SALIENCY IS A POSSIBLE RED HERRING WHEN DIAGNOSING POOR GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Poor generalization is one symptom of models that learn to predict target variables using spuriously-correlated image features present only in the training distribution instead of the true image features that denote a class. It is often thought that this can be diagnosed visually using saliency maps. We study if this is correct. In some prediction tasks, such as for medical images, one may have some images with masks drawn by a human expert, indicating a region of the image containing relevant information to make the prediction. We study multiple methods that take advantage of such auxiliary labels, by training networks to ignore distracting features which may be found outside of the region of interest, on the training images for which such masks are available. This mask information is only used during training and has an impact on generalization accuracy depending on the severity of the shift between the training and test distributions. Surprisingly, while these methods improve generalization performance in the presence of a covariate shift, there is no strong correspondence between the correction of saliency towards the features a human expert have labelled as important and generalization performance. These results suggest that the root cause of poor generalization may not always be spatially defined, and raise questions about the utility of masks as "attribution priors" as well as saliency for explainable predictions.

1 INTRODUCTION

A fundamental challenge when applying deep learning models stems from poor generalization due to covariate shift (Moreno-Torres et al., 2012) when the PAC learning i.i.d. assumption is invalid (Valiant, 1984) i.e. the training distribution is different from the test distribution. One explanation for this is *shortcut learning* or *incorrect feature attribution*, where the model during training overfits to a set of training-data specific decision rules to explain the training data instead of modelling the more general causative factors that generated the data (Goodfellow et al., 2016; Reed & Marks, 1999; Geirhos et al., 2020; Hermann & Lampinen, 2020; Parascandolo et al., 2020; Arjovsky et al., 2019; Zhang et al., 2016).

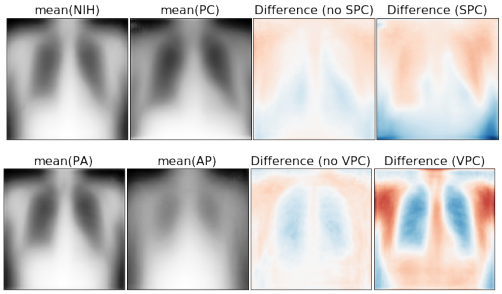

Figure 1: Illustration of the effect of Site-Pathology Correlations (SPC; top row) and View-Pathology Correlations (VPC; bottom row). The first two columns contain the mean of all X-Rays collected from a particular site/view. The third column shows the class-conditional mean difference between images taken from sick and healthy patients when pathology is uncorrelated with site/view, and the fourth column shows the same when pathology is correlated with site/view.

In medical imaging, test-set distribution shifts are common and this problem is exacerbated by small cohorts; previous work has identified the deleterious effect of confounding variables such as acquisition site or medically irrelevant aspects of the training images by demonstrating that saliency maps (Simonyan et al., 2014) produced by the model do not highlight features that a human expert would use to make a diagnosis (Zech et al., 2018; DeGrave et al., 2020; Badgeley et al., 2019; Zhao et al., 2019; Young et al., 2019). In Figure 1, we can see examples of the shift in distribution between imaging site (top) and imaging view (bottom). In the case that the prevalence of disease is in any way correlated with either of these confounders, the model is likely to learn to use these

differences to make a prediction regardless of their medical relevance (rightmost column). This can be a critical failure of the system, as explainable predictions are highly desirable for sensitive applications (Holzinger et al., 2017). A critical assumption underlying these aforementioned works is that properties of the saliency map are indicative of generalization performance.

In this work, we hypothesized that models can be regularized to ignore these confounders using *attribution priors* (Erion et al., 2019; Ross et al., 2017) to make predictions using the correct anatomy (as a doctor would), and this will improve out of distribution generalization. To evaluate this, we define *good saliency* to be a saliency map that agrees strongly with expert knowledge in the form of a binary mask on the input data.

We show that the existing and proposed feature-attribution-aware methods help facilitate generalization in the presence of a train-test distribution shift. However, while feature-attribution-aware methods change the saliency maps relative to baseline, there is no strong correlation between generalization performance and good saliency. This in turn challenges the assumption that the "incorrect" saliency maps were indicative of poor generalization performance. This suggests that A) efforts to validate model correctness using saliency maps may not be reliable, and B) that efforts to control feature attribution using masks on the input may not function as expected. All code and datasets for this paper are publicly available[1].

Our contributions include:

- Create a synthetic dataset that encourages models to overfit to an easy to represent confounder instead of a more complicated counting task.
- Create two new tasks constructed from open medical datasets which have a correlation between the pathology and either *imaging site* (site pathology correlation; SPC) or *view* (view pathology correlation; VPC), and we manipulate the nature of this correlation differently in the training and test distributions to create a distribution shift (Figure 1). This replicates the out of distribution feature attribution phenomena previously reported (Zhao et al., 2019; DeGrave et al., 2020).
- Explore many existing methods for controlling feature attribution using mask information; *right for the right reasons* (*RRR*; Ross et al. (2017)), *GradMask* (Simpson et al., 2019), adversarial domain invariance (Tzeng et al., 2017; Ganin & Lempitsky, 2015), as well as present a new method based on activation differences (*ActDiff*).
- Compare the real-life performance of feature-attribution-aware methods on these real-life out of distribution generalization tasks with traditional classifiers, and demonstrate the differences in their feature attributions using saliency maps.

2 Related work

It is a well-documented phenomenon that convolutional neural networks (CNNs), instead of building object-level representations of the input data, tend to find convenient surface-level statistics in the training data that are predictive of class (Jo & Bengio, 2017). Previous work has attempted to reduce the model's proclivity to use confounding features by randomly masking out regions of the input (DeVries & Taylor, 2017), forcing the network to learn representations that aren't dependent on single input feature. However, this regularization approach gives no control over the kinds of representations learned by the model, so we do not include it in our study.

