THINKING LIKE A MACHINE — GENERATING VISUAL RATIONALES THROUGH LATENT SPACE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpretability and small labelled datasets are key issues in the practical application of deep learning, particularly in areas such as medicine. In this paper, we present a semi-supervised technique that addresses both these issues simultaneously. We learn dense representations from large unlabelled image datasets, then use those representations to both learn classifiers from small labeled sets and generate visual rationales explaining the predictions.

Using chest radiography diagnosis as a motivating application, we show our method has good generalization ability by learning to represent our chest radiography dataset while training a classifier on an separate set from a different institution. Our method identifies heart failure and other thoracic diseases. For each prediction, we generate visual rationales for positive classifications by optimizing a latent representation to minimize the probability of disease while constrained by a similarity measure in image space. Decoding the resultant latent representation produces an image without apparent disease. The difference between the original and the altered image forms an interpretable visual rationale for the algorithm's prediction. Our method simultaneously produces visual rationales that compare favourably to previous techniques and a classifier that outperforms the current state-of-the-art.

1 INTRODUCTION

Deep learning as applied to medicine has attracted much interest in recent years as a potential solution to many difficult problems in medicine, such as the recognition of diseases on pathology slides or radiology images. However, adoption of machine learning algorithms in fields such as medicine relies on the end user being able to understand and trust the algorithm, as incorrect implementation and errors may have significant consequences. Hence, there has recently been much interest in interpretability in machine learning as this is a key aspect of implementing machine learning algorithms in practice. We propose a novel method of creating visual rationales to help explain individual predictions and explore a specific application to classifying chest radiographs.

There are several well-known techniques in the literature for generating visual heatmaps. Gradient based methods were first proposed in 2013 described as a saliency map in Simonyan et al. (2013), where the derivative of the final class predictions is computed with respect to the input pixels, generating a map of which pixels are considered important. However, these saliency maps are often unintelligible as convolutional neural networks tend to be sensitive to almost imperceptible changes in pixel intensities, as demonstrated by recent work in adversarial examples. In fact, obtaining the saliency map is often the first step in generating adversarial examples as in Goodfellow et al. (2014). Other recent developments in gradient based methods such as Integrated Gradients from Sundararajan et al. (2017) have introduced fundamental axioms, including the idea of sensitivity which helps focus gradients on relevant features.

Occlusion sensitivity proposed by Zeiler & Fergus (2013) is another method which covers parts of the image with a grey box, mapping the resultant change in prediction. This produces a heatmap where features important to the final prediction are highlighted as they are occluded. Another well-known method of generating visual heatmaps is global average pooling. Using fully convolutional neural networks with a global average pooling layer as described in Zhou et al. (2015), we can

examine the class activation map for the final convolutional output prior to pooling, providing a low resolution heatmap for activations pertinent to that class.

A novel analysis method by Ribeiro et al. (2016) known as locally interpretable model-agnostic explanations (LIME) attempts to explain individual predictions by simulating model predictions in the local neighbourhood around this example. Gradient based methods and occlusion sensitivity can also be viewed in this light — attempting to explain each classification by changing individual input pixels or occluding square areas.

However, sampling the neighbourhood surrounding an example in raw feature space can often be tricky, especially for image data. Image data is extremely complex and high-dimensional — hence real examples are sparsely distributed in pixel space. Sampling randomly in all directions around pixel space is likely to produce non-realistic images.

LIME's solution to this is to use superpixel based algorithms to oversegment images, and to perturb the image by replacing each superpixel by its average value, or a fixed pre-determined value. While this produces more plausible looking images as opposed to occlusion or changing individual pixels, it is still sensitive to the parameters and the type of oversegmentation used — as features larger than a superpixel and differences in global statistics may not be represented in the set of perturbed images. This difficulty in producing high resolution visual rationales using existing techniques motivates our current research.

2 METHODS

We introduce a novel method utilizing recent developments in generative adversarial networks (GANs) to generate high resolution visual rationales. We demonstrate the use of this method on a large dataset of frontal chest radiographs by training a classifier to recognize heart failure on chest radiographs, a common task for doctors.

