.2. CheXphoto

CheXphoto[2] is a dataset of natural or synthetic photos of chest x-rays from CheXpert, each following the same labeling scheme as CheXpert. The natural photos were taken using smartphone cameras in various lighting conditions. The synthetic examples were transformed to resemble photos of digital x-rays and x-ray films and were verified by physicians to still be viable for pathology detection 2. The public release of CheXphoto is composed of a training set of 32,521 images and a validation set of 702 images. As with CheXpert, we have held out the official CheXphoto validation set as our test set and used the same split to produce our training and validation sets (29,269 / 3,252).

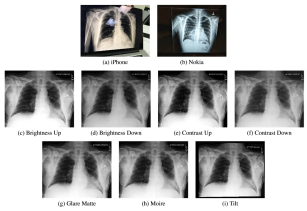

Figure 2. Samples from CheXphoto

3.3. Pre-processing and Splits

Each image was resized to 320x320 pixels, converted to grayscale, and normalized to the mean and standard deviation of the ImageNet training set. We have followed the U-Ones formulation from the CheXpert paper, where images that are labeled as "uncertain" for a particular pathology are treated as if they have the pathology [3]. Although Irvin et al. found this was not the absolute best approach, it still offered reasonably high performance and it provided us with a straightforward and consistent setup across all of our experiments.

[1]https://stanfordmlgroup.github.io/competitions/chexpert/

[2]https://stanfordmlgroup.github.io/competitions/chexphoto/

4. Methods

4.1. Model Architecture

In keeping with prior implementations of CXR interpretations models [3], where it was found to be the best architecture, all of our experiments use an underlying DenseNet-121 architecture. As detailed in Huang et al., Dense Convolutional Networks (DenseNet) connect each layer within a Dense Block to every subsequent layer within the block in a feed-forward fashion to make optimization feasible with deeper networks [2]. We replaced the final fully-connected layer with one with 5 outputs for our 5 pathologies of interest. To capture the notion that multiple pathologies may occur simultaneously in a given image, we apply a sigmoid non-linearity to the output of our model to produce the probabilities of each of the pathologies being present in the image, with 50% serving as the threshold for positivity for each condition.

4.2. Loss Function

For each experiment, we made use of binary cross-entropy loss, taken across the 5 pathologies and then averaged. Explicitly, we had:

$$
\begin{aligned}
L(x, y) &= mean(Z) \\
Z &= \{l_1, \ldots, l_5\} \\
l_n &= [y_n * \log x_n + (1 - y_n) * log(1 - x_n)]
\end{aligned}
$$

4.3. Baseline

Our baseline model was trained purely using CheXpert, with no additional augmentations or introduction of data from CheXphoto.

4.4. Training Strategies

4.4.1 Fine-tuned on CheXphoto

We fine-tuned the baseline model using the CheXphoto dataset, experimenting with freezing all but the last Dense Block and classification layer and leaving all layers unfrozen. This transfer learning-esque approach is meant to take the original learnings from the CheXpert dataset and refine them using CheXphoto to be robust against the distribution/dataset shift.

4.4.2 Mixed Training

We also tried training on a combination of the CheXphoto and CheXpert datasets, with no additional upsampling of CheXphoto. This is somewhat reminiscent of the successful "recalibrated" training from [4]. Here, the presence of both types of CXRs during the initial training may allow for the production of a model that is robust to variations in standard and photos of CXRs, while also allowing for a more efficient training process.

4.4.3 Mixed Training + Independent Classifiers

Lastly, we tried an approach that built off mixed training but tuned to each specific CXR domain. We froze all but the final classifier layer and produced two separate classifier heads by fine-tuning the model on either CheXpert or CheXphoto. We believed it may be possible that the distribution shift between the two types of CXRs was so vast that it could not be effectively bridged in a single classifier layer. With this approach, we hypothesized that we may get the benefits of a mixed-trained encoder that was robust to variations in the input CXRs coupled with a final domain-specific classification head.

4.5. MoCo Pre-training

Contrastive Learning is a form of self-supervised learning that aims to develop models that learn "good" representations of the input images by steering the model towards learning similar representations for similar images but different represents for different images[3].

MoCo Contrastive Learning approaches this goal by starting with augmenting a single image x using two different transformations to produce x_q and x_k. The former is passed through the encoder network to produce r_q and the latter goes through the momentum encoder network to produce r_k. A queue of images distinct from x, $r_{k,i}$, is maintained, and the model attempts to determine which of the images in $r_{k,i} + r_k$ corresponds with r_q, as driven by the contrastive loss function described in [1]. The gradient of the loss is backpropagated through the encoder, and the momentum encoder is given a corresponding momentum update so that its change is more gradual. Sowrirajan et al. showed that MoCo pre-training can be beneficial for CXR interpretation, producing better performance than a model initialized using ImageNet weights [12]. Moreover, it is possible that the strong encoder developed by this method may be more robust in the face of the domain shift brought by photos of CXRs than one that is trained off ImageNet weights using supervised learning.

Due to computational limits, we opted to leverage a DenseNet-121 model pre-trained by Facebook Research using MoCo rather than train one ourselves from scratch [13]. Facebook Research trained their model using both the CheXpert dataset and the MIMIC-CXR dataset - another large dataset of digital CXRs. Most of the training strategies described in the prior section were repeated using the MoCo weights. Additionally, we explored only fine-tuning the pre-trained model using CheXphoto to see if the precursory CheXpert step would even be unnecessary.

