Modifying LIME for Medical Images

Jiong Wei Lua JIONGWEI@MIT.EDU

Alexandru Socolov SOCOLOV@MIT.EDU

Andras Szep SZEP@MIT.EDU

1. Introduction

With massive breakthroughs of machine learning in medical settings in the past decade, there has been a growing attention to interpretability of the ever more complex models. Since many of them make decisions that may affect human well-being and are used in clinical contexts with humans in the loop, the importance of trust in such predictions is hard to overstate.

Currently available interpretability methods can be broadly classified into two categories: (i) saliency-based methods typically use the gradient information to infer salient features, while (ii) perturbation-based methods seek to query the model within a local neighborhood of the prediction of interest.

Amongst the latter, Local Interpretable Model Agnostic Explanations (LIME) method seeks to explain examples of interest by locally approximating the true, complex machine learning model with a more interpretable function. This is achieved by first generating simulated data points in the neighborhood of the original example, weighting these simulated data points based on their proximity to the original example, and then fitting a simple model locally to approximate the true model. For example, in Figure 1 below there is a highly complex non-linear decision boundary. The example which is to be explained is the bolded red cross, while neighbours are the non-bold crosses and blue circles, with their size corresponding to their proximity. The dotted line is the approximating linear decision boundary.

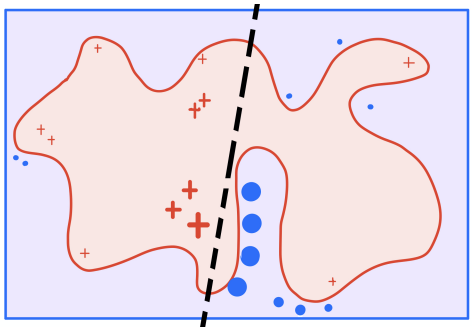

Figure 1: LIME In General

When applying LIME to images, the method segments the input into patches or "superpixels." The "superpixels" are then turned off and on to generate neighbours of the original image. Turning the "superpixels" off/on typically means substituting the segment with a default color like grey. This framework essentially asks the question - which patches or "superpixels" of the image, if removed, would most significantly affect the model output?

While this approach of generating neighbours is fairly intuitive, it has an important technical limitation. Namely, while thinking of superpixel contributions might be appropriate for image classification tasks in

general, visual explainability in the clinical context is significantly more nuanced. Consider the task of diagnosing Pneumothorax based on an X-ray. The air outside the lung in a Pneumothorax will surface as an increase in lucency (darkness) because of the lower air density relative to the lung. This is a non-trivial task even for radiologists as the lucency of the image can be affected by a wide-variety of factors, such as existing conditions, the size of the Pneumothorax and the positioning of the body. From our consultation with our mentor, Dr. Goehler, how clinicians typically diagnose Pneumothorax would involve being able to trace out the region where there is air collection. In our view, the ability to identify the region where there is a difference in lucency is significantly more sophisticated than asking, which random patch of the X-ray, if removed, would significantly change the model's predictions.

Thus, to address this limitation of LIME, we generated clinically-meaningful neighbours for medical images to improve the applicability of LIME. Specifically, we explored multiple options for finding clinically meaningful neighbours ranging from 'computing a low-dimensional embedding representation of each image and finding similar images in the latent space' to 'leveraging auxiliary information present in the medical reports accompanying the X-rays to find neighbours using NLP approaches'. We also developed and compared different LIME architectures trained on our clinically meaningful neighbours. We show that our approach generates neighbors that are significantly more clinically similar than the baseline of randomly selecting neighbors and that our LIME architecture addresses the original limitation of LIME by introducing clinical variability in neighbors instead of simply greying out superpixels.