Recently, multiple approaches have proposed to control feature representations by penalizing the model for producing saliency gradients outside of a regions of interest indicating the class-discriminative feature (Simpson et al., 2019; Zhuang et al., 2019; Rieger et al., 2019). These approaches were introduced by *Right for the Right Reasons* (*RRR*), which showed impressive improvements in attribution correctness on synthetic data (Ross et al., 2017). The follow-up work has generally demonstrated a small improvement in generalization accuracy on real data, and much more impressive results on synthetic data. Zhuang et al. (2019) was additionally designed to deal with class imbalances by increasing the impact of the gradients inside region of interest of the under-represented class.

[1] https://github.com/themaskedcrus4der/saliency-red-herring

An alternative to saliency-based methods, which can be noisy due to the ReLU activations allowing irrelevant features to pass through the activation function (Kim et al., 2019), would be to leverage methods that aim to produce *domain invariant* features in the latent space of the network. These methods regularize the network such that the latent representations of two or more "domains" are encouraged to be as similar as possible, often by minimizing a distance metric or by employing an adversary that is trained to distinguish between the different domains (Kouw & Loog, 2019; Ganin & Lempitsky, 2015; Tzeng et al., 2015; Liu & Tuzel, 2016). In this work, we view the masked version of the input as the training domain and the unmasked version of the input as the test domain, and compare these approaches with saliency-based approaches for the task of reducing the model's reliance on confounding features. To the best of our knowledge, these strategies have not been tried to control feature attribution.

3 METHODS

Domain Invariance with the Activation Difference and Adversarial Loss: Leveraging ideas from domain adaptation, we introduce two methods designed to make the model invariant to features arising from outside of the *attribution priors*. The first is the embarrassingly simple *Activation Difference* (*ActDiff*) approach, which simply penalizes, for each input, the l2-normed distance between the masked and unmasked input's latent representations. This loss is most similar to a style transfer approach which transforms a random noise image into one containing the same layer-wise activations as some target image (Gatys et al., 2015) to apply a visual aesthetic to some input semantic contents. This encourages the network to build features which appear inside the masked regions even though it always sees the full image during training. Therefore we minimize

$$\boldsymbol{\mathcal{L}_{act}} = \sum_{(\mathbf{X}_{masked},\mathbf{x})\in D} \mathcal{L}_{clf} + ||o_l(\mathbf{x}_{masked}) - o_l(\mathbf{x})||_2, \tag{1}$$

where $o_l(\mathbf{x})$ are the pre-activation outputs for layer l of the n-layer encoder $f(x)$ when the network is presented with the original data $\mathbf{x}$, $o_l(\mathbf{x}_{masked})$ are the pre-activations outputs for layer l when presented with masked data $\mathbf{x}_{masked}$, and $\mathcal{L}_{clf}$ is the standard cross entropy loss. We define $\mathbf{x}_{masked} = \mathbf{x}\cdot\mathbf{x}_{seg} + \text{shuffle}(\mathbf{x})\cdot(1-\mathbf{x}_{seg})$ where background pixels are shuffled uniquely for each presentation to the network in order to destroy any spatial information available in those regions of the image without introducing any consistent artefacts into the image or shifts in the distribution of input intensities.

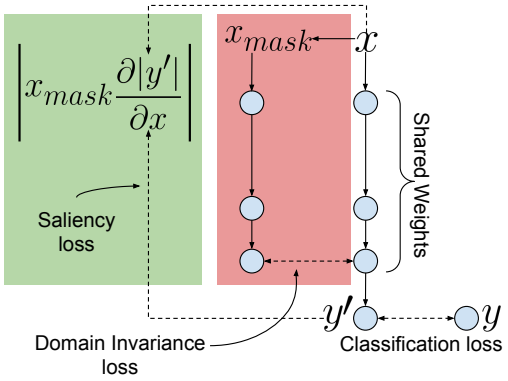

Figure 2: Schematic of the model used in all experiments. The backbone is an 18-layer ResNet. The *domain invariance* penalities, *ActDiff* and *Adversarial*, are applied to the linear layer after average pooling (although they could be applied at any location in the network). In contrast, the *saliency* penalties, *GradMask* and *Right for the Right Reasons*, are applied on the input space. All losses are denoted using standard dashed lines.

The second approach we explore employs a discriminator $\mathcal{D}(\cdot)$ optimized to distinguish between latent representations arising from passing $\mathbf{x}_{masked}$ or $\mathbf{x}$ through the encoder $f(\cdot)$ (Tzeng et al., 2017; Goodfellow et al., 2014). Simultaneously, we optimize the encoder $f(\cdot)$ to fool the discriminator and still produce representations that are predictive of the output class. Therefore, $\mathcal{D}$ and f are optimized using the $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_f$ respectively:

$$\boldsymbol{\mathcal{L}_{\mathcal{D}}} = \mathbb{E}_{\mathbf{x}_{masked}}[log\mathcal{D}(f(\mathbf{x}_{masked}))] + \mathbb{E}_{\mathbf{x}_{masked}}[log(1-\mathcal{D}(f(\mathbf{x})))] \tag{2}$$

$$\boldsymbol{\mathcal{L}_f} = \mathcal{L}_{clf} + \mathbb{E}_{\mathbf{x}}[log(1-\mathcal{D}(f(\mathbf{x}_{masked})))] + \mathbb{E}_{\mathbf{x}}[log\mathcal{D}(f(\mathbf{x}))] \tag{3}$$

This approach is similar to the one employed in (Janizek et al., 2020), where the authors encouraged the model to be invariant to the view of the X-Ray (Lateral vs PA), and here 'view' denotes whether the image is masked or not. $\mathcal{D}(\cdot)$ had three fully-connected hidden layers of size 1024 before outputting a binary prediction. To facilitate the stability of training, we updated $\mathcal{D}(\cdot)$ more frequently

than the encoder treating this ratio as a hyperparameter. We optimized $\mathcal{D}(\cdot)$ independently using Adam with a distinct learning rate, and applied spectral normalization to the hidden layers of $\mathcal{D}(\cdot)$.