Our method comprises of three main steps — we first use generative adversarial networks to train a generator on an unlabelled dataset. Secondly, we use the trained generator as the decoder section of an autoencoder. This enables us to encode and decode, to and from the latent space while still producing high resolution images. Lastly, we train simple supervised classifiers on the encoded representations of a smaller, labelled dataset. We optimize over the latent space surrounding each encoded instance with the objective of changing the instance's predicted class while penalizing differences in the resultant decoded image and the original reconstructed image. This enables us to visualize what that instance would appear as if it belonged in a different class.

Firstly, we use the Wasserstein GAN formulation by Arjovsky et al. (2017) and find that the addition of the gradient penalty term helps to stabilize training as introduced by Gulrajani et al. (2017). Our unlabelled dataset comprises of a set of 98,900 chest radiograph images, which are scaled to 128 by 128 pixels while maintaining their original aspect ratio through letterboxing, and then randomly translated by up to 8 pixels. We use a 100 dimensional latent space. Our discriminator and generator both use the DCGAN architecture while excluding the batch normalization layers and using Scaled Exponential Linear Units described in Klambauer et al. (2017) as activations except for the final layer of the generator which utilized a Tanh layer. We train the critic for 4 steps for each generator training step. The GAN training process was run for 200k generator iterations before visually acceptable generated images were produced. ADAM was used as the optimizer with the generator and discriminator learning rates both set to 5×10^{-5}.

In the next step, we use the trained generator as the decoder for the autoencoder. We fix the weights of the decoder during training and train our autoencoder to reproduce each of the images from the unlabelled dataset. The unlabelled dataset was split by patient in a 15 to 1 ratio into a training and validation set. We minimize the Laplacian loss between the input and the output, inspired by Bojanowski et al. (2017). Minimal overfitting was observed during the training process even when the autoencoder was trained for over 1000 epochs, the reconstruction loss on the validation set was similar to the reconstruction loss on the validation set.

We then train a classifier on a smaller labelled dataset consisting of 7,391 chest radiograph images paired with a B-type natriuretic peptide (BNP) blood test that is correlated with heart failure. This test is measured in nanograms per litre, and higher readings indicate heart failure. We perform a

natural logarithm on the actual value and divide the resultant number by 10 to scale these readings to between 0 and 1. We augment each labelled image and encode it into the latent space using our previously trained autoencoder. To prevent contamination, we separate our images by patient into a training and testing set with a ratio of 4 to 1 prior to augmentation and encoding. We demonstrate the success of simple classifiers upon this latent representation, including a 2 layer multilayer perceptron as well as a linear regressor.

To obtain image specific rationales, we optimize over the latent space starting with the latent representation of the given example. We fix the weights of the entire model and apply the ADAM optimizer on a composite objective comprising of the output value of the original predicted class and a linearly weighted mean squared error term between the decoded latent representation and the decoded original representation. We cap the maximum number of iterations at 5000 and set our learning rate at 0.1. We stop the iteration process early if the cutoff value for that class is achieved. The full algorithm is described in Algorithm 1. This generates a latent representation with a different prediction from the initial representation. The difference between the decoded generated representation and the decoded original representation is scaled and overlaid over the original image to create the visual rationale for that image. We use gradient descent to optimize the following objective:

$$z_{\text{target}} = \arg\min_{z} \; L_{\text{target}}(z) + \alpha \|X - G(z)\|^2 \tag{1}$$

$$X_{\text{target}} = G(z_{\text{target}}) \tag{2}$$

Where X is the reconstructed input image (having been passed through the autoencoder); X_{target} and z_{target} are the output image and its latent representation. G is our trained generator neural network. α is a coefficient that trades-off the classification and reconstruction objectives. L_{target} is a target objective which can be a negative class probability or in the case of heart failure, predicted BNP level. The critical difference between our objective and the one used for adversarial example generation is that optimization is performed in the latent space, not the image space.

Algorithm 1 Visual rationale generation

Require: α, learning rate
γ, image similarity penalty
ρ, cutoff value
Require: x, the initial input
$f : x \to z$, a function approximating the mapping between image and latent space
$g : z \to x$
$h(z)$, classifier predicting value from z

$z_0 \leftarrow z \leftarrow f(x)$
repeat
  $d \leftarrow \langle (g(z) - g(z_0))^2 \rangle$
  $y \leftarrow h(z)$
  $z \leftarrow z + \alpha * ADAM(z, y + \gamma d)$
until $y > \rho$
return $g(z_0) - g(z)$

We also apply our method to external datasets and demostrate good cross-dataset generalization, in particular the National Institutes of Health (NIH) ChestX-ray8 dataset recently released by Wang et al. (2017) We downsize the provided images to work with our autoencoder and split this by patient into a training, validation and testing set in the 7:1:2 ratio used by the dataset's authors. We encode these images into the latent space and apply a 6 layer fully connected neural network with 100 nodes in each layer utilizing residual connections. This architecture is fully described in figure 1.