[3]https://towardsdatascience.com/understanding-contrastive-learning-d5b19fd96607

4.6. Fixed-Budget Training

To measure how different proportions of digital CXRs and photos of CXRs in the training set affect the quality of the resultant model, we ran a fixed-budget study. Using a fixed training set size of 10,000 images, we tried progressively increasing the proportion of digital CXRs in it, trained a new model from scratch using the set, and measured its performance on our validation sets.

5. Experiments

All of our experiments were conducted on single-GPU GCP VMs.

5.1. Hyperparameters

All of our models were trained with a mini-batch size of 16 and using Adam optimization with default betas for consistency across experiments. We chose these hyperparameters based on their successful use in the CheXpert paper [3], suspecting they would be generally appropriate across all the different but related training regiments.

The learning rate for our baseline, mixed training, and fixed-budget experiments was set to $1e^{-4}$, again in keeping with the CheXpert paper. A brief hyperparameter search found this rate to be too high for fine-tuning experiments and we settled on a slightly lower $1e^{-5}$ for those experiments.

Our Training Strategy experiments were nearly all trained for 3 epochs (or an additional 3 epochs if the model was fine-tuned from another one of our models), once more in keeping with the CheXpert paper. The one exception was for the Mixed Training + CheXpert Head model, which was fine-tuned from the Mixed Training model for only one epoch, due to the large size of the CheXpert training set and because only one layer was unfrozen. All Fixed-Budget models were trained for one epoch for computational efficiency.

All none-MoCo models were initialized using weights of a DenseNet-121 pre-trained on ImageNet.

5.2. Metrics

In keeping with standard practice in medical classification literature, we use precision, recall, F1 score, accuracy,

and Area under the ROC Curve (AUC)[4]. Briefly, the higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

We also made use of the Matthew's Correlation Coefficient (MCC)[5], a statistical rate that ranges between a low of -1 and a high of 1, achieving high scores only if the prediction obtained strong results in all four confusion matrix categories (true positives, false negatives, true negatives, and false positives). MCC is viewed as a balanced measure as it is proportional to the number of positive and negative elements in the dataset - which is helpful in these often skewed medical datasets. In keeping with more recent medical and CXR literature [10] [11], MCC is the primary metric we focus on in our discussion. Data points using accuracy, precision, recall, and F1 score can be found in the appendix.

6. Results and Discussion

6.1. Training Strategies

Results on the validation sets for all of our training strategy experiments can be seen in 1.

6.1.1 Fine-tuned on CheXphoto

Fine-tuning the baseline model with CheXphoto causes the most dramatic improvement on overall CheXphoto performance (+0.323 on MCC relative to baseline). However, it comes at the expense of performance on CheXpert, which sees a small drop on both AUC/MCC compared to the baseline. Moreover, there is a sizable gap in the model's performance between standard and photo CXRs (particularly for Consolidation), suggesting it may be "overfit" to the photo domain and losing some knowledge gained from its training on CheXpert. Freezing layers on the baseline model before fine-tuning to try and avoid this loss ends up making performance worse on both CheXpert and CheXphoto compared to the unfrozen version, suggesting the fine-tuning in the frozen version ends up more disruptive to the classifier layer while also being handicapped from adapting the encoder portion of the model to the photos domain. Still, the large increase in MCC on photos for the unfrozen version compared to its relatively small drop on standard CXR shows the viability of this method.

6.1.2 Mixed Training

Mixed Training led to a significantly smaller gap in overall MCC performance between standard and photos of CXRs than fine-tuning, as well as a smaller drop on standard CXRs compared to the baseline (on AUC it even exceeded the baseline on standard CXRs). However, the gains on photo performance relative to the baseline were much more modest compared to the fine-tuning approach. Despite this, the still significant improvement on photos suggests this training regime does produce an encoder that is more robust to the artifacts introduced in such images, thought still somewhat at the expense of performance on standard CXRs.

6.1.3 Mixed Training + Independent Classifiers

Fine-tuning the classifier layers/heads of the Mixed-Training model separately for standard CXRs and photos produced the highest scoring model for standard CXRs and the second highest scoring for photos. The success of both of these independent classifier layers reaffirms our finding that the encoder portion of the model is made sufficiently robust to visual artifacts and general variations through the mixed training regime. However, the boost in performance on standard CXRs that was not seen in either regular Mixed Training or Fine-tuning seems to suggest that the domains between the two types of images are different enough that sharing a classifier layer is detrimental; indeed, the drop in performance on CheXpert by the standard Mixed-Training model suggests even the presence of CXR photos during training is enough to disrupt the classifier's performance on standard ones to some extent. This theory is strengthened by the larger drop seen with the fine-tuning strategy, which, by virtue of the photos-only training, appeared to deviate the classifier layer further away from properly handling standard CXRs. We revisit these concepts in subsequent sections.