Direct Saliency Control with the Right for the Right Reasons and GradMask Loss: These saliency-based feature attribution regularizers directly control which regions of the input are desirable for determining the class label by penalizing saliency outside of a defined input mask. Saliency maps, or "input feature attribution", can be calculated as $\frac{\partial|\hat{y}_i|}{\partial \mathbf{x}}$ for each input x (Zeiler & Fergus, 2013; Simonyan et al., 2014; Lo et al., 2015). *RRR* (Ross et al., 2017) calculates saliency of the summed *log probability* of the k output classes with respect to the input $\mathbf{x}$,

$$\boldsymbol{\mathcal{L}}_{rrr} = \sum_{(\mathbf{x}_{seg},\mathbf{x})\in D} \mathcal{L}_{clf} + \frac{\partial \left(\sum_k^K log(\hat{p_k})\right)^2}{\partial \mathbf{x}} \cdot (1 - \mathbf{x}_{seg}), \tag{4}$$

where $(1 - \mathbf{x}_{seg})$ is a binary mask that covers everything outside the defined regions of interest. *GradMask* (Simpson et al., 2019) similarly calculates saliency using the contrast between the two output logits (in the binary classification case),

$$\boldsymbol{\mathcal{L}}_{gradmask} = \sum_{(\mathbf{x}_{seg},\mathbf{x})\in D} \mathcal{L}_{clf} + \left| \frac{\partial \left|\hat{y}_0 - \hat{y}_1\right|}{\partial \mathbf{x}} \cdot (1 - \mathbf{x}_{seg}) \right|_2, \tag{5}$$

where $\hat{y}_0$ and $\hat{y}_1$ are the predicted outputs for our two classes.

Classify Masked: We evaluated the effect imposing the attribution prior by simply training a model using masked data (and evaluating it using unmasked data) as a control experiment. The data was masked by shuffling the pixels outside of the mask during training, as was done for the domain invariance experiments to produce $\mathbf{x}_{masked}$. We refer to these experiments as *Masked*.

4 EXPERIMENTS

Feature Attribution Analysis and Visualization: To evaluate our ability to control feature attribution, we first collected the absolute gradient g of the input with respect to the prediction made for all images of the positive class $g = |\frac{\partial \hat{y}_1}{\partial \mathbf{x}}|$ (Simonyan et al., 2014; Ancona et al., 2018). We next smoothed the resulting saliency map with a gaussian kernel ($\sigma = 1$). This output was used for all visualizations, and carried forward to calculate a binarized version such that the top p percentile saliency values were set to 1, where p is dynamically set to the number of pixels in the mask that is being compared to. We then calculated a measure of localization accuracy between binarized saliency map and the ground-truth segmentation using the intersection over union (IoU; $I_u(A, B) = \frac{A \cap B}{A \cup B}$).

Model, Optimization, Hyperparameters, and Search Parameters: The backbone of all experiments was the ResNet-18 model from Torchvision (He et al., 2016; Marcel & Rodriguez, 2010). For medical dataloaders we use the TorchXRayVision library (Cohen et al., 2020b;a). To see an overview of how the losses described above relate to the architecture, see Figure 2. All models we trained using the Adam optimizer (Kingma & Ba, 2014) using early stopping on the validation loss.

Hyperparamters were selected for all models using a Bayesian hyperparameter search with a single seed, 5 random initializations, 20 iterations, and trained for a maximum of 100 epochs with a patience of 20. Final models were trained using the best hyperparameters found for 100 epochs on 10 seeds.The learning rate for all models was searched on a log uniform scale between $[10^{-5}\ 10^{-2}]$. The *ActDiff*, *Adversarial*, *GradMask*, and *RRR* lambdas were all searched between $[10^{-4}\ 10]$, each on a log uniform scale. For *Adversarial*, we searched for the optimal discriminator : encoder training ratio of $[2 : 1\ 10 : 1]$, and the discriminator learning rate was searched on a log uniform scale between $[10^{-4}\ 10^{-2}]$. To retain context around the masked region in the Synthetic dataset, we dilated the mask by applying a Gaussian blur with a hyperparameter σ to the mask and then binarize the result; σ was searched on a uniform scale between $[0\ 2]$. For all experiments, the batch size was 16, weight of the classification cross entropy loss was 1. In performing our hyperparameter searches, we found that certain algorithms were much easier to tune than others. We present the final best validation AUCs encountered for all iterations for all experiments in Appendix Figure A.1.

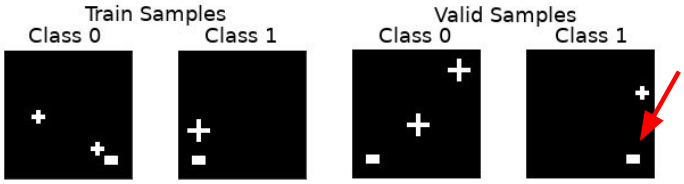

Figure 3: Example images from D_{train} and D_{valid} from both classes. In both distributions, cross size can vary between samples. In D_{train}, two crosses (denoting class 0) are always accompanied by a box x_c in the bottom right-hand corner, while a single cross (denoting class 1) is always accompanied by a confounder in the bottom left-hand corner. In D_{valid}, the relationship between classes and crosses remains the same, but the logic governing the location of the counfounder is reversed. The confounder is indicated with a red arrow.