To evaluate the usefulness of the generated visual rationales, we conduct an experiment where we compare visual rationales generated by a classifier to one which is contaminated. We train the classifier directly on the testing examples and over train until almost perfect accuracy on this set is achieved. We reason that the contaminated classifier will simply memorize the testing examples and hence will not be able to produce useful rationales.

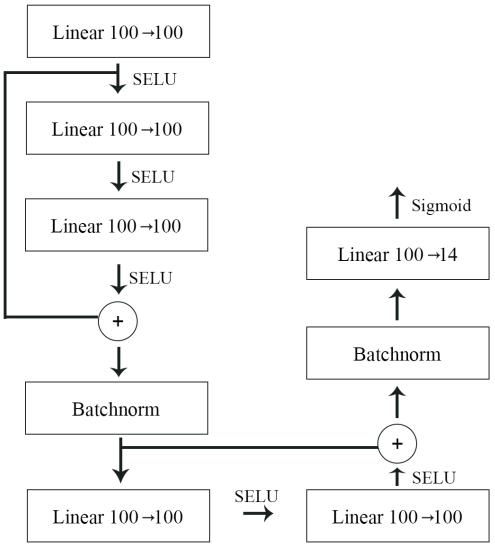

Figure 1: Classifier used for ChestX-ray8 dataset

We also apply our method to the well known MNIST dataset and apply a linear classifier with a 10 way softmax. In order to generate our visual rationales we select an initial class and a target class — we have chosen to transform the digit 9 to the digit 4 as these bear physical resemblance. We alter our optimization objective by adding a negatively weighted term for the predicted probability of the target class as described in Algorithm 2.

Algorithm 2 Visual rationale generation for multiclass predictors

Require: α, learning rate
γ, image similarity penalty
ρ, cutoff value
β, target class weighting
t, target class
Require: x, the initial input
$f : x \to z$, a function approximating the mapping between image and latent space
$g : z \to x$
$h_c(z) \to P(c|z)$, classifier predicting class probability from z
$z_0 \leftarrow z \leftarrow f(x)$
$m \leftarrow \text{argmin}_i \, h_i(z_0)$
repeat
  $d \leftarrow \langle (g(z) - g(z_0))^2 \rangle$
  $y_m \leftarrow h_m(z)$
  $y_t \leftarrow h_t(z)$
  $z \leftarrow z + \alpha * ADAM(z, y_m - \beta y_t + \gamma d)$
until $y_m > \rho$
return $g(z_0) - g(z)$

3 RESULTS

To illustrate the fidelity of our autoencoder we reconstruct each image in a smaller labelled set which has not been seen during training. The reconstructed images are show in Fig. 2. These images are obtained by simply encoding the input image into the latent representation and subsequently decoding this representation again.