6.1.4 MoCo Pre-Training

Our experiment with just fine-tuning the MoCo pre-trained model using CheXphoto had worse performance than the baseline on both types of images, indicating that even with pre-training, there is still a need for developing the encoder further using a supervised regiment with a large labeled dataset of digital CXRs like CheXpert. Our remaining experiments with MoCo closely mirrored the trends seen with the training strategies when using ImageNet initialization, except with reduced performance on MCC and AUC across the board. While its surprising the domain-specific pre-training was not helpful, this somewhat tracks with previous results seen with MoCo and CXRs, which found that the performance gain over ImageNet initialization rapidly diminished as more labeled data was used during supervised training [12]. However, actually seeing a performance drop relative to ImageNet initialization is unusual, but may also be an artifact of the Facebook Research weights or that the hyperparameters chosen were not suitable for MoCo models. More experimentation using other MoCo pre-trained models would be needed. One noteworthy result is that of the MoCo Mixed Training model, which had remarkably

[4]https://en.wikipedia.org/wiki/Receiver_operating_characteristic
[5]https://en.wikipedia.org/wiki/Matthews_correlation_coefficient

		Pleural Effusion		Edema		Atelectasis		Consolidation		Cardiomegaly		Average	
Training Strategy	Comparison	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC
Baseline	Standard	0.578	0.874	0.516	0.858	0.216	0.72	0.049	0.69	0.46	0.851	0.364	0.799
	Photos	0.546	0.869	0.473	0.85	0.141	0.715	0.002	0.669	0.345	0.84	0.301	0.789
	Standard - Photos	0.032	0.005	0.043	0.008	0.075	0.005	0.047	0.021	0.115	0.011	0.063	0.01
Baseline Fine-tuned on CheXphoto	Standard	0.503	0.841	0.446	0.822	0.227	0.682	0.154	0.651	0.381	0.798	0.342	0.759
	Photos	0.721	0.939	0.694	0.938	0.513	0.856	0.517	0.867	0.677	0.944	0.624	0.909
	Standard - Photos	-0.218	-0.098	-0.248	-0.116	-0.286	-0.174	-0.363	-0.216	-0.296	-0.146	-0.282	-0.15
Baseline Fine-tuned on CheXphoto (Frozen)	Standard	0.521	0.86	0.482	0.837	0.202	0.703	0.049	0.68	0.325	0.827	0.316	0.781
	Photos	0.617	0.895	0.558	0.876	0.27	0.747	0.091	0.716	0.454	0.876	0.398	0.822
	Standard - Photos	-0.096	-0.035	-0.076	-0.039	-0.068	-0.044	-0.042	-0.036	-0.129	-0.049	-0.082	-0.041
Mixed Training	Standard	0.596	0.878	0.497	0.858	0.236	0.721	0.009	0.7	0.437	0.848	0.355	0.801
	Photos	0.63	0.897	0.536	0.883	0.316	0.755	0.074	0.702	0.466	0.881	0.404	0.824
	Standard - Photos	-0.034	-0.019	-0.039	-0.025	-0.08	-0.034	-0.065	-0.002	-0.029	-0.033	-0.049	-0.023
MoCo Fine-tuned on CheXphoto Only	Standard	0.521	0.835	0.368	0.809	0.092	0.666	0	0.657	0.241	0.796	0.244	0.753
	Photos	0.544	0.845	0.419	0.831	0.098	0.7	0	0.674	0.261	0.825	0.264	0.775
	Standard - Photos	-0.023	-0.01	-0.051	-0.022	-0.006	-0.034	0	-0.017	-0.02	-0.029	-0.02	-0.022
MoCo Trained on CheXpert	Standard	0.568	0.872	0.481	0.842	0.16	0.701	0	0.686	0.388	0.841	0.319	0.788
	Photos	0.53	0.855	0.428	0.82	0.102	0.695	0	0.641	0.369	0.817	0.286	0.766
	Standard - Photos	0.038	0.017	0.053	0.022	0.058	0.006	0	0.045	0.019	0.024	0.033	0.022
MoCo Trained on CheXpert / Fine-tuned CheXphoto	Standard	0.53	0.855	0.428	0.82	0.102	0.695	0	0.641	0.369	0.817	0.286	0.766
	Photos	0.623	0.89	0.554	0.872	0.3	0.739	0.091	0.705	0.53	0.87	0.419	0.815
	Standard - Photos	-0.093	-0.035	-0.126	-0.052	-0.198	-0.044	-0.091	-0.064	-0.161	-0.053	-0.133	-0.049
MoCo Mixed Training	Standard	0.581	0.872	0.484	0.848	0.187	0.703	-0.005	0.685	0.415	0.841	0.333	0.79
	Photos	0.59	0.877	0.49	0.851	0.169	0.716	0	0.675	0.398	0.846	0.33	0.793
	Standard - Photos	-0.009	-0.005	-0.006	-0.003	0.018	-0.013	-0.005	0.01	0.017	-0.005	0.003	-0.003
Mixed Training + CheXpert Head	Standard	0.602	0.881	0.513	0.86	0.24	0.725	0.101	0.703	0.454	0.851	0.382	0.804
Mixed Training + CheXphoto Head	Photos	0.65	0.903	0.569	0.88	0.304	0.758	0.093	0.709	0.509	0.884	0.425	0.827
Mixed Training + Heads	Standard - Photos	-0.048	-0.022	-0.056	-0.02	-0.064	-0.033	0.008	-0.006	-0.055	-0.033	-0.043	-0.023

Table 1. Results of experiments on validation sets. Standard refers to digital CXRs from CheXpert and Photos refers to photographs of CXRs from CheXphoto. A more extensive reporting of results (including precision, recall, F1, etc.) can be found in the Appendix.