4.1 Synthetic Dataset

Protocol: Consider a classification problem where there exists some confounder feature x_c in the data (a vector of variables x) that is perfectly correlated with one of the output classes y in the training distribution D_{train} such that $p(y = 1|x_c) = 1$, while in the validation distribution D_{valid}, $p(y = 0|x_c) = 1$ (Figure 3). In this scenario, predicting using x_c is easier than predicting using the true features that would denote class membership and a classifier trained on D_{train} with traditional classification loss would predict the incorrect class with 100% probability on D_{valid}.

We generated a dataset with these conditions where class membership is denoted by the number of crosses present in the image (1 vs. 2 crosses) that can appear in any position, but there exists in all D_{train} images a confounder that always appears in the same position and is perfectly correlated with y (Figure 3). The logic governing the position of the confounder was inverted for D_{valid}. This dataset had 500 training, 128 validation, and 128 test examples respectively. Segmentations were generated for all crosses in the dataset to facilitate all feature-attribution controlling losses. In cases where the model relies on the confounder to make a class prediction, we expected 0.0 AUC for the validation and test sets.

Results: The mean and standard deviation of the validation AUC over all 10 models for the 100 epochs of training for the best-performing hyperparameters are shown in Figure 4, showing the variance in performance over model initilizations and data splits. We demonstrate the effect of controlling feature attribution by showing the mean saliency map calculated for 10 positive examples in the test set in Figure 5.

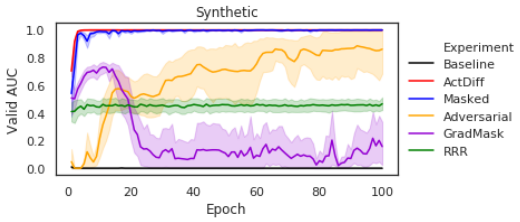

Figure 4: Synthetic Dataset Valid AUC over 100 epochs, averaged over 10 seeds.

The validation curves (Figure 4, demonstrate that the baseline model quickly learns the shortcut in the training set and shows below-chance generalization (AUC=0). The *Masked* approach is never given the opportunity to build a representation of the confounder and is therefore able to generalize. The domain invariance approaches, *ActDiff* and *Adversarial* also learn the task, although convergence is much more difficult for the *Adversarial* approach. The two saliency-based feature-attribution regularizers, *GradMask*, and *RRR*, show unstable training and never reach optimal performance, although they do prevent the model from overfitting to the confounder. The saliency maps for these models (Figure 5) show that all feature attribution controlling approaches (C-F) successfully refocus the model's attention away from the confounder, unlike the baseline and *Masked* models (A-B), leading to improved generalization. This

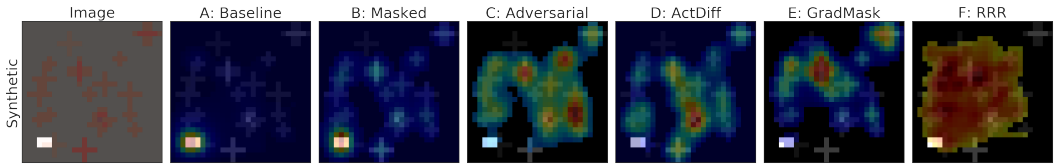

Figure 5: The average saliency maps from 10 randomly-selected D_{test} images for the Synthetic dataset. The first image shows the mean of the samples, with the mean of the masks superimposed in red. The remainder of the panels show the mean of the gradients after taking the absolute value, smoothing, and thresholding at the 50th percentile.

strong relationship between IoU and AUC for this dataset can be observed for all seeds in Figure 8 (note that this relationship does not hold for real world data).

4.2 X-RAY DATASET WITH SITE-PATHOLOGY CORRELATION

Protocol: We introduced a covariate-shift into a joint dataset of X-Rays drawn from two different imaging centres: the PadChest (Bustos et al., 2019) sample and the NIH Chestx-Ray8 (Wang et al., 2017) sample. The models were required to predict the presence of Cardiomegaly (enlarged heart). A site-driven overfitting signal has previously been reported when combining these datasets (Zech et al., 2018), where the model often attributes importance to the patient's shoulder when making a prediction. We also observed site-driven differences in regions far from the lungs (see the mean X-Ray from each dataset in Figure 1), and therefore hypothesized we could improve overfitting performance by masking out the edges of the image using a circular mask. To induce the covariate shift, we preferentially sampled the positive class from one of the two imaging centres in the training set to produce a site-pathology correlation (SPC), and then produce validation and test set where the reverse relationship is true. In the training set, 90% of the unhealthy patients were drawn from the PadChest dataset and the remaining 10% of the unhealthy patients were drawn from the NIH dataset, and the reverse logic was followed for the validation and test sets. In all splits the classes and site distributions were always balanced, making it tempting for the classifier to use a site-specific feature when predicting the class in the presence of site-pathology correlation ($N_{train} = 5542$, $N_{valid} = 2770$, $N_{test} = 2770$). We also constructed a version of this dataset where the classes are drawn equally from both sites for the training, validation, and test sets, i.e., a dataset where there exists no site pathology correlation (No SPC). All images were 244×244 pixels.