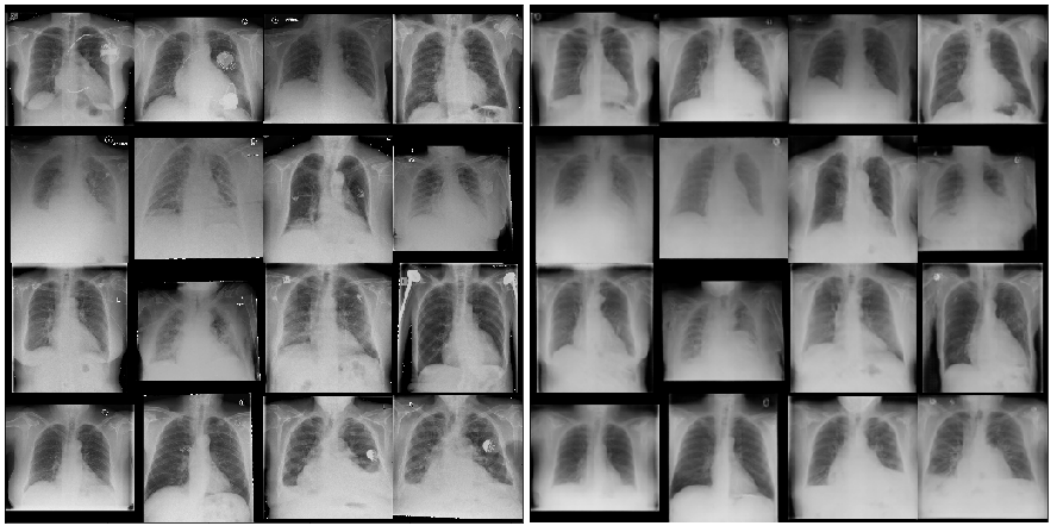

Figure 2: Left to right: Original images, reconstructed images

In the heart failure classification task, we threshold the known BNP values at 100ng/L to get binary labels as suggested by Lokuge et al. (2009). Our semi-supervised model achieves an AUC of 0.837 using a linear regressor as our final classifier with an AUC curve as shown in Fig 3. This is comparable to the AUC obtained by a multilayer perceptron.

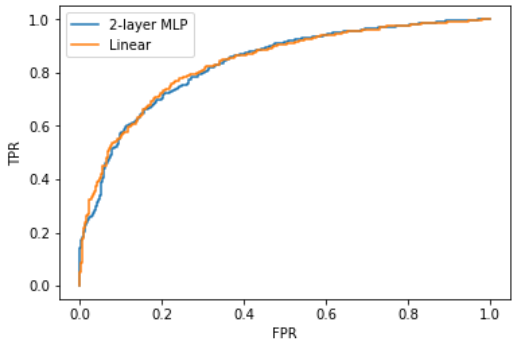

Figure 3: ROC plot for BNP prediction

In Fig 4 we demonstrate an example of the algorithm's reconstruction of a chest radiograph from a patient with heart failure, as well as the visualization of the same patient's chest radiograph without heart failure. We subtract the visualization of the radiograph without heart failure from the original reconstructed image and superimpose this as a heatmap on the original image to demonstrate the visual rationale for this prediction.

For the same image, we apply the saliency map method, integrated gradients, the occlusion sensitivity method with a window size of 8, as well as LIME to obtain Fig. 5 for comparison.

We apply our classifier as described above to the chest radiograph dataset released by the NIH recently and achieve results similar to or exceeding that of the baseline results reported in the original dataset. ROC curves are demonstrated in Fig 6. Comparison AUC results are reported in Table 1. We show that even without repeating the autoencoder or GAN training process on the new dataset, we are able to classify encoded representations of these chest radiographs with an accuracy comparable to or exceeding the performance of the published baseline network, which utilizes various state of the art network architectures as well as higher resolution images.