		Pleural Effusion		Edema		Atelectasis		Consolidation		Cardiomegaly		Average	
Training Strategy	Type	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC
Baseline	Standard	0.664	0.937	0.665	0.923	0.345	0.81	0	0.85	0.409	0.834	0.417	0.871
	Photos	0.575	0.916	0.473	0.887	0.211	0.79	0	0.847	0.288	0.77	0.309	0.842
Baseline Fine-tuned on CheXphoto	Standard	0.542	0.871	0.54	0.884	0.328	0.768	0.251	0.739	0.169	0.804	0.366	0.813
	Photos	0.551	0.852	0.472	0.883	0.243	0.737	0.149	0.706	0.342	0.764	0.351	0.788
MoCo Trained on CheXpert / Fine-tuned CheXphoto	Standard	0.62	0.881	0.535	0.902	0.456	0.831	0.162	0.874	0.31	0.743	0.416	0.846
	Photos	0.604	0.883	0.507	0.867	0.391	0.811	0.149	0.848	0.286	0.755	0.387	0.833
Mixed Training + CheXpert Head	Standard	0.695	0.936	0.625	0.935	0.342	0.824	0.172	0.894	0.401	0.841	0.447	0.886
Mixed Training + CheXphoto Head	Photos	0.605	0.918	0.519	0.913	0.395	0.823	0.149	0.861	0.368	0.821	0.407	0.867

Table 2. Table 2. Results of a few of the best models on the test set. See 1 for an explanation on terminology

consistent results between standard and photos of CXRs across all pathologies, even if they were not as high as those seen with the ImageNet models.

6.1.5 Evaluation on Test Set

2 shows the results of some of our best models (Baseline Fine-tuned on CheXphoto and the independent classifier head models) as well as the best performing MoCo model on our test sets. Some of the results are surprising and diverge from those seen on the validation set. Firstly, the Fine-tuned model's commanding boost on CheXphoto performance collapses on the test set, where it does only moderately above the baseline on overall MCC but actually does worse on AUC. Moreover, it continues to underperform the baseline on CheXpert / standard CXR. Secondly, the MoCo fine-tuned model has an unexpected resurgence: it performs close to the baseline on standard CXRs and exceeds its ImageNet counterpart overall on photos and on every pathology sans Cardiomegaly. Lastly, the Independent Classifier/Heads models now are the best performing models on both types of images. However, the CheXpert head's success on standard images relative to the baseline is less pronounced when looking at individual pathologies. In fact, it does worse than the baseline on standard CXR's for Edema, Atelectasis, and Cardiomegaly, primarily pulling ahead overall only because the baseline does particularly poorly on Consolidation. Still, the relatively consistent performance of this model on both the validation and test sets shows it has not overfit to its training set.

There are several takeaways from the results. Fine-tuning using CheXphoto and possibly even MoCo pre-training overall provide very inconsistent results, even when the models are evaluated on inputs from the same distribution they were fine-tuned on. Additionally, the success of the independent heads models seems to reaffirm that building a singular model for both types of CXRs may not be the best approach: there appears to be too large a distributional shift between the two domains such that a single classifier layer cannot effectively operate on both. Conversely, mixed training using images of both types is consistently beneficial at producing a more robust encoder that can handle both types of inputs.

Examining individual pathologies for both standard and

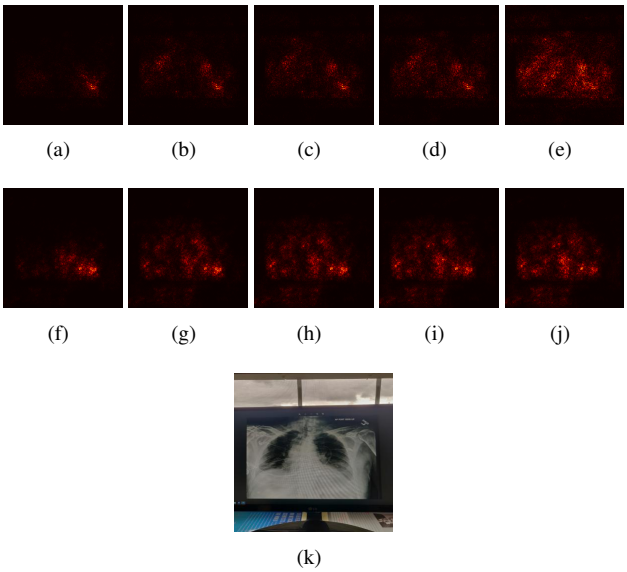

Figure 3. (a-e) Mixed Training + CheXphoto Head; (f-g) MoCo Trained on CheXpert / Fine-tuned CheXphoto; (k) actual photo from CheXphoto; Order of saliency maps is Pleural Effusion, Edema, Atelectasis, Consolidation, Cardiomegaly. All 5 pathologies are present in this image