2. Related Work

We conducted a review of the literature relating to modifying LIME for clinical setting. Zafar and Khan (2019) and Shankaranarayana and Runje (2019) both seek to address the problems of instability of LIME due to the random nature of introducing perturbations in LIME. Zafar and Khan (2019) propose DLIME, a deterministic version of LIME where neighbours are defined not by removing segments of the image, but by applying clustering techniques, and finding nearest-neighbours within clusters to identify data points which are within the local neighbourhood of the example of interest. Shankaranarayana and Runje (2019) propose ALIME, which aims to improve the local stability of LIME by incorporating an autoencoder as a weighting function. Concretely, this means that weighting of these neighbours is not based on the distance in the original feature space, but in the latent space of the autoencoder. Across both papers, both set of authors apply their methods to relatively low-dimensional healthcare datasets containing the physiological records and observations of patients. It is not clear that pure DLIME alone will extend well to medical imaging datasets which are typically high dimensional; applying clustering and nearest neighbour methods directly to the original space may not yield good quality neighbours due to the curse of dimensionality. Leveraging dimensionality reduction techniques such as autoencoder architectures as per ALIME could potentially deal with this challenge but Shankaranarayana and Runje (2019) do not address this in their work.

3. Methods, Data, and Experiment Setup

3.1. Overview of Approach

The overall goal of generating a more clinically meaningful explanation of a medical image can be broken down into two sub-problems.

First, for a given image, we need to find clinically meaningful neighbour images that are similar to the existing image. What exactly do we mean by *clinically meaningful neighbour images?* On one hand, the

neighbour images should be similar to the original image in the image space, and by this we mean visual similarity such as orientation, size, posture. This is to satisfy the locality criterion behind LIME. This similarity could also potentially be important in ensuring that non-medical features do not end up confounding the surrogate model. On the other hand, the neighbours should ideally exhibit some heterogeneity in their underlying diseases or clinical conditions. The diversity in underlying disease will allow for a model like LIME to detect and highlight what areas of the image differentiate the target image's disease from other medical conditions. If all the neighbours had the exact clinical conditions, no local surrogate model would be able to approximate the original black box model's decision boundary.

Once we can generate a set of clinically meaningful neighbors for a given target image, the second sub-problem is the question of how to extend the local surrogate methodology underlying LIME to these neighbors and how to best visualize areas of interest on the target image.

3.2. Data

We decided to focus on X-ray images primarily because our clinical advisor, Dr. Goehler, has expertise in Radiology. Thus, working with X-rays allowed us to rely on Dr. Goehlrer's insights throughout our project. We shortlisted the MIMIC-CXR and the ChestX-ray14 datasets to work with.

We selected the ChestX-ray14 dataset primarily because there was a pre-trained model publicly available (Wang et al. (2017)) for it that was successfully applied to predicting 14 different medical conditions. This significantly reduced our workload and helped us focus on building the surrogate model. We also found that the X-rays that we downloaded were mostly taken from the same anterior-posterior views. These identical views reduced the difficulty of identifying images that are similar.

The MIMIC-CXR dataset has the added benefit that the X-ray images are not only labeled but are also accompanied by medical reports written by clinicians. The information contained in these reports can be extracted through NLP techniques and leveraged to narrow the feasible set of nearest neighbours for a given image. For instance, if we know from a medical report that an X-ray was taken of a 34-year-old male patient who had pig-tail chest tubes inserted earlier, then good candidates for closest-neighbor images would be ones that also come from males in their 30s with pig-tail chest tubes.

3.3. Identifying Clinically Meaningful Neighbors with VAE & BERT

To address our first sub-problem of finding clinically meaningful neighbors, we attempted the following two approaches:

- Approach (1): Using variational autoencoders (VAEs) to compute a low-dimensional embedding representation of each image, finding similar images in the latent space to generate neighbours.
- Approach (2): Leveraging auxiliary information such as known diagnoses, physiological information and demographic information to narrow the feasible set of nearest neighbours for a given image.

3.3.1. Identifying Clinically Similar Images through VAEs

For approach (1), we turned to VAEs as the image feature space is extremely high dimensional; thus, trying to find similar images based on distance measures in the image space could be afflicted by the curse of dimensionality. Motivated by Shankaranarayana and Runje (2019), we trained VAE models with a 40- to 80-dimensional latent-dimensions to compute low-dimensional embeddings of the images, and then found similar X-rays by computing the similarity of images in the latent embedding space. Through

experimentation, we found that a 60 dimensional latent space worked best as it provided the ideal trade off between training time and compression. While our original images had a resolution of $3 \times 1024 \times 1024$, all RGB channels had the same values (black & white picture), and therefore we judged that using all three channels would create too many parameters to train for the neural network. Therefore, we used the torchvision transformer API to convert our images back to gray-scale and downscale our image to $1 \times 224 \times 224$. Our VAE was trained through mini-batch stochastic gradient descent for about 30 epochs.