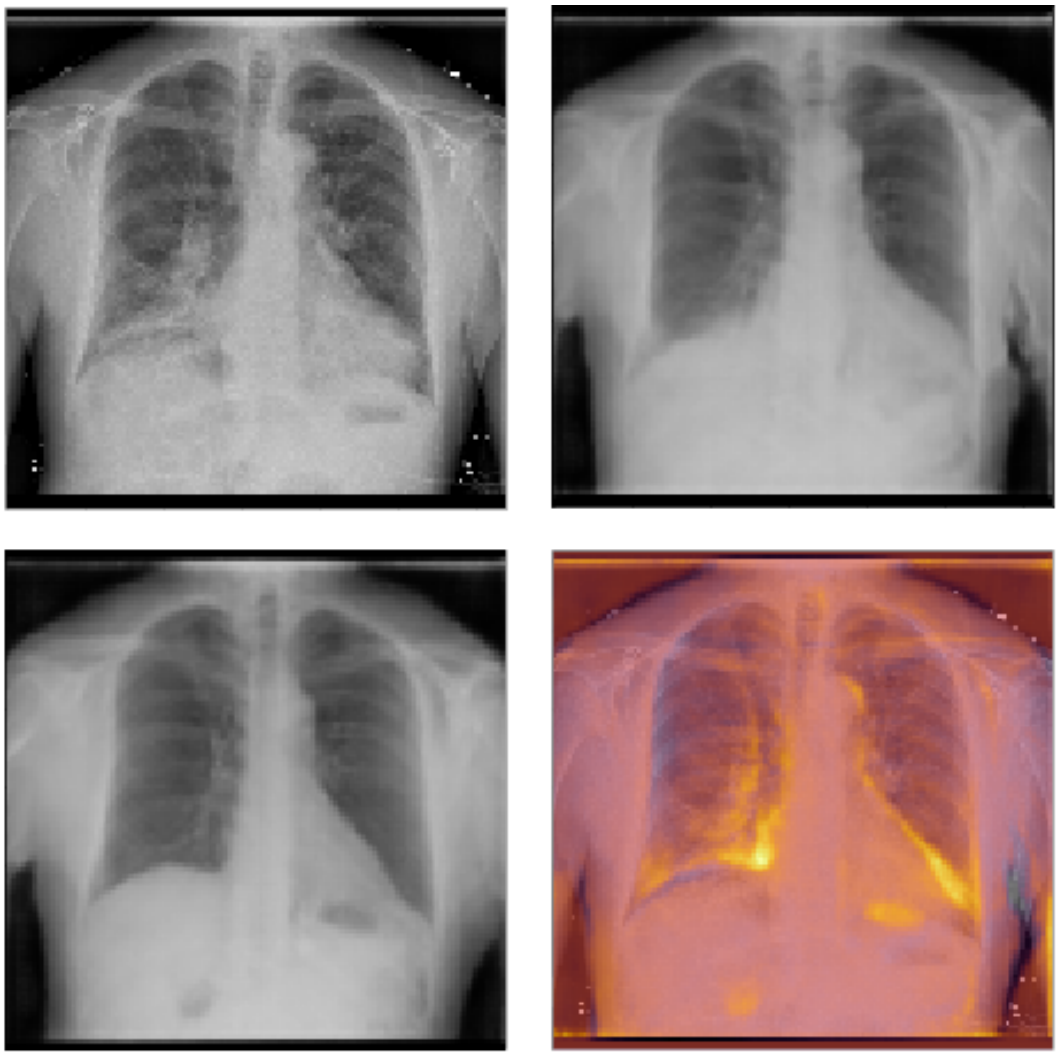

Figure 4: From top left to bottom right: Original image, reconstructed image, image visualized without heart failure, superimposed visual rationale on original image

	Ours	(Wang, 2017)
Atelectasis	**0.7546**	0.7069
Cardiomegaly	**0.8589**	0.8141
Effusion	**0.8243**	0.7362
Infiltration	**0.6945**	0.6128
Mass	**0.6958**	0.5644
Nodule	0.6247	0.**7164**
Pneumonia	**0.7346**	0.6333
Pneumothorax	**0.8164**	0.7891

Table 1: Comparison AUC results for ChestX-ray8 dataset

We apply our method to the MNIST dataset and demonstrate class switching between digits 9 and 4. Figure 7. demonstrates the visual rationales for why each digit has been classified as a 9 rather than a 4, as well as the transformed versions of each digit. As expected, the top horizontal line in the digit 9 is removed to make each digit appear as a 4. Interestingly, the algorithm failed to convert several digits into a 4 and instead converts them into other digits which are presumably more similar