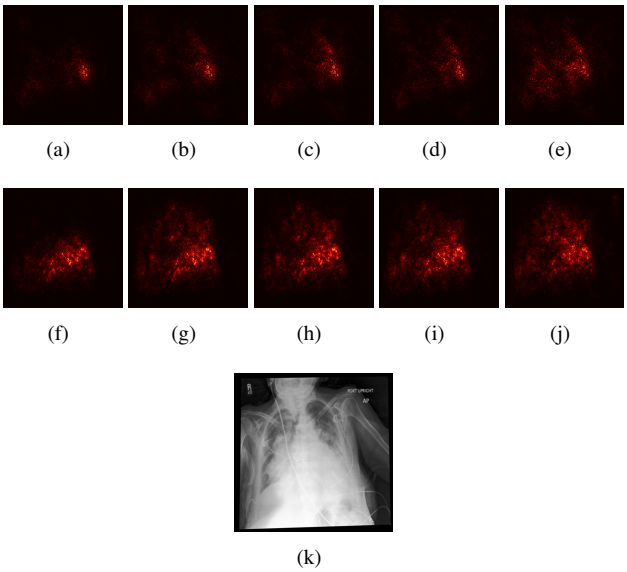

Figure 4. (a-e) Mixed Training + CheXphoto Head; (f-g) MoCo Trained on CheXpert / Fine-tuned CheXphoto; (k) actual photo from CheXphoto; Order of saliency maps is Pleural Effusion, Edema, Atelectasis, Consolidation, Cardiomegaly. All 5 pathologies are present in this image

photos of CXRs, we see that even our best model is inconsistent in its performance. Roughly, we see performance worsening as we go from Pleural Effusion, to Edema, to Atelectasis, to Cardiomegaly, and to Consolidation, with a rapid drop-off near the end. The CheXternal paper measured radiologist performance (using MCC) on CheXphoto in order of best to worst as Pleural Effusion, to Cardiomegaly, to Atelectasis, to Edema, to Consolidation [11]. Although our model does not perfectly track with human-level performance variation, it is clear that some of these pathologies are more difficult to detect than others, which reflects in the inconsistent performance of our models.

6.1.6 Qualitative Analysis

To further understand the performance of our two top performing models on the test set (Model A: Mixed Training + CheXphoto Head and Model B: MoCo Trained on CheXpert / Fine-tuned CheXphoto), we derived saliency maps from the scores the models produced for each pathology for a few photos of CXRs.

3 shows saliency maps for both models on a natural photo of a CXR that exhibits all 5 pathologies and has a prominent distractor in the form of a monitor. Both models performed poorly on this image: Model A only correctly predicted one pathology while Model B missed every one. As can be seen from the saliency maps, Model A does a reasonable job focusing on just the radiograph in the image, but is largely unsuccessful in detecting pathologies from the isolated CXR. Only for Pleural Effusion, which is characterized by liquid build-up in the lungs and manifests clearly in CXRs in a prominent white spot in the corner of the image, does Model A make an accurate prediction. For the more difficult pathologies, it is unable to detect anything in this image. The MoCo-derived Model B gets particularly distracted by the monitor stand, especially in (h-j). This speaks to a larger trend seen in both models' performances, where in images that have not only visual artifacts, but other objects like the monitor present, both seem to struggle to properly encode the actual radiograph and make detections.

4 shows a synthetic photograph of a CXR that exhibits all 5 pathologies. In this photograph, the overall weaker Moco-derived Model B correctly detects all 5 pathologies, while Model A fails to detect Atelectasis and Consolidation. This CXR exhibits a recurring difference between these two models: Model B, either due to its separate MoCo initialization or its different training strategy, tends to have a much wider area of the image it examines when making a prediction compared to Model A. In this particular example, that wider area of focus allows Model B to detect the extremely bright white region in the image that contribute to the Atelectasis and Consolidation diagnosis, something Model A's more restricted view results in it missing.

These two examples speak to the competing requirements that come to the forefront for models for interpretation of photos of CXRs. Models often derive much benefit from having a restricted receptive field when making predictions so they do not get distracted by visual artifacts like

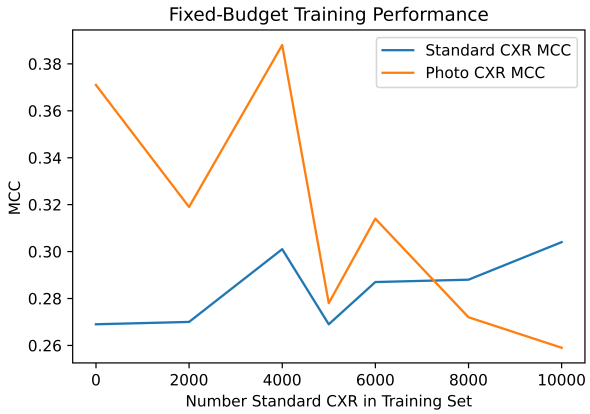

Figure 5. Fixed-Budget Training results on validation sets. A total of 10,000 images were used when training the models (all none-standard images in the set were photos of CXRs). MCC is average MCC across pathologies

the monitor in 3. However, as shown in 4, when visual artifacts are minimal or not present, this reduced receptive field can result in important details needed for proper classification being missed.

6.2. Fixed-Budget Training

5 shows the results from our Fixed-Budget training study. Gradually increasing the number of standard CXRs used in the training set does not cause a linear pattern of change in performance of the resultant model on either type of CXR. However, there is a sweet-spot when 4000 standard : 6000 photo are used in the training set, resulting in peak performance on CheXphoto's validation set and the 2nd best performance on CheXpert's. Interestingly, adding either more or less standard CXRs to the training set at that sweet-spot causes a significant drop-off in performance at both types of images.