3.3.2. Identifying Clinically Similar Examples through Medical Reports

The second approach for selecting clinically relevant neighbors was to rely on auxiliary report information present in the MIMIC-CXR dataset to find the nearest neighbors to a given X-ray study. Thus, our goal here was to utilize natural language processing (NLP) techniques to determine the similarity between the medical reports present in the MIMIC-CXR dataset that accompany each X-ray image.

To measure semantic similarity between two documents, there exists a wide array of approaches. Most of these approaches rely on creating some representation or embedding of a document (e.g., Bag of Words, Term Frequency - Inverse Document Frequency, Word2Vec, etc.) and then applying some similarity measure that can quantify the difference between the embeddings (e.g., K-means, Cosine Similarity, etc.).

A state-of-the-art approach of creating an embedding of a text is BERT, which relies on a bidirectional transformer architecture and on context on both sides of tokens during training to increase performance among other advantages. Thus, we chose to utilize a pre-trained BERT model to create embeddings of our medical reports. Since we were working with clinical data, we chose to implement a BERT model that was also trained on clinical data (the MED-STS dataset) from Wolf et al. (2019).

For both approaches, we compute the pairwise distance between the embedding of the image of interest and all other images in the database, and identify the K images which are most similar to the image to be explained. Based on an assessment of these two approaches, we identified approach (1) as the superior method of finding clinically meaningful neighbors that LIME can be applied to (see Results section).

3.4. Disease Classification

LIME approximates the prediction of a complex machine learning model locally with a more interpretable surrogate function. In our case, this more complex machine learning model is a deep convolutional neural network: the CheXNet model. Note that while the CheXNet model has been trained on the ChestX-ray14 dataset (Wang et al. (2017)), it has been successfully applied to the MIMIC data as well by Professor Szolovits and his team (Boag et al.).

To implement the CheXNet model, we built a Python algorithm that builds the same architecture as that of CheXNet model and then passes-in the pre-trained weights. Moreover, to make the CheXNet model work with our MIMIC-CXR data, we had to format the input X-ray images to have the same dimensions and number of channels as the set that CheXNet was trained on. This proved somewhat more challenging than initially anticipated as there were certain image pre-processing steps we initially missed but then corrected for after some careful debugging.

3.5. Our Proposed Modified LIME Architectures

Assuming that the above method allows us to identify images that are clinically meaningful neighbors (see Section 3.3), our second sub-problem was how to fit a LIME model to these neighbors and then best

visualize areas of interest on the input image. While in a traditional LIME the neighbors are all the same image with different super-pixels turned on/off, in our case the neighbors are similar in their latent space but still different images. Thus, it was unclear whether super-pixels defined on the image of interest would work well for the neighboring images or whether we could utilize the neighbors to define a better set of super-pixels. To explore these questions, we built three different LIME architectures.

3.5.1. Architecture 1: K*LIME

For our first architecture, aptly named K*LIME, we trained a different traditional LIME model for each of the K neighbors and output the image of interest explained with LIME alongside the other K X-rays' explanation. The idea here was to show a clinician how the areas of interest on the target image compare to those on X-rays that are similar but have different underlying diseases. Ideally, this approach would enable a clinical to understand why the model chose to classify the target image as a certain disease over other likely disease candidates.

3.5.2. Architecture 2: Σ-LIME

In the second approach, we first run the classical LIME individually on each of the K neighbours (Architecture 1) and store the explaining masks created. An important caveat here is making sure each neighbor LIME was interpreting the same diagnose as the input image. We then overlay the masks on top of each other to identify the regions of the image that were considered important across many neighbours and plot them on top of the original input image. We considered overlaying the explanation appropriate due to good performance of VAE in finding visually similar neighbours, e.g. similar position, brightness, body shapes.

Computationally, masks are just 1024×1024 matrices with entries $\in \{-1, 0, 1\}$, so summing them element-wise resulted in integer values in $[-K, K]$ range. The more positive a region is, the more it is associated with a positive diagnose, the more negative - the more likely is the lack thereof, according to the neural network that is being explained. Due to the summation operation, we refer to this architecture as Σ-LIME.