		With SPC		No SPC	
	Experiment	AUC	IoU	AUC	IoU
Baseline	Classification	0.70 ± 0.05	0.59 ± 0.03	0.93 ± 0.01	0.66 ± 0.03
	Masked	$\mathbf{0.77 \pm 0.09}$	0.60 ± 0.03	0.90 ± 0.01	0.63 ± 0.02
Domain Invariance	ActDiff	$\mathbf{0.73 \pm 0.03}$	$\mathbf{0.77 \pm 0.03}$	0.80 ± 0.18	0.58 ± 0.21
	Adversarial	0.68 ± 0.07	0.55 ± 0.03	0.92 ± 0.01	0.63 ± 0.03
Saliency Based	GradMask	$\mathbf{0.74 \pm 0.03}$	$\mathbf{0.74 \pm 0.01}$	$\mathbf{0.94 \pm 0.00}$	$\mathbf{0.71 \pm 0.02}$
	RRR	0.57 ± 0.09	0.57 ± 0.02	0.93 ± 0.01	0.64 ± 0.02

Table 1: X-Ray Cardiomegaly site-pathology correlation (SPC) test results for the best valid epoch over 100 epochs on the X-Ray dataset with and without an SPC. Results are averaged over 10 seeds, and we present the standard deviation. Results in bold both beat the Classification baseline and perform better than chance.

Results: We trained each model on both the SPC (90%) and No SPC datasets (Table 1). A ResNet trained on the No SPC dataset scores a test AUC of 0.93 ± 0.01, while one trained in the presence of a strong SPC scores a test AUC of 0.70 ± 0.05, indicating poorer generalization under a SPC. Validation curves across seeds show that model optimization was much easier and more consistent for all models when trained on the No SPC dataset (Appendix Figure A.2). The *ActDiff*, *Masked*, and *GradMask* approaches only improved classification performance over the baseline model in the presence of an SPC (*GradMask* performed only slightly better than baseline for the No SPC dataset). For *ActDiff* and *GradMask*, an improvement in classification performance was associated with an improvement in saliency (measured as an improvement in the IoU between the binarized saliency map and the masks), but this pattern was not observed for the masked experiment, even though the *Masked* experiment produced superior generalization (it is worth noting that the *Masked* performance was more variable across seeds). Furthermore, the *GradMask* approach improved saliency in the No SPC case, but this was not associated with large improvement in generalization performance. Figure 6 shows the mean saliency map calculated from a subset of all test images from all 10 models. Differences in the baseline model's saliency (A) in the SPC and No SPC case are apparent: in particular, the SPC model shows a greater reliance on the shoulder for producing a prediction. The *Masked* experiment (B) exacerbates this behavior but is accompanied by improved generalization. In contrast, *ActDiff* and *GradMask* appear to refocus the saliency away from the shoulder (D-E), reflected in the increased IoU scores for both the *ActDiff* and *GradMask* approaches for the SPC data. The weak relationship between IoU and AUC can be observed for all seeds in Figure 8. These results provide evidence that controlling feature attribution only facilitates generalization in the presence of

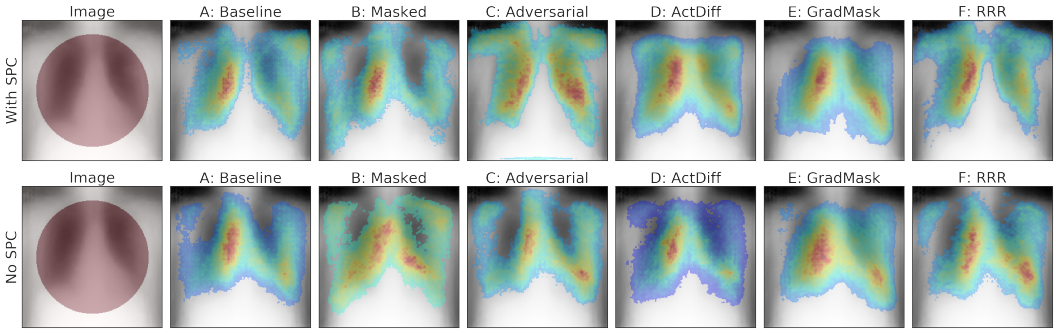

Figure 6: Mean saliency maps from 500 randomly-selected D_{test} images in the X-Ray dataset with a site-pathology correlation (SPC; top), and without (No SPC; bottom) after taking the absolute value, smoothing, and thresholding at the 50th percentile across all 10 seeds. The first image shows the mean input image of the model over all examples shown, with the mask overlaid in red.

a covarite shift, but improves saliency in both the presence or absence of a covariate shift, implying a weak relationship between good saliency and good generalization.

4.3 X-Ray Dataset with View-Pathology Correlation

Protocol: Here we made use of the RSNA Pneumonia challenge (Shih et al., 2019) dataset, which includes images taken from antero-posterior (AP) and posterior-anterior (PA) views, and where the task is to predict pneumonia. We introduced a covariate shift into this dataset by sampling the positive and negative classes during training such that 90% of the positive classes are sampled from one view in the training set and the opposite view in the validation and test sets. We call this a view-pathology correlation (VPC). After drawing the samples from both views and balancing the classes, the data in each split was as follows: $N_{train} = 2696$, $N_{valid} = 1348$, $N_{test} = 1348$. All images with Pneumonia have one or more bounding boxes indicating the predictive regions. In Figure 1 we can see the strong effect of the VPC, such that the lungs, shoulders, and torso show strong interclass differences in the VPC case. We also constructed a No VPC dataset in the same way as during the previous experiments, and all images were 224×224 pixels.