The inconsistency in the results reaffirm some previous findings from our Training Strategies experiments and offer new takeaways. First, there is not a clear winner on which of standard or photos of CXRs are better for driving model performance. Consistent with our prior results, training with primarily photos of CXRs leads to a classifier that is maladapted for standard ones, and vice versa. In other words, there is an evident domain mismatch. As such, where data should be collected in future efforts depends on the primary purpose of the model being created. Second, as evidenced by the observed sweet-spot, training with some pure CXRs is helpful, if not necessary, for peak model performance on photos of CXRs. We suspect these pure images allow the model to have an easier time learning what the key characteristics of each pathology are, without having to work around visual artifacts in the images. Lastly, given the sweet-spot model also achieved a high MCC on standard CXRs (although not the highest), this further reinforces our finding from our mixed training experiment that training with both types can yield a potentially more robust encoder.

7. Conclusion and Future Directions

In this work, we investigated which training strategies work best for building a model for the detection of pathologies in both digital CXRs and photographs of them. We trained using a variety of strategies, including Mixed Training, Fine-tuning, and MoCo Contrastive Learning. We found that many of these strategies result in models that, while performing significantly better on photos of CXRs, perform worse than a baseline model on standard CXRs. This indicates there is a large enough distribution shift between the two types of images that a single model with a single classifier layer is unable to serve both types at its peak possible performance. We additionally found that models that were first trained on standard CXRs and fine-tuned using photos of them provided inconsistent performance, both when initialized with MoCo or ImageNet weights. Our most consistent and highest performing models came from a mixed training strategy using both types of images, with independent classifier layers that were further fine-tuned using either of the two image types. As supported by these independent classifiers models and our fixed-budget training experiments, mixed training produces a robust encoder that can perform well on both types of CXR images. Training a separate classifier layer for the particular type of image of interest on top of this encoder can then achieve peak performance. Additionally, our fixed-budget experiments found that neither type of image is strictly better at producing higher performing models for both types of CXRs, due to the domain mismatch, but that training with both types of images is needed for peak performance on photographs of CXRs.

Steps for future work include repeating our Mixed Training + Independent Classifiers experiments using MoCo-derived weights instead of ImageNet and seeing how the resultant model performs. Furthermore, because of the inconsistent results we received using Facebook Research's MoCo weights, it would be valuable to run the Contrastive Learning process ourselves and generate weights from scratch or to use the weights recently released by the Stanford Machine Learning Group [12] and see if our results are replicated. Finally, given the knowledge this work has gleaned about the ideal training strategies for peak performance on photos of CXRs, a valuable next step would be to try building a state-of-the-art model using these findings, coupling the mixed training + independent classifier strategy with a more aggressive hyperparameter search and ensembling to produce a potentially very strong model.

8. Appendices

All code used in this work is available here: https://github.com/anfals/chexphoto

Tables 3 and 4 below show more extensive data from our training strategy experiments.

9. Contributions Acknowledgements

This work was produced solely by Anfal Siddiqui, who did all parts of it.

Much of the initial skeleton for our implementation was based off the code examples Stanford's CS 230 course graciously makes available here: https://github.com/cs230-stanford/cs230-code-examples. Saliency Map creation code was heavily inspired by Stanford's CS 231N material (https://cs231n.github.io/assignments2021/assignment3/).

MoCo weights were retrieved from Facebook Research, who have linked to them here: https://github.com/facebookresearch/CovidPrognosis. We particularly used the lr_0.01 checkpoint.

All implementation was done in native PyTorch [6], with Scikit-learn used for calculating metrics [7].

References

[1] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. 4

[2] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. 3

[3] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. 1, 2, 3, 4

[4] Po-Chih Kuo, Cheng Che Tsai, Diego M López, Alexandros Karargyris, Tom J Pollard, Alistair EW Johnson, and Leo Anthony Celi. Recalibration of deep learning models for abnormality detection in smartphone-captured chest radiograph. *NPJ digital medicine*, 4(1):1–10, 2021. 2, 4

[5] Hieu X. Le, Phuong D. Nguyen, Thang H. Nguyen, Khanh N. Q. Le, and Thanh T. Nguyen. Interpretation of smartphone-captured radiographs utilizing a deep learning-based approach, 2020. 2

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 9

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 9

[8] Nick A. Phillips, Pranav Rajpurkar, Mark Sabini, Rayan Krishnan, Sharon Zhou, Anuj Pareek, Nguyet Minh Phu, Chris Wang, Mudit Jain, Nguyen Duong Du, Steven QH Truong, Andrew Y. Ng, and Matthew P. Lungren. Chexphoto: 10,000+ photos and transformations of chest x-rays for benchmarking deep learning robustness, 2020. 1, 2

[9] Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Phil Chen, Amirhossein Kiani, Jeremy Irvin, Andrew Y. Ng, and Matthew P. Lungren. Chexpedition: Investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting, 2020. 2

[10] Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Jeremy Irvin, Andrew Y. Ng, and Matthew Lungren. Chexphotogenic: Generalization of deep learning models for chest x-ray interpretation to photos of chest x-rays, 2020. 1, 2, 5

[11] Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexternal. *Proceedings of the Conference on Health, Inference, and Learning*, Apr 2021. 1, 2, 5, 7

[12] Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar. Moco-cxr: Moco pretraining improves representation and transferability of chest x-ray models, 2021. 2, 4, 5, 8

[13] Anuroop Sriram, Matthew Muckley, Koustuv Sinha, Farah Shamout, Joelle Pineau, Krzysztof J. Geras, Lea Azour, Yindalon Aphinyanaphongs, Nafissa Yakubova, and William Moore. Covid-19 deterioration prediction via self-supervised representation learning and multi-image prediction, 2021. 2, 4