3.5.3. Architecture 3: KP-LIME

Figure 2 diagram illustrates our architecture combining the pre-trained VAE and modified LIME (KP-LIME) to generate explanations. Similarly to Architectures 1 and 2, we again find neighbors of an image of interest by projecting every image into the latent space using our pre-trained VAE and computing the pairwise distance between this latent representation of the target image and the existing database of latent representations to identify the K nearest neighbours.

Then, we pass the image to be explained and its set of K existing nearest neighbours into our modified LIME. We still apply a segmentation algorithm to the target image to partition it into different segments. However, unlike traditional LIME, we do not perform random masking of segments solely on the target image. Additionally, using the segments obtained from the target image, we also perform random masking on the K nearest neighbours. We do this for each neighbour P times, and the segments which are masked are determined randomly, effectively creating P different masked versions for each original neighbour.

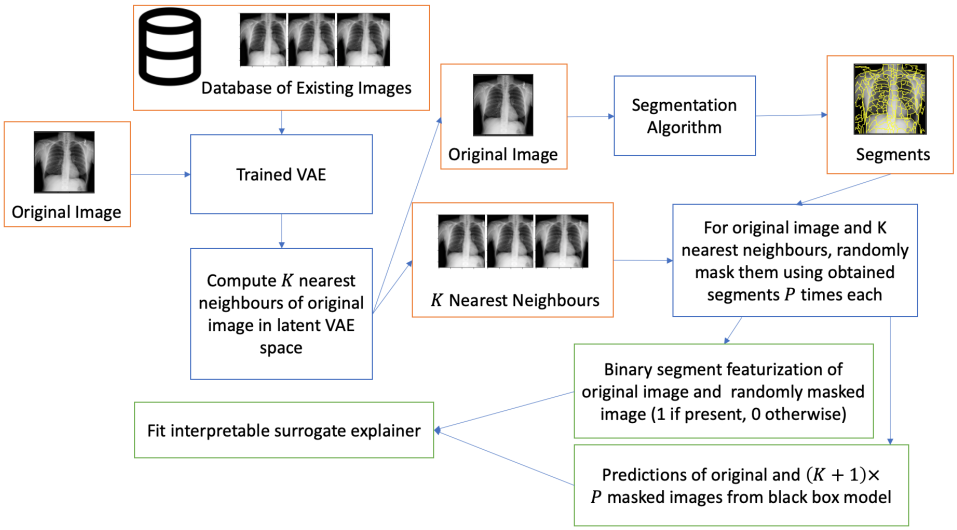

Figure 2: Architecture of KP-LIME

Our principal motivation for doing this is that the data generated from random masking exhibits more clinically meaningful variability, while preserving locality (because we use nearest neighbours). In traditional LIME, for a given segment in a masked image, it can be grey (switched off) or the corresponding segment of the target image. In our KP-LIME approach, the segment in the masked image can now potentially be the corresponding segment of any of the K neighbours - a genuine portion of an X-ray rather than a patch of grey. This helps us learn the black box model's decision boundaries better.

To train our surrogate classifier, the unmasked original image, as well as $(K+1) \times P$ randomly masked images, are individually featurized into a d-dimensional vector (where d is the number of segments). The i^{th} element of d is 1 if the i^{th} segment of the image is masked, and 0 if it is not. This forms our predictors for our local surrogate classifier. To obtain the labels for our local surrogate classifier, the predictions are obtained for all the $(K+1) \times P$ randomly masked images by passing them (in their original image space) through the original Disease Classification model outlined in Section 3.4. We then fit a local surrogate model on these features and labels to be our explainer as per the usual LIME.

4. Results

4.1. Clinically Meaningful Neighbors

To find clinically meaningful neighbors, we tested two approaches: (1) using variational autoencoders (VAEs), (2) using BERT on auxiliary medical texts - see Section 3.3.

4.1.1. Performance of the VAE Approach

For approach (1), we trained our VAE architecture on the CheXRay dataset about 10.5k training images for 30 epochs. We did not end up training a VAE for the MIMIC-CXR dataset. This was because in MIMIC-CXR, the X-ray images were taken from many different views (e.g. anterior-posterior, posterior-anterior, lateral) and there was no easy way of distinguishing the positions without reading in the DICOM images first. Given time and computing constraints, we were not confident that we would be able train a sufficiently complex VAE that would be able to learn high quality latent representations for all these views. Hence we focused on training the VAE for the ChestX-ray14 dataset.