		With VPC		No VPC	
	Experiment	AUC	IoU	AUC	IoU
Baseline	Classification	0.20 ± 0.03	0.13 ± 0.01	0.76 ± 0.02	0.26 ± 0.03
	Masked	$\mathbf{0.56 \pm 0.16}$	0.05 ± 0.03	0.50 ± 0.02	0.09 ± 0.04
Domain Invariance	ActDiff	$\mathbf{0.68 \pm 0.05}$	0.15 ± 0.02	0.60 ± 0.02	0.19 ± 0.01
	Adversarial	$\mathbf{0.62 \pm 0.10}$	0.11 ± 0.05	0.65 ± 0.02	0.14 ± 0.03
Saliency Based	GradMask	0.49 ± 0.06	$\mathbf{0.22 \pm 0.09}$	0.75 ± 0.01	0.25 ± 0.04
	RRR	0.21 ± 0.04	0.13 ± 0.02	0.75 ± 0.01	0.24 ± 0.05

Table 2: RSNA Pneumonia view-pathology correlation (VPC) test Results for the best valid epoch over 100 epochs on the RSNA dataset with and without a VPC. Results are averaged over 10 seeds, and we present the standard deviation. Results in bold both beat the Classification baseline and perform better than chance.

Results: Similarly to the SPC case, baseline AUC in the presence of VPC (0.20 ± 0.03) is substantially lower than the No VPC baseline (0.76 ± 0.02), and additionally far below chance performance (0.50), illustrating the strong covariate shift in this dataset. Appendix Figure A.2 shows that validation performance was more variable in the presence of VPC between models, and that performance was generally lower. The *ActDiff*, *Masked*, and *Adversarial* approaches improved classification performance over the baseline model and chance performance in the presence of VPC. For only the *GradMask* model was an improvement in classification performance associated with an improvement in saliency, and the *Masked* experiment produced IoU scores substantially worse than baseline (again, *Masked* performance was more variable across seeds). No model improved saliency or generalization performance over baseline in the No VPC case. Figure 7 shows the mean saliency map calculated from a subset of all test images from all 10 models. Differences in the *Baseline* model's saliency (A) in the VPC and No VPC case are apparent: in particular, the VPC model shows a greater reliance on the abdomen and shoulders for producing a prediction. The *Masked* experiment (B) again exacerbates this behavior but is accompanied by improved generalization over baseline. The *Adversarial*, *ActDiff*, and *GradMask* models (C-E) appear to refocus the saliency toward the

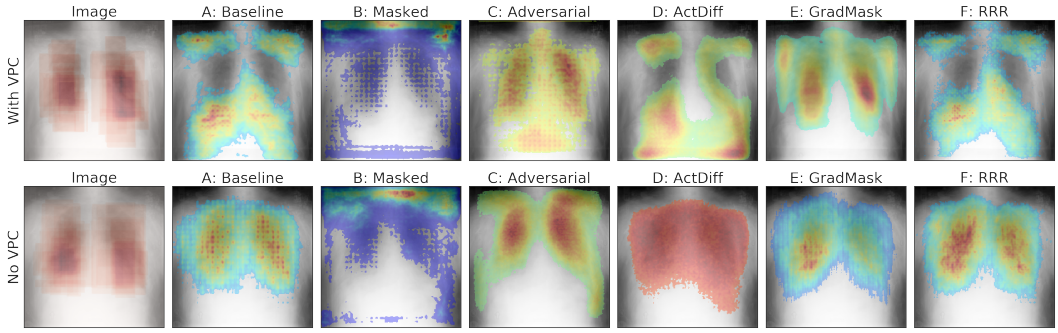

Figure 7: Mean saliency maps from 500 randomly-selected D_{test} images in the RSNA dataset with a site-pathology correlation (VPC; top), and without (No VPC; bottom) after taking the absolute value, smoothing, and thresholding at the 50th percentile across all 10 seeds. The first image shows the mean input image of the model over all examples shown, with the mask overlaid in red.

lungs for for the VPC case, but the mean *Adversarial* IoU score dropped over the full dataset, and *ActDiff* resulted in a minor IoU score increase. The *GradMask* approach produced the best saliency as measured by IoU but does not produce a model that outperforms random guessing on the test set. Again, these results suggest that the use of attribution priors can have a substantial effect on generalization performance in the absence of meaningful changes to the saliency towards what would be expected from the attribution priors, and vice versa. The weak relationship between IoU and AUC can be observed for all seeds in Figure 8.

5 Conclusion

We hypothesized that poor generalization performance could be partially attributable to classifiers exploiting spatially-distinct confounding features (or shortcuts) as have been previously diagnosed using saliency maps (Badgeley et al., 2019; Zech et al., 2018). We evaluated the performance of multiple mitigation strategies on a synthetic dataset and two real-world datasets exhibiting covariate shift, using attribution priors in the form of a segmentation of the discriminative features.

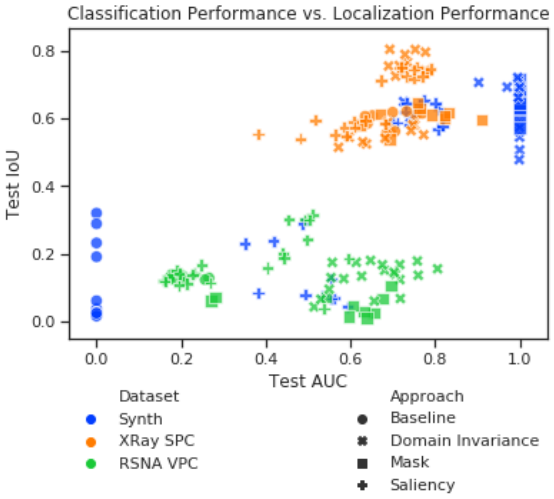

Figure 8: Test AUC vs. Test IoU (good saliency) for all seeds evaluated with covariate shift. Models presented grouped by "Approaches": baseline, input masking, domain-invariance, and saliency-based approaches.