[14] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images, 2020. 2

[15] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification, 2020. 2

		Pleural Effusion						Edema						Atelectasis						Consolidation						Cardiomegaly						Average		
Training Strategy	Comparison	MCC	AUC	P	R	F1	A	MCC	AUC	P	R	F1	A	MCC	AUC	P	R	F1	A	MCC	AUC	P	R	F1	A	MCC	AUC	P	R	F1	A	MCC	AUC	A
Baseline	Standard	0.578	0.874	0.785	0.725	0.754	0.793	0.516	0.858	0.65	0.663	0.656	0.8	0.216	0.72	0.56	0.239	0.335	0.714	0.049	0.69	0.463	0.012	0.023	0.811	0.46	0.851	0.639	0.44	0.521	0.871	0.364	0.799	0.798
	Photos	0.546	0.869	0.807	0.65	0.72	0.777	0.473	0.85	0.722	0.519	0.604	0.784	0.141	0.715	0.569	0.102	0.173	0.709	0.002	0.669	0.2	0.002	0.003	0.82	0.345	0.84	0.645	0.26	0.37	0.854	0.301	0.789	0.789
Baseline Fine-tuned on CheXphoto	Standard	0.503	0.841	0.798	0.591	0.679	0.756	0.446	0.822	0.515	0.791	0.623	0.725	0.227	0.682	0.445	0.51	0.475	0.66	0.154	0.651	0.319	0.299	0.308	0.747	0.381	0.798	0.57	0.369	0.448	0.855	0.342	0.759	0.749
	Photos	0.721	0.939	0.851	0.836	0.843	0.863	0.694	0.938	0.828	0.745	0.784	0.87	0.513	0.856	0.681	0.624	0.651	0.801	0.517	0.867	0.683	0.512	0.585	0.87	0.677	0.944	0.767	0.69	0.727	0.914	0.624	0.909	0.863
Baseline Fine-tuned on CheXphoto (Frozen)	Standard	0.521	0.86	0.814	0.596	0.688	0.764	0.482	0.837	0.565	0.752	0.645	0.762	0.202	0.703	0.513	0.274	0.357	0.702	0.049	0.68	0.456	0.012	0.024	0.811	0.325	0.827	0.653	0.225	0.335	0.858	0.316	0.781	0.779
	Photos	0.617	0.895	0.767	0.816	0.791	0.809	0.558	0.876	0.69	0.71	0.7	0.807	0.27	0.747	0.58	0.313	0.406	0.728	0.091	0.716	0.514	0.031	0.058	0.821	0.454	0.876	0.669	0.408	0.507	0.868	0.398	0.822	0.807
Mixed Training	Standard	0.596	0.878	0.741	0.824	0.781	0.798	0.497	0.858	0.699	0.558	0.621	0.804	0.236	0.721	0.541	0.306	0.391	0.712	0.009	0.7	0.316	0.001	0.003	0.811	0.437	0.848	0.66	0.383	0.485	0.87	0.355	0.801	0.799
	Photos	0.63	0.897	0.738	0.88	0.803	0.809	0.536	0.883	0.803	0.529	0.637	0.809	0.316	0.755	0.564	0.43	0.488	0.731	0.074	0.702	0.75	0.01	0.02	0.822	0.466	0.881	0.683	0.416	0.517	0.871	0.404	0.824	0.809
MoCo Fine-tuned on CheXphoto Only	Standard	0.521	0.835	0.74	0.713	0.726	0.765	0.368	0.809	0.684	0.356	0.469	0.768	0.092	0.666	0.496	0.074	0.129	0.698	0	0.657	0	0	0	0.811	0.241	0.796	0.659	0.125	0.211	0.85	0.244	0.753	0.778
	Photos	0.544	0.845	0.755	0.729	0.742	0.776	0.419	0.831	0.737	0.412	0.529	0.767	0.098	0.7	0.543	0.059	0.106	0.705	0	0.674	0	0	0	0.82	0.261	0.825	0.692	0.137	0.229	0.847	0.264	0.775	0.783
MoCo Trained on CheXpert	Standard	0.568	0.872	0.796	0.695	0.742	0.789	0.481	0.842	0.638	0.617	0.627	0.789	0.16	0.701	0.539	0.156	0.242	0.705	0	0.686	0.2	0	0	0.811	0.388	0.841	0.7	0.284	0.404	0.867	0.319	0.788	0.792
	Photos	0.53	0.855	0.777	0.671	0.72	0.769	0.428	0.82	0.732	0.433	0.544	0.77	0.102	0.695	0.518	0.074	0.13	0.704	0	0.641	0	0	0	0.82	0.369	0.817	0.648	0.293	0.404	0.856	0.286	0.766	0.784
MoCo Trained on CheXpert / Fine-tuned CheXphoto	Standard	0.53	0.855	0.777	0.671	0.72	0.769	0.428	0.82	0.732	0.433	0.544	0.77	0.102	0.695	0.518	0.074	0.13	0.704	0	0.641	0	0	0	0.82	0.369	0.817	0.648	0.293	0.404	0.856	0.286	0.766	0.784
	Photos	0.623	0.89	0.816	0.749	0.781	0.815	0.554	0.872	0.654	0.761	0.703	0.796	0.3	0.739	0.544	0.432	0.482	0.723	0.091	0.705	0.4	0.055	0.096	0.816	0.53	0.87	0.642	0.566	0.602	0.876	0.419	0.815	0.805
MoCo Mixed Training	Standard	0.581	0.872	0.757	0.777	0.767	0.793	0.484	0.848	0.658	0.592	0.623	0.794	0.187	0.703	0.532	0.215	0.306	0.706	-0.005	0.685	0	0	0	0.811	0.415	0.841	0.678	0.337	0.45	0.869	0.333	0.79	0.795
	Photos	0.59	0.877	0.759	0.792	0.775	0.797	0.49	0.851	0.724	0.546	0.622	0.79	0.169	0.716	0.517	0.193	0.281	0.706	0	0.675	0	0	0	0.82	0.398	0.846	0.667	0.323	0.435	0.861	0.33	0.793	0.795
Mixed Training + CheXpert Head	Standard	0.602	0.881	0.77	0.787	0.778	0.804	0.513	0.86	0.664	0.637	0.65	0.803	0.24	0.725	0.552	0.297	0.387	0.715	0.101	0.703	0.462	0.05	0.091	0.81	0.454	0.851	0.631	0.437	0.516	0.87	0.382	0.804	0.8
Mixed Training + CheXphoto Head	Photos	0.65	0.903	0.776	0.848	0.811	0.825	0.569	0.88	0.724	0.679	0.701	0.816	0.304	0.758	0.574	0.389	0.464	0.732	0.093	0.709	0.387	0.062	0.106	0.814	0.509	0.884	0.688	0.482	0.567	0.878	0.425	0.827	0.813