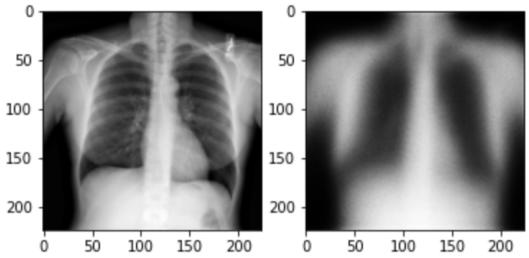

Figure 3: Original Image (left) and VAE reconstructed output (right)

Figure 3 shows a sample example of our original image and a reconstructed image using the VAE. Although the reconstruction seems to not be very good, our goal of using VAE was not to reconstruct images with fine resolution. Instead, it was to have a low-dimensional latent space in which nearest neighbours are similar in the image space, and we expect this to translate to similarity in visual features such as orientation, body shape, high level physiological features.

To evaluate this, we projected all ChestX-ray14 images into the latent dimensions of our VAE, and computed their pairwise Euclidean distance in the latent space. For each image, we identified their $K = 5$ nearest neighbours. We then qualitatively examined the visual similarity of the each image and their top 5 neighbours. We found broad qualitative evidence the latent space of the VAE was indeed able to encode visual and physiological similarities, and we provide two examples and their nearest neighbours in Figure 4.

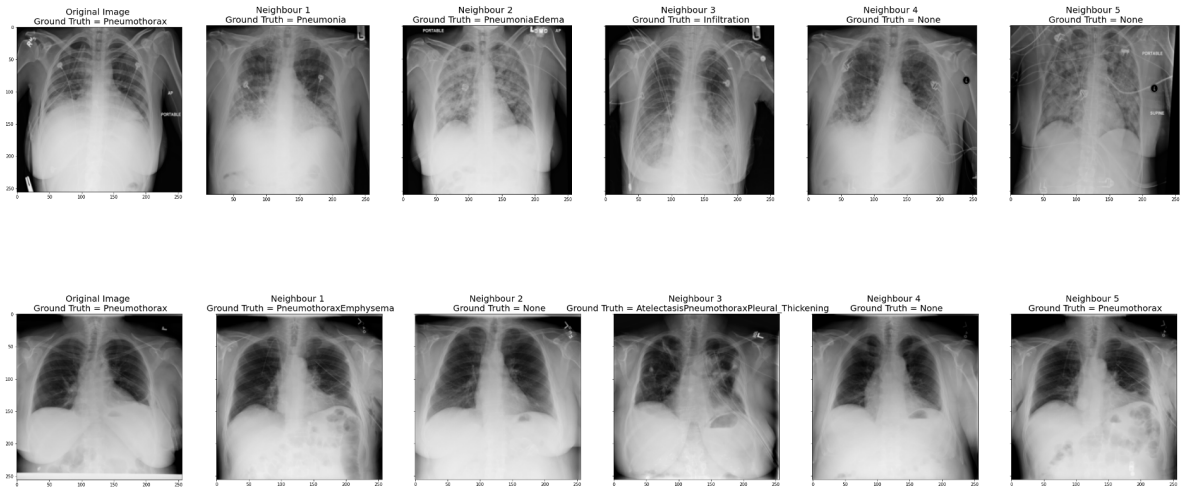

Figure 4: Examples of 2 Images and their 5 Nearest Neighbours

However, while we looked through several hundred images, the above qualitative assessment does not scale too well and is inevitably affected by an element of subjectivity. It also does not examine whether VAE can capture underlying clinical conditions. We then took a more quantitative approach to measure this. We compared an input's ground truth of clinical conditions with that of its $K = 5$ nearest neighbors and found that the set of diseases overlap an average of of 40.57% of the time. Moreover, we note that if we were to select the K neighbors randomly instead of using our VAE approach, then there would be significantly ($p < 0.01$) less overlap between the target image's set of diseases and its neighbors. The frequency of

how many neighbors of a target image have overlapping ground truth conditions for neighbors selected randomly vs. with our trained VAE is depicted on Figure 5.