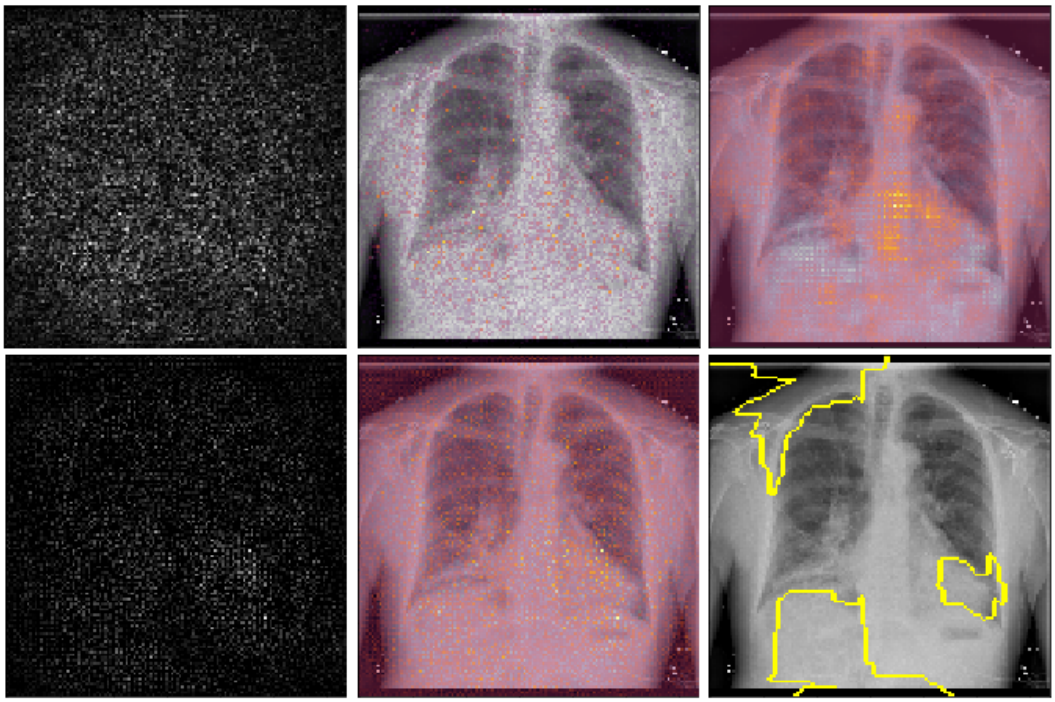

Figure 5: Top left to bottom right: Saliency map, saliency map overlaid on original image, heatmap generated via occlusion sensitivity method, Integrated gradients, integrated gradients overlaid on original image, LIME output

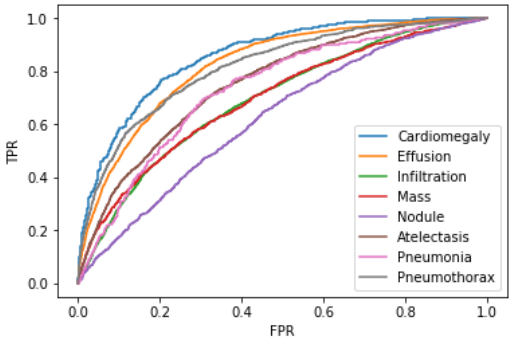

Figure 6: ROC curves for Chest X-Ray8 dataset

to that instance, despite the addition of the weighted term encouraging the latent representation to prefer the target class. This behaviour is not noted in our chest radiograph dataset as we are able to convert every image from the predicted class to the converse.

We compare this with the occlusion sensitivity and saliency map method demonstrated in Fig. 8.

Lastly, we contaminate our heart failure classifier as described in the methods section and compare visual rationales generated by the contaminated classifier with those generated previously. Fig 9. demonstrates images where both classifiers predict the presence of heart failure.

The rationales from the contaminated classifier focus on small unique aspects of the image and largely do not correspond to our notion of what makes a chest radiograph more likely to represent heart failure, namely enlarged hearts and congested lung fields.

Figure 7: From left to right: original images with visual rationale overlaid, transformed digits

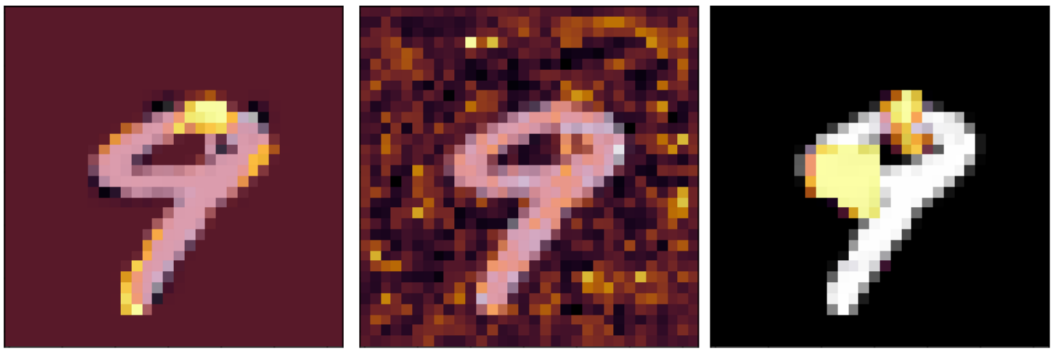

Figure 8: From left to right: visual rationale generated by our method, saliency map, occlusion sensitivity

4 DISCUSSION

We show in this work that using the generator of a GAN as the decoder of an autoencoder is viable and produces high quality autoencoders. The constraints of adversarial training force the generator to produce realistic radiographs for a given latent space, in this case a 100-dimensional space normally distributed around 0 with a standard deviation of 1.