In our synthetic dataset, we defined a spatially-distinct confounder that prevented generalization for standard classifiers, and demonstrated that imposing an attribution prior would facilitate generalization and improve saliency. In real data, we found that attribution priors also facilitate generalization when there exists a covariate shift between the training and testing distribution, but these methods hurt generalization when there was not a covariate shift. Surprisingly, improvement in generalization performance was not reliably accompanied by improvement in saliency: this is particularly apparent for the *Masked* condition which had a meaningful positive impact on AUC in both experiments but did not improve saliency. Figure 8 summarizes the relationship between generalization (AUC) and feature attribution (IoU) across all seeds for all experiments with a covariate shift: the positive correlation between AUC and IoU apparent for the synthetic data was not present in the real world datasets.

In summary, we find a tenuous relationship between good saliency and generalization performance in real datasets. Many models that exhibit good generalization performance do not obtain good saliency, and vice-versa. While our methods exert influence on the features the model constructs from the data, in real data we found no evidence of a relationship between improved generalization and improved saliency-based feature attribution. We now doubt the validity of using saliency maps for diagnosing whether a model is overfit because improving generalization using attribution priors is not accompanied by improved saliency.

REFERENCES

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. 2018 In *International Conference on Learning Representations*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2019.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv e-prints*, 2019 art. arXiv:1901.07441.

Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, 2020a.

Joseph Paul Cohen, Joseph Viviano, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. *https://github.com/mlmed/torchxrayvision*, 2020b.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. 2020 *medRxiv*. doi: 10.1101/2020.09.13.20193565.

Terrance DeVries and Graham W. Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. 2017 *arXiv:1708.04552*.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Learning explainable models using attribution priors. *arXiv:1906.10670*, 2019.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 2015.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv:1508.06576*, 2015.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv:2004.07780*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.90.

Katherine L Hermann and Andrew K Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *arXiv:2006.12433*, 2020.

Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv:1712.09923*, 2017.

Joseph D Janizek, Gabriel Erion, Alex J DeGrave, and Su-In Lee. An adversarial approach for the robust classification of pneumonia from chest radiographs. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020.

Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *CoRR*, 2017.

Beomsu Kim, Junghoon Seo, SeungHyun Jeon, Jamyoung Koo, Jeongyeol Choe, and Taegyun Jeon. Why are saliency maps noisy? cause of and solution to noisy saliency maps. *arXiv:1902.04893*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Wouter Marco Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, 2016.

Henry Z. Lo, Joseph Paul Cohen, and Wei Ding. Prediction gradients for feature extraction and analysis from convolutional neural networks. 2015 In *International Conference on Automatic Face and Gesture Recognition*. doi: 10.1109/FG.2015.7163154.

Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, 2010.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. 2012 *Pattern Recognition*. doi: 10.1016/J.PATCOG.2011.06.019.

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv:2009.00329*, 2020.

Russell D. Reed and Robert J. Marks. *Neural smithing : supervised learning in feedforward artificial neural networks*. MIT Press, 1999.

Laura Rieger, Chandan Singh, W James Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. *arXiv:1909.13584*, 2019.

Andrew Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *International Joint Conference on Artificial Intelligence*, 2017.

George Shih, Carol C. Wu, Safwan S. Halabi, Marc D. Kohli, Luciano M. Prevedello, Tessa S. Cook, Arjun Sharma, Judith K. Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C.B. Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N. Shah, Dharshan Vummidi, Kavitha Yaddanapudi, and Anouk Stein. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. 2019 *Radiology: Artificial Intelligence*. doi: 10.1148/ryai.2019180041.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *International Conference on Learning Representations (ICLR)*, 2014.

Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. GradMask: Reduce Overfitting by Regularizing Saliency. In *Medical Imaging with Deep Learning Workshop*, 2019.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

L. G. Valiant. A theory of the learnable. In *Proceedings of the Annual ACM Symposium on Theory of Computing*. 1984 Association for Computing Machinery. doi: 10.1145/800057.808710.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Kyle Young, Gareth Booth, Becks Simpson, Reuben Dutton, and Sally Shrapnel. Deep neural network or dermatologist? *arXiv:1908.06612*, 2019.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric K. Oermann. Confounding variables can degrade generalization performance of radiological deep learning models. *CoRR*, 2018.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, 2013.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*, 2016.

Qingyu Zhao, Ehsan Adeli, Adolf Pfefferbaum, Edith V Sullivan, and Kilian M Pohl. Confounder-Aware Visualization of ConvNets. *arXiv:1907.12727*, 2019.

Jiaxin Zhuang, Jiabin Cai, Ruixuan Wang, Jianguo Zhang, and Weishi Zheng. {CARE}: Class attention to regions of lesion for classification on imbalanced data. In *International Conference on Medical Imaging with Deep Learning – Full Paper Track*, 2019 London, United Kingdom.

A APPENDIX

A.1 ARCHITECTURE OF THE RESNET-18 MODEL

The backbone of the model used for all experiments was the 18-layer resnet available from torchvision (Marcel & Rodriguez, 2010) with a small architecture adjustment near the input of the network. The official model's input convolutional layer had a kernel size of 7, stride of 2, and padding of 3. In our experiments, we found this larger kernel size on the inputs led to less specific saliency maps (data not shown), and we therefore changed the kernel size to 3, stride to 1, and padding to 1 for this layer. Immediately following this layer, the official model employed a maxpooling operation. We found this operation to produce unusable noisy saliency maps, so we also disabled this operation for all experiments.