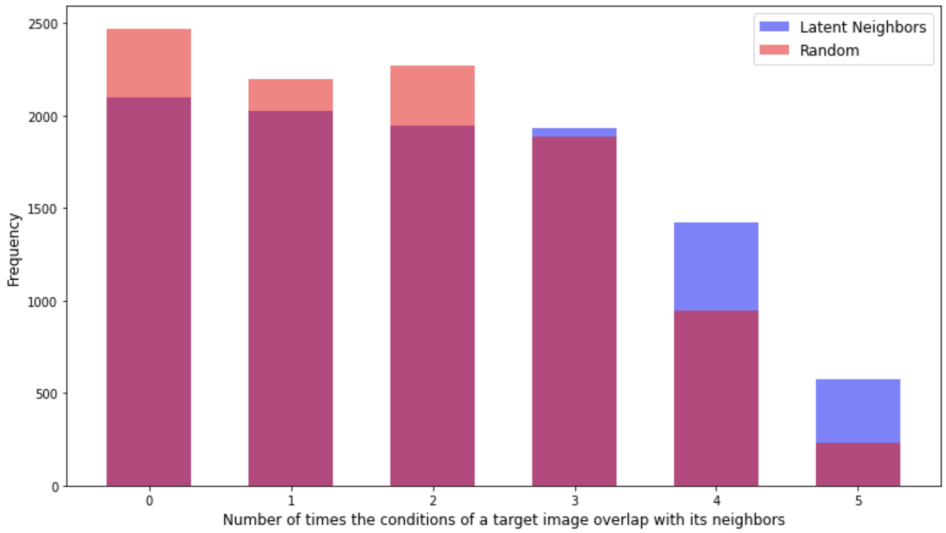

Figure 5: Frequency of disease overlap between a target image and its neighbors for neighbors selected with VAE (in blue) and randomly (in red) overlayed.

4.1.2. Performance of the BERT Approach

For our second approach, we used the clinical BERT model to find the most semantically similar reports to a given medical report. Table 1 shows an example result where we depict the $K = 3$ semantically closest reports to a given original medical report out of a sample of 24 MIMIC-CXR reports. While qualitatively the reports in Table 1 do seem similar, unfortunately, this approach does not give any guarantees in terms of similarity in the images corresponding to neighboring reports. In fact, as the X-ray images in the MIMIC-CXR dataset were taken from many different views, we found that corresponding images were in some cases taken from different angles altogether.

Original Report	The cardiac, mediastinal and hilar contours are normal. Pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is present. Multiple clips are again seen projecting over the left breast. Remote left-sided rib fractures are also re-demonstrated.
Nearest Semantic Neighbors	PA and lateral views of the chest provided. Lung volumes are somewhat low. There is no focal consolidation, effusion, or pneumothorax. Imaged osseousstructures are intact. No free air below the right hemidiaphragm is seen.
	Lung volumes are low. The heart size is normal. The mediastinal and hilarcontours are unremarkable. New nodular opacities are clustered within the left upper lobe, and to a lesser extent, within the right upper lobe. There is no pneumothorax or left-sided pleural effusion. Pulmonary vascularity is within normal limits. Postsurgical changes are noted in the right chest with partial resection of the right 6th rib, lateral right pleural thickening and chronic blunting of the costophrenic sulcus.
	PA and lateral views of the chest demonstrate low lung volumes. Tiny bilateral pleural effusions are new since ____. No signs of pneumonia or pulmonary vascular congestion. Heart is top normal in size though this is stable. Aorta is markedly tortuous, unchanged. Aortic arch calcifications are seen. There is no pneumothorax. No focal consolidation. Partially imaged upper abdomen is unremarkable.

Table 1: An original report and its nearest k = 3 semantic neighbors

4.1.3. How we identified neighbours for our LIME

We initially hoped to generate neighbours by taking into consideration both their similarity in the image space as well as incorporating auxiliary information in the medical reports. However, the medical reports were only available for images in the MIMIC-CXR dataset, and we did not manage to successfully train a VAE for this dataset for reason discussed above. Therefore, we only relied on VAE latent representations for selecting neighbors and focused on applying our methods to the ChestX-ray14 dataset.