This method bears resemblance to previous work done on inverting GANS done by Creswell & Bharath (2016), although we are not as concerned with recovering the exact latent representation but rather the ability to recreate images from our dataset. It is suggested in previous work in Kumar et al. (2017) that directly training a encoder to reverse the mapping learnt by the generator in a decoupled fashion does not yield good results as the encoder never sees any real images during training. By training upon the loss between the real input and generated output images we overcome this.

We further establish the utility of this encoder by using encoded latent representations to predict outcomes on unseen datasets, including one not from our institution. We achieve this without retraining our encoder on these unseen datasets, suggesting that the encoder has learnt useful features about chest radiographs in general.

Our primary contribution in this paper however is not the inversion of the generator but rather the ability to generate useful visual rationales. For each prediction of the model we generate a corre-

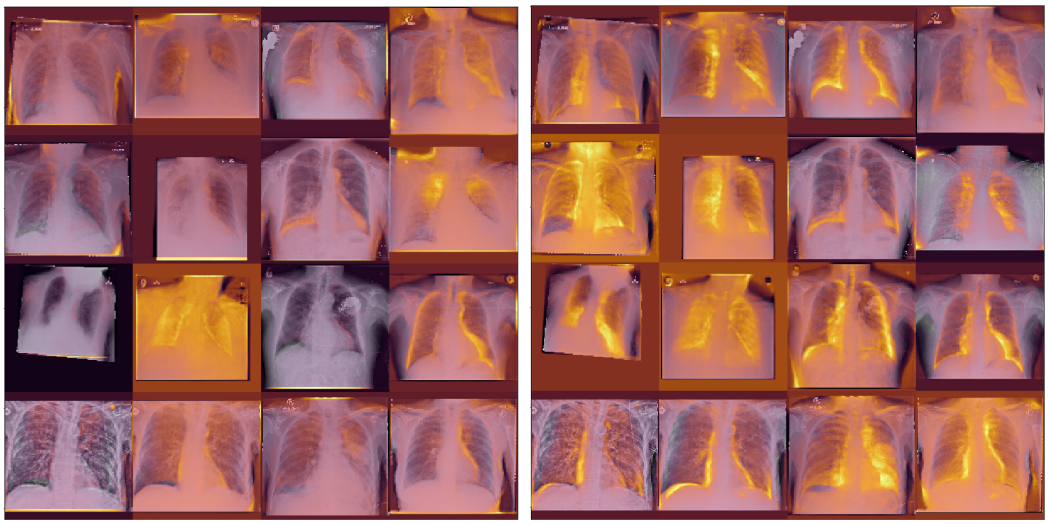

Figure 9: Left: Rationales from contaminated classifier. Right: rationales from normally trained classifier

sponding visual rationale with a target class different to the original prediction. We display some examples of the rationales this method produces and inspect these manually to check if these are similar to our understanding of how to interpret these images. The ability to autoencode inputs is essential to our rationale generation although we have not explored in-depth in this paper the effect of different autoencoding algorithms (for instance variational autoencoders) upon the quality of the generated rationales.

For chest radiographs, common signs of heart failure are an enlarged heart or congested lung fields, which appear as increased opacities in the parts of the image corresponding to the lungs. The rationales generated by the normally trained classifier in Fig 9 appear to be consistent with features described in the medical literature while the contaminated classifier is unable to generate these rationales.

We also demonstrate the generation of rationales with the MNIST dataset where the digit 9 is transformed into 4 while retaining the appearance of the original digit. We can see that the transformation generally removes the upper horizontal line of the 9 to convert this into a 4. Interestingly, some digits are not successfully converted. Even with different permutations of delta and gamma weights in Algorithm 2 some digits remain resistant to conversion. We hypothesize that this may be due to the relative difficulty of the chest radiograph dataset compared to MNIST - leading to the extreme confidence of the MNIST model that some digits are not the target class. This may cause vanishingly small gradients in the target class prediction, preventing gradient descent from achieving the target class.