A.2 HYPERPARAMETER SENSITIVITY

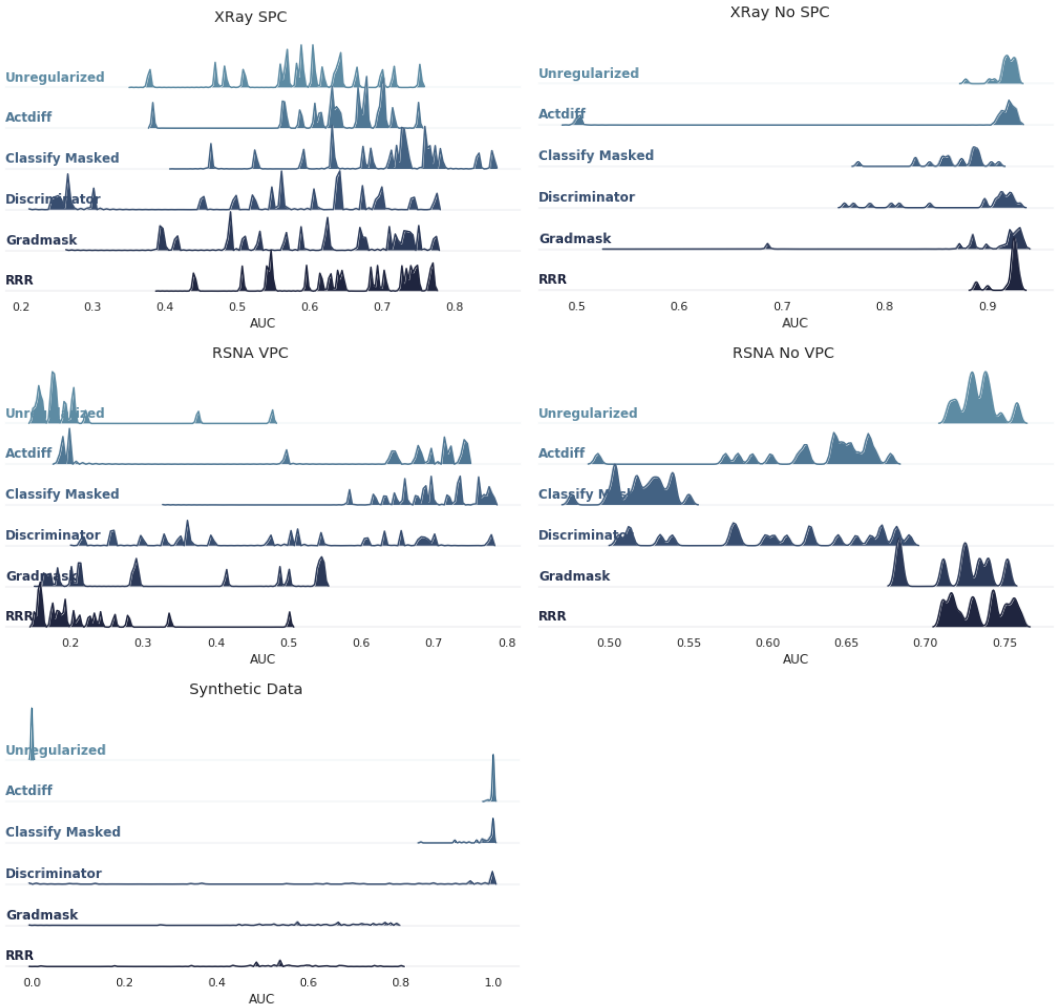

Figure A.1: The best validation AUC for all hyperparameter settings tested during the model tuning. On the synthetic dataset, domain-adaptation approaches were able to solve the task, as did the model that simply observed masked versions of the data. The saliency-based approaches were generally more sensitive to hyperparameters. In the presence of a covariate shift (SPC / VPC), hyperparameter tuning was more difficult on all real-world data.

A.3 VALIDATION CURVES FOR THE REAL WORLD DATASETS

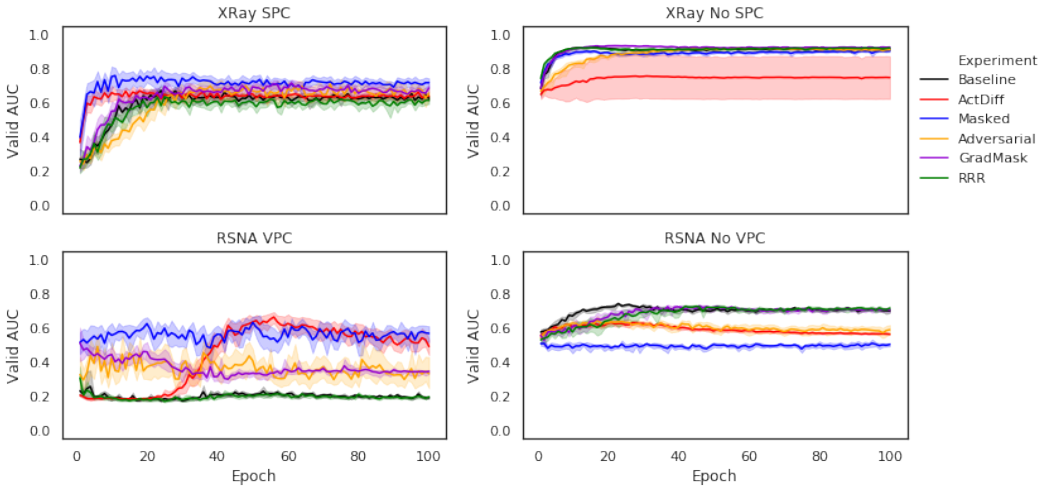

Figure A.2: All real dataset (X-Ray and RSNA) valid AUC across the 100 epochs of training. Mean and standard deviation presented over 10 seeds.