Table 3. Results of experiments on validation sets. Standard refers to digital CXRs from CheXpert and Photos refers to photographs of CXRs from CheXphoto. A = Accuracy; Average Accuracy actually refers to overall accuracy

		Pleural Effusion						Edema						Atelectasis						Consolidation						Cardiomegaly						Average		
Training Strategy	Type	MCC	AUC	P	R	F1	A	MCC	AUC	P	R	F1	A	MCC	AUC	P	R	F1	A	MCC	AUC	P	R	F1	A	MCC	AUC	P	R	F1	A	MCC	AUC	A
Baseline	Standard	0.664	0.937	0.821	0.687	0.748	0.868	0.665	0.923	0.708	0.756	0.731	0.893	0.345	0.81	0.786	0.275	0.407	0.726	0	0.85	0	0	0	0.859	0.409	0.834	1	0.221	0.361	0.774	0.417	0.871	0.824
	Photos	0.575	0.916	0.812	0.557	0.661	0.836	0.473	0.887	0.74	0.4	0.519	0.858	0.211	0.79	0.806	0.104	0.185	0.685	0	0.847	0	0	0	0.859	0.288	0.77	1	0.113	0.203	0.742	0.309	0.842	0.796
Baseline Fine-tuned on CheXphoto	Standard	0.542	0.871	0.795	0.522	0.631	0.825	0.54	0.884	0.487	0.867	0.624	0.799	0.328	0.768	0.577	0.512	0.543	0.705	0.251	0.739	0.353	0.364	0.358	0.816	0.169	0.804	0.636	0.103	0.177	0.722	0.366	0.813	0.774
	Photos	0.551	0.852	0.725	0.617	0.667	0.823	0.472	0.883	0.641	0.489	0.555	0.849	0.243	0.737	0.564	0.35	0.432	0.685	0.149	0.706	0.296	0.212	0.247	0.818	0.342	0.764	0.773	0.25	0.378	0.761	0.351	0.788	0.787
MoCo Trained on CheXpert / Fine-tuned CheXphoto	Standard	0.62	0.881	0.786	0.657	0.715	0.85	0.535	0.902	0.54	0.756	0.63	0.829	0.456	0.831	0.857	0.375	0.522	0.765	0.162	0.874	1	0.03	0.059	0.863	0.31	0.743	0.727	0.235	0.356	0.752	0.416	0.846	0.812
	Photos	0.604	0.883	0.76	0.662	0.707	0.843	0.507	0.867	0.587	0.622	0.604	0.843	0.391	0.811	0.727	0.4	0.516	0.744	0.149	0.848	0.625	0.051	0.093	0.862	0.286	0.755	0.671	0.24	0.354	0.745	0.387	0.833	0.807
Mixed Training + CheXpert Head	Standard	0.695	0.936	0.788	0.776	0.782	0.876	0.625	0.935	0.705	0.689	0.697	0.885	0.342	0.824	0.735	0.312	0.439	0.726	0.172	0.894	0.667	0.061	0.111	0.863	0.401	0.841	0.941	0.235	0.376	0.774	0.447	0.886	0.825
Mixed Training + CheXphoto Head	Photos	0.605	0.918	0.742	0.687	0.713	0.842	0.519	0.913	0.65	0.563	0.603	0.858	0.395	0.823	0.74	0.392	0.512	0.745	0.149	0.861	0.471	0.081	0.138	0.858	0.368	0.821	0.952	0.196	0.325	0.764	0.407	0.867	0.813

Table 4. Complete Results of a few of the best models on the test set. Standard refers to digital CXRs from CheXpert and Photos refers to photographs of CXRs from CheXphoto. A = Accuracy; Average Accuracy actually refers to overall accuracy