4.2. Applying LIME using different architectures

Having defined a set of neighbours, we apply the three architectures to explain a given input. Our output for each of the modified LIME methods are shown in the following figures: Figure 6 for K*LIME, Figure 7 for Σ-LIME, and Figure 8 for KP-LIME. For all three methods, we set $K = 5$, and for KP-LIME, we generated $P = 100$ random masks per neighbour.

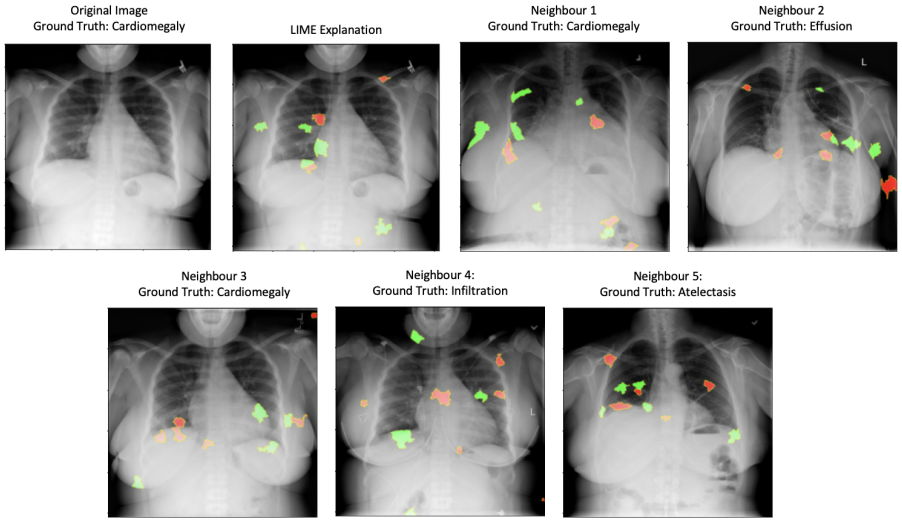

Figure 6: Sample K*LIME interpretation on a given input with 5 VAE neighbors

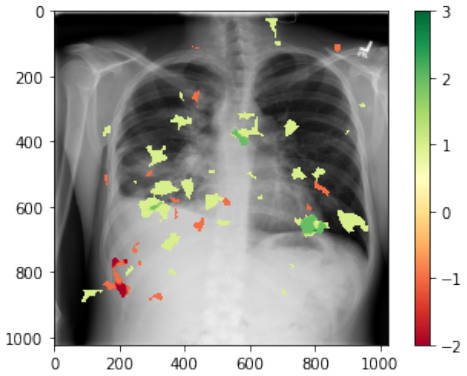

Figure 7: Sample Σ-LIME interpretation on a given input with 5 VAE neighbors

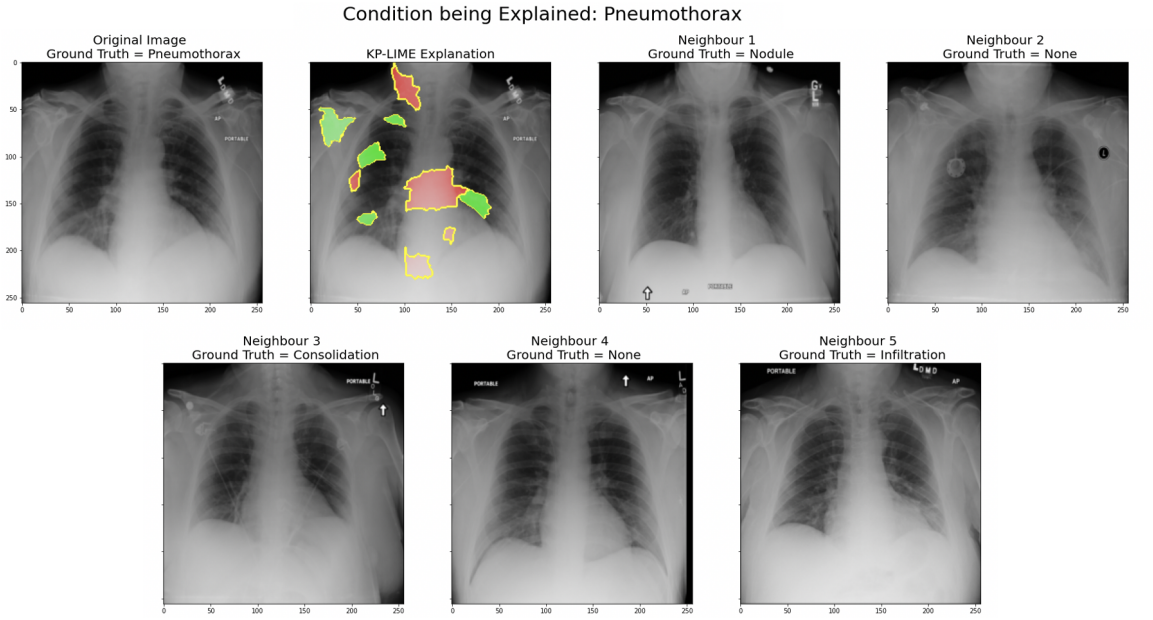

Figure 8: Sample KP-LIME Pneumothorax interpretation on a given input with 5 VAE neighbors

5. Discussion

In conclusion, we found that our VAE latent space representation seemed to be able to capture the image features and similarities, such as body position, image brightness as well as some body shapes. Computing the common class labels between nearest neighbours in the VAE space also showed that they have more in common in terms of underlying clinical conditions than the randomly selected neighbors. Upon the visual inspection, the NLP similar neighbors performed worse than VAE. Moreover, NLP neighbors were only available for MIMIC-CXR dataset that has both frontal and lateral X-Rays with no systematic way of distinguishing between the two.

Turning to LIME architectures, we believe the KP-LIME approaches resolves one of the original limitation of LIME, in which the data generating process for training the surrogate model only involved switching on and off segments of the original image. By integrating neighbour images, segments in the generated data can also be set to the corresponding segments of neighbour images, which we see as clinical variability. To the extent that the neighbours in the latent space are also similar in the image space, we also respect the locality requirement that underpins the idea of fitting a local surrogate model. Nonetheless, we have to acknowledge that we do not overcome one of the limitations of LIME - that we need to introduce some randomness in how segments are selected for masking, and the particular choice of segmentation algorithm can significantly affect the interpretation obtained since the segments will vary correspondingly.