We compare the visual rationale generated by our method to various other methods including integrated gradients, saliency maps, occlusion sensitivity as well as LIME in Fig. 5.

All of these methods share similarities in that they attempt to perturb the original image to examine the impact of changes in the image on the final prediction, thereby identifying the most salient elements. In the saliency map approach, each individual pixel is perturbed, while in the occlusion sensitivity method, squares of the image are perturbed. LIME changes individual superpixels in an image by changing all the pixels in a given superpixel to the average value. This approach fails on images where the superpixel classification is too coarse, or where the classification is not dependent on high resolution details within the superpixel. To paraphrase Sundararajan et al. (2017), attribution or explanation for humans relies upon counterfactual intuition — or altering the image to remove the cause of the predicted outcome. Model agnostic methods such as gradient based methods, while fulfilling the sensitivity and implementation invariance axioms, do not acknowledge the natural structure of the inputs. For instance, this often leads to noisy pixel-wise attribution as seen in Fig. 5. This does not fit well with our human intuition as for many images, large continuous

objects dominate our perception and we often do not expect attributions to differ drastically between neighbouring pixels.

Fundamentally these other approaches suffer from their inability to perturb the image in a realistic fashion, whereas our approach perturbs the image's latent representation, enabling each perturbed image to look realistic as enforced by the GAN's constraints.

Under the manifold hypothesis, natural images lie on a low dimensional manifold embedded in pixel space. Our learned latent space serves as a approximate but useful coordinate system for the manifold of natural images. More specifically the image (pardon the pun) of the generator $G[\mathbb{R}^d]$ is approximately the set of 'natural images' (in this case radiographs) and small displacements in latent space around a point z closely map into the tangent space of natural images around $G(z)$. Performing optimization in latent space is implicitly constraining the solutions to lie on the manifold of natural images, which is why our output images remain realistic while being modified under almost the same objective used for adversarial image generation.

Hence, our method differs from these previously described methods as it generates high resolution rationales by switching the predicted class of an input image while observing the constraints of the input structure. This can be targeted at particular classes, enabling us answer the question posed to our trained model — 'Why does this image represent Class A rather than Class B?'

There are obvious limitations in this paper in that we do not have a rigorous definition of what interpretability entails, as pointed out by Sundararajan et al. (2017). An intuitive understanding of the meaning of interpretability can be obtained from its colloquial usage — as when a teacher attempts to teach by example, an interpretation or explanation for each image helps the student to learn faster and generalize broadly without needing specific examples.

Future work could focus on the measurement of interpretability by judging how much data a second model requires when learning from the predictions and interpretations provided by another pre-trained model. Maximizing the interpretability of a model may be related to the ability of models to transfer information between each other, facilitating learning without resorting to the use of large scale datasets.

Other technical limitations include the difficulty of training a GAN capable of generating realistic images larger than 128 by 128 pixels. This limits the performance of subsequent classifiers in identifying small features. This can be seen in the poor performance of our model in detecting nodules, a relatively small feature, compared to the baseline implementation in the NIH dataset.

In conclusion, we describe a method of semi-supervised learning and apply this to chest radiographs, using local data as well as recent datasets. We show that this method can be leveraged to generate visual rationales and demonstrate these qualitatively on chest radiographs as well as the well known MNIST set.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. January 2017.

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. July 2017.

Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. November 2016.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. June 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. March 2017.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-Normalizing neural networks. June 2017.

Abhishek Kumar, Prasanna Sattigeri, and P Thomas Fletcher. Improved semi-supervised learning with GANs using manifold invariances. May 2017.

A. Lokuge, L. Lam, P. Cameron, H. Krum, de Villiers Smit, A. Bystrzycki, M. T. Naughton, D. Eccleston, G. Flannery, J. Federman, and et al. B-type natriuretic peptide testing and the accuracy of heart failure diagnosis in the emergency department. *Circulation: Heart Failure*, 3(1): 104–110, Nov 2009. doi: 10.1161/circheartfailure.109.869438. URL http://dx.doi.org/10.1161/CIRCHEARTFAILURE.109.869438.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. February 2016.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. December 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. URL http://arxiv.org/abs/1703.01365.

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. May 2017.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. November 2013.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. December 2015.