Unfortunately, due to the limited time and challenges we had, we were not able to come up with a systematic way to evaluate our various LIME architectures. Given the luxury of time, we would definitely have tried to come up with an evaluation strategy to evaluate the quality of the explanations from our modified architectures against that of default LIME more systematically.

Future Research

One of the directions that we could look at is to examine how auxiliary report information, in conjunction with the VAE, can help identify higher quality neighbours. Additionally, VAE has been trained on roughly 10k images, which is a small portion of what is available in ChestX-ray14, Thus, scaling the VAE training could potentially provide a more refined neighborhood understanding and/or higher quality neighbours.

6. Acknowledgements & Member Contributions

Acknowledgements

We want to thank Dr. Alex Goehler for guiding us through the medical interpretations of the X-rays as well as making sure our contribution can be valuable for the experts. We also thank Matthew McDermott, Monica Agrawal, David Sontag, and Peter Szolovits for helping us scope the problem and iteratively reviewing our work.

Member Contributions

Finally, while we worked closely together throughout the project, the contribution of our team members has been focused in the following areas:

- Jiong Wei Lua: pre-processing images for CheXNet, developing VAE and latent space similarity, developing KP-LIME
- Alexandru Socolov: building pre-trained CheXNet, applying classical LIME on ChestX-ray14, and MIMIC data and developing Σ-LIME
- Andras Szep: exploring NLP report similarity with BERT, training and evaluating VAEs, and developing K*LIME

References

William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for chest x-ray report generation.

Sharath M Shankaranarayana and Davor Runje. Alime: Autoencoder based approach for local interpretability. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 454–463. Springer, 2019.

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. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi: 10.1109/cvpr.2017.369. URL http://dx.doi.org/10.1109/CVPR.2017.369.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Muhammad Rehman Zafar and Naimul Mefraz Khan. DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *CoRR*, abs/1906.10263, 2019. URL http://arxiv.org/abs/1906.10263.